Darío Hurtado, Silvia Jaramillo, Viviana Torres y Andrea Rodríguez

Estadística
Aplicaciones en SPSS

UIDE
Powered by
Arizona State University

Estadística

Aplicaciones en SPSS

Darío Hurtado, Silvia Jaramillo, Viviana Torres y Andrea Rodríguez
Estadística. Aplicaciones en SPSS

Quito: Universidad Internacional del Ecuador, 2025
1.ª edición, 140 pp. Vol: 15 x 21 cm

CDU: 311 + 681.3
ISBN 978-9942-682-06-2
DOI: https://doi.org/10.33890/estadistica.aplicada.spss

1. Ciencia estadística
2. Computación. *Software*

Como citar: Hurtado Cuenca, Darío, Jaramillo Luzuriaga, Silvia,
Torres Díaz, Viviana y Rodríguez Aranda, Andrea. (2025). Estadística. Aplicaciones en SPSS. Universidad Internacional del Ecuador.
https://doi.org/10.33890/estadistica.aplicada.sps

Estadísticacales. Aplicaciones en SPSS

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

Directora 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*).

Estadística

Aplicaciones en SPSS

Darío Hurtado Cuenca
Universidad Internacional del Ecuador

Silvia Jaramillo Luzuriaga
Universidad Internacional del Ecuador

Viviana Torres Díaz
Universidad Internacional del Ecuador

Andrea Rodríguez Aranda
Investigadora independiente

UIDE
Powered by
Arizona State University

Índice

Introducción

En un mundo cada vez más orientado hacia los datos, la estadística se ha convertido en una disciplina esencial para una amplia gama de campos profesionales y académicos. Desde la economía y las ciencias sociales hasta la medicina y la ingeniería, la capacidad de recolectar, analizar e interpretar datos es fundamental para la toma de decisiones informadas. Los datos no solo proporcionan un espejo para observar el pasado, sino que también son una brújula para guiar el futuro. Sin embargo, para aprovechar todo su potencial, es necesario comprender los principios estadísticos y saber cómo usarlos de manera efectiva. Este libro está diseñado precisamente con ese objetivo: proporcionar una guía completa y accesible para el aprendizaje de la estadística y su aplicación práctica mediante herramientas como el SPSS.

Este texto se organiza de manera sistemática, al empezar con las generalidades y abordar conceptos fundamentales que sientan las bases para el entendimiento del análisis de datos. La introducción a la estadística incluye definiciones clave, la clasificación de los diferentes tipos de variables y una explicación sobre las escalas de medición, elementos esenciales para cualquier análisis cuantitativo. Con este primer capítulo, los lectores desarrollarán una comprensión sólida de qué es la estadística y cómo se clasifica en diferentes tipos, ya sea descriptiva o inferencial, así como una diferenciación clara entre las variables cualitativas y cuantitativas. Además, se profundiza en las fuentes y tipos de datos, lo que proporciona un marco teórico sobre cómo se originan y cómo deben tratarse para su análisis.

A medida que avanzamos en el texto, el segundo capítulo introduce al lector en el uso de SPSS, un *software* ampliamente utilizado para el análisis de datos estadísticos. Aquí, se explica la estructura del programa y el manejo de datos dentro del mismo, un paso crucial para cualquier persona que desee realizar análisis estadísticos de manera eficiente. Los usuarios aprenderán a navegar por las diferentes ventanas del *software*, a utilizar las barras de menú y herramientas, y a interactuar con los cuadros de diálogo. También se abordan aspectos más prácticos como la creación y edición de archivos de trabajo, así como la definición de las propiedades de las variables dentro del entorno SPSS. Este capítulo es esencial para aquellos que, además de entender la teoría estadística, buscan adquirir competencias prácticas que les permitan aplicar sus conocimientos en escenarios reales.

El tercer capítulo lleva al lector a un nivel más profundo de análisis con la tabulación de encuestas y la importación de bases de datos. Este apartado es particularmente útil para quienes trabajan con grandes volúmenes de datos provenientes de investigaciones de campo o estudios de mercado. Se detallan los procesos de codificación de instrumentos de medición, la creación de variables, el ingreso de datos y la manera de guardar y exportar archivos. Además, se incluyen secciones sobre cómo importar bases de datos externas, lo que permite a los lectores integrar datos de diferentes fuentes para un análisis más completo. Este capítulo es un puente entre la recolección de datos y su análisis estadístico, subrayando la importancia de una correcta organización y codificación de la información para garantizar la validez de los resultados.

El cuarto capítulo se enfoca en la transformación de datos, un paso crucial para preparar la información antes del análisis estadístico. La transformación de datos incluye técnicas para manejar respuestas múltiples, identificar casos duplicados o atípicos, calcular nuevas variables y recodificar las existentes. Este capítulo es fundamental para quienes buscan asegurar la integridad de sus datos antes de proceder con el análisis estadístico. Se detallan procedimientos como la recodificación de variables, ya sea en las mismas o en distintas variables, así como la creación de variables auxiliares y la agrupación de datos. Además, se abordan métodos para reemplazar valores perdidos, un problema común en los conjuntos de datos que, si no se maneja correctamente, puede llevar a conclusiones erróneas.

El quinto capítulo introduce las medidas numéricas, que son herramientas clave en el análisis estadístico. Aquí, los lectores aprenderán sobre las medidas

de tendencia central, como la media, la mediana y la moda, así como sobre las medidas de dispersión, incluyendo la desviación estándar, la varianza y el rango. Estas herramientas permiten a los analistas cuantificar las características de sus datos y obtener una comprensión más profunda de los patrones y tendencias dentro de ellos. Además, se discuten las propiedades de la distribución de datos, como la asimetría y la curtosis, que son fundamentales para la interpretación correcta de los resultados estadísticos.

En el sexto capítulo, se exploran las distribuciones de frecuencia, tanto para variables cualitativas como cuantitativas. Este capítulo es crucial para comprender cómo se distribuyen los datos y cómo esta distribución puede influir en las conclusiones que se derivan del análisis. La presentación de datos en forma de distribuciones de frecuencia relativa proporciona una perspectiva más detallada de cómo se agrupan los valores en un conjunto de datos.

El séptimo capítulo trata sobre la creación y el uso de gráficos en el análisis estadístico. Los gráficos son herramientas poderosas para visualizar datos y comunicar resultados de manera clara y efectiva. Se presentan diferentes tipos de gráficos, incluyendo barras, líneas, áreas, gráficos circulares, histogramas y diagramas de dispersión, entre otros. Cada tipo de gráfico se discute en términos de su aplicabilidad y cómo puede utilizarse para resaltar diferentes aspectos de los datos analizados.

El octavo capítulo aborda las tablas de contingencia, una herramienta esencial para analizar la relación entre dos o más variables cualitativas. Se explica la configuración y el uso de estas tablas, así como la interpretación de los resultados obtenidos. Este capítulo es particularmente útil para quienes necesitan analizar datos categóricos y entender cómo se relacionan diferentes variables entre sí.

Finalmente, el noveno capítulo introduce los conceptos básicos de la probabilidad, un tema fundamental en la estadística inferencial. Aquí se explican las reglas fundamentales de la probabilidad, incluyendo la regla de la suma y la regla del producto, así como el teorema de Bayes, que es clave para la interpretación de eventos dependientes.

Este libro, por tanto, no solo ofrece una base sólida en teoría estadística, sino que también equipa al lector con habilidades prácticas esenciales para el análisis de datos en un entorno real. Al finalizar esta obra, los lectores habrán adquirido una comprensión profunda de la estadística y estarán capacitados para aplicar sus conocimientos en la solución de problemas complejos en sus respectivos campos.

Capítulo 1

Generalidades

Generalidades

Introducción

La estadística es una herramienta fundamental en la recolección, análisis e interpretación de datos, que permite transformar la información en conocimiento útil para la toma de decisiones en diversas áreas como la economía, la administración, las ciencias sociales y la investigación científica. Este capítulo introduce los conceptos básicos de la estadística y proporciona un marco conceptual para comprender su importancia y aplicaciones.

Se comienza con una explicación de lo que es la estadística, considerando la diferencia entre la descriptiva y la inferencial. Se describen los distintos tipos de variables, esenciales para organizar y clasificar la información, y se discuten las escalas de medición, que determinan cómo se pueden interpretar y manipular los datos.

Además, se analiza la procedencia de los datos, que puede variar desde encuestas y experimentos hasta registros administrativos y fuentes externas. La comprensión de las fuentes de datos es crucial para garantizar la calidad y relevancia de los análisis estadísticos. Finalmente, se abordan los tipos de datos, como cualitativos y cuantitativos, y su impacto en el enfoque analítico que se debe adoptar.

Este capítulo establece las bases necesarias para el estudio de la estadística, ofrece una comprensión integral de los conceptos fundamentales que se aplicarán en análisis más avanzados. Es una introducción esencial para

cualquier persona que desee utilizar la estadística como una herramienta para interpretar datos y tomar decisiones informadas.

Objetivos del capítulo

- Identificar los diferentes tipos de variables y escalas de medición utilizados en la estadística.
- Explicar la clasificación de las variables según su relación y naturaleza.
- Utilizar las escalas de medición adecuadas para categorizar diferentes tipos de datos en un estudio.

1.1 La estadística

La estadística es la ciencia que tiene por objeto:

- Recolectar información por medio de fuentes primarias y/o secundarias
- Organizar la información recolectada
- Presentar la información organizada en tablas y/o gráficos
- Analizar la información presentada
- Interpretar la información para toma de decisiones.

Debido a la naturaleza de la ciencia estadística, esta es muy importante en áreas como administración, medicina, ingenierías, economía, entre otras. El uso que hace la estadística de información histórica permite conocer las variaciones existentes a lo largo del tiempo, que se pueden usar como materia prima para toma de decisiones.

Cabe resaltar que, en la actualidad, el gobierno de datos es muy apreciado, por lo que su manejo es cada vez mayor en las grandes, pequeñas y medianas empresas. Se resalta que su aplicación es importante para la identificación de patrones, lo que facilita las proyecciones y pronósticos en el desarrollo de los procesos de mejora de productos y servicios (Anderson et al. 2019).

1.2 Tipos de estadística

Descriptiva: es parte de la estadística que se enfoca en la recolección, orden, presentación y descripción de datos de manera numérica o gráfica. Su objetivo principal es resumir y analizar características básicas de un conjunto de datos, como la media, mediana, moda, desviación estándar, entre otros, para obtener una comprensión general de la distribución y la variabilidad de los datos. En resumen, la estadística descriptiva proporciona herramientas para resumir y describir datos de manera que sean más fáciles de interpretar y analizar.

Este proceso de simplificación de datos permite una comprensión más accesible de la información y facilita el proceso de toma de decisiones en áreas como los negocios y la administración, en que la visualización y organización de la información resultan esenciales para una comunicación efectiva de los resultados (Newbold et al., 2008).

Inferencial: se enfoca en hacer deducciones o conclusiones sobre una población a partir de información recopilada de una muestra de esa población. En otras palabras, se trata de sacar conclusiones sobre toda una población basándose en datos recolectados de una muestra representativa de la misma. Esto se logra mediante técnicas como la estimación de parámetros y la realización de pruebas de hipótesis, que permiten tomar decisiones y hacer predicciones basadas en la información disponible. La estadística inferencial es fundamental en la investigación científica, el análisis de datos y la toma de decisiones en diversos campos como la ciencia, la medicina, la economía y la sociología, entre otros (Webster, 2000).

1.3 Tipos de variables

La clasificación de las variables es un componente esencial en el análisis estadístico, debido a que define cómo se deben abordar los datos y qué métodos son los más apropiados para su estudio. Las variables se pueden clasificar según su relación y naturaleza, lo que permite identificar diferencias clave entre datos cualitativos y cuantitativos. Esta categorización no solo es crucial para la correcta interpretación de los resultados, sino que también guía la elección de técnicas analíticas que revelen relaciones significativas entre las variables en distintos contextos de investigación. Una comprensión sólida

de los tipos de variables es, por tanto, vital para la precisión y la validez de cualquier análisis de datos.

1.3.1 Clasificación de acuerdo con la relación

Las variables pueden tener una relación directa o indirecta dependiendo el tipo de fenómeno que se esté analizando. En el análisis estadístico, debido a la interdependencia existente entre vectores, siempre se considera el impacto que existe al mover determinada variable sobre un caso específico, diferenciándose así entre variables dependientes e independientes.

Dependientes: son aquellas que dependen de otra variable para su ejecución. También se las conoce como variable explicada, predicha, regresada, respuesta, endógena, resultado, controlada y, en el caso del SPSS, como objetivo.

Independientes: son aquellas que NO dependen de otra variable para su ejecución. También se las conoce como variable explicativa, predictora, regresora, impulso, exógena, covariante, de control y, en el caso del SPSS, como entrada.

1.3.2 Clasificación de acuerdo con la naturaleza

Según su naturaleza, pueden clasificarse en cualitativas y cuantitativas. Las variables cualitativas describen características o cualidades que no se pueden medir numéricamente, como el color, la nacionalidad o la marca. En cambio, las variables cuantitativas representan cantidades que pueden ser medidas de manera numérica y pueden ser continuas o discretas. Las primeras toman valores dentro de un rango, como la altura o el peso, mientras que las segundas toman valores específicos y separados, como la cantidad de hijos o el número de veces que ocurre un evento. Esta distinción es fundamental en el análisis estadístico, ya que determina el tipo de técnicas y herramientas que se utilizarán para estudiar y comprender los datos.

1.4 Escalas de medición

Se refiere a un sistema que clasifica y organiza datos de acuerdo con un conjunto de reglas o criterios específicos. En estadística y en investigación, las

escalas de medición son fundamentales porque determinan cómo los datos pueden ser analizados e interpretados. En la figura 1.4.1, se puede observar la clasificación de las variables según su escala de medición.

Figura 1.4.1 Clasificación de las etiquetas por escala de medición.

Etiqueta	Medida
Edad en años	Escala
Nivel de educación	Ordinal
Años con la empresa actual	Escala
Años en la dirección actual	Escala
Ingresos familiares en miles	Escala
Tasa de deuda sobre ingresos (x100)	Escala
Deuda de la tarjeta de crédito en miles	Escala
Otras deudas en miles	Escala
Impagos anteriores	Nominal
Impago pronosticado, modelo 1	Escala
Impago pronosticado, modelo 2	Escala
Impago pronosticado, modelo 3	Escala

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

1.4.1 Nominales

Representan categorías o grupos que no tienen un orden inherente entre ellos. Son etiquetas o nombres que se asignan a diferentes elementos de un conjunto de datos para clasificarlos, pero no implican ningún tipo de jerarquía. Por ejemplo, el color de los ojos, lugar de nacimiento.

1.4.2 Ordinales

Son aquellas que representan categorías o grupos que tienen un orden preestablecido, pero la diferencia entre ellos no es necesariamente uniforme o cuantificable. A diferencia de las variables nominales, las variables ordinales implican un cierto grado de jerarquía entre las categorías, pero no se puede determinar la magnitud exacta de la diferencia entre ellas. Por ejemplo, el nivel de educación (primaria, secundaria, superior).

1.4.3 Escala

Se refiere a un tipo de variable cuantitativa que se mide en niveles de intervalo o razón. Estas variables son numéricas y continuas, lo que significa que los datos pueden tomar cualquier valor dentro de un rango específico y que los intervalos entre los valores son iguales. Esta clasificación es adecuada para realizar análisis estadísticos que requieren la comparación de magnitudes, como correlaciones y regresiones, ya que permite interpretar las diferencias entre valores de manera consistente.

1.5 Fuente de datos

Se refiere al origen de la información usada en un análisis, investigación o cualquier proceso de toma de decisiones. Estas fuentes proporcionan los datos necesarios para comprender, analizar y tomar decisiones sobre un tema específico.

1.5.1 Primarias

Se refieren a recursos que proporcionan datos originales y no interpretados sobre un tema específico. Estas fuentes suelen ser documentos, registros, testimonios, artefactos u otros materiales que provienen directamente de la experiencia o la observación de eventos, sin intermediarios ni interpretaciones adicionales. Las fuentes primarias son fundamentales en la investigación y la academia, ya que ofrecen información de primera mano que puede ser analizada y evaluada de manera directa. Ejemplos comunes de fuentes primarias incluyen documentos históricos originales, entrevistas, experimentos científicos y encuestas directas.

1.5.2 Secundarias

Son recursos que recopilan, analizan y presentan datos e investigaciones previamente realizadas por otros. Estas fuentes proporcionan un análisis y síntesis de información existente, en lugar de recopilar datos de primera mano. Algunos ejemplos comunes de fuentes secundarias incluyen libros de texto,

enciclopedias, informes de investigación, artículos académicos y bases de datos bibliográficas. Estas fuentes son útiles para obtener una visión general o contexto sobre un tema, respaldar argumentos con evidencia ya establecida y realizar comparaciones entre diferentes estudios o puntos de vista.

1.6 Tipos de datos

Los datos pueden clasificarse según el momento y la forma en que son obtenidos, lo cual determina su estructura y las posibilidades analíticas que ofrecen.

1.6.1 Corte transversal

Se refieren a información recopilada en un momento específico en el tiempo, generalmente en una única ocasión. Este tipo de datos se utilizan en estudios que buscan analizar una población o muestra en un punto determinado, sin seguir a los individuos a lo largo del tiempo.

Tabla 1.6.1 Datos en corte transversal

ID	Nivel de educación (Años)	Años de experiencia	Ingresos anuales (USD)
1	16	3	40
2	18	5	55
3	14	20	60
4	16	10	50
5	12	35	58

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

Descripción de variables

- **ID:** identificador único para cada individuo
- **Nivel de educación (Años):** años de educación formal completados por el individuo
- **Años de experiencia:** años de experiencia laboral del individuo
- **Ingresos anuales (USD):** ingresos anuales en dólares estadounidenses

1.6.2 Series de tiempo

Son una colección de observaciones secuenciales tomadas en intervalos regulares a lo largo del tiempo. Estos datos capturan la evolución de una variable o fenómeno en función del tiempo, lo que permite analizar patrones, tendencias y ciclos temporales.

Tabla 1.6.2 Datos en series de tiempo

Año	Ventas anuales (USD)
2020	600
2021	650
2022	700
2023	750

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

Descripción de variables

- **Año:** año calendario
- **Ventas anuales (USD):** ventas totales anuales en dólares estadounidenses

1.6.3 Panel

Se refiere a un tipo de datos en el que se recopilan observaciones a lo largo del tiempo de las mismas unidades de estudio, como individuos, empresas o países.

Tabla 1.6.3 Datos en panel

ID	Año	Ingresos anuales (USD)	Años de educación	Años de experiencia
1	2020	50	16	5
1	2021	52	16	6
2	2020	45	14	10
2	2021	47	14	11

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

Descripción de variables:

- **ID:** identificador único para cada individuo
- **Año:** año calendario
- **Ingresos Anuales (USD):** ingresos anuales en dólares estadounidenses
- **Años de educación:** años de educación formal completados por el individuo
- **Años de experiencia:** años de experiencia laboral del individuo

1.6.4 Micropanel

Son un tipo específico de datos longitudinales que se recopilan de manera periódica de una muestra de individuos, hogares o empresas a lo largo del tiempo. Estos datos permiten el seguimiento y análisis de cambios y tendencias en variables clave a nivel individual o de grupo. En otras palabras, un micropanel es una serie temporal de datos detallados y específicos sobre un conjunto seleccionado de unidades de estudio, que se recopilan en intervalos regulares, como trimestral o anualmente, con el fin de estudiar cambios, comportamientos y relaciones a lo largo del tiempo.

Tabla 1.6.4 Datos en micropanel

ID	Año	Consumo (USD)	Ahorro (USD)
1	2020	20	5
1	2021	22	6
2	2020	18	4,5
2	2021	19	5

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

Descripción de variables

- **ID:** identificador único para cada individuo
- **Año:** año calendario
- **Consumo (USD):** gasto anual en consumo en dólares estadounidenses
- **Ahorro (USD):** ahorro anual en dólares estadounidenses

Nota. Los datos en micropanel son simplemente una versión a menor escala de los datos en panel.

Capítulo 2

Estructura del SPSS y manejo de datos

Estructura del SPSS y manejo de datos

Introducción

El SPSS (Statistical Package for the Social Sciences) es una de las herramientas más utilizadas en el análisis de datos en ciencias sociales, investigación de mercado y otras disciplinas. Este capítulo proporciona una introducción detallada a la estructura del SPSS para ayudar al lector a familiarizarse con el entorno de trabajo del *software*.

Se describe la funcionalidad de las diversas ventanas que componen el SPSS, como la de datos, que es el espacio principal donde se ingresan y visualizan los datos; la de variables, donde se definen las propiedades de las variables; y la de resultados, donde se muestran los *outputs* de los análisis realizados. Estas ventanas permiten una gestión eficiente de los datos y la ejecución de análisis complejos de manera intuitiva.

El capítulo también aborda las barras de menú principal, herramientas y estado, que facilitan el acceso a las distintas funciones del SPSS, lo que permite al usuario realizar tareas como el análisis estadístico, la creación de gráficos y la manipulación de datos con facilidad. Además, se explican los cuadros de diálogo, que son interfaces para configurar y ejecutar los diferentes comandos.

Finalmente, se proporciona una guía sobre la creación y edición de archivos de trabajo, incluyendo la definición de las propiedades de las variables,

un paso crítico para asegurar que los datos se manejen de manera coherente y precisa. Este capítulo es esencial para aquellos que buscan utilizar SPSS de manera efectiva para realizar análisis estadísticos rigurosos.

Objetivos del capítulo

- Reconocer las diferentes ventanas y barras de herramientas del software SPSS.
- Crear y guardar un archivo de trabajo en SPSS, incluyendo la definición de propiedades de variables.
- Revisar y editar archivos de datos en SPSS para garantizar la correcta configuración de las variables.

2.1 Tipo de ventanas

Cuando se inicia el programa SPSS, siempre aparecerá una ventana emergente que posee información de utilidad, tal como nuevos archivos, archivos recientes, novedades, módulos y guías de aprendizajes. Las guías de aprendizaje se refieren a temas como lectura de datos, uso del editor de datos, entre otros, los cuales pueden ser elegidos de acuerdo con las necesidades del investigador.

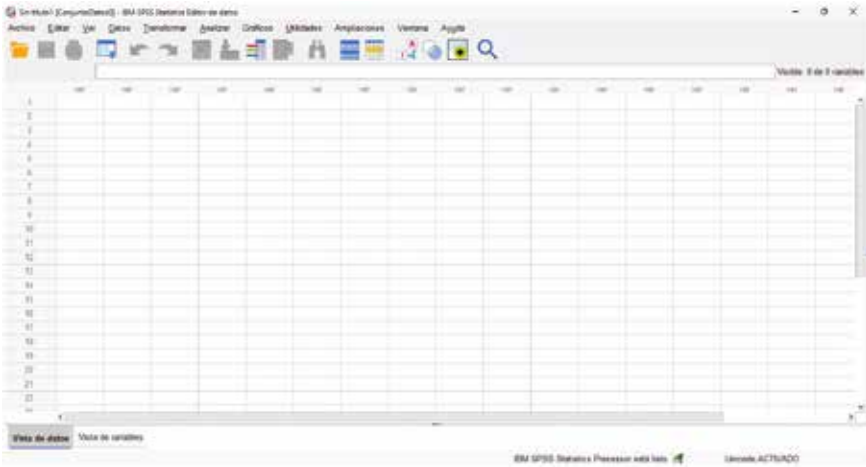
Figura 2.1.1 Ventana de bienvenida IBM SPSS Statistics

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Una vez que se ha dado paso a esta opción, aparecerá la ventana, **editor de datos**, la cual viene a ser la ventana principal del SPSS y consta de dos partes: vista de datos y vista de variables.

- **Vista de datos.** Es la sección en la cual se digitan las observaciones de una encuesta o donde se encuentra la base de datos con la que vamos a trabajar. En esta ventana, veremos los datos tabulados y codificados de acuerdo con nuestros requerimientos (véase la figura 2.1.2).

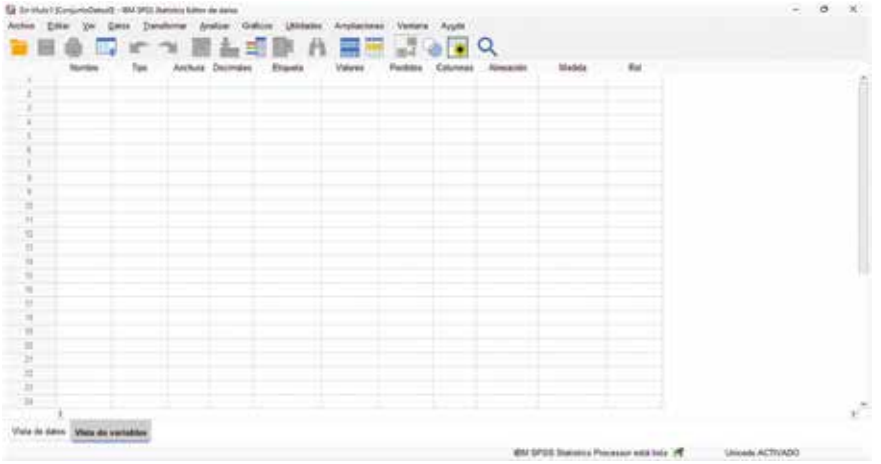
Figura 2.1.2 Ventana de edición de datos IBM SPSS Statistics



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Vista de variables.** En esta ventana se pueden observar ítems tales como Nombre, tipo, anchura, decimales, etiquetas, valores, perdidos, columnas, alineación, medidas y rol. Igualmente es importante indicar que, en esta ventana, se pueden editar las variables incluidas en los trabajos de investigación de mercado o académicos (véase la figura 2.1.3).

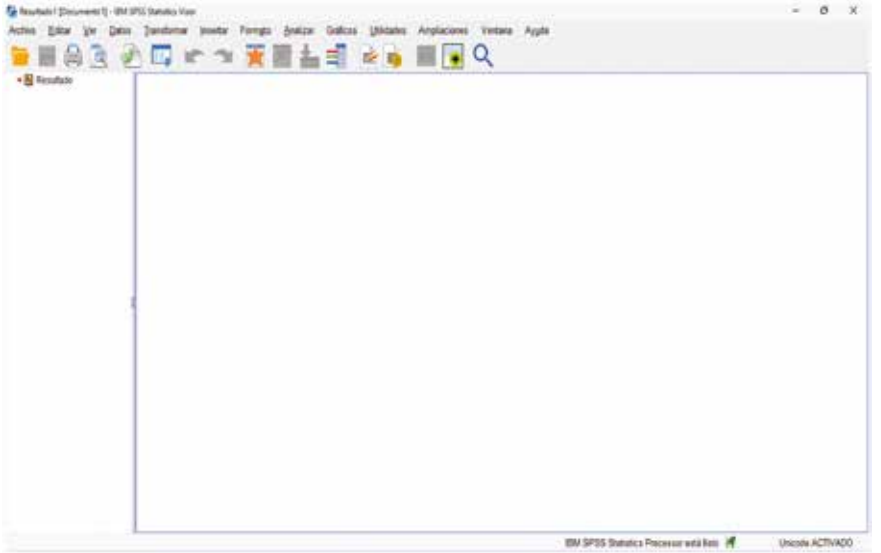
Figura 2.1.3 Ventana de edición de variables IBM SPSS Statistics



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Resultados.** En esta ventana, se pueden observar los resultados de las operaciones realizadas en la ventana de editor de datos y puede ser modificada y guardada para un posterior uso.

Figura 2.1.4 Ventana de resultados IBM SPSS Statistics



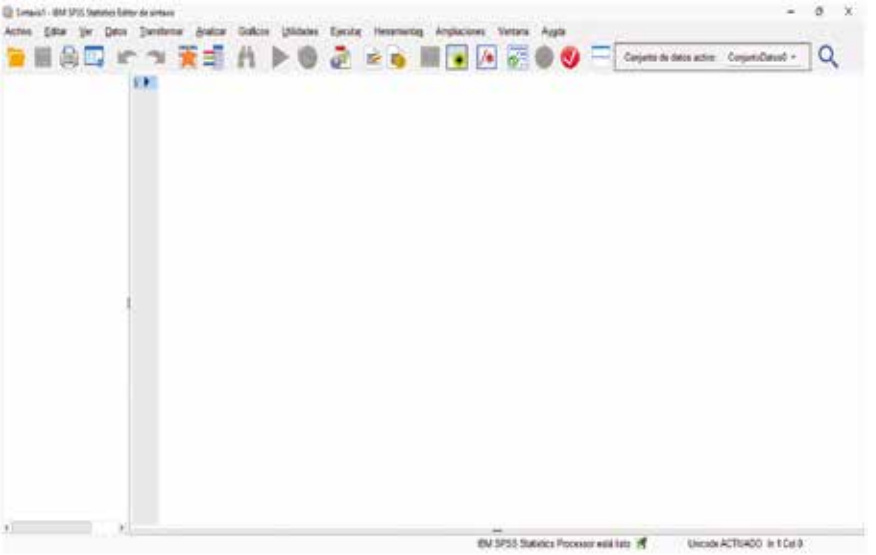
Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Internamente, en la ventana de resultados, se puede realizar lo siguiente:

- Editar tablas pivote. Al hacer doble clic sobre una tabla, se puede editar el texto, intercambiar filas, columnas, mostrar u ocultar resultados, etc.
- Editar gráficos. Permite llevar a cabo todo tipo de modificaciones en los gráficos, es decir, se puede cambiar color, ejes, etc.
- Editar texto de resultado. Se puede añadir texto como observaciones y este no aparece en las tablas pivote. Este se habilita ingresando a la pestaña insertar – nuevo texto.

- **Editor de sintaxis.** En esta ventana, se permite editar los comandos que se utilizan en cada procedimiento. Además, se puede guardar como un archivo de sintaxis para posteriormente utilizarlo.

Figura 2.1.5 Ventana de editor de sintaxis IBM SPSS Statistics



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En el presente documento, se hará uso de la sintaxis para el desarrollo de los ejercicios propuestos en el texto, copiándose los comandos y pegándose en lo posterior en esta ventana.

2.2 Barras de menú principal, herramientas, estado

Menú: la barra de menú contiene las funcionalidades del SPSS, por lo tanto, es necesario detallar el contenido de cada pestaña.

Figura 2.2.1 Barra de menú

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Archivo.** Tiene varias opciones como abrir un archivo existente, guardar, etc.
- **Editar.** Contiene opciones como, copiar datos, buscar, etc.
- **Ver.** Permite personalizar la barra de estado, visualizar las etiquetas de datos, entre otras opciones.
- **Datos.** Permite hacer cambios que afectan a todo el archivo de datos y cuyos cambios son temporales mientras no se guarden los valores (trasponer variables, fusionar archivos, entre otras).
- **Transformar.** Sirve para crear nuevas variables, series temporales, etc.
- **Analizar.** En esta pestaña, se ejecutan todos los procedimientos estadísticos según las necesidades presentadas.
- **Gráficos.** Brinda las opciones de gráficos que se pueden ejecutar de acuerdo con los datos.
- **Utilidades.** Proporciona información de las variables, cambiar fuentes, etc.
- **Ampliaciones.** Permite agregar paquetes adicionales para el análisis.
- **Ventana.** Dividir y/o minimizar la ventana.
- **Ayuda.** Tutorial, asistente estadístico, etc.

Herramientas: Se localiza situada debajo de la barra de menú principal. En esta barra, se encuentran algunos comandos básicos del SPSS.

Figura 2.2.2 Barra de herramientas

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Estado: Se sitúa en la base de la pantalla e indica el estado actual del proceso, considerando la siguiente información: estado del comando que se está ejecutando, filtro de datos, variable ponderada y segmentación de datos.

Figura 2.2.3 Barra de estado

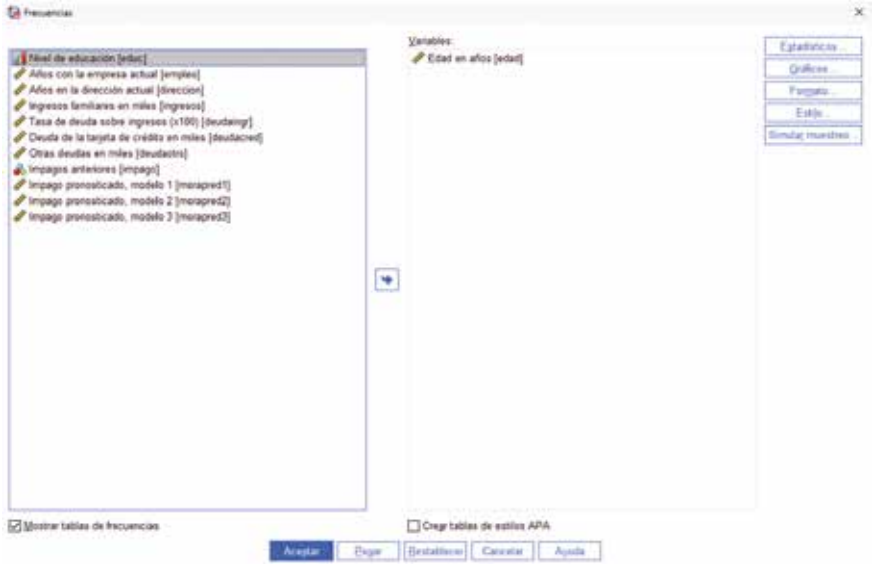
Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

2.3 Cuadros de diálogo

En esta ventana, se encuentra información de las variables que han sido ingresadas en la base, permitiendo, en lo posterior, su selección para obtener los estadísticos necesarios para el análisis, y presenta los siguientes items:

- Lista de variables
- Lista de variables seleccionadas
- Botones de comando

Figura 2.3.1 Cuadro de diálogo



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

2.4 Creación de un archivo de trabajo

Los archivos de trabajo pueden ser obtenidos de fuentes primarias o secundarias. En este caso, es importante considerar cualquiera de los siguientes pasos, conforme las necesidades:

- **Archivos nuevos.** Se va al menú archivo y se selecciona nuevo y luego datos o sintaxis.
- **Abrir archivo de datos.** Se selecciona un archivo preexistente; luego se elige abrir. En este caso, se selecciona el archivo que se requiera.
- **Abrir base de datos.** Ayuda a abrir un archivo que no esté en formato SPSS.
- **Leer datos de texto.** Lee archivos en formato de texto. En algunos casos, es interesante guardar los datos en tipo texto, esto es debido a que la información en tipo texto puede tener un peso menor al que se obtiene cuando se guarda los datos de manera tradicional.
- **Leer datos de cognos.** En este caso se debe contar con una base de datos en la web.

Posteriormente, una vez ejecutados los cálculos del apartado anterior, se puede guardar el archivo de datos para posteriormente ser utilizado en nuevos trabajos. Para ello es necesario tomar en consideración lo siguiente:

- **Guardar.** Guarda el archivo en formato SPSS.
- **Guardar como.** Permite elegir en que formato se quiere grabar el archivo.
- **Guardar todos los datos.**
- **Exportar.** Permite exportar los archivos al formato deseado.
- **Marcar archivo como solo lectura.**
- **Recopilar información de variable.**
- **Cambiar nombre de conjunto de datos.**
- **Mostrar información del archivo de datos.** En este caso, las variables se encuentran detalladas.
- **Gestionar conjunto de datos.**
- **Caché de los datos.**
- **Definir opciones de salida del visor (sintaxis).**

2.5 Editar un archivo

- **Borrar.** Elimina toda la información de la variable seleccionada.
- **Insertar variable.** Crea una variable estándar, la cual debe ser modificada con los requerimientos y la información de la que dispone el investigador.
- **Insertar caso.** Este comando abre camino dentro de una variable para incluir una nueva observación.
- **Reemplazar.**
- **Buscar archivo de datos.**
- **Ir a caso.** Ayuda a ir directamente a la observación que se requiera, en cualquier momento.
- **Ir a variable.** Ayuda a trasladarse directamente a la variable que se requiere analizar.
- **Ir a imputación.**
- **Opciones.** Permite cambiar la apariencia, el formato de presentación de los datos, entre otros.

2.6 Definición de propiedades de las variables

Cuando se ingresa una nueva variable se deben tomar en cuenta sus propiedades identificativas, tales como nombre, tipo de variable, anchura, decimales que se consideran en los datos, etiqueta, valores, datos perdidos, columnas, alineación, unidad de medida y su rol; esto es, con la finalidad de conocer el tipo de información con la que se cuenta y con la cual se va a trabajar.

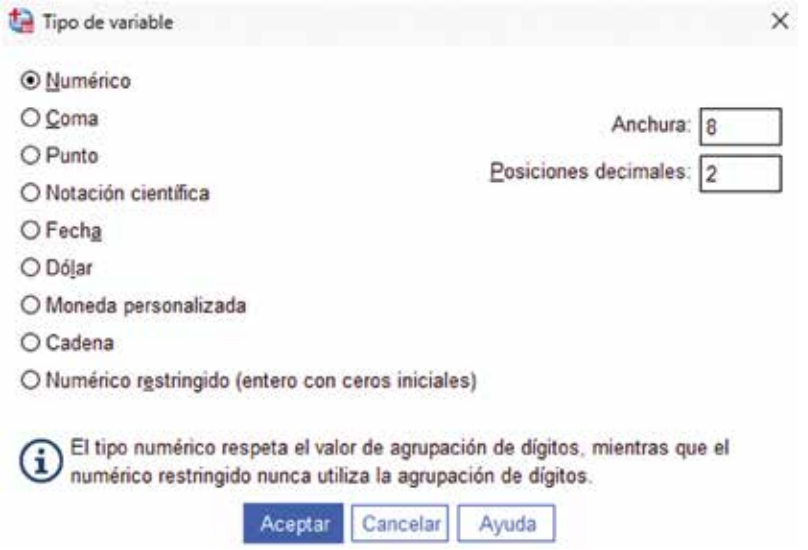
Figura 2.6.1 Barra de propiedades de la variable

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Nombre.** En este apartado se debe colocar un nombre identificativo corto, que guarde relación con la variable que se pretende analizar. No se pueden utilizar espacios ni caracteres especiales.

- **Tipo de variable.** Es necesario considerar cómo fue levantada la información para luego cargarla en la base. En este caso, se puede configurar como:
  - o **Numérico.** Los valores se muestran en formato numérico estándar.
  - o **Coma.** Cuando los valores están delimitados cada tres posiciones con coma y los decimales los delimita con punto (1,000.00).
  - o **Punto.** Cuando los valores están delimitados cada tres posiciones con punto y los decimales los delimita con coma (1.000,00).
  - o **Notación científica.** Cuando los valores se muestran con una E.
  - o **Fecha.** Las variables indican fechas.
  - o **Moneda personalizada.** Se representa la información en moneda.
  - o **Cadena.** Son variables alfanuméricas. Estas pueden contener cualquier carácter.
  - o **Anchura y decimales.** Se emplean para especificar la anchura y el número de decimales que están permitidos en las columnas (cuantos números se permiten por observación, Ejemplo. «2 anchura 1» o «2,2 anchura 2» o «2,22 anchura 3»), y el número de decimales que contienen las variables.

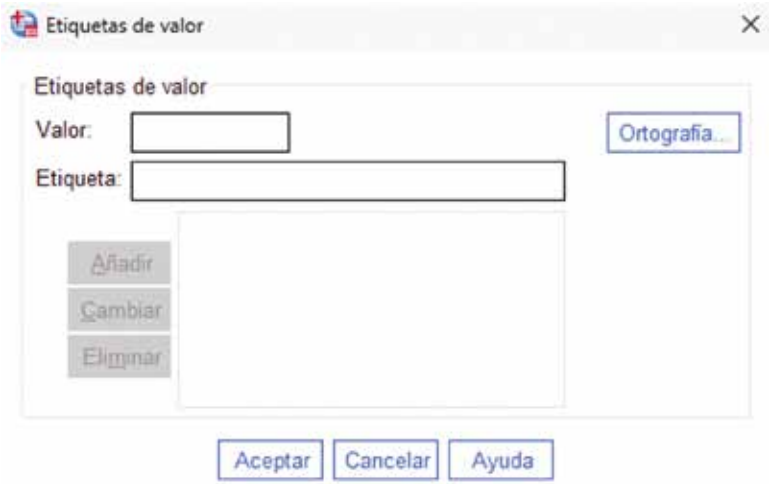
Figura 2.6.2 Ventana de edición del tipo de variable ingresada



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Para codificar los valores obtenidos, es necesario acceder a la ventana Etiquetas de valor. En esta sección, se puede ingresar los valores codificados correspondientes a la información recolectada, ya sea de fuentes primarias y/o secundarias.

Figura 2.6.3 Ventana de etiquetas de valor

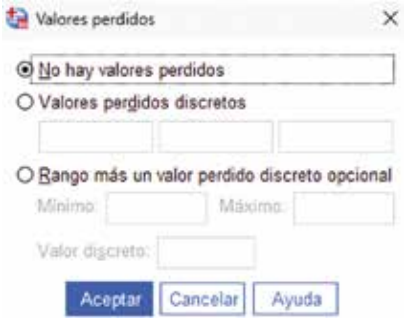


Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En la ventana de etiquetas de valor, se presenta lo siguiente:

- **Valores.** En esta pestaña se ingresa el número que corresponde según el código asignado previamente.
- **Etiqueta.** Va el nombre que corresponde al código asignado en las codificaciones de valores.

Figura 2.6.4 Ventana para ingresar valores perdidos



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Perdidos.** En esta ventana se detallan los valores que el sistema no debe tomar en consideración dentro del análisis. El analista determinará que valores son perdidos.
- **Columnas.** Determina la anchura de la columna
- **Alineación.** En qué posición de la celda se visualizan los datos en el editor.
- **Medida.**
  - **Nominal.** Es una variable cualitativa que permite la distinción entre elementos, pero no su orden. Denotan atributos. Ejemplo: nacionalidad, estado civil, etc.
  - **Ordinal.** Es una variable cualitativa y determina el orden. Ejemplo: escala de Likert.
  - **Escala.** Es una variable cuantitativa.
- **Rol.** Este ítem contiene información de actuación de las variables incluidas en la base de datos, es así como define su campo de actuación.

Dentro de lo revisado, el rol contiene la distinción de variables de entrada y objetivo como los principales; sin embargo, también se puede seleccionar ambos roles o a su vez, ninguno, al momento de definir los vectores. Igualmente, se tiene la opción para utilizar a la variable ingresada, como generadora de una partición o división de la base de datos.

Como ejemplo de uso de la clasificación en los roles más utilizados, se tiene la base de datos presentada en el software SPSS, con el nombre Bankloan.

Figura 2.6.5 Rol de la variable ingresada

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

  - **Entrada.** La variable se utiliza como una entrada (predictor).

- **Objetivo.** La variable se utiliza como un objetivo (predicha).
- **Ambas.** Las variables se desenvuelven como entrada y como objetivo.
- **Ninguno.** En este aspecto, se considera cuando las variables incluidas no desempeñan funciones.
- **Partición.** Indica un campo utilizado para dividir los datos en muestras independientes para entrenamiento, comprobación y validación (opcional).
- **División.** (Campos nominales, ordinales o marca únicamente) Especifica que se creará un modelo para cada valor posible del campo.

Figura 2.6.6 Datos obtenidos de Bankloan, como base para ejemplos

Nombre	Etiqueta	Rol
edad	Edad en años	Entrada
educ	Nivel de educación	Entrada
empleo	Años con la empresa actual	Entrada
direccion	Años en la dirección actual	Entrada
ingresos	Ingresos familiares en miles	Entrada
deudaingr	Tasa de deuda sobre ingresos (x100)	Entrada
deudacred	Deuda de la tarjeta de crédito en miles	Entrada
deudaotro	Otras deudas en miles	Entrada
impago	Impagos anteriores	Objetivo
morapred1	Impago pronosticado, modelo 1	Ninguna
morapred2	Impago pronosticado, modelo 2	Ninguna
morapred3	Impago pronosticado, modelo 3	Entrada

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En los datos presentados en la figura 2.6.6, es posible observar la clasificación por objetivo, a la variable impago y como entrada a las variables edad, educación, empleo, entre otras. En este caso, se tiene *a priori* que lo que se intenta investigar es el impago de una deuda considerando si el individuo tiene empleo, educación, etc. Este aspecto es importante para poder analizar a las personas en función si son sujetas de crédito o no lo son.

Capítulo 3

Tabulación de encuestas e importación de base de datos

Tabulación de encuestas e importación de base de datos

Introducción

El manejo adecuado de encuestas y bases de datos es fundamental para obtener resultados precisos y significativos en un análisis estadístico. Este capítulo aborda las técnicas necesarias para tabular encuestas e importar bases de datos en el SPSS para facilitar la integración de datos de diferentes fuentes para un análisis más robusto.

Se comienza con la codificación del instrumento de recolección de datos, un paso crucial para asegurar que la información obtenida sea manejada correctamente dentro del software. A continuación, se explica cómo crear y definir variables en SPSS, lo que permite organizar los datos de manera que se puedan analizar de forma efectiva. Este proceso incluye la asignación de nombres, etiquetas y valores a las variables, lo que garantiza que el análisis refleje con precisión la información recogida.

El capítulo también cubre el ingreso de datos, describiendo las mejores prácticas para asegurar la precisión y consistencia de los datos introducidos. Además se explora cómo guardar y proteger los archivos de trabajo, una medida esencial para preservar la integridad de los datos y facilitar el acceso en el futuro.

Uno de los aspectos más destacados de este capítulo es la importación de bases de datos externas. Se proporcionan instrucciones detalladas sobre cómo importar y ajustar los datos importados, asegurando su compatibilidad con el entorno de trabajo de SPSS. Este capítulo es vital para aquellos que necesitan manejar grandes volúmenes de datos y desean realizar análisis estadísticos confiables y precisos.

Objetivos del capítulo

- Enumerar los pasos para la codificación de instrumentos y creación de variables en SPSS.
- Ingresar y guardar datos en un archivo de SPSS, asegurando la integridad de la información.
- Importar una base de datos externa a SPSS, configurando correctamente las propiedades de las variables.

3.1 Codificación del instrumento

La codificación es un paso esencial que debe llevarse a cabo antes de iniciar el proceso de tabulación, ya que facilita y optimiza el ingreso de datos en el programa. Al contar con variables previamente codificadas, se garantiza una mayor precisión y eficiencia en la entrada de datos, lo que resulta crucial para un análisis adecuado. Por ejemplo, al disponer de información como categorías numéricas o códigos asignados a respuestas específicas, se simplifica el trabajo y se minimizan los errores durante la carga de los datos en el sistema.

I. Información General

1. Sexo: Masculino (1) Femenino (2) **Se codifica**
2. Fecha de nacimiento (año, mes, día): ____/__/__ **No se codifica**
3. Edad: _____ **No se codifica**

II. Información Específica: Se codifica la escala de Likert **Se codifica**

Pregunta:	Siempre (1)	Regular (2)	Nunca (3)
4. Participa usted en:			
4.1 Reuniones familiares			
4.2 Eventos estudiantiles			

5. Utiliza usted: (Pregunta de opción múltiple) **Se Codifica**

a. Celular (1)
b. Televisión (2)
c. Radio (3)

6. ¿Para que usted utiliza los medios de comunicación? (Pregunta abierta) **Se Codifica**
Utilizo los medios de comunicación para ver **noticias** y para conocer los adelantos de la **última moda**.

- Noticias 1
- Última moda 2

La información recolectada y codificada está completamente preparada para ser procesada e ingresada en SPSS. Para practicar y consolidar su aprendizaje, puede hacer uso de la sintaxis proporcionada al final del capítulo, lo que le permitirá aplicar los procedimientos de manera efectiva y asegurar la correcta ejecución del análisis de datos.

3.2 Creación de variables

Para ingresar los datos codificados en el SPSS es importante crear variables en la pestaña vista de variables. Por lo tanto, se debe tener consideración de cada una de las características de las observaciones y de las opciones presentes en el software. Por ejemplo: en la primera pregunta de la encuesta de prueba se tiene:

- Nombre. Sexo / Preg1
- Tipo. Numérico
- Anchura. 8
- Decimales. 0
- Etiqueta. Sexo
- Valores. 1-masculino; 2-Femenino
- Perdido. Ninguno
- Columnas. 8
- Alineación. Derecha
- Medida. Nominal
- Rol. Entrada

Figura 3.2.1 Barra de propiedades de la variable

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

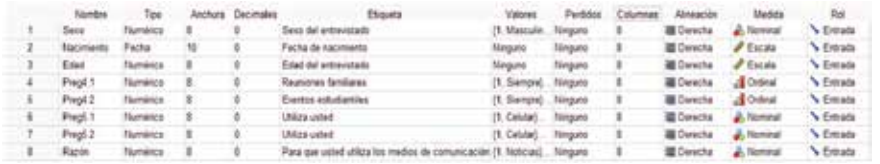
En este apartado, resulta esencial hacer clic en la pestaña de valores para ingresar los datos codificados correctamente. Este paso garantiza que los valores se registren de manera adecuada en el sistema, lo que permite una gestión eficiente de la información y asegura la precisión en los procesos posteriores de análisis y toma de decisiones.

Figura 3.2.2 Ventana de etiquetas de valor

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En las etiquetas de valor, se registran los códigos generados previamente, para asegurar que cada categoría de respuesta esté correctamente identificada antes de ingresar los datos en SPSS. En esta fase, se anotan cuidadosamente los códigos y sus correspondientes nombres, lo que facilita la organización y posterior análisis de los datos. Una vez que las preguntas han sido ingresadas en el sistema, se procede a introducir la información recolectada por medio de las encuestas. Así se garantiza que cada dato esté correctamente codificado y alineado con los valores establecidos previamente.

Figura 3.2.3 Ventana de vista de variables con información ingresada de datos obtenidos de la encuesta modelo



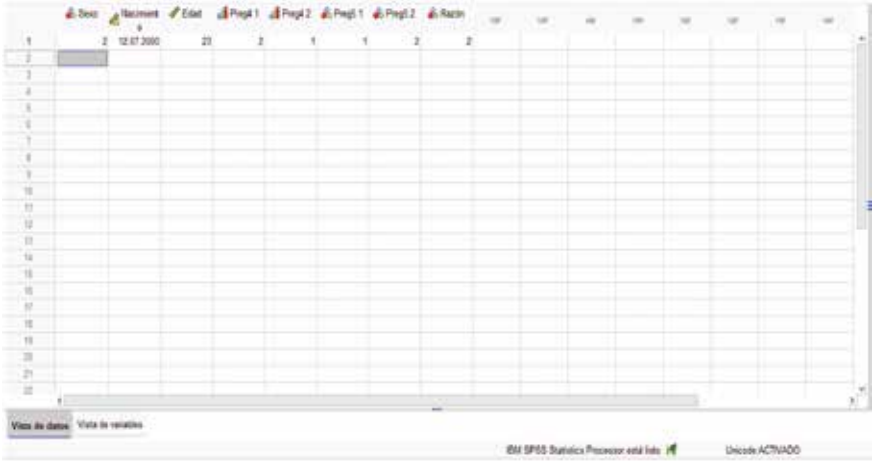
	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Sexo	Numérico	8	0	Sexo del entrevistado	{1, Masculin...	Ninguno	8	Derecha	Nominal	Entrada
2	Nacimiento	Fecha	10	0	Fecha de nacimiento	Ninguno	Ninguno	8	Derecha	Escala	Entrada
3	Edad	Numérico	8	0	Edad del entrevistado	Ninguno	Ninguno	8	Derecha	Escala	Entrada
4	Pregt.1	Numérico	8	0	Reuniones familiares	{1, Siempre}...	Ninguno	8	Derecha	Ordinal	Entrada
5	Pregt.2	Numérico	8	0	Eventos estudiantiles	{1, Siempre}...	Ninguno	8	Derecha	Ordinal	Entrada
6	PregS.1	Numérico	8	0	Utiliza usted	{1, Celular}...	Ninguno	8	Derecha	Nominal	Entrada
7	PregS.2	Numérico	8	0	Utiliza usted	{1, Celular}...	Ninguno	8	Derecha	Nominal	Entrada
8	Razón	Numérico	8	0	Para que usted utiliza los medios de comunicación	{1, Noticias}...	Ninguno	8	Derecha	Nominal	Entrada

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

3.3 Ingreso de datos

Para ingresar la información de la encuesta en SPSS, se debe acceder a la pestaña Vista de datos. Es fundamental que los datos sean introducidos conforme a la codificación previamente establecida, asegurando que cada respuesta se asocie correctamente con su código correspondiente. Esta codificación, realizada antes del ingreso de datos, facilita el análisis posterior y garantiza la coherencia en el procesamiento de la información. Al seguir este procedimiento, se minimizan errores y se asegura la precisión en los resultados obtenidos a partir de las encuestas.

Figura 3.3.1 Ventana de vista de datos con información ingresada de datos obtenidos de la encuesta modelo (vista de códigos)



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En la ventana, los datos ingresados pueden visualizarse de dos maneras distintas: utilizando los códigos numéricos asignados a cada categoría o mostrando los nombres codificados que corresponden a esos códigos. Esta flexibilidad es fundamental para adaptarse a las necesidades específicas del usuario, ya sea para una revisión rápida o para un análisis más detallado y preciso. Además, facilita la interpretación de los datos durante diferentes etapas del análisis. Las figuras 3.3.1 y 3.3.2 ilustran claramente estas opciones, proporcionando ejemplos de cómo se presentan los datos en cada formato, lo que contribuye a una mejor comprensión del proceso.

Figura 3.3.2 Ventana de vista de datos con información ingresada de datos obtenidos de la encuesta modelo (vista de nombres)

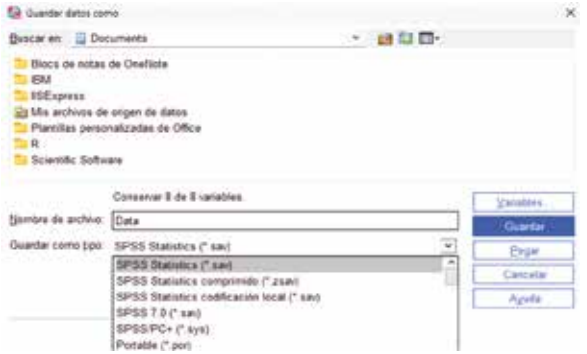
	Sexo	Nacimiento	Edad	Preg4.1	Preg4.2	Preg5.1	Preg5.2	Razón
1	Femenino	12.07.2000	23	Regular	Siempre	Celular	Televisión	Última moda
2	Masculino	13.09.2001	22	Siempre	Siempre	Televisión	Radio	Noticias
3	Femenino	21.08.2000	23	Regular	Regular	Radio	Radio	Noticias
4	Femenino	01.02.2004	20	Regular	Regular	Radio	Radio	Última moda
5	Femenino	01.12.2001	22	Nunca	Nunca	Televisión	Celular	Noticias
6	Masculino	12.07.2000	23	Nunca	Regular	Radio	Radio	Noticias
7	Masculino	14.03.2003	21	Regular	Siempre	Televisión	Televisión	Noticias
8	Masculino	12.10.2001	22	Siempre	Nunca	Radio	Radio	Noticias
9	Femenino	12.07.2000	23	Siempre	Nunca	Televisión	Televisión	Última moda
10	Masculino	19.01.2004	20	Regular	Nunca	Celular	Radio	Última moda
11								
12								
13								
14								
15								
16								

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

3.4 Guardar un archivo

Una vez finalizado el proceso de ingreso de información, o incluso durante el mismo, es posible guardar el documento en SPSS para asegurar la integridad de los datos. Para hacerlo, debe dirigirse a la pestaña Archivo y seleccionar la opción Guardar. Cabe destacar que el archivo puede guardarse en varios formatos, según las necesidades del usuario, lo que proporciona una gran flexibilidad. Esta funcionalidad no solo permite almacenar los datos de manera segura, sino que también facilita su acceso y recuperación en futuras sesiones de trabajo o análisis, lo que optimiza la eficiencia en la gestión de la información.

Figura 3.4.1 Ventana que muestra opciones de guardado

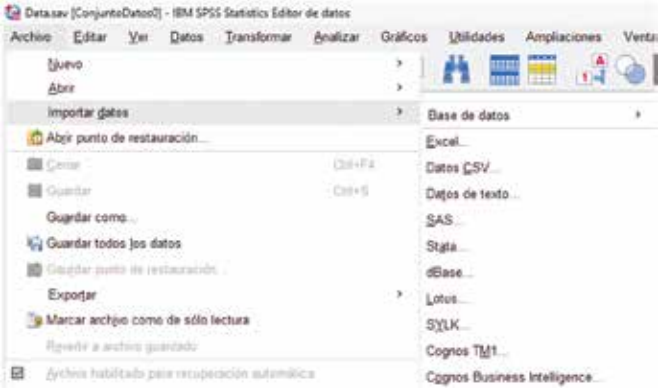


Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

3.5 Como importar una base de datos

En el manejo de datos, es fundamental la capacidad de importar bases de datos desde diversos formatos, lo que permite una flexibilidad considerable al trabajar con diferentes fuentes de información. Estos formatos pueden incluir archivos CSV, archivos de texto plano, bases de datos SQL, entre otros. Sin embargo, en el contexto del presente ejemplo, se ha optado por importar una base de datos en formato Excel, lo cual es una práctica común debido a la amplia utilización de este programa en la recopilación y manejo de datos.

Figura 3.5.1 Ventana que muestra opciones de importación

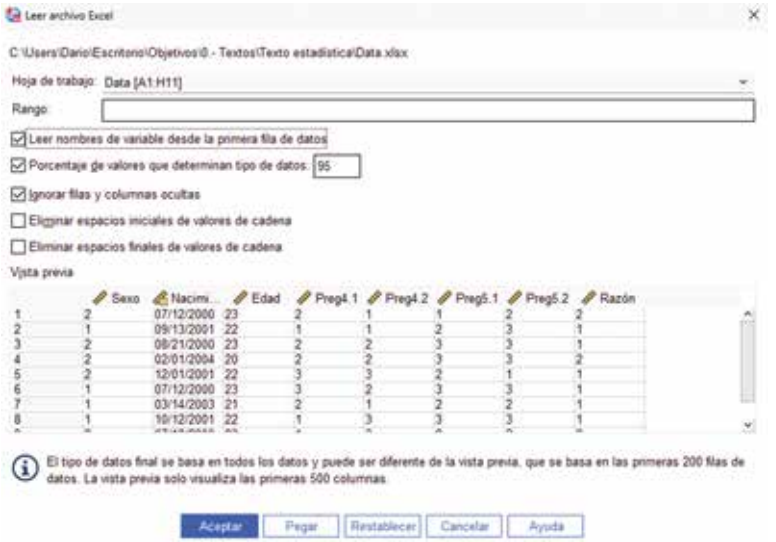


Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

El formato Excel es particularmente útil por su capacidad de organizar datos en hojas de cálculo, lo que facilita la visualización, edición y análisis de la información antes de importarla a programas estadísticos como SPSS, R o Python. Además, Excel permite trabajar con múltiples hojas dentro de un mismo archivo, lo que es ventajoso cuando se maneja un gran volumen de datos o se necesita estructurar la información en varias categorías.

Al importar datos desde Excel, es importante asegurarse de que las columnas estén correctamente formateadas y que no haya errores en la estructura de los datos, debido a que estos pueden generar problemas en los análisis posteriores. Esta opción es ideal para aquellos que buscan una integración sencilla y eficiente entre herramientas de análisis estadístico y las hojas de cálculo tradicionales.

Figura 3.5.2 Ventana de vista de variables con información ingresada de datos obtenidos de la encuesta modelo



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

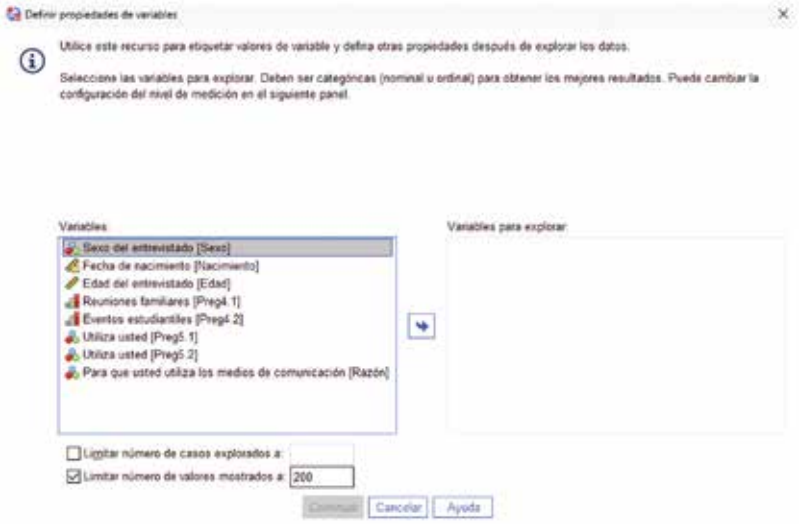
Al seleccionar la opción de importar datos y asegurarse de que la base de datos está en el formato adecuado, es crucial activar la casilla Leer nombres de variables desde la primera fila de datos. Si esta opción no se encuentra seleccionada, es probable que surjan inconvenientes durante la realización de los análisis de las variables, debido a que los nombres de las variables no

serán reconocidos correctamente por el software. Este paso es fundamental para garantizar que los datos sean interpretados de manera precisa y que los análisis subsiguientes reflejen correctamente la estructura de los datos (véase figura 3.5.2).

3.6 Propiedades de las variables

Las variables incluidas en el estudio pueden ser analizadas detalladamente una vez que los datos han sido ingresados. En este proceso, es posible examinar la composición de cada variable, el número de observaciones y la etiqueta asignada a cada una. Para realizar esta revisión, es necesario acceder al menú Datos y seleccionar la opción Definir propiedades de variables. Esta función permite asegurar que las variables están correctamente configuradas, lo que es esencial para garantizar la precisión en los análisis estadísticos posteriores.

Figura 3.6.1 Ventana para definir las propiedades de variables



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Finalmente, en la tabla 3.6.1, se presenta información en sintaxis. Para poder ejecutarlo en SPSS debe ir a:

- Archivo > Nuevo > Sintaxis.

- Copie su código, haga clic en la ventana de sintaxis, y pegue con Ctrl + V.
- Seleccione el código y haga clic en el botón Ejecutar (ícono «play» verde) o presione Ctrl + R.
- Revise la ventana de salida para ver los resultados.
- Vaya a Archivo > Guardar como para guardar el archivo con extensión .sps.

Tabla 3.6.1 Sintaxis con información modelo

```
DATA LIST LIST
 /Sexo (F1.0) Nacimiento (A11) Edad (F2.0) Preg4_1 (F1.0) Preg4_2 (F1.0) Preg5_1 (F1.0)
Preg5_2 (F1.0) Razón (F1.0).

BEGIN DATA
2 12-jul-00 23 2 1 1 2 2
1 13-sep-01 22 1 1 2 3 1
2 21-ago-00 23 2 2 3 3 1
2 01-feb-04 20 2 2 3 3 2
2 01-dic-01 22 3 3 2 1 1
1 12-jul-00 23 3 2 3 3 1
1 14-mar-03 21 2 1 2 2 1
1 12-oct-01 22 1 3 3 3 1
2 12-jul-00 23 1 3 2 2 2
1 19-ene-04 20 2 3 1 3 2
END DATA.

FORMATS Nacimiento (DATE11).
COMPUTE Nacimiento = DATE.DMY(DAY(Nacimiento), MONTH(Nacimiento),
YEAR(Nacimiento)).
VARIABLE LABELS Sexo 'Sexo' Nacimiento 'Fecha de Nacimiento' Edad 'Edad en años'
Preg4_1 'Pregunta 4.1' Preg4_2 'Pregunta 4.2'
Preg5_1 'Pregunta 5.1' Preg5_2 'Pregunta 5.2' Razón 'Razón de la encuesta'.
EXECUTE.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

Capítulo 4

Transformar datos

Transformar datos

Introducción

La transformación de datos es un paso crucial en el análisis estadístico, ya que permite preparar y ajustar la información recopilada para que sea analizada de manera efectiva. Este capítulo explora una variedad de técnicas de transformación de datos dentro de SPSS, lo que proporciona al lector las herramientas necesarias para optimizar el conjunto de datos antes de su análisis.

Se introducen conceptos como la respuesta múltiple, que permite manejar preguntas de encuesta con varias respuestas, y la identificación de casos duplicados y atípicos, que es crucial para garantizar la calidad de los datos. También se explican las técnicas para calcular nuevas variables y contar valores dentro de los casos, lo que permite una mayor flexibilidad en el análisis.

El capítulo aborda métodos para cambiar o recodificar variables, un proceso que puede ser necesario para adaptar los datos a los requisitos específicos del análisis. Se describen en detalle las opciones para recodificar variables en las mismas variables, en distintas variables y de manera automática, proporcionando ejemplos prácticos para cada caso.

Además, se discuten técnicas avanzadas como la creación de variables auxiliares y la agrupación visual y óptima, que permiten explorar y representar los datos de manera más efectiva. Finalmente, se aborda el reemplazo de valores perdidos, un aspecto crítico para mantener la integridad del conjunto de datos

y evitar sesgos en el análisis. Este capítulo es indispensable para quienes buscan mejorar la calidad y la usabilidad de sus datos, asegurando que estén en la mejor forma posible antes de proceder con análisis estadísticos más complejos.

Objetivos del capítulo

- Describir los métodos para transformar datos en SPSS, como recodificación y creación de variables auxiliares.
- Identificar y recodificar variables dentro de un conjunto de datos para adecuar el análisis.
- Generar variables nuevas a partir de las existentes utilizando funciones de transformación de datos en SPSS.

4.1 Respuesta múltiple

Cuando se trata de preguntas de opción múltiple, al momento de tabularlas y extraer información relevante de las variables, es posible agrupar cada respuesta en una sola variable. Para lograrlo, se debe acceder a la opción Analizar y seleccionar Respuesta múltiple. Esta funcionalidad permite manejar eficientemente las respuestas y obtener un análisis más cohesionado. Se recomienda utilizar la base de datos cargada en el capítulo 3 para practicar esta técnica.

Figura 4.1.1 Ventana para definir los conjuntos de variables que tienen más de una opción de respuesta

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Una vez activada la ventana correspondiente, se procede a seleccionar las variables del conjunto y a agruparlas en una sola variable. Es fundamental que la codificación del instrumento sea correcta, lo que implica que, al ingresar las variables en SPSS, todas deben compartir los mismos códigos de respuesta para cada pregunta. Este proceso es esencial para crear una nueva variable múltiple que pueda analizar preguntas con opciones múltiples y facilitar así la interpretación de los datos y la toma de decisiones informadas.

Nota. Esta agrupación sirve para cualquier tipo de variable; sin embargo, existe otra opción para agrupar solo variables escalares.

En código, se tiene:

Tabla 4.1.1 Agrupación de variables de opción múltiple

```
* Ingreso de datos.
DATA LIST FREE / ID Razon1 Razon2 Razon3.
BEGIN DATA
1 1 2 0
2 1 0 3
3 1 0 0
4 0 2 3
END DATA.

* Asignar etiquetas a las variables y valores.
VARIABLE LABELS ID 'Identificador del encuestado'
            Razon1 'Razón de preferencia'
            Razon2 'Razón de preferencia'
            Razon3 'Razón de preferencia'.
VALUE LABELS Razon1 Razon2 Razon3
          1 'Precio'
          2 'Calidad'
          3 'Reconocimiento'.

* Definir el conjunto de respuestas múltiples.
MULT RESPONSE GROUPS=$Razones 'Razones de preferencia' (razon1 razon2 razon3
(1,3))
  /FREQUENCIES=$Razones.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

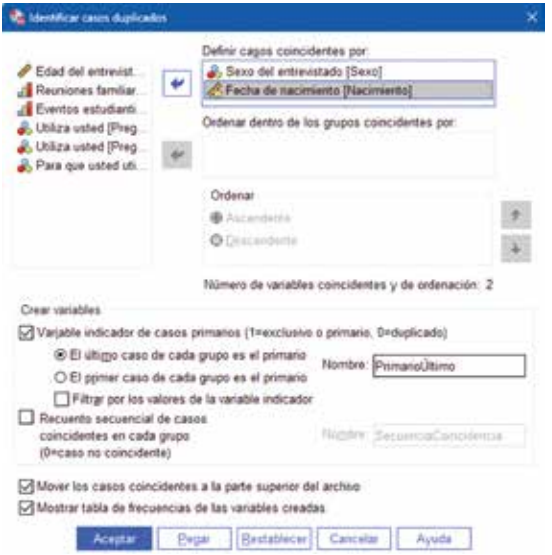
4.2 Identificar casos duplicados

Para asegurar la singularidad de las respuestas en un conjunto de datos, es fundamental utilizar la función de identificación de duplicados. Esta opción se encuentra en el menú de datos, bajo Identificar casos duplicados. Una vez allí, seleccionamos las variables que queremos analizar y ejecutamos la función.

Esta herramienta es especialmente útil en situaciones donde la repetición de observaciones puede generar problemas significativos. Por ejemplo, en un estudio de censos, cada persona debe ser registrada una sola vez para evitar sobreestimaciones en la población total. Otro caso es en la investigación médica, donde cada paciente debe tener un único registro para asegurar que los resultados de pruebas o tratamientos no se dupliquen, lo cual podría llevar a diagnósticos erróneos o a conclusiones incorrectas sobre la efectividad de un tratamiento. Del mismo modo, en el análisis financiero, cada transacción debe registrarse una sola vez para reflejar con precisión el estado financiero de una empresa.

La identificación de duplicados permite eliminar estas repeticiones, garantizando que cada dato en el análisis represente un caso único. Esto mejora la calidad y confiabilidad de las conclusiones obtenidas, asegurando que los resultados sean precisos y significativos.

Figura 4.2.1 Identificación de casos duplicados



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Como ejemplo, se puede considerar en este caso la variable sexo y fecha de nacimiento.

Figura 4.2.2 Indicador de cada último caso de coincidencia como primario

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Caso duplicado	1	10,0	10,0	10,0
	Caso primario	9	90,0	90,0	100,0
	Total	10	100,0	100,0	

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En la figura 4.2.2, se puede observar que, de un total de diez casos, se identificó que uno es un duplicado y nueve son casos primarios o únicos, lo que sugiere que, en el conjunto de datos analizado, el problema de duplicados es mínimo, debido a que únicamente el 10 % de las observaciones se repiten.

4.3 Identificar casos atípicos

Otro comando importante en la exploración de los datos es el Identificador de casos atípicos. Este nos ayuda a determinar los casos que tienen valores que se alejan de la media en proporciones mayores. Cuando el índice de anomalía es de 1 a 1,5, no se considera como anomalía, ya que su desviación es solo un poco superior a la media; sin embargo, los que pasan esos límites son considerados anómalos por presentar valores que al menos son el doble de la media.

Figura 4.3.1 Identificador de casos atípicos

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Si no existen valores atípicos, únicamente el SPSS da una advertencia; lo contrario ocurre cuando existen valores que pueden alterar el índice de anomalías. Por ejemplo:

Figura 4.3.2 Índice de anomalía

Caso	Índice de anomalía
3	5,000

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Un índice de anomalía de 5,000 indica que el caso 3 es significativamente diferente o inusual en comparación con los otros casos analizados. En un contexto de análisis de datos, este valor sugiere que el caso 3 presenta características que lo hacen destacar considerablemente, pudiendo ser un indicio de un valor atípico o una situación anómala que podría requerir una revisión más detallada.

Figura 4.3.3 Lista de ID de los homólogos de casos con anomalías

Caso	ID de homólogo	Tamaño de homólogo	Porcentaje de tamaño de homólogo
3	2	1	10,0%

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En la figura 4.3.3, se puede observar que, en el caso 3, se tiene un tamaño que es solo el 10 % del tamaño del caso con ID 2 (su homólogo). Este porcentaje sugiere que el caso 3 es significativamente más pequeño en comparación con su homólogo, lo que podría indicar que es menos representativo o de menor magnitud en el contexto del análisis realizado. Esta diferencia podría ser importante dependiendo del objetivo del análisis y podría justificar una investigación adicional para entender las razones detrás de esta discrepancia.

Figura 4.3.3 Lista de motivos de casos con anomalías

Razón: 1

Caso	Variable de razón	Impacto de variable	Valor de variable	Norma de variable
3	Edad	1,000	50	50,00

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

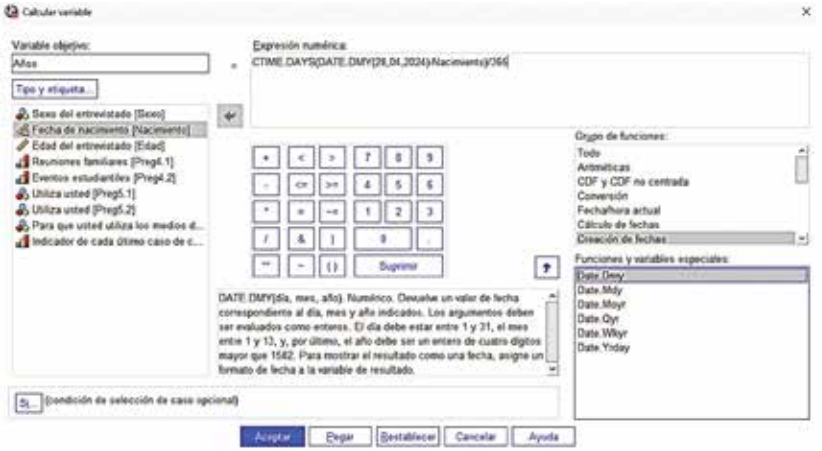
Según la figura 4.3.3, para el caso 3, la edad es cincuenta años; coincide perfectamente con la norma esperada de cincuenta años. El impacto de la variable es 1,000, lo que indica que la edad en este caso no está contribuyendo a ninguna anomalía. En otras palabras, la edad de cincuenta años es completamente normal y no es la causa de ninguna irregularidad en el caso 3. Es posible que, si hay una anomalía detectada en este caso, no sea debido a la variable de la edad, sino a otra variable que no se muestra en esta tabla.

4.4 Calcular variable

El cálculo de variables es importante cuando se está trabajando en bases de datos, por motivos de ordenamiento de información y/o por facilidad en la interpretación de los resultados. En Calcular variable, existen varias opciones para la generación y transformación de datos, tales como funciones aritméticas, funciones estadísticas, funciones de distribución, funciones de cadena, entre otras.

Por ejemplo, si tenemos una variable de fecha y queremos expresarla en años, lo que se puede realizar es seleccionar una de las opciones (extracción de tiempo – creación de fechas y variable) y generar la variable. El resultado final será una variable que nos da el dato en formato numérico.

Figura 4.4.1 Calcular nueva variable

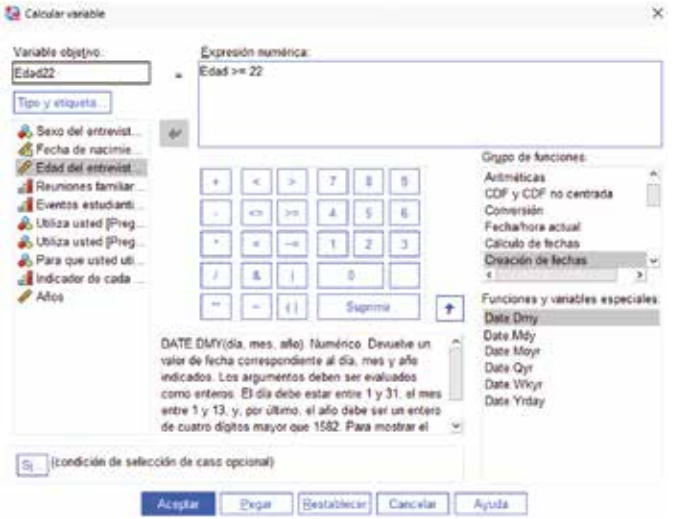


Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

De igual modo, si se necesita una variable condicional, lo que consideramos para el caso es determinar la variable sobre la cual se requiere restringir las observaciones. Por ejemplo, si queremos crear una variable donde se tenga el número de hijos que conforman una familia en personas mayores a cierta edad, se procede con la selección de la variable 1, en la cual se va a basar la prueba, y posteriormente se restringen las observaciones.

Este proceso es relevante cuando estamos filtrando las observaciones por presencia de ciertas características importantes dentro de los cálculos y/o modelos del objeto de estudio, lo que permite hacer uso exclusivamente de variables que cumplen las cualidades para describir cierto fenómeno.

Figura 4.4.2 Calcular nueva variable



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Para concluir con el cálculo de variables, se debe considerar que, para realizar esta operación en el SPSS, se pueden seguir los siguientes pasos:

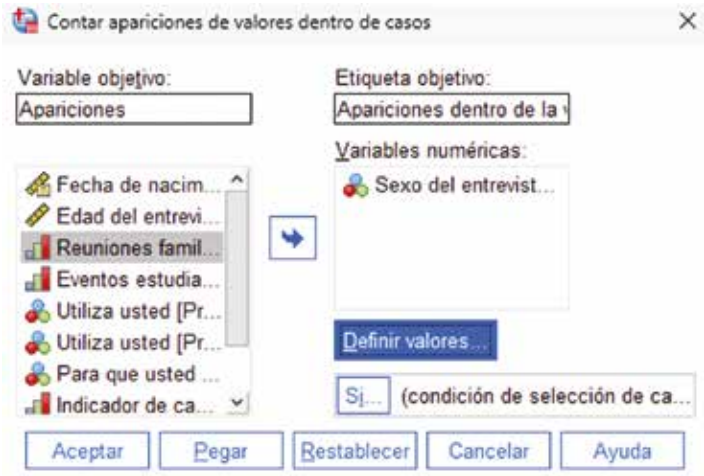
- Seleccione en el menú Transformar > Calcular variable.
- Escriba el nombre de la variable de destino.
- Para crear una expresión, se puede escribir la función directamente en dicho campo.
- Si los valores tienen decimales, debe utilizarse una coma como indicador decimal.

- Las variables en cadena consideran el tipo y etiqueta para especificar la estructura de los datos.
- Si el resultado de una expresión condicional es verdadero, se incluirá el caso en el subconjunto seleccionado; de lo contrario, no se incluirá
- Para las condiciones, la expresión debe utilizar al menos uno de los seis operadores de relación presentes en la calculadora.
- Dentro de las condicionales, es posible incluir nombres de variable, constantes, operadores, funciones, variables lógicas y operadores de relación

4.5 Contar valores dentro de los casos

Esta utilidad crea una variable que, para cada caso, realiza el conteo de apariciones del mismo valor, o valores, en una lista de variables. Por ejemplo, si se quiere conocer cuantas veces han calificado en un determinado rango a los servicios, se seleccionan las variables de calificación y finalmente nos indicará el número de respuestas con el valor seleccionado. Para ello se va a Transformar y luego se selecciona Contar valores dentro de los casos (véase figura 4.5.1) y finalmente se define el valor que se requiere contar.

Figura 4.5.1 Contador de valores



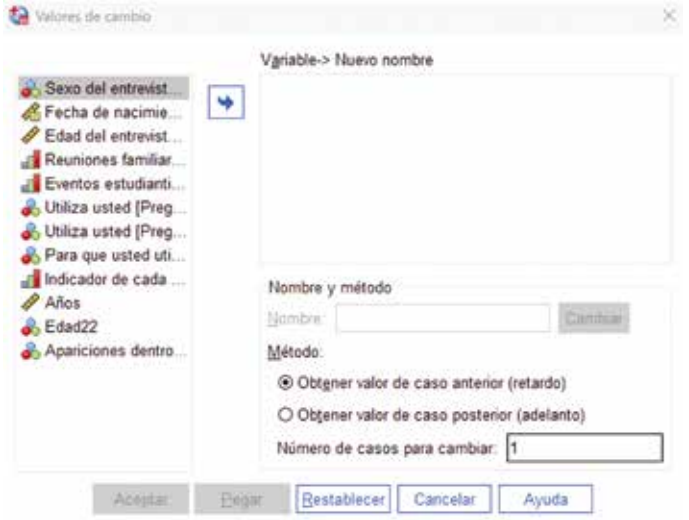
Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Finalmente, se puede realizar una tabla de frecuencias para analizar los conteos de los casos que deseemos analizar.

4.6 Valores de cambio

Los valores de cambio son herramientas cruciales en el análisis econométrico, ya que permiten generar retardos o adelantos en las series temporales, lo que facilita la captura de efectos dinámicos en los modelos. Estos valores son especialmente útiles para identificar relaciones de causa y efecto, así como para mejorar la precisión en las predicciones. Para acceder a esta funcionalidad, se debe ir a la opción Transformar en el software y seleccionar Valores de cambio. En la figura 4.6.1, se muestra la ventana que se despliega al seleccionar esta opción, ilustrando cómo se pueden configurar los parámetros para aplicar retardos o adelantos en los datos.

Figura 4.6.1 Modificador de valores de las variables



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

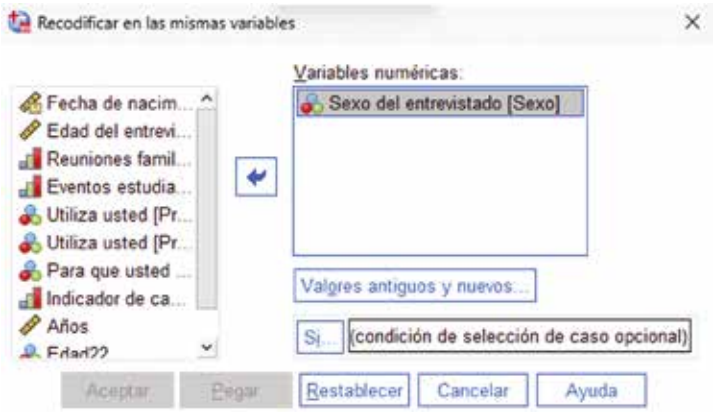
4.7 Recodificar variables

La recodificación permite ajustar los valores de los datos para facilitar la agrupación o combinación de categorías. Esta técnica puede aplicarse directamente sobre las variables originales o utilizarse para crear nuevas variables derivadas de los valores recodificados.

4.7.1 Recodificar en las mismas variables

El cuadro de diálogo Recodificar en las mismas variables se encuentra en la pestaña Transformar. Con esta opción, es posible reasignar los valores de las variables existentes o consolidar rangos de valores en nuevas categorías. Cabe destacar que las variables numéricas y de texto no pueden recodificarse de manera conjunta.

Figura 4.7.1 Recodificación en las mismas variables



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Para recodificar los valores de una variable

- Seleccione en los menús Transformar > Recodificar en las mismas variables.
- Seleccione las variables que desee recodificar. Si selecciona múltiples variables, todas deberán ser del mismo tipo.
- Pulse en Valores antiguos y nuevos y especifique cómo deben recodificarse los valores.

Figura 4.7.2 Ventana recodificación de valores en la misma variable

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

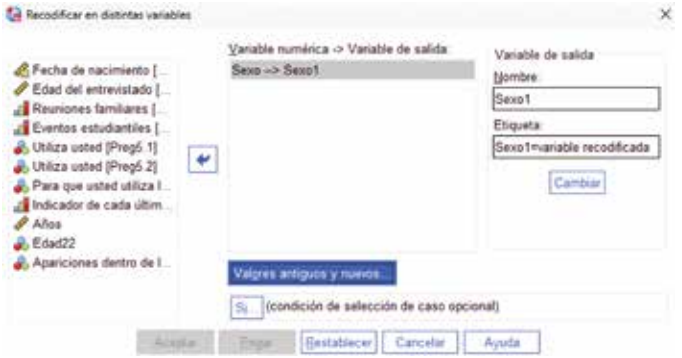
4.7.2 Recodificar en distintas variables

La opción Recodificar en distintas variables permite modificar los valores de las variables actuales o combinar rangos de valores en categorías para generar una nueva variable.

Este comando ofrece las siguientes funcionalidades:

- Permite trabajar tanto con variables numéricas como de texto.
- Admite la conversión de variables numéricas a texto y viceversa.
- Si se seleccionan varias variables, todas deben ser del mismo tipo, ya que no es posible recodificar simultáneamente variables numéricas y de texto.

Figura 4.7.3 Recodificación en distintas variables



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Para recodificar los valores de una variable en una nueva variable.

- Acceda al menú Transformar y seleccione la opción Recodificar en distintas variables.
- Escoja las variables que desea recodificar. Tenga en cuenta que, si selecciona varias variables, todas deben ser del mismo tipo (numéricas o de texto).
- Asigne un nombre para cada nueva variable en el campo correspondiente y haga clic en Cambiar.
- Haga clic en Valores antiguos y nuevos para definir las reglas de recodificación.
- Opcionalmente, puede establecer un subconjunto de casos para aplicar la recodificación. Esto se configura en el cuadro de diálogo relacionado con la opción Si los casos, similar a la utilizada en Contar apariciones.

4.7.3 Recodificación automática

Este comando en SPSS permite recodificar automáticamente variables que ya han sido previamente codificadas, reasignándoles nuevos códigos según sea necesario. Además, puede asignar códigos a variables que no hayan sido codificadas previamente, lo que es particularmente útil para variables como los años, donde es posible que desees transformar una secuencia continua en categorías o rangos específicos.

Esta funcionalidad es esencial cuando se trabaja con datos categóricos o cuando se necesita estandarizar la codificación de variables para su análisis. De esta manera, el comando facilita la manipulación y transformación de datos, asegurando consistencia y precisión en el proceso analítico.

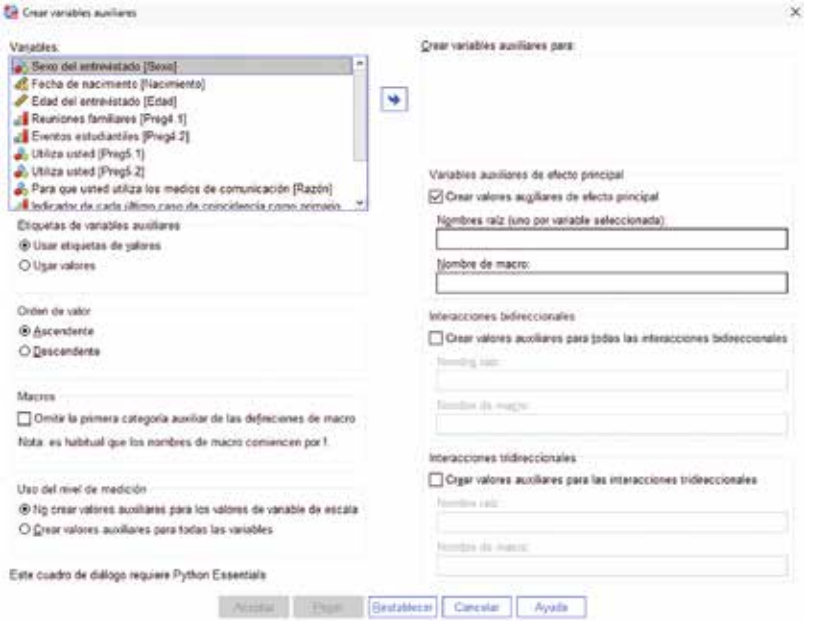
Figura 4.7.4 Ventana de vista de variables con información ingresada de datos obtenidos de la encuesta modelo

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

4.8 Crear variables auxiliares

Las variables auxiliares desempeñan un papel crucial en los análisis de regresión, especialmente cuando es necesario aislar o controlar el impacto de una observación específica sobre la variable dependiente. Estas variables permiten descomponer el efecto de diferentes factores y asegurar que el análisis se enfoque adecuadamente en las relaciones clave que se están estudiando. Una estrategia común en este contexto es la creación de variables diferenciadas dicotómicas, conocidas como variables *dummy*. Las variables *dummy* toman valores de 0 o 1, indicando la presencia o ausencia de una característica particular en una observación. Por ejemplo, si se desea evaluar el impacto de una política específica implementada en un solo año dentro de un análisis de series temporales, se puede crear una variable *dummy* que tome el valor 1 para ese año y 0 para los demás. Esto permite estimar de manera precisa cómo ese evento afecta la variable dependiente, separando su influencia de otros factores en el modelo.

Figura 4.8.1 Variables auxiliares



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Tabla 4.8.1Creación de variables dicotómicas

```
* Definir los datos.
DATA LIST FREE / año (F4.0) ventas (F8.2).
BEGIN DATA
2020 110
2021 115
2022 150
2023 130
END DATA.

* Crear una variable dummy para el año 2022.
COMPUTE promocion_2022 = (año = 2022).
EXECUTE.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

4.9 Agrupación visual

La agrupación visual es relevante cuando se requiere elaborar rangos. Por ejemplo, si tenemos información de edades, las tablas de frecuencias rara vez nos ayudan resumiendo la información, por el hecho de que existe gran cantidad de casos, lo cual dificultaría la interpretación de los resultados obtenidos. Para solucionar ese problema, se utiliza la opción Agrupación visual con la finalidad de poder organizar de mejor manera la información y con ello facilitar su interpretación.

Figura 4.9.1 Agrupación visual

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

4.10 Agrupación óptima

La agrupación óptima, al igual que la agrupación visual, genera rangos de datos sobre la base de una variable seleccionada. Sin embargo, a pesar de mantener la misma o similar funcionalidad, esta se caracteriza porque, a los rangos, se los puede crear considerando además una variable extra que conlleve una alta correlación entre sí.

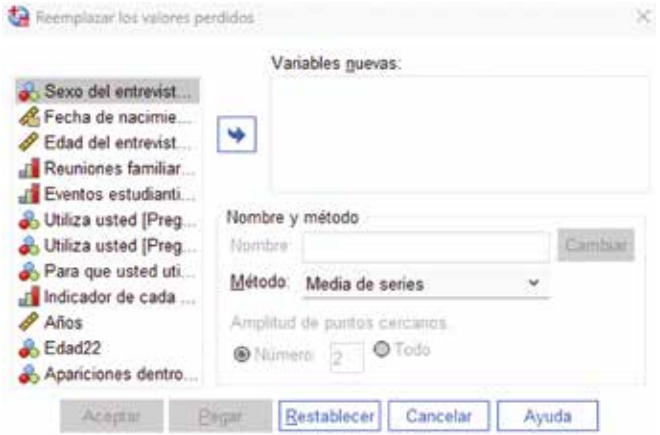
Figura 4.10.1 Agrupación óptima

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

4.11 Reemplazar valores perdidos

Cuando existen observaciones vacías, este comando permite llenarlo con aproximaciones sobre la base de los valores medios, moda, interpolación y tendencia lineal en el punto. Esta opción se encuentra en Transformar y luego en la opción Reemplazar valores perdidos.

Figura 4.11.1 Reemplazo de valores perdidos



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Capítulo 5

Medidas numéricas

Medidas numéricas

Introducción

Las medidas numéricas son herramientas fundamentales en el análisis estadístico, ya que permiten describir y resumir las características principales de un conjunto de datos. Este capítulo introduce las principales medidas numéricas usadas en la estadística descriptiva, proporcionando una base sólida para la comprensión de los datos y la toma de decisiones informadas.

El capítulo comienza con una explicación de los valores percentiles, que dividen un conjunto de datos en partes iguales y permiten identificar la posición relativa de un valor dentro de un conjunto de datos. Luego, se discuten las medidas de tendencia central, como la media, mediana y moda, que son esenciales para describir el valor típico de un conjunto de datos.

A continuación, se exploran las medidas de dispersión, que incluyen la desviación estándar, la varianza, y el rango, las cuales proporcionan información sobre la variabilidad de los datos. También se analizan los valores extremos, como el mínimo y el máximo, y se introduce el concepto de error estándar, que mide la precisión de la media muestral como estimación de la media poblacional.

Finalmente, se aborda la distribución de los datos, enfocándose en la asimetría y la curtosis, que describen la forma y la concentración de los datos alrededor de la media. Este capítulo es crucial para desarrollar una compren-

sión completa de cómo describir y analizar datos numéricos, y para preparar el terreno para análisis estadísticos más avanzados.

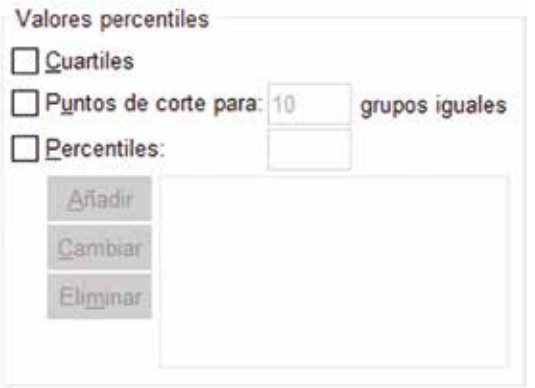
Objetivos del capítulo

- Definir medidas de tendencia central (media, mediana, moda) y de dispersión (desviación estándar, varianza).
- Calcular y comparar las medidas de tendencia central y dispersión para un conjunto de datos.
- Interpretar los resultados de las medidas numéricas y su significado en el contexto de un análisis estadístico.

5.1 Valores percentiles

Los valores percentiles son una medida estadística que se utiliza para clasificar y comparar datos en un conjunto de observaciones. Los percentiles dividen los datos ordenados en cien partes iguales, donde cada parte representa un percentil. Por ejemplo, el percentil 50 (también conocido como la mediana) divide los datos en dos partes iguales, con el 50 % de los datos por debajo de ese valor y el 50 % por encima.

Figura 5.1.1 Valores percentiles



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Los valores percentiles son útiles para entender la distribución de datos y para comparar la posición relativa de un valor dentro de un conjunto de datos. Por ejemplo, si alguien obtiene un percentil 75 en un examen, significa que su rendimiento está por encima del 75 % de los demás participantes en ese examen.

Este tipo de análisis es ampliamente utilizado en administración y economía, porque permite evaluar el rendimiento relativo y segmentar datos para identificar tendencias, lo que optimiza decisiones estratégicas en análisis de mercado y segmentación (Levin y Rubin, 2010).

En el SPSS, los percentiles pueden ser calculados de tres diferentes formas, según lo que se necesite mostrar. Por ejemplo, considerando la variable edad de la base de datos tomada como referencia (bankloan), se puede observar que, al seleccionar cuartiles, se parte la información en cuatro partes iguales, dando como resultado que el 25 % de las personas están por debajo de los veintinueve años o, el 75 % está por encima de los veintinueve años.

Figura 5.1.2 Percentiles edad en años (base Bankloan)

Edad en años

N	Válido	850
	Perdidos	0
Percentiles	25	29,00
	50	34,00
	75	41,00

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En resumen, los valores percentiles proporcionan una forma de expresar la posición relativa de un valor dentro de un conjunto de datos, lo que facilita la interpretación y comparación de datos en diferentes contextos.

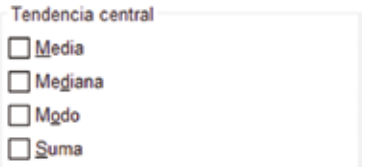
5.2 Tendencia central

La tendencia central se refiere a medidas estadísticas que representan el centro o la ubicación central de un conjunto de datos. Estas medidas son utilizadas para comprender la distribución de los datos y para describir la posición típica o más representativa de los mismos. Entre las medidas de tendencia central más utilizadas se encuentran la media aritmética, la mediana

y la moda. La media aritmética se calcula sumando todos los valores y dividiendo el resultado entre la cantidad total de datos. La mediana corresponde al valor central en un conjunto de datos organizados de menor a mayor. Por su parte, la moda es el valor que se repite con mayor frecuencia dentro del conjunto de datos.

Estas medidas son esenciales en economía y administración, ya que facilitan el análisis y la toma de decisiones al identificar patrones o valores típicos que resumen grandes conjuntos de datos, lo que permite detectar tendencias relevantes (Levin y Rubin, 2010; Newbold et al., 2008).

Figura 5.2.1 Medidas de tendencia central



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.2.1 Media

La media, también conocida como promedio, es un concepto estadístico que se utiliza para representar un valor típico en un conjunto de datos. Se calcula sumando todos los valores en el conjunto y dividiendo esta suma por el número total de valores. La media es una medida de tendencia central que proporciona información sobre el valor central o típico de un conjunto de datos, aunque puede verse afectada por valores extremos o atípicos en el conjunto. Es una herramienta fundamental en el análisis estadístico para entender patrones y tendencias en los datos. En la base tomada, se tiene que la edad media es de 35,03 años.

Figura 5.2.2 Media de la edad en años (base de datos Bankloan)

Edad en años

N	Válido	850
	Perdidos	0
Media		35,03

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Este tipo de promedio se usa para sintetizar grandes volúmenes de datos y apoyar la toma de decisiones en contextos de alta variabilidad (Newbold et al., 2008).

5.2.2 Mediana

La mediana es un concepto estadístico que se utiliza para representar el valor central de un conjunto de datos ordenados de menor a mayor. Para calcular la mediana, se deben seguir estos pasos:

- Ordenar los datos de menor a mayor.
- Si el número de datos es impar, la mediana es el valor que ocupa la posición central en la lista ordenada.
- Si el número de datos es par, la mediana se calcula promediando los dos valores centrales.

La mediana no se ve afectada por valores atípicos o muy alejados en los datos, a diferencia de la media aritmética. Esto la hace útil para describir la distribución de un conjunto de datos cuando hay valores atípicos como, por ejemplo, la mediana de la base de referencia es treinta y cuatro años. Esta información es válida para determinar los sesgos que se analizarán en apartados posteriores.

Figura 5.2.3 Mediana de la edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Mediana		34,00

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Así, la mediana permite identificar la concentración central real de los datos, especialmente en distribuciones asimétricas, optimizando su interpretación y utilidad en decisiones donde el equilibrio entre valores extremos y típicos es esencial (Anderson et al., 2019).

5.2.3 Moda (modo)

El concepto de moda se refiere al valor que ocurre con mayor frecuencia en un conjunto de datos. Es el dato que se repite más veces y, por lo tanto,

representa la categoría más común o popular en la muestra. La moda es una medida de tendencia central que puede ser útil para entender la distribución de los datos e identificar patrones o preferencias en una población o conjunto de observaciones.

Figura 5.2.4 Moda de la edad en años (base de datos Bankloan)

Edad en años

N	Válido	850
	Perdidos	0
Moda		29

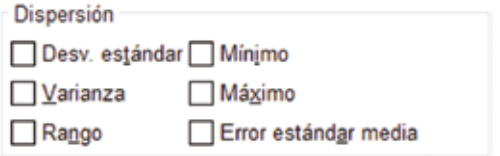
Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Además es particularmente efectiva en variables categóricas o cualitativas, donde otras medidas como la media y la mediana no son aplicables. En administración y economía, la moda ayuda a detectar las preferencias predominantes o productos más solicitados, siendo fundamental para análisis de demanda y segmentación de mercado (Lind, Marchal, y Wathen, 2020).

5.3 Dispersión

La dispersión se refiere a la medida de la variabilidad o la extensión en la distribución de un conjunto de datos. En otras palabras, indica qué tan alejados están los datos individuales de un conjunto de valores central, como la media o la mediana. Una mayor dispersión sugiere que los datos están más ampliamente distribuidos alrededor de la medida central, mientras que una menor dispersión indica que los datos están más cercanos entre sí y a la medida central. La dispersión es una medida importante en estadística para comprender la distribución y la variabilidad de los datos, lo que puede tener implicaciones significativas en análisis y toma de decisiones.

Figura 5.3.1 Medidas de dispersión



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En campos como la economía y administración, la dispersión permite evaluar el riesgo y la estabilidad de métricas clave, y ayuda a identificar tanto oportunidades como áreas de variabilidad significativa en el rendimiento o el mercado (Anderson et al., 2019).

5.3.1 Desviación estándar

La desviación estándar indica la dispersión o variabilidad de un conjunto de datos con respecto a su media (Anderson et al., 2019). En otras palabras, la desviación estándar muestra cuánto se alejan los valores individuales de un conjunto de datos de la media de ese conjunto. Un valor elevado de la desviación estándar refleja una mayor dispersión de los datos respecto a la media, mientras que un valor bajo indica que los datos están más concentrados en torno a la media. Es una herramienta importante en el análisis de datos para comprender la consistencia o la variabilidad de un conjunto de observaciones.

Figura 5.3.2 Desviación estándar de la edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Desv. Desviación		8,041

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.3.2 Varianza

La varianza es una medida estadística que representa la dispersión o variabilidad de un conjunto de datos en relación con su media; es decir, la varianza indica qué tan alejados están los valores individuales de los datos del valor promedio (Lind et al., 2020).

Una varianza alta significa que los datos están muy dispersos alrededor de la media, mientras que una varianza baja indica que los datos están más cerca de la media y son menos dispersos. La fórmula para calcular la varianza es la suma de los cuadrados de las diferencias entre cada dato y la media, dividida por el número total de datos. La varianza es una medida importante en estadística y se utiliza en muchos contextos para comprender la distribución y la variabilidad de los datos.

Figura 5.3.3 Varianza de la edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Varianza		64,665

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.3.3 Rango

Es la diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Es una medida sencilla, pero importante. Proporciona información sobre la variabilidad de los datos. Cuanto mayor sea el rango, mayor será la dispersión, lo que indica una mayor variabilidad o amplitud en la distribución de los valores (Lind et al., 2020). Por otro lado, un rango más pequeño sugiere una menor variabilidad y una concentración de valores cercanos entre sí. El rango es útil para tener una idea general de la dispersión de los datos y es especialmente útil cuando se trabaja con conjuntos de datos pequeños o cuando se desea una medida rápida de la variabilidad de los datos.

Figura 5.3.4 Rango de edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Rango		36

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.3.4 Mínimo

El concepto de mínimo se refiere al valor más pequeño dentro de un conjunto de datos. Es un término fundamental que se utiliza para analizar y describir la distribución de los datos, ya que proporciona información importante sobre el rango de valores que se están observando. El mínimo es útil en diversos contextos, como calcular la amplitud de un conjunto de datos (la diferencia entre el valor máximo y el mínimo), identificar valores atípicos o extremos, y en general para comprender la variabilidad y la dispersión de los datos en un conjunto de observaciones.

Figura 5.3.5 Mínimo de edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Mínimo		20

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.3.5 Máximo

Se refiere al valor más grande en un conjunto de datos. Es un concepto fundamental que se utiliza para analizar la distribución y la variabilidad de los datos. El máximo puede ser útil para identificar valores atípicos o destacados en un conjunto de datos, así como para determinar la capacidad máxima en un proceso o sistema.

Figura 5.3.6 Máximo de edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Máximo		56

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.3.6 Error estándar medio

El error estándar medio (SEM), en estadística, es una medida de dispersión que indica la precisión de la media de una muestra en relación con la media de la población. Se calcula dividiendo la desviación estándar de la muestra por la raíz cuadrada del tamaño de la muestra. El SEM es útil para determinar cuán confiable es la estimación de la media de la población basada en una muestra específica. Un SEM más pequeño indica que la media de la muestra es una estimación más precisa de la media de la población, mientras que un SEM más grande indica una estimación menos precisa. Esta medida permite realizar inferencias confiables y evaluar la representatividad de la muestra en relación con la población general (Anderson et al., 2019).

Figura 5.3.7 Error estándar de edad en años (base de datos Bankloan)

Edad en años

N	Válido	850
	Perdidos	0
Error estándar de la media		,276

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

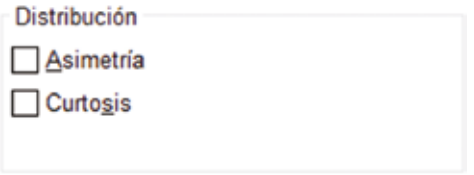
5.4 Distribución

La distribución en estadística se refiere a la manera en que los datos están dispersos o distribuidos en una muestra o población. En otras palabras, describe la forma en que los valores de una variable se distribuyen en un conjunto de datos.

Existen diferentes tipos de distribuciones, como la normal (o gaussiana), la distribución uniforme, la distribución exponencial, entre otras. Cada una de estas distribuciones tiene características específicas que ayudan a comprender mejor cómo se comportan los datos y qué patrones o tendencias pueden existir dentro de ellos. La distribución también es fundamental para realizar inferencias estadísticas y tomar decisiones basadas en los datos disponibles.

Conocer la distribución permite evaluar la variabilidad y simetría de los datos y aplicar modelos adecuados en el análisis estadístico (Anderson et al., 2019).

Figura 5.4.1 Media de distribución



Nota. Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

5.4.1 Asimetría

La asimetría se refiere a la medida de la falta de simetría en la distribución de datos. Una distribución simétrica es aquella en la que los valores están igualmente distribuidos a ambos lados de la media, lo que significa que la mitad de los datos están por encima de la media y la otra mitad está por debajo de ella, y la forma de la distribución es similar en ambos lados de la media.

La asimetría ocurre cuando esta simetría está ausente, lo que significa que la distribución no es igual en ambos lados de la media. Pueden haber tres tipos de asimetría:

Asimetría positiva (sesgo a la derecha): la cola derecha de la distribución es más larga que la cola izquierda. Esto indica que hay valores extremadamente altos que están alejados de la media, lo que provoca que la distribución se extienda más hacia la derecha.

Asimetría negativa (sesgo a la izquierda): la cola izquierda de la distribución es más larga que la cola derecha. Esto sugiere que hay valores extremadamente bajos que están alejados de la media, lo que provoca que la distribución se extienda más hacia la izquierda.

Asimetría cero (simetría): en este caso, la distribución es simétrica, lo que significa que no hay sesgo hacia la izquierda o hacia la derecha, y los valores están distribuidos de manera uniforme alrededor de la media.

Figura 5.4.2 Asimetría y error estándar de asimetría de la edad en años (base de datos Bankloan)

Edad en años

N	Válido	850
	Perdidos	0
Asimetría		,335
Error estándar de asimetría		,084

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

La asimetría es importante en estadística porque puede afectar la interpretación de los datos y la elección de las pruebas estadísticas adecuadas para analizarlos. Entender la asimetría es crucial para interpretar correctamente la distribución de datos y hacer inferencias precisas en análisis estadístico, ya que este parámetro afecta la elección de medidas de tendencia central y dispersión en distribuciones no simétricas (Levin y Rubin, 2010).

5.4.2 Curtosis

Se refiere a la medida de la «picudez» de la distribución de datos. Una distribución leptocúrtica (con curtosis positiva) tiene picos más altos y colas más pesadas, lo que indica que los datos tienen valores extremos más pronunciados. Por otro lado, una distribución platicúrtica (con curtosis negativa) tiene picos más bajos y colas menos pesadas, lo que indica que los datos tienen valores extremos menos pronunciados.

Figura 5.4.3 Curtosis y error estándar de curtosis de la edad en años (base de datos Bankloan)

Edad en años		
N	Válido	850
	Perdidos	0
Curtosis		-,658
Error estándar de curtosis		,168

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En resumen, la asimetría describe la falta de simetría en la distribución de los datos alrededor de la media, mientras que la curtosis describe la «picudez» de la distribución y la presencia de valores extremos. Ambas medidas son útiles para entender la forma y la naturaleza de los datos en un conjunto de datos.

Capítulo 6

Distribuciones de frecuencias

Distribuciones de frecuencias

Introducción

Las distribuciones de frecuencias son una herramienta clave en la estadística descriptiva, ya que permiten organizar y resumir grandes volúmenes de datos de manera clara y comprensible. Este capítulo introduce los conceptos básicos relacionados con las distribuciones de frecuencias, proporcionando una guía completa para su creación y análisis.

El capítulo comienza explicando qué son los datos en bruto y cómo pueden ser organizados en una distribución de frecuencias. Se destaca la importancia de ordenar los datos de manera sistemática, lo que facilita su análisis y la identificación de patrones. A continuación, se describen las distribuciones de frecuencia para variables cualitativas, que permiten resumir la frecuencia de las categorías, y para variables cuantitativas, que agrupan los datos en intervalos.

Además, se explora la distribución de frecuencia relativa, que expresa las frecuencias en términos porcentuales, facilitando la comparación entre diferentes conjuntos de datos. Este enfoque es especialmente útil cuando se analizan muestras de diferentes tamaños.

Este capítulo es esencial para aquellos que necesitan analizar y presentar datos de manera efectiva, ya que proporciona las herramientas necesarias para crear distribuciones de frecuencias claras y significativas. Al final del capítulo, el lector estará preparado para organizar y analizar datos de manera sistemática, sentando las bases para análisis más detallados y complejos.

Objetivos del capítulo

- Enumerar los pasos para crear una distribución de frecuencia para variables cualitativas y cuantitativas.
- Organizar un conjunto de datos en una distribución de frecuencia adecuada para el análisis.
- Analizar una distribución de frecuencia para identificar patrones o tendencias en los datos.

6.1 Datos en bruto

Los datos en bruto se refieren a la información recopilada que aún no ha sido procesada, organizada o analizada de ninguna manera. Estos datos son la materia prima de cualquier análisis estadístico o investigación. Pueden presentarse en diferentes formas, como números, texto, imágenes, sonidos, etc. Los datos en bruto son la base sobre la cual se aplican técnicas y herramientas para extraer información significativa y tomar decisiones informadas (base de datos tomada de Bankloan presentada en el SPSS como base de prueba). Este tipo de datos, al estar sin procesar, requiere de depuración y organización inicial para identificar patrones y tendencias relevantes, proporcionando un fundamento sólido para el análisis posterior (Levin y Rubin, 2010).

Figura 6.1.1 Base de datos de muestra (Bankloan)

	edad	educ	empleo	direccion	ingresos	deudaing	deudacred	deudaotro	impago	morapred1	morapred2	morapred3
1	41	Super...	17	12	176,00	9,30	11,36	5,01	Si	,80839	,78964	21304
2	27	No co...	10	6	31,00	17,30	1,36	4,00	No	,19830	,12845	43690
3	40	No co...	15	14	55,00	5,50	,86	2,17	No	,01054	,00299	14102
4	41	No co...	15	14	120,00	2,90	2,66	,82	No	,02214	,01027	10442
5	24	Título...	2	0	28,00	17,30	1,79	3,06	Si	,78159	,73798	43690
6	41	Título...	5	5	25,00	10,20	,39	2,16	No	,21671	,32819	23368
7	39	No co...	20	9	67,00	30,60	3,83	16,67	No	,18096	,17926	81709
8	43	No co...	12	11	38,00	3,60	,13	1,24	No	,01471	,01957	11336
9	24	No co...	3	4	19,00	24,40	1,36	3,28	Si	,74884	,61944	66390
10	36	No co...	0	13	25,00	19,70	2,78	2,15	No	,81506	,75723	51553
11	27	No co...	0	1	16,00	1,70	,18	,09	No	,35031	,51951	09056
12	26	No co...	4	0	23,00	5,20	,25	,94	No	,23906	,21902	13631
13	52	No co...	24	14	64,00	10,00	3,93	2,47	No	,06979	,06628	22890
14	37	No co...	6	9	25,00	16,30	1,72	3,01	No	,36449	,34947	40484
15	48	No co...	22	15	100,00	9,10	3,70	5,40	No	,01187	,00771	20866
16	36	Título...	9	6	49,00	8,60	,82	3,40	Si	,09678	,11384	19801
17	36	Título...	13	6	41,00	16,40	2,92	3,81	Si	,21206	,17502	40801
18	43	No co...	23	19	72,00	7,60	1,18	4,29	No	,00140	,00056	17793
19	39	No co...	6	9	61,00	5,70	,56	2,91	No	,10415	,09273	14424
20	41	Super...	0	21	26,00	1,70	,10	,34	No	,09192	,08691	09056

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

La importancia de los datos en bruto radica en su potencial para revelar patrones, tendencias y relaciones que pueden ser útiles para comprender fenómenos, identificar problemas, desarrollar hipótesis y tomar medidas correctivas o estratégicas. Sin embargo, es fundamental comprender que los datos en bruto por sí solos no proporcionan conocimiento; requieren un proceso de análisis y transformación para convertirse en información relevante y útil.

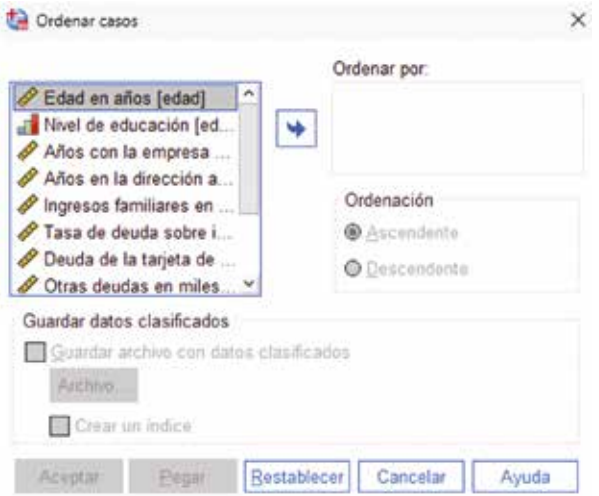
En el contexto de la ciencia de datos, el manejo de datos en bruto implica actividades como la limpieza de datos (eliminación de errores y valores atípicos), la transformación de datos (normalización, codificación, etc.) y la preparación de datos para su análisis posterior. Este proceso es crucial para garantizar la calidad y la integridad de los resultados obtenidos a partir de los datos. Mediante este tratamiento, los datos en bruto se convierten en insumos confiables y precisos, fundamentales para realizar análisis estadísticos sólidos y obtener conclusiones válidas (Levin y Rubin, 2010; Newbold et al., 2008).

6.2 Ordenamientos

El orden se refiere al proceso de organizar elementos o datos de acuerdo con algún criterio específico, como el tamaño, la importancia, la fecha, el orden alfabético, entre otros. Considerando el contexto de la estadística, el orden es útil para clasificar datos y facilitar su análisis y comprensión. Por ejemplo, en un conjunto de datos de ventas de productos, se pueden ordenar los productos de mayor a menor según sus ingresos generados. Esto permite identificar rápidamente los productos más rentables o populares.

Para ordenar los datos en SPSS, hay que ubicarse en la pestaña de Datos y se hace clic en Ordenar casos. En este aspecto, se considera la variable que vamos a tomar para que los casos/observaciones sean ordenados de forma ascendente o descendente, permitiendo así una mejor visualización y comprensión de los patrones dentro de un conjunto de datos crudos, base esencial para la exploración y análisis de tendencias en estudios cuantitativos (Anderson et al., 2019).

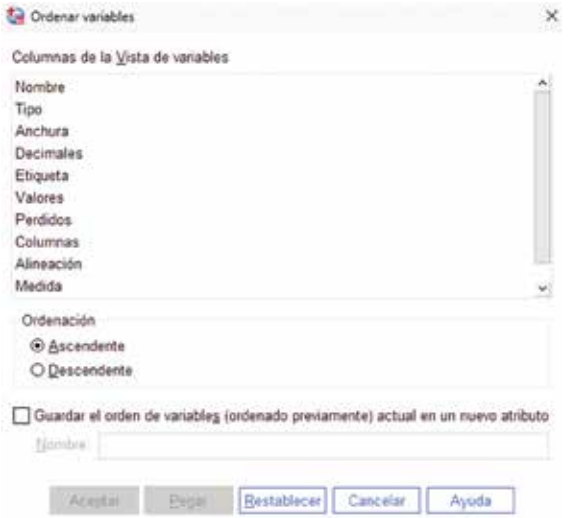
Figura 6.2.1 Ordenar casos



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

De la misma manera, se pueden ordenar las variables, para ello es necesario ubicarse en la pestaña de Datos y se hace clic en Ordenar variables. Aquí es posible ordenar las variables por Nombre, Tipo, Anchura, etc.

Figura 6.2.2 Ordenar variables



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

En resumen, el orden es un proceso clave para organizar y estructurar datos de manera que sean más accesibles, comprensibles y útiles para su posterior análisis o utilización.

6.3 Distribución de frecuencia para variables cualitativas

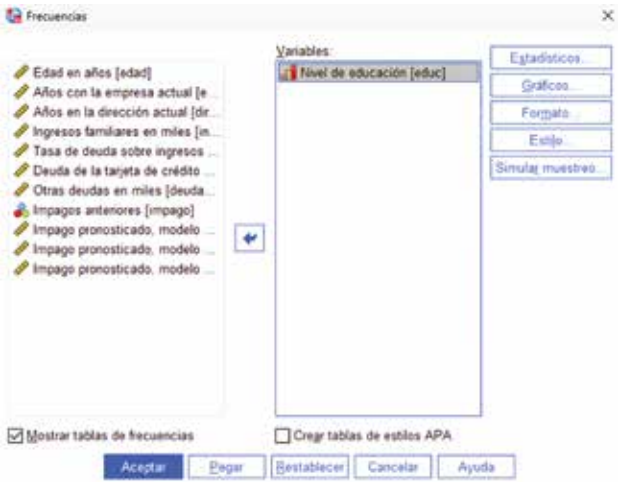
La distribución de frecuencia para variables cualitativas se refiere a la representación de la frecuencia con la que ocurren diferentes categorías o clases en un conjunto de datos. Las variables cualitativas, también conocidas como variables categóricas, son aquellas que representan características no numéricas o cualidades, como el color, la categoría, el tipo de producto, la opinión, etc.

Para crear una distribución de frecuencia para variables cualitativas, primero se identifican todas las categorías presentes en los datos. Luego, se cuenta la frecuencia de cada categoría, es decir, cuántas veces aparece en el conjunto de datos. Esta información se puede presentar en forma de tablas de frecuencia o gráficos, como histogramas de barras o gráficos circulares.

La distribución de frecuencia para variables cualitativas es útil para visualizar la distribución de diferentes categorías y entender la composición o la prevalencia de cada una en el conjunto de datos. Esto puede ser útil para detectar patrones, tendencias o preferencias dentro de una población o muestra.

En SPSS, se la obtiene al ingresar a Analizar, luego haciendo clic en Estadísticos descriptivos y finalmente en Frecuencias. Este análisis es fundamental para interpretar el comportamiento de categorías dentro de los datos, para facilitar la identificación de patrones en contextos de estudio específicos (Spiegel y Stephens, 2009).

Figura 6.3.1 Interfaz para seleccionar variables que se van a analizar



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

El resultado obtenido se puede visualizar en la tabla. Tal como se observa, la información hace referencia a:

- Frecuencia: valor numérico de las observaciones.
- Porcentaje: valor relativo expresado en porcentajes de las observaciones obtenidas.
- Porcentaje válido: es el porcentaje obtenido sin tomar en consideración los valores perdidos por cualquier motivo.
- Porcentaje acumulado: es el porcentaje que se obtiene de la suma de los porcentajes válidos.

Figura 6.3.2 Distribución de frecuencias para variables cualitativas

Nivel de educación

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	No completó el bachillerato	460	54,1	54,1	54,1
	Título de Bachiller	235	27,6	27,6	81,8
	Superiores iniciados	101	11,9	11,9	93,6
	Título Superior	49	5,8	5,8	99,4
	Título de Post-grado	5	,6	,6	100,0
	Total	850	100,0	100,0	

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

6.4 Distribución de frecuencia para variables cuantitativas

La distribución de frecuencia para variables cuantitativas se refiere a la representación de la frecuencia con la que se presentan diferentes valores numéricos en un conjunto de datos. Las variables cuantitativas son aquellas que representan cantidades numéricas o medidas, como la edad, el peso, la altura, el ingreso, etc.

Para crear una distribución de frecuencia para variables cuantitativas, primero se identifican los valores únicos presentes en los datos. Luego, se cuenta la frecuencia de cada valor, es decir, cuántas veces aparece en el conjunto de datos. Esta información se puede presentar en forma de tablas de frecuencia, histogramas, polígonos de frecuencia u otros gráficos adecuados (Anderson et al., 2019).

La distribución de frecuencia para variables cuantitativas es útil para comprender la distribución de los valores numéricos en un conjunto de datos y analizar su dispersión, centralidad y forma. Esto puede ser útil para identificar patrones, outliers (valores atípicos), calcular medidas de tendencia central y dispersión, y realizar análisis estadísticos más avanzados.

Por ejemplo, en un estudio sobre edades de una población, la distribución de frecuencia para la variable «edad» mostraría cuántas personas tienen cada edad específica, lo que permitiría comprender la distribución de edades en la población y realizar análisis demográficos o segmentaciones.

Figura 6.4.1 Distribución de frecuencias para variables cuantitativas

Edad en años

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	20	2	,2	,2	,2
	21	12	1,4	1,4	1,6
	22	14	1,6	1,6	3,3
	23	21	2,5	2,5	5,8
	24	30	3,5	3,5	9,3
	25	25	2,9	2,9	12,2
	26	30	3,5	3,5	15,8
	27	33	3,9	3,9	19,6
	28	38	4,5	4,5	24,1
	29	51	6,0	6,0	30,1
	30	30	3,5	3,5	33,6
	31	42	4,9	4,9	38,6
	32	30	3,5	3,5	42,1
	33	31	3,6	3,6	45,8
	34	38	4,5	4,5	50,2

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

6.5 Distribución de frecuencia relativa

La distribución de frecuencia relativa es una forma de presentar la frecuencia de diferentes categorías o valores en un conjunto de datos en relación con el total de observaciones o elementos. En lugar de mostrar la frecuencia absoluta (número de veces que aparece cada categoría), la distribución de frecuencia relativa muestra la proporción o el porcentaje que representa cada categoría en el conjunto de datos.

Para calcular la distribución de frecuencia relativa, se divide la frecuencia de cada categoría por el total de observaciones y se expresa como un porcentaje o decimal. Esto permite comparar la importancia relativa de cada categoría en el conjunto de datos y entender su contribución al total.

La distribución de frecuencia relativa es útil para destacar la proporción de cada categoría en relación con el conjunto completo de datos y facilitar comparaciones entre diferentes grupos o subconjuntos de datos. También es útil para identificar patrones o tendencias que no serían evidentes al observar solo la frecuencia absoluta.

Por ejemplo, en un estudio sobre preferencias de sexo en una muestra de personas, la distribución de frecuencia relativa mostraría el porcentaje de hombres y mujeres en la muestra en lugar de solo el número absoluto de cada categoría o clase, lo que permite comparaciones más claras y comprensibles (Anderson et al., 2019).

Capítulo 7

Gráficos

Gráficos

Introducción

Los gráficos son herramientas visuales poderosas que permiten comunicar datos de manera efectiva y comprensible. Este capítulo introduce las técnicas esenciales para crearlos y utilizarlos en el análisis estadístico. Se proporciona una guía práctica para seleccionar y diseñar los gráficos más adecuados para diferentes tipos de datos.

Se comienza con una descripción general de los tipos de gráficos más comunes usados en la estadística, incluyendo gráficos de barras, de líneas, circulares y diagramas de dispersión. Cada tipo de gráfico se discute en detalle, se explican sus aplicaciones específicas y cómo interpretar la información que presentan.

El capítulo también aborda la creación de gráficos más avanzados, como diagramas de cajas y barras de error, que son útiles para representar la variabilidad y la distribución de los datos. Además, se exploran los gráficos tridimensionales (3D), que pueden ofrecer una perspectiva adicional en ciertos tipos de análisis, aunque se debe tener cuidado con su interpretación para evitar malentendidos.

Finalmente, se presenta la pirámide de población y los histogramas, que son herramientas fundamentales para representar distribuciones de frecuencia y tendencias demográficas. Este capítulo es crucial para cualquier persona que necesite presentar datos de manera clara y efectiva, ya que proporciona las

habilidades necesarias para seleccionar y crear gráficos que comuniquen de manera precisa y atractiva.

Objetivos del capítulo

- Identificar los diferentes tipos de gráficos disponibles en SPSS, como barras, líneas e histogramas.
- Crear gráficos adecuados para representar diferentes tipos de datos en SPSS.
- Interpretar la información visual representada en los gráficos y su relevancia para el análisis estadístico.

7.1 Gráficos

La información representada mediante gráficos es importante en los análisis descriptivos, predictivos y prescriptivos. Estos gráficos no solo facilitan la interpretación de grandes volúmenes de datos, sino que también permiten identificar patrones y tendencias de manera visual, lo cual es fundamental para la toma de decisiones informadas.

Los gráficos, al traducir datos numéricos en representaciones visuales, ayudan a comunicar hallazgos complejos de manera más accesible y comprensible para una audiencia diversa. Por ejemplo, en un análisis descriptivo, un gráfico de barras puede mostrar de forma clara cómo se distribuyen las ventas por regiones y destacrar las áreas de mayor y menor rendimiento. En un análisis predictivo, un gráfico de líneas puede ilustrar las proyecciones futuras basadas en datos históricos, lo que ayuda a prever comportamientos y preparar estrategias adecuadas. Por último, en un análisis prescriptivo, gráficos como los de dispersión pueden mostrar la relación entre diferentes variables y sugerir las mejores acciones que seguir para optimizar resultados.

Además, los gráficos permiten una comparación rápida y efectiva entre diferentes conjuntos de datos, lo cual es invaluable para identificar discrepancias y oportunidades de mejora. La visualización de datos no solo embellece los informes, sino que enriquece el análisis al facilitar una comprensión más profunda y una comunicación más efectiva de los *insights* obtenidos.

Basado en ello, en el programa SPSS, podemos visualizar en la barra de herramientas, la opción: Gráficos, Cuadros de diálogo antiguos, las siguientes opciones:

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

7.2 Barras

Las barras, en el contexto de representaciones gráficas y estadísticas, se refieren a un tipo de gráfico donde la información se presenta mediante barras rectangulares de longitudes proporcionales a los valores que representan. Estas barras se utilizan comúnmente para comparar cantidades o categorías diferentes.

Figura 7.2.1 Gráfico de barras

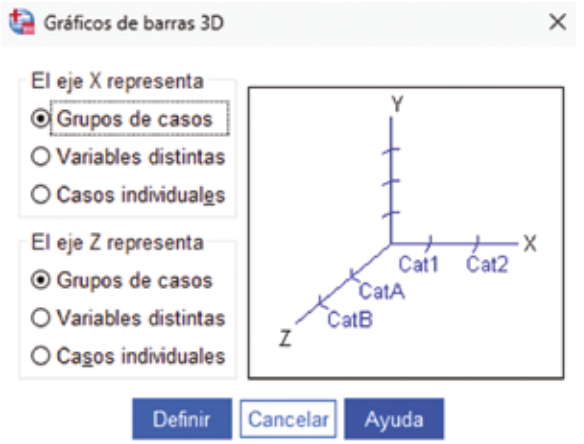
Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- Gráfico de barras simple: puede ser utilizado para mostrar las ventas mensuales de un producto, proporcionando una visión clara de cómo fluctúan las ventas de un mes a otro.
- Gráfico de barras agrupadas: permite hacer comparaciones entre diferentes categorías, ideal para analizar cómo se desempeñan varios productos o divisiones dentro de un mismo período.
- Gráfico de barras apiladas: es útil para mostrar variaciones en el tiempo y cómo se distribuyen los datos entre diferentes subcategorías, lo que facilita la visualización de la composición y cambios en las proporciones a lo largo del tiempo.

7.3 Barras 3D

El gráfico de barras 3D permite una visualización más compleja y detallada de los datos, al añadir una tercera dimensión que puede representar otra variable o categoría. Este tipo de gráfico es especialmente útil cuando se necesita mostrar la interacción entre múltiples variables o grupos de datos.

Figura 7.3.1 Gráfico de barras 3D



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Este tipo de grafico permite:

- **Comparación entre categorías:** al usar grupos de casos en ambos ejes X y Z, se pueden comparar diferentes categorías entre sí en un mismo gráfico, lo cual es útil para ver cómo diferentes categorías se relacionan entre sí.
- **Análisis de múltiples variables:** si se seleccionan variables distintas para los ejes X y Z, el gráfico mostrará cómo estas variables interactúan entre sí, proporcionando una visión más rica y detallada del conjunto de datos.
- **Visualización de casos individuales:** esto es útil cuando se necesita ver datos detallados para casos específicos en lugar de agrupaciones o resúmenes.

7.4 Líneas

En el contexto de representaciones gráficas, las líneas se refieren a un tipo de gráfico donde los puntos de datos se conectan mediante líneas rectas. Este tipo de gráfico se utiliza comúnmente para mostrar tendencias y cambios en los datos a lo largo del tiempo o en relación con una variable independiente. Por ejemplo, un gráfico de líneas podría mostrar la evolución de las temperaturas diarias a lo largo de un mes.

Figura 7.4.1 Gráfico de Líneas

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

El tipo de grafico de línea:

- **Simple:** es un gráfico donde se representa una sola línea que muestra la tendencia de una variable.
- **Múltiple:** permite representar varias líneas en un solo gráfico para mostrar comparaciones entre diferentes grupos o categorías.
- **Vertical:** se usa para mostrar líneas verticales, típicamente para comparar diferentes puntos en el tiempo o diferentes categorías.

Dentro del grafico de línea, se tiene lo siguiente:

- **Resúmenes para grupos de casos:** en esta opción, agrupa los datos antes de graficar, generalmente usando estadísticas descriptivas como el promedio, la mediana, etc., para cada grupo.
- **Resúmenes para distintas variables:** utiliza los valores de diferentes variables, resumiendo cada una antes de graficar.
- **Valores individuales de los casos:** permite graficar los valores de cada caso individual sin ningún tipo de resumen o agrupación.

7.5 Áreas

Los gráficos de áreas son similares a los gráficos de líneas, pero en lugar de conectar los puntos de datos con líneas, el área entre los puntos y el eje horizontal se rellena para resaltar la magnitud o la proporción de los datos representados. Estos gráficos son útiles para visualizar la contribución relativa de diferentes categorías a un total o para mostrar la evolución de una variable a lo largo del tiempo de manera más visual.

Figura 7.5.1 Gráfico de áreas

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

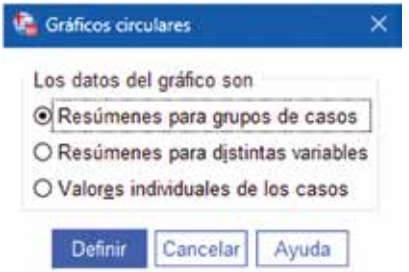
El tipo de grafico de áreas:

- **Simple:** representa una única serie de datos como un área llena, que muestra la magnitud de los valores en un gráfico acumulativo.
- **Apilado:** es un gráfico de áreas en el cual varias series de datos se apilan una sobre la otra, permitiendo ver tanto la contribución individual de cada serie como el total acumulado.

7.6 Circular

Los gráficos circulares, también conocidos como gráficos de pastel, representan datos en forma de un círculo dividido en sectores proporcionales a los valores que representan. Cada sector representa una categoría o una parte de un todo, y la suma de todos los sectores completa el círculo. Estos gráficos son efectivos para mostrar la distribución de categorías o porcentajes en relación con un total.

Figura 7.6.1 Gráficos circulares

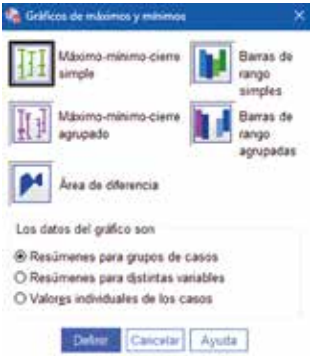


Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

7.7 Máximos y mínimos

En el contexto de análisis de datos, los máximos y mínimos se refieren a los valores más altos y bajos respectivamente en un conjunto de datos. Estos valores son importantes para comprender la variabilidad y la gama de los datos, así como para identificar tendencias o puntos destacados, como picos o valles en una serie temporal.

Figura 7.7.1 Gráficos de máximos y mínimos



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

El tipo de grafico de máximos y mínimos presenta opciones como:

- **Máximo-mínimo-cierre simple:** el cual presenta un gráfico que muestra el valor máximo, mínimo y de cierre de una variable, generalmente utilizado para analizar series temporales o datos financieros.
- **Máximo-mínimo-cierre agrupado:** este es similar al gráfico anterior, pero permite mostrar varias series de datos agrupadas, lo que facilita la comparación entre diferentes grupos o categorías.
- **Barras de rango simples:** es un gráfico de barras que muestra el rango de variación entre un valor máximo y mínimo para cada categoría o grupo.
- **Barras de rango agrupadas:** es similar a las barras de rango simples, pero con la posibilidad de agrupar varias series, lo que permite comparaciones más detalladas entre diferentes grupos.
- **Área de diferencia:** es un gráfico que muestra la diferencia entre dos series de datos, representada como un área sombreada. Este tipo de gráfico es útil para resaltar la diferencia entre dos series a lo largo del tiempo.

7.8 Diagramas de cajas

Los diagramas de cajas, también conocidos como boxplots, son gráficos que representan la distribución de un conjunto de datos mediante cuartiles, medianas, valores máximos y mínimos. Estos diagramas proporcionan información sobre la dispersión y la simetría de los datos, así como la presencia de valores atípicos. Son útiles para comparar distribuciones y detectar patrones o anomalías.

Figura 7.8.1 Diagramas de cajas

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Los tipos de diagramas de cajas:

- **Simple:** muestra la distribución de una variable, incluyendo la mediana, los cuartiles y los posibles valores atípicos. Es útil para visualizar la dispersión y la simetría de los datos.
- **Agrupado:** permite comparar la distribución de una variable entre diferentes grupos o categorías. Cada grupo tendrá su propio diagrama de caja, lo que facilita la comparación directa.

7.9 Barras de error

Las barras de error se utilizan en gráficos estadísticos para representar la variabilidad o incertidumbre asociada con los valores representados. Estas barras se extienden desde cada punto de datos y muestran la gama probable de valores que podrían tener los datos, lo que ayuda a interpretar la precisión de las estimaciones o mediciones.

Figura 7.9.1 Barras de error

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

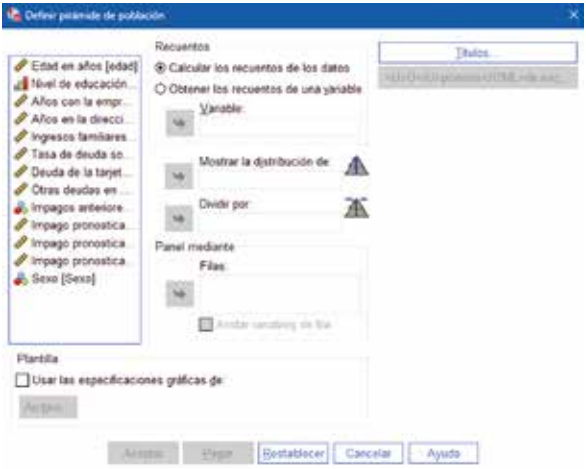
Los tipos de gráfica de barras de error:

- **Simple:** muestra la media y el error estándar o el intervalo de confianza de una variable. Es útil para visualizar la precisión de las estimaciones de la media.
- **Agrupados:** permite comparar diferentes grupos. Cada uno tendrá su propia barra de error, lo que facilita la comparación entre las medias y la variabilidad entre grupos.

7.10 Pirámide de población

Una pirámide de población es un tipo de gráfico que representa la distribución de la población por grupos de edad y sexo. La forma de la pirámide refleja la estructura demográfica para mostrar la proporción de personas en cada grupo de edad y resalta tendencias como el envejecimiento de la población o la presencia de una población joven.

Figura 7.10.1 Gráfico de pirámides poblacionales



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

7.11 Dispersión

Los gráficos de dispersión muestran la relación entre dos variables mediante la representación de puntos en un plano cartesiano, donde cada uno representa un par de valores de las variables. Estos gráficos son útiles para identificar patrones, correlaciones o tendencias en los datos y para visualizar la dispersión o la concentración de los puntos alrededor de una línea de tendencia.

Figura 7.11.1Gráfico de dispersión/puntos

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

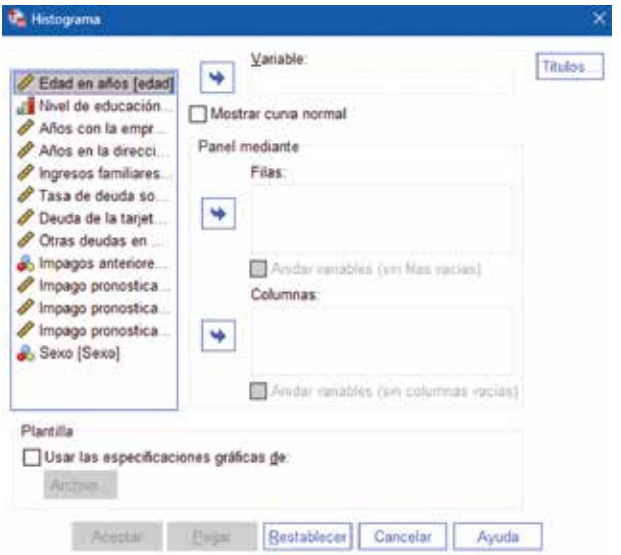
Tipos de gráficos de dispersión:

- **Dispersión simple:** muestra la relación entre dos variables mediante puntos dispersos en el plano cartesiano. Es útil para identificar patrones, tendencias o correlaciones entre las variables.
- **Dispersión superpuesta:** es similar al gráfico de dispersión simple, pero permite superponer varias series de datos en el mismo gráfico para compararlas.
- **Dispersión matricial:** muestra las relaciones entre varias combinaciones de variables en una sola vista, facilitando la identificación de correlaciones entre múltiples variables.
- **Dispersión 3D:** agrega una tercera dimensión, permitiendo visualizar la relación entre tres variables simultáneamente.
- **Puntos simples:** es un gráfico básico que representa puntos simples sin ninguna otra estructura adicional. Suele ser utilizado para visualizar la distribución de una variable.

7.12 Histogramas

Los histogramas son gráficos que representan la distribución de frecuencias de un conjunto de datos mediante barras rectangulares donde la altura de cada barra representa la frecuencia o la cantidad de datos en un intervalo específico. Estos gráficos son útiles para visualizar la distribución y la forma de los datos, identificar patrones como sesgos o concentraciones, y comparar la frecuencia de diferentes intervalos.

Figura 7.12.1 Histograma



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Capítulo 8

Tablas de contingencia

Tablas de contingencia

Introducción

Las tablas de contingencia son herramientas estadísticas esenciales para analizar la relación entre dos o más variables categóricas. Este capítulo proporciona una introducción completa a las tablas de contingencia y explica cómo se construyen, interpretan y usan en el análisis de datos.

El capítulo comienza definiendo qué es una tabla de contingencia y cómo se utiliza para resumir la relación entre variables categóricas. Se presentan diferentes tipos de tablas de contingencia, como las cruzadas, que permiten comparar la frecuencia de ocurrencia de diferentes combinaciones de categorías. Estas tablas son fundamentales para identificar patrones y asociaciones entre variables, lo que puede proporcionar información valiosa en investigaciones sociales, estudios de mercado, y otros campos.

Además, se aborda la configuración de las tablas de contingencia en SPSS, incluyendo la selección de variables, la disposición de las casillas y la interpretación de los resultados. También se discuten las pruebas exactas y los estadísticos usados para evaluar la significancia de las asociaciones observadas, con lo que se proporciona al lector una comprensión sólida de cómo interpretar los resultados.

Este capítulo es vital para aquellos que buscan analizar datos categóricos y entender las relaciones entre variables. Al final del capítulo, el lector estará

equipado para crear, interpretar y usar tablas de contingencia en sus propios análisis estadísticos.

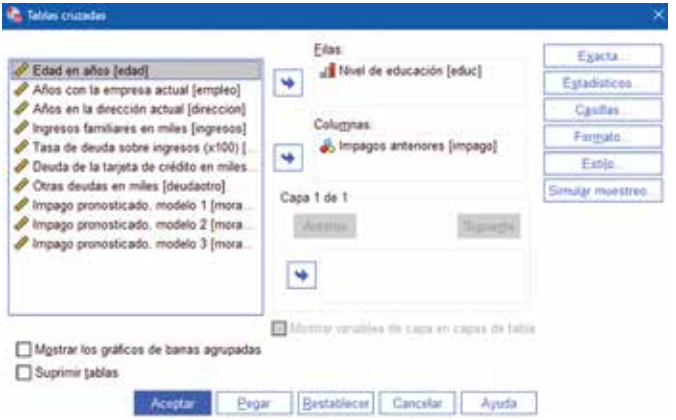
Objetivos del capítulo

- Definir una tabla de contingencia y sus componentes principales.
- Construir una tabla de contingencia para analizar la relación entre dos o más variables categóricas.
- Interpretar los resultados obtenidos de una tabla de contingencia y su significancia en el análisis estadístico.

8.1 Tablas cruzadas o de contingencia

Una tabla de contingencia organiza los datos en una matriz rectangular, donde las filas y columnas representan las diferentes categorías de las variables en estudio. Cada celda en la tabla muestra la frecuencia o el conteo de ocurrencias para la combinación específica de categorías correspondientes a esa fila y columna. Para crear una tabla de contingencia en el SPSS es necesario ir a Analizar, luego Estadísticos descriptivos y finalmente se selecciona Tablas cruzadas. Para que pueda dar seguimiento al ejercicio, puede utilizar la base de ejemplo Bankloan.

Figura 8.1.1 Tablas cruzadas



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Durante la creación de tablas de contingencia, es fundamental seleccionar adecuadamente las variables que se asignarán a las filas y columnas. Las variables independientes (predictoras) deben ubicarse en las filas, ya que son las que se desean analizar, mientras que la variable dependiente se coloca en las columnas para evaluar cómo es influenciada por las variables predictoras.

Figura 8.1.2 Tabla de contingencia o tabla cruzada

Tabla cruzada Nivel de educación*Impagos anteriores

Recuento

		Impagos anteriores		
		No	Sí	Total
Nivel de educación	No completó el bachillerato	293	79	372
	Título de Bachiller	139	59	198
	Superiores iniciados	57	30	87
	Título Superior	24	14	38
	Título de Post-grado	4	1	5
Total		517	183	700

Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

8.1.1 Tipos de tablas de contingencia

Existen diferentes tipos de tablas de contingencia, dependiendo del número de variables involucradas:

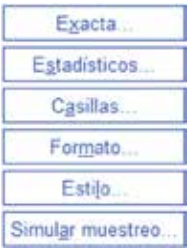
- **Tablas de contingencia simples (2x2):** son las más básicas, con dos variables y dos categorías en cada variable. Son útiles para analizar la relación entre dos variables binarias.
- **Tablas de contingencia múltiples:** tienen más de dos categorías o variables, lo que permite analizar interacciones más complejas. Por ejemplo, una tabla de 3x4 analizaría la relación entre dos variables, una con tres categorías y otra con cuatro.

8.2 Configuración de las tablas de contingencia

Para realizar un análisis estadístico completo y personalizado, es crucial entender las diversas herramientas y opciones disponibles en el software

que utilizamos. Dentro de la interfaz del software estadístico, se presentan varios botones que permiten configurar y ajustar el análisis según nuestras necesidades específicas. Estos botones no solo nos facilitan la obtención de resultados precisos, sino que también nos permiten adaptar la presentación de esos resultados para una mejor interpretación. A continuación, exploraremos la funcionalidad de cada uno de estos botones y cómo pueden ser utilizados para optimizar nuestro análisis estadístico.

Figura 8.2.1 Opciones presentes en la tabla de contingencia o tabla cruzada

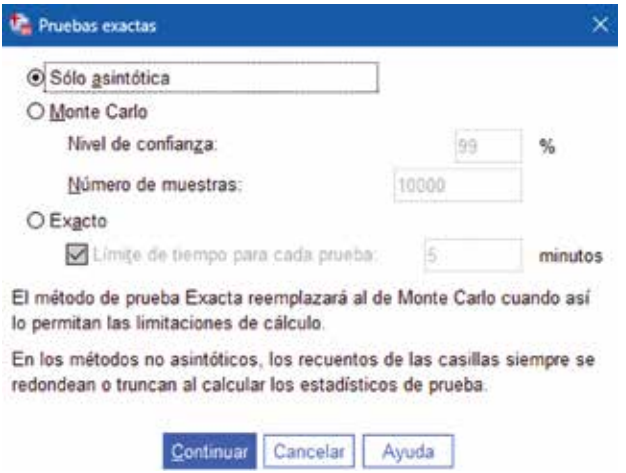


Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

8.2.1 Pruebas exactas

Permite seleccionar el método que se utilizará para calcular los estadísticos de prueba, específicamente en el contexto de análisis de tablas de contingencia o pruebas de hipótesis.

Figura 8.2.2 Pruebas exactas



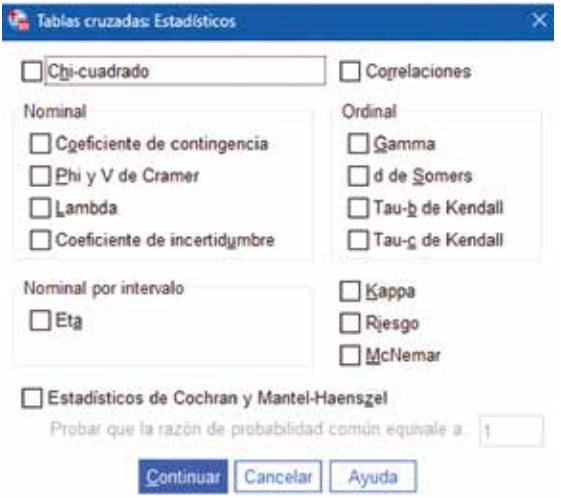
Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Solo asintótica:** esta opción calcula las pruebas usando métodos asintóticos, que se basan en aproximaciones que son válidas cuando el tamaño de la muestra es grande.
- **Monte Carlo:** es útil cuando las condiciones para los métodos asintóticos no se cumplen, por ejemplo, en muestras pequeñas o cuando los supuestos teóricos no se ajustan bien a los datos.
    - **Nivel de confianza:** porcentaje que indica la confianza en los intervalos calculados a partir de las simulaciones. Se suele establecer en 90 %, 95 % o 99 %.
    - **Número de muestras:** indica el número de simulaciones que se van a realizar. Un mayor número de muestras puede dar un resultado más preciso, pero también requerirá más tiempo de cálculo.
- Exacto: este intenta calcular un valor exacto del estadístico de prueba, sin depender de aproximaciones asintóticas o simulaciones. Esto es factible y generalmente se prefiere cuando se trabaja con tablas pequeñas o cuando se desea la mayor precisión posible en la prueba.
    - **Límite de tiempo para cada prueba:** esta opción permite establecer un límite de tiempo máximo para que el software intente calcular una prueba exacta. Si el tiempo límite es superado, el software podría no completar el cálculo, o podría revertir a un método menos intensivo, como Monte Carlo o asintótico.

8.2.2 Estadísticos

Cuando se realiza un análisis de tablas cruzadas, es esencial elegir los estadísticos adecuados para evaluar la relación entre las variables categóricas. La ventana que se muestra proporciona una serie de opciones que permiten seleccionar los estadísticos específicos que se desea calcular. Estos estadísticos ayudan a entender la naturaleza y la fuerza de las relaciones entre las variables en estudio, ya sea mediante pruebas de independencia, medidas de asociación o análisis más complejos. Estos resultados permiten obtener una visión clara de la interacción entre categorías, lo que facilita la toma de decisiones informadas al identificar patrones significativos y dependencias estadísticas en el contexto analizado (Spiegel y Stephens, 2009).

Figura 8.2.3 Tablas cruzadas: Estadísticos



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

- **Chi-cuadrado:** es una prueba estadística que evalúa si existe una asociación significativa entre las variables categóricas en la tabla de contingencia. Es uno de los métodos más comunes para analizar tablas cruzadas.
- **Correlaciones:** aunque generalmente se asocian con variables numéricas, las correlaciones para variables categóricas (usualmente ordinales) también pueden calcularse, ayudando a medir la fuerza y la dirección de la relación entre ellas.

Nominal:

- **Coeficiente de contingencia:** mide la asociación entre dos variables categóricas. Es similar al chi-cuadrado, pero ofrece una medida normalizada.
- **Phi y V de Cramer:** son estadísticos usados para medir la fuerza de la asociación en tablas de contingencia. El V de Cramer una adaptación del coeficiente phi para tablas de más de 2x2.
- **Lambda:** es una medida de asociación que indica la reducción del error al predecir la categoría de una variable dependiente tras conocer la categoría de la variable independiente.

- **Coeficiente de incertidumbre:** proporciona una medida de la proporción de incertidumbre en una variable que se explica por la otra.

Nominal por intervalo:

- **Eta:** mide la relación entre una variable nominal y una variable de intervalo, a menudo utilizada en ANOVA.

Ordinal:

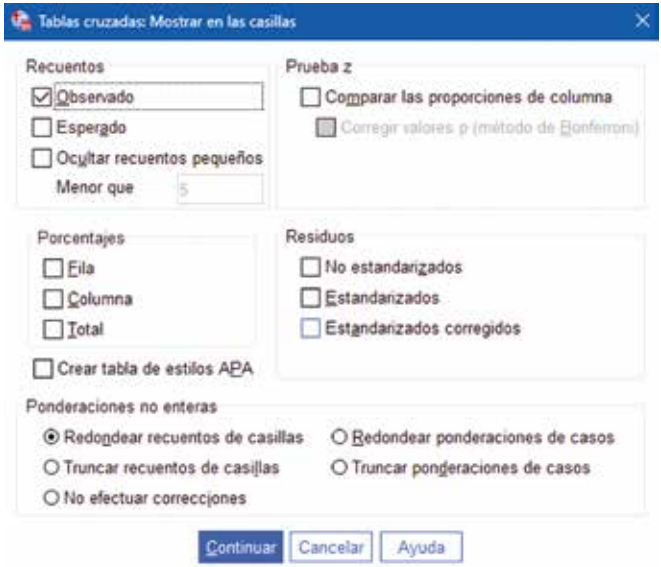
- **Gamma, d de Somers, Tau-b y Tau-c de Kendall:** son estadísticos que miden la asociación entre variables ordinales, ajustando por posibles empates en los datos.
- **Kappa:** mide la concordancia entre dos evaluadores o métodos, ajustada por la concordancia que podría ocurrir por azar.
- **Riesgo:** evalúa el riesgo relativo o la *odds ratio* entre las categorías de dos variables, comúnmente usado en estudios epidemiológicos.
- **McNemar:** es una prueba estadística utilizada para comparar proporciones en tablas 2x2 con datos pareados, útil en estudios de casos y controles.
- **Estadísticos de Cochran y Mantel-Haenszel:** son utilizados para evaluar la asociación entre variables categóricas mientras se controla por variables de confusión o estratos.

8.2.3 Casillas

En las tablas de contingencia (cruzadas) es fundamental tener control sobre cómo se presentan y visualizan los datos en las celdas de la tabla. La configuración de visualización no solo afecta la claridad de los resultados, sino que también permite a los analistas destacar los aspectos más relevantes del análisis. La ventana que se presenta a continuación ofrece diversas opciones para personalizar cómo se mostrarán los datos en las celdas de una tabla cruzada, asegurando que la información clave se resalte y sea fácil de interpretar.

Esta capacidad de personalización ayuda a enfocar el análisis en patrones o relaciones significativas, facilitando la toma de decisiones basada en la estructura visual optimizada de la tabla (Anderson et al., 2019).

Figura 8.2.4 Tablas cruzadas: Mostrar en las casillas



Obtenido de IBM SPSS Statistics for Windows, Version 27.0. Armonk, NY: IBM Corp

Recuentos:

- **Observado:** esta opción está seleccionada por defecto y muestra los recuentos reales de casos en cada celda de la tabla.
- **Esperado:** muestra los valores esperados bajo la hipótesis de independencia entre las variables, lo que permite comparar con los valores observados para evaluar discrepancias.
- **Ocultar recuentos pequeños:** permite ocultar recuentos menores a un valor especificado, lo que puede ser útil para enfocar el análisis en datos más significativos y evitar interpretar frecuencias muy bajas.

Prueba z:

- **Comparar las proporciones de columna:** ofrece la posibilidad de comparar las proporciones entre columnas y determinar si existen diferencias significativas. La opción de corregir los valores p mediante el método de Bonferroni se puede seleccionar para ajustar por múltiples comparaciones y reducir el riesgo de obtener resultados significativos por azar.

Porcentajes: permite calcular y mostrar los porcentajes basados en:

- **Fila:** para ver la distribución relativa dentro de cada fila.
- **Columna:** para comparar la distribución relativa dentro de cada columna.
- **Total:** para entender la proporción de cada celda respecto al total general de la tabla.

Residuos:

- **No estandarizados:** muestra las diferencias simples entre los recuentos observados y esperados, que pueden ser difíciles de interpretar sin normalización.
- **Estandarizados:** ajusta los residuos para facilitar la identificación de celdas que se desvían significativamente de lo esperado.
- **Estandarizados corregidos:** proporciona una medida aún más precisa al corregir los residuos por posibles sesgos, destacando las celdas con discrepancias significativas.

Ponderaciones no enteras: ofrece opciones para manejar recuentos y ponderaciones que no son números enteros, con la posibilidad de redondear, truncar o no aplicar correcciones, dependiendo de la precisión requerida en el análisis.

Capítulo 9

Introducción a la probabilidad

Introducción a la probabilidad

Introducción

Este capítulo presenta los principios fundamentales de la probabilidad, esenciales para cuantificar la incertidumbre y realizar análisis estadísticos. Se comienza definiendo la probabilidad y su importancia en la toma de decisiones bajo incertidumbre, para introducir el concepto de espacio muestral, que incluye todos los posibles resultados de un experimento aleatorio. Se diferencian tipos de eventos, como los mutuamente excluyentes y no excluyentes, cruciales para el cálculo de probabilidades.

Se abordan las reglas fundamentales de la probabilidad, como la regla de la suma y la del producto, herramientas clave para calcular probabilidades en combinaciones de eventos. Además, se explora la probabilidad condicional, que permite determinar la probabilidad de un evento dado que otro ha ocurrido.

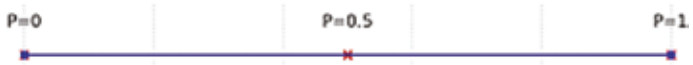
El capítulo también cubre leyes fundamentales como la ley de los grandes números y el teorema del límite central, que son pilares en la teoría de la probabilidad. Finalmente, se introduce el teorema de Bayes, una herramienta poderosa para actualizar probabilidades a medida que se obtiene nueva información. Este capítulo proporciona las bases necesarias para comprender y aplicar la probabilidad en análisis estadísticos, preparando al lector para abordar problemas más complejos en estudios futuros.

Objetivos del capítulo

- Definir conceptos básicos de probabilidad, incluyendo espacio muestral y eventos.
- Explicar las reglas fundamentales de la probabilidad, como la regla de la suma y la regla del producto.
- Calcular la probabilidad de eventos simples y compuestos utilizando las reglas básicas de probabilidad.

9.1 Probabilidad

La probabilidad es una medida cuantitativa que expresa la incertidumbre asociada con la ocurrencia de un evento. Esta medida toma valores en un rango que va de 0 a 1, o del 0 % al 100 %, dependiendo de la representación utilizada. Entre estos extremos, los valores de probabilidad reflejan el grado de confianza en que ocurra el evento, lo que es útil para evaluar riesgos y hacer predicciones en estudios estadísticos y decisiones informadas en diversos contextos (Lind et al., 2020; Anderson et al., 2019; Webster, 2000; Mohammad y Garcia Dunna, 1996).



En una recta numérica que representa la probabilidad, se pueden identificar tres puntos clave:

P=0: este punto representa la probabilidad de que un evento ocurra cuando es imposible. En otras palabras, no existe ninguna posibilidad de que el evento se realice.

P=0,5: este punto se sitúa en el centro de la recta, lo que representa una probabilidad de 0,5 o 50 %. Aquí, el evento tiene la misma probabilidad de ocurrir que de no ocurrir, lo que implica una incertidumbre máxima con un 50 % de posibilidades para cada resultado posible.

P=1: este punto indica que el evento es absolutamente seguro de ocurrir. Con una probabilidad de 1 o 100 %, no hay ninguna incertidumbre, y el evento se realizará con certeza.

Existen diferentes enfoques para entender la probabilidad, como la probabilidad clásica, frecuencia relativa o empírica y subjetiva.

- **Probabilidad clásica:** se usa cuando todos los resultados posibles de un experimento son igualmente probables. Por ejemplo, al lanzar un dado, la probabilidad de obtener un número específico (digamos 3) es 1/6, ya que hay seis resultados posibles y todos son igualmente probables.
- **Probabilidad empírica o frecuencia relativa:** se basa en la observación de la frecuencia con la que ocurre un evento en una gran cantidad de repeticiones del experimento. Por ejemplo, si lanzas una moneda 100 veces y se obtiene cara cuarenta y ocho veces, se puede estimar la probabilidad de obtener cara como 48/100=0,48.
- **Probabilidad subjetiva:** se refiere a la probabilidad basada en el juicio personal o creencia de una persona sobre la ocurrencia de un evento, que puede no estar basada en datos empíricos.

En el contexto de las ciencias sociales, la economía o la administración, la probabilidad se utiliza para tomar decisiones bajo incertidumbre, realizar predicciones y modelar situaciones en las que los resultados no son determinísticos. Este concepto es fundamental en la estadística y en la teoría de decisiones, donde se emplea para evaluar riesgos, hacer pronósticos y optimizar recursos. Su aplicación permite valorar distintos escenarios posibles y anticipar comportamientos o eventos, para facilitar la planificación estratégica y la administración eficiente de recursos en situaciones donde la certeza no es alcanzable (Newbold at al., 2008).

9.2 Espacio muestral y tipos de eventos

El espacio muestral es el conjunto de todos los posibles resultados de un experimento aleatorio, denotado por S. Los eventos, que pueden ser simples o compuestos, son subconjuntos de este espacio. Eventos mutuamente excluyentes no pueden ocurrir simultáneamente, mientras que los no excluyentes pueden coincidir. Estos conceptos son fundamentales en el estudio de la probabilidad, ya que definen el marco en el cual se calculan las probabilidades de ocurrencia de distintos eventos y permiten estructurar modelos para experimentos complejos. La comprensión de estos principios facilita el análisis de

fenómenos aleatorios y la toma de decisiones en situaciones de incertidumbre (Anderson et al., 2019; Spiegel y Stephens, 2009).

9.2.1 Espacio muestral

Es el conjunto de todos los posibles resultados de un experimento aleatorio. El espacio muestral se denota comúnmente por la letra S y puede ser finito o infinito según el experimento. Por ejemplo, en un lanzamiento de un dado, el espacio muestral es S={1,2,3,4 ,5,6}, debido a que esos son los únicos resultados posibles. Comprender el espacio muestral es esencial en probabilidad, pues permite definir todos los eventos posibles y calcular sus probabilidades, lo cual es fundamental en modelos de análisis y toma de decisiones bajo incertidumbre (Lind et al., 2020).

9.2.2 Tipos de eventos

- **Evento simple:** es aquel que consiste en un solo resultado del espacio muestral. Por ejemplo, al lanzar un dado, obtener un 3 es un evento simple.
- **Evento compuesto:** es aquel que consiste en dos o más resultados del espacio muestral. Por ejemplo, obtener un número par al lanzar un dado es un evento compuesto, ya que incluye los resultados {2,4,6}

9.2.3 Eventos mutuamente excluyentes y no excluyentes

- **Mutuamente excluyentes:** existen eventos que no pueden ocurrir al mismo tiempo. A estos se les llama mutuamente excluyentes. Por ejemplo, en el lanzamiento de un dado, los eventos «obtener un 2» y «obtener un 5» son mutuamente excluyentes porque no pueden suceder simultáneamente.
- **No mutuamente excluyentes:** existen eventos que pueden ocurrir al mismo tiempo, llamados no mutuamente excluyentes. Por ejemplo, en un conjunto de cartas, los eventos «sacar una carta roja» y «sacar un número par» no son mutuamente excluyentes, debido a que es posible sacar una carta roja que sea un número par, como el 2 de corazones.

9.3 Reglas fundamentales de la probabilidad

Las reglas fundamentales de la probabilidad proporcionan las bases para el cálculo de la probabilidad de diferentes eventos en un experimento aleatorio. Estas reglas incluyen la adición, la multiplicación y el complemento de eventos, lo que permite determinar la probabilidad conjunta, marginal y condicional, lo que es esencial para analizar situaciones de incertidumbre y tomar decisiones informadas. Estas reglas son fundamentales para construir modelos estadísticos que permiten predecir y gestionar riesgos en diversos contextos de estudio (Levin y Rubin, 2010).

9.3.1 Regla de la suma (adición)

Existen dos formas principales de aplicar la regla de la suma:

- Para eventos mutuamente excluyentes: Si los eventos A y B son mutuamente excluyentes, la probabilidad de que ocurra A o B es la suma de sus probabilidades individuales: $P(A \cup B) = P(A) + P(B)$

Tabla 9.3.1 Sintaxis de probabilidad mutuamente excluyente

```
DATA LIST FREE / ID (F1) Opción (A1).
BEGIN DATA.
1 A
2 B
3 A
4 B
5 A
END DATA.

FREQUENCIES VARIABLES=Opción
  /STATISTICS=FREQ.

* Calcular las probabilidades directamente.
INPUT PROGRAM.
COMPUTE P_A = 3/5.
COMPUTE P_B = 2/5.
COMPUTE P_A_or_B = P_A + P_B.
END CASE.
END FILE.
END INPUT PROGRAM.
* Mostrar un único resultado.
LIST VARIABLES = P_A P_B P_A_or_B.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

- Para eventos no mutuamente excluyentes: si los eventos A y B no son mutuamente excluyentes, la probabilidad de que ocurra A o B es la suma de sus probabilidades individuales menos la probabilidad de que ocurran ambos eventos: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Tabla 9.3.2 Sintaxis de probabilidad no mutuamente excluyentes

```
* Paso 1: Definir los datos.
DATA LIST FREE / Opción_A (F1) Opción_B (F1).
BEGIN DATA.
1 1
1 0
0 1
1 1
0 1
END DATA.
* Paso 2: Calcular las probabilidades.
COMPUTE Opción_A_y_B = Opción_A * Opción_B.
AGGREGATE /OUTFILE=* /N = N /P_A = MEAN(Opción_A) /P_B =
MEAN(Opción_B) /P_A_y_B = MEAN(Opción_A_y_B).
COMPUTE P_A_or_B = P_A + P_B - P_A_y_B.
* Mostrar las probabilidades.
LIST VARIABLES = P_A P_B P_A_y_B P_A_or_B.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

9.3.2 Regla del producto (multiplicación)

Existen dos formas principales de aplicar la regla del producto:

- Para eventos independientes: si los eventos A y B son independientes, la probabilidad de que ambos ocurran simultáneamente es el producto de sus probabilidades individuales: $P(A \cap B) = P(A) * P(B)$
- Para eventos dependientes: Si los eventos A y B son dependientes, la probabilidad de que ambos ocurran es el producto de la probabilidad de uno de ellos por la probabilidad condicional del otro dado que el primero ha ocurrido: $P(A \cap B) = P(A) * P(B|A)$

9.3.3 Probabilidad condicional

Es la probabilidad de que un evento ocurra dado que otro evento ha ocurrido. Se denota por P(B | A) y se calcula usando la fórmula:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Esta regla se utiliza en situaciones donde la ocurrencia de un evento afecta la probabilidad de que ocurra otro.

9.4 Leyes fundamentales de la probabilidad

Las leyes fundamentales de la probabilidad son principios esenciales que permiten comprender y predecir el comportamiento de eventos aleatorios. Estas leyes incluyen conceptos clave que explican cómo se comportan las medias y distribuciones muestrales en experimentos repetidos, sentando las bases para análisis y aplicaciones en estadística e inferencia.

9.4.1 Ley de los grandes números

La Ley de los Grandes Números establece que, a medida que el número de repeticiones de un experimento aleatorio aumenta, la media de los resultados observados se aproxima a la media esperada. Esto significa que cuanto más se repite un experimento, más cerca estará la probabilidad empírica de la probabilidad teórica.

Tabla 9.4.1 Sintaxis de la Ley de los grandes números

```
* Paso 1: Generar una muestra aleatoria de tamaño 1000.
NEW FILE.
INPUT PROGRAM.
LOOP id = 1 TO 1000.
   COMPUTE muestra = RV.NORMAL(50, 10). /* Genera datos con media 50 y
desviación estándar 10 */
   END CASE.
END LOOP.
END FILE.
END INPUT PROGRAM.
EXECUTE.

* Paso 2: Crear un índice acumulativo para calcular medias crecientes.
SORT CASES BY id.
COMPUTE sample_size = $CASENUM. /* Índice acumulativo del tamaño de la muestra
*/
EXECUTE.

* Paso 3: Calcular las medias acumulativas.
AGGREGATE
 /OUTFILE=* MODE=ADDVARIABLES
 /BREAK=sample_size
 /mean_acumulativa = MEAN(muestra).
```

```
* Paso 4: Graficar las medias acumulativas.
GGRAPH
  /GRAPHDATASET NAME="grafico" VARIABLES=sample_size mean_acumulativa
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("grafico"))
  DATA: sample_size=col(source(s), name("sample_size"))
  DATA: mean_acumulativa=col(source(s), name("mean_acumulativa"))
  GUIDE: axis(dim(1), label("Tamaño de la muestra"))
  GUIDE: axis(dim(2), label("Media acumulativa"))
  GUIDE: text.title(label("Ley de los Grandes Números"))
  ELEMENT: line(position(sample_size*mean_acumulativa), color.interior(color.black))
END GPL.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

9.4.2 Teorema de límite central

El teorema del límite central afirma que, si tomamos muestras suficientemente grandes de una población con cualquier distribución, la distribución de la media muestral se aproximará a una distribución normal, independientemente de la forma de la distribución original. Este teorema es fundamental en la inferencia estadística, ya que permite el uso de la distribución normal para estimaciones y pruebas de hipótesis.

9.4.3 Teorema de Bayes

El teorema de Bayes es una herramienta fundamental en la probabilidad y la estadística que permite actualizar la probabilidad de un evento en función de nueva información o evidencia. Este teorema es crucial para la toma de decisiones en situaciones de incertidumbre, al combinar la probabilidad previa con la probabilidad condicional para obtener una probabilidad posterior más precisa.

Tabla 9.4.2 Sintaxis de la Ley de los grandes números

```
* Crear datos simulados correctamente.
DATA LIST FREE / Evento (F1.0) Evidencia (A1).
BEGIN DATA
1 A
1 A
0 A
1 B
0 B
0 C
1 C
1 C
0 C
END DATA.

* Frecuencias de cada categoría.
FREQUENCIES VARIABLES=Evento Evidencia.

* Crear tabla cruzada para calcular P(Evidencia|Evento).
CROSSTABS
  /TABLES=Evidencia BY Evento
  /CELLS=ROW COUNT.

* Definir probabilidades para el cálculo del Teorema de Bayes.
COMPUTE P_Evento = 0.5.
* Probabilidad previa de Evento.

COMPUTE P_Evidencia = 0.3.
* Probabilidad de la Evidencia "A".

COMPUTE P_Evidencia_given_Evento = 0.6.
* P(Evidencia=A|Evento=1).

* Teorema de Bayes: P(Evento|Evidencia) = [P(Evidencia|Evento) * P(Evento)] /
P(Evidencia).
COMPUTE P_Evento_given_Evidencia =
    (P_Evidencia_given_Evento * P_Evento) / P_Evidencia.
EXECUTE.
```

Nota. La información proporcionada es ficticia y es utilizada con fines académicos.

Índice de figuras

Índice de tablas

Referencias

Anderson, D. R., Sweeney, D. J. y Williams, T. A. (2019). *Estadística para negocios y economía* (13.a ed.). Cengage.

Levin, R. I., y Rubin, D. S. (2010). *Estadística para Administración y Economía* (7.a ed.). Pearson.

Lind, D. A., Marchal, W. G. y Wathen, S. A. (2020). *Estadística aplicada a negocios y economía* (17.a ed.). McGraw-Hill Interamericana.

Mohammad, A. R. y García Dunna, E. (1996). *Simulación y análisis de modelos estocástivos*. McGraw-Hill Interamericana.

Newbold, P., Carlson, W. L. y Thorne, B. M. (2008). *Estadística para administración y economía* (3.a ed.). Pearson.

Spiegel, M. R. y Stephens, L. J. (2009). *Estadística* (4.a ed.). McGraw-Hill Interamericana.

Webster, A. L. (2000). *Estadística aplicada a los negocios y la eco*nomía (3.a ed.). Bogotá: McGraw-Hill Interamericana.

En un mundo cada vez más orientado hacia los datos, la estadística se ha convertido en una disciplina esencial para una amplia gama de campos profesionales y académicos. Desde la economía y las ciencias sociales hasta la medicina y la ingeniería, la capacidad de recolectar, analizar e interpretar datos es fundamental para la toma de decisiones informadas. Los datos no solo proporcionan un espejo para observar el pasado, sino que también son una brújula para guiar el futuro. Sin embargo, para aprovechar todo su potencial, es necesario comprender los principios estadísticos y saber cómo usarlos de manera efectiva. Este libro está diseñado precisamente con ese objetivo: proporcionar una guía completa y accesible para el aprendizaje de la estadística y su aplicación práctica mediante herramientas como el SPSS.

Este texto se organiza de manera sistemática, al empezar con las generalidades y abordar conceptos fundamentales que sientan las bases para el entendimiento del análisis de datos. La introducción a la estadística incluye definiciones clave, la clasificación de los diferentes tipos de variables y una explicación sobre las escalas de medición, elementos esenciales para cualquier análisis cuantitativo. Con este primer capítulo, los lectores desarrollarán una comprensión sólida de qué es la estadística y cómo se clasifica en diferentes tipos, ya sea descriptiva o inferencial, así como una diferenciación clara entre las variables cualitativas y cuantitativas. Además, se profundiza en las fuentes y tipos de datos, lo que proporciona un marco teórico sobre cómo se originan y cómo deben tratarse para su análisis.

UIDE
Powered by
Arizona State University

ISBN: 978-9942-682-06-2
9 789942 682062