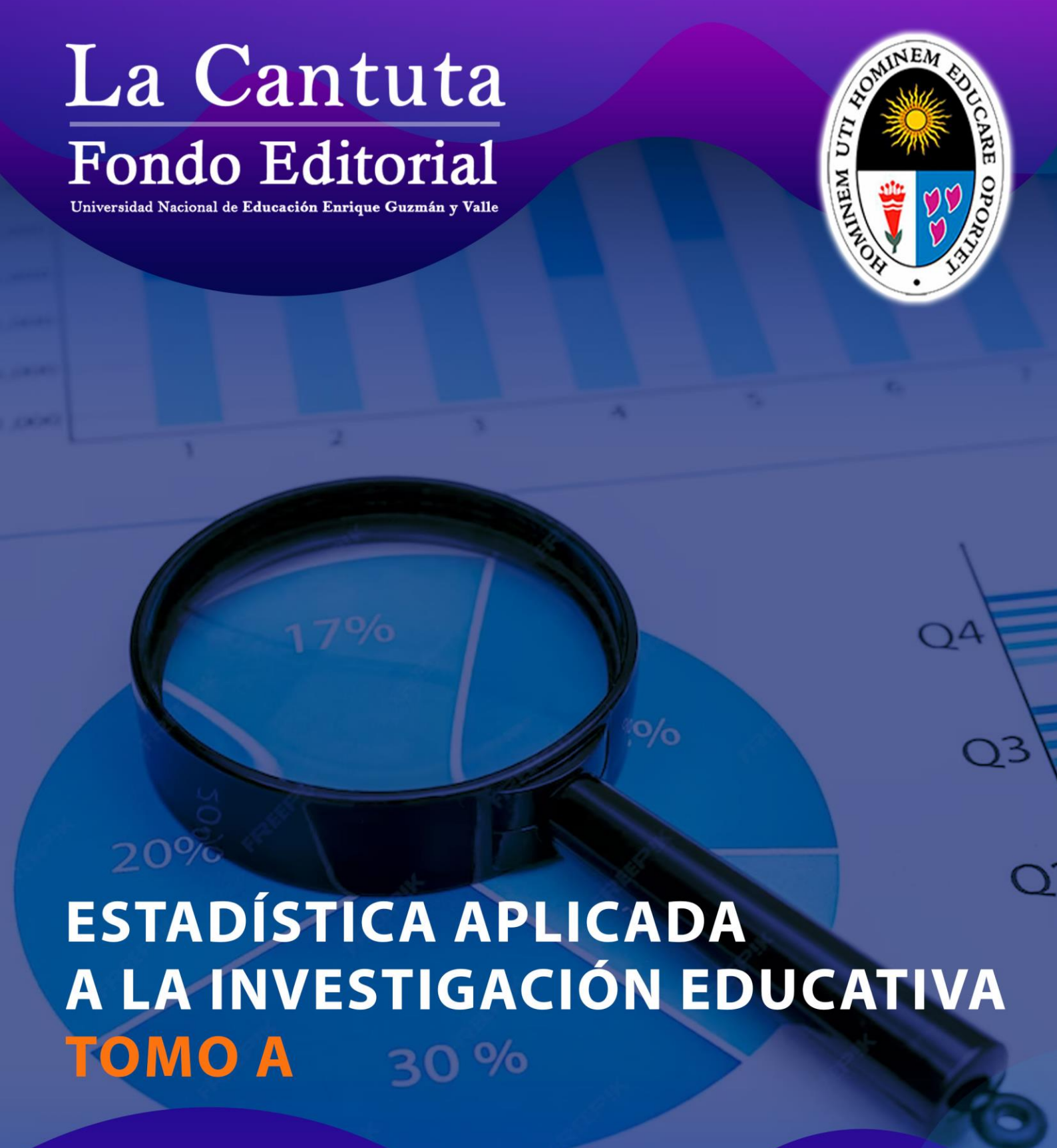
La Cantuta
Fondo Editorial
Universidad Nacional de Educación Enrique Guzmán y Valle



ESTADÍSTICA APLICADA A LA INVESTIGACIÓN EDUCATIVA

TOMO A

John Peter Castillo Mendoza
Ketty Magali Munguía Miguel
Narcio Felimon Vilcapoma Lara

fondoeditorial.une.edu.pe

Universidad Nacional de Educación Enrique Guzmán y Valle (UNE)

ESTADÍSTICA APLICADA A LA INVESTIGACIÓN EDUCATIVA

TOMO A

John Peter Castillo Mendoza
Ketty Magali Munguía Miguel
Narcio Felimon Vilcapoma Lara

Lima – Perú
2023

ESTADÍSTICA APLICADA A LA INVESTIGACIÓN EDUCATIVA

TOMO A

© **John Peter Castillo Mendoza**
jcastillo@une.edu.pe

Ketty Magali Munguía Miguel
kmunguia@une.edu.pe

Narcio Felimon Vilcapoma Lara
nvilcapoma@une.edu.pe

Editada por:

© Universidad Nacional de Educación Enrique Guzmán y Valle (UNE) - **Fondo Editorial "La Cantuta"**
Dirección: Enrique Guzman y Valle N° 951, Lurigancho-Chosica 15472, Perú
ISNI: 0000 0000 8534 4267
fondoeditorial@une.edu.pe
Teléf. móvil: +51 999 140 920
Portal Web: https://www.une.edu.pe/

Primera edición digital: Octubre 2023
Libro digital disponible en: https://fondoeditorial.une.edu.pe/

Hecho el Depósito Legal en la Biblioteca Nacional del Perú N° 2023-10230
ISBN: 978-612-4148-52-1
DOI: https://doi.org/10.54942/lacantuta.32

Corrección de estilo: Luis Pablo Diaz Tito
luisp.diaz@upsjb.edu.pe / Tel. de contacto: +51 955 129 801
Diseño y Diagramación: Gráfica "imagen"
Manuel Enrique Sampen Antonio
sampen25@gmail.com / Tel. de contacto: +51 990 064 589

Revisión por pares ciegos aprobado por el **Consejo Editorial del Fondo Editorial "La Cantuta".**

Libro resultado de Investigación y con revisión por pares doble ciego.
Sello editorial: Fondo Editorial (978-612-4148)

TABLA DE CONTENIDO

PRESENTACIÓN

El propósito del presente texto es presentar metodologías prácticas del uso de la estadística en la investigación científica en educación, mediante la aplicación de casos de estudio, técnicas estadísticas de análisis de datos, ejercicios propuestos, comentados y resueltos.

El texto está dirigido a los estudiantes y docentes, en los cursos de estadística e investigación como texto de consulta; así también para que los estudiantes puedan realizar sus proyectos de investigación o desarrollar su tesis al finalizar sus estudios. A los docentes les ayudará a tener un material para que puedan obtener ejemplos en el desarrollo de sus clases y en su trabajo de investigación.

Debido a que la cantidad de contenidos en estadística aplicada a la investigación es bastante amplia, hemos dividido el material en dos partes. Una de ellas es el presente texto, que trata del análisis estadístico descriptivo. La otra es un texto adicional que abarca el análisis inferencial.

Este texto busca introducirlos en los conceptos básicos de la investigación científica en educación y la estadística aplicada.

En el Capítulo 1 se presentan los fundamentos generales de la estadística, resaltando su importancia como herramienta para el análisis de datos en la investigación educativa.

El Capítulo 2 profundiza en dos conceptos clave: población y muestra. Se explican distintos métodos de muestreo y cómo determinar el tamaño adecuado de una muestra.

A continuación, el Capítulo 3 se enfoca en el análisis descriptivo y exploratorio de datos, detallando el proceso de recolección de datos, tipos de variables, detección de valores atípicos y técnicas de presentación de datos.

Finalmente, se estudian en profundidad las medidas descriptivas como tendencia central, dispersión, forma y posición no central.

Esperamos que encuentren en estas páginas conceptos y herramientas útiles para fortalecer sus conocimientos en estadística aplicada a la educación.

Los Autores.

CAPÍTULO 1

INTRODUCCIÓN A LA INVESTIGACIÓN CIENTÍFICA EN EDUCACIÓN Y LA ESTADÍSTICA

1.1. Fundamentos de la Estadística

La estadística se define como el estudio de los datos. Su objetivo es recopilar, organizar, analizar e interpretar datos para comprender y describir fenómenos presentes en la sociedad y en la ciencia.

La estadística es importante porque nos ayuda a entender situaciones de incertidumbre. Por ejemplo, si queremos evaluar la efectividad de una nueva estrategia educativa, la estadística permite diseñar estudios, recoger datos, realizar pruebas de hipótesis y detectar patrones y tendencias. Esto sustenta la decisión sobre si la estrategia es efectiva. Además, facilita comunicar resultados de investigación al identificar relaciones causales, estimar probabilidades y cuantificar la incertidumbre de los fenómenos estudiados. Esto permite presentar los hallazgos de forma clara y precisa.

Es así como la estadística tiene un rol crucial en el avance de la sociedad al aportar herramientas para describir situaciones de incertidumbre en investigaciones científicas. Su aprendizaje es clave en la sociedad actual, generando la necesidad de enseñar conceptos y prácticas de probabilidad y estadística en todos los niveles educativos.

Con el tiempo, las técnicas estadísticas han evolucionado en función de nuevos planteamientos teóricos y el desarrollo de software de análisis de datos (Castro & Lizasoain, 2012). Este desarrollo ha sido más rápido que la capacidad de las instituciones educativas para responder a la creciente demanda de su enseñanza, la cual se ha vuelto cada vez más trascendental en la sociedad moderna.

Es importante revisar contenidos y estrategias de enseñanza de probabilidad y estadística, áreas complejas y en rápido desarrollo. Su estrecha relación con otras áreas del conocimiento las hace más difíciles. Además, en el sistema educativo actual se están implementando cambios como nuevas formas de aprendizaje y uso de tecnología, lo que requiere revisar cuidadosamente las metodologías utilizadas (Batanero et al., 2000).

En educación secundaria, la mayoría de las aplicaciones de probabilidad se relacionan con juegos de azar, por ser familiares a los estudiantes y tener espacios muestrales finitos. No obstante, para valorar su rol, se debe ampliar la visión abordando diferentes áreas del conocimiento.

En educación superior, hay desafíos y cambios necesarios cuando se introducen conocimientos estadísticos provenientes de expertos no académicos (personas con alto nivel de conocimiento en estadística, pero no directamente afiliados a instituciones educativas). Esto puede afectar la estructura

de los programas educativos, los objetivos de enseñanza y la relación entre profesores y alumnos, por la cual es importante adoptar enfoques pedagógicos adecuados para enseñar estos conocimientos altamente especializados.

La estadística se ha incorporado en los programas educativos de todos los niveles por su rápido desarrollo, como tema dentro de matemáticas. Las metodologías para enseñar conceptos estadísticos y probabilísticos se basaban en las de conceptos matemáticos. Por ejemplo, en secundaria se crean tablas de frecuencia y se calculan medidas de tendencia central y dispersión. En educación superior, para abordar problemas de inferencia estadística, se usan herramientas basadas en teoría de probabilidades. Sin embargo, este enfoque en fundamentos teóricos de la ciencia limita la enseñanza de la estadística.

Para modificar y ampliar la forma de enseñar estadística, se debe considerar el aprendizaje del estudiante como factor relevante. Por ejemplo, se pueden utilizar distintas representaciones para comprender el concepto de función de distribución con relación a variables aleatorias, como gráficos, lenguaje coloquial y símbolos, esto involucrarían estructuras cognitivas con imágenes mentales, propiedades y procesos asociados. Para la cual, es importante considerar las ideas previas de los estudiantes sobre representaciones gráficas y aleatoriedad para construir la función de distribución sistemáticamente. Desde esta perspectiva se estaría considerando la forma de aprender de los estudiantes y los procesos propios de sus aprendizajes, caracterizando una enseñanza adaptada a sus necesidades.

El análisis de la adquisición de conceptos probabilísticos, el estudio histórico-epistemológico de las variables aleatorias y la introducción a las pruebas de hipótesis, son temas de investigación que ayudan a decidir cómo enseñar estos conceptos, considerando el papel del conocimiento en la discusión. Estos conceptos evidencian aspectos a considerar al determinar cómo enseñarlos. Además, se debe analizar cómo se construye socialmente el conocimiento, enfocándose en las prácticas relacionadas con los conceptos y considerando el impacto de la tecnología, las ciencias y los medios de comunicación.

Para Tamayo (2002), la investigación es un proceso que involucra la búsqueda de respuestas a preguntas específicas, la sistematización del conocimiento y la verificación de teorías. La investigación no se limita únicamente al análisis de datos. Aunque este proceso es una parte crucial de una investigación, la investigación en sí misma es un proceso más amplio y complejo, que incluyen la formulación de preguntas de investigación, la recopilación de datos, la revisión de literatura, el diseño de experimentos, la interpretación de resultados, entre otros.

Por lo tanto, al enseñar estadística en el contexto de la investigación educativa, se debe enfatizar en los procesos que tiene la investigación y que la estadística es una herramienta importante para abordar preguntas de investigación y tomar decisiones informadas en el campo educativo. Esto permitirá a los estudiantes aplicar eficazmente sus habilidades estadísticas en un contexto educativo y contribuir al avance de la educación a través de la investigación.

CAPÍTULO 2

POBLACIÓN Y MUESTRA

2.1. Población y muestra

Unidad de análisis

En el campo de la estadística, la unidad de análisis o unidad de observación hace referencia a la entidad o elemento que está siendo observado o estudiado dentro de un conjunto de datos (Vílchez, 2011, p.3). Más específicamente, es el objeto o cosa sobre el cual se recopila información. Por ejemplo, en un estudio que analiza el desempeño académico de los estudiantes en una escuela, la unidad de análisis sería cada uno de los estudiantes, pero también podría ser la clase, la escuela o incluso el distrito escolar.

Es fundamental seleccionar la unidad de análisis adecuada para que los resultados sean precisos y relevantes respecto al problema en cuestión. Asimismo, la unidad de análisis es importante para determinar el tamaño de la muestra necesario para obtener resultados estadísticamente significativos y para aplicar los métodos de análisis apropiados.

Ejemplos

- Un profesor mayor de 40 años, perteneciente a la UGEL Nº 2.
- Un colegio de adultos de la zona de La Victoria.

Las unidades de análisis están asociadas a los objetivos o preguntas del estudio, por ejemplo:

Tabla 1

Ejemplos de unidad de análisis

Pregunta del estudio	Unidad de análisis
¿Cómo influye la gestión educativa en la imagen de las instituciones educativas públicas del distrito de San Juan de Lurigancho?	Docentes y administrativos de las Instituciones educativas del distrito de San Juan de Lurigancho.
¿De qué manera la gestión de conflictos tiene implicancia con el clima organizacional en los docentes y administrativos de la Facultad de Ciencias Sociales y Humanidades de la Universidad Nacional de Educación?	Docentes y administrativos de la Facultad de Ciencias Sociales y Humanidades de la UNE (en este caso se considera *unidad de análisis compuesta*).

Pregunta del estudio	Unidad de análisis
¿De qué manera el clima organizacional tiene una relación con la satisfacción académica de los docentes de la Facultad de Ciencias Sociales y Humanidades de la Universidad Nacional de Educación?	Docentes que laboran en la Facultad de Ciencias Sociales y Humanidades de la Universidad Nacional de Educación.
¿De qué manera el uso de internet se relaciona con el rendimiento escolar en los estudiantes del V ciclo de educación primaria de la IE Experimental de Aplicación de la UNE?	Estudiantes, docentes y padres de familia (*Unidad de análisis compuesta*).
¿La aplicación del aprendizaje por descubrimiento permitirá el desarrollo de capacidades de los estudiantes en el área de Ciencia Tecnología y Ambiente del 1º grado de educación secundaria del Colegio Nº 0055 Manuel Gonzales Prada, Ugel 06 Chosica?	Estudiantes del 1º grado de educación secundaria del Colegio Nº 0055 Manuel Gonzales Prada.

Población

Es el conjunto total de las personas, objetos o medidas que se están estudiando y que tienen características comunes. Puede ser finita o infinita. La población infinita es un concepto que se toma en estadística cuando la población es demasiado grande (Martínez, 2003, p.7).

Ejemplos

- Población finita: Profesores mayores de 40 años, pertenecientes a la UGEL Nº 2.
- Población infinita: Estudiantes universitarios de Latinoamérica.

La población debe limitarse claramente entorno a sus características de contenido, lugar, espacio, volumen y tiempo.

Ejemplos

- Docentes de la especialidad de matemática e informática de la Facultad de Ciencias de la UNE, 2017.
- Egresados de la promoción 2010 de la Facultad de Ciencias de la UNE.
- Alumnos del turno diurno de la I.E. Juan Velasco de Lima año 2017.
- Estudiantes de la Facultad de Ciencias Sociales de la UNE.

Es importante definir claramente la población, ya que esta definición establecerá el alcance de las inferencias que se pueden derivar de los datos recopilados. Además, la elección de la población puede afectar la muestra, la definición de variables y cómo se analizan los datos.

Unidad de Muestreo

Es el elemento utilizado para seleccionar una muestra. Es decir, la unidad de muestreo es la entidad u objeto que se está seleccionando de la población para ser incluido en la muestra. En muchos casos la unidad de muestreo y la unidad de análisis son las mismas. Sin embargo, la unidad de análisis y la unidad de muestreo pueden ser diferentes según los objetivos y el diseño del estudio.

Ejemplo

Si se quiere estudiar el desarrollo del pensamiento lógico matemático en niños del nivel primario, es posible que no se cuente con información sobre cuántos son, donde viven, como se llaman, por lo que es difícil extraer una muestra de todos los niños (*unidad de análisis*). El investigador deberá tomar como muestra a las I.E. de nivel primario, para llegar a niños del nivel primario (*unidad de muestreo*).

Sin embargo, para el ejemplo anterior en el estudio del desarrollo del pensamiento lógico matemático en niños del nivel primario, también se debe considerar como factores a los docentes y a los padres. Por esta razón debe hacerse un muestreo estratificado en cada I.E. seleccionada.

Entonces, a la I.E. se le llama: *Unidad de muestreo compuesta o primaria*.

A los niños, docentes y padres: *Unidades de muestreo elementales*.

Muestra

Es un subconjunto seleccionado de una población con el objetivo de hacer inferencias (deducciones) sobre esta población (Diaz, 2015, p. 237).

Este subconjunto debe ser representativa y tener un tamaño mínimo apropiado, para hacer dichas inferencias. Puede contener Unidades de Muestreo Primarias (UMP). Las UMP son elementos individuales que se seleccionan directamente de la población para formar la muestra.

Ejemplo

Si se está realizando un estudio sobre el rendimiento académico de los estudiantes en una institución educativa, las UMP podrían ser los estudiantes tomados al azar.

También están las Unidades de Muestreo Secundarias (UMS), que se utilizan cuando no es posible o práctico seleccionar directamente las UMP de la población de estudio.

Ejemplo

Si se desea estudiar el rendimiento académico de los estudiantes de una ciudad. La población de estudio serían todos los estudiantes de esa ciudad, por lo que hacer una muestra de toda esa población de estudiantes sería costoso y difícil, por lo que se pueden utilizar las UMS.

En este caso, se podría elegir las instituciones educativas como UMS, seleccionando al azar algunas de estas Instituciones, luego se seleccionarían al azar algunos de estos estudiantes de cada una de estas Instituciones para formar la muestra.

Cuando la población es muy dinámica la muestra debe ser mayor y debe ser representativa, es decir las distintas variedades y matices de la población deben estar presentes proporcionalmente en la muestra.

Parámetro

Es una medida numérica que describe una característica de una población. Los parámetros pueden ser medidas de tendencia central, como la media o la mediana, o medidas de dispersión, como la desviación estándar o el rango (Martínez, 2003, p.9). También pueden ser medidas de asociación entre dos variables, como la correlación o la covarianza.

Ejemplo

Si se está interesado en la altura promedio de los estudiantes universitarios de una población, el parámetro sería la media de alturas de todos los estudiantes en la población de estudio.

Estadístico

Llamada también *estimador* o *estadígrafo*, es una medida numérica que se calcula en base a los datos de una muestra y se utilizan para hacer inferencias sobre la población (Martínez, 2003, p.9). Los estadísticos pueden ser medidas de tendencia central, como la media o la mediana, o medidas de dispersión, como la desviación estándar o el rango. También pueden ser medidas de asociación entre dos variables, como la correlación o la covarianza.

El valor del estadístico es conocido y varía de acuerdo con la muestra tomada.

Población (Parámetros)	Muestra (Estadísticos)
N	n
μ	$\bar{x}$
σ^2	s^2
…	…

Donde:

N : Tamaño de la población.

n : Tamaño de la muestra.

μ *:* Media poblacional.

$\bar{x}$: Media muestral.

σ^2 : Varianza de la población.

s^2 : Varianza de la muestra.

Ejemplo

Si se tiene la muestra de 100 estudiantes universitarios y se desea conocer la altura promedio de la población en total, entonces la media de las alturas de la muestra es un estadístico, y es un estimador de la media de la población es un parámetro.

Error de muestreo

Es la diferencia entre el valor del estadístico calculado en la muestra y el valor del parámetro en la población completa (Diaz, 2015, p. 240).

El error de muestreo es una consecuencia de la selección de una muestra, debido a que si la muestra representa solo una parte de la población entonces es posible el estadístico de la muestra difiera del parámetro de la población.

El error de muestreo está determinado por el tamaño de la muestra, es decir cuanto mayor sea la muestra menor será el error de muestreo.

En estadística el error de muestreo es fuente de incertidumbre y una forma de reducir este error es aumentar el tamaño de la muestra. La reducción del error de muestreo aumenta la probabilidad de obtener un resultado exacto.

Ejemplo

Población (Parámetro)	Muestra (Estadístico)
$\sigma^2=0.35$	$s^2=0.36$

0.01 (error de muestreo)

Existen fórmulas estadísticas que permiten calcular el tamaño de la muestra en función del error permisible y el nivel de confianza, como se verá posteriormente.

El error máximo permisible depende del nivel de confianza y el tamaño de la muestra. Por ejemplo, en el ámbito educativo, el nivel de confianza suele establecerse en 95%. Esto implica que el máximo error permisible debe ser del 5% (0.05). A partir de esta información, se selecciona el tamaño de la muestra adecuado para obtener resultados con una confianza del 95%.

Determinar el tamaño de la muestra es esencial para que los resultados del estudio sean confiables y representativos de la población objetivo. Con una muestra adecuada, se puede estimar con mayor precisión las características de la población y realizar inferencias válidas a partir de los datos obtenidos.

Error no muestral

Es un tipo de error que no está relacionado con el proceso de muestreo. Puede ser causado por diversos factores, como errores de medición, errores de registro de datos, errores de cálculo, sesgos de selección de la muestra, sesgos de no respuesta, entre otros (Fernández & Ruiz,2004, p.12).

Ejemplo

En una encuesta tomada a los estudiantes, muchos ítems quedaron sin responder, además en el ingreso de los datos en la computadora, se equivocaron e ingresaron un carácter en vez de un número.

Obtención de la muestra

La determinación de la muestra en el proceso de la investigación se realiza de manera sistemática luego de la definición de las variables en función a los objetivos de la investigación. Es decir, se tiene que:

1. Definir del problema.
2. Definir de los objetivos.
3. Definir de las variables.
4. Determinar la muestra en relación con los objetivos de la investigación.

El muestreo es la técnica para seleccionar una muestra representativa de una población. Cuando el tamaño de la muestra es menor que el de la población, se pueden extraer múltiples muestras de esa población. El conjunto de todas las posibles muestras que se pueden seleccionar de una población se conoce como *espacio muestral*. Cada muestra individual representa un subconjunto de elementos de la población total.

2.2. Métodos de muestreo

Existen diferentes métodos de muestreo, que se realizan en función a las características de la población, objetivos, tiempo y recursos (Cochran, 1982, p. 28). Se agrupan en:

Muestreo no probabilístico

Es un método de selección en la que no se conoce o no se puede calcular la probabilidad de que una unidad de muestreo sea seleccionada para formar parte de la muestra. En este método el investigador selecciona las unidades de muestreo según considera representativas o que cree que puedan representar relevante para el estudio (Hernández et al.,2014, p.176).

Se debe tener en cuenta que el muestreo no probabilístico no permite estimar el error muestral y puede llevar a resultados sesgados.

El muestreo no probabilístico incluye:

- *Muestro por conveniencia*: Se refiere a la selección de unidades de muestreo que están disponibles y son fáciles de acceder para el investigador.

  Por ejemplo, si un investigador desea estudiar la satisfacción de los estudiantes respecto a los servicios de la universidad, puede seleccionar a los estudiantes que están disponibles en ese momento de la investigación.

- *Muestreo por juicio*: Se refiere a la selección de unidades de muestreo en base a la opinión o criterio subjetivo del investigador.

Por ejemplo, si el investigador desea estudiar la calidad del servicio educativo en una Institución educativa, puede seleccionar en su opinión a los estudiantes que hayan recibido una atención de calidad.

- *Muestreo de cuotas*: Se refiere a la selección de unidades de muestreo en base a las características específicas de la población, como la edad, el género o la ubicación geográfica.

  Por ejemplo, si el investigador desea estudiar la opinión de los habitantes de una ciudad sobre un tema determinado, puede seleccionar una muestra de acuerdo con la cuota de edad, género y ubicación geográfica de la población.

Muestreo probabilístico

Es un método de muestreo que se basa en la probabilidad conocida de que cada unidad de la población pueda ser seleccionada para formar parte de la muestra. Es decir, cada unidad de la población tiene la misma oportunidad de ser elegida. Es un método más riguroso y recomendable para estudios científicos (Hernández,2014, p.175).

Algunos métodos de muestreos son los siguientes:

- *Muestreo aleatorio simple*: Este tipo de muestreo es ideal cuando existen poblaciones homogéneas y pueden ser aplicadas a poblaciones infinitas o finitas. Es un mecanismo ideal para elegir la mejor muestra posible; donde cada elemento de la población tiene la misma probabilidad de ser escogido.

  Tienen dos formas:

  - Muestreo aleatorio con reposición.
  - Muestreo aleatorio sin reposición.

- *Muestreo estratificado:* En este muestreo la población se divide en estratos o subgrupos homogéneos en términos de una o varias características relevantes. Seguidamente se selecciona una muestra aleatoria de cada estrato.

- *Muestreo sistemático:* Implica seleccionar una unidad de muestreo aleatoria de la población y a partir de ahí seleccionar a cada k-ésima unidad de la población hasta alcanzar el tamaño de muestra deseado. La elección de k depende del tamaño de la población y del tamaño de la muestra.

- *Muestreo por conglomerado:* Implica dividir la población en grupos más grandes o conglomerados, seleccionar al azar algunos de estos conglomerados, y luego tomar una muestra de las unidades dentro de los conglomerados seleccionados. Esta técnica es útil cuando es costoso o difícil acceder a las unidades individuales de la población, pero es posible acceder a los conglomerados.

2.3. Tamaño de la muestra

El tamaño de la muestra es el número de unidades de muestreo que se seleccionan para realizar un estudio. Hay varios factores que determinan el tamaño de la muestra como son: el tamaño de la población, el nivel de confianza, la variabilidad de la población, entre otros.

Se espera que un tamaño de muestra mayor conduzca a resultados más precisos y menos sesgadas, sin embargo, pueden afectar el tiempo y costo requerido para llevar a cabo el estudio. Por lo que se debe de elegir un tamaño de la muestra adecuado y que sea representativo de la población, sin incurrir en costos excesivos y tiempos prolongados.

Lo importante es elegir el tamaño de la muestra no la proporción que representa la muestra del total de la población.

Ejemplo

Si se tiene 100 individuos, se debe tomar al menos el 30%. Pero si la población tiene 40000 individuos, una muestra de 30% representará el 12000, el 10% representará 4000 casos y el 1% dará una muestra de 400. En este caso lo más adecuado sería una muestra del 1%.

Si la población es homogénea se requiere pocos elementos en la muestra y si la población es heterogénea se requiere un mayor número de elementos.

El ***nivel de confianza*** es la certeza o grado que se tiene en una estimación o intervalo de confianza obtenida a partir de una muestra de datos. Se expresa en porcentajes, y es asumida por el investigador antes de realizar el tratamiento estadístico. Los valores más comunes son: 90%, 95% (en estudios de educación) y 99% (en estudios de medicina). Cuanto mayor es el nivel de confianza se aumenta el intervalo de confianza por lo tanto hay mayor variabilidad para que una estimación sea más precisa.

Z es el valor crítico de la distribución normal estándar, utilizado para calcular los límites inferior y superior de un intervalo de confianza para un parámetro poblacional como la media o proporción. El valor de Z depende del nivel de confianza deseado, este valor puede ser encontrado en la tabla de valores críticos de la distribución normal.

Nivel de confianza	90%	95%	99%
Z	1,64	1,96	2,57

Para establecer una fórmula del tamaño de muestra se requiere saber el tipo de parámetro que se desea estimar (Media o Proporción), es decir, si el interés es estimar una media aritmética se requiere una fórmula específica y si se quiere estimar una proporción se considera otra fórmula.

Para los estudios en educación se suelen hacer uso de las fórmulas para calcular el tamaño de la muestra para estimar *proporción* del parámetro.

A. Cuando la varianza muestral (s^2) es conocida:

A.1. Para poblaciones infinitas o tamaños de población desconocida:

$$n = \frac{Z^2 s^2}{E^2}$$

n : Tamaño de la muestra.

E : Margen de error permisible.

Z : Valor crítico de la distribución normal estándar correspondiente al nivel de confianza deseado.

Para un nivel de confianza del 95%, Z es igual a 1.96.

s^2 : Varianza muestral.

Ejemplo

Supongamos que queremos estimar la proporción de estudiantes universitarios que están a favor de un nuevo programa de becas. No conocemos el tamaño de la población universitaria, pero sabemos que la varianza de la proporción de estudiantes a favor en la población es de 0,06.

Entonces tomando un nivel de confianza al 95%, se tiene que:

E = 0,02 (Margen de error permisible 2%, asumida por el investigador)

Z = 1,96 (Para un nivel de confianza del 95%)

s^2 = 0,06 (Varianza muestral)

Sustituyendo los valores en la fórmula, se obtiene:

$$n = \frac{Z^2 s^2}{E^2} = \frac{(1{,}96)^2(0{,}06)}{(0{,}02)^2} \approx 576$$

Por lo tanto, necesitaríamos una muestra de al menos 576 estudiantes, para estimar la proporción de estudiantes estarían a favor de las becas, con un nivel de confianza del 95% y error muestral máximo permitido del 2%.

A.2. Para poblaciones finitas (N) o conocidas:

$$n = \frac{NZ^2 s^2}{E^2(N-1) + Z^2 s^2}$$

Ejemplo

Supongamos que queremos estimar la cantidad de estudiantes de una universidad que tienen un empleo de medio tiempo. Sabemos que el tamaño de la población es de 10000 estudiantes y que la varianza de la proporción de estudiantes con empleo de medio tiempo en la población es de 0,04.

Entonces tomando un nivel de confianza al 95%, se tiene que:

E = 0,02 (Margen de error permisible 2%, asumida por el investigador)

Z = 1,96 (Para un nivel de confianza del 95%)

s^2 = 0,04 (Varianza muestral)

N = 10000

Sustituyendo los valores en la fórmula, se obtiene:

$$n = \frac{NZ^2 s^2}{E^2(N-1) + Z^2 s^2} = \frac{(10000)(1{,}96)^2(0{,}04)}{(0{,}02)^2(10000-1) + (1{,}96)^2(0{,}04)} \approx 370$$

Por lo tanto, necesitaríamos una muestra de al menos 370 estudiantes, con un nivel de confianza del 95% y error muestral máximo permitido del 2%.

B. Cuando la varianza muestral (s^2) es desconocida

B1. Para poblaciones infinitas o tamaños de población desconocida:

$$n = \frac{Z^2 pq}{E^2}$$

Donde:

n : Tamaño de la muestra.

E : Margen de error permisible.

p : Proporción esperada de la población.

q : Complemento de la proporción esperada en la población (*1-p*).

Z : Valor crítico de la distribución normal estándar correspondiente al nivel de confianza deseado. Para un nivel de confianza del 95%, Z es igual a 1.96.

Ejemplo

Supongamos que se desea estimar la proporción de estudiantes de una determinada ciudad que aprueban un examen de matemáticas en una escuela pública. Se sabe que, en la población general, el 60% de los estudiantes aprueban el examen.

Entonces tomando un nivel de confianza al 95%, se tiene que:

E = 0,05 (Margen de error permisible 5%)

p = 0,6 (Proporción de estudiantes que aprueban un examen 60%)

q = 0,4 (Los que no aprueban el examen 40%, complemento de p)

Z = 1,96 (Para un nivel de confianza del 95%)

Sustituyendo los valores en la fórmula, se obtiene:

$$n = \frac{Z^2 pq}{E^2} = \frac{(1{,}96)^2(0{,}6)(0{,}4)}{0{,}05^2} \approx 369$$

Por lo tanto, se necesitaría una muestra de al menos 369 estudiantes para estimar la proporción de estudiantes que aprueban el examen de matemáticas con un nivel de confianza del 95% y un margen de error del 5%.

B2. Para poblaciones finitas (N) o conocidas:

$$n = \frac{NZ^2 pq}{E^2(N-1) + Z^2 pq}$$

Donde:

N : Tamaño de la población.

Ejemplo

Estimar el tamaño de la muestra de estudio para una población de 120 estudiantes de primer año de educación secundaria.

Entonces no se conoce la variación muestral. La población es finita. Tomando un nivel de confianza al 95%, se tiene que:

N = 120

E = 0,05 (Margen de error permisible 5%)

p = 0,5 (Asumida por el investigador)

q = 0,5 (Asumida por el investigador)

Z = 1,96 (Para un nivel de confianza del 95%)

Sustituyendo los valores en la fórmula, se obtiene:

$$n = \frac{120(1{,}96)^2(0{,}5)(0{,}5)}{(0{,}05)^2(120-1)+(1{,}96)^2(0{,}5)(0{,}5)} \approx 92$$

Por lo tanto, se necesitaría una muestra de 92 estudiantes para estimar la proporción de estudiantes con un nivel de confianza del 95% y un margen de error del 5%.

Cuestionario

1. *La unidad de análisis o unidad de observación hace referencia a:*
   a) El conjunto total de las personas u objetos que se están estudiando.
   b) El elemento que se está seleccionando de la población para la muestra.
   c) La entidad o elemento que está siendo observado o estudiado.
   d) El grupo al que pertenece la población de estudio.
2. *La población infinita en estadística es:*
   a) Una población muy grande donde se desconoce el total de elementos.
   b) Una población con características desconocidas.
   c) Una población donde todos los elementos son accesibles.
   d) Una población heterogénea.
3. *Los parámetros son:*
   a) Medidas que describen una muestra.
   b) Medidas que describen una población.
   c) Errores en la medición.
   d) Variables cualitativas.
4. *La muestra debe ser:*
   a) Mayor que la población.
   b) Menor que la población.
   c) Igual a la población.
   d) Representativa y con un tamaño mínimo apropiado.
5. *En el muestreo probabilístico:*
   a) Se desconoce la probabilidad de selección de los elementos.
   b) La selección se basa en criterios subjetivos.
   c) Todos los elementos tienen la misma probabilidad de ser elegidos.
   d) Solo se eligen elementos accesibles.
6. *El nivel de confianza del 95% implica que:*
   a) El 5% de los datos son incorrectos.
   b) Hay un 95% de certeza en los resultados.
   c) El error máximo permisible es 5%.
   d) Se tomó el 95% de la población.
7. *Si en una población desconocida la proporción esperada es 0.6, ¿cuál es el valor de q?*
   a) 0.4
   b) 0.6
   c) 1
   d) No se puede calcular.
8. *Si la población es muy heterogénea, el tamaño de muestra debe ser:*
   a) Pequeño.
   b) Grande.
   c) Igual al tamaño de la población.
   d) No importa el tamaño.
9. *La precisión de los resultados:*
   a) Aumenta al reducir el tamaño de muestra.
   b) Disminuye al aumentar la variabilidad de la población.
   c) No depende del nivel de confianza.
   d) Es máxima cuando la muestra es menor al 30%.

Alternativas correctas: c, a, b, d, c, c, a, b, b

Ejercicios

1. Se desea estimar el porcentaje de estudiantes de secundaria de Lima Metropolitana que han recibido clases virtuales durante la pandemia. Se conoce que la proporción esperada en la población es de 0.8. Si se desea un nivel de confianza del 95% y un error máximo permisible del 4%, ¿cuál es el tamaño mínimo de muestra requerido?
2. Se desea realizar un estudio para estimar la proporción de docentes de primaria satisfechos con las capacitaciones recibidas el último año en una UGEL que tiene 450 docentes. Si se desea un nivel de confianza del 95% y un error máximo permisible del 5%, ¿cuál es el tamaño mínimo de muestra requerido si la población es finita?

CAPÍTULO 3

ANÁLISIS DESCRIPTIVO Y EXPLORATORIO DE LOS DATOS

3.1. Recopilación de los datos

Una vez que hemos elegido el diseño de investigación adecuado y una muestra apropiada que se ajuste a nuestro problema de estudio y, en caso de haberse establecido, nuestras hipótesis, el siguiente paso implica la obtención de los datos relevantes sobre los atributos, conceptos o variables de las unidades de muestreo (Hernández et al., 2014, p. 198).

Para llevar a cabo la recolección de datos en una investigación educativa de manera efectiva, se pueden seguir los siguientes pasos:

1. **Definición detallada de variables**: Comenzamos por definir de manera precisa las variables o aspectos que deseamos medir, basándonos en los objetivos de la investigación. Por ejemplo, podríamos considerar variables como rendimiento académico, motivación y estrategias de aprendizaje.
2. **Selección y desarrollo de instrumentos**: Luego, elegiremos o crearemos los instrumentos más apropiados para recolectar datos relacionados con estas variables. Ejemplos de tales instrumentos incluyen cuestionarios, escalas de actitudes, pruebas de conocimiento, guías de observación y guiones de entrevistas. Estos instrumentos se ajustarán específicamente a las características de las variables que estamos estudiando.
3. **Determinación de muestra representativa**: A continuación, seleccionaremos una muestra representativa de la población que deseamos estudiar. Esto podría implicar el uso de técnicas de muestreo específicas para garantizar que la muestra sea verdaderamente representativa de la población objetivo. Nuestra muestra comprenderá estudiantes, docentes, padres u otros grupos relevantes según el enfoque de nuestra investigación.
4. **Aplicación de instrumentos con consideraciones éticas**: Una vez que tengamos nuestra muestra, aplicamos los instrumentos de recolección de datos. Esto podría realizarse de manera presencial, virtual o incluso mediante un enfoque mixto, dependiendo del contexto y las circunstancias. Es fundamental seguir protocolos éticos en todo momento y obtener consentimientos informados antes de la recolección de información.
5. **Organización y almacenamiento de datos**: A medida que recopilamos los datos, los organizamos y almacenamos de manera adecuada en bases de datos, tablas, matrices u otros formatos que sean apropiados para el tipo de análisis que realizaremos más adelante. Esto nos ayudará a mantener la información ordenada y lista para el análisis.

6. **Limpieza y depuración de datos**: Una vez que tengamos todos los datos, verificamos la calidad, completitud y consistencia de la información para asegurarnos de que sea confiable y precisa. Esta fase es crucial para garantizar un análisis posterior preciso y significativo.
7. **Análisis cuantitativo y cualitativo**: Analizamos los datos cuantitativos aplicando técnicas estadísticas pertinentes que nos permitan extraer información valiosa. Por otro lado, los datos cualitativos serán sometidos a análisis de contenido y codificación para identificar patrones y temas emergentes.
8. **Interpretación y contextualización de resultados**: Al interpretar los hallazgos, los relacionamos directamente con los objetivos y preguntas de investigación planteados al principio. Esto nos permitirá entender la relevancia y el impacto de los resultados en el contexto educativo en estudio.
9. **Elaboración de conclusiones y recomendaciones aplicables**: Basándonos en los resultados obtenidos, elaboramos conclusiones sólidas y recomendaciones aplicables al contexto educativo investigado. Explicamos cómo estos resultados pueden ser implementados en la práctica, brindando un valor práctico a los profesionales y partes interesadas.

3.2. Técnicas de Análisis de Datos

Hay varias técnicas de análisis de datos que se pueden utilizar, su elección dependerá de los datos que se estén analizando y de los objetivos de la investigación.

Las técnicas de análisis de datos se clasifican en 2 tipos:

1. Técnicas de análisis cualitativo.
2. Técnicas de análisis cuantitativo.

3.2.1. Técnicas de análisis cualitativo

El análisis cualitativo es un enfoque de investigación centrada en el análisis la comprensión de los datos sin utilizar técnicas estadísticas.

El objetivo de las técnicas de análisis cualitativo es resumir, analizar e interpretar la información obtenida mediante métodos cualitativos (entrevistas, análisis documental, etc.). A continuación, presentamos algunas técnicas más comunes:

- Análisis de discurso: Utilizada para construir significados y representaciones sociales a partir del análisis del lenguaje. Implica la identificación de patrones.
- Análisis de contenido: Esta técnica analiza datos no estructurados como entrevistas o documentos escritos. Permite la identificación de patrones y temas en el texto.
- Análisis de caso: Utilizada para estudiar un caso específico en profundidad con el objetivo de entender su complejidad. Permite la identificación de patrones, causas y perspectivas.
- Análisis fenomenológico: Utilizada para entender la experiencia subjetiva de las personas. Implica la identificación de patrones y su interpretación en términos de como las personas experimentan el mundo.
- Teoría fundamentada: Utilizada para desarrollar teorías a partir de los datos en lugar de aplicar una teoría existente para analizar los datos.

3.2.2. Técnicas de análisis cuantitativo

El objetivo de esta técnica es describir, graficar, comparar, relacionar y resumir datos utilizando herramientas estadísticas. A continuación, se mencionan las más comunes:

Tabla 2

Técnicas de análisis cuantitativo

Usos	Descripción	Técnicas
Describir variables	Caracterizar una muestra variable por variable	Distribución de frecuencias
		Porcentajes
		Promedios, desviación estándar
		Gráficos (de barra, de sectores, histogramas)
Comparar grupos	Se compara la diferencia entre grupos de la muestra según las variables seleccionadas	T de student
		Análisis de varianzas
		Kruskall-Wallis
		Gráficos de barras múltiples
Analizar la relación entre variables	Determinar la relación entre 2 o más variables	R de Pearson
		Rho de Spearman
		Chi-Cuadrado
		Análisis de regresión (lineal, logística, multinominal)
		Análisis de correspondencia
		Gráfico de dispersión
Analizar la validez	Analizar la validez de constructo de los instrumentos de medición	Análisis factorial
		Análisis de Clusters o conglomerados
		Escalamiento multidimensional
Analizar la confiabilidad	Analizar la confiabilidad de los instrumentos de medición	Coeficiente de Alfa de Crombach
		Coeficiente Omega de McDonald
		Análisis de componentes principales
		Teoría de respuesta al ítem (TRI)
		Análisis de confiabilidad test-retest

3.3. Recolección de datos

La recolección de datos es esencial en una investigación y se realiza mediante medición o conteo de variables de interés en la población o muestra. Los datos se clasifican en **secundarios**, obtenidos de fuentes previas como revistas y registros, y **primarios**, generados por el investigador directamente de unidades de observación. Las técnicas para recolectar datos primarios incluyen cuestionarios, entrevistas, observaciones y experimentos (Martínez, 2003, p.16; Rivera, 2017, p. 8).

Tabla 3

Técnicas de recolección de datos primarios

Técnica	Descripción	Ejemplo
Observación	Recopilación de datos a través de la observación directa de comportamientos y eventos.	- *Simple: S*e capta información intentando pasar desapercibido. - *Participante:* El observador se integra al entorno observado.
Entrevista estructurada	Basada en un listado fijo de preguntas y la interacción se limita a realizar la pregunta y anotar la respuesta. En estadística se acostumbra a utilizar este tipo de entrevista debido a que permite un procesamiento matemático.	- Encuestas. - Cuestionarios.
Entrevista no estructurada	Hay un margen de libertad al formular las preguntas y registrar las respuestas.	- Entrevistas en profundidad, grupos de discusión.
Experimentos	Prueba de hipótesis bajo condiciones controladas	- Experimentos de laboratorio, experimentos de campo.
Estudio de casos	Estudio exhaustivo y detallado de un caso específico.	- Estudios de casos individuales, estudios de casos múltiples.

3.4. Variables

Las variables son las características o propiedades que se investigan y que pueden variar o tomar diferentes valores. Estos valores son los datos recopilados y a través del análisis de los datos es posible examinar como las variables se relacionan, influyen o responden en diferentes condiciones (Mendenhall et al., 2006, p.8). Las variables se vuelven significativas en la investigación científica cuando se vinculan con otras variables, es decir, cuando forman parte de una hipótesis o teoría. En estas situaciones, comúnmente se les refiere como constructos o construcciones hipotéticas (Hernández et al., 2014, p. 105).

En las investigaciones cualitativas, donde no se requiere la formulación de hipótesis, resulta difícil identificar y medir variables de la misma manera que en las investigaciones cuantitativas. En la investigación cuantitativa, se suelen formular hipótesis, que son afirmaciones específicas sobre las relaciones entre variables. Estas hipótesis luego se prueban mediante métodos estadísticos para determinar si son válidas o no (Ñaupas,2018, p. 256).

Las variables se clasifican en: **cuantitativas** o **cualitativas**.

Las variables cualitativas pueden ser: ordinal y nominal (categóricas).

Las variables cuantitativas pueden ser: discretas o continuas.

Tabla 4

Tipo de variables

Variable	Tipo	Características	Ejemplo
Cualitativo	Nominales	Representan categorías o grupos discretos	Género (masculino/femenino), Estado civil (soltero/casado/divorciado).
	Ordinales	Tienen un orden o jerarquía inherente	Escala de calificación (muy insatisfecho/insatisfecho/neutro/satisfecho/muy satisfecho), Nivel de satisfacción (bajo/moderado/alto), Nivel educativo (primaria/secundaria/universidad).
Cuantitativo	Discretas	Representan valores numéricos individuales o contables	Número de hijos, Número de exámenes, Cantidad de aprobados.
	Continuas	Representan valores numéricos en una escala continua	Edad, Peso, Altura, Promedio, Tiempo transcurrido.

Clasificación de las variables desde la perspectiva de la medición

Para clasificar una variable desde la perspectiva de la medición se utiliza una escala de medición. Una escala de medición es un sistema de clasificación utilizado para asignar valores a las observaciones o respuestas en una variable. Las escalas de medición más comunes son: nominal, ordinal, intervalo y de razón (Sánchez & Reyes, 2009, p.158).

Tabla 5

Tipo de variables según su escala de medición

Tipo de variable	Descripción	Ejemplo
Variables Nominales	Representan categorías o grupos sin un orden específico.	Género (masculino/femenino), Estado civil (soltero/casado/divorciado). Aunque se numeren los valores no se pueden operar aritméticamente.
Variables Ordinales	Tienen un orden o jerarquía inherente, pero las diferencias entre los valores no son uniformes.	Escala de calificación (muy insatisfecho/insatisfecho/neutro/satisfecho/muy satisfecho)

Tipo de variable	Descripción	Ejemplo
Variables de Intervalo	Tienen un orden y las diferencias entre los valores son uniformes, pero no hay un punto de referencia absoluto (El 0 es arbitrario).	Rendimiento académico, Inteligencia, Promedio (con un punto de referencia arbitrario)
Variables de Razón	Tienen un orden, las diferencias entre los valores son uniformes y tienen un punto de referencia absoluto (El 0 es absoluto, significa ausencia o inexistencia).	Edad, Peso, Ingresos, Tiempo transcurrido, Tiempo de reacción.

Según el tipo de variable y la escala, si la variable de estudio es cuantitativa y de escala de intervalo o de razón, si además se confirma la normalidad de los datos y su homogeneidad se debe de utilizar una estadística paramétrica, en caso contrario se utiliza una estadística no paramétrica.

Si se tiene una variable cualitativa ya sea con escala nominal u ordinal se utiliza una estadística no paramétrica. Las pruebas estadísticas paramétricas o no paramétricas que se deben de elegir dependiendo del diseño de la investigación.

En el diseño se determinan los casos de estudio, en la Tabla 6 se puede observar los casos, las pruebas estadísticas que se realizan y los supuestos que se deben de verificar y evaluar antes de aplicar la prueba. Si los supuestos no se aplican entonces podría llevar a resultados poco confiables y a conclusiones erróneas.

Otra forma de categorizar a las variables es como **variables dependientes** y **variables independientes**. Kerlinger & Lee (2002) definen la variable independiente como la variable que un investigador manipula o controla en un experimento para observar su efecto en la variable dependiente. La variable independiente es la causa o el factor que se considera que tiene un impacto sobre la variable dependiente. La variable dependiente es la variable que se mide o se observa en respuesta a la manipulación de la variable independiente, representa el resultado o el efecto que se está estudiando en la investigación (p. 42).

Ejemplo

Pregunta de investigación: ¿Cómo afecta el tiempo dedicado al estudio (*variable independiente*) al rendimiento académico de los estudiantes (*variable dependiente*) en matemáticas?

Variable Independiente: Tiempo dedicado al estudio. Esta es la variable que el investigador puede manipular o controlar. En este caso, el investigador podría dividir a un grupo de estudiantes en subgrupos y asignarles diferentes cantidades de tiempo de estudio, como 1 hora, 3 horas y 5 horas por día.

Variable Dependiente: Rendimiento académico en matemáticas. Esta es la variable que se observa o mide para evaluar su relación con la variable independiente. En este ejemplo, el rendimiento académico podría medirse mediante calificaciones en exámenes de matemáticas, puntuaciones en

tareas y proyectos, o calificaciones finales en un curso.

Sánchez & Reyes (2009) también definen la **variables extrañas** o variables de confusión, como factores que pueden influir en la relación entre la variable independiente y la variable dependiente de una manera que el investigador debe considerar (p.75).

Ejemplo

En el ejemplo anterior, una variable interviniente podría ser la motivación de los estudiantes, ya que puede influir tanto en el tiempo dedicado al estudio como en el rendimiento académico. Si no se tiene en cuenta esta variable, podría confundir la relación entre el tiempo de estudio y el rendimiento académico. Otras variables intervinientes podrían ser: calidad de enseñanza, estrategias de estudio, factores socioemocionales, apoyo familiar, etc.

Además, existen las **variables de control**, las cuales se describen como aquellas variables que se mantienen constantes o se modifican de manera intencionada para prevenir que tengan un impacto en la relación entre la variable independiente y la variable dependiente. Su propósito principal es garantizar que cualquier alteración observada en la variable dependiente se pueda atribuir a la variable independiente y no a otras variables no deseadas o influencias externas que pudieran afectar los resultados de un experimento o estudio (Namakforoosh, 2008, p. 67).

Ejemplo

En el ejemplo anterior, una posible variable de control podría ser el **nivel de conocimientos previos matemáticos de los estudiantes**. Esta variable se mantendría constante o controlada al asegurarse de que todos los grupos de estudiantes elegidos para el estudio tengan un nivel parecido de conocimientos matemáticos antes de comenzar a asignarles distintas cantidades de tiempo de estudio.

Tabla 6

Estadísticas paramétricas y no paramétricas para variables cuantitativas

Diseño	Paramétricas	No paramétricas
Relación entre 2 variables numéricas	Correlación de Pearson (Linealidad, normalidad, homocedasticidad).	Correlación de Spearman (Monotonicidad, No asume distribuciones específicas). Correlación de Kendall (Monotonicidad, No asume distribuciones específicas).
	Regresión lineal (Linealidad, Independencia de errores, homocedasticidad, normalidad de errores, independencia de variables independientes).	
Comparación de 2 promedios independientes	Prueba t de Student para muestras independientes (Normalidad, homogeneidad de varianzas)	Prueba U de Mann Whitney (No asume distribuciones específicas).
	Prueba T de Welch (Normalidad, no se asume homogeneidad de varianzas)	
Comparación de 2 promedios relacionados	Prueba t de Student para muestras relacionadas (Normalidad de las distribuciones, homogeneidad de varianzas)	Prueba T de Wilcoxon (No asume distribuciones específicas).
Comparación de n promedios grupos independientes	Análisis de varianza (ANOVA). ANOVA de un factor. ANOVA de dos factores. (Normalidad, homogeneidad de varianzas).	Kruskal – Wallis (No asume distribuciones específicas).
Comparación de n promedios grupos relacionados	Anova de mediciones repetidas (Normalidad, igualdad de matrices de covarianza).	Friedman (No asume distribuciones específicas)

En la tabla anterior se observa las pruebas estadísticas para variables cuantitativas, si no se cumplen los supuestos, entonces se puede elegir una prueba no paramétrica.

Por ejemplo, cuando se quiere evaluar la relación entre dos variables mediante r de Pearson, pero estas variables no cumplen los supuestos de linealidad, normalidad, homocedasticidad de los datos, entonces podría utilizarse la Correlación de rho de Spearman.

Se debe tener en cuenta que tanto Pearson como Spearman miden la correlación y Chi-cuadrado mide la asociación entre variables. La correlación y asociación refieren una relación entre dos variables, pero se utilizan en contextos diferentes. La asociación se refiere a que, si existe una relación, mientras que la correlación es una medida de fuerza y dirección de la relación lineal entre estas dos variables.

Para las variables cualitativas se tienen las siguientes pruebas.

Tabla 7

Estadísticas para variables cualitativas

Diseño	Prueba estadística	Descripción
Estudios de asociación entre variables cualitativas.	Prueba de Chi-cuadrado (χ^2)	Se utiliza para evaluar la asociación entre dos variables cualitativas. Compara la frecuencia observada con la frecuencia esperada bajo la hipótesis nula de independencia entre las variables.
Estudios de asociación entre variables cualitativas con tamaños de muestra pequeños.	Prueba exacta de Fisher	Es una alternativa a la prueba de Chi-cuadrado cuando las condiciones de aplicación de esta última no se cumplen. Se utiliza para evaluar la asociación entre dos variables cualitativas en tablas de contingencia pequeñas.
Estudios con muestras apareadas o repetidas.	Prueba de McNemar	Se utiliza para evaluar la asociación entre dos variables cualitativas en un diseño de muestras apareadas o repetidas.
Estudios que involucran variables cualitativas binarias.	Coeficiente de correlación de phi (ϕ)	Mide la asociación o relación entre dos variables cualitativas binarias. Se basa en la tabla de contingencia 2x2 y varía entre -1 y 1.
Estudios que involucran variables cualitativas.	Coeficiente de contingencia (C)	Mide la fuerza de asociación entre dos variables cualitativas a través de la tabla de contingencia. También varía entre -1 y 1.
Comparación de más de 2 muestras relacionadas	Q de Cochran	Recomendado para observaciones >20.

Debe tenerse en cuenta también que el tamaño de los datos es un factor importante para elegir una prueba estadística. Entendiendo que el tamaño de los datos es el número de observaciones o casos incluidos en el estudio.

Para evaluar este factor debe de aplicarse pruebas adicionales como: poder estadístico, sensibilidad de la prueba y estabilidad de resultados. Estas pruebas se realizan generalmente utilizando los

paquetes estadísticos como el Lenguaje R, SAS, SPSS y G*Power. Cada uno de las cuales tienen sus propias sintaxis y procedimientos.

Sin embargo, una muestra grande de datos proporciona un mayor poder estadístico. Si las muestras son pequeñas entonces debiera hacerse pruebas de sensibilidad y también pruebas de estabilidad de los resultados.

3.5. Análisis exploratorio de los datos. Detección de datos ausentes y atípicos.

La exploración estadística inicial de los datos sirve para obtener estadísticos descriptivos y representaciones gráficas, tanto del conjunto total de datos como de subgrupos. Esto permite identificar valores atípicos, obtener descripciones de los datos, evaluar supuestos y caracterizar diferencias entre subpoblaciones.

La inspección puede indicar la presencia de valores extremos, discontinuidades u otras peculiaridades, lo que ayuda a determinar si las técnicas de análisis posteriores son adecuadas. También puede señalar la necesidad de transformar los datos o utilizar pruebas no paramétricas.

En resumen, la exploración inicial es útil para:

- Inspeccionar y describir los datos
- Identificar valores atípicos
- Comprobar supuestos estadísticos
- Caracterizar diferencias entre subgrupos

Existen paquetes estadísticos que ayudan en para esta fase exploratoria, indicamos al lector que revise el texto práctico en donde se muestran ejemplos utilizando el SPSS.

3.5.1. Distribución de frecuencias

Cuando los datos de una variable están dispersos, es difícil comprender su estructura observando los valores brutos. Sin embargo, al dividirlos en clases o categorías y crear una distribución de frecuencias, se puede revelar el patrón de dispersión y entender mejor cómo están distribuidos los datos a lo largo de la variable en cuestión. Esto es fundamental en estadísticas y análisis de datos para obtener información significativa a partir de conjuntos de datos dispersos (Pérez, 2013, p. 128).

Tabla de frecuencias

La tabla de frecuencias es una forma de agrupación de datos, cualitativos o cuantitativos, que facilita la lectura y el análisis de estos. Sirve para observar la frecuencia con la cual los datos adoptan ciertos valores.

Ejemplo

Tabla 8

Frecuencias de nivel de rendimiento

Nivel de rendimiento	Frecuencia absoluta (fi)	Frecuencia acumulada (Fi)	Frecuencia relativa (hi%)	Frecuencia relativa acumulada (Hi%)
Excelente	10	10	25	25
Bueno	15	25	37.5	62.5
Regular	8	33	20	82.5
Deficiente	7	40	17.5	100
Total	40		100	

En la tabla de frecuencias como la anterior se puede hallar:

- **Frecuencia absoluta (fi):** Es simplemente el número de veces que un valor específico se repite en un conjunto de observaciones que estamos estudiando.
- **Frecuencia absoluta acumulada (Fi):** Es la suma total de las frecuencias absolutas hasta cierto valor o intervalo en la lista de observaciones.
- **Frecuencia relativa (hi%):** Esta es la proporción de veces que aparece un valor específico en relación con el total de observaciones. Pero en lugar de expresarlo como un número, se presenta como un porcentaje para que sea más fácil de entender.
- **Frecuencia relativa acumulada (Hi%):** Aquí se suma las frecuencias relativas en orden para obtener una proporción acumulada.

En las tablas de frecuencia, cuando agrupamos los valores en intervalos, encontramos algunos términos adicionales: **clase, marca de clase** y **límites reales**. Todos ellos se explicarán más adelante, en el *caso de las variables cuantitativas continuas.*

Ejemplo

Si se tiene un conjunto de datos que representa las edades de un grupo de estudiantes: 18, 20, 22, 18, 25, 18, 22, 20, 24. La tabla de frecuencias es la siguiente:

Tabla 9

Frecuencias de las edades de los estudiantes

Edad	Frecuencia absoluta (fi)	Frecuencia acumulada (Fi)	Frecuencia relativa (hi%)	Frecuencia relativa acumulada (Hi%)
18	3	3	33.3	33.3
20	2	5	22.2	55.5
22	2	7	22.2	77.7
25	1	8	11.1	88.8
24	1	9	11.1	100
Total	9		100	

En esta tabla, la edad 18 años tiene una frecuencia absoluta de 3 (aparece tres veces en el conjunto de datos), una frecuencia acumulada de 3. La edad de 20 años tiene una frecuencia absoluta de 2, y una frecuencia acumulada de 5 (sumando los que tienen 18 y 20 años).

La frecuencia relativa se calcula dividiendo la frecuencia absoluta de cada valor entre el total de observaciones (en este caso, 9). Por ejemplo, la edad 18 tiene una frecuencia absoluta de 3, entonces su frecuencia relativa se calcula dividiendo 3 entre 9, lo que da como resultado 0.333 multiplicado por 100 es 33.3. La frecuencia relativa acumulada para los de 18 años es 33.3 y para los de 20 años es 55.5 (33.3 +22.2). De manera similar, se calculan los valores para el resto de los elementos en la tabla.

A continuación, se explicará la estructuración de una tabla de frecuencias para cada tipo de variable: cualitativa, cuantitativa discreta y cuantitativa continua.

a) Caso de las variables cualitativas

Para el caso de variables cualitativas se realiza un conteo de las observaciones.

Ejemplo

Supongamos que estamos investigando las preferencias de 20 estudiantes en cuanto a su asignatura favorita. Se tiene la siguiente tabla de tabulación:

Asignatura favorita	Frecuencia
Matemáticas	7
Ciencias	4
Lenguaje	5
Historia	2
Arte	2

La Tabla de frecuencias es la siguiente:

Tabla 10

Frecuencias de las preferencias de asignaturas de los estudiantes

Asignatura favorita	Frecuencia (fi)	Frecuencia acumulada (Fi)	Frecuencia relativa (hi%)	Frecuencia relativa acumulada (Hi%)
Matemáticas	7	7	35	35
Ciencias	4	11	20	55
Lenguaje	5	16	25	80
Historia	2	18	10	90
Arte	2	20	10	100
Total	20		100	

Interpretación: En base a la tabla de frecuencias, se puede observar que la asignatura favorita con la mayor frecuencia es Matemáticas, con 7 estudiantes seleccionándola. Aproximadamente el 35% de los estudiantes eligió Matemáticas como su asignatura favorita. Luego le siguen Ciencias y Lenguaje con 4 y 5 estudiantes respectivamente. El 90% de los estudiantes eligió alguna de estas cuatro asignaturas como su favorita. La asignatura menos preferida parece ser Historia y Arte, con solo 2 estudiantes seleccionándolas como su favorita.

b) Caso de las variables cuantitativas discretas

Las variables cuantitativas discretas son aquellas representadas sólo por números enteros. Ejemplo: número de libros leídos en un mes (para evaluar los hábitos de lectura), número de curso desaprobados (para evaluar estrategias de aprendizaje o mejoras de rendimiento), etc.

Ejemplo

Se tiene los datos de 15 estudiantes y la cantidad de cursos que desaprobaron. La tabla de frecuencias sería la siguiente:

Tabla 11

Frecuencias de cursos desaprobados por los estudiantes

Cursos Desaprobados	Frecuencia (fi)	Frecuencia acumulada (Fi)	Frecuencia relativa (hi%)	Frecuencia relativa acumulada (Hi%)
0	3	3	20	20
1	3	6	20	40
2	4	10	26.66	66.66
3	2	12	13.33	79.99
4	2	14	13.33	93.32
5	1	15	6.66	100
Total	15		100	

Interpretación: En base a la tabla de frecuencias, se puede observar que la mayoría de los estudiantes (67%) desaprobaron 2 o menos cursos. Además, el 40% de los estudiantes desaprobó solo 1 curso. La cantidad de estudiantes que desaprobaron 3, 4 o 5 cursos es menor, representando el 13% de los estudiantes en cada caso. Alrededor del 20% de los estudiantes no desaprobó ningún curso.

c) Caso de las variables cuantitativas continuas

En el caso de las variables continuas los datos pueden tener decimales, para este caso se utilizan intervalos para representarlos en una tabla de frecuencias.

Ejemplo

Los promedios ponderados de calificaciones de los estudiantes (los promedios se hallan con decimales y se pueden tener mínimos y máximos), tiempo dedicado al estudio en horas (desde muy bajos 0,3 horas hasta muy altas 5,30 horas), nivel de lectura, nivel de escritura, etc.

Así también, cuando se manejan más de 30 observaciones es necesario usar *intervalos* que permitan ordenar de forma práctica los valores.

Para crearlos existe todo un procedimiento e implica la aparición de 3 nuevas columnas:

- **Clase:** indica el número de intervalo del que se trata.
- **Marca de clase (X_i):** es un promedio de los límites del intervalo de clase i. Es el número representativo del intervalo.

- **Límites reales:** cada intervalo tiene números que representan sus límites, pero lo límites reales indican los verdaderos valores que toma una medición, ya que en los límites nominales son aparentes.

Ejemplo

Se tiene las edades de 30 profesores. Los datos son los siguientes:

30	43	58	61	70	42	58	39	60	55
57	49	61	69	43	46	69	44	59	62
66	71	70	65	37	40	61	65	56	38

Para realizar la tabla de frecuencias se debe de realizar lo siguiente:

1. Calcular el rango de los datos:

   Rango = Valor máximo - Valor mínimo

   Rango = 71 - 30 = 41

2. Calcular el número de intervalos:

   Utilizamos la regla de Sturges para determinar el número aproximado de intervalos para 30 datos:

   Número de intervalos = $1 + \log_2(30) = 1 + 4.91 = 5.91 \approx 6$

   Donde: $\log_2(30)$ es el logaritmo en base 2 de 30.

   Para 30 datos, el número de intervalos sería aproximadamente: 5.91, redondeamos a 6.

3. Calcular la amplitud del intervalo:

   Amplitud del intervalo = Rango / Número de intervalos

   Amplitud del intervalo = 41 / 6 ≈ 6.83.

   Redondeamos la amplitud del intervalo a 7 para simplificar los cálculos.

4. Calcular los límites de los intervalos:

   Límite inferior del primer intervalo es: Valor mínimo = 30

   Límite superior del último intervalo es: Valor máximo = 71

   A partir de estos límites, podemos calcular los límites para los otros intervalos sumando la amplitud del intervalo sucesivamente. Tenga en cuenta que debe de restar 1 al límite superior para evitar superposición, ya que se está tomando intervalos cerrados.

   30 +7=37 → [30,36]

5. Calcular la marca de clase (Xi)

   La marca de clase es la semisuma de los límites en cada intervalo.

6. Contar las frecuencias:

   En cada intervalo, contar cuántos datos caen en ese rango y registrar la frecuencia(fi).

Calcular la frecuencia acumulada (Fi) sumando las frecuencias de los intervalos anteriores.

7. Calcular las frecuencias relativas y las frecuencias relativas acumuladas:

   La frecuencia relativa (hi%) se calcula dividiendo la frecuencia (fi) en cada intervalo por el número total de datos y multiplicándolo por 100.

   La frecuencia relativa acumulada (Hi%) se calcula sumando las frecuencias relativas de los intervalos anteriores.

Tabla 12

Frecuencias de la edad de 30 profesores universitarios

Clase	Intervalo	Xi	fi	Fi	hi%	Hi%
1	[30,36]	33	1	1	3.3	3.3
2	[37,43]	40	7	8	23.3	26.6
3	[44,50]	47	3	11	10.0	36.6
4	[51,57]	54	3	14	10.0	46.6
5	[58,64]	61	8	22	26.7	73.3
6	[65,71]	68	8	30	26.7	100

Ejemplo

Supongamos que queremos analizar las calificaciones de un grupo de 20 estudiantes en un examen de matemáticas. Los datos son los siguientes:

16.5 17.2 8.3 17.9 6.8 9.1 17.6 18.9 7.4 8.2

17.7 16.9 8.5 8.1 17.8 16.7 18.7 17.3 19.0 18.4

En este caso:

1. Organizar los datos en orden ascendente:

   6.8 7.4 8.1 8.2 8.3 8.5 9.1 16.5 16.7 16.9

   17.2 17.3 17.6 17.7 17.8 17.9 18.4 18.7 18.9 19

2. Determinar el rango de los datos:

   Rango = Valor máximo - Valor mínimo = 19.0 - 6.8 = 12.2

3. Calcular el número de intervalos:

   Número de intervalos = $1 + \log_2(20) = 1+4.32 = 5.32 \approx 6$

   Para 20 datos, el número de intervalos sería aproximadamente: 5.32, redondeamos a 6.

4. Calcular la amplitud del intervalo:

   Amplitud del intervalo = Rango / Número de intervalos

   Amplitud del intervalo = 12.2 / 5 = 2.44

5. Calcular los límites de los intervalos:

   Límite inferior del primer intervalo es: Valor mínimo = 6.8

   Límite superior del último intervalo es: Valor máximo = 19.0

   A partir de estos límites, podemos calcular los límites para los otros intervalos sumando la amplitud del intervalo sucesivamente.

   Tenga en cuenta que el límite superior del primer intervalo debe ser tomado en el segundo intervalo, por lo tanto, dejamos el límite del primer intervalo abierto.

   6.8 + 2.44 = 9.24 → [6.8, 9.24>

   9.24 + 2.44 = 11.27 → [9.24, 11.27>

6. Calcular la marca de clase (Xi)

   La marca de clase es la semisuma de los límites en cada intervalo.

7. Contar las frecuencias:

   En cada intervalo, contar cuántos datos caen en ese rango y registrar la frecuencia(fi).

   Calcular la frecuencia acumulada (Fi) sumando las frecuencias de los intervalos anteriores.

8. Calcular las frecuencias relativas y las frecuencias relativas acumuladas:

   La frecuencia relativa (hi%) se calcula dividiendo la frecuencia (fi) en cada intervalo por el número total de datos y multiplicándolo por 100.

   La frecuencia relativa acumulada (Hi%) se calcula sumando las frecuencias relativas de los intervalos anteriores.

Tabla 13

Frecuencias de las calificaciones de 20 estudiantes en matemáticas

Clase	Intervalo	Xi	fi	Fi	hi%	Hi%
1	[6.8, 9.24>	8.02	7	7	35	35
2	[9.24, 11.27>	10.255	0	7	0	35
3	[11.27, 13.3>	12.285	0	7	0	35
4	[13.3, 15.33>	14.315	0	7	0	35
5	[15.33, 17.36>	16.345	5	12	25	60
6	[17.36, 19.39]	18.375	8	20	40	100

3.5.2. Presentación de datos

La presentación de los datos se hace fundamentalmente utilizando dos métodos: el método tabular y el método gráfico.

A. Método tabular

Consiste en una presentación resumida de la información usando tablas, pudiendo ser estos univariantes o bivariantes. Cada fila representa una unidad de observación o caso, y cada columna

corresponde a una variable o atributo específico. En este tipo de representación se recomienda evitar la sobrecarga de información, se debe mostrar solo la información relevante y necesaria, demasiados datos pueden dificultar la comprensión y el análisis.

Ejemplo

En una tabla univariante, la información es sobre una única variable. A continuación, se muestra la variable nivel educativo.

Tabla 14

Cantidad de estudiantes por nivel educativo

Nivel educativo	Cantidad de estudiantes
Primaria	150
Secundaria	200
Universitario	300

Ejemplo

En una Tabla bivariante se presenta la información sobre dos variables y su relación. En este tipo de tabla, se muestran los valores de dos variables diferentes y se exploran las relaciones entre ellas. En el ejemplo, se muestra la cantidad de estudiantes de una escuela según su nivel de *grado* y *género*.

Tabla 15

Cantidad de estudiantes por género y grado

Grado	Masculino	Femenino
Primero	30	25
Segundo	40	35
Tercero	25	30

B. Método gráfico

Un gráfico estadístico es la *presentación de la información por medio de figuras geométricas*. El objetivo primordial de un gráfico es dar una impresión visual de conjunto para una rápida y fácil comprensión. A continuación, se presenta los tipos de variables y los gráficos estadísticos comúnmente utilizados:

Tabla 16

Tipo de variables y gráficos estadísticos

Tipo de variable	Variable	Ejemplo
Cualitativa	Nominal	Gráfico de barras, Gráfico de sectores
	Ordinal	Gráfico de barras, Gráfico de sectores
Cuantitativa	Discretas	Histograma, Polígono de Frecuencia, Diagrama de Tallos y Hojas, Gráfico de Barras, Gráfico de Puntos
	Continuas	Histograma, Polígono de Frecuencia, Boxplot, Gráfico de Dispersión, Gráfico de Series de Tiempo

Representación gráfica de variables cualitativas

Gráficos de barras

Consisten en barras rectangulares de diferentes alturas, donde la altura de cada barra representa la frecuencia o el porcentaje de observaciones en cada categoría.

Ejemplo

A continuación, se muestra la variable nominal Género y se registra la cantidad de estudiantes en cada género. El gráfico de barras permite visualizar y comparar fácilmente la cantidad de estudiantes en cada género.

Tabla 17

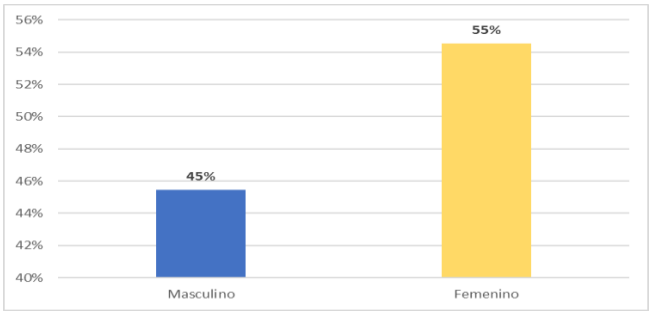
Frecuencia del género de estudiantes

Género	f_i	h_i *(%)*
Masculino	50	45%
Femenino	60	55%
Total	110	100%

La representación gráfica en barras se recomienda realizar en porcentajes para su mejor interpretación.

Figura 1

Cantidad de estudiantes por género



Interpretación: Según los resultados el 45% de los estudiantes son de género masculino, lo que equivale a 50 estudiantes. El 55% de los estudiantes son de género femenino, lo que equivale a 60 estudiantes.

Ejemplo

En el ejemplo se tiene una variable ordinal Nivel de Satisfacción y se registra la cantidad de estudiantes en cada nivel de satisfacción. Para representar estos datos en un gráfico de barras, se construye un eje vertical que representa la cantidad de estudiantes y un eje horizontal que muestra los niveles de satisfacción.

Tabla 18

Cantidad de estudiantes según el nivel de satisfacción

Nivel de Satisfacción	Cantidad de Estudiantes
Bajo	15
Medio	30
Alto	25
Muy alto	20

Para la representación hallamos los porcentajes respectivos.

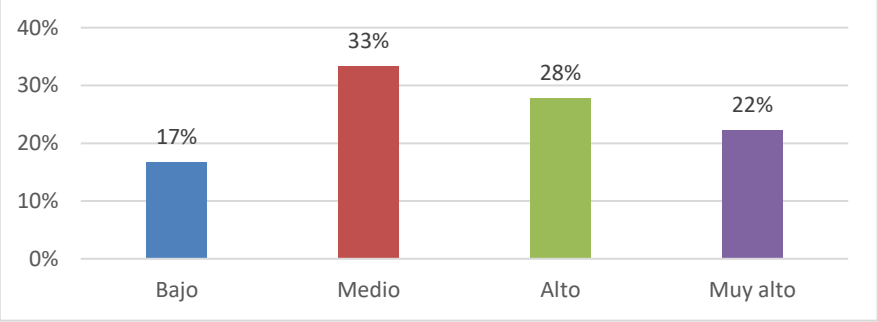
Tabla 19

Frecuencias de la satisfacción de estudiantes

Nivel de Satisfacción	f_i	h_i *(%)*
Bajo	15	17%
Medio	30	33%
Alto	25	28%
Muy alto	20	22%
Total	90	100%

Figura 2

Gráfico de barras de la cantidad de estudiantes por nivel de satisfacción



Interpretación: Según los resultados un tercio de los estudiantes (33%) se ubicaron en el nivel medio, mientras que un poco más de una cuarta parte se ubicaron en el nivel Alto (28%), un poco más de una quinta parte se ubicaron en el nivel Muy Alto (22%). Un 17% se ubicaron en el nivel bajo. Se observa que prácticamente la mitad de los estudiantes (50%) se ubicaron en el nivel Alto y Muy Alto.

Gráficos de sectores

Es un gráfico que es llamado también gráfico de pastel, este gráfico se representa mediante un círculo dividido en segmentos, el tamaño de dichos segmentos es proporcional al valor o frecuencia la que representa. Este gráfico es fácil de comprender y es útil cuando se requiere destacar una proporción en referencia a los demás segmentos. Se utiliza generalmente en variables nominales.

Ejemplo

Se tiene la información sobre la cantidad de datos de estudiantes por niveles de una institución educativa. Se requiere representarlo mediante un gráfico de sectores e interpretarlo.

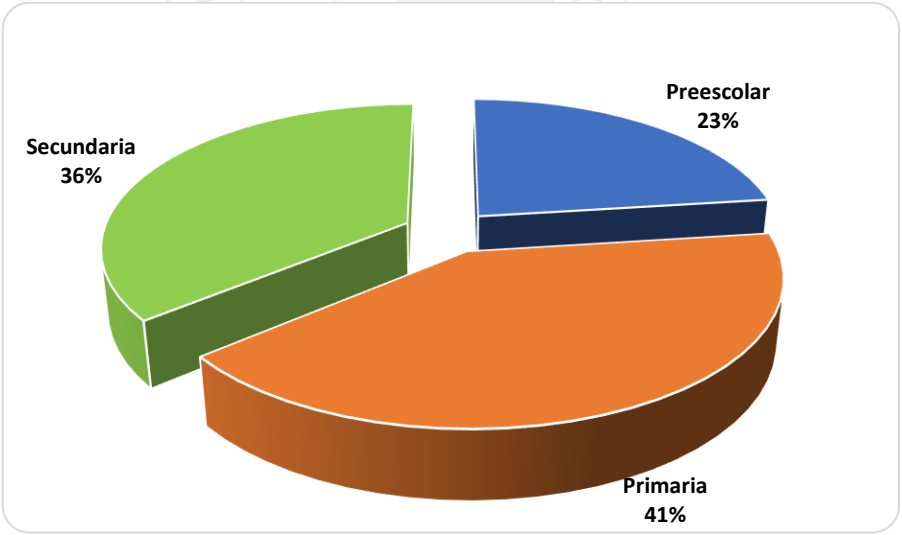
Tabla 20

Frecuencias de estudiantes por niveles

Nivel educativo	fi	hi(%)
Preescolar	500	23%
Primaria	900	41%
Secundaria	800	36%
Total	2200	100%

Figura 3

Porcentaje de estudiantes por niveles



Interpretación: Según los resultados más de una tercera parte de los estudiantes estaban en el nivel primaria (41%), un poco más de una tercera parte se encontraba en el nivel secundario (36%) y prácticamente una quinta parte estaban en el nivel Preescolar (23%).

Histograma

Es una representación gráfica que muestra la distribución de frecuencias de un conjunto de datos de una variable cuantitativa. El histograma resume y visualiza cómo los datos están distribuidos a lo largo de un rango de valores. Esta representación gráfica permite comprender la forma de cómo se distribuyen los datos (Devore,2008, p. 13).

Consideraciones:

1. Si se incluye los valores atípicos, estos van a influir en la representación gráfica y su interpretación sería afectada.
2. Se debe de elegir la cantidad adecuada de clases (intervalos). Demasiadas clases pueden ocultar patrones, mientras que muy pocas pueden perder detalles importantes.

Ejemplo

Se tiene la tabulación de las edades de los estudiantes. Estas edades están agrupadas por intervalos. Se requiere la representación en un histograma y su interpretación.

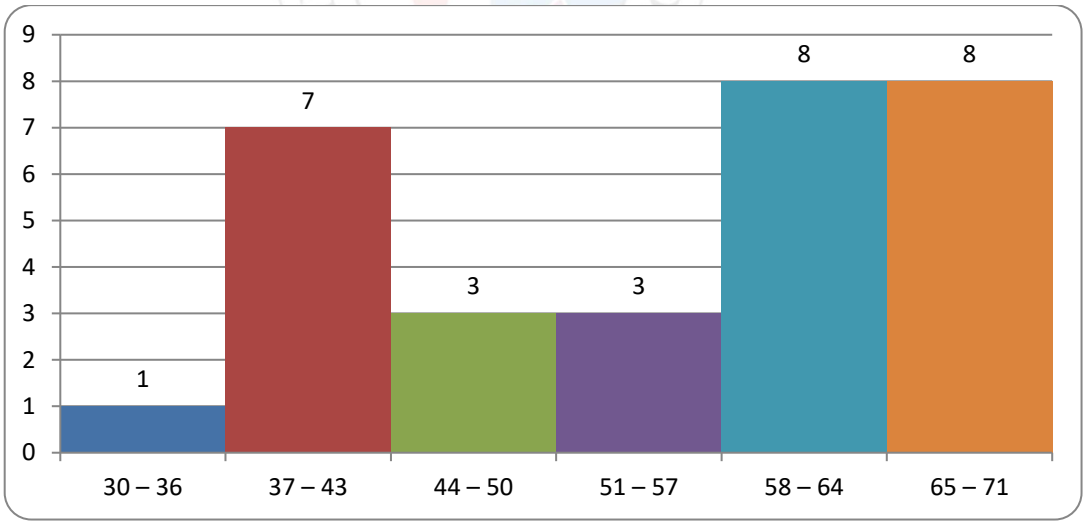
Tabla 21

Frecuencias por intervalos

Clase	*Edad*	*xi*	*fi*	*Fi*	*hi%*	*Hi%*	*Límites reales*
1	30 – 36	33	1	1	3.3	3.3	29.5 – 36.5
2	37 – 43	40	7	8	23.3	26.6	36.5 – 43.5
3	44 – 50	47	3	11	10.0	36.6	43.5 – 50.5
4	51 – 57	54	3	14	10.0	46.6	50.5 – 57.5
5	58 – 64	61	8	22	26.7	73.3	57.5 – 64.5
6	65 – 71	68	8	30	26.7	100	64.5 – 71.5
			30				

Figura 4

Histograma de las edades de los estudiantes



Interpretación: Según el histograma, podemos observar que la distribución de los datos no es simétrica, debido a que los datos no se distribuyen de manera uniforme a ambos lados. Se puede observar que la distribución no tiene la forma de la campana de Gauss (distribución normal). La mayoría de las edades de los estudiantes están entre los intervalos de 58 a 64 y 65 a 71.

3.6. Análisis descriptivo

3.6.1. Medidas descriptivas

Cuando se trabaja con tablas y gráficos o estos ya han sido construidos a partir de una colección de datos, se requieren medidas más exactas que faciliten el análisis de los datos. En la Tabla 22 se muestran las medidas de resumen más utilizadas en estadística.

Tabla 22

Tipo de variables y las medidas descriptivas

Tipo de medida	Medida de resumen	Descripción
Medidas de tendencia central	Media	Promedio de los valores en un conjunto de datos.
	Mediana	Valor central en un conjunto de datos ordenados.
	Moda	Valor o valores que ocurren con mayor frecuencia en un conjunto de datos.
Medidas de dispersión	Rango	Diferencia entre el valor máximo y el valor mínimo en un conjunto de datos. Indica la amplitud total de los datos.
	Desviación estándar	Medida de dispersión que indica cuánto varían los valores respecto a la media.
	Varianza	Medida de dispersión que representa la media de los cuadrados de las desviaciones de los valores respecto a la media.
	Coeficiente de variación	Medida de variabilidad relativa que se calcula dividiendo la desviación estándar por la media y multiplicando por 100.
Medidas de forma	Simetría	Valor que indica la igualdad o equilibrio en la forma de una distribución de datos.
	Curtosis	Valor que indica la forma o apuntamiento de una distribución de datos.
Medidas de posición no central	Cuartiles, quintiles, deciles y percentiles	Valores que dividen un conjunto de datos ordenados en n partes iguales.

Estas medidas descriptivas se utilizan de acuerdo con el tipo de variable que se estudia. En la Tabla 23, se muestran los tipos de variables y las medidas descriptivas que se pueden ser aplicados.

Tabla 23

Medidas descriptivas en los tipos de variables

Tipo de medida	Nominal	Ordinal	Intervalo/Razón
Medidas de tendencia central	Moda	Moda Mediana	Moda Mediana Media
Medidas de dispersión			Mínimo Máximo Desviación estándar Varianza Rango
Medidas de posición no central		Todos	Todos
Medidas de forma			Asimetría Curtosis

3.6.2. Medidas de tendencia central

Son medidas que reflejan la ubicación central de los datos, es decir, son aquellas que permiten hallar un solo valor numérico alrededor del cual los datos parecen agruparse de cierta manera. A partir de estos valores se puede analizar y comparar grupos de datos. En educación son las medidas más utilizadas, por ejemplo, para evaluar el rendimiento o desempeño académico se utiliza los promedios (media aritmética).

a) Media

Recibe también el nombre de *promedio*. Es un valor representativo de un conjunto de datos, el cual caracteriza a toda una distribución. En el caso de ser un estadístico, es decir hallado sobre la muestra, se representa con el siguiente símbolo: $\bar{x}$; mientras que, si se trata de un parámetro, es decir hallada a partir de la población se representa con: μ.

Media aritmética para datos no agrupados

La media se obtiene dividiendo la sumatoria de todos los datos en estudio sobre el número total de datos en estudio.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Donde:

$\bar{x}$: Media aritmética.

x_i : Es el i-ésimo dato.

n : Número total de datos (muestra).

$\sum_{i=1}^{n} x_i$: Suma de todos los datos.

Ejemplo

Se desea calcular la media aritmética de las calificaciones obtenidas por un grupo de estudiantes en un examen. Las notas son las siguientes: 15, 09, 15, 12, 18, 14, 13, 17, 12, 14.

La media aritmética será la siguiente:

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{15 + 9 + 15 + 12 + 18 + 14 + 13 + 17 + 12 + 14}{10} = \frac{139}{10} = 13{,}9$$

Consideraciones:

1. La media aritmética es sensible a datos extremos, es decir que si hay datos que son demasiados altos o demasiados bajos entonces el resultado podría afectar al resultado y tener una interpretación sesgada. Por lo cual antes de aplicar la media aritmética se debe de verificar si hay datos atípicos.
2. La media aritmética supone que se está trabajando con una distribución normal o una distribución aproximada a la distribución normal. Si los datos están sesgados o no tienen distribución normal entonces se recomienda la interpretación de los datos mediante la mediana.
3. Cuanto mayor es el número de datos, mayor es la precisión y confiabilidad de la interpretación de la media.
4. La interpretación de la media se complementa con otras medidas como la desviación estándar, mediana, percentiles, etc. para tener una visión más completa de los datos.

Media aritmética para datos agrupados

La media para datos agrupados es una medida de tendencia central que representa el promedio de los datos cuando están agrupados por intervalos.

Se utiliza la siguiente fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{k} X_i . f_i}{n}$$

Donde:

$\bar{x}$: Media aritmética.

n : Número total de datos (muestra).

X_i : Marca de clase del i-ésimo intervalo (semisuma de los valores extremos del intervalo).

f_i : Frecuencia de clase del i-ésimo intervalo.

k : Número de intervalos.

$\sum_{i=1}^{k} X_i . f_i$: Suma del producto de la marca de clase y la frecuencia absoluta.

Ejemplo:

En la Tabla 24, se tiene los datos sobre las edades de los docentes. Se desea hallar media aritmética.

Tabla 24

Frecuencias de la edad de 30 profesores universitarios

Clase	Intervalo	Xi	fi
1	[30,36]	33	1
2	[37,43]	40	7
3	[44,50]	47	3
4	[51,57]	54	3
5	[58,64]	61	8
6	[65,71]	68	8
		Total	30

Recordando que Xi representa la semisuma de los extremos del intervalo, obtenemos que:

$$\bar{x} = \frac{\sum_{i=1}^{k} X_i . f_i}{n} = \frac{\sum_{i=1}^{6} X_i . f_i}{30}$$

$$\bar{x} = \frac{(33 \times 1) + (40 \times 7) + (47 \times 3) + (54 \times 3) + (61 \times 8) + (68 \times 8)}{30}$$

$$\bar{x} = \frac{1648}{30} = 54{,}9\hat{3}$$

Consideraciones:

1. Si los datos se agrupan en intervalos, entonces se pierde precisión debido a que se están considerando solo puntos medios.
2. Si los intervalos son muy amplios, entonces hay una mayor agrupación de los datos en estos intervalos, perdiéndose información como una cierta variabilidad en los datos.
3. Al igual que la anterior, esta medida puede estar afectada por valores extremos.
4. Se debe complementar con otras medidas.
5. Su interpretación depende del contexto, y el objetivo del estudio.

Media aritmética ponderada

La media aritmética ponderada conocida también como promedio ponderado, se obtiene multiplicando un peso a cada dato del estudio y se divide entre la suma total de estos pesos.

Se utiliza la siguiente fórmula:

$$\overline{x_p} = \frac{\sum_{i=1}^{n} p_i . x_i}{n}$$

Donde:

$\overline{x_p}$: Media aritmética ponderada.

n : Número total de datos (muestra).

p_i : Es la ponderación para el i-ésimo dato.

x_i : Es el i-ésimo dato.

Ejemplo

El promedio final de las calificaciones de una asignatura universitaria tiene los siguientes criterios y pesos:

Prácticas (P)	30%
Investigación(I)	30%
Evaluación parcial (EP)	20%
Evaluación final (EF)	20%

Por lo tanto, el promedio final tendrá la siguiente fórmula:

$$PF = \frac{(30\% \times P) + (30\% \times I) + (20\% \times EP) + (20\% \times EF)}{100\%}$$

$$PF = \frac{(3 \times P) + (3 \times I) + (2 \times EP) + (2 \times EF)}{10}$$

En este caso, se observa que el peso de las prácticas(P) e investigación (I) es de 3 para cada criterio y el peso de las evaluaciones (EP y EF) es de 2.

Si un estudiante obtuviera las siguientes calificaciones:

P=18 *I=16* *EP=14* *EF=16*

Su promedio sería el siguiente:

$$PF = \frac{(3 \times 18) + (3 \times 16) + (2 \times 14) + (2 \times 16)}{10} = \frac{162}{10} = 16{,}2$$

El promedio ponderado se utiliza ampliamente en educación para el cálculo de las calificaciones finales, pero también se utiliza para hallar el resultado de los cuestionarios aplicados a individuos, en donde se han determinado dimensiones y pesos para la interpretación de una variable.

Ejemplo

El cuestionario Escala de Ansiedad Generalizada GAD-7 (Kroenke, Williams y Löwe, 2006) asigna un peso diferente a cada uno de los siete ítems que la componen, de acuerdo con la intensidad de la sintomatología que se evalúa. Los siete ítems de la GAD-7 y los pesos asignados se observan en la Tabla 25.

Tabla 25

Ítems de la Escala de Ansiedad Generalizada GAD-7 y pesos

Nro	Ítems	Peso
1	Nerviosismo o sensación de estar muy tenso/a	0,5
2	Preocupación excesiva por diferentes cosas	0,5
3	Dificultad para relajarse	0,5
4	Inquietud o impaciencia	0,5
5	Dificultad para concentrarse	0,5
6	Miedo de que algo terrible pueda suceder	0,5
7	Problemas para conciliar o mantener el sueño, o tener sueño inquieto	0,5

Como puede observarse todos los pesos son iguales por lo tanto indica que todos los ítems contribuyen equitativamente en el puntaje final de ansiedad. El rango de los puntajes va desde 0 a 21. Sin embargo, los pesos asignados pueden variar dependiendo de la adaptación o versión de la escala GAD-7.

Existen varios métodos para asignar pesos en un cuestionario de investigación, a saber:

1. Juicio de expertos: Basados en su experiencia y conocimientos los expertos pueden considerar la relevancia, impacto o importancia de cada dimensión y asignar pesos.
2. Métodos estadísticos: Si se tiene bastantes datos, es posible utilizar por ejemplo el Análisis Factorial exploratorio o confirmatorio, a partir de la cual se dan los pesos según los coeficientes factoriales o cargas factoriales que resultaron a partir del análisis.

   Si se tienen datos de estudios anteriores se podrían utilizar otras técnicas como: análisis de regresión lineal, regresión logística o análisis de discriminante, con la cual identificamos las dimensiones de mayor importancia y dar mayor peso.
3. Según modelos teóricos: Se otorgan los pesos según la jerarquía en la que se presentan en el modelo teóricos, las dimensiones que son fundamentales o amplios reciben mayores pesos.

La asignación de pesos puede llevar a resultados más rigurosos, pero requieren experiencia, tiempo y costo adicional.

Ejemplo

Un investigador podría construir un cuestionario de investigación, agrupar los ítems en dimensiones y dar pesos a las dimensiones para obtener un promedio ponderado.

Tabla 26

Ejemplo de plantilla para tabulación de los datos

	Dimensión1(3)				**Dimensión2(4)**					**Dimensión3(4)**					**Dimensión4(2)**				
	It1	*It2*	*It3*	*SD1*	*It4*	*It5*	*It6*	*It7*	*SD2*	*It8*	*It9*	*It10*	*It11*	*SD3*	*It12*	*It13*	*It14*	*SD4*	*EF*
1																			
2																			
3																			
4																			
5																			
6																			

Donde:

Dimensión1(3) : Dimensión 1 con peso 3

SD1 : Suma de los ítems de la dimensión 1.

It1 : Ítem 1.

EF : Evaluación final.

Para cada sujeto su evaluación final sería:

$$EF = \frac{(3 \times SD1) + (4 \times SD2) + (4 \times SD3) + (2 \times SD4)}{13}$$

b) Mediana

Es una medida de tendencia central que divide a una distribución ordenada de datos en dos partes iguales. La mediana es una medida ampliamente empleada en el campo de las ciencias sociales, ya que muchas de las variables examinadas en esta disciplina poseen un nivel de medición ordinal (Landero & González, 2006, p. 173).

Mediana para datos no agrupados

Para calcular la mediana, se siguen los siguientes pasos:

1. Ordenar los datos de forma ascendente o descendente.
2. Si el número de datos es par, la mediana es la semisuma de los dos datos centrales.
3. Si el número de datos es impar, la mediana es el dato que se encuentra en el centro.

Ejemplo

Supongamos que tenemos los datos de las calificaciones de los estudiantes en un examen de matemáticas:

15 12 18 17 12 16 17 13

- Ordenamos los datos:

$$12 \quad 12 \quad 13 \quad \mathbf{15} \quad \mathbf{16} \quad 17 \quad 17 \quad 18$$

- Como el número de datos es par (8 en total) entonces la media es la semisuma de los datos centrales, en este caso son 15 y 16.
- La mediana será semisuma de los valores centrales.

$$Me = \frac{15 + 16}{2} = 15{,}5$$

Interpretación: El 50 % de los estudiantes obtuvieron una calificación encima de 15,5 y el otro restante 50% obtuvieron por debajo de este valor.

Ejemplo

Supongamos que tenemos la talla de 7 estudiantes, las cuales son en metros:

$$1{,}56 \quad 1{,}58 \quad 16{,}00 \quad 1{,}62 \quad 1{,}64 \quad 1{,}68 \quad 1{,}65$$

- Ordenamos los datos: $1{,}56 \quad 1{,}58 \quad 1{,}62 \quad \mathbf{1{,}64} \quad 1{,}65 \quad 1{,}68 \quad 16$
- Como el número de datos es impar (7 en total) entonces la mediana es el dato central.

$$Me = 1{,}64$$

Interpretación: El 50% de los estudiantes miden encima de 1,64m y el otro 50% debajo de este valor.

Consideraciones:

1. La mediana es menos sensible a los datos atípicos, sin embargo, estos datos atípicos deben tenerse en cuenta en la interpretación de la mediana.
2. Si la media aritmética y la mediana son aproximadamente iguales, puede indicar que los datos tienden a ser homogéneos o que la distribución de los datos es simétrica, aunque depende también del contexto y naturaleza de los datos.
3. Si los datos tienen una distribución simétrica, entonces la mediana puede ser considerada con un valor central, pero si tiene una distribución asimétrica (como puede ser la presencia de valores atípicos) la mediana es posible que no pueda ser considerara como valor central.

Ejemplo

Supongamos que tenemos las notas de 11 estudiantes de comunicación:

$$10 \quad 12 \quad 20 \quad 13 \quad 14 \quad 12 \quad 11 \quad 30 \quad 15 \quad 12 \quad 16$$

Como puede observarse en los datos es posible que se halla ingresado por error el valor atípico 30, y también hay un dato 20 de un estudiante sobresaliente, que es considerado atípico.

De los datos se tiene que la media es: $\bar{x} = 15$

Ordenando los datos:

$$10 \quad 11 \quad 12 \quad 12 \quad 12 \quad \mathbf{13} \quad 14 \quad 15 \quad 16 \quad 20 \quad 30$$

La mediana es: $\mathrm{Me} = 13$

El promedio 15 daría la impresión de que los estudiantes tienen un rendimiento regular dado que los valores atípicos elevaron la media aritmética, sin embargo, la mediana indicaría que el 50% de los estudiantes tiene su nota debajo de 13, es decir deficiente. Por lo tanto, no sería coherente la interpretación.

Por lo expuesto, se recomienda que antes de hallar las medidas de tendencia central se analicen los datos atípicos.

Si es debido a un error de digitación, simplemente se corrige, en cambio si es debido a otros factores el retiro de los datos atípicos debe de tomarse con precaución.

La decisión de retirar los datos atípicos o excluir a los sujetos en una investigación que proporcionaron datos atípicos debe basarse en fundamentos científicos sólidos. Además de reducir el tamaño de la muestra traería afectaciones a los resultados y conclusiones.

Así también, la exclusión de los sujetos de una investigación por su condición (como puede ser déficit de atención) podría llevar a una exclusión injusta y discriminatoria, como también excluir a los estudiantes sobresalientes no sería ético ni justo científicamente.

Mediana para datos agrupados

Cuando se analizan una gran cantidad de datos y están agrupados, los pasos para hallar la mediana son los siguientes:

1. En una tabla de frecuencias se debe de tener los intervalos, frecuencia absoluta(f_i) y frecuencia acumulada ($\mathrm{F_i}$) (Ver el apartado de distribución de frecuencias).
2. Identificar el intervalo mediano, el cual es el primer intervalo cuya frecuencia acumulada es mayor o igual a la mitad del número de datos, es decir:

   Identificar el intervalo, tal que $F_i \geq \frac{n}{2}$
3. Calcular la mediana mediante la fórmula:

$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} \times a$$

Donde:

L_i : Es el límite inferior del intervalo mediano.

n : Número total de datos (muestra).

F_{i-1} : Frecuencia acumulada anterior al intervalo mediano.

f_i : Frecuencia absoluta del intervalo mediano.

a : Amplitud del intervalo mediano, que se calcula restando el límite superior menos el límite inferior del intervalo.

Ejemplo

Se requiere determinar la mediada de las edades de un total de 170 estudiantes universitarios, para la cual se ha recopilado la siguiente distribución de datos:

Intervalo de Edad	Frecuencia
[18,20]	25
[21,23]	40
[24,26]	55
[27,29]	30
[30,32]	20

Para calcular la mediana, seguimos los pasos mencionados anteriormente.

1. Obtener la tabla de frecuencias con frecuencia absoluta(fi) y frecuencia acumulada $(\mathrm{F_i})$.

Clase	Intervalo	fi	Fi
1	[18,20]	25	25
2	[21,23]	40	65
3	[24,26]	55	120
4	[27,29]	30	150
5	[30,32]	20	170

2. Identificamos el intervalo mediano, para la cual se tiene en cuenta el primer intervalo que cumpla:

$$F_i \geq \frac{170}{2} = 85$$

Por lo tanto, el intervalo mediano es $[24{,}26]$, dado que $F_3 = 120 \geq 85$

Clase	Intervalo	fi	Fi
1	[18,20]	25	25
2	[21,23]	40	65
3	[24,26]	55	120
4	[27,29]	30	150
5	[30,32]	20	170

3. Aplicamos la fórmula correspondiente:

$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} \times a = 24 + \frac{\frac{170}{2} - 65}{55} \times 2 = 24{,}72$$

Interpretación: La mediana calculada para la edad de los estudiantes universitarios en este estudio sería aproximadamente 24,72 años. Esto significa que el valor central o medio de las edades se encuentra alrededor de los 24 años.

c) Moda

La moda es el dato que aparece con mayor frecuencia en un conjunto de datos, es decir el dato que más se repite.

Sin embargo, en un conjunto de datos puede haber múltiples modas (bimodal o multimodal), debido a que puede que se repitan varios datos con la misma frecuencia.

También, podría existir un conjunto de datos que no tengan moda, debido a que no hay datos repetidos o que los datos tengan la misma frecuencia, por lo tanto, se afirma que no tendrían una moda definitiva.

Ejemplo

En un examen, los estudiantes obtuvieron las siguientes calificaciones:

18 19 15 17 12 **18** 19 12 **18** 14

La moda en este conjunto de datos es 18, porque es la que tiene mayor frecuencia.

Interpretación: La calificación más común o frecuente en el examen es de 18. Esto significa que, un número significativo de estudiantes obtuvo una calificación de 18.

Ejemplo

Supongamos que se realiza un estudio en el que se recopilan las calificaciones de un examen. Se tienen los siguientes datos:

12 15 14 8 17 16 **10** **12** 13 8

19 15 **10** **12** **10** 13 14 15 8 **10**

En este conjunto de datos, hay múltiples modas, es decir, varios valores que se repiten con la misma frecuencia máxima. Al observar los datos, se puede ver que los valores 10 y 12 se repiten con mayor frecuencia, ambos aparecen 4 veces. Por lo tanto, las modas en este conjunto de datos son 10 y 12 (Bimodal).

Interpretación: En el conjunto de datos de calificaciones del examen, se observan múltiples modas, es decir, varios valores que se repiten con la misma frecuencia máxima. Las modas encontradas fueron 10 y 12, lo que indica que estas calificaciones son las más comunes o frecuentes entre los estudiantes evaluados.

Consideraciones:

1. No todos los conjuntos de datos tienen una moda clara, en tales casos se afirma que no hay una moda definitiva.
2. Son utilizadas generalmente para variables categóricas o cualitativas.

Moda para datos agrupados

Cuando los datos están agrupados, es posible hallar la moda mediante los siguientes pasos:

1. En una tabla de frecuencias se debe de tener los intervalos, marca de clase (Xi) y frecuencias absolutas(f_i) (Ver el apartado de distribución de frecuencias).
2. La moda aproximada será la marca de clase con la frecuencia más alta. Al intervalo que contiene la moda aproximada se llama intervalo modal.
3. La moda exacta se halla mediante la fórmula:

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \times a$$

Donde:

L_i : Es el límite inferior del intervalo modal, es decir el intervalo con mayor frecuencia absoluta.

f_i : Frecuencia del intervalo modal.

f_{i-1} : Frecuencia del intervalo anterior al intervalo modal.

f_{i+1} : Frecuencia del intervalo posterior al intervalo modal.

a : Amplitud del intervalo modal, que se calcula restando el límite superior menos el límite inferior del intervalo.

4. Si hay más de un intervalo con la frecuencia más alta, entonces el conjunto de datos puede ser multimodal.

Ejemplo

En un estudio sobre educación primaria, se requiere determinar el número de horas de mayor frecuencia en cuanto a horas de estudio semanales de un grupo de estudiantes de primaria. Se tiene los siguientes datos agrupados por horas de estudio.

Clase	Intervalo	fi
1	[1,3]	5
2	[4,6]	12
3	[7,9]	8
4	[10,12]	6
5	[13,15]	3

Seguimos los pasos para determinar la moda:

1. En la tabla de frecuencias completamos las marcas de clase (Xi). Recordando que la marca de clase es la semisuma del límite superior y el límite inferior del intervalo.

Clase	Intervalo	Xi	fi
1	[1,3]	2	5
2	[4,6]	5	12
3	[7,9]	8	8
4	[10,12]	11	6
5	[13,15]	14	3

2. En la tabla observamos que el intervalo [4,6] tiene la mayor frecuencia con 12 horas. Por lo tanto, la moda aproximada es 5 y el intervalo modal es [4,6].
3. Para hallar la moda exacta aplicamos la fórmula:

$$Mo = L_i + \frac{f_i - f_{i-1}}{(f_i - f_{i-1}) + (f_i - f_{i+1})} \times a$$

$$Mo = 4 + \frac{12 - 5}{(12 - 5) + (12 - 8)} \times 2 = 6{,}091$$

Interpretación: La mayor frecuencia de horas de estudio para este grupo de estudiantes de primaria es de 6,091 horas semanales.

Consideraciones:

1. En muchos casos la moda aproximada es suficiente para interpretar los datos.
2. Para interpretaciones más precisas se hace necesario calcular la moda exacta, por ejemplo, para aquellos casos donde hallan frecuencias similares en intervalos adyacentes.
3. Cuando se van a tomar decisiones educativas, identificación de patrones, tendencia de comportamientos, etc. la moda exacta puede ser relevante para hacer análisis de datos más precisos de los resultados.

3.6.3. Medidas de dispersión

Reciben también el nombre de medidas de variabilidad y su utilidad radica en que miden el grado de separación de los datos respecto a un valor central. Estas medidas serán grandes si las observaciones están distantes de la media y pequeñas si están cerca (Devore, 2008, p.32). Específicamente, son útiles porque:

1. Permiten juzgar la confiabilidad de la medida de tendencia central.

2. Los datos demasiados dispersos tienen un comportamiento especial.

3. Es posible comparar dispersión de diversas muestras.

A continuación, se profundizará en aquellas medidas de dispersión que indican cómo las observaciones se separan de la media aritmética.

a) Rango

Es la medida más simple de dispersión que indica la amplitud total de la distribución de datos. El rango se define como la diferencia entre el valor máximo y el mínimo presente en el conjunto de datos.

$$A = Valor_{max} - Valor_{min}$$

Consideraciones:

1. Se debe tener en cuenta que el rango es sensible a valores atípicos extremos, por lo que en algunos casos no reflejan adecuadamente la dispersión de los datos.

2. El rango es una medida inicial y rápida cuando se requiere ver la dispersión de los datos, pero deben de utilizarse medidas más robustas y como la desviación estándar o rango intercuartílico.

Ejemplo

En un estudio cuasiexperimental donde se evalúa el desempeño académico de dos grupos de estudiantes, Grupo A y Grupo B, en matemáticas. Se tienen los siguientes datos de las calificaciones obtenidas por los estudiantes de cada grupo:

Grupo A:	17	15	18	15	19
Grupo B:	16	14	17	18	13

Para obtener el rango hallamos el máximo y el mínimo dato para cada grupo:

Grupo A: Valor máximo: 19 Valor mínimo: 15

Grupo B: Valor máximo: 18 Valor mínimo: 13

A continuación, calculamos el rango para cada grupo:

Rango del Grupo A = 19-15 = 4

Rango del Grupo B = 18-13 = 5

En este caso el rango del Grupo B es mayor, por lo tanto, indica que hay una mayor dispersión en las calificaciones del grupo B.

b) La varianza

Es una medida de dispersión que cuantifica la variabilidad de los datos con respecto a la media aritmética. Se le puede definir como un promedio de las distancias elevadas al cuadrado, entre cada variable y la media.

La fórmula para hallar la varianza poblacional es la siguiente:

$$\delta^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Donde:

N : Tamaño de la población.

x_i : Es el i- ésimo dato.

μ : Media aritmética poblacional.

Cuando se trabaja con una muestra se utiliza la fórmula siguiente:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

Donde:

n : Tamaño de la muestra.

x_i : Es el i- ésimo dato.

$\bar{x}$: Media aritmética muestral.

Consideraciones:

1. Los datos tienen que ser numéricos y deben estar en la misma escala.
2. La varianza se mide en unidades al cuadrado, lo que puede dificultar su interpretación directa. Para tener una idea más clara, es recomendable complementarla con la desviación estándar, que se encuentra en las mismas unidades originales. La desviación estándar es simplemente la raíz cuadrada de la varianza.
3. La varianza es sensible a los datos atípicos, es decir los datos demasiado grandes, demasiado pequeños afectan a la varianza. Por lo que debe hacerse un análisis para su eliminación o inclusión en el estudio.

Ejemplo

Se tiene una muestra de las calificaciones obtenidas por 6 estudiantes en un examen de matemáticas. Las calificaciones son las siguientes:

10 15 10 15 18 19

Para hallar la varianza, primero calculamos la media aritmética:

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{10 + 15 + 10 + 15 + 18 + 19}{6} = \frac{87}{6} = 14{,}5$$

Aplicamos la fórmula:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1} = \frac{(10-14{,}5)^2 + (15-14{,}5)^2 + (10-14{,}5)^2 + (15-14{,}5)^2 + (18-14{,}5)^2 + (19-14{,}5)^2}{6-1}$$

$$s^2 = 14{,}7$$

Interpretación: La varianza muestral obtenida es 14,7. Por lo tanto, la variabilidad u heterogeneidad de las calificaciones de los estudiantes es de 14,7.

c) La desviación estándar

La desviación estándar es otra medida de dispersión en estadística, complementaria a la varianza. Es una medida que indica cuánto se alejan, en promedio, los valores de un conjunto de datos respecto a la media aritmética. La desviación estándar es una medida de la dispersión y la variabilidad de los datos con respecto a la media.

La desviación estándar se calcula como la raíz cuadrada positiva de la varianza.

La fórmula para la desviación estándar poblacional es la siguiente:

$$\delta = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N}}$$

Donde:

σ : Desviación estándar poblacional.

N : Tamaño de la población.

x_i : Es el i- ésimo dato.

$\bar{x}$: Media aritmética.

La fórmula para la desviación estándar muestral es la siguiente:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

Donde:

s : Desviación estándar muestral.

n : Tamaño de la muestra.

x_i : Es el i- ésimo dato.

$\bar{x}$: Media aritmética.

Consideraciones:

1. Cuanto mayor sea el tamaño de la muestra, más confiable será la estimación de la desviación estándar.
2. La desviación estándar es sensible a valores extremos, la cual significa que los valores atípicos pueden afectarlo.
3. Los datos tienen que estar en la misma escala de medición.
4. La desviación estándar no proporciona información sobre la forma de distribución de los datos.
5. La desviación estándar se utiliza cuando se tiene distribuciones normales o simétricas.

Ejemplo

Se tiene un conjunto de datos que representa el porcentaje de desnutrición en niños de diez escuelas primarias de la ciudad de Lima, durante el año 2019.

Los datos son los siguientes:

5% 8% 10% 12% 6% 4% 9% 7% 11% 13%

Se solicita hallar la desviación estándar muestral e interpretarla.

Para hallar la desviación estándar, primero calculamos la media aritmética:

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = \frac{5+8+10+12+6+4+9+7+11+13}{10} = \frac{85}{10} = 8{,}5$$

Aplicamos la fórmula:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(5-8{,}5)^2 + (8-8{,}5)^2 + (10-8{,}5)^2 + (12-8{,}5)^2 + (6-8{,}5)^2 + (4-8{,}5)^2 + (9-8{,}5)^2 + (7-8{,}5)^2 + (11-8{,}5)^2 + (13-8{,}5)^2}{10-1}}$$

$$s = 2{,}87$$

Interpretación: La desviación estándar sobre la desnutrición de niños en edad escolar es aproximadamente 2.87. Esto nos indica que, en promedio, los porcentajes de desnutrición se alejan alrededor de 2.87 puntos del promedio del 8.5%.

d) El coeficiente de variación (CV)

El coeficiente de variación es una medida estadística que expresa la variabilidad de los datos en relación con la media y la desviación estándar de una población o muestra. Se representa en forma de porcentaje y se distingue por permitir la comparación de la variabilidad entre dos o más conjuntos de datos, incluso cuando están expresados en diferentes unidades.

En el caso de datos no agrupados, la fórmula para hallar el coeficiente de variación es:

Para una población: $\boldsymbol{CV = \frac{\sigma}{\mu} \times 100}$

Para una muestra: $\boldsymbol{CV = \frac{s}{\bar{x}} \times 100}$

Donde:

CV : Coeficiente de variación.

σ : Desviación estándar poblacional.

s : Desviación estándar muestral.

$\bar{x}$: Media aritmética.

La interpretación del CV se realiza de acuerdo con el contexto de los datos. A continuación, se muestra una tabla que resume los porcentajes y su interpretación.

Tabla 27

Interpretación del Coeficiente de Variación

Porcentaje CV	Denominación	Interpretación
0%	Nulo	Datos totalmente homogéneos y consistentes. Todos los valores son iguales.
Menos del 15%	Bajo	Baja dispersión relativa; datos homogéneos y cercanos al valor medio.
Entre 15% y 30%	Moderado	Moderada dispersión relativa; cierta variabilidad, pero aún hay homogeneidad.
Más del 30%	Alto	Alta dispersión relativa; datos heterogéneos y menos consistentes.

Otra clasificación es la siguiente:

Tabla 28

Interpretación de la del Coeficiente de Variación y la distribución de los datos

Porcentaje CV	Denominación	Interpretación	Distribución
$CV < 10\%$	Poca dispersión	Datos con poca dispersión relativa.	Homogénea
$10\% \leq CV < 33\%$	Dispersión aceptable	Datos con dispersión moderada.	Generalmente homogénea
$33\% \leq CV < 50\%$	Dispersión alta	Datos con alta dispersión relativa.	Heterogénea
$CV \geq 50\%$	La dispersión es muy alta	Datos con muy alta dispersión.	Muy heterogénea

Ejemplo

Se tiene dos grupos de estudiantes: Grupo A y Grupo B. Cada grupo ha tomado un curso de comunicación y han obtenido las siguientes notas finales:

Grupo A : 18 17 19 18 16

Grupo B : 17 19 15 12 10

Se requiere comparar la variabilidad de las notas de ambos grupos. Es decir que grupo tiene mayor dispersión de los datos en cuanto a su media aritmética.

Calculamos la media aritmética de ambos grupos:

$$\overline{x_A} = \frac{18 + 17 + 19 + 18 + 16}{5} = \frac{88}{5} = 17{,}6$$

$$\overline{x_B} = \frac{17 + 19 + 15 + 12 + 10}{5} = \frac{73}{5} = 14{,}6$$

Seguidamente calculamos la desviación estándar de los grupos:

$$s_A = \sqrt{\frac{(18 - 17{,}6)^2 + (17 - 17{,}6)^2 + (19 - 17{,}6)^2 + (18 - 17{,}6)^2 + (16 - 17{,}6)^2}{5 - 1}}$$

$s_A = 1{,}14$

$$s_B = \sqrt{\frac{(17-14{,}6)^2 + (19-14{,}6)^2 + (15-14{,}6)^2 + (12-14{,}6)^2 + (10-14{,}6)^2}{5-1}}$$

$s_B = 3{,}65$

Finalmente calculamos el CV de cada grupo:

$$CV_A = \frac{s_A}{\overline{x_A}} \times 100 = \frac{1{,}14}{17{,}6} \times 100 \approx \mathbf{6{,}48}\%$$

$$CV_B = \frac{s_B}{\overline{x_B}} \times 100 = \frac{3{,}65}{14{,}6} \times 100 \approx \mathbf{25}\%$$

Interpretación:

Grupo A: El coeficiente de variación es aproximadamente 6.48%, lo que indica que la dispersión de las notas finales en este grupo es baja. Los estudiantes tienen notas cercanas a la media (17,6), lo que sugiere una homogeneidad en el desempeño académico en el curso de comunicación.

Grupo B: El coeficiente de variación es aproximadamente 25%, es decir los datos tienen una distribución heterogénea, lo que indica que la dispersión de las notas finales en este grupo es más alta que en el Grupo A. Hay una mayor variabilidad en las notas, lo que sugiere que algunos estudiantes tuvieron un rendimiento significativamente diferente en comparación con otros.

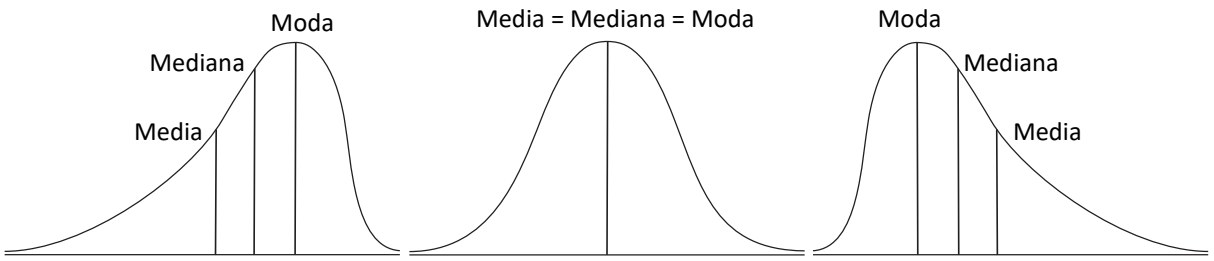
3.6.4. Medidas de forma

Las medidas de forma son un conjunto de estadísticos que describen la forma o apariencia de una distribución de datos. Estas medidas proporcionan información sobre cómo se distribuyen los valores alrededor de la media y cómo se comporta la distribución en términos de apuntamiento o achatamiento (Landero & González, 2006, p. 181).

Relación entre media, mediana y moda

La relación entre la media, la mediana y la moda proporciona información sobre la forma y la simetría de la distribución de los datos. Si la media, la mediana y la moda son iguales, entonces la distribución es simétrica; en cambio, cuando existen diferencias, la distribución de los datos puede ser sesgada hacia la derecha o hacia la izquierda.

En una distribución asimétrica, la media se desplaza hacia una de las colas de la distribución, mientras que la mediana y la moda pueden estar en diferentes posiciones. Si la media es mayor que la mediana y la moda, entonces la distribución tiene un sesgo positivo; por el contrario, si la media es menor que la mediana y la moda, entonces la distribución tiene un sesgo negativo (Martín & Cabera, 2008, p.39).



Distribución asimétrica negativa | Distribución simétrica | Distribución asimétrica positiva

Si una distribución de datos no es simétrica, entonces se considera asimétrica. Sin embargo, tienen fórmulas diferentes debido a que estas dos medidas buscan cuantificar conceptos opuestos en la distribución de datos.

a) Simetría

Una distribución es simétrica cuando los datos se distribuyen equitativamente con igual frecuencia y distancia tanto por debajo como por encima de la media aritmética.

En estas distribuciones, el valor de las medidas de tendencia central - media, moda y mediana - coinciden. La simetría indica que la población es homogénea en relación con la variable en estudio.

La fórmula para calcular la simetría es conocida como coeficiente de simetría de Pearson (coeficiente de asimetría de tercer momento). Si el coeficiente es aproximado a 0 entonces se considera simétrica.

$$Sk = \frac{\bar{x} - Me}{s}$$

Donde:

Sk : Coeficiente de simetría de Pearson.

$\bar{x}$: Media aritmética.

Me : Mediana.

s : Desviación estándar.

b) Asimetría

Se clasifica como asimétrica la distribución donde los datos por debajo de la media son más frecuentes que aquellos por encima de la media, o viceversa. En este caso, se establece que la población es heterogénea para la variable en estudio.

Para calcular la asimetría de una distribución, se utiliza el coeficiente de asimetría de Fisher.

$$g1 = \frac{3 \times (\bar{x} - Me)}{s}$$

Donde:

$g1$: Coeficiente de asimetría de Fisher.

$\bar{x}$: Media aritmética.

Me : Mediana.

s : Desviación estándar.

Según el valor del coeficiente de asimetría g1, se determina:

Si g1 = 0 la distribución es simétrica

Si g1 < 0 la distribución es asimétrica a la izquierda

Si g1 > 0 la distribución es asimétrica a la derecha

Consideraciones:

1. Cuanto mayor sea el tamaño de la muestra, más confiable será la estimación de la simetría y asimetría de la distribución de los datos.
2. Los valores atípicos pueden influir en la simetría y asimetría.
3. La interpretación de la simetría y asimetría depende del contexto. Los resultados deben ser explicados en función al contexto.

Ejemplo

En un estudio, se requiere analizar el rendimiento académico de los estudiantes en un curso de matemáticas en función del promedio ponderado de cada uno de ellos. Se busca interpretar estos resultados considerando el contexto del lugar donde viven, ya que el acceso a recursos educativos, apoyo y oportunidades podría influir en los resultados.

Si los datos tienen una distribución simétrica, se podría afirmar que los estudiantes tienen acceso equitativo a recursos educativos, apoyo y oportunidades.

Por otro lado, si los datos tienen una distribución asimétrica hacia la derecha, esto podría indicar que algunos estudiantes han obtenido puntajes más altos, mientras que la mayoría de los estudiantes se agrupan en puntajes más bajos. En este contexto, podrían existir factores externos que influyan en el rendimiento académico de los estudiantes. Por ejemplo, factores socioeconómicos, acceso limitado a recursos educativos o dificultades personales que afecten el rendimiento podrían explicar la asimetría hacia la derecha. Algunos estudiantes podrían tener más oportunidades de apoyo académico, acceso a tecnología o un ambiente de aprendizaje más favorable, lo que se reflejaría en sus puntajes más altos.

c) Curtosis

La curtosis es una medida estadística que se utiliza para describir la forma o apuntamiento de una distribución de datos. El mayor o menor número de valores de la variable alrededor de la media, dará lugar a una distribución de más o menos apuntamiento.

Al tomar como referencia la distribución normal, se dice que una distribución es más apuntada que la distribución normal (leptocúrtica) o menos apuntada (platicúrtica). Las distribuciones que se asemejan

a la distribución normal se les denomina mesocúrticas (Martín & Cabera, 2008, p.41).

Para calcular la curtosis, se utiliza la siguiente fórmula:

$$g2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n.s^4} - 3$$

Donde:

$g2$: Coeficiente de curtosis.

$\bar{x}$: Media aritmética.

n : Tamaño de la muestra.

s : Desviación estándar.

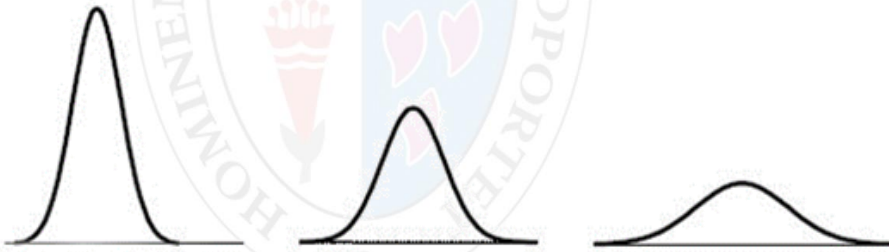
x_i : Es el i- ésimo dato.

Dependiendo entonces del valor del coeficiente de curtosis g2, se determina:

Si g2 = 0 la distribución es mesocúrtica.

Si g2 < 0 la distribución es platicúrtica.

Si g2 > 0 la distribución es Leptocúrtica.



Leptocúrtica Mesocúrtica Platicúrtica

Ejemplo

Se tiene las calificaciones de un grupo de 20 estudiantes en un examen. Se requiere hallar los coeficientes de asimetría y curtosis e interpretarlos.

Los datos son:

12	13	14	15	16	16	17	17	17	18
18	18	19	19	19	19	20	20	20	20

Para calcular la asimetría y la curtosis, primero es necesario obtener la media ($\bar{x}$), la desviación estándar (s) y la mediana (Me) de las calificaciones.

Media $\bar{x} = 17,35$

Desviación estándar s = 2,39

Mediana Me= 18

Cálculo de la asimetría:

$$g1 = \frac{3 \times (\bar{x} - Me)}{s} = \frac{3 \times (17{,}35 - 18)}{2{,}39} = -0{,}82$$

Cálculo de la curtosis:

$$g2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n.s^4} - 3 = \frac{(12 - 17{,}35)^4 + (13 - 17{,}35)^4 + \cdots + (20 - 17{,}35)^4}{20 \times 2{,}39^4} - 3$$

$g2 =$ -0,59

Interpretación: La asimetría es negativa, significa que la distribución tiene una cola larga hacia la izquierda, lo que indica que hay algunos estudiantes que obtuvieron calificaciones muy bajas y la mayoría de los estudiantes se agrupan en el extremo superior del rango.

Si la curtosis es negativa, la distribución es menos apuntada (platicúrtica) que la distribución normal, lo que significa que los valores están menos concentrados alrededor de la media.

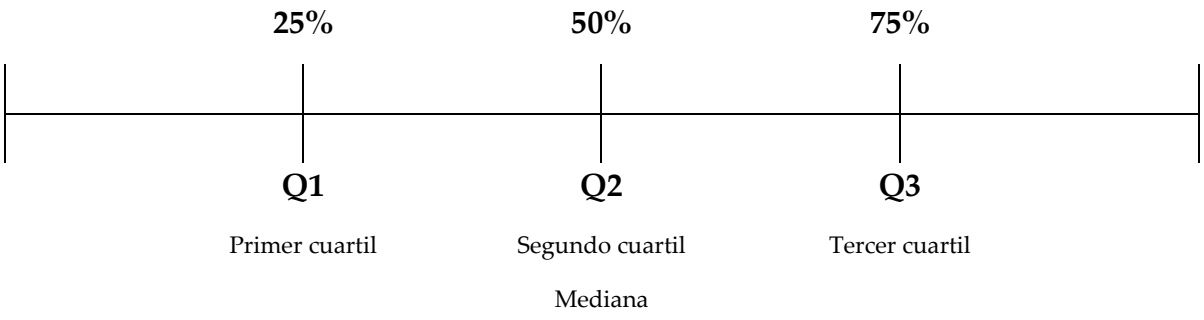
3.6.5. Medidas de posición no central

Las medidas de posición no central (no típicas) se denominan de esa manera porque no representan un valor central, como la media, mediana o moda.

Estas medidas dividen los datos en partes iguales y son útiles para comprender cómo se distribuyen los datos en diferentes porcentajes o fracciones. Para calcular estas medidas, los datos deben ordenarse de manera ascendente o descendente.

a) Cuartiles

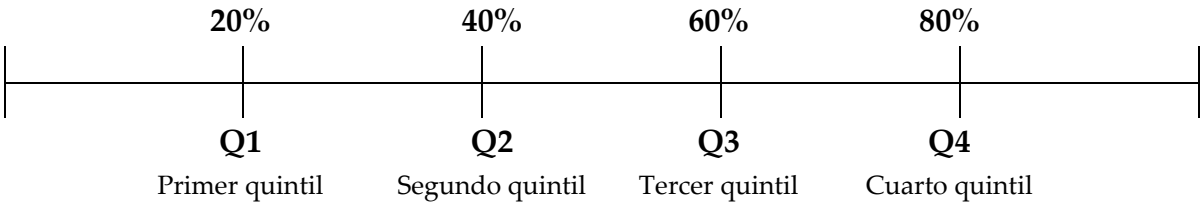
Los cuartiles dividen un conjunto de datos ordenado en cuatro partes iguales. Cada parte contiene el 25% de los datos. Tiene tres valores: Q1, Q2 y Q3.



Las fórmulas para calcular un cuartil son las siguientes:

b) Quintiles

Los quintiles dividen un conjunto de datos ordenado en cinco partes iguales. Cada parte contiene el 20% de los datos.



Las fórmulas para calcular un quintil son las siguientes:

Datos agrupados	Datos no agrupados
Para un conjunto de datos par: $$Q_k = \frac{k.n}{5}$$ Para un conjunto de datos impar: $$Q_k = \frac{k(n+1)}{5}$$ Donde: Q_k : k-ésimo quintil. k : Número de quintil (1,2,3 ó 4). n : Número total de elementos.	$$Q_k = L_i + \left(\frac{\frac{k.n}{5} - F_{i-1}}{f_i}\right) \times a$$ Donde: Q_k : k-ésimo cuartil. k : Número de quintil (1,2,3 ó 4). n : Número total de elementos. L_i : Límite inferior del intervalo Q_k. f_i : Frecuencia absoluta del intervalo Q_k. F_{i-1} : Frecuencia acumulada anterior al intervalo Q_k. a : Amplitud del intervalo Q_k.

c) Deciles

Los deciles dividen un conjunto de datos ordenado en diez partes iguales. Cada parte contiene el 10% de los datos. Al igual que la anterior las fórmulas son:

Datos agrupados	Datos no agrupados
Para un conjunto de datos par: $$D_k = \frac{k.n}{10}$$ Para un conjunto de datos impar: $$D_k = \frac{k(n+1)}{10}$$ Donde: D_k : k-ésimo decil. k : Número de decil (1,2, 3,..,9). n : Número total de elementos.	$$D_k = L_i + \left(\frac{\frac{k.n}{10} - F_{i-1}}{f_i}\right) \times a$$ Donde: D_k : k-ésimo decil. k : Número de decil (1,2,3,…,9). n : Número total de elementos. L_i : Límite inferior del intervalo D_k. f_i : Frecuencia absoluta del intervalo D_k. F_{i-1} : Frecuencia acumulada anterior al intervalo D_k. a : Amplitud del intervalo D_k.

d) Percentiles

Los percentiles dividen un conjunto de datos ordenado en cien partes iguales. Cada parte contiene el 1% de los datos. Las fórmulas son:

Datos agrupados	Datos no agrupados
Para un conjunto de datos par: $P_k = \frac{k.n}{100}$ Para un conjunto de datos impar: $P_k = \frac{k(n+1)}{100}$ Donde: P_k : k-ésimo percentil. k : Número de centil (1,2,3,..,99). n : Número total de elementos.	$P_k = L_i + \left(\frac{\frac{k.n}{100} - F_{i-1}}{f_i}\right) \times a$ Donde: P_k : k-ésimo percentil. k : Número de percentil (1,2,3,..,99). n : Número total de elementos. L_i : Límite inferior del intervalo P_k. f_i : Frecuencia absoluta del intervalo P_k. F_{i-1} : Frecuencia acumulada anterior al intervalo P_k.

Ejemplo

Se tiene las notas de las calificaciones de 30 estudiantes de Historia. Se desea determinar el tercer quintil y su interpretación. Las notas son:

18	14	15	12	12	13	14	15	19	9
10	11	4	17	12	13	14	15	11	12
11	8	8	7	13	14	11	12	16	19

Para hallar el cálculo del quintil, primero ordenamos los datos:

4	7	8	8	9	10	11	11	11	11
12	12	12	12	12	13	13	13	14	14
14	14	15	15	15	16	17	18	19	19

Seguidamente calculamos la posición del tercer quintil:

$$Q_3 = \frac{3.n}{5} = \frac{3 \times 30}{5} = 18$$

Interpretación del quintil 3:

El tercer quintil (Q3) es el valor 18 y el dato que se encuentra en esta posición es 13. Por lo tanto, significa que al menos el 60% de los estudiantes tienen una calificación de 13 o menos en el curso. En otras palabras, el 40% superior de los estudiantes obtuvo una calificación más alta que 13, mientras que el 60% restante obtuvo una calificación igual o menor a 13.

Baremación

El cálculo de las medidas de posición puede ser útil para realizar la baremación o clasificación de las calificaciones según categorías. Es una forma de dividir los datos en grupos basados en su posición relativa dentro del grupo.

Ejemplo

Por ejemplo, para el conjunto de datos anterior, podemos clasificarlos en: excelente, bueno, regular, deficiente y malo.

De las notas:

4	7	8	8	9	10	11	11	11	11
12	12	12	12	12	13	13	13	14	14
14	14	15	15	15	16	17	18	19	19

Podemos calcular directamente los quintiles, o también podemos agrupar los datos puntualmente:

X	fi	Fi	
4	1	1	
7	1	2	
8	2	4	
9	1	5	
10	1	6	Q1
11	4	10	
12	5	15	Q2
13	3	18	Q3
14	4	22	
15	3	25	Q4
16	1	26	
17	1	27	
18	1	28	
19	2	30	

En este caso, para ubicar la posición se tiene en cuenta la frecuencia acumulada (Fi). Las posiciones de los quintiles son:

$Q_1 = \frac{1.n}{5} = \frac{1\times30}{5} = 6$ (Dato en esta posición es 10)

$Q_2 = \frac{2.n}{5} = \frac{2\times30}{5} = 12$ (Dato en esta posición es 12)

$Q_3 = \frac{3.n}{5} = \frac{3\times30}{5} = 18$ (Dato en esta posición es 13)

$Q_4 = \frac{4.n}{5} = \frac{4\times30}{5} = 24$ (Dato en esta posición es 15)

Por lo tanto, el baremo teniendo en cuenta los quintiles sería el siguiente:

Malo : De 0 a 10
Deficiente : De 11 a 12
Regular : De 13 a 15
Excelente : De 16 a 20

Cuestionario

1. *Para determinar el número de intervalos en una tabla de frecuencias con datos agrupados, se utiliza generalmente:*

   a) Prueba t de Student

   b) Regla de Sturges

   c) Análisis de Varianza

   d) Kruskal Wallis

2. *Si en un conjunto de datos la media es mayor que la mediana y la moda, se puede afirmar que la distribución tiene:*

   a) Sesgo negativo

   b) Sesgo positivo

   c) Es simétrica

   d) No tiene sesgo

3. *La desviación estándar es una medida de dispersión que indica:*

   a) La diferencia entre el valor máximo y mínimo

   b) El valor más frecuente

   c) El valor central de los datos

   d) Cuánto varían los valores respecto a la media

4. *Para una investigación cuantitativa, la mejor técnica para recoger datos es:*

   a) Grupos focales

   b) Entrevista no estructurada

   c) Observación participante

   d) Encuesta

5. *En una distribución de frecuencias, la amplitud de clase se calcula:*

   a) Sumando todas las frecuencias

   b) Dividiendo el rango entre el número de intervalos

   c) Restando el límite inferior del límite superior

   d) Multiplicando la frecuencia por la marca de clase

6. *La mediana es una medida de tendencia central que:*

   a) Indica la moda de los datos

   b) Divide los datos por la mitad

   c) Es el promedio de los datos

   d) Indica la dispersión de los datos

7. *Para comparar la dispersión entre dos grupos es recomendable utilizar:*

   a) Rango

   b) Varianza

   c) Desviación estándar

   d) Coeficiente de variación

8. *La observación participante es una técnica cualitativa de recolección de datos donde:*

   a) El investigador interactúa con los participantes

   b) Se aplican cuestionarios estructurados

   c) Se manipulan variables experimentalmente

   d) Se utilizan grupos de enfoque o discusión

9. *Los percentiles son medidas de posición que dividen los datos en:*

   a) 2 partes iguales

   b) 4 partes iguales

   c) 10 partes iguales

   d) 100 partes iguales

10. *La confiabilidad de un instrumento de medición puede evaluarse mediante:*

    a) Varianza

    b) Curtosis

    c) Coeficiente alfa de Cronbach

    d) Análisis factorial

    Alternativas correctas: b, b, d, d, c, b, d, a, d,c

Ejercicios

1. *Se tiene los datos de pesos (kg) de 8 estudiantes:*
   68, 70, 76, 80, 82, 90, 95, 100

   Calcular:

   a) El rango

   b) La media aritmética

   c) La mediana

   d) La moda

   e) Interpretar los resultados

2. *Las calificaciones de 10 estudiantes en una evaluación fueron:*
   15, 12, 10, 18, 20, 5, 16, 14, 19, 17

   Determinar:

   a) La media aritmética

   b) La desviación estándar

   c) El coeficiente de variación

   d) Interpretar los resultados

3. *En una encuesta realizada a 100 estudiantes sobre su asignatura favorita, se obtuvieron los siguientes resultados:*

   Matemáticas: 25 estudiantes

   Comunicación: 20 estudiantes

   Ciencias: 30 estudiantes

   Historia: 15 estudiantes

   Artes: 10 estudiantes

   Construir:

   a) La tabla de frecuencias

   b) La representación gráfica (diagrama de barras)

   c) Interpretar los resultados

4. *Los tiempos (min) para resolver un examen de 10 estudiantes fueron:*
   35, 55, 37, 40, 39, 60, 57, 59, 54, 45

   Calcular:

   a) Los cuartiles

   b) Percentil 70

   c) Interpretar los percentiles calculados

5. *Se tiene los datos del rendimiento académico (promedio ponderado) de un grupo de estudiantes durante dos ciclos académicos:*

   Ciclo 1: 12.5, 10.2, 15.6, 13.8, 11.4, 9.7, 14.3, 10.5

   Ciclo 2: 11.8, 12.5, 14.2, 11.4, 10.9, 8.6, 13.7, 11.2

   Analizar si hay diferencias en la dispersión de las calificaciones entre ambos ciclos calculando:

   a) La desviación estándar de cada ciclo

   b) El coeficiente de variación de cada ciclo

   c) Interpretar los resultados

Referencias

Alba, M. & Ruiz, N. (2005). Muestreo estadístico (7ª Ed.). Oviedo: Septem Ediciones.

Batanero, C., Garfield, J., Ottaviani, M. & Truran, J. (2000). Investigación en Educación Estadística: Algunas Cuestiones Prioritarias. *Statistical Education Research Newsletter1*(2). https://www.ugr.es/~batanero/pages/ARTICULOS/Investiga.pdf

Castro, M. & Lizasoain, L. (2012). Las técnicas de modelización estadística en la investigación educativa: minería de datos, modelos de ecuaciones estructurales y modelos jerárquicos lineales. *Revista Española de Pedagogía, 70*(251), 131–148. http://www.jstor.org/stable/23766443

Cochran, W. (1982). *Técnicas de muestreo*. México: Compañía Editorial Continental S.A.

Devore, L. (2008). *Probabilidad y Estadística para Ingeniería y Ciencias*. Séptima Edición. California: California Polytechnic State University.

Díaz, S. C. (2015). *Metodología de la investigación científica: pautas metodológicas para diseñar y elaborar el proyecto de investigación*. Editorial San Marcos.

Hernández, R., Fernández, C. & Baptista, L. (2014). *Metodología de la investigación* (6ª ed.). México:Mc Graw.

Kerlinger, F. & Lee, H. (2002). *Investigación del comportamiento. Métodos de investigación en ciencias sociales* (4ª ed.). México: McGraw-Hill.

Landero, R. & González, M. (2006). Estadística con SPSS y metodología de la Investigación. México: Trillas.

Martín, Q. & Cabera, T. (2008). *Tratamiento estadístico de datos con SPSS*. Madrid: Thomson.

Martinez, C. (2003). *Estadística y muestreo*. Colombia: Ecoe Ediciones.

Mendenhall, M., Beaver, J. & Beaver, B. (2006). *Introducción a la probabilidad y estadística* (13ª ed.). México: Cengage Learning.

Namakforoosh, M. (2008). *Metodología de la Investigación*. México: Limusa.

Ñaupas, H., Valdivia, M., Palacios, J. , & Romero, H. (2018). *Metodología de la Investigación. Cuantitativa – Cualitativa y Redacción de la Tesis* (5ª Ed.). Bogotá: Ediciones de la U.

Pérez, L. (2013). *IBM SPSS. Estadística Aplicada. Conceptos y Ejercicios Resueltos.* Madrid: Ibergarceta Publicaciones.

Rivera, R. (2017). *Estadística y Probabilidad.* Arequipa: Universidad Autónoma San Francisco.

Sánchez, H. & Reyes, C. (2009). *Metodología y Diseño de la Investigación Científica*. Lima: Editorial Visión Universitaria.

Tamayo, M. (2002). *El proceso de la Investigación Científica*. México: Editorial Limusa.

Vilchez, J. (2011). *Inferencia Estadística para Investigadores*. Lima: Editorial Carvil.