UNIVERSIDAD
NACIONAL
DE COLOMBIA
SEDE PALMIRA
FACULTAD DE INGENIERÍA
Y ADMINISTRACIÓN

Estadística descriptiva para ingeniería ambiental con SPSS

Viviana Vargas Franco

ESTADÍSTICA DESCRIPTIVA PARA INGENIERÍA AMBIENTAL CON SPSS

VIVIANA VARGAS FRANCO

ESTADÍSTICA DESCRIPTIVA PARA INGENIERÍA AMBIENTAL CON SPSS

CALI, JULIO DE 2007

Vargas Franco, Viviana
Estadística descriptiva para ingeniería ambiental con SPSS / Viviana Vargas Franco. -- Editora Viviana Vargas Franco. -- Cali: Impresora Feriva, 2007.
312 p.: il.; 24 cm.
ISBN 978-958-33-9319-3
1. Estadística descriptiva. 2. Análisis de datos. 3. Estadística con ayuda de computador. 4. SPSS para Windows (Programa para computador) - Métodos estadísticos. 5. Medio ambiente - Métodos estadísticos I. Tít.
519.53 cd 21 ed.
A1131724

CEP-Banco de la República-Biblioteca Luis Ángel Arango

vvargasf@palmira.unal.edu.co
Julio de 2007

ISBN 978-958-33-9319-3

Universidad Nacional de Colombia - Sede Palmira
Facultad de Ingeniería y Administración

Foto carátula: Carlos Carrillo

Impreso en los talleres gráficos
de Impresora Feriva S.A.
Calle 18 No. 3-33
PBX: 5249009
www.feriva.com
Cali, Colombia

A

Diana y David, mis hijos

Agradecimientos

La autora expresa sus más sinceros agradecimientos a las diversas personas e instituciones que han colaborado en la elaboración de este libro, entre las que se destacan las siguientes:

Adela Parra Romero. Estadística - Universidad del Valle.
Juan José Castillo. Ingeniero Ambiental - Universidad Nacional de Colombia, Sede Palmira.
Mauricio Rojas Delgado. Estudiante Ingeniería Agrícola - Universidad Nacional de Colombia, Sede Palmira.
Natalia Tamayo González. Ingeniera Ambiental - Universidad Nacional de Colombia, Sede Palmira.
Rafael Domínguez Lasso. Ingeniero Agroindustrial - Universidad Nacional de Colombia, Sede Palmira.
Ricardo Alberto Londoño Saldaña. Ingeniero Agroindustrial - Universidad Nacional de Colombia, Sede Palmira.

Instituciones

Instituto Cinara de la Universidad del Valle. Santiago de Cali
Departamento Administrativo de Gestión del Medio Ambiente de Cali-DAGMA.
Corporación Autónoma Regional del Valle del Cauca-CVC.
Universidad Nacional de Colombia – Sede Palmira

Contenido

Pág.

Introducción 1

Capítulo 1
Fundamentos de los métodos estadísticos

1.1 Modelos estadísticos 4
1.2 Aspectos generales del método científico 5
1.3 Los datos como materia prima de los métodos estadísticos 8
1.4 Aspectos relacionados con la calidad del dato 9
1.5 Conceptos en la aplicación de los métodos estadísticos 11
1.6 Estadística descriptiva vs estadística inferencial 13
1.7 Definición de variables 14
1.7.1 Variables cualitativas o categóricas 14
1.7.2 Variables cuantitativas 15
1.7.3 Otras clasificaciones 17
1.8 Métodos paramétricos y no paramétricos 17
1.9 Métodos estadísticos por tipo de variable 18
1.10 Etapas generales en la construcción de un modelo estadístico 20

Capítulo 2
Medidas descriptivas

2.1 Medidas de tendencia central 23
2.1.1 Media 24
2.1.2 Mediana 36
2.1.3 Moda 38
2.2 Medidas de dispersión 41
2.2.1 Rango 41
2.2.2 Desviación media 42
2.2.3 Varianza 44
2.2.4 Desviación estándar 46
2.2.5 Coeficiente de variación 48

Capítulo 3
Distribución de frecuencias

3.1 Distribución de frecuencias univariadas 53
3.1.1 Distribución de frecuencias univariadas para una variable discreta 54
3.1.2 Distribución de frecuencias univariadas para una variable continua 61
3.2. Distribuciones bidimensionales de frecuencia 89
3.2.1 Distribución bidimensional en variables discretas 89
3.2.2 Distribución bidimensional para variables continuas 93

Capítulo 4
Medidas y gráficas de posición

4.1 Cuartiles 98
4.2 Deciles 103
4.3 Percentiles 106
4.4 Medidas de dispersión para indicadores de posición 110
4.5 Representación gráfica de las medidas de posición 110
4.5.1 Diagramas de cajas y alambres 110
4.5.2 Diagrama de tallos y hojas 120

Capítulo 5
Modelos de regresión

5.1 Modelo de regresión lineal simple 127
5.2 Supuestos del modelo de regresión lineal simple 131
5.3 Diagrama de dispersión 132
5.4 Otros modelos de regresión 136
5.5 Coeficiente de correlación 147
5.6 Coeficiente de determinación 155

Capítulo 6
Planeación estadística en un proyecto de investigación

6.1 Objetivos del proyecto 159
6.2 Descripción del sistema 159
6.3 Codificación del sistema 161
6.4 Definición de variables, sitios y frecuencia de muestreo 162
6.5 Formatos de muestreo 164

6.6 Flujo de información 165
6.7 Sistema de información 167

Capítulo 7
Evaluación de sistemas para tratamiento de agua potable

7.1 Estadísticas descriptivas 171
7.2 Gráficos de medias, mínimos y máximos 173
7.3 Histogramas 180
7.4 Tablas cruzadas 182
7.5 Gráficos de frecuencias acumuladas 185
7.6 Gráficos de tallos y hojas 186
7.7 Percentiles 190
7.8 Diagrama de cajas y alambres 193

Capítulo 8
Calidad de aire

8.1 Gráficos de estadísticas descriptivas 204
8.2 Histogramas 211
8.3 Tablas cruzadas 214
8.4 Gráficas de frecuencias acumuladas 217
8.5 Percentiles 220
8.6 Contaminación del aire en Ciudad de México 224

Capítulo 9
Calidad de agua en una fuente superficial

9.1 Estadísticas descriptivas 237
9.2 Presentación gráfica 239
9.3 Histogramas 245
9.4 Tablas cruzadas 248
9.5 Frecuencias acumuladas 251
9.6 Percentiles 252

Capítulo 10
Instrucciones en SPSS

10.1 Ingresando los datos a SPSS 257
10.2 Importando archivos de Excel 259
10.3 Estadísticas descriptivas 263

10.4 Histograma 268
10.5 Gráfico de frecuencias acumuladas 270
10.6 Gráficos en tres dimensiones 271
10.7 Gráficos de barras en tres dimensiones 273
10.8 Gráfico de tallos y hojas 274
10.9 Gráfico de cajas y alambres 276
10.10 Percentiles 277
10.11 Tablas cruzadas o distribución de frecuencias con dos variables 280

Capítulo 11
Gráficas en Excel

11.1 Gráfico para la media, desviación estándar y el máximo 283
11.2 Gráfico para media, máximo y mínimo 288
11.3 Gráfico de series de tiempo 291

Bibliografía 295

Introducción

Este libro tiene como objetivo proporcionar aspectos conceptuales de la estadística descriptiva con aplicaciones en estudios de la Ingeniería Sanitaria y Ambiental. Está diseñado como texto de consulta en cursos de estadística o para el uso de estudiantes o profesionales que desarrollen un estudio o una investigación donde se requiera aplicar técnicas de estadística descriptiva para el análisis de datos y la toma de decisiones.

En él se exponen aspectos conceptuales de los principales métodos de la estadística descriptiva en lo relacionado con la organización, presentación, estimación y análisis de indicadores estadísticos aplicados en estudios o investigaciones en la Ingeniería Sanitaria y Ambiental. Este trabajo se constituye en un aporte al uso de los métodos estadísticos descriptivos, considerando que se han escrito muchos textos sobre métodos estadísticos pero pocos en el ámbito nacional y regional con aplicaciones a la Ingeniería Sanitaria y Ambiental.

Si bien es cierto que el espectro de desarrollo de la Ingeniería Sanitaria y Ambiental es amplio, se han seleccionado casos sobre evaluación de la calidad de agua en una fuente superficial, comparación de sistemas de tratamiento para agua potable y evaluación de la contaminación del aire en una región específica. Otras aplicaciones pueden seguir la metodología estadística utilizada en los casos estudiados en el presente libro.

Debido al avance de los recursos informáticos, en cuanto a *hardware y software*, los cuales han permitido una utilización intensiva de los métodos estadísticos, en este libro se presentan los procesos o rutinas para la estimación de los indicadores estadísticos en la hoja electrónica Excel (Microsoft Office) y el programa estadístico SPSS (Statistical Package for the Social Sciences) versión 11.5.

La forma como se expone el libro se presenta a continuación: Los primeros cinco capítulos contienen los aspectos conceptuales de la estadística descriptiva. El capítulo 1 presenta los fundamentos de los métodos estadísticos; el capítulo 2, medidas de tendencia central y medidas de dispersión; el capítulo 3, distribuciones univariadas

y bivariadas; el capítulo 4, medidas y gráficas de posición, y el capítulo 5, modelos de regresión lineal. En cada uno de estos capítulos se desarrollan ejemplos que ilustran los procesos estadísticos relacionados con estudios sobre ingeniería sanitaria y ambiental.

Del capítulo 6 al capítulo 9 se presenta la aplicación de los métodos estadísticos descriptivos a casos documentados de la Ingeniería Sanitaria y Ambiental. El capítulo 6 desarrolla la planeación estadística de un proyecto de investigación; el capítulo 7 analiza la evaluación de plantas de tratamiento de agua; el capítulo 8 presenta un estudio de calidad de aire, y el capítulo 9, un estudio sobre la calidad de agua en una fuente superficial.

Los capítulos 10 y 11 presentan las instrucciones para utilizar el software SPSS y Excel, respectivamente.

Las bases de datos de los casos de aplicación fueron recolectadas en diversas investigaciones y estudios desarrollados por varias instituciones, entre las que se destacan: Instituto Cinara de la Universidad del Valle, Corporación Autónoma Regional del Valle del Cauca (CVC), Universidad Nacional de Colombia, sede Palmira y Departamento Administrativo de Gestión del Medio Ambiente de la ciudad Santiago de Cali (DAGMA).

Capítulo 1

Fundamentos de los métodos estadísticos

Los procesos de recolección, organización, presentación, procesamiento, análisis e interpretación de datos numéricos son aspectos fundamentales en el desarrollo de un estudio o una investigación en general, y en particular en los estudios relacionados con la Ingeniería Sanitaria y Ambiental, considerando que generalmente en estos últimos los datos son la herramienta básica para la consolidación de las investigaciones y la toma de decisiones.

Los datos generan información para la toma de decisiones en condiciones de certeza o de incertidumbre. Para la toma de decisiones en condiciones de certeza se utilizan modelos matemáticos determinísticos y la toma de decisiones en condiciones de incertidumbre, medida por la teoría de la probabilidad, se realiza a través de los modelos estadísticos estudiados en la ciencia Estadística.

La estadística es la ciencia que se encarga de la recopilación, organización, presentación, análisis e interpretación de datos numéricos, con el fin de tomar decisiones con criterios de incertidumbre y confiabilidad. Los métodos estadísticos tratan de la presentación gráfica y resumen de datos a través de indicadores, estimación de parámetros poblacionales, pruebas de hipótesis en relación con parámetros poblacionales, determinación de la exactitud de las estimaciones, estudio de la variación, estudio de correlación y el diseño de experimentos, de forma univariada y multivariada, entre otros.

1.1 Modelos estadísticos

Un modelo estadístico es una representación simplificada, formal y abstracta de un fenómeno de la naturaleza o de un sistema, éste puede representar la estructura, el comportamiento o el funcionamiento de una parte de interés o el conjunto del fenómeno o del sistema. La representación se hace a través de símbolos matemáticos que corresponden a relaciones entre parámetros y variables.

Un modelo se considera adecuado si efectiva y objetivamente representa la realidad que pretende estudiar y conocer. El elemento básico para juzgar un modelo es su confrontación con la realidad, esto implica que para juzgar el modelo debe hacerse una observación empírica del objeto de estudio y con base en ella juzgar la bondad del modelo (Quiroga).

La construcción y aplicación de un modelo estadístico se define a través de los elementos básicos de la teoría estadística: datos, aleatoriedad, variabilidad, teoría de probabilidad, selección muestral, estimación de parámetros y docimasia de hipótesis, entre otros.

No existe un modelo perfecto, pero se debe preferir un modelo simple, donde no se pierda información, considerando los componentes sistémicos y aleatorios del fenómeno.

Los métodos estadísticos proporcionan criterios y modelos matemáticos para realizar los procesos de recolección, procesamiento y análisis de datos requeridos en estudios donde una componente fundamental son los datos, con características de variabilidad y aleatoriedad. La aplicación de los métodos estadísticos permite generar conclusiones objetivas con criterios de confiabilidad y riesgo en la toma de decisiones. Los métodos estadísticos son un medio y no un fin y como tal deben ser utilizados; los resultados estadísticos deben ser contrastados con análisis de las teorías y modelos conceptuales o modelos matemáticos que permitan suministrar avances significativos en las diferentes áreas de su aplicación.

La estadística como ciencia independiente es un desarrollo del siglo XX. Sir Ronald Aymer Fischer (1890-1962) fue el principal representante, el transformador de ideas que cohesionó y estableció los fundamentos teóricos de la inferencia estadística como método de razonamiento inductivo que da un nuevo sentido al procesamiento de datos e intenta medir su grado de incertidumbre. Sus resultados le dieron a la estadística estatus de disciplina científica, reafirmado por los innumerables campos de aplicación de sus metodologías (Yáñez, 2001).

El avance del análisis estadístico en los últimos años ha sido rápido y su uso se constituye en una valiosa herramienta para la toma de decisiones. La actualización

permanente de los recursos informáticos en cuanto a *hardware* y *software* ha permitido una utilización intensiva de los métodos estadísticos.

Existen dos fases en el procesamiento estadístico de un conjunto de datos: una parte relacionada con la estadística descriptiva o estadística deductiva y otra relacionada con la estadística inferencial o estadística inductiva. La estadística descriptiva consiste en resumir el conjunto de datos de una investigación en indicadores estadísticos que permiten estimar el grado de centralidad, dispersión, posición y distribución de frecuencias. El análisis descriptivo es una etapa importante en la comprensión de un fenómeno, pues permite estudiar las tendencias generales del conjunto de datos.

Generalmente después del proceso descriptivo se hace la estimación de la inferencia estadística o estadística inferencial. Esta consiste, a partir de los resultados estadísticos de una muestra representativa de una población, en realizar generalizaciones o inducciones a parámetros de la población, considerando criterios de riesgo y confiabilidad, estimados a partir de la teoría de la probabilidad, tal como se observa en la Figura 1.1.

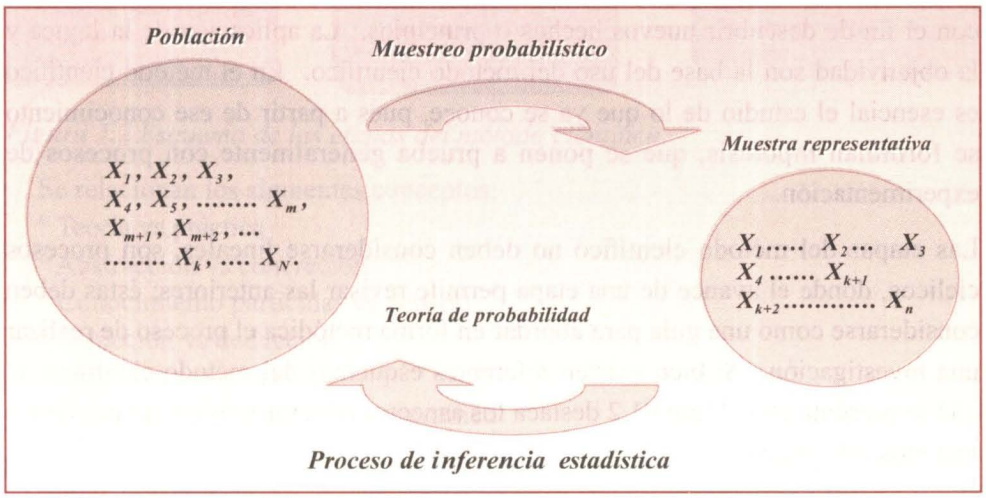


Figura 1.1 *Esquema del proceso de inferencia estadística*

Los métodos estadísticos están relacionados con el método científico en las etapas de recolección, organización, presentación y análisis de datos, para la deducción de conclusiones y la toma de decisiones razonables de acuerdo con los análisis estadísticos.

1.2 Aspectos generales del método científico

El conocimiento científico es aquel que se realiza mediante la aplicación del método científico; permite el uso de la razón, la lógica, la objetividad y tiende a evitar que

el conocimiento surja de la pasión o la emoción. Por medio de la investigación científica el hombre ha alcanzado una reconstrucción conceptual del mundo que es cada vez más amplia, profunda y exacta (Bunge). El conocimiento científico puede caracterizarse como conocimiento racional, sistemático, exacto, verificable y por consiguiente falible.

El método científico es una guía para desarrollar una investigación o estudio con resultados de carácter científico. La palabra método viene del griego: "meta", que significa "con" y "odos" que significa "camino", es decir, es la forma de proceder encaminada hacia un objetivo donde lo que se va desarrollando guarda orden y coherencia. El método científico puede concebirse como un modelo general de acercamiento a la realidad; es una pauta o matriz abstracta y amplia, dentro de la cual están los procedimientos y técnicas específicas que se emplean en una investigación.

Una investigación puede definirse como el estudio sistemático de un sujeto u objeto con el fin de descubrir nuevos hechos o principios. La aplicación de la lógica y la objetividad son la base del uso del método científico. En el método científico es esencial el estudio de lo que ya se conoce, pues a partir de ese conocimiento se formulan hipótesis, que se ponen a prueba generalmente con procesos de experimentación.

Las etapas del método científico no deben considerarse lineales, son procesos cíclicos, donde el avance de una etapa permite revisar las anteriores; éstas deben considerarse como una guía para abordar en forma metódica el proceso de realizar una investigación. Si bien existen diferentes esquemas del método científico, el que se presenta en la Figura 1.2 destaca los aspectos relacionados con el uso de los métodos estadísticos.

Entre las características básicas del proceso de investigación se destacan los siguientes aspectos:

- Un producto de la investigación: nuevo conocimiento
- Es un proceso sistemáticamente organizado
- Es un proceso en espiral del conocimiento
- Genera saltos cualitativos del conocimiento por acumulación de pequeños cambios cuantitativos
- Permite replicabilidad de los resultados
- Operan la lógica y la objetividad

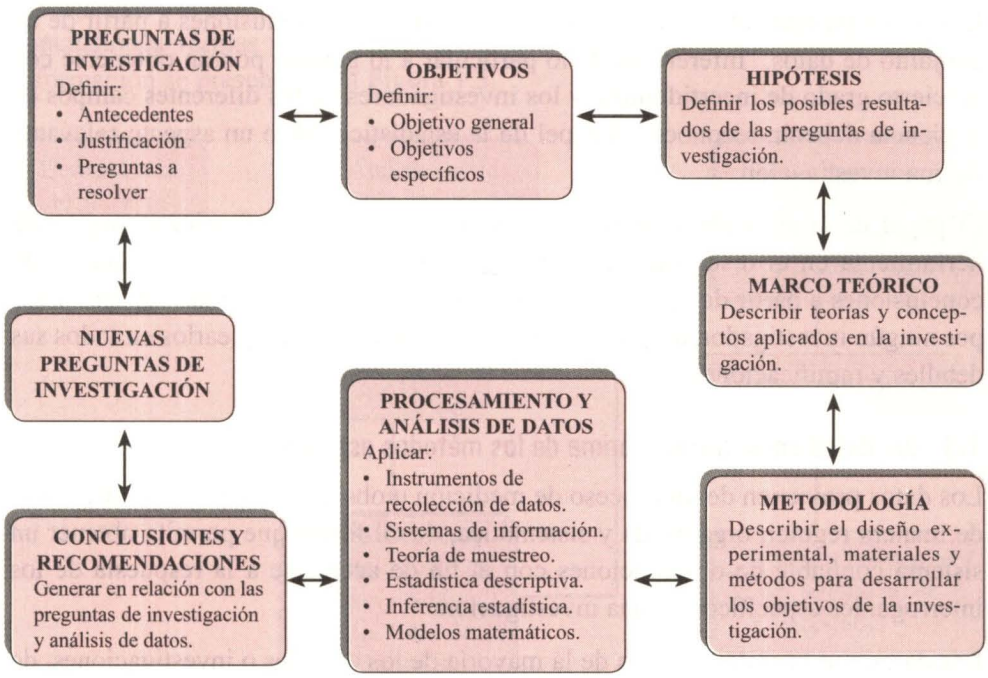


Figura 1.2 Esquema de las etapas del método científico.

- Se relacionan los siguientes conceptos:
  - Teoría vs práctica
  - Abstracción vs concreción
  - Conocimiento particular vs general
  - Inducción vs deducción
  - Análisis vs síntesis
  - Conocimiento heurístico vs científico

La estadística es un conjunto de herramientas útiles en la investigación en las fases de planeación, análisis e interpretación de los resultados de una investigación, apoyando el desarrollo del método científico en la descripción y la predicción. Por la naturaleza de los métodos estadísticos los resultados son parciales y fragmentados más que completos y definitivos.

En una investigación debe haber concordancia lógica entre los objetivos, el diseño de la investigación, el análisis de los resultados y las conclusiones; generalmente los conceptos y métodos estadísticos juegan un papel importante únicamente en el análisis e interpretación de datos, lo cual conduce con frecuencia a investigaciones en las que no hay una buena concordancia entre los objetivos, el diseño de la investigación y las conclusiones.

Los procesos estadísticos proporcionan información y conclusiones a partir de un conjunto de datos. Inferencias de lo particular a lo general podrán obtenerse con un cierto grado de incertidumbre y los investigadores en los diferentes campos de la ciencia deberán reconocer el papel de la estadística como un aspecto relevante de una investigación.

El papel de la estadística en la investigación es, entonces, funcionar como una herramienta en el diseño de ésta, en el análisis de datos y en la extracción de conclusiones a partir de ellos. Los métodos estadísticos no deberían ser ignorados por ningún investigador, aun cuando no tengan ocasión de emplearlos en todos sus detalles y ramificaciones.

1.3 Los datos como materia prima de los métodos estadísticos

Los datos provienen de un proceso de medición u observación que debe realizarse de manera regular, organizada y sistemática, de tal forma que permita obtener un sistema confiable de observaciones con el fin de acercarse a la respuesta de los interrogantes específicos de una investigación.

Los datos son la materia prima de la mayoría de los estudios o investigaciones, de ellos depende en buena medida el aprovechamiento de los métodos estadísticos para su posterior análisis. De nada vale acumular datos sobre una investigación si no existen criterios para su organización y procesamiento estadístico.

En un estudio donde los resultados generan un conjunto de datos, es casi indispensable resumirlos en indicadores de carácter estadístico que faciliten su presentación, interpretación y análisis. Un conjunto de datos no genera información por sí mismo, es a través del procesamiento matemático o estadístico *significativo* donde se pueden encontrar indicadores y medidas de tendencia que generen información:

$$\text{Datos} \neq \text{Información}$$

No se puede caer en la frase *"ricos en datos, pobres en información"*. En general los textos de métodos estadísticos no mencionan o suponen que el proceso de recolección y calidad del dato es un aspecto conocido por los investigadores o profesionales que realizan estudios, sin embargo es una de las fases de la experimentación que generalmente no se planea con el cuidado que se requiere.

La recolección de datos y su posterior análisis no son la finalidad principal de una investigación o un estudio, es necesario realizar procesos de modelación matemática y estadística que permitan generar información sobre las preguntas de la investigación. La información que se genere del proceso de análisis debe

incorporarse a teorías y marcos conceptuales, de tal forma que se consigan conclusiones válidas y objetivas. Un proceso que permite transformar datos en información se presenta en la Figura 1.3.

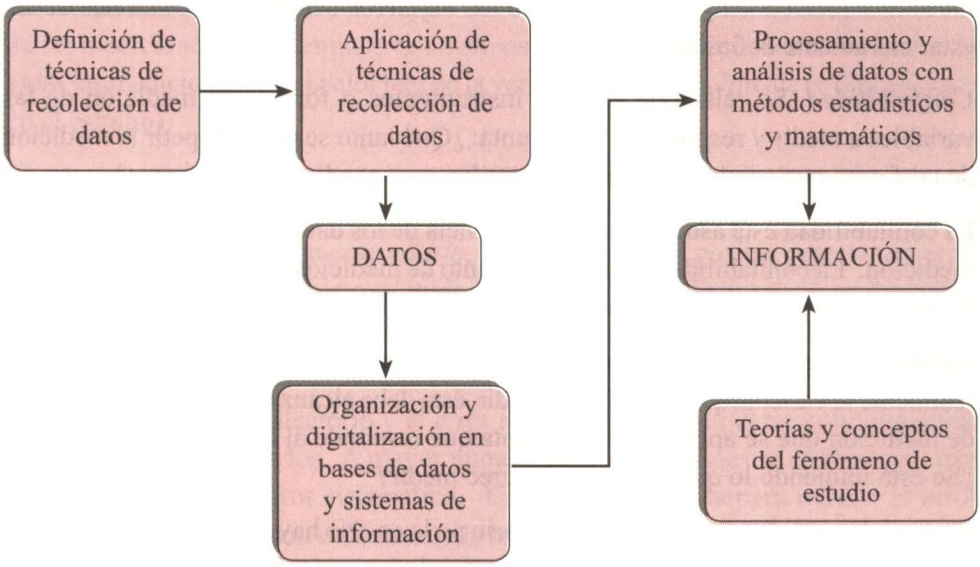


Figura 1.3 *Un esquema metodológico para convertir datos en información.*

1.4 Aspectos relacionados con la calidad del dato

La calidad de los datos es uno de los aspectos importantes que se deben planear antes de las etapas de recolección y aplicación de los métodos estadísticos, pues los procesos estadísticos generalmente no verifican ni corrigen deficiencias en la calidad de los datos. Varios componentes se deben estudiar sobre la calidad de un conjunto de datos: confiabilidad, validez y representatividad, entre otros.

Representatividad. Está relacionada con el tamaño de la muestra y la forma como se seleccionan los individuos u observaciones a ser analizados y responde a la pregunta: ¿Los resultados de la muestra pueden aplicarse o generalizarse a la población objeto de estudio?

El tamaño de la muestra depende del grado de variabilidad del fenómeno a estudiar, el nivel de precisión deseado y el nivel de confiabilidad requerido, así como de los costos de personal, reactivos y equipos, entre otros.

La forma de selección del número de muestras, es decir, el tipo de muestreo a utilizar, puede ser probabilístico (cada elemento tiene una probabilidad conocida de ser seleccionado en la muestra), o no probabilístico (no todos los elementos tienen

probabilidad de ser incluidos en la muestra). Se deben seleccionar los individuos sin sesgo y que haya participación de los diversos elementos del fenómeno a estudiar.

La representatividad está ligada a la definición de la población objetivo y a la muestra seleccionada y estas a su vez a los objetivos del estudio, los cuales deben estar claramente definidos

Confiabilidad. Se relaciona con los instrumentos o formas de medición de las variables a medir y responde a la pregunta: ¿Qué tanto se puede repetir la medición de tal forma que produzca resultados similares en condiciones similares?

La confiabilidad está asociada a la consistencia de los datos con los instrumentos de medición. La confiabilidad de un instrumento de medición se refiere al grado en que su aplicación, repetida al mismo sujeto u objeto, produce resultados iguales.

Validez. Se refiere al grado en que un instrumento, concepto o indicador mide realmente la variable que se pretende medir, ésta debe alcanzarse en todo instrumento de medición que se aplica. Una pregunta que responde al concepto de validez es: ¿Se está midiendo lo que realmente se cree medir?

Si es así, la medida es válida, de lo contrario no lo es. No hay medición perfecta, pero es necesario que haya una representación fiel de las variables a observar, mediante el instrumento de medición.

Un instrumento de medición puede ser confiable, pero no necesariamente válido. Por eso es conveniente que los resultados de una investigación demuestren ser ***confiables y válidos***.

Factores que afectan la confiabilidad y la validez. Algunos factores que afectan la confiabilidad y la validez de un conjunto de datos:

- Improvisación
- Instrumentos de medición utilizados en diferentes contextos y sin adaptación
- Falta de validación de los instrumentos de medición
- Instrumentos inadecuados para las variables seleccionadas
- Condiciones inadecuadas en las que se aplica el instrumento
- Capacitación deficiente al personal de apoyo
- Instrucciones deficientes

Fuentes de error. Algunas fuentes de error en las mediciones son: error aleatorio, error sistemático, normalidad y anormalidad.

Error aleatorio. Es el producido por el sistema de mediciones, es un error constante que está presente en cada una de las mediciones que se efectúan. Su valor no afecta

al valor real ni al valor promedio del conjunto de datos. En términos estadísticos es igual a la diferencia entre una medición y la media de todas las mediciones.

Error sistemático. Es el producido por la medición de cada una de las componentes del sistema, no es constante, es el error de redondeo que se lleva a cabo en cada una de las mediciones. En términos estadísticos es igual a la diferencia de la media de todas las mediciones con el valor real de la variable (que normalmente es desconocido en el estudio).

El error sistemático normalmente permanecerá cuando se repita la medición. De ahí que sea difícil detectarlo en un estudio. Éste también indica que el instrumento de medida no es completamente *válido*. Algunas veces es posible detectar un error sistemático si el mismo objeto se mide con dos métodos distintos. Si se descubre, se elimina por *corrección* de mediciones (por ejemplo, por *normalización* de las mismas) o por *calibración de la escala* del instrumento de medida.

En un estudio el error aleatorio y el error sistemático pueden darse conjuntamente y es importante detectarlos. A mayor número de observaciones se controla el error aleatorio, pero no el error sistemático. Entre las estrategias para reducir el error sistemático se encuentran: calibración de los instrumentos y realización de medidas ocultas. En general, los fabricantes de instrumentos de medición suelen garantizar que el error total (aleatorio + sistemático) de su equipo es inferior a cierto límite, siempre y cuando el instrumento sea usado con las especificaciones definidas.

Normalidad y anormalidad. Se dice que los datos son normales si el patrón sigue la forma de una curva normal o en forma de campana, en caso contrario se habla de datos con anormalidad. En el caso de datos normales, se pueden estimar intervalos de confianza alrededor de indicadores estadísticos de interés; en caso de anormalidad se pueden estimar niveles percentiles, que pueden estar alrededor del 95% y 97,5%, que depende del estudio que se esté realizando.

1.5 Conceptos en la aplicación de los métodos estadísticos

A continuación se describen algunos conceptos fundamentales para la aplicación de los métodos estadísticos.

Población. Se define de acuerdo con los objetivos del estudio, y está determinada por condiciones ambientales, de tiempo y espacio, entre otras. La población se define como la totalidad de los elementos o individuos que tienen características similares y sobre los cuales se desean realizar inferencias o generalizaciones. Se deben definir claramente quiénes y qué características deben tener los objetos o sujetos del estudio, es decir, la población.

Muestra. Es una parte seleccionada de la población objeto de estudio y sobre la cual se van a realizar las mediciones. La muestra debe ser representativa con el fin de dar confiabilidad a las inferencias o generalizaciones a la población. La muestra puede ser seleccionada con criterios probabilísticos o criterios no probabilísticos. En general, para el uso de la inferencia estadística se requiere una muestra probabilística. Para la selección de una muestra probabilística se deben considerar los siguientes aspectos:

- Definir en forma precisa la población
- Considerar el marco muestral (fuente de extracción de unidades)
- Seleccionar el tipo de muestreo (depende de la población, puede ser aleatorio, estratificado, por conglomerados, sistemático, entre otros)
- Estimar el tamaño de muestra (con criterios estadísticos, definir: nivel de confiabilidad deseado, nivel de precisión en la estimación y nivel de variabilidad de las variables de interés)
- Definir un procedimiento de muestreo (cómo seleccionar los elementos de la población)
- Seleccionar la muestra

Una población puede ser finita o infinita, pero la muestra siempre será finita. La muestra puede ser de interés inmediato, pero importa principalmente describir la población de la cual se tomó. La escogencia de la muestra debe reflejar estrechamente las posibles características de la población.

Parámetro. Se refiere a un indicador estadístico que es calculado a través de las observaciones o datos de la población. El valor del parámetro es constante y generalmente desconocido, el cual se estima a través de los datos de la muestra.

Estadístico o estadígrafo. Se refiere a un indicador estadístico que es calculado de las observaciones o datos de la muestra. El valor del estadístico es conocido y varía con la muestra. En general estos indicadores son los que se pretenden generalizar a la población a través del proceso de inferencia estadística. Los más utilizados son: media aritmética, desviación estándar, momentos, coeficientes de correlación, entre otros. La media muestral es un estadístico que permite estimar la media poblacional, que es un parámetro.

Estimación. Es el proceso estadístico mediante el cual se infieren o generalizan los datos de un estadístico a un parámetro, utilizando la teoría de la probabilidad. Es decir, se generalizan los valores de los resultados muestrales a valores poblacionales.

Distribución de probabilidades. Es la forma de agrupación de los datos. Existe un gran número de distribuciones asociadas a la forma de agrupación y al tipo de variable de los datos. Algunos ejemplos de distribuciones son: normal, Poisson, geométrica,

hipergeométrica, entre otras. Si los datos se aproximan a una de estas distribuciones, su modelo teórico se puede utilizar para propósitos de toma de decisiones.

1.6 Estadística descriptiva vs estadística inferencial

Los métodos estadísticos se pueden clasificar en dos fases: estadística descriptiva y estadística inferencial. No es que existan dos estadísticas, las primeras son técnicas descriptivas y las segundas inferenciales, estas últimas se apoyan en los resultados de las técnicas descriptivas y permiten generalizar de una muestra a una población, utilizando la teoría de la probabilidad, tal como se observa en la Figura 1.4.

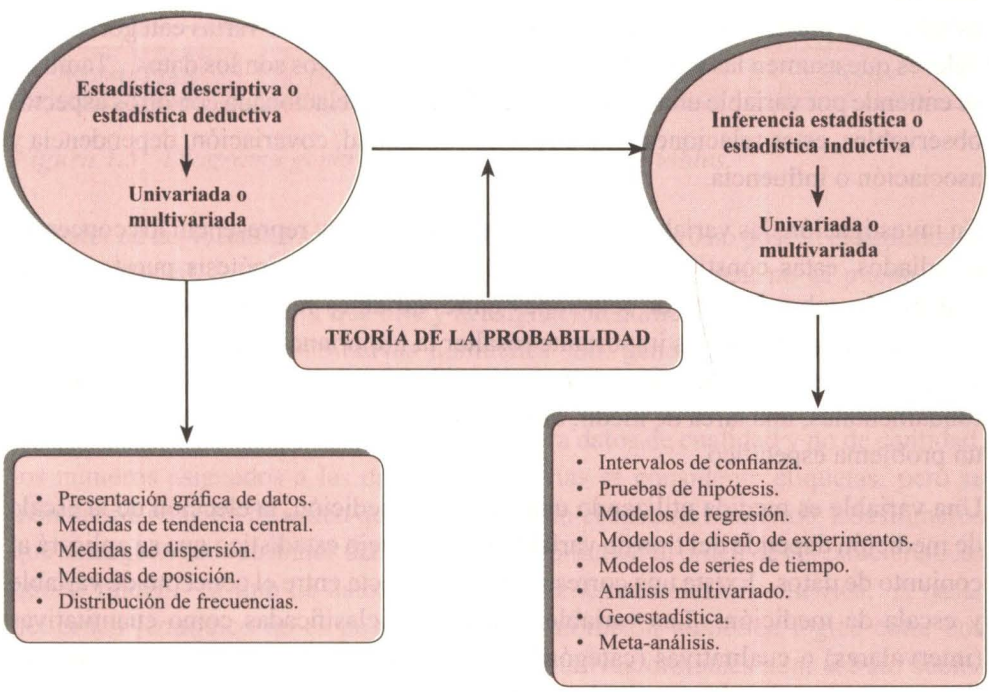


Figura 1.4 *Esquema de la relación entre estadística descriptiva e inferencial y sus principales procesos.*

La estadística descriptiva, como su nombre lo indica, permite describir *significativamente* un conjunto de datos mediante la presentación, organización y resumen en indicadores estadísticos. Las técnicas con las cuales se resume el conjunto de datos son: las medidas de tendencia central, de dispersión, de posición y el análisis de distribución de frecuencias; estos métodos pueden ser de carácter univariado o multivariado, de acuerdo con los requerimientos del estudio. Generalmente después del análisis descriptivo se desarrolla el análisis inferencial.

El análisis estadístico inferencial permite hacer un proceso inductivo para inferir sobre una medida estadística, generalmente la media aritmética, a la población con base en observaciones de una muestra seleccionada en el estudio. Este tipo de análisis utiliza la teoría de la probabilidad para cuantificar el nivel de confianza de las conclusiones obtenidas (Behar, 1996). Algunos métodos para realizar el proceso de inferencia están conformados por modelos de diseño de experimentos, modelos de regresión, intervalos de confianza y pruebas de hipótesis.

1.7 Definición de variables

Una variable es una característica observable o medible en un objeto o sujeto de estudio, que puede adoptar diferentes valores o expresarse en varias categorías. Los valores que asumen las variables en cada uno de los sujetos son los datos. También se entiende por variable una característica observable relacionada con otros aspectos observables, estas relaciones pueden ser de causalidad, covariación, dependencia y asociación o influencia.

En investigación, las variables son los aspectos a medir y representan los conceptos estudiados, estas constituyen un elemento básico de las hipótesis puesto que se construyen sobre la base de relaciones entre variables referentes a determinadas unidades de medición. Es importante resaltar la importancia de las variables como elementos básicos del método científico, ya que la investigación es, en ciertos aspectos fundamentales, una tarea de medir, analizar y concluir sobre variables de interés en un problema específico.

Una variable es medida utilizando una escala de medición, la elección de la escala de medición depende del tipo de variable y del manejo estadístico que se aplicará al conjunto de datos. Existe una correspondencia directa entre el concepto de variable y escala de medición. Las variables pueden ser clasificadas como cuantitativas (intervalares) o cualitativas (categóricas), dependiendo si los valores presentados tienen o no un orden de magnitud natural (cuantitativas), o simplemente un atributo no sometido a cuantificación (cualitativa). Un diagrama donde se presentan la clasificación de los principales tipos de variables y la relación con la escala de medición se presenta en la Figura 1.5.

1.7.1 *Variables cualitativas o categóricas*

Son aquellas cuyos valores tienen un carácter de cualidad no susceptible, naturalmente de variación numérica. Se clasifican en ordinales y nominales.

Nominal, se denomina a la variable cualitativa que genera valores de cualidad, sin tener ellos ningún orden o jerarquía. Los números asignados a las diversas categorías

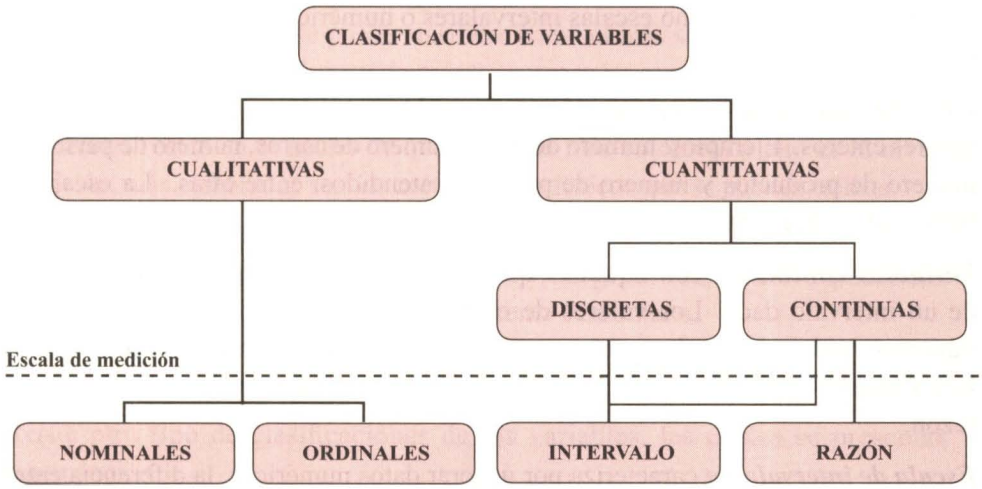


Figura 1.5 *Diagrama general de clasificación de variables.*

del valor de las variables se consideran como etiquetas, pero no poseen el significado numérico usual, los valores tienen una naturaleza no-métrica, no se puede decir que una categoría es mejor que otra y la asignación numérica es arbitraria. Algunos ejemplos de variables cualitativas nominales son: género, raza, profesión, credo religioso, color de ojos, partidos políticos y estado civil.

Ordinal, se denomina a una variable que genera datos de cualidad y no de cantidad, los números asignados a las diversas categorías se consideran etiquetas, pero se genera una relación de orden que se preserva en el sistema numérico. Los números que se asignan a los atributos deben respetar o conservar el orden de las características que se miden. El tipo de datos que resulta tiene naturaleza no-métrica. A pesar del orden jerárquico no es posible obtener valoración numérica lógica entre dos valores. Algunos ejemplos de variables cualitativas ordinales son: estrato socio-económico, nivel de satisfacción (acuerdo-total, acuerdo-parcial, desacuerdo-parcial y desacuerdo-total) y calificación (E-excelente, S-satisfactorio, A-aceptable, D-deficiente, I-insuficiente).

Las funciones de distribución asociadas a una variable discreta son: uniforme discreta, Bernoulli, binomial, hypergeométrica, Poisson, geométrica, binomial negativa, Beta-binomial y logarítmica.

1.7.2 *Variables cuantitativas*

Son aquellas donde las características o propiedades pueden presentarse en diversos grados o intensidad y poseen un carácter numérico. Las escalas cuantitativas son

reconocidas también como escalas intervalares o numéricas. Estas se clasifican en continuas y discretas.

Variables discretas, los valores de estas variables son enumerables y toman sólo valores enteros. Ejemplos: número de hijos, número de carros, número de personas, número de productos y número de pacientes atendidos, entre otras. La escala de medición es de intervalo.

Variables continuas, son aquellas que pueden tomar infinitos valores dentro de un intervalo dado. Los valores de estas variables están relacionados con los números reales. Ejemplos: peso, estatura, salario y temperatura, entre otros. Las variables continuas presentan dos escalas de medición: de intervalo y de razón.

Escala de intervalo, se caracteriza por generar datos numéricos, la diferencia entre dos medidas es significativa. En esta escala tienen sentido la suma y la resta de valores, pero no existe un cero absoluto ni las distancias entre los valores generan noción de equivalencia. En esta escala no tiene sentido el concepto de división. Algunos ejemplos: puntuaciones en una prueba de razonamiento (IQ) y temperatura del agua.

Por ejemplo, en esta escala es posible decir el mejor desempeño (IQ) que tuvo un estudiante en una prueba frente a otro; un niño con un IQ de 150 es mejor que un niño que obtuvo 75, pero no se puede decir que el primero tiene el doble de inteligencia que el segundo. En esta escala no hay un cero verdadero. El cero en temperatura Fahrenheit es una temperatura seleccionada al azar. El cero en centígrados corresponde a otra temperatura muy diferente. El resultado es que, a pesar de que 100°C es el doble de 50°C, en una temperatura de 100°C no hace el doble de calor que en una de 50°C.

Escala de razón, es el nivel más complejo en las escalas, tiene un origen natural, el cero absoluto, y al igual que en la escala de intervalo se generan medidas numéricas y las diferencias son valores significativos. La resta y la división entre dos valores de esta escala tienen significado. Ejemplos: peso, estatura y edad, entre otros. Aquí tiene sentido hablar de que una persona pesa el doble de otra, o que alguien tiene el doble de años que otra persona.

En general las medidas dan origen a datos continuos, mientras que las enumeraciones o conteos originan datos discretos. Es siempre posible pasar de una escala a otra menos exigente. Ejemplo: los estudiantes pueden medirse en metros (variable continua-razón), pero pueden también ordenarse de mayor a menor, convirtiéndose en una variable ordinal.

En nivel de complejidad se puede clasificar como el más simple, la escala nominal, seguido de la escala ordinal, posteriormente aparecen las escalas de intervalo y la escala de más alto nivel de complejidad es la de razón. La importancia de esta clasificación por niveles reside en el hecho de que mientras más complejo o alto es el nivel de medición, más elaborados son los métodos estadísticos que se pueden utilizar.

Las funciones de distribución asociadas a una variable continua son: uniforme, normal, exponencial, gamma, beta, Cauchy, Log normal, doble exponencial o Laplace, Weibull, Logística, Gumbel y sistema Personiano.

1.7.3 *Otras clasificaciones*

Existe otro tipo de clasificaciones de las variables, las cuales se presentan a continuación:

Variables dependientes (Y): Reciben este nombre las variables a explicar, o sea, el objeto de una investigación que se trata de explicar en función de otros elementos.

Variables independientes (X): Son las variables explicativas, es decir, los factores o elementos susceptibles de explicar las variables dependientes (*Y*); en una investigación de tipo experimental son las variables que se manipulan.

Variables intermedias o intervinientes: En algunos casos de análisis de relación causa-efecto, se introducen una o más variables de enlace interpretativo entre las variables dependientes e independientes.

Variables explicatorias: Son las propiedades que interesan directamente al investigador en términos de su modelo.

Variables externas: Son las que están fuera del interés teórico inmediato y pueden afectar los resultados de la investigación empírica.

La clasificación de las variables depende de cada investigación en particular.

1.8 Métodos paramétricos y no paramétricos

Dentro de los métodos estadísticos se pueden distinguir los métodos paramétricos y no paramétricos. La estadística paramétrica se aplica principalmente a datos de tipo cuantitativo y cada técnica tiene supuestos estadísticos que se deben cumplir para poder aplicar el método; uno de los principales supuestos se refiere a la normalidad de la población de la cual fue extraída la muestra, si no se cumple este supuesto, sobre todo en los casos en que la muestra es de tamaño menor de 30 unidades, las conclusiones a las que se llegue podrían ser erróneas. Cuando las variables que se manejan no son de tipo cuantitativo o cuando no se cumplen

los supuestos estadísticos requeridos para las diferentes pruebas, se utilizan los métodos no paramétricos.

Los métodos utilizados para las variables de tipo cuantitativo (intervalo o razón) son los métodos paramétricos, los cuales presentan buenos niveles de confiabilidad en la predicción. En las escalas cualitativas (nominales u ordinales) se utilizan los métodos estadísticos no paramétricos, que no son tan precisos en su predicción. En la Tabla 1.1 se presentan las principales características de los métodos paramétricos y no paramétricos.

Tabla 1.1 *Principales características de los métodos paramétricos y no paramétricos.*

Métodos paramétricos	Métodos no paramétricos
• Se requieren conocimientos de teoría de la probabilidad, pruebas de hipótesis y funciones de distribución, entre otros.	• Se requieren conocimientos elementales a nivel matemático. Son fáciles de usar y entender.
• Se deben cumplir varios supuestos sobre los datos de la población: distribución normal, varianzas iguales, entre otros.	• Se tienen pocos supuestos, los datos pueden o no tener distribución, es decir, libre distribución.
• Las variables deben ser cuantitativas, con escala de medición de intervalo o de razón.	• Se pueden utilizar con variables de tipo cualitativo con escalas de medición ordinal o nominal. También se pueden utilizar en variables cuantitativas.
• Se pueden realizar análisis multivariados.	• Presenta limitaciones en el análisis multivariado.
• Generalmente se requieren tamaños de muestra grandes ($n > 30$).	• Se pueden trabajar con muestras pequeñas ($n < 30$).
• Se utiliza el total del conjunto de datos.	• Solo se utiliza parte del conjunto de datos.
• Son métodos eficientes y confiables estadísticamente.	• No son tan eficientes estadísticamente, presentan una mayor probabilidad de rechazar una hipótesis nula falsa (error Tipo II).

1.9 Métodos estadísticos por tipo de variable

Un aspecto a considerar en una investigación es definir el tipo de análisis estadístico que se debe realizar dependiendo de las variables y su escala de medición. Como una guía se presentan en la Tabla 1.2 los diversos métodos estadísticos que se pueden aplicar según el tipo de variable y su escala de medición.

Tabla 1.2 *Clasificación de métodos estadísticos dependiente del tipo de variable y su escala de medición.*

Tipo de variable		Método a utilizar		
		Estadística descriptiva	Estadística inferencial paramétrica	Estadística inferencial no paramétrica
Cualitativa	nominal	moda frecuencias	Análisis de correspondencias. Análisis de correlación canónica no lineal. Análisis de homogeneidad. Modelos de regresión de elección discreta.	Tabulación cruzada: Chi-cuadrado, Mcnemar, Cochran, Coeficiente de contingencia, Phi, Cramer's V, Lambda Rachas.
	ordinal	moda frecuencias mediana	Análisis de correspondencias. Análisis de correlación canónica no lineal. Análisis de homogeneidad. Análisis de componentes principales categórico. Regresión categórica. Modelos de regresión de elección discreta–ordenados.	Tabulación cruzada: Chi-cuadrado, Gamma, Somer's d, Kendall's, Tau-b, Kendall's tau-c. Kruskal-Wallis. Prueba de la mediana. Friedman. Mann-Whitney. Wilcoxon. Rachas.
Cuantitativa	discreta	moda frecuencias mediana	Análisis de correspondencias. Análisis de correlación canónica no lineal. Análisis de homogeneidad. Análisis de componentes principales categórico. Regresión categórica. Modelos de regresión de elección discreta–ordenados.	Tabulación cruzada: Chi-cuadrado, Gamma, Somer's d, Kendall's, Tau-b, Kendall's tau-c. Kruskal-Wallis. Prueba de la mediana. Friedman. Mann-Whitney. Wilcoxon. Rachas.
	continua	Todas	Estimación puntual y por intervalo. Pruebas de hipótesis. ANOVA. MANOVA. Análisis de componentes principales. Modelo de regresión lineal simple y múltiple.	Kruskal-Wallis. Prueba de la mediana. Mann-Whitney. Wilcoxon. Signo. Rachas. Chi-cuadrado.

1.10 Etapas generales en la construcción de un modelo estadístico

Como una guía y no como una norma inflexible, se pueden delinear las siguientes etapas en la construcción de un modelo o procesamiento estadístico (Quiroga).

• ***Caracterización del problema***

En esta etapa se deben definir los diferentes aspectos del problema, con el fin de lograr una idea global del mismo, considerando en lo posible ir de lo simple a lo complejo, de las partes al todo. En este aspecto se pueden seguir los siguientes pasos:

El sistema. Definición del sistema y los diversos componentes del sistema, de acuerdo con el problema, su delimitación, los diversos componentes y sus relaciones.

Justificación. Se debe definir el porqué y el para qué de la investigación y del estudio del sistema, aclarando los elementos teóricos sobre el problema y sus fuentes, realizando una revisión del estado del arte. Se deben definir el tipo de parámetros, variables y supuestos sobre sus relaciones; de causalidad o de correlación. Así mismo, se deben definir variables de respuesta, variables de estado, variables endógenas y/o exógenas y la caracterización de información disponible, en inventario y tamaño.

• ***Definición de objetivos e hipótesis***

Se deben plantear los objetivos e hipótesis generales en relación con el problema objeto de la investigación. Las hipótesis deben basarse principalmente en la naturaleza misma del fenómeno o sistema, apoyadas en teorías, experiencias y criterios de personas que conozcan la problemática estudiada. Se deben definir alternativas de modelos y su aplicación.

• ***Marco teórico***

De acuerdo con las hipótesis, se deben exponer los elementos teóricos fundamentales de la investigación y de carácter estadístico que permitirán la construcción, el desarrollo y aplicación de los modelos estadísticos.

• ***Diseño de metodologías estadísticas***

Se debe caracterizar el proceso de muestreo o el diseño experimental utilizado para la obtención de las observaciones, definiendo limitaciones y cobertura (población y muestra). Así mismo, definir los parámetros y las variables, su caracterización y su nivel de importancia: ¿cuáles variables se observan?, ¿cómo se observan?, ¿cuáles se generan? y ¿cómo se generan? Las variables deben clasificarse según diferentes criterios (aleatoria, determinística, de respuesta, independiente, dependiente, observable, no observable, generada, endógena, exógena, de estado, controlada, no

controlada y covariable, entre otras). Debe juzgarse su grado de variabilidad, los posibles factores que la determinan y definir sus categorías.

En la caracterización de parámetros deben explicarse su interpretación y su papel en el sistema o fenómeno. Del mismo modo, describir los métodos de estimación de parámetros, propiedades, errores estándar y criterios para evaluarlos. Se deben describir y explicar la docimasia de hipótesis estadísticas. ¿Qué supuestos se deben validar ? ¿Cuál es su importancia? ¿Cómo validarlos? Se deben describir y explicar los métodos y formas de aplicación del modelo construido y validado, sus alcances, limitaciones y ventajas.

CAPÍTULO
2

Medidas descriptivas

Este capítulo presenta las principales medidas descriptivas de tendencia central y dispersión utilizadas para el resumen de un conjunto de datos. Una medida descriptiva es un valor que caracteriza las observaciones resumiéndolas en medidas de tendencia central, dispersión o variabilidad y forma o asociación.

Las medidas de tendencia central describen valores típicos que se encuentran entre el valor mínimo y el valor máximo observado en el conjunto de datos. Las medidas de dispersión o variabilidad describen en qué medida los valores de un conjunto de datos son distintos entre sí o con respecto a una medida de centralidad. Las medidas de forma describen las características de una distribución de frecuencias de un conjunto de datos. Las medidas de asociación, para el caso de dos o más variables, muestran el grado de asociación entre estas variables y cómo están relacionadas.

2.1 Medidas de tendencia central

Estas medidas permiten describir el grado de centralidad de un conjunto de datos. Son valores que representan un valor central hacia el cual tiene tendencia a concentrarse el conjunto de datos. Entre las medidas de tendencia central se destacan:

- Media:
  aritmética
  geométrica

armónica
cuadrática
rango medio
ponderada

- Mediana
- Moda

Las medidas de centralidad más utilizadas son la media aritmética, mediana y moda. En algunos textos al cálculo de estas tres medidas se le denomina *promedio*.

2.1.1 *Media*

2.1.1.1 Media aritmética

Es la medida más utilizada en el análisis de un conjunto de datos, es un valor central que toma en cuenta todos los valores que aparecen en el conjunto de datos y las distancias relativas a estos valores. Los valores tienen la misma importancia en el grupo de datos.

Su analogía física se puede comparar con el centro de masa de una colección de masas de una dimensión, tal como se presenta en la Figura 2.1

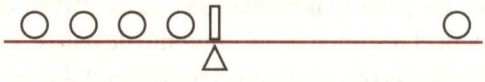


Figura 2.1 *Representación gráfica del concepto de media.*

La media aritmética es la suma de los valores de la variable sobre el número de datos en análisis, la notación en la muestra es diferente que en la población.

Si $X_1, X_2, X_3, \ldots\ldots\ldots, X_{n-1}, X_n$ *representan los valores de una variable en una muestra, entonces la media aritmética se calcula por medio de la ecuación 2.1.*

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum X}{n} \qquad (2.1)$$

$\overline{X}$: *(se lee "X barra" o "X trazo"): media de un conjunto de datos provenientes de una muestra*

n : *número de datos de una muestra*

$\sum$: *(es la letra griega mayúscula sigma): signo de sumatoria (se lee "suma de")*

Cuando los datos representan el total de la población, la notación de la media es diferente de la media de los datos muestrales.

Si $X_1, X_2, X_3, \ldots\ldots, X_{N-1}, X_N$ *representan los valores de una variable en una población, entonces la media aritmética se calcula por medio de la ecuación 2.2.*

$$\mu = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum X}{N} \qquad (2.2)$$

μ *: (es la letra griega minúscula mu): media de un conjunto de datos provenientes de una población*

N *: número de datos de una población*

La media aritmética poblacional se estima a partir de la media aritmética muestral utilizando la teoría de la probabilidad.

En estudios ambientales o de ingeniería sanitaria en muy pocas oportunidades se cuenta con los datos poblacionales, muy frecuentemente se tienen conjuntos de datos provenientes de una muestra, considerando que generalmente los fenómenos naturales tienen población infinita, lo cual impide obtener los datos de la población. Por ejemplo, para estimar la calidad de agua de una fuente de agua o la calidad del aire en una determinada zona, tener la población es equivalente a analizar "toda" el agua del río o "todo" el aire de la zona de estudio, lo cual no es posible. Esto refuerza la importancia de la estimación de la media poblacional a partir de la media muestral.

La media aritmética no siempre tiene sentido conceptual o validez real. Por ejemplo, si en un muestreo de calidad de agua se tiene un valor de pH de 4 unidades, es decir ácido, y un valor de pH de 8 unidades, es decir básico, el promedio del agua daría un pH de 6 unidades, es decir neutro, lo cual no tendría sentido desde el punto de vista real, por lo anterior es necesario analizar la validez lógica y real de esta medida antes de ser utilizada.

La media aritmética sólo tiene sentido para datos cuantitativos, ya sean estos de carácter discreto o continuo, pues no se puede promediar el sexo, que toma categorías de femenino y masculino, así estas estén categorizadas como 0 y 1, debido a que la media daría 0,5, que no tiene sentido ni representación real. ***En el presente texto la media aritmética se denominará media o promedio.*** En la Tabla 2.1 se presentan algunas ventajas y limitaciones de la media aritmética.

Tabla 2.1 *Ventajas y limitaciones de la media aritmética.*

Ventajas	Limitaciones
• Es la medida estadística más comúnmente empleada. • Es fácil de calcular y entender. • Se pueden realizar cálculos algebraicos. • En su cálculo se incluye cada uno de los datos de la muestra o la población. • Es un valor único para cada conjunto de datos. • Las unidades son las mismas de la variable analizada. • La distribución de las medias que se obtienen de muestreos repetidos de una población se conoce y es de gran utilidad en el proceso de inferencia. Generalmente es la distribución normal.	• Es fuertemente afectada por los valores extremos, ya sean valores máximos o mínimos y por consiguiente puede estar lejos de ser una representación de la muestra. • No es conveniente utilizarla en: conjunto de datos demasiado heterogéneos, cuando los datos sean proporcionales o estén en progresión geométrica. • Se debe analizar junto con medidas de dispersión. • Se debe acompañar por otras medidas de tendencia central, tales como la mediana y la moda. • Sólo tiene sentido en variables cuantitativas.

Ejemplo 2.1 *Un monitoreo de la calidad de agua en una fuente superficial, en la variable turbiedad, presenta los siguientes resultados:*

Datos primer muestreo: 5; 4; 5; 4; 8; 10; 9 (UNT) → $\bar{X} = \frac{\sum_{i=1}^{7} x_i}{7} = 6{,}4$ **(UNT)**

Con una muestra adicional: 12 (UNT) → $\bar{X} = \frac{\sum_{i=1}^{8} x_i}{8} = 7{,}1$ **(UNT)**

Con otra muestra adicional: 150 (UNT) → $\bar{X} = \frac{\sum_{i=1}^{9} x_i}{9} = 23$ **(UNT)**

Con otra muestra adicional: 320 (UNT) → $\bar{X} = \frac{\sum_{i=1}^{10} x_i}{10} = 52{,}7$ **(UNT)**

(UNT: Unidades Nefelométricas de Turbiedad)

Considerando el primer muestreo, la media de turbiedad para la fuente superficial es 6,4 UNT, valor que indica el centro del conjunto de datos. A medida que se adicionan valores extremos de turbiedad, la media incrementa su valor significativamente. Un solo dato extremo altera el valor de la media de manera significativa.

El valor de la media para datos homogéneos es un buen indicador del grado de centralidad de un conjunto de datos; sin embargo, es una medida fuertemente afectada por valores extremos, y esto es una gran limitación para el uso de este indicador estadístico sin el análisis conjunto de otras medidas de centralidad o dispersión.

2.1.1.2 Propiedades del operador sumatoria

A continuación se presentan las principales propiedades del operador sumatoria, las cuales permiten comprobar algunas propiedades de la media.

- $\sum_{i=1}^{n} c = nc$ donde c es constante y n el número de datos
- $\sum_{i=1}^{n} c X_i = c \sum_{i=1}^{n} X_i$
- $\sum_{i=1}^{n} \overline{X} = n \overline{X}$
- $\sum_{i=1}^{n} (a X_i \pm b Y_i) = a \sum_{i=1}^{n} X_i \pm b \sum_{i=1}^{n} Y_i$
- $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \longrightarrow \sum_{i=1}^{n} X_i = n \overline{X}$

2.1.1.3 Propiedades de la media

- *La suma de las desviaciones de los datos con respecto a la media es cero. Esta propiedad surge del hecho de que la media es el punto de equilibrio de la distribución, tal como se presenta en la ecuación 2.3. La media es la única medida de tendencia central que cumple esta propiedad.*

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0 \qquad (2.3)$$

Demostración: Aplicando propiedades del operador sumatoria se tiene el siguiente proceso:

$$\sum_{i=1}^{n}(X_i-\overline{X}) = \sum_{i=1}^{n}X_i-\sum_{i=1}^{n}\overline{X} = \sum_{i=1}^{n}X_i - n\overline{X} = n\overline{X} - n\overline{X} = 0$$

- *Las sumas de los cuadrados de las desviaciones a partir de la media aritmética es menor que la suma de cuadrados de las desviaciones a partir de cualquier otro valor. En forma algebraica:*

$$\sum(X_i-\overline{X})^2 \quad \textit{es mínima.}$$

- *Si cada uno de los datos de una variable toma valores constantes (k), la media será igual al valor de la constante. En términos algebraicos:*

  Si $\overline{X} = k$, *para todo* $i = 1, 2, \ldots n$, *entonces* $\overline{X} = k$.

- *Si cada uno de los datos de una variable es afectado aditivamente (negativamente) por una constante (k), la media de la nueva variable es equivalente a sumar (restar) la constante a la media de la variable original. En forma algebraica:*

  Si $Y_i = k \pm X_i$, *para todo* $i = 1, 2, \ldots n$, *entonces* $\overline{Y} = k \pm \overline{X}$.

- *Si cada uno de los datos de una variable es afectado multiplicativamente por una constante (k), la media de la nueva variable es equivalente a multiplicar la constante por la media de la variable original. En forma algebraica:*

  Si $Y_i = kX_i$, *para todo* $i = 1, 2, \ldots n$, *entonces* $\overline{Y} = k\overline{X}$.

- *Si cada uno de los datos de una variable es dividido por una constante (k), entonces la media de la nueva variable es la media de la variable original, dividida por la constante. Algebraicamente:*

  Si $Y_i = \dfrac{X_i}{k}$, *para todo* $i = 1, 2, \ldots n$, *entonces* $\overline{Y} = \dfrac{\overline{X}}{k}$

- *Si se genera una variable como la combinación lineal de dos variables, la media de la nueva variable será la combinación lineal de las medias de las variables originales. Algebraicamente:*

  Si $Z_i = aX_i + bY_i$, *para todo* $i = 1, 2, \ldots n$, *entonces* $\overline{Z} = a\overline{X} + b\overline{Y}$.

- *En general, de todas las medidas utilizadas para calcular la tendencia central de una población, la media es la menos sujeta a variación debida a cambios en la muestra.*

La media es la medida de tendencia central más utilizada en estadística, pues emplea los datos disponibles de una variable y tiene una fuerte aplicabilidad en el proceso de inferir de una muestra a una población, debido a que las distribuciones de medias que se obtienen de muestreos repetidos de una población se conocen y son de gran utilidad en el proceso de inferencia.

2.1.1.4 *Media geométrica*

Esta es una medida de centralidad que se utiliza generalmente cuando los valores dependen del tiempo; varían de manera no lineal o cuando existe un alto grado de heterogeneidad en el conjunto de datos.

La media geométrica de un conjunto de datos $X_1, X_2, X_3, \ldots\ldots\ldots, X_{n-1}, X_n$ *de una muestra se define como la raíz n-ésima de la multiplicación del conjunto de datos y se calcula como se presenta en la ecuación 2.4.*

$$\overline{X}_g = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdot \ldots \cdot X_{n-1} \cdot X_n} \qquad (2.4)$$

Para facilitar el cálculo se aplica la función log a ambos lados de la ecuación:

$$\log \overline{X}_g = \log \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n}$$

$$= \log (X_1 \cdot X_2 \cdot \ldots \cdot X_n)^{1/n}$$

$$= \frac{1}{n} \log (X_1 \cdot X_2 \cdot \ldots \cdot X_n)$$

$$= \frac{1}{n} (\log X_1 + \log X_2 + \ldots + \log X_n)$$

generando la ecuación 2.5.

$$\log \overline{X}_g = \frac{\sum_{i=1}^{n} \log(X_i)}{n} \qquad (2.5)$$

Entonces para hallar la media geométrica se aplica la función exponencial en base 10, a ambos lados de la igualdad, generando:

$$\overline{X}_g = 10^{\log \overline{X}_g} = 10^{\frac{\sum \log X}{n}}$$

Cuando los datos representan el total de la población la notación de la media geométrica se presenta a continuación.

La media geométrica de un conjunto de datos $X_1, X_2, X_3, \ldots\ldots\ldots, X_{N-1}, X_N$ de una población, se define como la raíz N-ésima de la multiplicación del conjunto de datos y se calcula como se presenta en la ecuación 2.6.

$$\mu_g = \sqrt[N]{X_1 \cdot X_2 \cdot X_3 \cdot \ldots \cdot X_{N-1} \cdot X_N} \qquad (2.6)$$

El empleo de la media geométrica es equivalente a realizar una transformación de la variable original ***X***, en ***log(X)*** y el posterior cálculo de la media aritmética a la nueva variable, para obtener el logaritmo de la media geométrica. Por ejemplo, si la variable abarca un campo de variación muy grande, tal como el porcentaje de impureza de un producto químico (por lo general alrededor del 0.1%, pero en ocasiones llega incluso al 1% o más); en este caso es conveniente el empleo de ***log X*** en lugar de ***X*** para obtener una distribución más simétrica y una aproximación más cercana a la curva normal. En la Tabla 2.2 se presentan algunas ventajas y limitaciones de la media geométrica.

Tabla 2.2 *Ventajas y limitaciones de la media geométrica*

Ventajas	Limitaciones
• Es una medida resistente a datos extremos, permite detectar en un conjunto muy heterogéneo, una medida de tendencia central confiable. • Las unidades de la media geométrica son las mismas de la variable. • Se pueden realizar cálculos algebraicos. • En su cálculo se incluye cada uno de los datos de la muestra. • Es un valor único para un conjunto de datos. • Es muy útil cuando el conjunto de datos representa aumentos o disminuciones porcentuales. • Se utiliza para promediar valores cuyo crecimiento sea en progresión geométrica.	• No es fácil de calcular y para un número considerable de datos ($n > 150$), se presentan limitaciones en el programa Excel. En el programa SPSS no está considerada dentro de las rutinas más comunes. • Puede presentar limitaciones en su interpretación. • Cuando existe uno o varios valores de la variable iguales a cero, el valor de la media geométrica toma automáticamente el valor de cero. • Sólo se puede calcular cuando la raíz n-ésima exista. • Programas como Excel no validan el signo del producto y siempre que hay valores negativos no la calcula. • Sólo tiene sentido en variables de carácter cuantitativo. • El desarrollo algebraico de esta medida puede tener un grado de complejidad mayor que el desarrollo de la media aritmética.

Ejemplo 2.2 *Considerando la situación del ejemplo 2.1 se calcula la media geométrica:*

Datos del primer muestreo:

5; 4; 5; 4; 8; 10; 9 (UNT) → $\overline{X}_g = \sqrt[7]{x_1 \cdot x_2 \cdot \cdot x_7} = 6$ *(UNT)*

Considerando una muestra adicional:

12 (UNT) → $\overline{X}_g = \sqrt[8]{x_1 \cdot x_2 \cdot \cdot x_8} = 6{,}6$ *(UNT)*

Considerando otra muestra adicional:

150 (UNT) → $\overline{X}_g = \sqrt[9]{x_1 \cdot x_2 \cdot \cdot x_9} = 9{,}3$ *(UNT)*

Considerando otra muestra adicional:

320 (UNT) → $\overline{X}_g = \sqrt[10]{x_1 \cdot x_2 \cdot \cdot x_{10}} = 13{,}2$ *(UNT)*

La media geométrica para los datos del primer muestreo es 6 UNT y a medida que se incorporan datos extremos la media geométrica se incrementa levemente en comparación con la alteración que presentan las medias aritméticas calculadas en el ejemplo 2.1.

El valor de la media geométrica es considerablemente menos afectado por valores extremos en comparación con los valores de la media aritmética, generando una medida más cercana a la centralidad del conjunto de datos cuando el conjunto de datos es heterogéneo.

2.1.1.5 Media armónica

Equivale a la transformación del conjunto de datos originales en el recíproco de cada dato, $1/X$, y luego se calcula la media de los datos transformados, es el recíproco de $\overline{X}$. Su campo de aplicación es bastante restringido. Es útil al promediar velocidades, volúmenes de ventas y cuando la variable crece en progresión armónica.

La media armónica de un conjunto de datos $X_1, X_2, X_3, \ldots\ldots\ldots, X_{n-1}, X_n$ *provenientes de una muestra se define como la media de los recíprocos del conjunto de datos, tal como se presenta en la ecuación 2.7.*

$$\overline{X}_h = \frac{1}{\dfrac{\left(\sum_{i=1}^{n} \frac{1}{X_i}\right)}{n}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} \qquad (2.7)$$

Siempre que $X_i \neq 0$

Para un conjunto de datos provenientes de una población se calcula como se presenta a continuación.

La media armónica de un conjunto de datos $X_1, X_2, X_3, \ldots\ldots\ldots, X_{N-1}, X_N$ *provenientes de una población se define como la media de los recíprocos del conjunto de datos, tal como se presenta en la ecuación 2.8.*

$$\mu_h = \frac{1}{\frac{\left(\sum_{i=1}^{N} \frac{1}{X_i}\right)}{N}} = \frac{N}{\sum_{i=1}^{N} \frac{1}{X_i}} \qquad (2.8)$$

Siempre que $X_i \neq 0$

La relación entre las medias aritmética, geométrica y armónica se presenta en la desigualdad 2.9.

$$\overline{X}_h \leq \overline{X}_g \leq \overline{X} \qquad (2.9)$$

La media armónica es la más resistente a valores extremos, seguida por la media geométrica y luego la media aritmética. Las fortalezas de la media aritmética son sus propiedades, las cuales permiten desarrollos algebraicos y propiedades importantes para la inferencia estadística y la distribución normal que presenta la familia de medias de un estudio.

Ejemplo 2.3 *Considerando la situación del ejemplo 2.1 se calcula la media armónica:*

Datos del primer muestreo:

5; 4; 5; 4; 8; 10; 9 (UNT) → $\overline{X}_h = \frac{7}{\sum_{i=1}^{7} \frac{1}{X_i}} = 5{,}7\ (UNT)$

Con una muestra adicional:

12 (UNT) → $\overline{X}_h = \frac{8}{\sum_{i=1}^{8} \frac{1}{X_i}} = 6{,}1\ (UNT)$

Con otra muestra adicional:

150 (UNT) → $\overline{X}_h = \frac{9}{\sum_{i=1}^{9} \frac{1}{X_i}} = 6{,}8\ (UNT)$

Con otra muestra adicional:

$$320\ (UNT) \rightarrow \bar{X}_h = \frac{10}{\sum_{i=1}^{10} \frac{1}{X_i}} = 7{,}5\ (UNT)$$

El valor de la media armónica para turbiedad en el primer muestreo es 5,7 UNT, y a medida que se adicionan valores extremos a la muestra el valor de la media armónica no se incrementa significativamente.

Como se puede observar, a través de los ejemplos 2.1, 2.2 y 2.3, se cumple la relación de desigualdad presentada en la ecuación 2.9 entre las medias armónica, geométrica y aritmética. La media armónica genera los menores valores de centralidad del conjunto de datos y es la que menor impacto presenta por valores extremos. Sin embargo, esta medida presenta limitaciones en su manejo algebraico y no existe cuando algún dato toma el valor de cero. Así mismo no posee ventajas en su distribución.

2.1.1.6 Media cuadrática

Es otra medida de tendencia central, que consiste en elevar al cuadrado los valores y generar la raíz cuadrada de la media aritmética de estos nuevos valores, es poco afectada por valores extremos, pero presenta pocas ventajas algebraicas y de distribución.

La media cuadrática de un conjunto de datos $X_1, X_2, X_3, \ldots\ldots, X_{n-1}, X_n$ *provenientes de una muestra se define como se presenta en la ecuación 2.10.*

$$\bar{X}^2 = \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}} \qquad (2.10)$$

$\bar{X}^2$ *es la notación para la media cuadrática muestral*

Cuando los datos representan la totalidad de una población la definición de la media cuadrática se presenta a continuación.

La media cuadrática de un conjunto de datos $X_1, X_2, X_3, \ldots\ldots, X_{N-1}, X_N$ *provenientes de una población se define como se presenta en la ecuación 2.11.*

$$\mu^2 = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N}} \qquad (2.11)$$

μ^2 *es la notación para la media cuadrática poblacional*

Ejemplo 2.4 *Considerando la situación del ejemplo 2.1 se calcula la media cuadrática:*

Datos del primer muestreo:

$$5;\ 4;\ 5;\ 4;\ 8;\ 10;\ 9\ (UNT) \rightarrow \overline{X}^2 = \sqrt{\frac{\sum_{i=1}^{7} X_i^2}{7}} = 6{,}8\ (UNT)$$

Con un dato adicional:

$$12\ (UNT) \rightarrow \overline{X}^2 = \sqrt{\frac{\sum_{i=1}^{8} X_i^2}{8}} = 7{,}7\ (UNT)$$

Con otro dato adicional:

$$150\ (UNT) \rightarrow \overline{X}^2 = \sqrt{\frac{\sum_{i=1}^{9} X_i^2}{9}} = 50{,}5\ (UNT)$$

Con otro dato adicional:

$$320\ (UNT) \rightarrow \overline{X}^2 = \sqrt{\frac{\sum_{i=1}^{10} X_i^2}{10}} = 112\ (UNT)$$

El valor de la media cuadrática para turbiedad en el primer muestreo es 6,8 UNT, pero a medida que se adicionan valores extremos el valor de la media cuadrática aumenta significativamente.

La media cuadrática presenta más variabilidad que la media aritmética. Esta medida es fuertemente afectada por valores extremos.

2.1.1.7 Rango medio o semirrango

Otro valor representativo de importancia, sobre todo cuando se necesita rápidamente una medida de centralidad es el rango medio o semirrango.

El rango medio se define como la media aritmética del valor máximo y el valor mínimo de un conjunto de datos y se calcula como se presenta en la ecuación 2.12.

$$RM = \frac{X_{mín} + X_{máx}}{2} \qquad (2.12)$$

Donde $X_{mín}$ *es el valor mínimo y* $X_{máx}$ *es el valor máximo del conjunto de datos.*

Aunque el rango medio se calcula fácil y rápidamente, a menudo es ineficiente porque ignora la información contenida en los términos intermedios. Así mismo puede que no sea representativo, en el caso de que alguno de los valores máximo o mínimo, sean valores especiales o atípicos dentro del conjunto de datos.

Ejemplo 2.5 *Considerando la situación del ejemplo 2.1 se calcula el rango medio:*

Datos del primer muestreo:

$$5;\ 4;\ 5;\ 4;\ 8;\ 10;\ 9\ (UNT) \rightarrow RM = \frac{X_{mín} + X_{máx}}{2} = 7\ (UNT)$$

Con un dato adicional:

$$12\ (UNT) \rightarrow RM = \frac{X_{mín} + X_{máx}}{2} = 8\ (UNT)$$

Con otro dato adicional:

$$150\ (UNT) \rightarrow RM = \frac{X_{mín} + X_{máx}}{2} = 77\ (UNT)$$

Con otro dato adicional:

$$320\ (UNT) \rightarrow RM = \frac{X_{mín} + X_{máx}}{2} = 162\ (UNT)$$

El rango medio para turbiedad en el primer muestreo es 7 UNT; sin embargo, cuando se adicionan datos extremos esta media aumenta significativamente.

El valor del rango medio presenta una variación similar al valor de la media aritmética, por su definición es afectada por los valores extremos.

2.1.1.8 Media ponderada

Cuando se conoce la media de varios grupos de datos y el número de datos en cada grupo, se puede calcular la media global que se conoce como la media ponderada, mediante la ecuación 2.13.

$$\overline{X}_p = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \ldots\ldots + n_k\overline{X}_k}{n_1 + n_2 + \ldots\ldots + n_k} \qquad (2.13)$$

En el siguiente ejemplo se ilustra su uso.

Ejemplo 2.6 *Se ha realizado un monitoreo de 4 meses sobre la calidad de agua en sólidos suspendidos (mg/l), en el afluente de una planta de tratamiento de agua potable. Las medias mensuales se presentan a continuación:*

Sólidos suspendidos (mg/l)	Mes 1	Mes 2	Mes 3	Mes 4
$\overline{X}$	9,8	11,4	7,5	10,5
n	13	18	20	15

Para el cálculo de la media se utiliza la media ponderada, descrita en la ecuación 2.13

$$\overline{X}_p = \frac{(13 \cdot 9{,}8) + (18 \cdot 11{,}4) + (20 \cdot 7{,}5) + (15 \cdot 10{,}5)}{13 + 18 + 20 + 15}$$

$$\overline{X}_p = 9{,}7 \text{ mg/l}$$

Es decir, la media de sólidos suspendidos en el afluente de la planta durante los 4 meses fue de 9,7 mg/l

2.1.2 *Mediana*

Es la segunda medida más utilizada después de la media aritmética para estimar el centro de un conjunto de datos. Para hallar la mediana de un conjunto de datos estos deben ser inicialmente puestos en orden de magnitud, de manera creciente o decreciente. La mediana es el elemento central del conjunto de datos, es una medida de posición; hay el mismo número de observaciones a la derecha y a la izquierda del valor de la mediana.

La mediana divide la distribución de los datos en el punto medio; el 50% de los datos está por encima de la mediana y el otro 50% está por debajo de la mediana, es decir, es el valor que divide el conjunto de datos en dos grupos iguales.

Si $X_1, X_2, X_3, \ldots\ldots, X_{n-1}, X_n$ *representan los valores ordenados de forma ascendente o descendente de una variable seleccionada de una muestra, entonces la mediana se calcula mediante la ecuación 2.14.*

$$M_e = \begin{cases} X_{\frac{n+1}{2}} & si\ n\ es\ impar \\ \dfrac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} & si\ n\ es\ par \end{cases} \qquad (2.14)$$

Cuando los datos representan la totalidad de una población la fórmula de la mediana se presenta a continuación:

Si $X_1, X_2, X_3, \ldots\ldots\ldots, X_{N-1}, X_N$ *representan los valores ordenados de forma ascendente o descendente de una variable seleccionada de una población, entonces la mediana se calcula mediante la ecuación 2.15.*

$$M_e = \begin{cases} X_{\frac{N+1}{2}} & si\ N\ es\ impar \\ \dfrac{X_{\frac{N}{2}} + X_{\frac{N}{2}+1}}{2} & si\ N\ es\ par \end{cases} \qquad (2.15)$$

Si el número de datos es impar, la mediana es el dato del centro del conjunto de datos. Una vez los datos se ordenen en forma ascendente o descendente. Los datos que se repiten deben ser ordenados, también, en su secuencia lógica. Si el número de datos es par, la mediana es la media de los dos datos del centro. En la Tabla 2.3 se presentan algunas ventajas y limitaciones de la mediana.

Tabla 2.3 *Ventajas y limitaciones de la mediana.*

Ventajas	Limitaciones
• Su valor no se ve afectado por datos extremos y por lo tanto es una medida de importancia cuando se presenta esta situación en un conjunto de datos.	• Es afectada por el número de observaciones, pero no por su magnitud.
• Es fácil de calcular y entender.	• En general la mediana es menos estable que la media de una muestra a otra, por lo tanto no es tan útil en la estadística inferencial.
• Las unidades de la mediana son las mismas de la variable.	• Los datos deben ser ordenados antes de calcular la mediana.
• Se puede hallar en variables cualitativas y cuantitativas.	• Su definición no permite realizar procesos algebraicos.
• Es un valor único para un conjunto de datos.	
• Cuando los datos tienen una marcada asimetría, es mejor representar la tendencia central con la mediana que con la media.	

Ejemplo 2.7 *Considerando la situación del ejemplo 2.1 se calcula la mediana:*

-

Datos del primer muestreo (n=7):

5; 4; 5; 4; 8; 10; 9 (UNT) → $M_e = X_{\frac{7+1}{2}} = X_4 = 5\ (UNT)$

Con un dato adicional (n=8):

12 (UNT) → $M_e = \dfrac{X_{\frac{8}{2}} + X_{\frac{8}{2}+1}}{2} = \dfrac{X_4 + X_5}{2} = 6{,}5\ (UNT)$

Con otro dato adicional (n=9):

150 (UNT) → $M_e = X_{\frac{9+1}{2}} = X_5 = 8\ (UNT)$

Con otro dato adicional (n=10):

320 (UNT) → $M_e = \dfrac{X_{\frac{10}{2}} + X_{\frac{10}{2}+1}}{2} = \dfrac{X_5 + X_6}{2} = 8{,}5\ (UNT)$

La mediana para la turbiedad en el primer muestreo es 5 UNT, es decir, el 50% de los datos son menores a 5 UNT y el 50% son mayores a 5 UNT. A medida que se adicionan datos extremos esta medida varía levemente.

El valor de la mediana es el valor central de la distribución de datos, es una medida bastante resistente a valores extremos, por lo tanto es una buena medida de centralidad del conjunto de datos.

2.1.3 *Moda*

Como su nombre lo indica, representa el valor o valores que tienen la mayor frecuencia en el conjunto de datos; son los valores que más se repiten, ya sean estos muestrales o poblacionales. En un conjunto de datos puede no existir un valor modal o existir una o más modas. Cuando hay una moda, el conjunto de datos se denomina unimodal, en el caso de dos modas se denomina bimodal, en el caso de tres modas se denomina trimodal y en el caso de más modas se denomina multimodal. La moda se representa como M_o para datos muestrales o poblacionales. En la Tabla 2.4 se muestran algunas ventajas y limitaciones de la moda.

Tabla 2.4 *Ventajas y limitaciones de la moda.*

Ventajas	Limitaciones
• Es fácil de calcular y entender. • Las unidades de la moda son las mismas de la variable. • No requiere cálculo. • Puede utilizarse para datos cualitativos y datos cuantitativos. • No es afectada por datos extremos aislados.	• La moda no necesariamente ocurrirá como un valor central. • La moda no siempre existe. • No se pueden realizar procesos algebraicos. • No presenta mucha utilidad con pocos datos en el conjunto de análisis. • En general cuando el conjunto de datos no resulta unimodal se debe a posibles fallas en el muestreo o falta de homogeneidad de los mismos. • A pesar de describirse como una medida de centralidad, cuando los datos no son simétricos, no la representa.

Ejemplo 2.8 *Considerando la situación del ejemplo 2.1, se estima la moda:*

Primer muestreo:

$$5;\ 4;\ 5;\ 4;\ 8;\ 10;\ 9\ (UNT) \rightarrow M_{o1} = 4\ (UNT) \text{ y } M_{o2} = 5\ (UNT)$$

Con un dato adicional:

$$12\ (UNT) \rightarrow M_{o1} = 4\ (UNT) \text{ y } M_{o2} = 5\ (UNT)$$

Con otro dato adicional:

$$150\ (UNT) \rightarrow M_{o1} = 4\ (UNT) \text{ y } M_{o2} = 5\ (UNT)$$

Con otro dato adicional:

$$320\ (UNT) \rightarrow M_{o1} = 4\ (UNT) \text{ y } M_{o2} = 5\ (UNT)$$

Los datos del primer muestreo presentan dos modas, es decir, es un conjunto de datos bimodal; los valores que mayor frecuencia presentan en turbiedad son 4 UNT y 5 UNT. A medida que se incorporan datos extremos al conjunto de datos las modas se mantienen constantes, en este caso específico.

Si se obtienen diferentes muestras de una población en forma aleatoria, la media varía en cada una de ellas, lo mismo sucede con la mediana y la moda. Sin embargo, la media varía menos que la mediana y la moda, lo cual es muy importante en la estadística inferencial y es una de las principales razones del uso de la media en

esta rama de la estadística. Una media muestral con seguridad está más cerca de la media poblacional que la mediana o la moda de la muestra.

La media, la mediana y la moda proporcionan una parte de la descripción del conjunto de datos. Sin embargo, es necesario definir indicadores que permitan estimar el grado de variación o dispersión de los datos con relación a las medidas de tendencia central y del conjunto de datos en general. Estas medidas por sí solas no son suficientes para analizar y tomar decisiones en relación con un fenómeno en estudio, como se ilustra en el siguiente ejemplo.

Ejemplo 2.9 *Se evalúa el efluente de dos reactores en paralelo para tratamiento de agua potable, en la variable color real medida en Unidades de Platino Cobalto (UPC), generando las siguientes medias:*

Reactor 1: $\overline{X}_1 = 10$ **UPC**

Reactor 2: $\overline{X}_2 = 10$ **UPC**

En el análisis y comparación de estos dos reactores se estaría muy tentado a concluir la igualdad en el efluente para color real. Sin embargo, los datos con los cuales se calcularon las medias se presentan a continuación:

Reactor 1: *10; 12; 10; 12; 8; 10; 8 UPC*
Reactor 2: *58; 2; 2; 2; 2; 2; 2 UPC*

Como se puede apreciar, los datos arrojados por los dos reactores en color real difieren significativamente, factor que no se puede evidenciar sólo a través del valor de la media. Por lo tanto, a pesar de ser la media una de las medidas más utilizadas para resumir y analizar un conjunto de datos, es necesario acompañar esta medida con otras medidas de centralidad y dispersión, las cuales permitan estimar el grado de variación del conjunto de datos.

En la Tabla 2.5 se presentan otras medidas de centralidad que permiten analizar de forma más integral la calidad de agua en color real de los dos reactores. Se puede apreciar, a través de estas medidas, que el reactor 1 tiene más homogeneidad en el conjunto de datos, en comparación con el reactor 2, debido a que en el primero las medidas de tendencia central son muy similares, mientras que en el segundo difieren significativamente.

Tabla 2.5 *Medidas de tendencia central de dos reactores para potabilización de agua en color real.*

Medidas de tendencia central	Color Real (UPC)	
	Reactor 1	Reactor 2
Media	10	10
Mediana	10	2
Media geométrica	9,9	3,2
Moda	10	2

Se puede generalizar que un conjunto de datos es homogéneo cuando la media, la mediana y la media geométrica presentan valores similares, en caso contrario se presenta heterogeneidad en el conjunto de datos. Sin embargo, existen medidas descriptivas que miden en forma adecuada el grado de dispersión o variabilidad del conjunto de datos, denominadas medidas de dispersión.

2.2 Medidas de dispersión

Las medidas de dispersión o variabilidad permiten generar criterios sobre el grado de homogeneidad o heterogeneidad del conjunto de datos que se está analizando, en relación con una medida de centralidad, o con respecto a los datos en sí. Las medidas estadísticas más utilizadas para medir el grado de variabilidad o dispersión son: rango, desviación media, varianza, desviación estándar y coeficiente de variación.

2.2.1 *Rango*

Es la diferencia entre el valor máximo y el valor mínimo del conjunto de datos. Mide la longitud en la cual se encuentran los datos, en general a mayor longitud mayor dispersión de los datos; sin embargo, es necesario analizar la variable y las unidades en las cuales se está midiendo, con el fin de hacer un análisis adecuado de esta medida de dispersión.

El rango de una muestra aleatoria o de una población se define por la ecuación 2.16.

$$R = X_{máx} - X_{mín} \qquad (2.16)$$

En la Tabla 2.6 se presentan algunas ventajas y limitaciones de esta medida de dispersión.

Tabla 2.6 *Ventajas y limitaciones del rango.*

Ventajas	Limitaciones
• Es la medida de variación más fácil de calcular y entender. • Las unidades coinciden con las de la variable de análisis.	• No se pueden realizar cálculos algebraicos. • Sólo incluye dos datos para su cálculo: el valor máximo y el valor mínimo, ignorando los valores intermedios. • Es fuertemente afectada por los valores extremos. • Se debe acompañar de otras medidas de dispersión para su análisis.

Ejemplo 2.10 *Considerando la situación del ejemplo 2.1 se calcula el rango:*

Datos del primer muestreo:

5; 4; 5; 4; 8; 10; 9 (UNT) → $R = X_{máx} - X_{mín} = 6$ *(UNT)*

Con un dato adicional: 12 (UNT) → $R = X_{máx} - X_{mín} = 8$ *(UNT)*

Con otro dato adicional: 150 (UNT) → $R = X_{máx} - X_{mín} = 146$ *(UNT)*

Con otro dato adicional: 320 (UNT) → $R = X_{máx} - X_{mín} = 316$ *(UNT)*

Para el primer muestreo el rango es 6 UNT, es decir, la diferencia entre el valor mínimo y el valor máximo es 6 UNT. A medida que se incorporan datos extremos el rango aumenta considerablemente, evidenciando el grado de dispersión de los datos.

Como se puede apreciar, a medida que el conjunto de datos presenta más variación o heterogeneidad, el rango incrementa su valor de forma significativa. El rango es una buena medida del grado de dispersión de un conjunto de datos.

2.2.2 *Desviación media*

Se define como la media aritmética de los valores absolutos de las desviaciones de los datos, con respecto a la media; también se puede calcular en relación con la mediana, en este último caso la desviación media representa un valor menor. Una limitación de esta medida es su poca facilidad para el desarrollo algebraico.

En la obtención de esta medida intervienen todos los valores del análisis; por lo tanto, permite una información relativa de todos ellos, y da mejor conocimiento del grado de variabilidad de la distribución de los datos que el rango.

Si $X_1, X_2, X_3, \ldots\ldots, X_{n-1}, X_n$ representan los valores de una variable en una muestra, entonces la desviación media se calcula por medio de la ecuación 2.17.

$$dm = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} \qquad (2.17)$$

Si los datos son el total de la población, la notación de la desviación media se presenta a continuación:

Si $X_1, X_2, X_3, \ldots\ldots, X_{N-1}, X_N$ representan los valores de una variable en una población, entonces la desviación media se calcula por medio de la ecuación 2.18.

$$DM = \frac{\sum_{i=1}^{N} |X_i - \mu|}{N} \qquad (2.18)$$

Ejemplo 2.11 *Considerando la situación del ejemplo 2.1 se calcula la desviación media:*

Datos del primer muestreo:

5; 4; 5; 4; 8; 10; 9 (UNT) → $dm = \frac{\sum_{i=1}^{7} |X_i - \bar{X}|}{7} = 2{,}2\ (UNT)$

Con un dato adicional:

12 (UNT) → $dm = \frac{\sum_{i=1}^{8} |X_i - \bar{X}|}{8} = 2{,}6\ (UNT)$

Con otro dato adicional:

150 (UNT) → $dm = \frac{\sum_{i=1}^{9} |X_i - \bar{X}|}{9} = 28{,}2\ (UNT)$

Con otro dato adicional:

320 (UNT) → $dm = \frac{\sum_{i=1}^{10} |X_i - \bar{X}|}{10} = 72{,}9\ (UNT)$

La desviación media para el primer conjunto de datos toma el valor de 2,2 UNT, que indica el nivel de dispersión de los datos con relación al valor medio, que es 6,4 UNT. Cuando se introducen datos extremos al muestreo, la desviación media aumenta evidenciando el grado de dispersión del conjunto de datos.

A medida que el conjunto de datos presenta mayor variabilidad la desviación media aumenta su valor y permite medir el grado de variabilidad del conjunto de datos.

2.2.3 *Varianza*

Debido a las limitaciones algebraicas que evidencian el rango y la desviación media, se origina el concepto de varianza, que mide las variaciones del conjunto de datos con respecto a su media aritmética y se define como la media aritmética de los cuadrados de las desviaciones de cada dato a la media aritmética. En general, cuanto menor sea el valor de la varianza, menor es el grado de variación o heterogeneidad del conjunto de datos con respecto a su media aritmética. Sin embargo, es necesario contextualizar el análisis de esta medida a la variable y las unidades en que está medida.

Si $X_1, X_2, X_3, \ldots\ldots\ldots, X_{n-1}, X_n$ *representan los valores de una variable seleccionada de una muestra, entonces se define la varianza muestral como la ecuación 2.19.*

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + (X_3 - \overline{X})^2 + \ldots. + (X_n - \overline{X})^2}{n-1} \qquad (2.19)$$

El cociente ***(n -1)*** se utiliza en reemplazo de n, debido a que con esta definición se obtiene una mejor estimación de la variable poblacional, es decir, el valor esperado de S^2 es igual a σ^2, en términos matemáticos:

$$E(S^2) = \sigma^2$$

Además, S^2 cumple con la propiedad de ser un estimador insesgado, una característica deseable para un estimador.

En el caso de que los datos sean el total de la población, la notación se presenta a continuación:

Si $X_1, X_2, X_3, \ldots\ldots\ldots, X_{N-1}, X_N$ *representan los valores de una variable seleccionada de una población, entonces se define la varianza poblacional como la ecuación 2.20.*

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N} = \frac{(X_1 - \mu)^2 + (X_2 - \mu)^2 + (X_3 - \mu)^2 + \ldots. + (X_N - \mu)^2}{N} \qquad (2.20)$$

σ: *es la letra griega "sigma"*

En la Tabla 2.7 se muestran algunas ventajas y limitaciones de la varianza.

Tabla 2.7 *Ventajas y limitaciones de la varianza.*

Ventajas	Limitaciones
• Es de las medidas de variación, la más utilizada. • Se pueden realizar cálculos algebraicos. • Se incluyen todos los datos en su cálculo.	• Las unidades de esta medida son las unidades de la variable al cuadrado. • No es fácil su interpretación debido a sus unidades. • Se debe acompañar de otras medidas de dispersión para su análisis.

Ejemplo 2.12 *Considerando la situación del ejemplo 2.1 se puede calcular el valor de la varianza:*

Datos del primer muestreo:

5; 4; 5; 4; 8; 10; 9 (UNT) →
$$S^2 = \frac{\sum_{i=1}^{7} (X_i - \bar{X})^2}{7-1} = 6{,}3\ (UNT)^2$$

Con un dato adicional:

12 (UNT) →
$$S^2 = \frac{\sum_{i=1}^{8} (X_i - \bar{X})^2}{8-1} = 9{,}3\ (UNT)^2$$

Con otro dato adicional:

150 (UNT) →
$$S^2 = \frac{\sum_{i=1}^{9} (X_i - \bar{X})^2}{9-1} = 2276{,}3\ (UNT)^2$$

Con otro dato adicional:

320 (UNT) →
$$S^2 = \frac{\sum_{i=1}^{10} (X_i - \bar{X})^2}{10-1} = 10844{,}3\ (UNT)^2$$

Como se puede apreciar la varianza genera una idea significativa del grado de variabilidad de un conjunto de datos, pues a medida que aumenta el grado de heterogeneidad esta medida aumenta sustancialmente, aunque sus unidades elevadas al cuadrado limitan fuertemente su interpretación.

2.2.3.1 Propiedades de la varianza

- *El valor de la varianza es siempre positivo o igual a cero, esto es:* $S^2 \geq 0$, *para cualquier conjunto de datos.*

- *Si todos los valores de un conjunto de datos son constantes, el valor de la varianza es igual a cero. Algebraicamente:*

  Si $X_i = k$, *para todo* $i = 1, 2, \ldots n$, *entonces* $S^2 = 0$.

- *La varianza no se altera cuando a cada uno de los datos se le suma o se le resta una constante. En términos algebraicos:*

  Si $Y_i = X_i \pm k$, *para todo* $i = 1, 2, \ldots n$, *entonces* $S_y^2 = S_x^2$.

- *Si cada uno de los datos en análisis se multiplica por una constante, la varianza resultará multiplicada por la constante al cuadrado. Algebraicamente:*

  Si $Y_i = kX_i$, *para todo* $i = 1, 2, \ldots n$, *entonces* $S_y^2 = k^2 S_x^2$.

- *Si se divide por un mismo número a cada uno de los datos en análisis, la varianza quedará multiplicada por el cuadrado de dicho divisor. En este caso la constante debe ser diferente de cero. Algebraicamente:*

  Si $Y_i = \frac{1}{k} X_i$, *para todo* $i = 1, 2, \ldots n$, *entonces* $S_y^2 = \frac{1}{k^2} S_x^2$; $k \neq 0$

- *Una ecuación alternativa para el cálculo* ***aproximado*** *de la varianza se presenta a continuación:*

$$S^2 \cong \frac{\sum_{i=1}^{n} X_i^2}{n} - \bar{X}^2$$

2.2.4 *Desviación estándar*

La forma de superar una de las limitaciones de la varianza, sus unidades al cuadrado, es a través del uso de la raíz cuadrada, dando origen al concepto de desviación estándar.

La desviación estándar muestral se define como la raíz cuadrada positiva de la varianza muestral, tal como se presenta en la ecuación 2.21.

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \qquad (2.21)$$

La desviación estándar poblacional se define como la raíz cuadrada positiva de la varianza poblacional, tal como se presenta en la ecuación 2.22.

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}} \qquad (2.22)$$

En la Tabla 2.8 se presentan algunas ventajas y limitaciones de la desviación estándar.

Tabla 2.8 *Ventajas y limitaciones de la desviación estándar.*

Ventajas	Limitaciones
• Es, junto con la varianza, una de las medidas de variación más utilizadas. • Tiene las mismas unidades de la variable analizada. • Se pueden realizar cálculos algebraicos. • Se incluyen todos los datos en su cálculo.	• Se debe acompañar de otras medidas de dispersión para su análisis. • Para su cálculo primero debe calcularse la varianza.

Ejemplo 2.13 *Considerando la situación del ejemplo 2.1 se calcula el valor de la desviación estándar:*

Datos del primer muestreo:

$$5;\ 4;\ 5;\ 4;\ 8;\ 10;\ 9\ (UNT) \rightarrow S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{7}(X_i - \bar{X})^2}{7-1}} = 2{,}5\ (UNT)$$

Con un dato adicional:

$$12\ (UNT) \rightarrow S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{8}(X_i - \bar{X})^2}{8-1}} = 3\ (UNT)$$

Con otro dato adicional:

$$150\ (UNT) \rightarrow S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{9}(X_i - \bar{X})^2}{9-1}} = 47{,}7\ (UNT)$$

Con otro dato adicional:

$$320\ (UNT) \rightarrow S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{10}(X_i - \bar{X})^2}{10-1}} = 104{,}1\ (UNT)$$

Para los datos del primer muestreo la desviación estándar es 2,5 UNT, que indica poca variación entre los datos, es decir, los datos se alejan de la media (6,4 UNT) en una desviación estándar en 2,5 UNT hacia adelante y en 2,5 UNT hacia atrás de la media. A medida que el conjunto de datos se vuelve heterogéneo, la desviación estándar toma valores bastante grandes. Por ejemplo, con todo el conjunto de datos analizados, el valor de la desviación estándar es 104,1 UNT, lo que significa que los datos se alejan en promedio 104,1 UNT del valor medio (6,4 UNT).

Ésta es una buena medida del grado de dispersión del conjunto de datos; a medida que aumenta el grado de variación de los datos esta medida aumenta, en las mismas unidades de la variable de origen.

2.2.5 *Coeficiente de variación*

El coeficiente de variación permite estimar la relación porcentual entre el valor de la media y la desviación estándar. A medida que se presenta mayor heterogeneidad en el conjunto de datos, el valor del coeficiente de variación es mayor. Esta medida puede tomar valores negativos sólo cuando la media tiene un valor negativo, por ejemplo, en el caso de la variable temperatura o nivel de pérdidas. En este caso se sugiere tomar el valor absoluto para una mejor interpretación del coeficiente de variación.

El coeficiente de variación muestral consiste en expresar la desviación estándar muestral como un porcentaje de la media muestral, tal como se presenta en la ecuación 2.23

$$CV = \frac{S}{\overline{X}} x\ 100\% \qquad (2.23)$$

Siempre que $\overline{X} \neq 0$

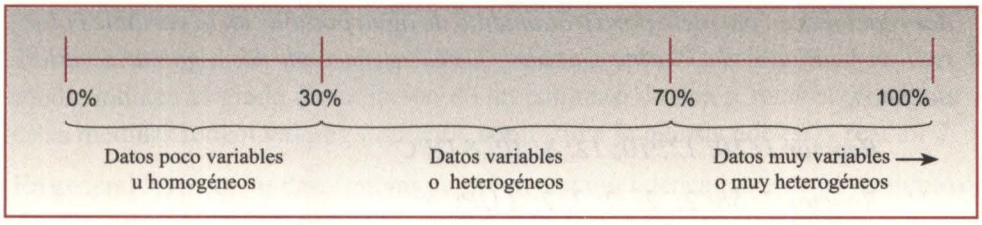
El coeficiente de variación poblacional consiste en expresar la desviación estándar poblacional como un porcentaje de la media poblacional, tal como se presenta en la ecuación 2.24.

$$CV = \frac{\sigma}{\mu} x\ 100\% \qquad (2.24)$$

Siempre que $\mu \neq 0$

Esta medida es adimensional, sus unidades están dadas en porcentaje, por lo tanto es un buen indicador de comparación entre dos o más diferentes variables o dos o más diferentes poblaciones.

Como una guía para su interpretación se puede tomar el siguiente esquema:



De otra forma:

Si $S \leq 0.3\overline{X}$ entonces el conjunto de datos es poco variable u homogéneo con relación a la media.

Si $0.3\overline{X} < S \leq 0.7\overline{X}$ entonces el conjunto de datos es variable o heterogéneo con relación a la media.

Si $S > 0.7\overline{X}$ entonces el conjunto de datos es muy variable o muy heterogéneo con relación a la media.

Ejemplo 2.14 *Considerando la situación del ejemplo 2.1, se calcula el coeficiente de variación:*

Datos del primer muestreo:

5; 4; 5; 4; 8; 10; 9 (UNT) → $CV = \frac{2,5}{6,4} x100\% = 39,1\%$ *(UNT) o* $S = 0,39\,\overline{X}$

Con un dato adicional:

12 (UNT) → $CV = \frac{3}{7,1} x100\% = 42,3\%$ *(UNT) o* $S = 0,42\,\overline{X}$

Con otro dato adicional:

150 (UNT) → $CV = \frac{47,7}{23} x100\% = 207,4\%$ *(UNT) o* $S = 2,07\,\overline{X}$

Con otro dato adicional:

320 (UNT) → $CV = \frac{104,1}{52,7} x100\% = 197,6\%$ *(UNT) o* $S = 1,97\,\overline{X}$

Para el primer conjunto de datos el CV = 39%, indica que los datos presentan variación con relación a la media. A medida que se consideran datos extremos en el muestreo, el CV toma valores de 207% y 197%, que indica una gran variación de los mismos con relación a la media.

El coeficiente de variación aumenta considerablemente a medida que la distancia entre la media y la desviación estándar crecen.

Ejemplo 2.15 *Considerando los datos presentados en el ejemplo 2.8: Se evalúan dos reactores en paralelo para tratamiento de agua potable, en la variable color real en Unidades de Platino Cobalto (UPC), generando las siguientes series de datos:*

Reactor 1: 10; 12; 10; 12; 8; 10; 8 UPC

Reactor 2: 58; 2; 2; 2; 2; 2; 2 UPC

En la Tabla 2.9 se presentan las principales medidas de tendencia central y dispersión para este conjunto de datos.

Tabla 2.9 *Medidas descriptivas para la comparación de dos reactores para potabilización de agua, en color real.*

Medidas descriptivas	Símbolo matemático	Reactor 1	Reactor 2
Media	$\overline{X}$	10 UPC	10 UPC
Mediana	M_e	10 UPC	2 UPC
Media geométrica	$\overline{X}_g$	9,9 UPC	3,2 UPC
Moda	M_0	10 UPC	2 UPC
Rango	R	4 UPC	56 UPC
Varianza	S^2	2,7 UPC^2	448 UPC^2
Desviación estándar	S	1,6 UPC	21,2 UPC
Coeficiente de variación	$C.V.$	16,3 %	211,7 %

A pesar de tener los mismos promedios en color real, los dos reactores presentan eficiencias bastante diferentes, tal como se puede evidenciar en las medidas de dispersión. El rango para el primer reactor es 4 UPC y para el segundo es 56 UPC, lo cual evidencia que en los datos del segundo reactor la distancia entre el valor mínimo y el valor máximo es mucho mayor que la del reactor 1.

La desviación estándar, esto es, el promedio de la distancia de los datos con respecto a la media, es 1,6 UPC para el primer reactor y 21,2 UPC para el segundo reactor. Es decir, los datos se alejan de la media en 1,6 UPC para el primer reactor y se alejan 21,2 UPC para el segundo reactor, lo cual permite concluir que existe una mayor variación en el reactor 2.

El coeficiente de variación es también un buen indicador del grado de variación de los datos en relación con la media; para el reactor 1 es 16,3% y para el reactor 2 es 211,7%. Un CV=16,3% significa que el conjunto de datos es homogéneo para el caso

del reactor 1; sin embargo, un CV= 211,7% significa gran variación o heterogeneidad en el conjunto de datos, para el caso del reactor 2.

También, las medidas de centralidad, como la mediana, la media geométrica y la moda, indican el grado de variación de un conjunto de datos, pues en el reactor 1 estas medidas toman valores similares, contrario a lo que sucede en el reactor 2.

En general las medidas descriptivas permiten resumir adecuadamente un conjunto de datos en medidas de centralidad y medidas de dispersión que permiten caracterizar el fenómeno en estudio. Adicionalmente es necesario estudiar la distribución del conjunto de datos, tal como se desarrolla en el próximo capítulo.