Organización de datos y variables estadísticas

Breve descripción:

Este componente orientado al nivel técnico ofrece un enfoque sobre los principios y metodologías para la organización de datos estadísticos, abarcando la clasificación y el agrupamiento de variables cualitativas y cuantitativas. Incluye técnicas específicas para evitar errores y optimizar la precisión de los datos. Proporciona herramientas que mejoran la claridad y confiabilidad de los resultados en análisis estadísticos.

Noviembre 2024

Tabla de contenido

Introducción

En el ámbito de la estadística aplicada, la correcta organización de la muestra de datos siguiendo metodologías estadísticas debe asegurar que el análisis sea representativo, preciso y relevante. La estructura de los datos recolectados y la aplicación de técnicas estadísticas adecuadas son determinantes en la interpretación de resultados, puesto que una muestra desorganizada o sin criterios claros puede llevar a conclusiones imprecisas o sesgadas.

¿Cómo se organiza una muestra de datos para garantizar la coherencia y utilidad en el análisis estadístico? Este componente formativo profundiza en los principios y metodologías para la organización de datos en muestras, explorando técnicas de clasificación, agrupación y segmentación de acuerdo con el tipo de variable y el contexto del estudio. Se abarcarán enfoques prácticos que permiten estructurar los datos de manera que respalden un análisis eficaz y riguroso.

A lo largo de este componente, el aprendiz desarrollará habilidades para organizar muestras de datos de acuerdo con principios estadísticos, identificando las mejores prácticas para cada tipo de variable y objetivo analítico. Se aprenderán técnicas que optimizan la organización de datos y mejoran la precisión y claridad en los resultados estadísticos.

La organización de la muestra de datos es una etapa indispensable para cualquier análisis exitoso, pues un principio básico en estadística señala que "la exactitud de las conclusiones es tan buena como la calidad de la organización de los datos".

¡Bienvenido a recorrer este camino por la organización estructurada y eficiente de muestras de datos para el análisis estadístico!

Video 1. Organización de datos y variables estadísticas

Enlace de reproducción del video

Síntesis del video: Organización de datos y variables estadísticas
En el componente formativo «Organización de datos y variables estadísticas» se exploran los principios esenciales para clasificar y estructurar datos, optimizando su precisión y claridad en el análisis estadístico.

Durante el desarrollo de este componente, se busca que el aprendiz adquiera un conocimiento sólido sobre la clasificación de variables, el uso de escalas de medición y su relación con los métodos estadísticos aplicables.

Se incluyen técnicas de agrupación de datos, tanto para variables cualitativas como cuantitativas, detallando métodos que permiten evitar errores en la organización de los datos y asegurar una estructura coherente y confiable.

El componente también profundiza en cómo organizar datos en estudios con múltiples variables, utilizando matrices de datos y aplicando segmentación por dimensiones para facilitar el análisis de interacciones entre variables.

La organización de datos en niveles de medición apropiados facilita el uso de técnicas estadísticas adecuadas, mejorando la calidad y relevancia de los resultados obtenidos.

La estandarización de formatos desempeña un papel importante en la organización de datos, permitiendo una comparación precisa entre variables de diferentes contextos.

Este componente proporciona las herramientas para que el aprendiz pueda realizar análisis estadísticos efectivos y tomar decisiones informadas basadas en una estructura de datos organizada y bien fundamentada.

A lo largo del componente, se proporcionan ejemplos que permiten al aprendiz asociar y más tarde aplicar estos conceptos en situaciones reales de análisis estadístico.

¡Bienvenidos al mundo de la organización de datos!

1. Definición de variable y clasificación

1.1. Concepto de variable en estadística

En estadística, una **variable** es cualquier característica, atributo o propiedad que puede asumir diferentes valores dentro de una población o muestra. Las variables se usan en la recolección y análisis de datos, dado que permiten identificar y medir los fenómenos de interés en un estudio.

Una variable representa cualquier aspecto que pueda variar en un estudio. Los valores que toman las variables pueden ser cualitativos (descriptivos) o cuantitativos (numéricos). Las variables se estudian para identificar patrones, realizar comparaciones y extraer conclusiones estadísticas sobre una población.

Las variables en estadística se pueden clasificar de diferentes maneras según el tipo de valores que toman, su escala de medición o su función en un análisis. Esta clasificación es importante porque el tipo de variable determina los métodos estadísticos que se aplicarán para su análisis.

1.2. Variables cualitativas (categóricas)

Las variables cualitativas, también conocidas como categóricas, son aquellas que representan características o atributos de los individuos o elementos de un estudio que no se expresan en términos numéricos. En lugar de valores numéricos, estas variables asignan a los individuos en categorías o grupos con base en sus atributos. Las variables cualitativas son fundamentales en investigaciones donde el interés principal radica en clasificar y entender distribuciones de características específicas dentro de una población.

- **Variables nominales**: este tipo de variable cualitativa clasifica los datos en categorías sin un orden o jerarquía específica entre ellas. En otras palabras, las categorías son mutuamente exclusivas y colectivamente exhaustivas, pero no tienen un nivel de superioridad o inferioridad respecto a otras. Las variables nominales se emplean comúnmente en estudios donde se necesita una clasificación básica sin intención de establecer un orden.

  Ejemplos: algunas variables nominales comunes incluyen el género (masculino, femenino, otro), el tipo de sangre (A, B, AB, O), el color de ojos (azul, marrón, verde, negro) y el estado civil (soltero, casado, divorciado, viudo). En estos casos, cada categoría es única y no existe una relación jerárquica entre ellas.
- **Variables ordinales**: las variables ordinales, a diferencia de las nominales, no solo agrupan datos en categorías, sino que también establecen un orden o jerarquía entre estas categorías. Sin embargo, las diferencias cuantitativas entre las categorías no son precisas ni consistentes. Aunque se puede identificar un orden, no es posible medir numéricamente las distancias entre categorías de manera uniforme.

  Ejemplos: en el contexto educativo, una variable ordinal puede ser la denominación (primaria, secundaria, terciaria), donde existe una progresión en el grado de educación alcanzado, y el nivel de satisfacción (bajo, medio, alto), que indica una jerarquía en la percepción del usuario. En ambos casos, las categorías siguen un orden, pero no podemos cuantificar la diferencia exacta entre cada nivel.

Las variables cualitativas se usan en estudios de mercado, encuestas de satisfacción y ciencias sociales, ya que permiten segmentar a la población y analizar la distribución de ciertos atributos o características en un conjunto de datos.

1.3. Variables cuantitativas (numéricas)

Las variables cuantitativas son aquellas que se expresan en términos numéricos y permiten realizar operaciones matemáticas. Estas variables representan cantidades o medidas que pueden manipularse aritméticamente para proporcionar información sobre la distribución, tendencia central y dispersión de los datos. Las variables cuantitativas suelen aparecer en estudios donde se busca medir y analizar fenómenos con precisión numérica.

- **Variables discretas**: son aquellas que pueden tomar solo un conjunto finito o contable de valores, generalmente números enteros. Estas variables suelen representar conteos o eventos que no permiten fraccionamientos, lo que significa que los valores que asumen son enteros. Las variables discretas se utilizan comúnmente en estudios donde se cuantifican eventos o entidades individuales.

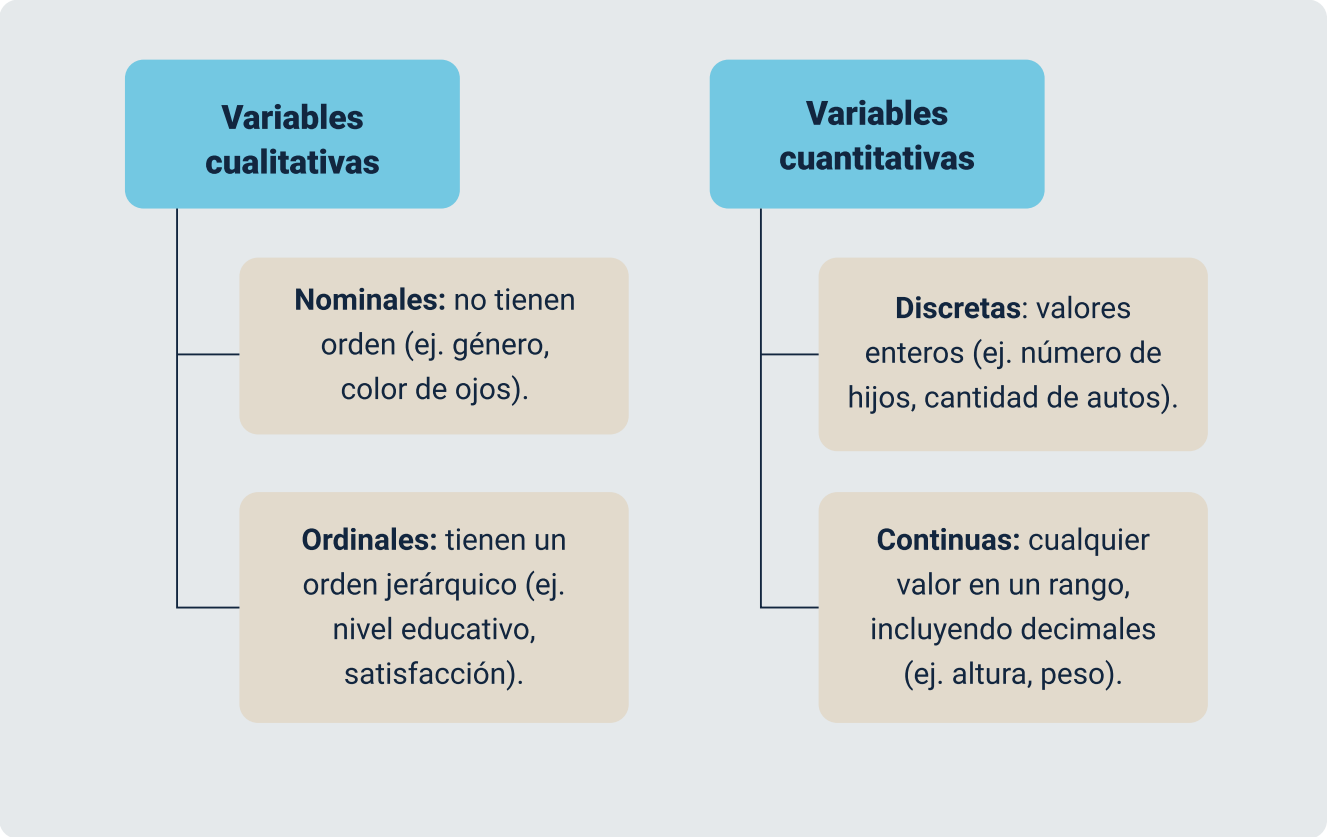
  Ejemplos: una representación de variable discreta es el número de hijos en una familia (0, 1, 2, ...), la cantidad de productos vendidos en una tienda (1, 2, 3, ...), y el número de autos que posee una familia (1, 2, 3, ...). En estos casos, las cantidades representadas no pueden dividirse en fracciones.
- **Variables continuas**: las variables continuas, a diferencia de las discretas, pueden tomar cualquier valor dentro de un rango específico, incluidos valores decimales. Estas variables son empleadas para medir fenómenos que pueden fraccionarse y que requieren un alto grado de precisión. Las

variables continuas permiten medir con gran exactitud, y son fundamentales en estudios donde es importante captar variaciones sutiles.

Ejemplos: una representación de variable continua es el peso de una persona (como 70.5 kg), la altura (como 1.75 m), el tiempo (como 12.34 segundos), y la temperatura (como 36.7°C). En estos casos, los valores pueden tomar infinitos puntos en el rango de medición, lo cual proporciona una precisión que es particularmente útil en ciencias experimentales, medicina y física.

Las variables cuantitativas se usan en estudios que requieren análisis estadísticos precisos, como el cálculo de la media, mediana, moda y desviación estándar. La posibilidad de realizar operaciones matemáticas permite que los investigadores obtengan información detallada sobre la dispersión y centralización de los datos en la muestra o población estudiada.

Figura 1. Clasificación de variables



Fuente. OIT, 2024.

1.4. Escalas de medición

Las variables también se pueden clasificar según la escala de medición que se utilice. Las escalas de medición determinan qué tipo de operaciones matemáticas o comparaciones se pueden realizar con los datos.

- **Escala nominal**: clasifica los datos en categorías que no tienen un orden. Es usada para variables cualitativas nominales.

  Ejemplo: estado civil (soltero, casado, divorciado).

- **Escala ordinal**: ordena los datos en categorías, pero no permite cuantificar las diferencias entre estas categorías. Es usada para variables cualitativas ordinales.

  Ejemplo: nivel de satisfacción (bajo, medio, alto).
- **Escala de intervalo**: mide la diferencia entre valores de manera cuantitativa, pero no tiene un "cero absoluto". Es usada para variables cuantitativas continuas.

  Ejemplo: temperatura en grados Celsius o Fahrenheit.
- **Escala de razón**: es similar a la escala de intervalo, pero con un "cero absoluto", lo que permite multiplicar o dividir los valores. Es usada para variables cuantitativas continuas.

  Ejemplo: peso, altura, ingresos.

1.5. Variables dependientes e independientes

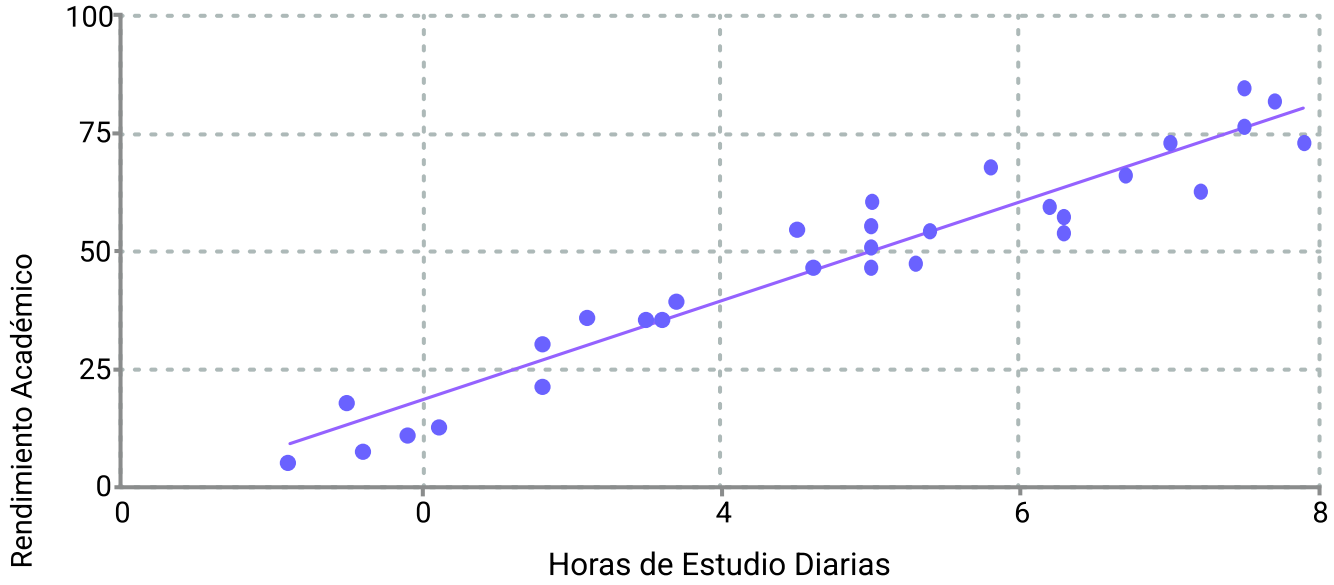
En los estudios experimentales o de investigación, las variables pueden clasificarse según su papel en el análisis:

- **Variable independiente**: es la variable que se manipula o cambia para observar su efecto en otra variable. También se la conoce como "predictora" o "explicativa".
- **Variable dependiente**: es la variable que se mide para ver cómo responde a los cambios en la variable independiente. También se la conoce como "respuesta" o "resultado".

Ejemplo: en una investigación sobre el efecto de horas de estudio en el rendimiento académico, la variable independiente sería las "horas de estudio", mientras que la variable dependiente sería el "rendimiento académico".

Al representar esta información en una figura, se puede observar lo siguiente:

Figura 2. Ejemplo variables dependientes e independientes



Fuente. OIT, 2024.

a) Variable Independiente (Eje X - Horas de Estudio)

- Es la variable que el estudiante puede controlar o manipular directamente.
- Se representa en el eje horizontal (X).
- Va aproximadamente de 1 a 8 horas de estudio diarias.
- Es independiente porque su valor no está condicionado por otros factores en nuestro estudio.

- El estudiante decide libremente cuántas horas dedicar al estudio.

b) Variable Dependiente (Eje Y - Rendimiento Académico)

- Es la variable que se ve afectada o "depende" de la variable independiente.
- Se representa en el eje vertical (Y).
- Se mide en una escala de rendimiento (puntos).
- Es dependiente porque su valor está influenciado por las horas de estudio.
- El rendimiento "depende" de cuánto tiempo estudia el alumno.

1.6. Relación entre los tipos de variables y los métodos estadísticos

El tipo de variable influye directamente en los métodos estadísticos que se pueden aplicar. Por ejemplo:

a) Para **variables cualitativas nominales**, se usan frecuencias y porcentajes para resumir los datos, y pruebas como el chi-cuadrado para analizar relaciones.

b) Para **variables ordinales**, se pueden usar medidas de tendencia central (como la mediana) y pruebas no paramétricas como la prueba de rangos de Wilcoxon.

c) Para **variables cuantitativas discretas y continuas**, se usan medidas como la media, la desviación estándar, y pruebas estadísticas paramétricas como la t de Student o ANOVA, dependiendo del contexto.

2. Niveles de medición

Los niveles de medición se usan ampliamente en estadística, ya que determinan el tipo de operaciones matemáticas que se pueden realizar con los datos y qué tipos de análisis estadísticos son apropiados. Cada variable se mide en uno de los cuatro niveles de medición: nominal, ordinal, de intervalo y de razón.

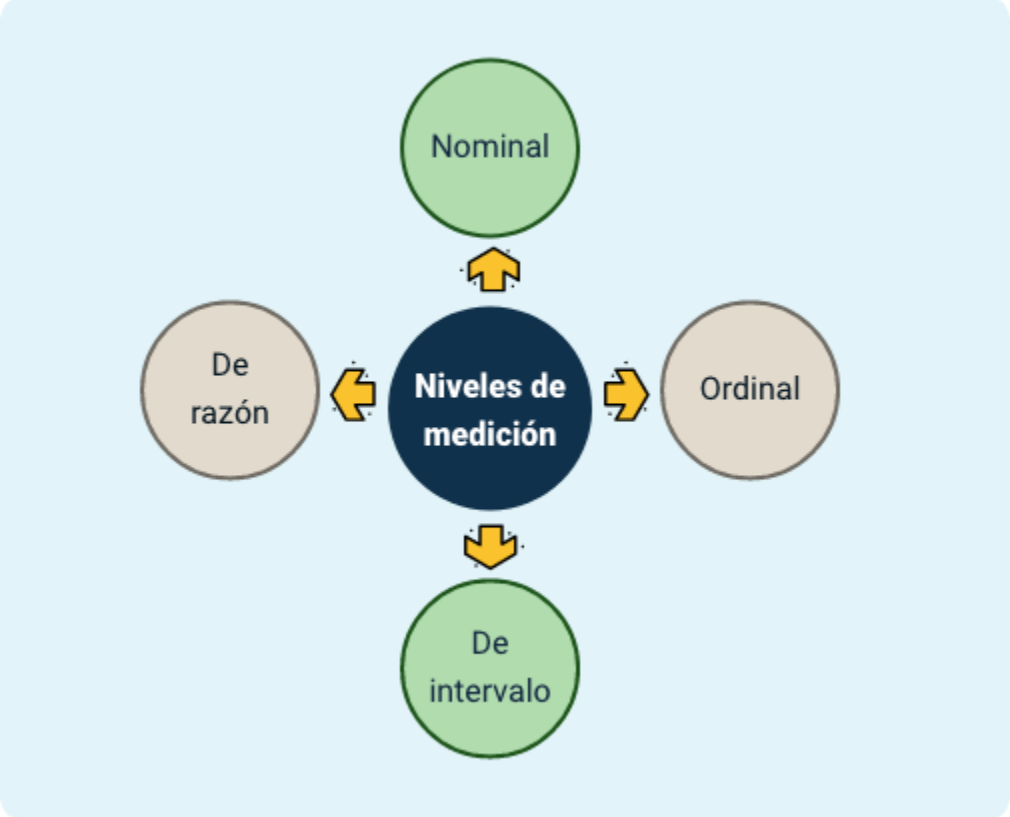
Cada nivel de medición posee sus propias limitaciones y fortalezas en cuanto a las operaciones matemáticas y los tipos de pruebas estadísticas que se pueden aplicar. Por ejemplo, analizar datos de nivel nominal con técnicas diseñadas para niveles de intervalo o de razón puede conducir a errores significativos en la interpretación, ya que no hay una base cuantitativa válida para realizar operaciones matemáticas avanzadas en datos categóricos.

Una elección incorrecta del nivel de medición puede llevar a conclusiones erróneas o a malinterpretar las relaciones entre variables, lo cual puede tener serias repercusiones, especialmente en ámbitos como la investigación científica, la economía y la toma de decisiones empresariales. Por ejemplo, si se trata una variable ordinal como si fuera de intervalo (sin considerar que las distancias entre categorías no son iguales), se podría concluir equivocadamente que las diferencias entre categorías son equivalentes, lo cual afectaría la validez de los resultados y podría influir negativamente en las decisiones basadas en estos datos.

Por otro lado, una selección precisa del nivel de medición asegura que los análisis estadísticos sean apropiados y que los resultados obtenidos sean robustos y significativos. Al utilizar el nivel de medición correcto, se garantiza que las técnicas estadísticas aplicadas correspondan a las características de los datos, maximizando así la precisión y confiabilidad de los hallazgos. Además, una elección adecuada permite un

uso más eficiente de los datos, ya que cada nivel de medición abre la puerta a distintos métodos analíticos, desde simples conteos hasta modelos complejos de regresión y análisis de varianza. En última instancia, una correcta selección del nivel de medición permite obtener insights más profundos y relevantes, promoviendo decisiones informadas y basadas en evidencia.

Figura 3. Niveles de medición



Fuente. OIT, 2024.

2.1. Nivel de medición nominal

El nivel nominal es el más básico de los cuatro niveles. Se utiliza para variables cualitativas que agrupan a los individuos u objetos en categorías sin un orden

intrínseco. No existe un valor numérico asociado con estas categorías, y no es posible realizar operaciones matemáticas con ellas.

a) Características

- Los datos se agrupan en categorías o etiquetas.
- No existe un orden ni jerarquía entre las categorías.
- No es posible hacer cálculos matemáticos entre los valores.

b) Ejemplos

- **Género**: masculino, femenino, otro.
- **Color de ojos**: azul, marrón, verde, negro.

El análisis de las variables nominales se basa en la frecuencia de aparición de cada categoría. Por ejemplo, en un estudio de mercado, podríamos observar cuántos clientes prefieren una marca específica.

2.2. Nivel de medición ordinal

El nivel ordinal también se utiliza para variables cualitativas, pero a diferencia del nivel nominal, las categorías tienen un orden o jerarquía. Sin embargo, las diferencias entre las categorías no son cuantificables. Aunque se puede establecer un rango, no se pueden realizar cálculos precisos con las diferencias entre las categorías.

a) Características

- Los datos se agrupan en categorías con un orden o jerarquía.
- No se puede medir la magnitud exacta de la diferencia entre las categorías.
- No se pueden realizar operaciones aritméticas entre los valores.

b) Ejemplo

- **Nivel de satisfacción**: insatisfecho, neutral, satisfecho.
- **Clasificación en una carrera**: primer lugar, segundo lugar, tercer lugar.

El análisis estadístico de las variables ordinales puede incluir el cálculo de la mediana o el uso de pruebas no paramétricas, como la prueba de rangos de Wilcoxon. En el caso de una encuesta sobre la satisfacción del cliente, se puede calcular el porcentaje de clientes en cada nivel de satisfacción.

2.3. Nivel de medición de intervalo

El **nivel de intervalo** es el primer nivel que se usa para variables cuantitativas. En este nivel, los datos no solo tienen un orden, sino que las diferencias entre los valores son medibles y significativas. Sin embargo, no existe un "cero absoluto", lo que significa que no se pueden realizar operaciones multiplicativas entre los valores.

a) Características

- Las diferencias entre los valores son consistentes y medibles.
- No existe un "cero absoluto"; el valor cero es arbitrario.
- Se pueden realizar operaciones aritméticas como la suma o la resta, pero no la multiplicación o división.

b) Ejemplo

- **Temperatura en grados Celsius**: 20 °C, 30 °C, 40 °C. Aquí, la diferencia entre 20 °C y 30 °C es la misma que entre 30 °C y 40 °C, pero 0 °C no representa una ausencia de temperatura.
- **Fechas en un calendario**: la diferencia entre el año 2000 y el año 1990 es de 10 años, pero el año "0" no es un punto de referencia absoluto.

En el análisis estadístico de variables de intervalo, se pueden calcular medidas como la media o la desviación estándar. Sin embargo, la falta de un "cero absoluto" impide que se realicen ciertas operaciones, como comparar razones entre valores.

2.4. Nivel de medición de razón

El nivel de razón es el nivel más alto de medición y se aplica a variables cuantitativas que, además de tener diferencias medibles entre los valores, cuentan con un "cero absoluto". Esto significa que se pueden realizar todas las operaciones aritméticas, incluidas la multiplicación y la división, ya que el cero en este nivel indica la ausencia total de la característica medida.

a) Características

- Existen diferencias consistentes y medibles entre los valores.
- Existe un "cero absoluto", lo que permite realizar operaciones multiplicativas.
- Se pueden realizar todas las operaciones aritméticas: suma, resta, multiplicación y división.

b) Ejemplos

- **Peso**: un objeto que pesa 0 kg no tiene masa. Es decir, un objeto que pesa 4 kg es el doble de pesado que uno que pesa 2 kg.
- **Ingresos**: si una persona tiene 0 ingresos, no tiene ganancias. Si una persona gana $1000, puede decirse que gana el doble que alguien que gana $500.

El análisis estadístico de las variables de razón permite realizar cualquier tipo de operación matemática y aplicar técnicas avanzadas de análisis, como regresión y

correlación. Por ejemplo, se puede calcular el promedio de ingresos de un grupo de personas o determinar la relación entre el peso de un grupo de personas y su altura.

2.5. Importancia de los niveles de medición en el análisis estadístico

El nivel de medición de una variable tiene un impacto directo en el tipo de análisis que se puede realizar. Seleccionar el método estadístico adecuado depende de conocer correctamente el nivel de medición de las variables involucradas.

Ejemplo

- **Variables nominales**: se analizan usando frecuencias y proporciones.
- **Variables ordinales**: se pueden resumir utilizando la mediana y realizar análisis no paramétricos.
- **Variables de intervalo y de razón**: se utilizan para análisis estadísticos avanzados que incluyen medidas de tendencia central, dispersión y pruebas paramétricas.

Comprender el nivel de medición es fundamental para evitar errores en la interpretación de los resultados y para garantizar que las operaciones estadísticas aplicadas sean las correctas.

3. Técnicas de agrupación de datos

3.1. Definición de datos agrupados y no agrupados

- **Datos no agrupados**: son aquellos datos presentados individualmente, tal como se recolectan. Cada valor permanece sin clasificar, lo cual es útil en conjuntos de datos pequeños o cuando es importante mantener el valor exacto de cada observación.
- **Datos agrupados**: consisten en datos organizados en clases o intervalos, lo que reduce el número de categorías a observar y permite una visión más clara de las tendencias generales. Este método es ideal para grandes volúmenes de datos y facilita la creación de tablas de frecuencias y gráficos.

3.2. Criterios para agrupar datos

La agrupación de datos debe realizarse de manera coherente, siguiendo ciertos criterios para garantizar su eficacia y precisión:

- **Naturaleza de la variable**: las variables cuantitativas continuas se agrupan fácilmente en intervalos, mientras que las variables cualitativas suelen agruparse en categorías definidas.
- **Rango de datos**: determinar el rango ayuda a decidir el número de clases o intervalos. El rango se calcula como la diferencia entre el valor máximo y el mínimo en el conjunto de datos.
- **Número de clases**: generalmente, se elige entre 5 y 20 clases para equilibrar la simplificación con el nivel de detalle.

- **Amplitud de los intervalos**: debe ser constante para cada clase en la mayoría de los casos, lo cual facilita la comparación entre intervalos.

3.3. Métodos de agrupación de datos cuantitativos

Para las variables cuantitativas, las técnicas de agrupación más comunes incluyen:

a) **Intervalos de clase**: consiste en dividir el rango de los datos en intervalos de igual amplitud. Cada intervalo representa un rango de valores en el conjunto de datos.
   - **Ejemplo**: en un conjunto de edades que va de 18 a 60, podríamos definir intervalos de 5 en 5 (18-22, 23-27, etc.).

b) **Agrupación de datos en frecuencias absolutas y relativas**: se cuenta el número de observaciones en cada intervalo, generando una tabla de frecuencias absolutas (número de casos en cada intervalo) y una tabla de frecuencias relativas (proporción o porcentaje de casos en cada intervalo).

c) **Intervalos de clase no uniformes**: a veces se eligen intervalos de distinta amplitud si ciertas partes del conjunto de datos requieren mayor detalle. Esto es útil en conjuntos con una gran concentración de datos en ciertos rangos.

3.4. Métodos de agrupación de datos cualitativos

Para las variables cualitativas, la agrupación se realiza mediante categorías o clases que representan cada valor único o grupo de valores en los datos. Las técnicas incluyen:

- **Clasificación por categorías**: los datos se agrupan en categorías o clases según el valor que representan. Por ejemplo, la variable “nivel educativo” puede agruparse en categorías como primaria, secundaria y superior.
- **Agrupación por frecuencias de categorías**: similar a los datos cuantitativos, se cuenta el número de veces que cada categoría aparece y se presentan en tablas de frecuencias, que pueden ser absolutas o relativas.

4. Organización de la muestra de datos

Es importante tener en cuenta que la organización adecuada de los datos facilita el procesamiento, reduce errores y contribuye a obtener resultados confiables.

4.1. Criterios para organizar datos según el tipo de variable

La organización de datos depende del tipo de variable (cualitativa o cuantitativa) y su nivel de medición (nominal, ordinal, de intervalo o de razón). Estos criterios determinan cómo agrupar y presentar los datos:

- **Variables cualitativas**: usualmente, se organizan en categorías o clases, facilitando su representación en tablas de frecuencias o gráficos de barras.
- **Variables cuantitativas**: pueden organizarse en intervalos o rangos, según su naturaleza (discreta o continua), y representarse en histogramas o polígonos de frecuencias.

4.2. Técnicas de agrupación de datos para análisis estadístico

Agrupar datos sirve para simplificar su análisis y destacar patrones. Las principales técnicas de agrupación incluyen:

- **Intervalos de clase**: se utilizan para variables cuantitativas continuas, agrupando los valores en rangos (por ejemplo, edades de 20-30, 30-40).
- **Categorías**: para variables cualitativas, se agrupan en categorías predefinidas (por ejemplo, nivel educativo: primaria, secundaria, superior).
- **Agrupación por niveles de importancia**: según el impacto de las variables en el estudio, se pueden agrupar en conjuntos prioritarios.

4.3. Importancia de una organización coherente de los datos

Una organización coherente facilita la interpretación y análisis, garantizando que los datos sean accesibles y claros. Los beneficios de una organización adecuada incluyen:

- **Claridad en el análisis:** para identificar tendencias y patrones.
- **Precisión en los resultados**: que minimiza errores y evita interpretaciones erróneas.
- **Eficiencia en el procesamiento de datos:** que agiliza el análisis y reduce la posibilidad de duplicados o inconsistencias.

4.4. Métodos para evitar errores en la organización de datos

La organización de datos puede ser susceptible a errores, especialmente en conjuntos grandes o complejos. Los métodos para reducir errores incluyen:

Tabla 1. Métodos para evitar errores en la organización de datos

Método	Descripción	Ejemplos de aplicación	Beneficios
Estandarización de formatos.	Aplicar un formato consistente en nombres, fechas, unidades, etc., para evitar errores de interpretación y entrada.	• Usar el formato ISO para fechas (AAAA-MM-DD). • Unificar el uso de unidades, como kg o cm.	Facilita la comprensión y reduce errores en el procesamiento de datos.
Validación de entradas.	Implementar reglas que limiten los valores aceptables para cada campo o variable, evitando entradas erróneas.	• Definir un rango válido para la edad (ej. 0-120 años). • Limitar los valores de género a opciones definidas	Ayuda a detectar y corregir entradas inusuales o incorrectas.

Método	Descripción	Ejemplos de aplicación	Beneficios
		(masculino, femenino, otro).	
Control de duplicados.	Verificar la existencia de datos duplicados para evitar la redundancia y posibles errores en los resultados del análisis.	• Identificar registros duplicados por ID o nombres en bases de datos. • Usar funciones de duplicación en herramientas como Excel o SQL.	Previene la sobrestimación de resultados y mejora la precisión.
Manejo de datos faltantes.	Establecer procedimientos para identificar, manejar y documentar datos ausentes, evitando que afecten el análisis.	• Rellenar valores nulos con la media o mediana del conjunto. • Usar la imputación predictiva para completar valores faltantes.	Aumenta la integridad de los datos y reduce el sesgo.
Automatización de procesos.	Utilizar scripts o programas para automatizar la limpieza y organización de datos, reduciendo errores manuales.	• Automatizar la eliminación de espacios en blanco en nombres. • Programar scripts de Python para limpieza sistemática de datos.	Reduce el tiempo y minimiza errores humanos.
Control de versiones.	Usar herramientas de control de versiones para registrar cambios y revertir a versiones anteriores en caso de error.	• Usar Git para mantener un historial de cambios en los datos. • Crear respaldos automáticos de bases de datos.	Permite rastrear errores y revertir cambios problemáticos.

Método	Descripción	Ejemplos de aplicación	Beneficios
Verificación cruzada	Comparar los datos ingresados con fuentes o registros originales para confirmar su exactitud.	• Comparar datos ingresados con documentos originales. • Verificar datos financieros contra reportes de contabilidad.	Asegura la precisión de los datos en los registros.
Documentación de procesos.	Registrar todos los pasos de limpieza y organización para mantener un registro claro del tratamiento de datos.	• Documentar cambios y reglas de transformación en una guía. • Registrar métodos usados para manejar datos faltantes o anómalos.	Facilita la reproducción y auditoría del proceso.
Uso de herramientas de calidad de datos.	Utilizar software especializado para identificar y corregir errores comunes en los datos.	• Herramientas como Talend o OpenRefine para identificar inconsistencias. • Plugins en Excel para validar entradas.	Mejora la precisión y reduce el tiempo de limpieza de datos.
Análisis de valores atípicos.	Identificar y revisar valores extremos o inusuales para confirmar su validez antes de su inclusión en el análisis.	• Detectar valores fuera de los percentiles 1 y 99. • Revisar valores anómalos con gráficos de dispersión.	Reduce el riesgo de incluir datos incorrectos o sesgados.

Fuente. OIT, 2024.

4.5. Organización de datos en estudios con múltiples variables

En investigaciones que involucran múltiples variables, la correcta organización de los datos es fundamental para permitir un análisis exhaustivo y coherente. Cuando los estudios incluyen diversas dimensiones y aspectos de análisis, estructurar los datos de manera adecuada facilita la identificación de patrones, interacciones y relaciones entre variables.

- **Matrices de datos**: las matrices o tablas de datos son una herramienta para estructurar múltiples variables asociadas a cada observación o individuo dentro de un estudio. En estas matrices, cada fila representa una observación o caso, mientras que cada columna contiene los valores de una variable específica. Esta organización tabular permite analizar múltiples variables de forma simultánea, facilitando el uso de métodos estadísticos multivariados y el análisis comparativo entre observaciones. Este formato es empleado para estudios longitudinales o de cohortes, donde se necesita hacer un seguimiento de varias características a lo largo del tiempo.
- **Segmentación por dimensiones**: la segmentación de datos según dimensiones relevantes, como el tiempo, el lugar, o cualquier otra categoría contextual, facilita el análisis detallado y la exploración de interacciones complejas entre variables. Por ejemplo, en un estudio que evalúa el impacto de distintos tratamientos médicos en diferentes ubicaciones geográficas, los datos pueden segmentarse por región y por periodos de tiempo, lo cual permite observar cómo los efectos de los tratamientos varían según estas dimensiones. Esta segmentación es útil en

análisis factoriales y análisis de correlación donde se exploran relaciones entre variables en contextos específicos.

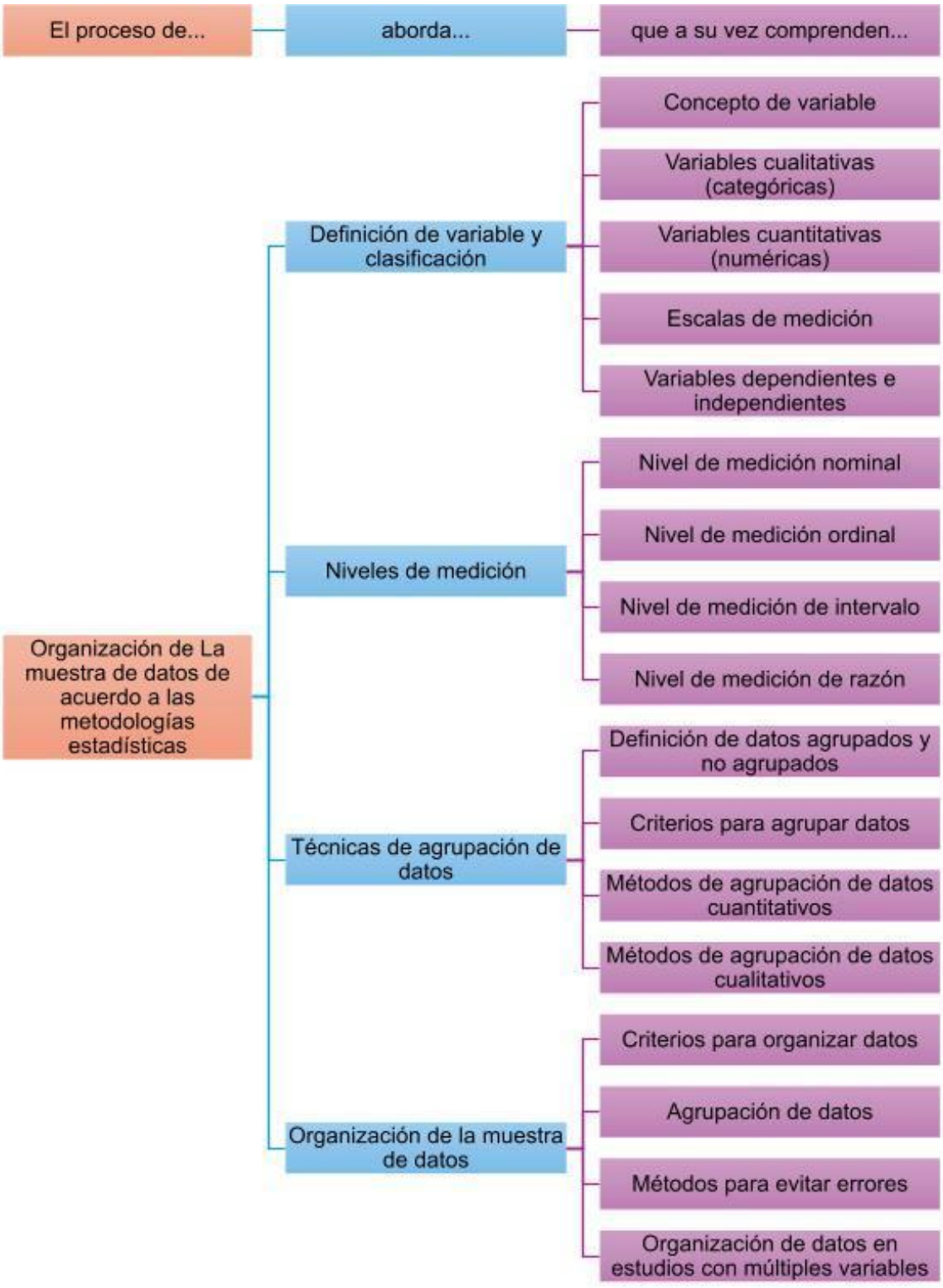
- **Estandarización de formatos**: la estandarización se usa cuando se manejan datos de múltiples variables, especialmente si provienen de distintas fuentes o si se recolectan en diferentes condiciones. Estandarizar el formato de los datos significa aplicar una uniformidad en la forma en que se registran las unidades de medida, los formatos de fecha, la nomenclatura de las categorías, entre otros aspectos. Esta uniformidad permite reducir errores, evitar ambigüedades y facilita el procesamiento de los datos. La estandarización asegura que todas las variables sean compatibles para comparaciones directas y análisis conjuntos, optimizando la precisión en los resultados y permitiendo una interpretación más confiable de las interacciones entre variables.

Síntesis

El siguiente diagrama proporciona una visión general sintetizada de los principales temas abordados en este componente sobre la organización de la muestra de datos de acuerdo a las metodologías estadísticas. Este mapa está diseñado para ayudar al lector a visualizar la interrelación entre las diversas áreas que constituyen el proceso de organización y estructuración de datos en el contexto estadístico, permitiendo un análisis más coherente y efectivo.

En el origen del diagrama se encuentra el concepto principal de "Organización de la muestra de datos de acuerdo a las metodologías estadísticas", del cual se derivan temas fundamentales: definición de variable y clasificación, niveles de medición, técnicas de agrupación de datos y organización de la muestra de datos. Cada una de estas áreas se desglosa en conceptos clave, reflejando la estructura y el contenido del componente.

Este diagrama actúa como una guía visual para explorar los conceptos presentados en el texto, permitiendo al lector comprender rápidamente la amplitud y la organización de los temas tratados, así como sus conexiones. Al revisar este mapa, el lector podrá observar cómo los diferentes aspectos de la estadística se entrelazan para formar un proceso integral y sistemático en la organización de datos. Se invita a utilizar este diagrama como un complemento al contenido detallado del componente, sirviendo como una referencia rápida y un recordatorio visual de los conceptos esenciales en la organización y manejo de datos en estudios estadísticos.



Fuente. OIT, 2024.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
El muestreo en procesos estadísticos	Ecosistema de Recursos Educativos Digitales SENA. (2022a, agosto 26). El muestreo en procesos estadísticos.	Video	https://www.youtube.com/watch?v=yGbtOWCHY4I
Estadística descriptiva, gráficas e informes estadísticos	Ecosistema de Recursos Educativos Digitales SENA. (2023b, marzo 24). Estadística descriptiva, gráficas e informes estadísticos.	Video	https://www.youtube.com/watch?v=v5UMlXHe2nM
Estadísticas básicas	Ecosistema de Recursos Educativos Digitales SENA. (2022c, diciembre 9). Estadísticas básicas.	Video	https://www.youtube.com/watch?v=UKdYYtCdvKw
Etapas del procesamiento de datos y métodos estadísticos - Introducción	Ecosistema de Recursos Educativos Digitales SENA. (2023d, marzo 26). Etapas del procesamiento de datos y métodos estadísticos - Introducción.	Video	https://www.youtube.com/watch?v=ndzj15PQEVw
Introducción a la aplicación de herramientas estadísticas en la presentación de datos.	Ecosistema de Recursos Educativos Digitales SENA. (2023c, marzo 24). Introducción a la aplicación de herramientas estadísticas en la presentación de datos.	Video	https://www.youtube.com/watch?v=M9q9zxX8Evc
Introducción a la estadística.	Ecosistema de Recursos Educativos Digitales SENA.	Video	https://www.youtube.com/watch?v=wMCDkknpUVw

Tema	Referencia	Tipo de material	Enlace del recurso
	(2023e, septiembre 20). Introducción a la estadística.		
Modelos matemáticos y estadísticos aplicados en Big Data – Introducción	Ecosistema de Recursos Educativos Digitales SENA. (2023a, marzo 23). Modelos matemáticos y estadísticos aplicados en Big Data – Introducción.	Video	https://www.youtube.com/watch?v=eH2X2oWgggk
Principales elementos de la estadística	Ecosistema de Recursos Educativos Digitales SENA. (2022b, octubre 26). Principales elementos de la estadística.	Video	https://www.youtube.com/watch?v=Ad5gxB9PhKQ

Glosario

Agrupación de datos: proceso de organizar datos en categorías o intervalos para facilitar su análisis.

Amplitud de intervalos: rango o tamaño de cada clase en una agrupación de datos cuantitativos.

Análisis estadístico: estudio detallado de datos mediante técnicas estadísticas para extraer conclusiones.

Categorización: clasificación de datos en grupos específicos según sus características.

Clasificación: organización de datos o variables en grupos o tipos.

Coherencia de datos: consistencia en la organización de los datos para evitar errores en el análisis.

Control de calidad: procedimientos para asegurar la precisión y confiabilidad de los datos.

Datos agrupados: datos organizados en clases o intervalos, ideales para volúmenes grandes.

Datos cualitativos: datos que representan categorías o atributos sin valores numéricos.

Datos cuantitativos: datos numéricos que representan cantidades y permiten operaciones matemáticas.

Escala de intervalo: nivel de medición para variables cuantitativas sin un cero absoluto; permite suma y resta.

Escala de medición: tipo de escala utilizada para medir una variable, como nominal, ordinal, intervalo o razón.

Escala de razón: nivel de medición con un cero absoluto que permite todas las operaciones matemáticas.

Escala nominal: nivel de medición que clasifica datos en categorías sin orden ni jerarquía.

Escala ordinal: nivel de medición que organiza datos en categorías con un orden pero sin cuantificar las diferencias.

Frecuencia: número de veces que ocurre una categoría o valor en un conjunto de datos.

Intervalo: rango de valores dentro de un conjunto de datos agrupados.

Medición: proceso de asignar valores a una variable de acuerdo con una escala específica.

Variable dependiente: variable cuyo valor depende de la manipulación de otra variable, llamada independiente.

Variable independiente: variable que se manipula para observar su efecto en otra variable dependiente.

Referencias bibliográficas

Batanero, C. (2001). Didáctica de la estadística. Granada: Universidad de Granada.

Cochran, W. G. (1980). Técnicas de muestreo (3.ª ed.). México: CECSA.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). Metodología de la investigación (6.ª ed.). México: McGraw-Hill.

Martínez, J. (2004). Muestreo estadístico. Madrid: Alianza Editorial.

Montgomery, D. C., & Runger, G. C. (2015). Probabilidad y estadística aplicada a la ingeniería (5.ª ed.). México: McGraw-Hill.

Scheaffer, R. L., Mendenhall, W., & Ott, R. L. (2007). Elementos de muestreo (6.ª ed.). México: Thomson.

Triola, M. F. (2018). Estadística (12.ª ed.). México: Pearson Educación.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). Probabilidad y estadística para ingenieros (9.ª ed.). México: Pearson Educación.

Créditos

Elaborado por:

Organización Internacional del Trabajo