Fuentes y técnicas de recolección de datos para análisis

Breve descripción:

Este componente aborda la recolección de datos, desde conceptos básicos hasta métodos avanzados de muestreo. Explora la diferencia entre población y muestra, tipos de muestreo, y la importancia de elegir fuentes confiables. Introduce herramientas de inferencia estadística para obtener datos representativos. Orientado a nivel técnico, ofrece una visión completa y práctica del proceso de recolección de datos para análisis estadístico.

Noviembre 2024

Tabla de contenido

Introducción

En el ámbito del análisis de datos, contar con fuentes confiables y aplicar técnicas adecuadas de recolección de información es crucial para obtener resultados precisos y valiosos. La calidad de los datos recopilados y su adecuada selección determinan en gran medida la utilidad y precisión de los análisis posteriores, ya que datos incompletos, sesgados o mal interpretados pueden llevar a conclusiones erróneas o poco útiles.

¿De dónde provienen los datos y cómo se asegura su validez y pertinencia para un análisis efectivo? Este componente formativo profundiza en estas cuestiones esenciales, explorando las fuentes de datos más comunes, las técnicas de recolección, y los métodos para evaluar su confiabilidad. A través de una combinación de teoría y práctica, se cubrirán tanto fuentes tradicionales como datos obtenidos de plataformas digitales y redes sociales, bases de datos públicas, encuestas y estudios de campo.

Durante el desarrollo de este curso, el aprendiz adquirirá habilidades para identificar, seleccionar y validar diversas fuentes de información, adaptando las técnicas de recolección a diferentes contextos y tipos de datos. Mediante estudios de caso y ejercicios prácticos, se abordarán las mejores prácticas para la recolección de datos en el entorno actual, con herramientas digitales y enfoques estructurados que garantizan datos de calidad para el análisis.

El proceso de recolección de datos es la base sobre la cual se construye cualquier análisis exitoso. Como señala un principio fundamental en el campo del análisis: la precisión del análisis es tan buena como la precisión de los datos que lo sustentan.

¡Te damos una maravillosa bienvenida a las fuentes y técnicas de recolección de datos!

Video 1. Fuentes y técnicas de recolección de datos para análisis

Enlace de reproducción del video

Síntesis del video: Fuentes y técnicas de recolección de datos para análisis
En el componente formativo «Fuentes y técnicas para recolección de datos» se exploran los métodos y estrategias esenciales para obtener datos de manera confiable y representativa, fundamentales para un análisis estadístico efectivo y riguroso.

Durante el desarrollo del componente, se pretende la comprensión integral del proceso de recolección de datos, desde la selección de muestras y fuentes hasta la aplicación de técnicas de inferencia.

La recolección de datos se apoya en la comprensión de conceptos de estadística, incluyendo la distinción entre población y muestra, así como la importancia de la representatividad en los resultados.

El componente profundiza en las diferentes técnicas de muestreo como el muestreo aleatorio simple, estratificado y por conglomerados, cada una adaptada a distintos contextos y necesidades de precisión.

Las fuentes de datos son otro pilar de este componente. Estas se clasifican en fuentes primarias, como encuestas y experimentos, y fuentes secundarias, como bases de datos públicas y reportes previos.

La elección de las fuentes depende del objetivo del estudio y requiere evaluar la confiabilidad y validez de los datos obtenidos.

La inferencia estadística se usa para generar los hallazgos de una muestra a la población. Este proceso incluye técnicas como la estimación puntual, intervalos de confianza y pruebas de hipótesis.

El componente también aborda los requerimientos para la recolección de datos, incluyendo la identificación de variables y la distinción entre datos cuantitativos y cualitativos.

¡Bienvenidos al mundo de la recolección de datos para análisis estadístico!

1. Conceptos generales de estadística

La estadística es una disciplina matemática que se ocupa de la recolección, análisis, interpretación y presentación de datos. Es esencial en diversos campos como la economía, la biología, la ingeniería, y las ciencias sociales, ya que proporciona las herramientas para realizar investigaciones basadas en datos y tomar decisiones informadas.

1.1. Definición de estadística y su propósito

La estadística se puede definir como la ciencia que estudia cómo recoger, organizar, analizar e interpretar datos para extraer conclusiones y tomar decisiones. Los datos pueden provenir de diversas fuentes y representan observaciones cuantitativas o cualitativas sobre algún fenómeno o conjunto de fenómenos.

El propósito principal de la estadística es transformar estos datos en información útil para facilitar la toma de decisiones, describir situaciones o fenómenos y hacer predicciones sobre comportamientos futuros. La estadística proporciona métodos tanto para describir lo que ocurre en un conjunto de datos (estadística descriptiva) como para hacer inferencias y conclusiones más generales a partir de esos datos (estadística inferencial).

1.2. Clasificación de la estadística: descriptiva e inferencial

La estadística se divide en dos ramas principales: la **estadística descriptiva** y la **estadística inferencial**.

Tabla 1. Clasificación de la estadística

Rama de la Estadística	Descripción	Objetivo	Herramientas Principales	Resultados
Estadística Descriptiva	Se refiere a los métodos que permiten resumir y describir las características principales de un conjunto de datos. No hace generalizaciones o inferencias más allá de los datos analizados; se limita a describir lo que se observa.	Resumir y describir datos de una muestra o población.	**Medidas de tendencia central**: Media, mediana, moda. **Medidas de dispersión**: Desviación estándar, varianza. **Representaciones gráficas**: Histogramas, diagramas de barras, tablas de frecuencias.	Proporciona información exacta y precisa sobre los datos en cuestión.
Estadística Inferencial	Va más allá de la descripción de los datos y se centra en hacer generalizaciones sobre una población a partir de una muestra. Utiliza métodos probabilísticos para estimar parámetros poblacionales, realizar pruebas de hipótesis y hacer predicciones.	Hacer generalizaciones y predicciones sobre la población a partir de una muestra representativa.	Intervalos de confianza. Pruebas de hipótesis. Estimación de parámetros.	Trabaja con probabilidades e incertidumbre, ya que las conclusiones se basan en una muestra y no en toda la población. Permite hacer predicciones y generalizaciones aplicables a la población en general.

Fuente. OIT, 2024.

Ejemplo: **diferencia entre estadística descriptiva e inferencial.**

Supongamos que estamos analizando las edades de los empleados en una empresa y tenemos los siguientes datos de una muestra de 10 empleados:

Edades = {25, 30, 35, 40, 45, 50, 55, 60, 65, 70}

- **Estadística descriptiva**: podemos calcular medidas como la media y la desviación estándar de la muestra.

$\bar{x} = (25 + 30 + 35 + 40 + 45 + 50 + 55 + 60 + 65 + 70) / 10 = 47.5$ años

$\sigma = \sqrt{\sum(xi - \bar{x})^2 / n} = \sqrt{((25 - 47.5)^2 + ... + (70 - 47.5)^2 / 10} \approx 15.14$ años

- **Estadística inferencial**: si tomamos una muestra similar de otra empresa con 100 empleados, podemos usar un intervalo de confianza para estimar la media de la población de empleados.

Si la media muestral es $\bar{x} = 47.5$, con un error estándar de $SE = 15.14 / \sqrt{10} \approx 4.79$, el intervalo de confianza del 95% sería:

$IC = 47.5 \pm 1.96 * 4.79 = (38.12, 56.88)$

Esto significa que con un 95% de confianza, la edad promedio de la población de empleados está entre 38.12 y 56.88 años

1.3. Aplicaciones prácticas de la estadística en la recolección de datos

La estadística se aplica en numerosas áreas del conocimiento y en diversas situaciones del mundo real. Algunas de sus aplicaciones más comunes incluyen:

- **Investigación médica**: en estudios clínicos, la estadística se utiliza para determinar la efectividad de un tratamiento a partir de datos obtenidos de

un grupo de pacientes. Los investigadores recolectan datos sobre el estado de salud de los pacientes antes y después de recibir el tratamiento, y luego usan métodos estadísticos para comparar los resultados.

- **Ciencia social**: en estudios de encuestas o investigaciones sociales, la estadística descriptiva puede resumir las respuestas de los encuestados sobre diversos temas, como la opinión política o el nivel de satisfacción con servicios públicos. La estadística inferencial se utiliza para hacer inferencias sobre el comportamiento de toda la población a partir de las respuestas de una muestra representativa.
- **Marketing**: las empresas utilizan la estadística para analizar patrones de compra de los consumidores y predecir tendencias futuras. Los análisis descriptivos permiten entender el comportamiento de compra actual, mientras que los análisis inferenciales pueden ayudar a predecir futuras ventas y adaptar estrategias de marketing.
- **Producción industrial**: en control de calidad, la estadística es fundamental para monitorear procesos y asegurar que los productos cumplan con los estándares requeridos. Se recolectan datos sobre muestras de productos fabricados, y a partir de los análisis estadísticos se decide si el proceso debe ajustarse o continuar.
- **Investigación científica**: en campos como la biología o la física, la estadística es vital para interpretar los resultados de experimentos. Los científicos recopilan datos experimentales, los analizan y hacen inferencias sobre teorías o hipótesis a partir de esos datos.

2. Población y muestra en estadística

En el contexto de la estadística, tanto la población como la muestra juegan roles clave en la recolección y análisis de datos. Entender estos conceptos es esencial para garantizar la precisión y relevancia de los estudios estadísticos.

2.1. Diferencias entre población y muestra

- **Población**: en estadística, una población es el conjunto completo de individuos, objetos, eventos o medidas de interés sobre los que se desea obtener información. La población puede ser finita (como el número total de estudiantes en una universidad) o infinita (como el número de tiradas de un dado en un experimento).

  Ejemplo: si se desea investigar el comportamiento de compra de todos los consumidores de un país, todos los consumidores registrados conforman la población.

- **Muestra**: una muestra es un subconjunto de la población seleccionada para ser estudiada. Debido a que, en la mayoría de los casos, es imposible o impráctico estudiar a toda la población, se selecciona una muestra representativa para extraer conclusiones. Una muestra adecuada debe reflejar las características de la población para que los resultados puedan ser generalizados.

  Ejemplo: en un estudio sobre el comportamiento de compra, un grupo de 1,000 consumidores seleccionados al azar podría ser la muestra de una población de millones de consumidores.

2.2. Criterios para definir una muestra representativa

Para que los resultados de un estudio estadístico sean válidos, la muestra debe ser representativa de la población. Esto significa que debe reflejar las características clave de la población, como la distribución de edad, género, ingresos, etc. Para garantizar que una muestra sea representativa, es importante seguir estos criterios:

- **Aleatoriedad**: la selección de los elementos de la muestra debe ser aleatoria para evitar sesgos. En una muestra aleatoria, cada elemento de la población tiene la misma probabilidad de ser seleccionado, lo que minimiza la posibilidad de errores sistemáticos.
- **Tamaño adecuado**: el tamaño de la muestra es un aspecto importante a tener en cuenta. Si la muestra es demasiado pequeña, puede no representar adecuadamente la variabilidad de la población. Si es demasiado grande, el análisis puede ser costoso y complicado sin un aumento significativo en la precisión de los resultados.
- **Heterogeneidad de la población**: si la población es muy diversa, es importante asegurarse de que la muestra incluya subgrupos que representen esa diversidad. Técnicas como el muestreo estratificado se utilizan para garantizar que se incluyan todas las categorías importantes de la población.

2.3. Criterios para seleccionar una muestra representativa

Existen varias técnicas y criterios que se deben seguir para seleccionar una muestra que realmente represente a la población. Algunos de los métodos más comunes incluyen:

- **Muestreo aleatorio simple**: en este método, todos los elementos de la población tienen la misma probabilidad de ser seleccionados. Es uno de los métodos más fáciles y efectivos para obtener una muestra representativa.
- **Muestreo sistemático**: consiste en seleccionar cada k-ésimo elemento de la población después de un punto de partida aleatorio. Este método es útil cuando los datos están organizados de manera secuencial.
- **Muestreo estratificado**: la población se divide en grupos o "estratos" basados en una característica relevante (como género o edad), y luego se selecciona una muestra aleatoria de cada estrato. Este método asegura que cada subgrupo esté representado en la muestra.
- **Muestreo por conglomerados**: la población se divide en grupos o conglomerados, y luego se seleccionan algunos conglomerados al azar. Se estudia a todos los miembros de los conglomerados seleccionados. Este método es útil cuando la población está dispersa geográficamente.

2.4. Relación entre tamaño de la muestra y precisión de los resultados

El tamaño de la muestra tiene un impacto directo en la precisión de los resultados obtenidos. A mayor tamaño de la muestra, menor es el error estándar y mayor es la probabilidad de que la muestra refleje con precisión las características de la población. Sin embargo, hay un punto de rendimientos decrecientes: aumentar el tamaño de la muestra más allá de cierto umbral no necesariamente mejora de manera significativa la precisión, pero sí incrementa los costos y la complejidad del análisis.

- **Tamaño mínimo**: hay fórmulas estadísticas que permiten calcular el tamaño mínimo de muestra necesario para un estudio, dependiendo del nivel de confianza deseado y el margen de error permitido.

- **Error estándar**: es una medida de la variabilidad en las estimaciones obtenidas a partir de una muestra. Un tamaño de muestra mayor reduce el error estándar, lo que implica mayor precisión en las estimaciones.
- **Intervalos de confianza**: un intervalo de confianza es un rango dentro del cual es probable que se encuentre un parámetro poblacional, basado en una muestra. El tamaño de la muestra afecta el ancho del intervalo de confianza: muestras más grandes generan intervalos más estrechos, lo que indica mayor precisión.

Ejemplo: selección de una muestra representativa

Supongamos que tenemos una población de 500 estudiantes en una universidad y queremos estudiar su rendimiento académico. Seleccionamos una muestra de 50 estudiantes utilizando un muestreo estratificado por nivel educativo (pregado y posgrado)

- La población está compuesta por 400 estudiantes de pregado y 100 estudiantes de posgrado.
- Queremos que la muestra sea proporcional por lo que seleccionamos:

$$\frac{400}{500} \times 50 = 40 \text{ estudiantes de pregrado.}$$

$$\frac{400}{500} \times 50 = 10 \text{ estudiantes de posgrado.}$$

De esta manera, la muestra de 50 estudiantes mantiene las proporciones de la población en términos de niveles educativos.

3. Procesos estadísticos

El proceso estadístico es un conjunto sistemático de pasos que se siguen para recolectar, analizar y presentar datos con el fin de obtener información útil para la toma de decisiones. Este proceso involucra varias fases que garantizan la validez y precisión de los resultados obtenidos.

3.1. Fases del proceso estadístico: recolección, análisis e interpretación

El proceso estadístico generalmente se divide en tres grandes fases:

a) **Recolección de datos**: es el primer paso en cualquier proceso estadístico y consiste en obtener la información necesaria para el análisis. Los datos pueden recolectarse de diversas fuentes, como encuestas, experimentos o registros históricos. Esta fase es crítica, ya que la calidad y precisión de los datos recolectados impactan directamente en los resultados finales. Durante esta fase, es importante definir las variables que se van a estudiar, el tipo de datos que se necesitan (cuantitativos o cualitativos), y el método de recolección (observación, encuestas, entrevistas, etc.). La recolección puede realizarse mediante:

- **Muestreo**: selección de una muestra representativa de la población.
- **Censos**: recolección de datos de toda la población.
- **Registros administrativos**: uso de datos ya existentes como los registros médicos, registros financieros, etc.

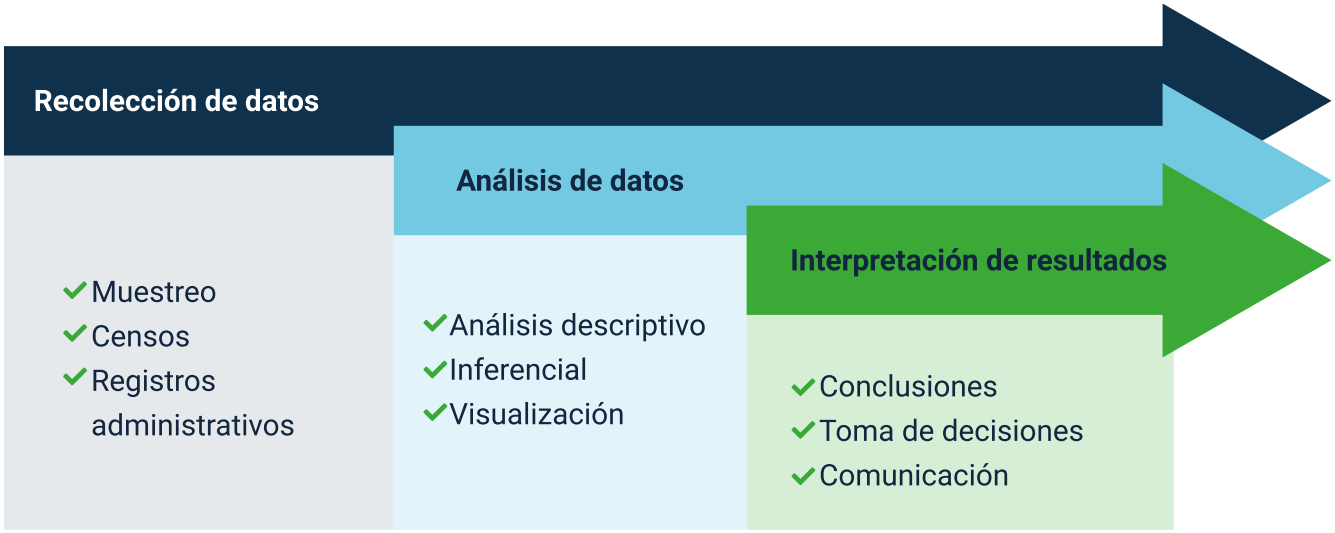
b) **Análisis de datos**: una vez que se han recolectado los datos, el siguiente paso es organizarlos y analizarlos. Esto implica ordenar los datos, clasificarlos y aplicar técnicas estadísticas adecuadas. Se pueden utilizar métodos de análisis descriptivo, como el cálculo de medidas de tendencia

central (media, mediana, moda) y de dispersión (varianza, desviación estándar), o métodos inferenciales que permiten hacer generalizaciones sobre la población.

En esta fase, se busca identificar patrones, tendencias o relaciones entre las variables. Las técnicas estadísticas empleadas dependen del tipo de datos y de los objetivos del estudio. Por ejemplo, los gráficos como los histogramas o los diagramas de dispersión ayudan a visualizar los datos, mientras que los modelos estadísticos permiten hacer predicciones.

c) **Interpretación de resultados**: finalmente, una vez que se han analizado los datos, es necesario interpretar los resultados obtenidos. Esta fase consiste en darle sentido a los resultados y responder las preguntas de investigación o los objetivos planteados al inicio del estudio. Se utilizan conclusiones basadas en los resultados estadísticos para tomar decisiones informadas o formular nuevas hipótesis. La interpretación debe realizarse con cuidado, teniendo en cuenta posibles sesgos en los datos y limitaciones del estudio. Además, es importante que los resultados se comuniquen de manera clara, generalmente a través de informes o presentaciones que incluyan gráficos y tablas para facilitar la comprensión.

Figura 1. Fases del proceso estadístico



Fuente. OIT, 2024.

La recolección de datos es esencial en el proceso estadístico, ya que los datos constituyen la base de cualquier análisis. Existen varios métodos para recolectar datos, y la elección de un método dependerá del tipo de información que se necesita y de los recursos disponibles.

a) **Encuestas**: las encuestas son uno de los métodos más comunes para recolectar datos. Consisten en formular una serie de preguntas estandarizadas a un grupo de personas (la muestra) con el fin de obtener información cuantitativa o cualitativa. Las encuestas pueden ser presenciales, telefónicas o en línea. Una de sus ventajas es que permiten obtener datos de una gran cantidad de personas en un periodo corto de tiempo.

Tipos de encuestas:

- **Encuestas estructuradas**: tienen preguntas cerradas con respuestas predefinidas.
- **Encuestas semiestructuradas**: combinan preguntas cerradas y abiertas, ofreciendo mayor flexibilidad.

b) **Observación**: este método implica la recolección de datos a través de la observación directa de comportamientos, eventos o fenómenos. La observación puede ser participante (el investigador forma parte del entorno que está observando) o no participante (el investigador observa desde fuera sin interactuar con el entorno). Es un método útil cuando se necesita estudiar fenómenos en su contexto natural.

Ejemplo: en estudios de comportamiento animal, los investigadores pueden observar el comportamiento de una población de animales en su hábitat natural.

c) **Experimentación**: en un experimento, el investigador manipula una o más variables independientes para observar el efecto sobre una variable dependiente. Este método permite establecer relaciones de causa y efecto. Los experimentos suelen realizarse en entornos controlados, como laboratorios, para minimizar la influencia de variables externas.

Ejemplo: en un estudio de un nuevo medicamento, los investigadores podrían dar el medicamento a un grupo de pacientes y un placebo a otro grupo, observando las diferencias en los resultados de salud entre ambos grupos.

3.2. Definición y objetivos de cada fase del proceso estadístico

Cada fase del proceso estadístico tiene objetivos claros que contribuyen a la integridad del análisis y a la validez de los resultados:

- **Recolección de datos**: el objetivo de esta fase es obtener datos precisos y confiables que respondan a las preguntas de investigación. La recolección adecuada de datos minimiza errores y sesgos, lo que asegura la calidad del análisis posterior.
- **Análisis de datos**: el objetivo de esta fase es identificar patrones, tendencias y relaciones entre las variables. A través del análisis estadístico, se extrae información significativa que ayudará a interpretar el fenómeno en estudio. Además, el análisis permite realizar comparaciones y predicciones.
- **Interpretación de resultados**: el objetivo aquí es proporcionar una explicación clara y precisa de los hallazgos del análisis de datos. La interpretación correcta de los resultados es fundamental para que las conclusiones del estudio sean útiles y relevantes para la toma de decisiones.

3.3. Importancia de la correcta recolección de datos para evitar sesgos

La correcta recolección de datos es esencial para evitar errores y sesgos que podrían comprometer la validez de los resultados. Algunos tipos de sesgos que pueden ocurrir durante la recolección de datos son:

- **Sesgo de selección**: ocurre cuando la muestra no es representativa de la población. Por ejemplo, si se estudia una muestra de estudiantes

universitarios para hacer inferencias sobre toda la población adulta, se incurriría en un sesgo de selección.

- **Sesgo de respuesta**: sucede cuando las respuestas de los participantes son influenciadas por la forma en que se formulan las preguntas o por el entorno en el que se realiza la encuesta. Esto puede ocurrir, por ejemplo, si las preguntas están redactadas de manera que sugieren una respuesta particular.
- **Sesgo de no respuesta**: ocurre cuando una parte de la muestra seleccionada no participa en el estudio, lo que puede afectar la representatividad de los datos recolectados.

3.4. Control de calidad en la recolección de datos estadísticos

El control de calidad en la recolección de datos es una etapa crítica para garantizar que los datos obtenidos sean precisos, completos y confiables. Algunas prácticas comunes para asegurar la calidad de los datos incluyen:

- **Capacitación del personal**: es importante que las personas encargadas de recolectar los datos estén capacitadas en los métodos y herramientas que van a utilizar. Esto asegura que la recolección sea consistente y libre de errores.
- **Supervisión y monitoreo**: durante la recolección de datos, es esencial supervisar el proceso para identificar y corregir errores en tiempo real. El monitoreo constante garantiza que se sigan los procedimientos establecidos.

- **Pruebas piloto**: antes de realizar un estudio completo, se puede hacer una prueba piloto con un pequeño grupo para detectar posibles problemas en los cuestionarios, herramientas o procedimientos.
- **Revisión de datos**: después de la recolección, es importante revisar los datos para identificar valores atípicos o inconsistencias que podrían indicar errores en la recolección.

Ejemplo: **proceso estadístico para una encuesta de satisfacción**

Una empresa realiza una encuesta de satisfacción a sus clientes y obtiene los siguientes resultados sobre una escala de 1 a 10 (donde 1 es muy insatisfecho y 10 es muy satisfecho):

Respuestas = {7, 8, 6, 9, 7, 8, 5, 7, 6, 9}

Fase de recolección de datos: Se recolectan las respuestas de los clientes mediante un formulario en línea.

Fase de análisis:

- **Media**: $\bar{x} = (7 + 8 + 6 + 9 + 7 + 8 + 5 + 7 + 6 + 9) / 10 = 7.2$
- **Desviación estándar**: $\sigma \approx 1.35$

Fase de interpretación: La media de satisfacción es de 7.2, lo que indica que, en general, los clientes están satisfechos. Sin embargo, la desviación estándar muestra cierta variabilidad en las respuestas, por lo que algunos clientes están menos satisfechos.

4. Técnicas de muestreo

El muestreo es una técnica utilizada en estadística para seleccionar una parte representativa de una población con el fin de realizar un estudio. Dado que en muchas situaciones es impracticable estudiar a toda la población, se selecciona una muestra que, si se elige adecuadamente, reflejará las características de la población. Las técnicas de muestreo garantizan que esta selección sea lo más representativa posible, minimizando sesgos y errores.

4.1. Muestreo aleatorio simple: definición y aplicación

El muestreo aleatorio simple es uno de los métodos de selección más básicos y efectivos. En este tipo de muestreo, cada individuo u objeto de la población tiene la misma probabilidad de ser seleccionado, lo que garantiza que no haya preferencias o sesgos en la elección de la muestra.

- **Definición**: el muestreo aleatorio simple se basa en la premisa de que todos los elementos de la población son equiprobables para ser seleccionados. Esto puede lograrse utilizando una lista numerada de la población y seleccionando los elementos mediante métodos aleatorios, como un sorteo o el uso de generadores de números aleatorios.
- **Aplicación**: este método es ampliamente utilizado cuando se tiene acceso a una lista completa de la población y se desea obtener una muestra representativa. Sin embargo, es menos práctico en poblaciones grandes o dispersas, ya que puede ser difícil asegurar la equiprobabilidad para cada miembro de la población.

Ejemplo: si una universidad quiere estudiar el rendimiento académico de sus estudiantes, puede utilizar un muestreo aleatorio simple seleccionando al azar un conjunto de estudiantes de la lista completa de matriculados.

4.2. Muestreo estratificado: ventajas y procedimientos

El muestreo estratificado es una técnica en la que la población se divide en subgrupos homogéneos llamados estratos, y luego se selecciona una muestra aleatoria de cada estrato. Este método garantiza que todos los subgrupos de la población estén representados en la muestra, lo cual es particularmente útil cuando la población es heterogénea.

a) **Definición**: en el muestreo estratificado, los estratos se forman en función de alguna característica relevante (por ejemplo, edad, género, nivel educativo). Posteriormente, se selecciona una muestra aleatoria de cada estrato, asegurando que la representación de cada subgrupo en la muestra sea proporcional al tamaño del estrato en la población.

   Ventajas:

   - Aumenta la precisión de las estimaciones, ya que cada estrato está representado adecuadamente.
   - Reduce la variabilidad dentro de los estratos, ya que los miembros de un estrato comparten características comunes.
   - Mejora la representación de subgrupos pequeños en la población, que de otro modo podrían estar subrepresentados en un muestreo aleatorio simple.

b) **Procedimientos**:

   - Dividir la población en estratos según una o varias características clave.

- Calcular el tamaño de la muestra que se tomará de cada estrato (puede ser proporcional o igual para todos los estratos).
- Seleccionar una muestra aleatoria simple dentro de cada estrato.

Ejemplo: si un investigador desea estudiar los hábitos de lectura de estudiantes de diferentes niveles de educación (primaria, secundaria, y universitaria), podría estratificar a los estudiantes por nivel educativo y seleccionar una muestra aleatoria de cada nivel.

4.3. Muestreo por conglomerados: características y ejemplos de uso

El **muestreo por conglomerados** es una técnica en la que la población se divide en grupos heterogéneos, o conglomerados, y luego se selecciona al azar uno o varios de estos conglomerados para ser estudiados en su totalidad. A diferencia del muestreo estratificado, los conglomerados no son homogéneos, sino que cada uno representa una "miniatura" de la población completa.

a) **Definición**: los conglomerados son subgrupos naturales de la población que pueden representar una mezcla diversa de características. En lugar de seleccionar individuos directamente, se seleccionan conglomerados completos, y todos los miembros de los conglomerados seleccionados son incluidos en la muestra.

b) **Características**:

- Es útil cuando es difícil acceder a toda la población, pero los conglomerados son más accesibles.
- Se utiliza cuando los conglomerados representan adecuadamente a la población en su conjunto.

- Puede ser más barato y eficiente que otras técnicas de muestreo, ya que reduce costos de desplazamiento o logística.

c) **Ejemplos de uso**:

- Estudio sobre el rendimiento académico de estudiantes en una región. En lugar de seleccionar estudiantes individualmente de toda la región, se pueden seleccionar al azar varias escuelas (conglomerados), y luego estudiar a todos los estudiantes de esas escuelas.
- Investigación sobre la satisfacción de empleados en una empresa con múltiples sucursales. Se puede seleccionar un conjunto de sucursales (conglomerados) y encuestar a todos los empleados de esas sucursales.

4.4. Comparación entre diferentes técnicas de muestreo

Las diferentes técnicas de muestreo tienen ventajas y desventajas, y la elección del método depende del contexto del estudio y las características de la población.

a) **Muestreo aleatorio simple**:

- **Ventajas**: fácil de aplicar y entender, garantiza que todos los elementos de la población tienen la misma probabilidad de ser seleccionados.
- **Desventajas**: no siempre es representativo si la población es muy heterogénea o si es difícil acceder a una lista completa de la población.

b) **Muestreo estratificado**:

- **Ventajas**: mejora la precisión del estudio y garantiza la representación adecuada de todos los subgrupos.
- **Desventajas**: requiere conocimientos previos sobre la estructura de la población para formar los estratos, lo cual puede ser costoso o difícil de implementar.

c) **Muestreo por conglomerados**:

- **Ventajas**: menor costo y mayor eficiencia en términos de tiempo y recursos cuando los conglomerados son fáciles de identificar y acceder.
- **Desventajas**: puede generar estimaciones menos precisas si los conglomerados no son representativos de la población.

4.5. Importancia del tamaño de la muestra en cada técnica de muestreo

El tamaño de la muestra es un factor que afecta la precisión de los resultados obtenidos mediante el muestreo. Independientemente de la técnica de muestreo elegida, el tamaño de la muestra debe ser suficiente para minimizar el error estándar y garantizar la representatividad de la población.

- **Muestreo aleatorio simple**: un tamaño de muestra mayor reduce la variabilidad y el error de muestreo, mejorando la precisión de las estimaciones.
- **Muestreo estratificado**: el tamaño de la muestra en cada estrato debe ser proporcional al tamaño del estrato en la población, garantizando que todos los subgrupos estén representados adecuadamente.
- **Muestreo por conglomerados**: aunque esta técnica puede reducir costos, a menudo requiere un mayor tamaño de muestra o más conglomerados para compensar la menor precisión que puede derivarse de la heterogeneidad dentro de los conglomerados.

Ejemplo: muestreo aleatorio simple y sistemático

- **Muestreo aleatorio simple**: seleccionamos una muestra de 5 empleados de una población de 20 empleados utilizando un generador de números aleatorios. Supongamos que los números seleccionados son: 3, 7, 9, 14, 18.
- **Muestreo sistemático**: de la misma población de 20 empleados, seleccionamos un empleado cada 4 (comenzando con el empleado 3). Los empleados seleccionados serán: 3, 7, 11, 15, 19.

Esto asegura una distribución uniforme en la selección de empleados.

5. Inferencia estadística

La **inferencia estadística** es una rama de la estadística que permite sacar conclusiones sobre una población a partir de una muestra de datos. Mediante el uso de probabilidades y modelos estadísticos, la inferencia nos permite estimar características de una población, probar hipótesis y realizar predicciones basadas en los datos observados. Es fundamental para la toma de decisiones en contextos donde no es posible o práctico estudiar a toda la población.

5.1. Concepto de inferencia estadística y su relevancia

La inferencia estadística se refiere a los métodos utilizados para extraer conclusiones sobre una población basándose en una muestra de esa población. La inferencia es esencial en situaciones donde sería demasiado costoso o imposible obtener datos de todos los elementos de una población (por ejemplo, una encuesta nacional o un estudio médico con miles de pacientes).

La relevancia está dada por:

- **Ahorro de recursos**: en lugar de estudiar a toda la población, la inferencia permite hacer generalizaciones basadas en una muestra, lo que ahorra tiempo, dinero y esfuerzo.
- **Generalización**: al usar técnicas inferenciales, es posible hacer predicciones y conclusiones más amplias sobre la población con un margen de error controlado.
- **Toma de decisiones**: la inferencia es clave para tomar decisiones basadas en datos, como en investigaciones científicas, estudios de mercado, o procesos industriales de control de calidad.

Ejemplo: en una empresa que quiere conocer la satisfacción de los clientes, la inferencia estadística permite estimar la satisfacción global basándose en una encuesta a una muestra representativa de clientes.

5.2. Diferencia entre parámetros y estadísticos

Para comprender la inferencia estadística, es importante distinguir entre **parámetros** y **estadísticos**:

- **Parámetro**: un parámetro es un valor numérico que describe una característica de toda la población. Como es difícil o imposible conocer el valor exacto del parámetro (por ejemplo, la media de ingresos de todos los habitantes de un país), se utilizan muestras para estimarlo.
- **Estadístico**: un estadístico es un valor numérico que describe una característica de una muestra. Los estadísticos se calculan a partir de los datos recolectados y se utilizan para hacer inferencias sobre los parámetros de la población.

Ejemplo: si tomamos una muestra de 100 estudiantes de una universidad y calculamos su promedio de calificaciones, ese promedio es un estadístico. Si intentamos hacer una inferencia sobre el promedio de calificaciones de todos los estudiantes de la universidad, estamos haciendo una estimación del parámetro poblacional.

5.3. Inferencia en la toma de decisiones basada en datos

La inferencia estadística es fundamental para la toma de decisiones en situaciones de incertidumbre. Al aplicar técnicas inferenciales, podemos hacer predicciones y estimaciones sobre una población o evento futuro con un margen de

error cuantificado. Las decisiones basadas en datos tienen una mayor probabilidad de éxito porque están respaldadas por evidencia numérica y análisis objetivo.

- **Toma de decisiones en investigación médica**: los ensayos clínicos que prueban la efectividad de un nuevo medicamento utilizan muestras de pacientes para hacer inferencias sobre cómo ese medicamento funcionará en la población general. Las decisiones sobre la aprobación o no de un fármaco se basan en la inferencia estadística obtenida de la muestra.
- **Toma de decisiones empresariales**: las empresas utilizan la inferencia para hacer predicciones sobre ventas futuras, satisfacción del cliente o efectividad de campañas publicitarias, lo que les permite tomar decisiones estratégicas informadas.
- **Predicción de tendencias**: la inferencia estadística se usa en estudios de mercado para predecir las tendencias de consumo, lo que ayuda a las empresas a ajustar sus estrategias de producción y marketing.

5.4. Tipos de estimación en inferencia: puntual y por intervalos

Existen dos tipos principales de estimaciones que se utilizan en la inferencia estadística: **estimaciones puntuales** y **estimaciones por intervalos**.

a) **Estimación puntual**: es un solo valor numérico calculado a partir de una muestra que se utiliza para estimar un parámetro poblacional. Por ejemplo, si se toma una muestra aleatoria de salarios de empleados y se calcula la media, ese valor de la media es una estimación puntual del salario promedio en la población.

- **Ventajas**: es fácil de calcular y proporcionar.

- **Desventajas**: no proporciona información sobre la variabilidad o incertidumbre de la estimación.

b) **Estimación por intervalos**: en lugar de un solo valor, se proporciona un rango (intervalo) dentro del cual es probable que se encuentre el parámetro poblacional con un cierto nivel de confianza. Un intervalo de confianza es un ejemplo de estimación por intervalos.

- **Ventajas**: proporciona una medida de la precisión de la estimación y refleja la incertidumbre inherente en el muestreo.
- **Desventajas**: más complejo de calcular y comunicar que la estimación puntual.

Ejemplo: un intervalo de confianza del 95% para el salario promedio puede ser $30,000 a $35,000, lo que indica que estamos un 95% seguros de que el salario promedio de la población está dentro de ese rango.

5.5. Aplicación de pruebas de hipótesis en la inferencia estadística

Una de las herramientas más poderosas de la inferencia estadística es la prueba de hipótesis. Este proceso permite evaluar la validez de una afirmación sobre un parámetro poblacional, basándose en los datos de una muestra.

a) **Hipótesis nula (H_o)**: es la afirmación que se asume verdadera al comienzo de la prueba. Normalmente, la hipótesis nula es una afirmación de "no efecto" o "no diferencia". Por ejemplo, en un estudio médico, la hipótesis nula podría ser que un nuevo tratamiento no es más efectivo que el tratamiento actual.

b) **Hipótesis alternativa (H_1)**: es la afirmación opuesta a la hipótesis nula. Es lo que el investigador espera demostrar. En el caso del tratamiento médico, la hipótesis alternativa sería que el nuevo tratamiento sí es más efectivo que el tratamiento actual.

c) **Proceso de prueba de hipótesis**:

- **Formular la hipótesis nula (H_o) y la hipótesis alternativa (H_1)**.
- **Elegir un nivel de significancia (α)**: este valor (generalmente 0.05) es el umbral que se utiliza para decidir si los resultados observados son lo suficientemente inusuales como para rechazar la hipótesis nula.
- **Calcular una estadística de prueba**: esta estadística se utiliza para comparar los resultados observados con lo que se esperaría si la hipótesis nula fuera cierta.
- **Decisión**: si la estadística de prueba indica que los resultados son improbables bajo la hipótesis nula (es decir, el valor p es menor que α), se rechaza H_o en favor de H_1.

d) **Valor p**: El valor p es la probabilidad de obtener un resultado igual o más extremo que el observado, suponiendo que la hipótesis nula es cierta. Si el valor p es menor que el nivel de significancia se concluye que hay suficiente evidencia para rechazar la hipótesis nula.

Ejemplo: si una empresa está probando si una nueva campaña publicitaria mejora las ventas en comparación con la campaña anterior, la hipótesis nula podría ser que no hay diferencia en las ventas. Al realizar la prueba de hipótesis, si el valor p es menor que 0.05, la empresa puede concluir que la nueva campaña es significativamente más efectiva.

Ejemplo: prueba de hipótesis

Supongamos que queremos probar si el salario promedio de los empleados en una empresa es mayor a $50,000. Tenemos una muestra de 30 empleados con una media de $\bar{x} = 52{,}000$ y una desviación estándar de $\sigma = 8{,}000$.

- **Hipótesis nula (H_o)**: El salario promedio es $\mu = 50{,}000$.
- **Hipótesis alternativa (H_1)**: El salario promedio es mayor a $\mu = 50{,}000$.

Utilizamos una **prueba t** para una media muestral. El estadístico de prueba t es:

$t = (\bar{x} - \mu) / (\sigma/\sqrt{n}) = (52{,}000 - 50{,}000) / (8{,}000/\sqrt{30}) \approx 1.37$

Con un nivel de significancia $\alpha = 0.05$, el valor crítico para t con 29 grados de libertad es 1.699.

Como $t = 1.37 < 1.699$, no podemos rechazar H_o, lo que indica que no hay suficiente evidencia para afirmar que el salario promedio es mayor a $50,000.

6. Definición de requerimientos para la recolección de datos

La definición de requerimientos para la recolección de datos es una fase crítica en cualquier estudio estadístico, ya que determina el tipo de información que se necesita, la forma en que se obtendrá, y las características que deben cumplir los datos para asegurar su utilidad en el análisis posterior. Los requerimientos bien definidos aseguran que los datos recolectados sean adecuados para cumplir con los objetivos del estudio y que la recolección de datos se realice de manera eficiente y efectiva.

6.1. Tipos de requerimientos en proyectos estadísticos

Antes de recolectar datos, es necesario definir qué tipo de datos se necesitan para cumplir con los objetivos del estudio. Los requerimientos pueden variar dependiendo del contexto y el tipo de análisis que se va a realizar. Los requerimientos pueden dividirse en las siguientes categorías:

- **Requerimientos cuantitativos**: se refieren a datos numéricos que pueden ser medidos y expresados en números. Estos incluyen, por ejemplo, medidas de ingresos, peso, altura, cantidades vendidas, etc. Los datos cuantitativos son esenciales para realizar análisis numéricos y estadísticos que permitan obtener resultados precisos.

  Ejemplo: en un estudio de ventas, los datos requeridos pueden incluir la cantidad de unidades vendidas, los ingresos generados, o el precio promedio por unidad.

- **Requerimientos cualitativos**: se refieren a datos no numéricos que describen características o atributos. Estos datos suelen ser categóricos y se utilizan para clasificar o agrupar elementos. Incluyen variables como género, opiniones, preferencias, tipos de productos, entre otros.

Ejemplo: en una encuesta de satisfacción del cliente, los datos cualitativos pueden incluir comentarios sobre la experiencia de compra o la clasificación de los productos en categorías de satisfacción (muy satisfecho, satisfecho, insatisfecho).

- **Requerimientos temporales**: establecen la necesidad de recolectar datos en intervalos de tiempo específicos o durante un periodo particular. Los datos temporales son importantes cuando se quiere analizar cambios o tendencias a lo largo del tiempo.

  Ejemplo: un estudio sobre el comportamiento de los consumidores puede requerir datos de ventas mensuales durante los últimos cinco años para identificar patrones estacionales.

- **Requerimientos geográficos**: se refieren a la necesidad de recolectar datos en una ubicación o conjunto de ubicaciones específicas. Los estudios geográficos o regionales necesitan definir claramente las áreas de donde se tomarán los datos.

  Ejemplo: un análisis de mercado puede requerir datos específicos de ventas de diferentes ciudades o regiones para comparar los comportamientos de compra en diversas zonas geográficas.

6.2. Requerimientos cuantitativos y cualitativos en la estadística

Los requerimientos cuantitativos y cualitativos son fundamentales para determinar el enfoque y los métodos que se utilizarán en el estudio estadístico. Ambos tipos de datos son igualmente importantes, pero tienen diferentes aplicaciones y usos dentro del análisis.

a) **Requerimientos cuantitativos**:

- **Datos numéricos**: permiten realizar análisis estadísticos como la estimación de medias, la varianza, la correlación, entre otros.
- **Escalas de medición**: pueden clasificarse en datos de razón (con un valor cero absoluto, como ingresos) o datos de intervalo (sin un valor cero absoluto, como la temperatura en grados Celsius).
- **Ejemplos de análisis**: los datos cuantitativos se utilizan para calcular tendencias centrales (media, mediana, moda), medidas de dispersión (desviación estándar, varianza), y para realizar pruebas de hipótesis o análisis de regresión.

b) **Requerimientos cualitativos**:

- **Datos categóricos**: permiten clasificar y agrupar elementos en categorías, y se utilizan comúnmente en la estadística descriptiva para describir frecuencias o proporciones.
- **Escalas de medición**: pueden clasificarse en variables nominales (sin un orden intrínseco, como género) o variables ordinales (con un orden intrínseco, como niveles de satisfacción).
- **Ejemplos de análisis**: los datos cualitativos se utilizan para calcular frecuencias, porcentajes y realizar análisis de tablas cruzadas, análisis de chi-cuadrado, o análisis cualitativos más detallados.

6.3. Identificación de las variables clave a recolectar

Una vez que se han definido los requerimientos cuantitativos y cualitativos, es importante identificar cuáles serán las variables clave que se recolectarán durante el

estudio. Estas variables son los datos específicos que se medirán o recolectarán y que serán relevantes en el análisis posterior.

a) **Variables dependientes e independientes**:
   - La **variable dependiente** es la que se quiere estudiar o predecir (por ejemplo, los ingresos o el rendimiento académico).
   - Las **variables independientes** son aquellas que pueden influir o tener un efecto sobre la variable dependiente (por ejemplo, el nivel educativo o la cantidad de horas trabajadas).

b) **Criterios para identificar las variables clave**:
   - Relevancia para los objetivos del estudio.
   - Capacidad para ser medidas de manera confiable.
   - Capacidad para influir en las decisiones o conclusiones que se derivarán del estudio.

Ejemplo: en un estudio sobre la efectividad de un nuevo medicamento, la variable dependiente podría ser la mejora en los síntomas del paciente, mientras que las variables independientes podrían incluir la dosis del medicamento, la edad y el estado de salud inicial del paciente.

6.4. Proceso de validación y ajuste de los requerimientos iniciales

Después de identificar las variables clave y definir los requerimientos para la recolección de datos, es necesario validar y, si es necesario, ajustar los requerimientos antes de comenzar la recolección. Este proceso es esencial para asegurarse de que los datos recolectados cumplirán con los objetivos del estudio y serán útiles para el análisis.

a) **Validación de requerimientos**:

- Asegurarse de que los datos a recolectar sean suficientes para responder a las preguntas de investigación o hipótesis planteadas.
- Comprobar la disponibilidad de los datos. Por ejemplo, puede ser que algunos datos sean difíciles de recolectar o no existan en el formato necesario.
- Evaluar la viabilidad de la recolección de datos en términos de tiempo, recursos y acceso a las fuentes de datos.

b) **Ajuste de requerimientos**:

- Si se encuentran limitaciones o problemas en la recolección de ciertos datos, los requerimientos pueden ser ajustados. Esto puede implicar la simplificación del estudio o la redefinición de algunas variables.
- Los ajustes pueden incluir cambios en los métodos de recolección de datos (pasar de encuestas presenciales a encuestas en línea, por ejemplo) o la recolección de datos adicionales que no se consideraron inicialmente.

6.5. Impacto de los requerimientos en la precisión y relevancia de los datos

Los requerimientos establecidos para la recolección de datos tienen un impacto directo en la calidad, precisión y relevancia de los datos obtenidos. Si los requerimientos son demasiado amplios o ambiguos, es probable que se recojan datos innecesarios o que falte información crítica. Por el contrario, requerimientos muy estrictos pueden limitar la cantidad de datos disponibles y llevar a conclusiones insuficientes.

- **Precisión de los datos**: los requerimientos deben ser claros para asegurar que los datos recolectados sean lo suficientemente precisos como para permitir un análisis detallado y significativo. Los datos imprecisos o incompletos pueden llevar a resultados inexactos y análisis poco confiables.
- **Relevancia de los datos**: solo deben recolectarse los datos que son relevantes para el estudio. Los requerimientos deben estar alineados con los objetivos del análisis para evitar recolectar datos innecesarios o irrelevantes, lo cual solo incrementaría la complejidad del análisis sin agregar valor.
- **Coherencia y comparabilidad**: al definir los requerimientos, es importante asegurarse de que los datos sean coherentes entre las diferentes fuentes y comparables a lo largo del tiempo o entre diferentes subgrupos.

Ejemplo: recolección de datos cualitativos y cuantitativos

En un estudio sobre hábitos alimenticios, definimos los siguientes requerimientos:

a) **Datos cuantitativos:**
   - Ingreso mensual de los encuestados (en dólares).
   - Número de comidas al día.

b) **Datos cualitativos:**
   - Preferencias alimenticias (vegetariano, carnívoro, etc.).
   - Satisfacción con los hábitos alimenticios (alta, media, baja).

Al definir estos requerimientos, garantizamos que los datos recolectados sean relevantes y suficientes para analizar los hábitos alimenticios y su relación con los ingresos y la satisfacción.

7. Fuentes de datos

Las fuentes de datos son los orígenes desde los cuales se obtienen los datos necesarios para realizar análisis estadísticos. Se requiere identificar correctamente las fuentes de datos adecuadas, ya que la calidad, confiabilidad y relevancia de los datos dependen directamente de su procedencia. Existen diferentes tipos de fuentes de datos, y cada una tiene sus características particulares, ventajas y desventajas. Además, evaluar la confiabilidad de las fuentes es fundamental para garantizar la validez de los resultados obtenidos a partir de los datos.

7.1. Clasificación de fuentes de datos: primarias y secundarias

Las fuentes de datos se clasifican en dos grandes categorías: fuentes primarias y fuentes secundarias. Esta clasificación depende de si los datos fueron recolectados directamente por el investigador o si fueron obtenidos de fuentes ya existentes.

a) **Fuentes de datos primarias**:

- **Definición**: son datos recolectados de primera mano, es decir, directamente por el investigador o la entidad que lleva a cabo el estudio. Este tipo de fuente se utiliza cuando los datos necesarios no existen previamente o cuando es importante tener control total sobre el proceso de recolección de datos.
- **Ejemplos**: encuestas, entrevistas, experimentos, observación directa.
- **Ventajas**:
  - Control total sobre el proceso de recolección de datos.
  - Datos personalizados y específicos para los objetivos del estudio.
  - Mayor grado de precisión y relevancia.

- **Desventajas**:
    - El proceso de recolección puede ser costoso y llevar mucho tiempo.
    - Puede ser difícil acceder a ciertos tipos de datos de manera directa (por ejemplo, en estudios con grandes poblaciones).

b) **Fuentes de datos secundarias**:

- **Definición**: son datos que ya han sido recolectados por otra persona o entidad para otros fines, pero que se pueden reutilizar en el estudio actual. Los datos secundarios están disponibles en informes, bases de datos, libros, publicaciones académicas, entre otros.
- **Ejemplos**: informes gubernamentales, datos de censos, estadísticas de instituciones, investigaciones anteriores, bases de datos públicas.
- **Ventajas**:
    - Los datos ya están recolectados y disponibles, lo que ahorra tiempo y recursos.
    - Se pueden obtener datos históricos o de poblaciones grandes que serían difíciles de recolectar de manera primaria.
    - Es más accesible y económico.
- **Desventajas**:
    - Menor control sobre la calidad de los datos y el proceso de recolección.
    - Los datos pueden no ser completamente relevantes o específicos para el estudio actual.
    - Pueden no estar actualizados o ser obsoletos.

7.2. Métodos para evaluar la confiabilidad y validez de las fuentes

La confiabilidad y validez de las fuentes de datos son aspectos críticos que deben evaluarse antes de usar los datos en un análisis. Los datos confiables son aquellos que pueden replicarse y que producen resultados consistentes, mientras que los datos válidos son aquellos que realmente miden lo que se supone que deben medir.

a) **Criterios para evaluar la confiabilidad de una fuente de datos**:

- **Procedencia**: identificar el origen de los datos. Los datos provenientes de instituciones reconocidas, agencias gubernamentales o entidades de investigación respetadas suelen ser más confiables.
- **Método de recolección**: revisar cómo se obtuvieron los datos. Métodos rigurosos y estandarizados de recolección aumentan la confiabilidad.
- **Actualidad**: evaluar si los datos son recientes o si han quedado obsoletos. Datos más recientes son preferibles, especialmente en estudios que dependen de información actual.
- **Consistencia**: verificar si los datos son consistentes a lo largo del tiempo y entre diferentes fuentes. Inconsistencias pueden indicar problemas en la recolección o procesamiento de los datos.

b) **Criterios para evaluar la validez de una fuente de datos**:

- **Pertinencia**: los datos deben ser relevantes y directamente aplicables a los objetivos del estudio. Datos irrelevantes o fuera de contexto pueden sesgar los resultados.
- **Exactitud**: se refiere a qué tan bien los datos reflejan la realidad que se pretende medir. Las definiciones utilizadas en la recolección de datos deben ser claras y adecuadas.

- **Transparencia**: las fuentes confiables suelen ser transparentes sobre los métodos y procedimientos utilizados para recolectar los datos. La falta de información sobre el origen y la metodología de los datos puede ser una señal de baja calidad.
- **Imparcialidad**: es importante que la fuente no tenga conflictos de interés o sesgos que puedan haber influido en la recolección o presentación de los datos. Fuentes imparciales y objetivas proporcionan datos más confiables.

7.3. Uso de fuentes de datos primarias en encuestas y estudios

Las **fuentes de datos primarias** son especialmente útiles en estudios que requieren datos personalizados y específicos, como encuestas o investigaciones que buscan medir una característica particular de una población. Algunos aspectos importantes al usar fuentes primarias incluyen:

a) **Diseño de encuestas**: una encuesta bien diseñada es clave para recolectar datos primarios de alta calidad. Las preguntas deben ser claras, precisas y estar alineadas con los objetivos del estudio.
   - **Preguntas cerradas**: ofrecen opciones de respuesta predefinidas y son fáciles de analizar.
   - **Preguntas abiertas**: permiten respuestas más detalladas, aunque son más difíciles de procesar y analizar.

b) **Control de calidad**: durante la recolección de datos primarios, es fundamental implementar mecanismos para garantizar la calidad de los datos, como la capacitación de los encuestadores, la supervisión de las entrevistas y la realización de pruebas piloto.

c) **Ventajas del uso de fuentes primarias**:

- Proporcionan datos actualizados y específicos para el estudio.
- Mayor flexibilidad para ajustar el diseño de la recolección de datos según las necesidades del proyecto.

d) **Desafíos**:

- Requiere más tiempo y recursos que el uso de fuentes secundarias.
- La recolección de datos primarios puede estar sujeta a sesgos si no se diseña adecuadamente el proceso (por ejemplo, sesgo de selección o sesgo de respuesta).

7.4. Fuentes de datos secundarias: bases de datos públicas, informes y publicaciones

Las **fuentes de datos secundarias** incluyen una amplia gama de materiales que pueden ser reutilizados para investigaciones actuales. Algunos ejemplos de estas fuentes incluyen:

a) **Bases de datos públicas**: muchos gobiernos, organizaciones no gubernamentales (ONGs) y organismos internacionales proporcionan acceso gratuito a bases de datos que contienen información estadística y registros administrativos. Estas bases de datos pueden ser especialmente útiles para estudios a gran escala.

Ejemplos: datos de censos, bases de datos de salud, datos económicos, estadísticas educativas.

b) **Informes**: los informes de instituciones académicas, agencias gubernamentales, empresas privadas y organismos internacionales son

otra fuente importante de datos secundarios. Estos informes a menudo contienen análisis exhaustivos y datos ya procesados.

Ejemplos: informes de desarrollo humano, informes de salud pública, reportes de mercado.

c) **Publicaciones académicas**: los artículos de investigación publicados en revistas científicas a menudo presentan datos que se pueden reutilizar en estudios posteriores, especialmente en revisiones de literatura o estudios comparativos.

Ejemplos: estudios de caso, revisiones sistemáticas, análisis estadísticos previos.

d) **Ventajas del uso de fuentes secundarias**:

- Son de fácil acceso y generalmente menos costosas que la recolección de datos primarios.
- Proporcionan datos históricos que permiten realizar estudios longitudinales o analizar tendencias a lo largo del tiempo.

e) **Limitaciones**:

- La relevancia de los datos secundarios puede no coincidir completamente con los objetivos del estudio.
- Los datos secundarios pueden estar desactualizados o no haber sido recolectados con el rigor necesario.

7.5. Estrategias para combinar fuentes de datos múltiples en un análisis estadístico

En muchos estudios estadísticos, se pueden combinar fuentes de datos primarias y secundarias para obtener una visión más completa y enriquecida del fenómeno que se está analizando. Algunas estrategias para combinar estas fuentes incluyen:

- **Triangulación**: consiste en utilizar múltiples fuentes de datos para verificar y validar los resultados obtenidos. Por ejemplo, si un estudio de mercado utiliza encuestas primarias y datos secundarios de informes del sector, la triangulación puede ayudar a confirmar si las tendencias observadas son coherentes entre las diferentes fuentes.
- **Integración de datos**: en algunos estudios, es posible combinar bases de datos provenientes de distintas fuentes para realizar análisis más detallados. Esto es común en estudios de salud, donde se pueden combinar registros de pacientes con datos demográficos o socioeconómicos.
- **Uso secuencial de fuentes**: en algunos casos, los datos secundarios se utilizan como una primera aproximación para identificar tendencias o áreas de interés, y luego se recolectan datos primarios más específicos para profundizar en esas áreas.
- **Análisis comparativo**: comparar los datos obtenidos de distintas fuentes (primarias y secundarias) para analizar discrepancias y mejorar la comprensión del fenómeno. Las diferencias entre los resultados pueden proporcionar información valiosa sobre las limitaciones de las distintas fuentes de datos o sobre variaciones en el fenómeno estudiado.

Las fuentes de datos, tanto primarias como secundarias, son fundamentales para la realización de estudios estadísticos. Las fuentes primarias permiten recolectar datos específicos y actualizados, mientras que las fuentes secundarias proporcionan acceso a grandes volúmenes de datos ya disponibles. Evaluar la confiabilidad y validez de las fuentes de datos es necesario para asegurar la calidad del análisis y las conclusiones obtenidas. La combinación de diversas fuentes de datos puede enriquecer los estudios y permitir un análisis más completo y preciso del fenómeno en estudio.

Ejemplo: evaluación de fuentes de datos

Supongamos que estamos utilizando las siguientes fuentes de datos para un estudio:

- **Datos primarios**: encuestas aplicadas a una muestra de clientes.
- **Datos secundarios**: reportes de satisfacción del cliente proporcionados por la empresa de los últimos cinco años.

Evaluamos la confiabilidad de las fuentes:

a) **Los datos primarios son confiables** porque controlamos el proceso de recolección. Esto significa que nosotros mismos diseñamos las preguntas, seleccionamos a los participantes y recolectamos la información directamente. De esta manera, tenemos mayor control sobre la calidad y la relevancia de los datos.

b) **Los datos secundarios son útiles**, pero pueden no estar actualizados o tener sesgos relacionados con el método de recolección anterior.

Figura 2. Tipos de Fuentes de Datos

Según el origen	Según la naturaleza	Según el formato	Según la procedencia	Según la accesibilidad
Datos Primarios Datos Secundarios	Datos Cualitativos Datos Cuantitativos	Datos Estructurados Datos No Estructurados	Datos Internos Datos Externos	Datos Abiertos Datos Privados

Fuente. OIT, 2024.

8. Determinación de la muestra

La **determinación de la muestra** es un paso clave en cualquier estudio estadístico que implica hacer inferencias sobre una población. La muestra es el subconjunto de la población que se selecciona para ser estudiado, y su calidad y representatividad determinan en gran medida la validez de los resultados obtenidos. Definir el tamaño de la muestra y los métodos de selección adecuados es esencial para asegurar que las conclusiones se puedan generalizar correctamente a la población completa.

Métodos para seleccionar una muestra adecuada

Existen varios métodos para seleccionar una muestra adecuada, y la elección del método depende del tipo de estudio, los recursos disponibles y las características de la población. Estos métodos se dividen principalmente en dos categorías: **muestreo probabilístico** y **muestreo no probabilístico**.

Tabla 2. Métodos de muestreo

Tipo de Muestreo	Descripción	Métodos Principales
Muestreo probabilístico	Es un tipo de muestreo donde cada individuo de la población tiene una probabilidad conocida y no nula de ser seleccionado. Es el método preferido para asegurar que la muestra sea representativa de la población.	**Muestreo aleatorio simple**: cada elemento de la población tiene la misma probabilidad de ser seleccionado. Ideal cuando la población es homogénea. **Muestreo sistemático**: se selecciona un elemento inicial al azar y luego a intervalos regulares. **Muestreo estratificado**: la población se divide en subgrupos (estratos) y se selecciona una muestra aleatoria de cada uno.

Tipo de Muestreo	Descripción	Métodos Principales
		Muestreo por conglomerados: se divide en grupos heterogéneos, seleccionando uno o varios para estudio.
Muestreo no probabilístico	En este tipo de muestreo, no todos los elementos de la población tienen una probabilidad conocida de ser seleccionados. Puede ser menos representativo, pero es útil en estudios exploratorios o cuando no es posible aplicar un muestreo probabilístico.	**Muestreo por conveniencia**: se elige una muestra fácil de acceder, como los primeros individuos disponibles. **Muestreo por cuotas**: se seleccionan sujetos en proporciones específicas para representar ciertas características de la población, sin aleatoriedad en cada proporción. **Muestreo intencional o dirigido**: el investigador selecciona a los individuos que considera más adecuados para el estudio.

Fuente. OIT, 2024.

8.1. Criterios para seleccionar una muestra representativa

Una muestra representativa es aquella que refleja fielmente las características de la población completa. Para asegurar que una muestra sea representativa, es necesario seguir ciertos criterios durante su selección:

- **Diversidad de la población**: la muestra debe incluir a individuos con diferentes características, como edad, género, nivel educativo, ingresos, etc., en proporciones similares a las de la población. Esto asegura que todos los subgrupos de la población estén representados en la muestra.

- **Tamaño de la muestra**: un tamaño de muestra más grande tiende a ser más representativo, ya que permite capturar una mayor variabilidad dentro de la población. Sin embargo, el tamaño adecuado depende del tipo de análisis que se realizará y de los recursos disponibles.
- **Muestreo aleatorio**: siempre que sea posible, se deben usar métodos de muestreo probabilísticos para evitar sesgos en la selección de los individuos. Los métodos no probabilísticos solo deben usarse cuando no se dispone de otra opción.
- **Accesibilidad y viabilidad**: en algunos casos, ciertas partes de la población pueden ser difíciles de acceder o estar dispersas geográficamente. Es importante que la muestra sea lo suficientemente accesible para recolectar datos de manera eficiente sin comprometer la representatividad.

8.2. Consideraciones al seleccionar una muestra para minimizar el error

El **error muestral** es la diferencia entre el valor estimado a partir de la muestra y el valor real de la población. Para minimizar este error y mejorar la precisión de los resultados, es necesario tener en cuenta varias consideraciones al seleccionar una muestra:

- **Tamaño de la muestra**: el tamaño de la muestra es uno de los factores más importantes para reducir el error muestral. Un tamaño de muestra mayor generalmente disminuye el error muestral, pero hay un punto en el que los beneficios adicionales se reducen.
- **Aleatoriedad**: la selección aleatoria de los elementos de la muestra asegura que todos los individuos de la población tengan la misma

probabilidad de ser seleccionados, lo que reduce el sesgo y aumenta la precisión de la estimación.

- **Homogeneidad de la población**: si la población es muy heterogénea (con mucha variabilidad interna), se debe tener especial cuidado en la selección de la muestra para que incluya a todos los subgrupos relevantes. El uso de muestreo estratificado es particularmente útil en este caso.
- **Métodos de recolección de datos**: es importante que los métodos utilizados para recolectar datos de la muestra sean consistentes y que no introduzcan sesgos. La capacitación del personal encargado de la recolección de datos y la estandarización de los procedimientos son claves para minimizar el error.

8.3. Impacto del tamaño de la muestra en los resultados estadísticos

El tamaño de la muestra tiene un impacto directo en la precisión y validez de los resultados obtenidos en un análisis estadístico. Un tamaño de muestra adecuado permite obtener estimaciones confiables sobre la población, mientras que un tamaño de muestra demasiado pequeño puede producir resultados sesgados o imprecisos.

- **Error estándar**: el error estándar disminuye a medida que el tamaño de la muestra aumenta. Un error estándar menor implica una mayor precisión en las estimaciones realizadas sobre la población.

  Ejemplo: si se estima el ingreso promedio de una población con una muestra pequeña, el intervalo de confianza será más amplio y menos preciso. Con una muestra más grande, el intervalo se estrechará, lo que indicará una estimación más precisa del parámetro poblacional.

- **Intervalos de confianza**: los intervalos de confianza son más amplios con muestras pequeñas y más estrechos con muestras grandes. Un intervalo de confianza estrecho indica que hay mayor certeza de que el parámetro poblacional se encuentra dentro del rango estimado.
- **Poder estadístico**: el poder estadístico de una prueba es la capacidad de detectar un efecto o diferencia real en la población si es que existe. Un tamaño de muestra mayor aumenta el poder estadístico, lo que significa que es más probable que se detecten diferencias significativas cuando realmente existen.
- **Costo-beneficio**: aunque un tamaño de muestra más grande mejora la precisión, también aumenta los costos y el tiempo necesario para recolectar y procesar los datos. Por lo tanto, es importante encontrar un equilibrio entre el tamaño de la muestra y los recursos disponibles.

8.4. Técnicas para validar la representatividad de la muestra seleccionada

Una vez que se ha seleccionado una muestra, es fundamental validar su representatividad para garantizar que los resultados del estudio puedan generalizarse a la población completa. Existen varias técnicas para evaluar la representatividad de una muestra:

- **Comparación con la población**: comparar las características clave de la muestra (como edad, género, nivel educativo) con las de la población completa. Si las características de la muestra se alinean con las de la población, es más probable que sea representativa.

Ejemplo: en un estudio sobre hábitos de consumo en una ciudad, la muestra debe reflejar la distribución de edades, niveles de ingresos y otros factores demográficos de la población total de esa ciudad.

- **Pruebas de hipótesis**: las pruebas estadísticas pueden utilizarse para evaluar si las características de la muestra son significativamente diferentes de las características conocidas de la población. Si no se detectan diferencias significativas, la muestra puede considerarse representativa.
- **Ajustes ponderados**: si se detecta que ciertos grupos están subrepresentados o sobrerrepresentados en la muestra, se pueden aplicar ponderaciones a los datos para corregir estos desequilibrios y mejorar la representatividad.
- **Verificación mediante métodos alternativos**: comparar los resultados obtenidos a partir de la muestra con estudios o investigaciones anteriores que utilizaron diferentes métodos o datos secundarios. Si los resultados coinciden, es una indicación de que la muestra es representativa.

Ejemplo: cálculo del tamaño de la muestra

Queremos estimar el ingreso promedio de una ciudad con un margen de error de $500 y un nivel de confianza del 95%. Sabemos que la desviación estándar del ingreso en estudios anteriores es de $4,000.

El tamaño de la muestra se calcula usando la fórmula para el tamaño de muestra de una media:

$$n = (Z * \sigma / E)^2 = (1.96 * 4000 / 500)^2 = (7.840/500)^2 = 245.86$$

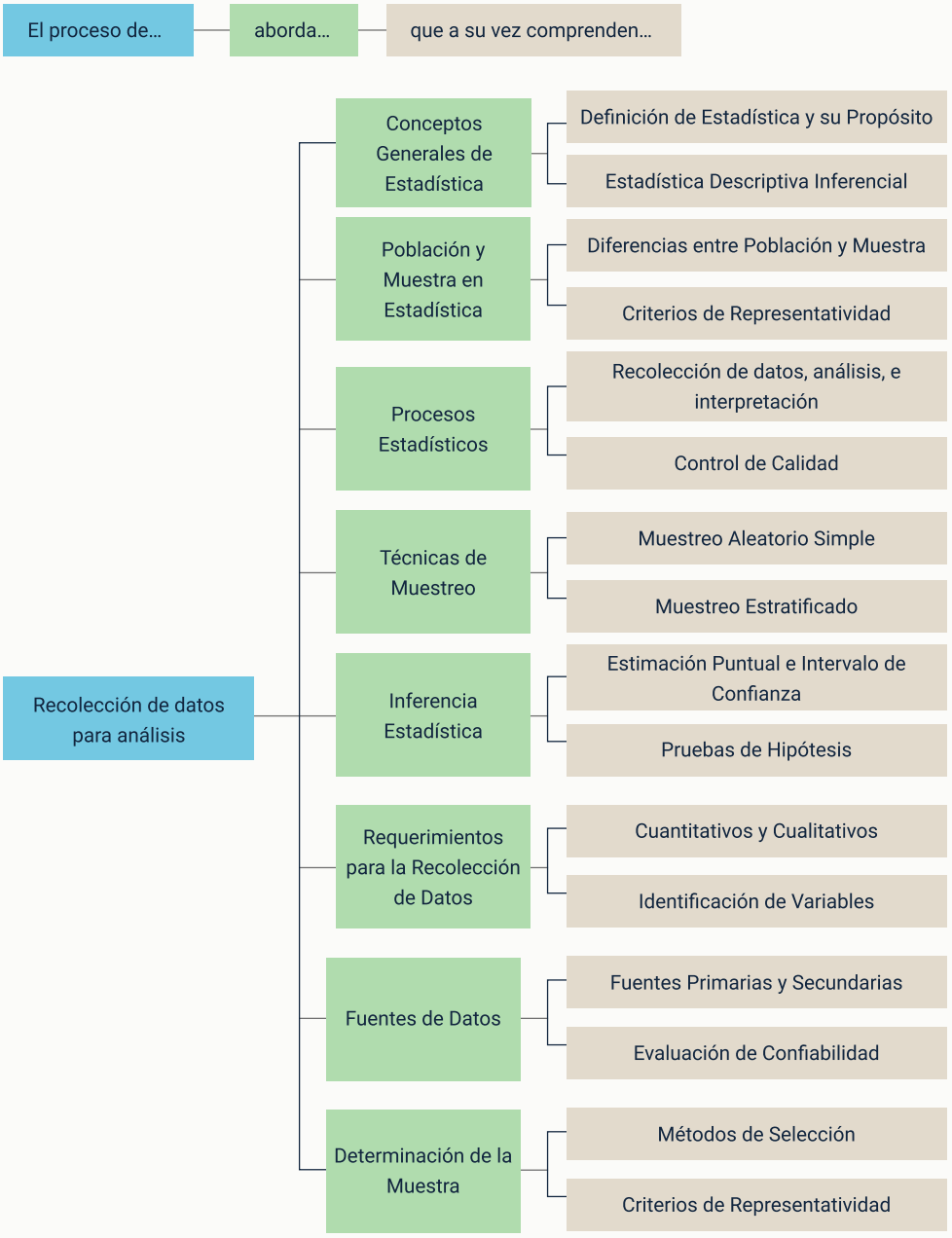
Redondeando, necesitamos una muestra de al menos 246 personas para estimar el ingreso promedio con el margen de error deseado.

Síntesis

El siguiente diagrama proporciona una visión general de los principales temas abordados en este componente, centrado en el uso de medidas estadísticas y probabilísticas para el análisis de datos. Este mapa conceptual está diseñado para facilitar al lector la visualización de las interrelaciones entre las diferentes áreas que componen el proceso de análisis estadístico, contribuyendo a una comprensión estructurada y práctica de los conceptos clave.

En el origen del diagrama se encuentra el concepto principal de "Recolección de datos para análisis", del cual se derivan temas fundamentales: conceptos generales de estadística, población y muestra en estadística, procesos estadísticos, técnicas de muestreo, inferencia estadística, requerimientos para la recolección de datos, fuentes de datos y determinación de la muestra. Cada una de estas áreas se subdivide 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 aprendiz podrá observar cómo los diferentes aspectos de la estadística y la recolección de datos se entrelazan para formar un proceso integral y sistemático. 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 recolección de datos para análisis estadístico.



Fuente. OIT, 2024.

Material complementario

Tema	Referencia	Tipo de material	Enlace del recurso
Conceptos generales de estadística	Ecosistema de Recursos Educativos Digitales SENA. (2023c, septiembre 20). Introducción a la estadística.	Video	https://www.youtube.com/watch?v=wMCDkknpUVw
Población y muestra en estadística.	Ecosistema de Recursos Educativos Digitales SENA. (2022b, octubre 26). Principales elementos de la estadística.	Video	https://www.youtube.com/watch?v=Ad5gxB9PhKQ
Procesos estadísticos.	Ecosistema de Recursos Educativos Digitales SENA. (2022a, julio 27). Herramientas de información estadística básica.	Video	https://www.youtube.com/watch?v=AW1LM-d0YWE
Procesos estadísticos.	Ecosistema de Recursos Educativos Digitales SENA. (2023a, marzo 24). Estadística descriptiva, gráficas e informes estadísticos.	Video	https://www.youtube.com/watch?v=v5UMlXHe2nM
Técnicas de muestreo.	Ecosistema de Recursos Educativos Digitales SENA. (2023b, marzo 26). Etapas del procesamiento de datos y métodos estadísticos - Introducción.	Video	https://www.youtube.com/watch?v=ndzj15PQEVw

Glosario

Análisis de datos: paso del proceso estadístico donde los datos recolectados se organizan y estudian para identificar patrones y relaciones.

Control de calidad: prácticas implementadas durante la recolección de datos para asegurar precisión y confiabilidad.

Dato: unidad básica de información sin procesar, obtenida a través de observaciones, encuestas u otras fuentes.

Estadística: ciencia que se dedica a la recolección, organización, análisis e interpretación de datos para la toma de decisiones.

Estadística descriptiva: rama de la estadística que se enfoca en resumir y describir las características principales de un conjunto de datos.

Estadística inferencial: rama de la estadística que permite hacer generalizaciones y predicciones sobre una población a partir de una muestra.

Fuente primaria: datos recolectados directamente por el investigador específicamente para el estudio en cuestión.

Fuente secundaria: datos previamente recopilados por otros y utilizados en el análisis actual.

Interpretación: fase en la que se analizan los resultados para sacar conclusiones y responder preguntas de investigación.

Intervalo de confianza: rango de valores dentro del cual se espera que se encuentre un parámetro poblacional con un nivel de confianza especificado.

Muestra: subconjunto representativo de la población, utilizado para hacer inferencias sobre el total.

Muestreo aleatorio simple: técnica de muestreo en la que todos los elementos de la población tienen la misma probabilidad de ser seleccionados.

Muestreo estratificado: método de muestreo en el que la población se divide en subgrupos homogéneos, seleccionando una muestra de cada uno.

Muestreo por conglomerados: técnica en la que la población se agrupa en conglomerados y se seleccionan algunos para ser estudiados en su totalidad.

Parámetro: valor numérico que representa una característica de la población, como la media o la desviación estándar.

Población: conjunto total de individuos, objetos o eventos de interés en un estudio estadístico.

Prueba de hipótesis: procedimiento estadístico utilizado para evaluar si una afirmación sobre un parámetro poblacional es consistente con los datos de la muestra.

Sesgo: error sistemático en la recolección de datos que distorsiona los resultados y afecta la validez de las conclusiones.

Variable: característica o atributo que puede ser medido en los individuos de un estudio; puede ser cualitativa o cuantitativa.

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