Estadística Descriptiva Aplicada en Python

Para la Investigación Científica en Ciencias Sociales y Educativas

1era Edición

SciELa

Autores:

- Marcelo Bernave Chancusig López
- Guido Euclides Yauli Chicaiza
- Guadalupe de las Mercedes López Castillo
- José Antonio Andrade Valencia
- Jhon Eduardo López Velasco

Descriptive Statistics Applied in Python for Scientific Research in Social and Educational Sciences

ISBN: 978-9942-7221-6-4

Primera Edición, Junio 2024

Estadística Descriptiva Aplicada en Python para la Investigación Científica en Ciencias Sociales y Educativas

ISBN digital: 978-9942-7221-6-4
DOI: https://doi.org/10.62131/978-9942-7221-6-4

Editado por:

Sello editorial:

Quevedo, Los Rios, Ecuador

E-mail: admin@editorial-sciela.org
Código Postal: 120303
WEB: https://editorial-sciela.org

Este libro se sometió a arbitraje bajo el sistema de doble ciego (peer review) y antiplágio. Este producto investigativo cumple con la Declaración de Principios de Budapest, San Francisco, México, Helsinki y Firma del Marco del MIT

Dirección editorial:
Lic. Alexander Fernando Haro, MSI.

Revisor (1):
Ing. Carlota Delgado Vera, Mg.

Revisor (2):
Ing. Julián Coronél Reyes, Mg.

Sistema de clasificación decimal DEWEY

519.5 - Matemáticas estadísticas

Clasificación comercial internacional - THEMA

P - Matemáticas y ciencias
PD - Ciencia: cuestiones generales
PDM - Investigación científica

AUTORES

Marcelo Bernave Chancusig López

https://orcid.org/0009-0006-8794-4232

mbchancusig@espe.edu.ec

Universidad de las Fuerzas Armadas ESPE

Ingeniero en Electrónica Control y Redes Industriales, con una maestría en Electricidad con mención en Sistemas Eléctricos de Potencia. Trabajo como docente en el Departamento de Ciencias Exactas de la Universidad de las Fuerzas Armadas ESPE. He realizado diversas publicaciones en los campos de estadística y educación, aportando a ambos a través de mi investigación y enseñanza.

Marcelo Chancusig

Guido Euclides Yauli Chicaiza

https://orcid.org/0009-0007-3508-678X

guido.yauli@utc.edu.ec

Universidad Técnica de Cotopaxi

Ingeniero Agrónomo, en la Universidad Técnica de Ambato en el año 1994, Master en Ciencias de la Educación, Mención Planeamiento de Instituciones de Educación Superior en la Universidad Técnica de Cotopaxi, Magister en Agronomía Mención Sistemas Agropecuarios en la Universidad Estatal Amazónica; posee 29 años de Experiencia Docente Universitario dictando cátedras como Riegos y Drenajes, Estadística, Metodología de la Investigación, Avalúos y Peritajes Agropecuarios, además ha ocupado cargos Directivos en la Universidad Técnica de Cotopaxi como Decanaturas y Vicerrectorado.

Guido Yauli

Guadalupe de las Mercedes López Castillo

https://orcid.org/0009-0001-9829-0493

guadalupe.lopez@utc.edu.ec

Universidad Técnica de Cotopaxi

Tengo una maestría en Gestión de la Producción y en Agronomia con Mención en Sistemas Agropecuarios. Laboro 27 años en la Universidad Técnica de Cotopaxi, Facultad de Ciencias Agropecuarias y Recursos Naturales, Carrera de Agronomia, como Docente Investigador. Coordino el grupo de Investigación de Cultivos Andinos, soy Investigadora del grupo de investigación de Conservación de suelos. He sido parte del proyecto Generativo denominado Granos Andinos hasta la actualidad. Escrito proyectos Formativos, el que esta vigente es el de Sustento Andino. He sido Directora y lectora de proyectos de titulación de Grado y Posgrado. He coordinado la Maestría de Sanidad como titular y la Maestría de Agroindustrias como encargada. Tengo algunos artículos científicos, he Coordinado y he sido Ponente en algunos eventos Científicos Técnicos de la Universidad.

Guadalupe López

José Antonio Andrade Valencia

https://orcid.org/0000-0003-4289-2855

jose.andrade@utc.edu.ec

Universidad Técnica de Cotopaxi

PhD. En conservación y restauración del medio natural, Magister en Gestión Ambiental, Magíster en seguridad y prevención de riesgos del trabajo, docente investigador, Coordinador del Proyecto “Recuperación de germoplasma de especies vegetales” de la Universidad Técnica de Cotopaxi. Coordinador del programa de Maestría de Gestión Ambiental.

José Andrade

Jhon Eduardo López Velasco

https://orcid.org/0009-0001-7694-2665

academiasuperiordocente18@gmail.com

HIGHER TEACHING ACADEMY
"MARÍA TERESA LÓPEZ VELASCO"

Master en Psicología General,en la Universidad Isabel I de España, Magister en Ciencias de la Educación de la Pontificia Universidad Católica Del Ecuador, Diplomado Superior en Docencia universitaria de la Pontificia Universidad Católica Del Ecuador, licenciado en ciencias de la Educación Universidad Nacional de Chimborazo, Profesor en ciencias de la Educación Instituto Superior Pedagógico Jaime Roldós Aguilera, Especialista en Neurología, psicología del Aprendizaje Universidad Pública del Alto Bolivia, Especialista en Redacción Profesional Enfoque Perú , Máster Instructor Lightouse English Center San Pedro Sula Honduras. Consultor HIGHER TEACHING ACADEMY "MARÍA TERESA LÓPEZ VELASCO"

Jhon López V.

ÍNDICE

Prefacio

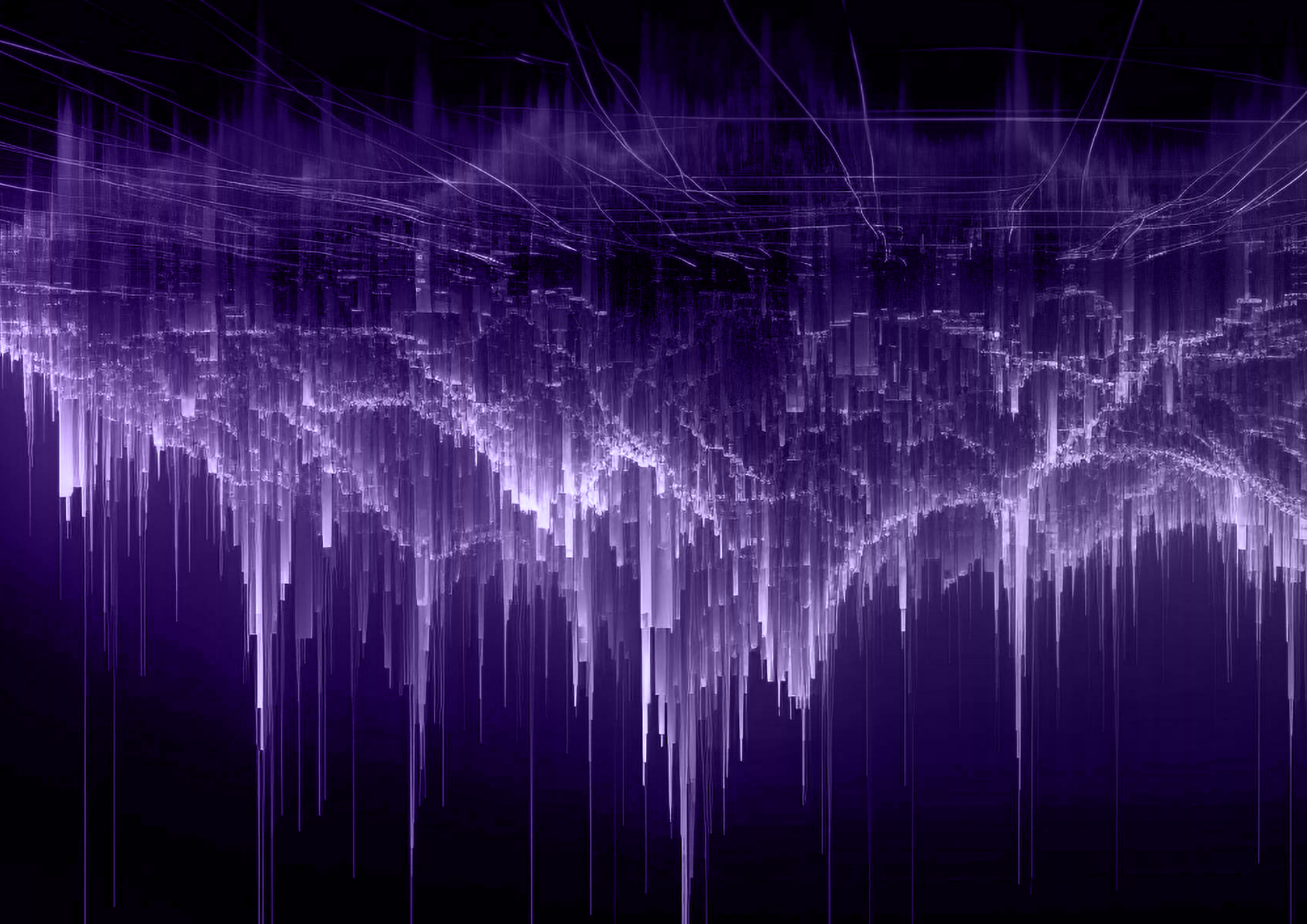
Capítulo I.
Introducción a la Estadística Descriptiva y Python

CAPÍTULO II.
MEDIDAS DE TENDENCIA CENTRAL

CAPÍTULO III.
MEDIDAS DE DISPERSIÓN

Capítulo IV.
Distribuciones de Frecuencias y Tablas

Capítulo V.
Correlación y Covarianza



Prefacio

Datos aplicados a las ciencias sociales y educativas

En la era de la información, el conocimiento es poder. Pero este poder no radica únicamente en la acumulación de datos, sino en la capacidad de analizarlos, interpretarlos y convertirlos en decisiones informadas. Este libro, "Estadística Descriptiva Aplicada en Python para las Ciencias Sociales y Educativas", se concibe como una herramienta fundamental para estudiantes, investigadores y profesionales de las ciencias sociales y educativas que desean adquirir o reforzar habilidades en el análisis estadístico utilizando Python, uno de los lenguajes de programación más populares y poderosos en el campo del análisis de datos.

El propósito de este texto es doble. En primer lugar, aspiramos a proporcionar una comprensión sólida de los principios y métodos de la estadística descriptiva, que son la base para cualquier tipo de análisis estadístico. En segundo lugar, buscamos enseñar cómo aplicar estos conceptos utilizando Python, haciendo que los procesos de análisis sean más eficientes y reproducibles. Abordamos este desafío presentando ejemplos relevantes y actuales que resuenan

con problemas y situaciones típicas en las ciencias sociales y la educación.

Nos enfocaremos en las medidas de tendencia central y dispersión, la construcción de tablas de frecuencias, y la interpretación de los resultados estadísticos a través de la visualización de datos. Asimismo, discutiremos cómo la correlación y la covarianza pueden revelar relaciones interesantes entre variables sociales y educativas.

Python se ha elegido por su simplicidad y potencia. A través de sus bibliotecas especializadas como NumPy, Pandas, Matplotlib y Seaborn, Python facilita la realización de análisis complejos y la visualización de datos de forma accesible para aquellos sin un fondo extenso en programación.

Cada capítulo de este libro está diseñado para construir su conocimiento de manera incremental, comenzando con los fundamentos y avanzando hacia técnicas más sofisticadas de análisis de datos. Además, hemos incluido un subtema de comprensión lectora al final de cada capítulo para reforzar el aprendizaje y asegurar la asimilación del material cubierto.

Este libro es más que un texto académico; es un compañero en su viaje hacia el dominio de las habilidades de análisis de datos en las ciencias sociales y educativas. Ya sea que se encuentre dando sus primeros pasos en estadística descriptiva o buscando profundizar su comprensión existente, "Estadística Descriptiva Aplicada en Python para las Ciencias Sociales y Educativas" está diseñado para ser un recurso valioso que le acompañará en su desarrollo profesional y académico.

Capítulo I.

Introducción a la Estadística Descriptiva y Python

Capítulo I.

Introducción a la Estadística Descriptiva y Python

1. *Fundamentos de la estadística descriptiva*

La estadística descriptiva es una rama fundamental de las matemáticas aplicadas que se enfoca en organizar, resumir y presentar datos de manera informativa y comprensible. Su objetivo principal es describir las características principales de un conjunto de datos, proporcionando una visión general que permita entender su comportamiento y distribución. Para ello, utiliza diversas herramientas como medidas de tendencia central, medidas de dispersión y representaciones gráficas.

Una de las principales herramientas de la estadística descriptiva es la medida de tendencia central, que incluye el cálculo de la media, la mediana y la moda. La media aritmética es el promedio de los valores de un conjunto de datos y proporciona una indicación de su valor central. La mediana es el valor que divide al conjunto de

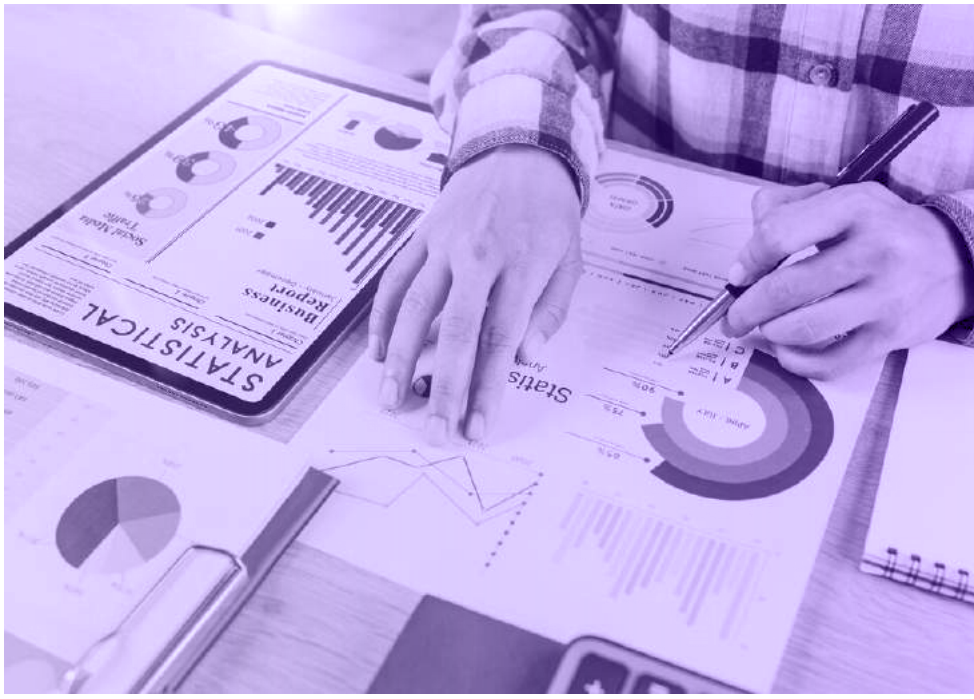
datos en dos partes iguales cuando están ordenados, mientras que la moda representa el valor que aparece con mayor frecuencia en el conjunto de datos.

Además de las medidas de tendencia central, la estadística descriptiva también utiliza medidas de dispersión para entender la variabilidad de los datos. Entre estas medidas se encuentran la desviación estándar, la varianza y el rango intercuartílico. La desviación estándar y la varianza miden la dispersión de los datos con respecto a la media, mientras que el rango intercuartílico proporciona una medida de dispersión que se basa en los cuartiles del conjunto de datos.

Por último, las representaciones gráficas son una herramienta poderosa de la estadística descriptiva para visualizar la distribución de los datos. Algunos ejemplos de estas representaciones son los histogramas, los diagramas de barras, los diagramas de caja y los gráficos de dispersión. Estas representaciones permiten identificar patrones, tendencias y valores atípicos en los datos de manera intuitiva y efectiva. Se detallan fundamentos específicos:

- Descripción y Resumen de Datos: La estadística descriptiva facilita la descripción y resumen de datos, utilizando representaciones gráficas como histogramas, gráficos de caja, o medidas numéricas como tendencias centrales y variabilidad (Cooksey,

2020). Estos métodos ayudan a visualizar y entender los patrones o tendencias dentro de los datos.

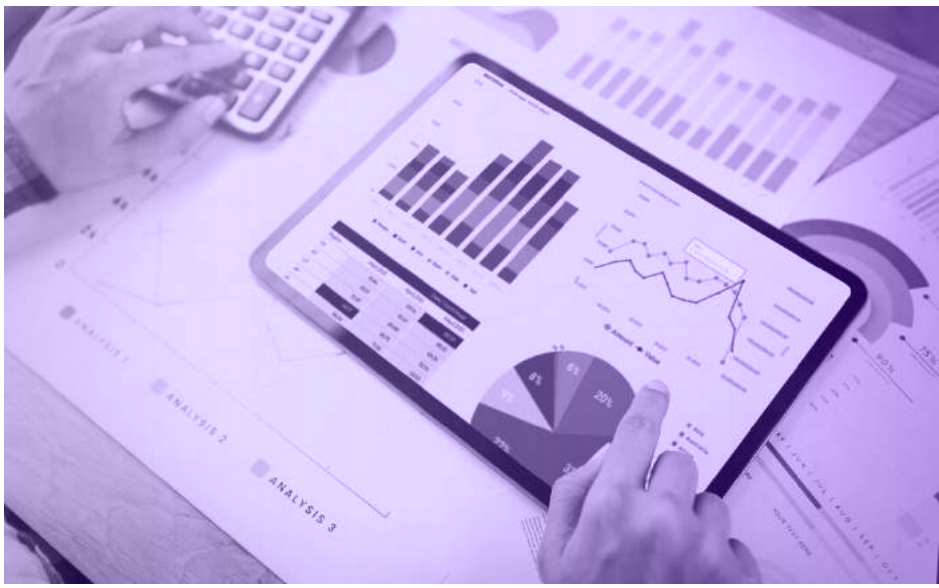


- Medidas de Centralidad y Dispersión: Las medidas comunes incluyen la media, mediana, modo (para tendencia central), y la desviación estándar y rango (para dispersión). Estas medidas son fundamentales para entender las características generales de los conjuntos de datos (Nick, 2007).

- Aplicaciones en Diversos Campos: La estadística descriptiva se utiliza en una amplia gama de aplicaciones, desde el análisis de precios de vivienda hasta estudios sobre el mercado de valores, demostrando su versatilidad y capacidad para facilitar la comprensión de relaciones complejas entre variables (Dong, 2023).

- Importancia en la Investigación: Antes de proceder a análisis estadísticos más complejos, es crucial realizar una descripción estadística para entender la naturaleza básica de los datos. Esto incluye la identificación de valores atípicos, la comprensión de la distribución de los datos y la preparación de los mismos para análisis posteriores (Bryant, 2019).

2. Introducción al lenguaje de programación Python



Python es un lenguaje de programación versátil y poderoso que ha ganado una gran popularidad en la comunidad de desarrollo en los últimos años. Destacado por su sintaxis simple y legible, Python es una excelente opción tanto para principiantes como para programadores experimentados. Su filosofía de diseño, que enfatiza la legibilidad del código y la simplicidad en el desarrollo, lo convierte en una herramienta ideal para una amplia gama de aplicaciones, desde desarrollo web hasta análisis de datos y aprendizaje automático.

Una de las características distintivas de Python es su amplia biblioteca estándar, que proporciona una gran cantidad de módulos y funciones predefinidas que facilitan tareas comunes de programación. Esto incluye módulos para manipulación de archivos, manejo de datos, redes, GUI (interfaz gráfica de usuario), y mucho más. Además, Python cuenta con una comunidad activa y vibrante que contribuye constantemente con paquetes adicionales a través del Python Package Index (PyPI), lo que amplía aún más su funcionalidad y utilidad.

Python es un lenguaje interpretado, lo que significa que no necesita

ser compilado antes de ejecutar el código. Esto facilita el desarrollo rápido y la experimentación, ya que los programas pueden ejecutarse línea por línea en un intérprete interactivo o mediante la ejecución de scripts completos. Además, Python es multiplataforma, lo que significa que puede ejecutarse en una variedad de sistemas operativos, incluyendo Windows, macOS y Linux, lo que lo hace ideal para desarrollar aplicaciones que deben ser ejecutadas en diferentes entornos.

La sintaxis clara y legible de Python, que utiliza espacios en blanco para delimitar bloques de código en lugar de llaves o palabras clave, hace que sea fácil de aprender y entender incluso para aquellos que son nuevos en la programación. Esto lo convierte en una excelente opción para enseñar conceptos fundamentales de programación, así como para prototipar rápidamente ideas y proyectos. Basándonos en fuentes bibliográficas destacamos a los siguientes autores:

- Facilidad de Aprendizaje y Enseñanza: Python es conocido por su sintaxis clara y legible, lo que facilita su aprendizaje y enseñanza. Es ampliamente utilizado como un primer lenguaje de programación en muchos entornos académicos debido a su simplicidad y poderosa funcionalidad. Esta característica lo hace ideal para quienes se inician en la programación, proporcionando

una base sólida en los conceptos fundamentales de la programación (Samimi, 2013).

- Aplicaciones en Computación Científica: A pesar de no ser diseñado específicamente para la computación numérica, Python se ha convertido en una herramienta favorita en este campo gracias a paquetes como NumPy y SciPy, que extienden sus capacidades para el manejo eficiente de operaciones numéricas y científicas. Esto demuestra la adaptabilidad de Python para satisfacer necesidades específicas más allá de sus funcionalidades básicas (Johansson, 2018).

- Uso Interactivo y Modular: Python permite un enfoque modular en el desarrollo de programas, lo cual es útil tanto en la educación como en aplicaciones profesionales. La interactividad de Python, junto con su capacidad de ser extendido a través de módulos y bibliotecas, facilita la experimentación y el desarrollo rápido de aplicaciones. Además, el lenguaje apoya la reutilización de código, lo que permite a los programadores construir aplicaciones complejas con menos esfuerzo (Dierbach, 2014).

3. Configuración del entorno de trabajo en Python

Una de las formas más comunes de configurar un entorno de trabajo es mediante el uso de un entorno virtual. Los entornos virtuales permiten aislar las dependencias y bibliotecas de Python de un proyecto específico, lo que garantiza que cada proyecto tenga su propio conjunto de paquetes sin interferir con otros proyectos en el mismo sistema.

Para crear un entorno virtual en Python, se puede utilizar la herramienta virtualenv o el módulo venv que viene incluido en la instalación estándar de Python a partir de la versión 3.3. Estos permiten crear un entorno virtual en un directorio específico, donde se pueden instalar paquetes y dependencias de manera aislada.

Una vez creado el entorno virtual, se puede activar utilizando comandos específicos dependiendo del sistema operativo. En sistemas Unix y macOS, se utiliza el comando source para activar el entorno virtual, mientras que en Windows se utiliza el comando activate. Una vez activado el entorno virtual, el prompt de comandos cambiará para indicar que el entorno virtual está activo.

Con el entorno virtual activo, se puede instalar las bibliotecas y paquetes necesarios para el proyecto utilizando el gestor de paquetes pip, que es la herramienta estándar para instalar y gestionar paquetes en Python. Se pueden instalar paquetes específicos uti-

lizando el comando pip install, seguido del nombre del paquete.

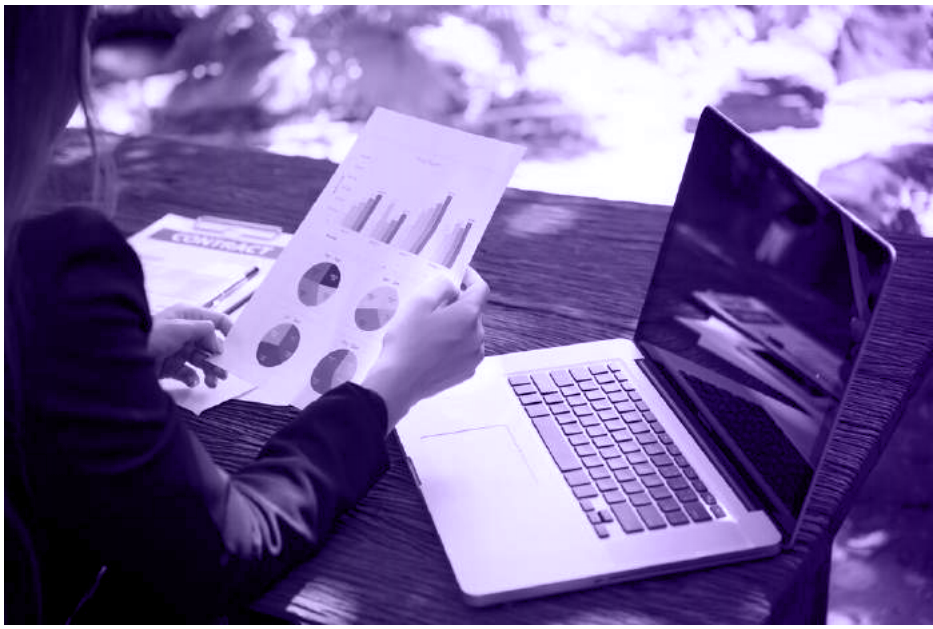
Además de configurar un entorno virtual, también es importante elegir un editor de texto o un entorno de desarrollo integrado (IDE) para escribir código Python. Algunos de los editores de texto populares para Python incluyen Visual Studio Code, Sublime Text y Atom, mientras que algunos IDE populares incluyen PyCharm, Spyder, Colab y Jupyter Notebook. Estos editores e IDEs ofrecen características como resaltado de sintaxis, completado de código, depuración y gestión de proyectos que facilitan el desarrollo en Python.

4. Bibliotecas estadísticas en Python: NumPy y Pandas

NumPy es una biblioteca que ofrece soporte para arrays y matrices grandes y multidimensionales, junto con una colección de funciones matemáticas para operar eficientemente en estos arrays. La eficacia de NumPy proviene de su capacidad para realizar operaciones vectorizadas, lo que reduce la necesidad de bucles explícitos y aprovecha las capacidades de hardware subyacentes para acelerar las operaciones. NumPy no solo es el núcleo de muchas otras bibliotecas de análisis de datos en Python, sino que también es utilizado directamente para implementar funcionalidades matemáticas y es-

tadísticas complejas con un rendimiento optimizado (Nelli, 2015).

Pandas, por otro lado, es una biblioteca diseñada específicamente para la manipulación y el análisis de datos estructurados. Ofrece estructuras de datos como DataFrame y Series, que son esenciales para la manipulación de datos tabulares típicos de muchas aplicaciones de ciencia de datos. Pandas facilita tareas comunes como la carga, manipulación, modelado, y análisis de datos complejos con una sintaxis intuitiva y potente. Además, Pandas se construye sobre NumPy y, por lo tanto, integra bien las capacidades numéricas de NumPy con sus propias capacidades de manipulación de datos estructurados, lo que permite a los usuarios realizar un análisis de datos completo sin necesidad de cambiar entre herramientas (McKinney, 2010).



5. Tipos de datos y escalas de medición

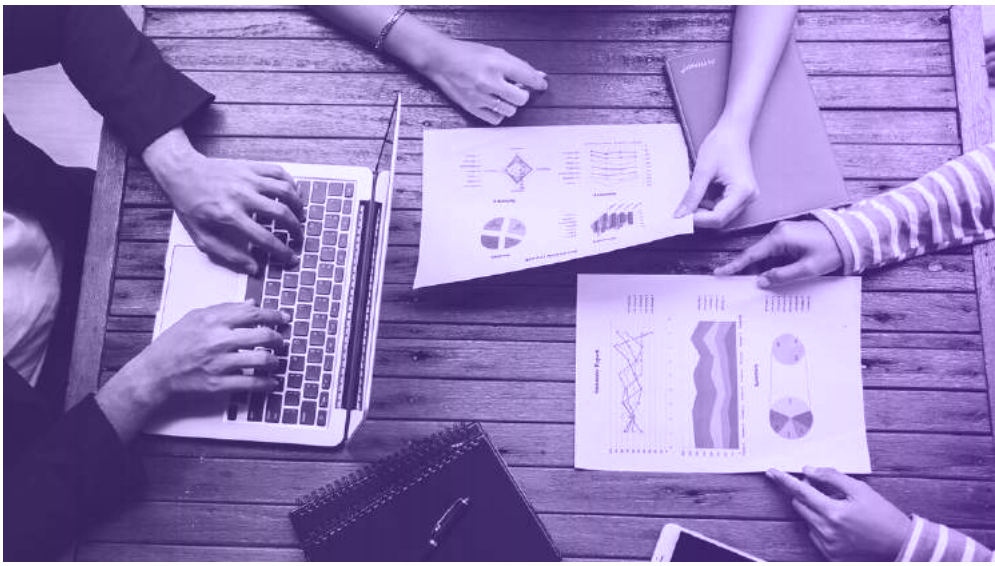
En Python, el tratamiento de los tipos de datos y las escalas de medición es una parte fundamental del análisis de datos y la estadística. Los tipos de datos en Python se pueden dividir en varias categorías, y cada uno se asocia con diferentes escalas de medición:

Tipos de Datos Básicos en Python:

- Booleanos (bool): Verdadero o Falso. Son útiles para pruebas lógicas.
- Numéricos: Incluyen enteros (int), flotantes (float) y complejos (complex).
- Cadenas de texto (str): Para datos textuales.
- Listas (list): Colecciones ordenadas y mutables.
- Tuplas (tuple): Colecciones ordenadas e inmutables.
- Diccionarios (dict): Colecciones no ordenadas de pares clave-valor.
- Conjuntos (set): Colecciones no ordenadas de elementos únicos.

Escala de Medición:

- Nominal: Categorías sin ningún orden inherente. En Python, se suelen representar con cadenas de texto (str) o con conjuntos (set) si se trata de categorías únicas.



- Ordinal: Categorías con un orden o rango específico. A menudo se representan con listas (list) de cadenas de texto, donde el orden de los elementos refleja la jerarquía, o con enteros que representan esa jerarquía.

- Intervalo: Números que tienen sentido en términos de distancia o diferencia, pero no tienen un verdadero cero. Se representan comúnmente con números flotantes (float) y a veces con enteros (int), dependiendo de la naturaleza de los datos.

- Razón: Similar a la escala de intervalo, pero con un cero significativo que permite la comparación de razones. También se representan con números (int o float), pero la interpretación de los ceros es lo que distingue a esta escala.

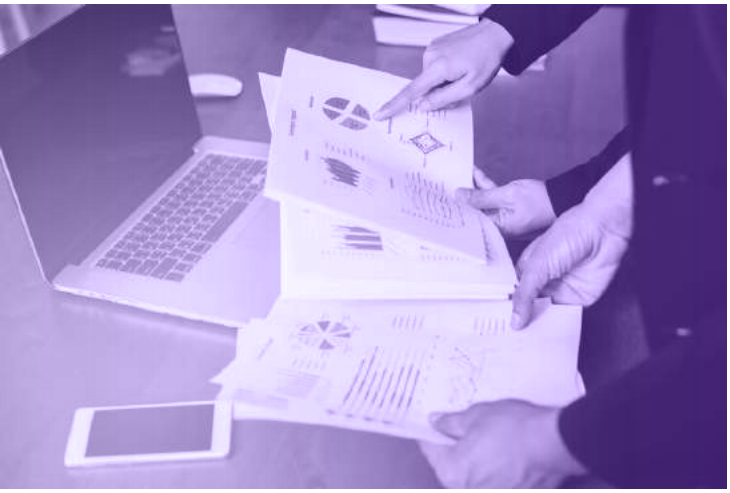
Librerías de Python para Datos y Escalas de Medición:

- NumPy: Proporciona arrays (numpy.array) que pueden contener datos numéricos de tipo entero o flotante, ideales para cálculos matemáticos.

- Pandas: Ofrece estructuras de datos como DataFrame y Series, que son perfectas para trabajar con datos tabulares. Pandas maneja internamente las distintas escalas de medición mediante la asignación de tipos de datos adecuados (dtype).

- SciPy y Statsmodels: Librerías que proporcionan funciones para el análisis estadístico que pueden manejar datos en diferentes escalas.

6. Importación y manejo de datos con Python



La importación y manejo de datos con Python es una tarea fundamental para cualquier proyecto de análisis de datos o ciencia de datos. Python ofrece numerosas herramientas y bibliotecas que facilitan este proceso, desde la lectura de archivos hasta la manipulación y visualización de datos.

Una de las bibliotecas más utilizadas para la importación de datos en Python es Pandas. Pandas proporciona estructuras de datos flexibles y eficientes, como el DataFrame, que permite cargar datos tabulares desde una variedad de fuentes, como archivos CSV, Excel, bases de datos SQL, y más. Para importar datos usando Pandas, primero se debe instalar la biblioteca y luego importarla en el script utilizando el comando import pandas as pd.

Una vez que Pandas está importado, se puede utilizar la función pd.read_csv() para cargar datos desde un archivo CSV, especificando la ruta del archivo como argumento. De manera similar, existen funciones para

leer datos desde otros tipos de archivos, como pd.read_excel() para archivos Excel y pd.read_sql() para bases de datos SQL.

Una vez que los datos están cargados en un DataFrame de Pandas, se pueden realizar una variedad de operaciones de manipulación de datos, como seleccionar columnas específicas, filtrar filas basadas en ciertos criterios, agregar nuevas columnas, y más. Pandas también proporciona métodos para limpiar y transformar datos, como eliminar valores nulos, cambiar tipos de datos y agrupar datos para realizar análisis agregados.

Además de Pandas, Python ofrece otras bibliotecas poderosas para el manejo y manipulación de datos, como NumPy para operaciones numéricas y matriciales, Matplotlib y Seaborn para visualización de datos, y Scikit-learn para aprendizaje automático y análisis predictivo.

```
##Importar un archivo CSV

import pandas as pd

# Cargar un archivo CSV

df_csv = pd.read_csv('ruta/del/archivo.csv')

# Mostrar las primeras filas del DataFrame

print(df_csv.head())
```

```
##Importar un archivo Excel
# Cargar un archivo Excel
df_excel = pd.read_excel('ruta/del/archivo.xlsx')
# Mostrar las primeras filas del DataFrame
print(df_excel.head())
```

```
##Importar un archivo JSON
# Cargar un archivo JSON
df_json = pd.read_json('ruta/del/archivo.json')
# Mostrar las primeras filas del DataFrame
print(df_json.head())
```

```
##Importar un URL
# Cargar datos desde una URL
url = 'http://ejemplo.com/datos.csv'
df_url = pd.read_csv(url)
```

```
# Mostrar las primeras filas del DataFrame

print(df_url.head())
```

```
##Leer desde una base de datos SQL

from sqlalchemy import create_engine

# Crear conexión con la base de datos

engine = create_engine('sqlite:///ruta/de/la/base_de_datos.
db')

# Leer los datos de una tabla específica

df_sql = pd.read_sql_table('nombre_de_la_tabla', engine)

# Mostrar las primeras filas del DataFrame

print(df_sql.head())
```

```
##Importar un archivo de texto delimitado por tabulaciones
(TSV)

# Cargar un archivo TSV

df_tsv = pd.read_csv('ruta/del/archivo.tsv', delimiter='\t')
```

```
# Mostrar las primeras filas del DataFrame

print(df_tsv.head())
```

7. Ética en el análisis de datos

La ética en el análisis de datos es un tema crucial dado el creciente acceso a grandes cantidades de información y la capacidad de analizar estos datos para influir en decisiones importantes. Los aspectos éticos del análisis de datos se extienden desde la recopilación de datos hasta su manipulación y los efectos que estos tienen sobre las personas y la sociedad.

Uno de los principales desafíos éticos en el análisis de datos es el manejo adecuado de la información personal y la protección de la privacidad. La capacidad de los analistas de datos para extraer información detallada sobre individuos plantea preocupaciones significativas sobre la privacidad y el consentimiento. Los individuos a menudo no son conscientes de cómo se recopilan, almacenan, analizan y utilizan sus datos. Las prácticas responsables requieren que los analistas se aseguren de que los datos se manejen de manera transparente y con el consentimiento informado de las personas implicadas (Herschel & Miori, 2017).

Además, la integridad de los análisis de datos es fundamental para mantener la confianza en las investigaciones y conclusiones derivadas de grandes conjuntos de datos. La manipulación de datos para obtener resultados específicos, conocida como p-hacking, y otras prácticas cuestionables como no reportar todos los resultados de

los análisis para destacar solo aquellos que apoyan una hipótesis previa, socavan la objetividad del proceso de investigación. Estas prácticas pueden llevar a conclusiones erróneas y potencialmente perjudiciales.

Por último, es esencial considerar las implicaciones a largo plazo del análisis de datos, incluyendo cómo los resultados pueden afectar a diferentes grupos dentro de la sociedad. Por ejemplo, los algoritmos de decisión automatizados pueden perpetuar o incluso exacerbar desigualdades existentes si no se diseñan con una consideración cuidadosa de sus impactos sociales. Los analistas de datos y los científicos tienen la responsabilidad de evaluar críticamente estos aspectos y trabajar hacia métodos que promuevan la equidad y la justicia (Floridi & Taddeo, 2016).

8. Comprensión lectora

¿Cuál es el objetivo principal de la estadística descriptiva?

A. Inferir propiedades de una población

B. Describir las características principales de un conjunto de datos

C. Predecir futuros resultados basados en datos históricos

D. Validar hipótesis estadísticas

ANSWER: **B**

¿Qué representa la mediana en un conjunto de datos?

A. El valor más frecuente

B. El valor mínimo

C. El valor que divide al conjunto en dos partes iguales

D. El promedio de los valores

ANSWER: **C**

¿Cuál de las siguientes NO es una medida de dispersión utilizada en estadística descriptiva?

A. Varianza

B. Mediana

C. Desviación estándar

D. Rango intercuartílico

ANSWER: **B**

¿Qué tipo de representación gráfica no fue mencionada como herramienta en la estadística descriptiva?

A. Histogramas

B. Diagramas de caja

C. Mapas de calor

D. Gráficos de dispersión

ANSWER: **C**

¿Cuál de las siguientes afirmaciones describe mejor Python?

A. Es un lenguaje compilado que requiere ejecución línea por línea.

B. Es conocido por su complejidad y sintaxis confusa.

C. Es un lenguaje de programación que enfatiza la legibilidad del código.

D. Es un lenguaje que no soporta la manipulación de datos.

ANSWER: **C**

¿Qué herramienta de Python se utiliza para crear un entorno virtual?

A. pip

B. virtualenv o venv

C. PyPI

D. NumPy

ANSWER: **B**

¿Qué es NumPy?

A. Una biblioteca para el desarrollo web

B. Una biblioteca para análisis textual

C. Una biblioteca que soporta arrays y matrices grandes y multidimensionales

D. Un entorno de desarrollo integrado (IDE)

ANSWER: **C**

¿Cuál de las siguientes es una estructura de datos proporcionada por Pandas?

A. Arrays

B. DataFrame

C. Booleanos

D. Funciones

ANSWER: **B**

¿Qué función de Pandas permite leer datos de un archivo CSV?

A. pd.read_csv()

B. pd.read_excel()

C. pd.read_sql()

D. pd.DataFrame()

ANSWER: **A**

¿Cuál es una escala de medición que no tiene un verdadero cero?

A. Nominal

B. Ordinal

C. Intervalo

D. Razón

ANSWER: **C**

¿Qué tipo de datos se representa comúnmente con cadenas de texto en Python?

A. Datos numéricos

B. Datos booleanos

C. Datos nominales

D. Datos de intervalo

ANSWER: **C**

¿Cuál es una práctica ética importante en el análisis de datos?

A. Manipulación de datos para obtener resultados específicos

B. Uso de datos sin el consentimiento de los individuos

C. Protección de la privacidad y manejo transparente de datos

D. Limitar el análisis a pequeños conjuntos de datos

ANSWER: **C**

¿Qué afirmación es falsa respecto a la estadística descriptiva?

A. Utiliza el análisis predictivo como su principal herramienta.

B. Se centra en organizar y resumir datos.

C. Desarrolla representaciones gráficas de los datos.

D. Proporciona medidas de tendencia central.

ANSWER: **A**

¿Cuál de las siguientes herramientas no está relacionada con Python?

A. virtualenv

B. Pandas

C. PyCharm

D. HTML5

ANSWER: **D**

¿Qué librería de Python se utiliza principalmente para la manipulación y análisis de datos tabulares?

A. SciPy

B. Matplotlib

C. Pandas

D. Flask

ANSWER: **C**

Capítulo II.

Medidas de Tendencia Central

Capítulo II.

Medidas de Tendencia Central

1. *Concepto de tendencia central*

El concepto de tendencia central en estadística se refiere a la medida que busca representar el valor típico o central de un conjunto de datos. En otras palabras, es una forma de resumir la distribución de los datos en un solo valor que sea representativo de la ubicación central de los mismos.

Las medidas de tendencia central más comunes son la media, la mediana y la moda. La media aritmética es el promedio de todos los valores del conjunto de datos y se calcula sumando todos los valores y dividiendo el resultado entre el número total de observaciones. Es sensible a los valores extremos y puede no ser representativa si hay datos atípicos.

La mediana es el valor que ocupa la posición central cuando los

datos están ordenados de manera ascendente o descendente. Es menos sensible a los valores atípicos que la media y proporciona una indicación más robusta de la ubicación central de los datos.

La moda es el valor que aparece con mayor frecuencia en el conjunto de datos. Puede haber una moda (unimodal) o más de una moda (multimodal). Es útil principalmente para datos cualitativos o discretos.

Estas medidas de tendencia central proporcionan información importante sobre la distribución y la forma de los datos, lo que ayuda a comprender mejor su estructura y a tomar decisiones informadas en el análisis estadístico. Sin embargo, es importante considerar el contexto y las características específicas del conjunto de datos al elegir la medida de tendencia central más apropiada.

2. Calculando la media aritmética en Python

Para calcular la media aritmética en Python, uno puede emplear una lista de números y utilizar funciones básicas del lenguaje. Se proporciona un ejemplo detallado de cómo realizar este cálculo.

```
#Sin librerias
numeros = [5, 10, 15, 20, 25]
suma_total = sum(numeros)
cantidad_numeros = len(numeros)
media_aritmetica = suma_total / cantidad_numeros
```

```
#Con librería Numpy
import numpy as np
numeros_np = np.array(numeros)
media_aritmetica_np = np.mean(numeros_np)
print("La media aritmética es:", media_aritmetica_np)
```

3. Mediana y su significado en distribuciones sesgadas

La mediana es una medida de tendencia central que describe el valor que se encuentra justo en el medio de un conjunto de datos cuando estos están ordenados en secuencia. Para encontrar la mediana, se ordenan los datos y se selecciona el número que divide el conjunto en dos mitades iguales. Si el número total de observaciones es impar, la mediana es el valor central. Si es par, la mediana es el promedio de los dos valores centrales.

En distribuciones sesgadas, la mediana puede ser especialmente significativa como medida de tendencia central. En contraste con la media, que puede ser influenciada fuertemente por valores extremos o atípicos (outliers), la mediana es más resistente a estos valores y puede proporcionar una mejor representación del "centro" de los datos.

- Resistencia a Valores Atípicos: En una distribución con un sesgo

significativo, ya sea hacia la derecha (sesgo positivo) o hacia la izquierda (sesgo negativo), los valores extremos pueden distorsionar la media, desplazándola hacia el lado del sesgo. La mediana, al ser el valor medio, no se ve afectada de la misma manera por estos extremos. Por ejemplo, en un conjunto de ingresos donde unos pocos individuos ganan cantidades exorbitantemente altas, la mediana reflejaría mejor el ingreso típico que la media.

- Representación Más Fiable del Centro: En distribuciones sesgadas, la mediana a menudo se considera una representación más fiable del centro de la distribución que la media. Esto se debe a que indica el punto bajo el cual cae el 50% de los datos, proporcionando una visión clara de la distribución general sin la influencia de los extremos.

- Uso en Descripciones de Datos Reales: La mediana es frecuentemente utilizada en informes estadísticos sobre ingresos, precios de vivienda, y otros datos económicos y sociales, donde la asimetría es común. Esto ayuda a evitar conclusiones erróneas que podrían surgir de simplemente reportar la media.

La mediana en Python, se puede hacerlo fácilmente utilizando nuevamente la biblioteca numpy. ejemplo:

```
import numpy as np
datos = [10, 20, 30, 40, 500]
datos_np = np.array(datos)
mediana_np = np.median(datos_np)
print("La mediana es:", mediana_np)
```

4. Moda y su utilidad en datos categóricos

La moda es la medida de tendencia central que identifica el valor o valores que aparecen con mayor frecuencia en un conjunto de datos. A diferencia de la media y la mediana, la moda puede ser utilizada con datos tanto numéricos como categóricos, lo que la hace especialmente útil en análisis de datos cualitativos.

- Identificación de Tendencias Comunes: En datos categóricos, la moda permite identificar la categoría más común o popular. Por ejemplo, si se recopilan datos sobre las marcas de automóviles preferidas en una encuesta, la moda indicará la marca que más gente prefiere.

- Análisis de Frecuencias: La moda es útil para entender la distribución de las frecuencias en datos categóricos. Esto puede ser importante en estudios de mercado, encuestas de opinión pública, y otros campos donde saber qué categoría es más común puede influir en decisiones comerciales o políticas.

- Simplicidad y Aplicabilidad: Calcular la moda no requiere de operaciones matemáticas complejas y puede aplicarse a cualquier tipo de dato que pueda ser clasificado en categorías distintas, incluyendo variables nominales y ordinales.

- Comparación entre Grupos: La moda puede ayudar a comparar la prevalencia de categorías entre diferentes grupos o segmentos. Por ejemplo, podría usarse para comparar las preferencias de producto entre diferentes demografías dentro de un estudio de consumidores.

Para calcular la moda en Python, se puede utilizar la biblioteca

statistics, que es parte de la librería estándar de Python, o scipy, que proporciona herramientas más robustas para el análisis estadístico:

```
from statistics import mode

colores = ["rojo", "azul", "azul", "verde", "rojo", "rojo", "verde"]

moda_colores = mode(colores)

print("La moda es:", moda_colores)
```

5. Comparación de medidas: media, mediana y moda

La media, mediana y moda son medidas de tendencia central que se utilizan para resumir datos con un solo valor que representa el centro de la distribución. Cada una tiene sus propias características y aplicaciones, dependiendo de la naturaleza y distribución de los datos. Aquí se realiza una comparación de estas tres medidas:

Media

- Definición: Es el promedio de un conjunto de valores, calculado como la suma de todos los valores dividida por el número de valores.
- Utilidad: Es útil para datos cuantitativos que están uniformemente distribuidos sin valores atípicos.

- Sensibilidad: Está muy influenciada por valores extremos. Un solo valor atípico puede alterar significativamente la media.
- Aplicaciones: Se usa ampliamente en análisis financieros, investigación científica, y siempre que los datos sean relativamente simétricos.

Mediana

- Definición: Es el valor que divide un conjunto de datos ordenados en dos partes iguales. Si el número de observaciones es impar, es el valor central; si es par, es el promedio de los dos valores centrales.
- Utilidad: Es más representativa para distribuciones asimétricas o cuando los datos incluyen valores atípicos.
- Sensibilidad: Menos sensible a valores atípicos en comparación con la media.
- Aplicaciones: Frecuentemente utilizada en la descripción de ingresos, precios de bienes raíces, y otros datos que pueden ser sesgados o tener valores atípicos.

Moda

- Definición: Es el valor o valores que aparecen con mayor frecuencia en un conjunto de datos.
- Utilidad: Es la única medida de tendencia central que se puede usar con datos categóricos, además de ser útil para datos numéricos.
- Sensibilidad: No se ve afectada por valores atípicos, pero puede

no ser única; un conjunto de datos puede tener más de una moda o no tener ninguna.

- Aplicaciones: Útil en el análisis de datos cualitativos como encuestas, preferencias de consumidores y cualquier situación donde se necesite identificar el valor más común.

Comparación en la Práctica

Cuando se compara la utilidad de estas medidas, la elección depende del tipo de datos y del propósito del análisis. Por ejemplo:

- Datos con Valores Atípicos: La mediana es generalmente preferida porque refleja mejor el centro de la distribución sin ser distorsionada por extremos.
- Datos Categóricos: La moda es la medida aplicable ya que permite identificar la categoría más frecuente.
- Datos Simétricos y Sin Atípicos: La media es adecuada y proporciona una medida precisa del centro.

Ejemplo Ilustrativo

Supongamos un conjunto de datos de salarios en una empresa pequeña: [35,000, 37,000, 40,000, 41,000, 500,000]. Aquí, la media se vería fuertemente influenciada por el valor atípico (500,000), mientras que la mediana, que sería 40,500, proporcionaría una mejor idea del salario típico. La moda en este caso no sería informativa, ya que todos los salarios, excepto el atípico, son únicos.

En conclusión, la elección entre media, mediana y moda debe basarse en el análisis del conjunto de datos específico y en la naturaleza de la pregunta estadística que se desea responder.

6. Aplicaciones en Ciencias Sociales y Educativas

Imaginemos que tenemos un conjunto de datos recopilados de una encuesta realizada a estudiantes universitarios, donde se evaluaron diversos aspectos de la enseñanza en línea. Los datos incluyen calificaciones en una escala de 1 a 10 para varios ítems como claridad de explicaciones, interacción con el profesorado y apoyo técnico.

Variables

- Clarity: Claridad de las explicaciones en las clases en línea.
- Interaction: Interacción con el profesorado.
- Support: Apoyo técnico durante las clases.

Objetivo

Nuestro objetivo será calcular las medidas de tendencia central (media, mediana y moda) para cada variable y discutir qué nos indican sobre la percepción estudiantil.

Primero, vamos a simular algunos datos para este ejemplo y luego aplicaremos las medidas de tendencia central.

```
import numpy as np

import pandas as pd

from scipy import stats
```

```
np.random.seed(42)  # Para reproducibilidad
data = {
    "Clarity": np.random.randint(1, 11, 100),
    "Interaction": np.random.randint(1, 11, 100),
    "Support": np.random.randint(1, 11, 100)
}
df = pd.DataFrame(data)
# Cálculos de tendencia central
means = df.mean()
medians = df.median()
modes = df.mode().loc[0]
# Resultados
print("Medias:\n", means)
print("Medianas:\n", medians)
print("Modas:\n", modes)
```

Análisis de los resultados

- Media: Nos proporciona el promedio de las respuestas, indicando una visión general del nivel de satisfacción.
- Mediana: Al ser menos sensible a valores extremos, la mediana

puede ofrecer una mejor representación de la tendencia central si la distribución de los datos es muy sesgada.

- Moda: Nos indica el valor que más frecuentemente fue seleccionado por los estudiantes, lo que puede ser útil para identificar las percepciones más comunes.

Con los resultados obtenidos, podemos analizar si hay áreas específicas donde los estudiantes se sienten más satisfechos o insatisfechos. Por ejemplo, si la media de la interacción es significativamente más baja que las otras, eso podría indicar un área de mejora necesaria en la enseñanza en línea.

Este análisis inicial puede ser profundizado con técnicas estadísticas adicionales y utilizarse para informar decisiones sobre mejoras en los métodos de enseñanza. También se puede expandir para incluir otras medidas estadísticas como desviación estándar, análisis de correlaciones, o incluso análisis multivariables para entender mejor las dinámicas entre diferentes aspectos evaluados.

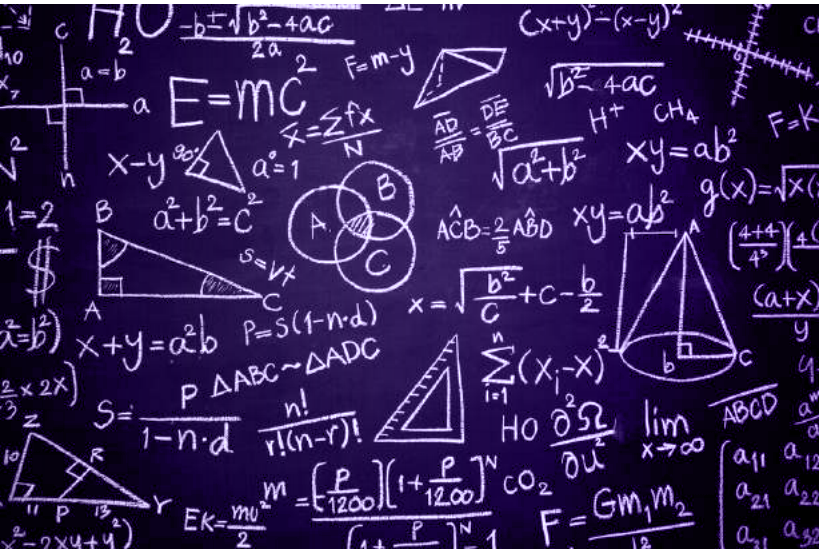
7. Problemas comunes y cómo solucionarlos

Datos Faltantes

- Problema: La falta de respuestas en algunas variables puede distorsionar las medidas de tendencia central.

- Solución: Se pueden imputar los valores faltantes utilizando diferentes técnicas, como la imputación por la media, la mediana o métodos más sofisticados como la imputación múltiple, dependiendo de la naturaleza de los datos.

Valores Atípicos (Outliers)



- Problema: Valores extremadamente altos o bajos pueden sesgar las medias y no representar adecuadamente el conjunto de datos.

- Solución: Identificar y tratar los outliers. Una opción es eliminarlos o sustituirlos, aunque una mejor práctica podría ser utilizar la mediana o la moda, que son menos sensibles a los outliers.

Distribuciones Sesgadas

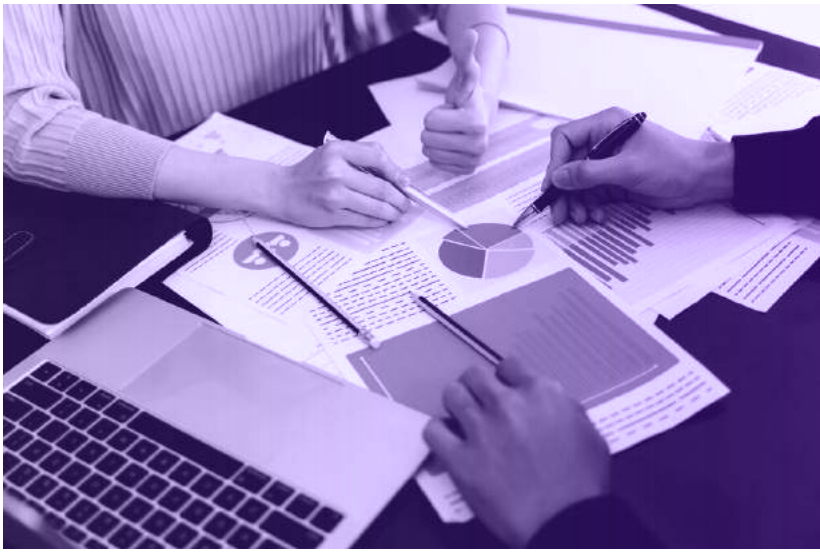
- Problema: Las distribuciones no simétricas pueden hacer que la media no refleje el centro de los datos.

- Solución: Utilizar la mediana como una medida más robusta de tendencia central, o aplicar transformaciones a los datos para normalizar la distribución antes de analizarlos.

Tamaño de la Muestra Inadecuado

- Problema: Muestras muy pequeñas pueden no ser representativas y pueden llevar a estimaciones inestables de la media.

- Solución: Aumentar el tamaño de la muestra, si es posible, o utilizar técnicas de estimación que sean menos sensibles al tamaño de la muestra, como métodos de bootstrapping para estimar la variabilidad de la media.



Errores de Codificación de Datos

- Problema: Errores en la entrada de datos pueden resultar en cálculos incorrectos de las medidas de tendencia central.

- Solución: Realizar una limpieza y validación de datos exhaustiva antes del análisis. Automatizar este proceso con scripts de Python puede ayudar a detectar errores comunes como valores no válidos o inconsistencias.

Dependencia entre Observaciones

- Problema: En ciencias sociales, las respuestas de los participantes pueden estar influenciadas por las de otros (por ejemplo, en encuestas grupales), lo que podría sesgar las medidas.

- Solución: Analizar la estructura de los datos para identificar y

ajustar por cualquier dependencia, usando técnicas estadísticas adecuadas como modelos jerárquicos.

Uso Incorrecto de Medidas

- Problema: Aplicar incorrectamente las medidas de tendencia central sin considerar la naturaleza de los datos, como usar la media en datos nominales.
- Solución: Elegir la medida de tendencia central correcta según el nivel de medida de los datos: moda para datos nominales, mediana para datos ordinales y media para datos de intervalo o de razón.

8. Comprensión lectora

¿Cuál es una medida de tendencia central que es particularmente sensible a los valores extremos?

A. Mediana

B. Moda

C. Media

D. Rango

ANSWER: C

¿Qué medida de tendencia central es el valor que divide un conjunto de datos en dos partes iguales?

A. Media

B. Mediana

C. Moda

D. Varianza

ANSWER: **B**

¿Cuál de las siguientes afirmaciones es verdadera sobre la moda?

A. Es siempre numérica.

B. No puede ser utilizada con datos categóricos.

C. Puede haber más de una en un conjunto de datos.

D. Es el valor más bajo en un conjunto de datos.

ANSWER: **C**

Si un conjunto de datos es multimodal, ¿qué significa esto?

A. Tiene múltiples medias.

B. No tiene una moda definida.

C. Tiene múltiples modas.

D. Está distribuido uniformemente.

ANSWER: **C**

¿Qué medida de tendencia central es más útil en distribuciones sesgadas?

A. Media

B. Mediana

C. Moda

D. Desviación estándar

ANSWER: **B**

En Python, ¿cómo se calcula la media aritmética usando la librería NumPy?

A. np.median(numeros_np)

B. np.mean(numeros_np)

C. np.mode(numeros_np)

D. np.sum(numeros_np)

ANSWER: **B**

¿Cuál de las siguientes NO es una ventaja de la mediana sobre la media?

A. Es menos afectada por valores atípicos.

B. Es fácil de calcular en conjuntos de datos grandes.

C. Representa mejor el centro en distribuciones sesgadas.

D. Es la única medida de tendencia central aplicable a datos ordinales.

ANSWER: **D**

¿Qué medida de tendencia central es especialmente relevante para datos cualitativos o categóricos?

A. Media

B. Mediana

C. Moda

D. Varianza

ANSWER: **C**

Cuando los datos están ordenados, ¿qué representa la mediana?

A. El promedio de los datos

B. El valor más común

C. El valor central

D. La suma de todos los valores

ANSWER: **C**

¿Qué propiedad de la media se discute en el contexto de los valores extremos?

A. Resistencia

B. Sensibilidad

C. Variabilidad

D. Inmutabilidad

ANSWER: **B**

¿Cuál es una característica clave de la moda en análisis estadístico?

A. Puede ser calculada solo para datos numéricos.

B. Indica el valor que más frecuentemente aparece en un conjunto de datos.

C. Es menos útil en el análisis de datos cualitativos.

D. Siempre se encuentra en el centro de un conjunto de datos.

ANSWER: **B**

¿Cómo se denomina una distribución con más de una moda?

A. Unimodal

B. Bimodal

C. Trimodal

D. Multimodal

ANSWER: **D**

¿Qué problema puede afectar a la mediana cuando se trata de grandes conjuntos de datos?

A. Es más complicada de calcular que la moda.

B. Puede cambiar drásticamente con la adición de más datos.

C. No es afectada por valores atípicos.

D. Puede ser difícil determinarla con precisión en conjuntos de datos pares.

ANSWER: **D**

¿En qué escenario es menos adecuado utilizar la media como medida de tendencia central?

A. Cuando los datos están uniformemente distribuidos.

B. Cuando hay valores extremos en el conjunto de datos.

C. Cuando todos los valores son muy similares.

D. Cuando los datos son de una escala nominal.

ANSWER: **B**

¿Qué medida de tendencia central se debe preferir en encuestas de opinión donde las respuestas son categorías como "Satisfactorio", "Neutral" y "Insatisfactorio"?

A. Media

B. Mediana

C. Moda

D. Variabilidad

ANSWER: **C**

Capítulo III.

Medidas de Dispersión

Capítulo III.

Medidas de Dispersión

1. Importancia de la dispersión en estadística

La dispersión nos da una idea de cuán "extendidos" están los datos. Por ejemplo, dos conjuntos de datos pueden tener la misma media, pero uno puede consistir en valores que están muy cerca de la media, mientras que el otro puede tener valores que están muy dispersos. Este concepto se cuantifica a través de medidas como el rango, la desviación estándar, y la varianza.

El rango es la diferencia entre el valor máximo y mínimo en un conjunto de datos, ofreciendo una visión rápida de la dispersión. Sin embargo, como solo considera los valores extremos, puede no dar una imagen completa, especialmente en conjuntos de datos grandes o con valores atípicos. La desviación estándar y la varianza, que se basan en las diferencias cuadráticas de todos los valores respecto a la media, proporcionan una medida más robusta de la

dispersión interna de los datos (Bhardwaj & Sharma, 2013).

En el mundo de los negocios y la economía, la dispersión puede influir en decisiones críticas. Por ejemplo, un inversor puede mirar la varianza de los retornos de una acción para evaluar su riesgo. Una alta dispersión indica una mayor volatilidad, lo que podría disuadir a los inversores que prefieren estabilidad.

En las ciencias de la salud, la dispersión puede indicar la variabilidad en las respuestas al tratamiento entre diferentes pacientes. Un tratamiento que muestra poca dispersión en los resultados probablemente será más confiable y predecible en su efectividad.

Una comprensión profunda de la dispersión también es esencial para realizar pruebas estadísticas y construir modelos predictivos. En los análisis de regresión, por ejemplo, una alta dispersión en las variables puede afectar la precisión de las estimaciones de los coeficientes y hacer más difícil identificar relaciones significativas entre variables.

La dispersión no solo ayuda a comprender la variabilidad, sino que también influye en la robustez de las conclusiones estadísticas. Las pruebas de hipótesis, que a menudo dependen de suposiciones sobre la dispersión (como la homogeneidad de varianzas), pueden llevar a conclusiones erróneas si se ignoran las propiedades de dispersión de los datos.

Aunque las medidas de dispersión son herramientas valiosas, también presentan desafíos. Por ejemplo, la presencia de valores atípicos puede exagerar la dispersión percibida, llevando a interpretaciones engañosas. Además, en distribuciones no normales, las medidas estándar de dispersión como la varianza y la desviación estándar pueden no ser adecuadas, y pueden requerirse métodos más sofisticados para una correcta interpretación (Anghelache et al., 2016).

2. Rango y rango intercuartílico en Python

El rango y el rango intercuartílico (IQR) son herramientas valiosas en la investigación científica, especialmente en las ciencias sociales y educación. En estas disciplinas, los investigadores a menudo enfrentan conjuntos de datos con características particulares como distribuciones no normales, valores extremos y variables que pueden ser categóricas u ordinales, además de cuantitativas. Estas medidas de dispersión ayudan a entender mejor las características de los datos y a tomar decisiones informadas sobre los métodos analíticos y las interpretaciones de los resultados.

Aplicación en Ciencias Sociales y Educación:

- Contexto: Cuando se evalúan programas educativos, es común analizar cuestionarios de satisfacción o desempeño académico de

los estudiantes.

- Rango: Puede mostrar la variabilidad total en las respuestas de los estudiantes, indicando qué tan divergentes pueden ser las opiniones sobre la eficacia del programa.
- IQR: Ofrece una mirada más detallada a donde se concentra la mayoría de las respuestas, ayudando a identificar si la mayoría de los estudiantes se sienten moderadamente satisfechos o no, independientemente de unos pocos valores extremadamente altos o bajos.

Estudios de Desigualdad Social:

- Contexto: Investigaciones sobre ingresos, acceso a recursos educativos o salud entre diferentes grupos socioeconómicos o regiones geográficas.
- Rango: Proporciona una vista rápida de la disparidad total dentro de un grupo o entre grupos.
- IQR: Ayuda a comprender la disparidad dentro del rango medio de la población, potencialmente indicando desigualdades menos extremas pero más comunes.

Análisis de Encuestas sobre Actitudes y Comportamiento:

- Contexto: Encuestas sobre actitudes hacia temas sociales, políticos o comportamientos educativos.
- Rango: Revela los extremos de las opiniones o comportamientos estudiados, mostrando la diversidad de respuestas.
- IQR: Centra el análisis en la mayoría de las respuestas, lo que

puede ser crucial para diseñar políticas o programas que apunten a las actitudes o comportamientos más comunes.

Supongamos que un investigador está evaluando la eficacia de un nuevo método pedagógico en la enseñanza de matemáticas en escuelas secundarias. Las evaluaciones se recogen a través de cuestionarios que califican la satisfacción de los estudiantes con el método en una escala del 1 al 10.

```
import pandas as pd

# Simular datos de evaluaciones de satisfacción

data = {

    "Satisfaction": [4, 5, 5, 6, 7, 7, 8, 9, 2, 10, 5, 6, 7, 8, 9, 4, 4, 5, 6, 7]

}

df = pd.DataFrame(data)

# Calculando el rango

data_range = df['Satisfaction'].max() - df['Satisfaction'].min()
```

```
# Calculando el IQR
Q1 = df['Satisfaction'].quantile(0.25)
Q3 = df['Satisfaction'].quantile(0.75)
IQR = Q3 - Q1
print("Rango de Satisfacción:", data_range)
print("Rango Intercuartílico de Satisfacción:", IQR)
```

3. Varianza y desviación estándar: cálculo e interpretación

La varianza y la desviación estándar son dos de las medidas estadísticas más importantes para entender la dispersión o la variabilidad en un conjunto de datos, siendo especialmente útiles en la investigación científica en ciencias sociales y educación. Estas medidas proporcionan información crítica sobre cómo los datos se distribuyen alrededor de la media, lo que puede ser fundamental para interpretar encuestas, evaluaciones y otros tipos de datos relacionados con comportamientos y percepciones humanas.

La varianza mide cuán dispersos están los datos respecto a la media. Una varianza alta indica que los datos están muy esparcidos, mientras que una varianza baja sugiere que los datos están más agrupados cerca de la media.

La desviación estándar es la raíz cuadrada de la varianza y proporciona una medida de dispersión que está en las mismas unidades

que los datos originales, facilitando la interpretación.

Evaluaciones de Programas Educativos

- Contexto: Al evaluar la efectividad de diferentes métodos de enseñanza, las puntuaciones de los estudiantes en pruebas o cuestionarios pueden ser analizadas para determinar la consistencia de un método en diferentes contextos o poblaciones.
- Varianza/Desviación Estándar Alta: Indica que hay una gran variabilidad en cómo diferentes estudiantes responden al método, lo que podría sugerir que el método no es uniformemente efectivo para todos los estudiantes.
- Varianza/Desviación Estándar Baja: Sugeriría que la mayoría de los estudiantes tienen rendimientos similares, lo que puede ser indicativo de una eficacia más uniforme del método.

Estudios sobre Inequidad Social

- Contexto: Al estudiar la distribución de ingresos o el acceso a recursos educativos, la varianza y la desviación estándar pueden ayudar a identificar el grado de inequidad.
- Alta Varianza: Puede indicar una gran disparidad en el acceso o los ingresos entre diferentes grupos.
- Baja Varianza: Sugeriría una distribución más equitativa de los recursos o ingresos.

Investigaciones sobre Comportamiento y Actitudes

- Contexto: Al analizar las respuestas de encuestas sobre actitudes hacia políticas sociales, cuestiones de género, educación, etc.

Interpretación:

- Desviación Estándar Alta: Refleja opiniones muy variadas sobre el tema, indicando una falta de consenso.
- Desviación Estándar Baja: Implica un consenso más generalizado o una opinión común sobre el tema.

Supongamos que estamos evaluando las calificaciones de un nuevo curso en línea. Podemos calcular la varianza y la desviación estándar de las calificaciones para evaluar cuán consistentemente el curso está siendo recibido por diferentes estudiantes.

```
import numpy as np
import pandas as pd
# Generar datos de calificaciones
np.random.seed(42)
grades = np.random.normal(75, 10, 200)  # 200 calificaciones con media 75 y desvío estándar 10
# Crear DataFrame
df = pd.DataFrame(grades, columns=['Grades'])
# Calcular la varianza y la desviación estándar
variance = df['Grades'].var()
std_deviation = df['Grades'].std()
print("Varianza de las calificaciones:", variance)
print("Desviación estándar de las calificaciones:", std_deviation)
```

4. Coeficiente de variación y su aplicación

El coeficiente de variación (CV) es una medida estadística que proporciona una indicación de la variabilidad relativa de los datos en relación con la media del conjunto de datos. Es especialmente útil en situaciones donde se desean comparar las dispersiones de distribuciones que tienen diferentes unidades de medida o medias muy distintas. Este coeficiente se expresa como un porcentaje y se calcula como la razón entre la desviación estándar y la media, multiplicada por 100.

Fórmula del Coeficiente de Variación

$$CV = \frac{desviación\ estándar}{media} = \frac{\sigma}{\mu} \qquad CV = \frac{desviación\ estándar}{media} = \frac{s}{\bar{x}}$$

Aplicación en Ciencias Sociales y Educación

El coeficiente de variación tiene múltiples aplicaciones en la investigación científica en ciencias sociales y educación, especialmente en los siguientes contextos:

Comparación de Variabilidad entre Grupos Diferentes

- Contexto: Investigadores pueden estar interesados en comparar la consistencia de las puntuaciones en pruebas estandarizadas entre diferentes escuelas o distritos, que podrían tener diferentes medias.
- Aplicación del CV: El CV permite comparar la variabilidad de las puntuaciones independientemente de las medias. Esto es crucial cuando las medias son incomparables debido a diferencias

contextuales o demográficas.

Estudios de Diversidad Económica o Social

- Contexto: Al analizar la distribución de ingresos o la utilización de servicios públicos en diferentes regiones.
- Aplicación del CV: Un CV alto en el ingreso de una región indicaría una gran desigualdad económica. Por otro lado, un CV bajo en la utilización de servicios podría indicar equidad en el acceso a estos servicios entre la población.

Evaluación de la Homogeneidad en Respuestas de Encuestas

- Contexto: Al estudiar las actitudes o percepciones sobre cuestiones políticas, sociales o educativas.
- Aplicación del CV: El CV ayuda a determinar qué tan homogéneas o dispersas son las opiniones o comportamientos dentro de un grupo. Un CV bajo sugiere una opinión o comportamiento cohesivo, mientras que un CV alto podría indicar diversidad o división de opiniones.

Imaginemos que queremos analizar la variabilidad de las calificaciones en dos cursos diferentes para determinar cuál de ellos presenta una mayor consistencia en las calificaciones de los estudiantes.

```
import numpy as np
import pandas as pd
# Datos simulados: calificaciones de dos cursos diferentes
data = {
    "Curso_A": np.random.normal(70, 8, 100),  # Media 70, desviación estándar 8
    "Curso_B": np.random.normal(70, 12, 100)  # Media 70, desviación estándar 12
}
df = pd.DataFrame(data)
# Cálculo del Coeficiente de Variación
cv_a = (df['Curso_A'].std() / df['Curso_A'].mean()) * 100
cv_b = (df['Curso_B'].std() / df['Curso_B'].mean()) * 100
print("Coeficiente de Variación para el Curso A:", round(cv_a, 2), "%")
print("Coeficiente de Variación para el Curso B:", round(cv_b, 2), "%")
```

5. Visualización de la dispersión: box plots e histogramas

Este ejemplo supondrá que estamos analizando datos de una encuesta sobre la satisfacción de estudiantes universitarios con varios aspectos de un programa educativo online. Los aspectos a analizar incluirán la satisfacción con el material del curso, la interacción

con los profesores y el soporte técnico.

Supongamos que tenemos respuestas de 300 estudiantes, donde calificaron su satisfacción en una escala del 1 al 10 para cada uno de los tres aspectos mencionados.

Objetivos:

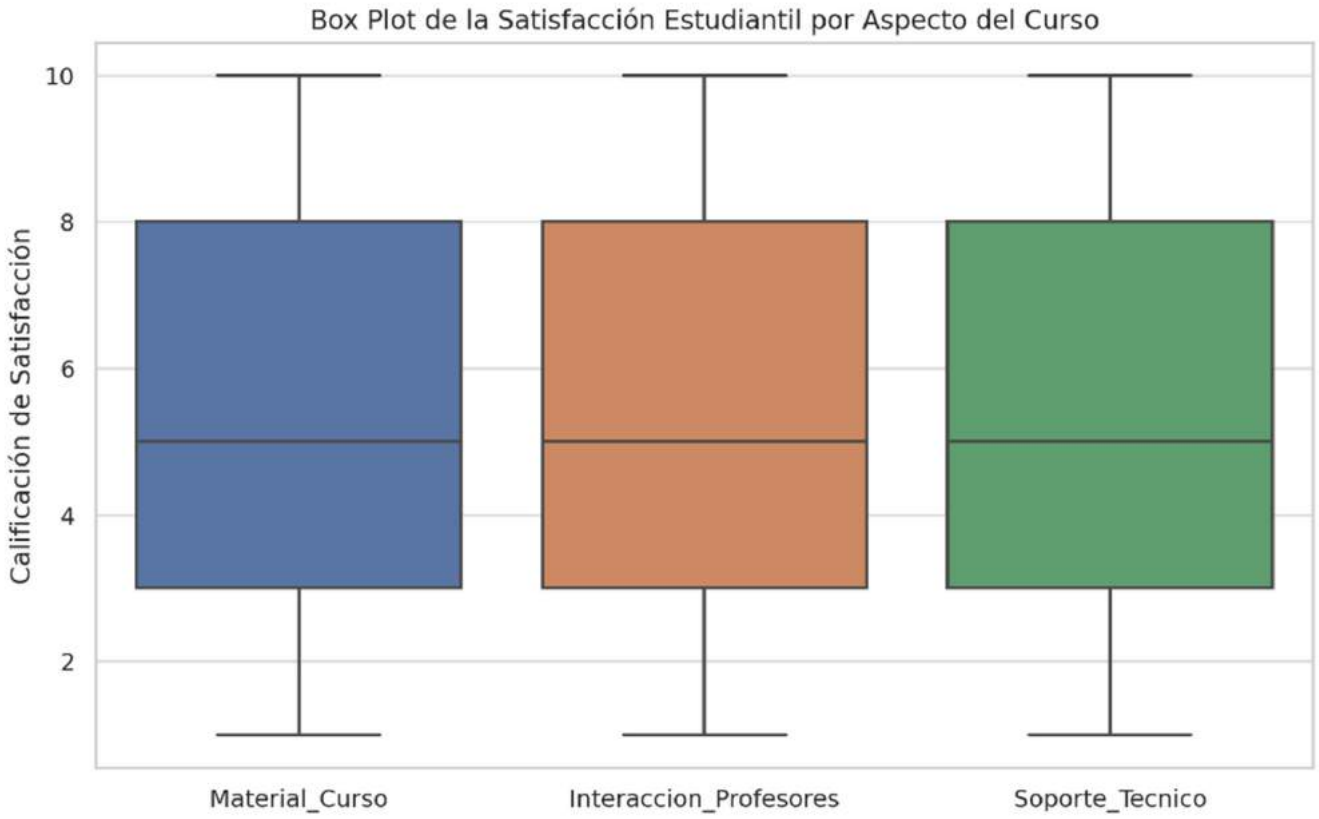
- Visualizar la distribución de las calificaciones para cada aspecto.
- Identificar la presencia de valores atípicos.
- Comparar la centralidad y la dispersión de las calificaciones entre los aspectos.

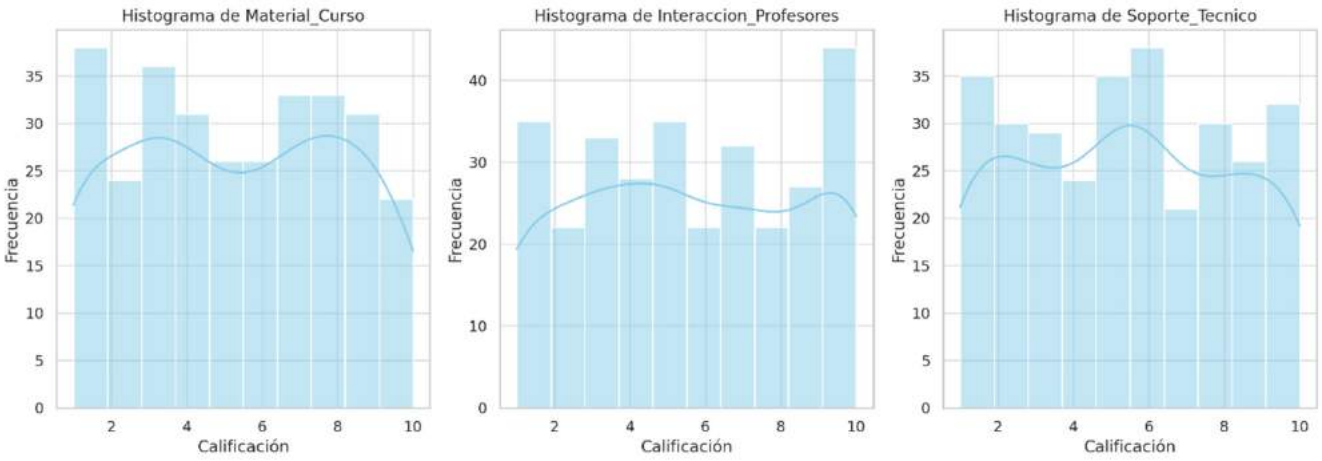
Primero, generaremos datos simulados y luego utilizaremos matplotlib y seaborn para crear los box plots y los histogramas.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Configurar estilo de seaborn
sns.set(style="whitegrid")
# Generar datos simulados
np.random.seed(42)  # Para reproducibilidad
data = {
```

```
    'Material_Curso': np.random.randint(1, 11, 300),
    'Interaccion_Profesores': np.random.randint(1, 11, 300),
    'Soporte_Tecnico': np.random.randint(1, 11, 300)
}
# Crear DataFrame
df = pd.DataFrame(data)
# Crear box plots
plt.figure(figsize=(10, 6))
sns.boxplot(data=df)
plt.title('Box Plot de la Satisfacción Estudiantil por Aspecto del Curso')
plt.ylabel('Calificación de Satisfacción')
plt.show()
# Crear histogramas
plt.figure(figsize=(14, 5))
for i, column in enumerate(df.columns, 1):
    plt.subplot(1, 3, i)
  sns.histplot(df[column], bins=10, kde=True, color='skyblue')
    plt.title(f'Histograma de {column}')
    plt.xlabel('Calificación')
```

```
    plt.ylabel('Frecuencia')
plt.tight_layout()
plt.show()
```





6. Dispersión en datos no paramétricos

El análisis de la dispersión en datos no paramétricos es particularmente relevante en las ciencias sociales y educativas, donde las variables de interés a menudo no cumplen con las suposiciones de las pruebas paramétricas. Estas variables pueden ser ordinales, categóricas o no seguir una distribución normal, por lo que los métodos no paramétricos son esenciales para su análisis correcto y ético.

¿Qué son los Datos No Paramétricos?

Los datos no paramétricos incluyen aquellos que no se ajustan a una distribución normal, así como datos categóricos y ordinales. En ciencias sociales y educación, este tipo de datos es común en encuestas, cuestionarios y otros instrumentos de medición donde las respuestas se dan en escalas de Likert, clasificaciones, o categorías nominales.

Uso de Medidas de Tendencia Central No Paramétricas:

- Mediana: Una medida de tendencia central que es útil cuando los datos son sesgados o tienen valores atípicos.
- Moda: Especialmente útil para datos categóricos, donde se identifica la categoría más frecuentemente observada.

Medidas de Dispersión:

- Rango: La diferencia entre el valor máximo y mínimo en el conjunto de datos.
- Rango Intercuartílico (IQR): Muestra la dispersión de la mitad central de los datos y es menos sensible a valores extremos.

Pruebas No Paramétricas para Comparar Grupos:

- Prueba de Kruskal-Wallis: Utilizada para comparar las medianas entre dos o más grupos independientes cuando los datos no son normales.
- Prueba de Friedman: Para comparar grupos relacionados usando datos no paramétricos.

Visualización:

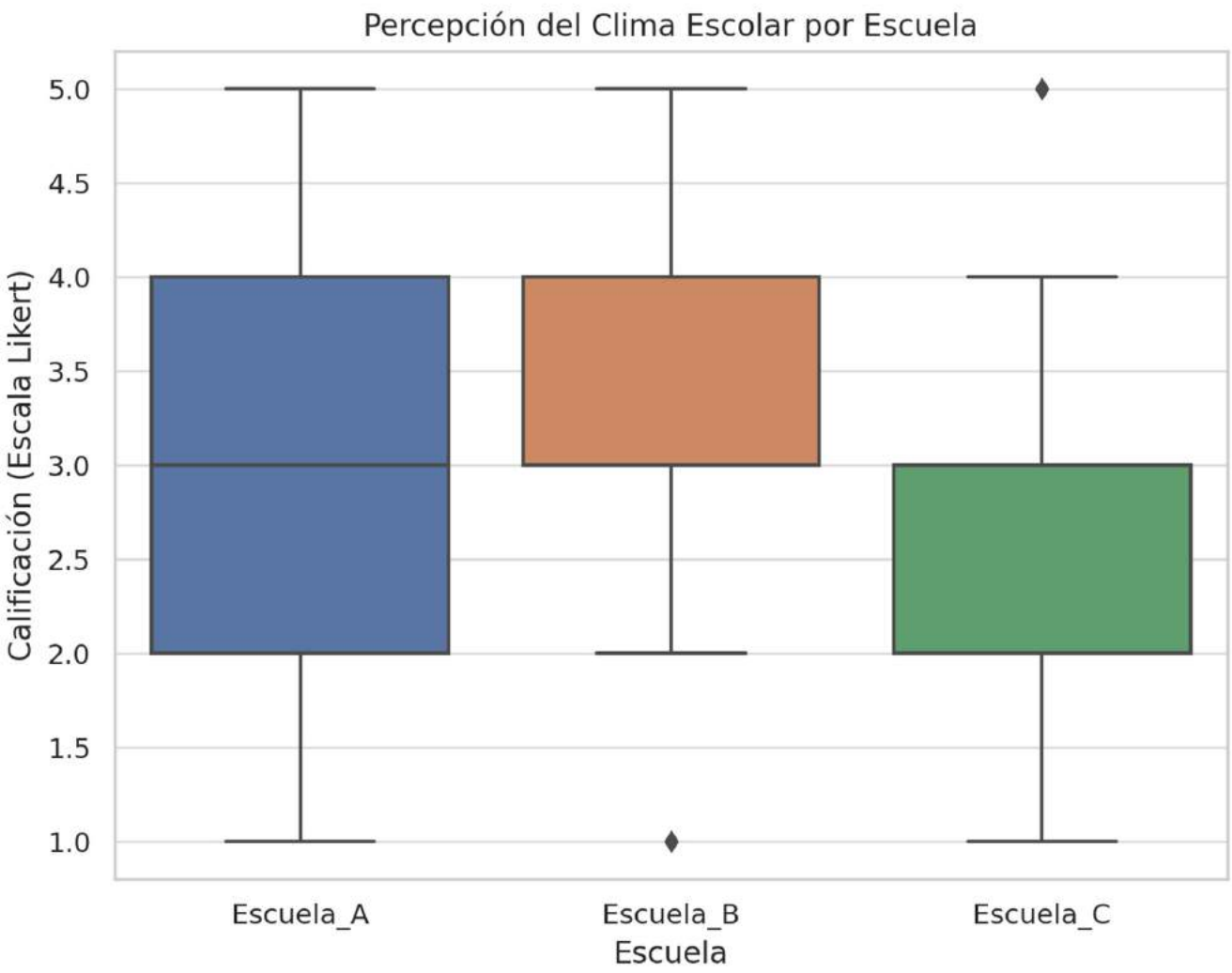
- Box plots: Útiles para visualizar la dispersión, la mediana, y los valores atípicos.
- Histogramas y gráficos de barras: Para datos categóricos, mostrando la frecuencia de cada categoría.

Supongamos que realizamos una encuesta en diferentes escuelas utilizando una escala de Likert (1 a 5) para evaluar el clima escolar. Queremos analizar la dispersión de las respuestas para entender variaciones en la percepción del clima escolar.

Generaremos datos simulados y utilizaremos métodos no paramétricos para analizar la dispersión:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
# Generar datos simulados
np.random.seed(42)
datos = {
    'Escuela_A': np.random.choice([1, 2, 3, 4, 5], 100, p=[0.1, 0.2, 0.4, 0.2, 0.1]),
   'Escuela_B': np.random.choice([1, 2, 3, 4, 5], 100, p=[0.05, 0.15, 0.50, 0.20, 0.10]),
    'Escuela_C': np.random.choice([1, 2, 3, 4, 5], 100, p=[0.15, 0.35, 0.30, 0.15, 0.05])
}
df = pd.DataFrame(datos)
# Visualización con Box Plot
sns.boxplot(data=df)
plt.title('Percepción del Clima Escolar por Escuela')
plt.xlabel('Escuela')
plt.ylabel('Calificación (Escala Likert)')
plt.show()
```



7. Comprensión lectora

¿Qué medida de dispersión es la diferencia entre el valor máximo y mínimo en un conjunto de datos?

A. Varianza

B. Desviación estándar

C. Rango

D. Coeficiente de variación

ANSWER: **C**

¿Cuál de las siguientes afirmaciones describe mejor la desviación estándar?

A. Es una medida de dispersión que no tiene en cuenta la media.

B. Proporciona una visión rápida de la dispersión sin detalles.

C. Mide la dispersión de los datos en relación con la media.

D. Se calcula sin considerar los valores extremos.

ANSWER: **C**

¿Qué medida de dispersión es particularmente útil cuando se analizan datos con valores atípicos?

A. Varianza

B. Rango

C. Rango intercuartílico (IQR)

D. Media aritmética

ANSWER: **C**

¿Cómo influye la dispersión en la toma de decisiones en el mundo de los negocios según el texto?

A. Una menor dispersión siempre indica una menor rentabilidad.

B. La alta dispersión en los retornos de una inversión indica mayor volatilidad y riesgo.

C. La dispersión no afecta las decisiones de inversión.

D. El rango es la única medida de dispersión considerada para la

toma de decisiones financieras.

ANSWER: **B**

En el contexto de la salud, ¿por qué es importante entender la dispersión?

A. Indica cuán efectivos son los tratamientos en todos los pacientes.

B. Muestra la relación directa entre los tratamientos y sus costos.

C. Determina la precisión diagnóstica de los equipos médicos.

D. Previene la aparición de nuevas enfermedades.

ANSWER: **A**

¿Cuál es un desafío mencionado en el texto respecto al uso de medidas de dispersión como la varianza y la desviación estándar?

A. Son demasiado simples para ser útiles.

B. Pueden ser engañosas en presencia de valores atípicos.

C. Solo aplican a datos cuantitativos.

D. Son irrelevantes en estudios científicos.

ANSWER: **B**

¿Qué afirmación es cierta sobre el rango intercuartílico (IQR)?

A. Es igual a la suma de todos los cuartiles.

B. Es la diferencia entre el tercer y el primer cuartil.

C. Se calcula sumando el rango y dividendo entre dos.

D. Es más sensible a valores atípicos que la varianza.

ANSWER: **B**

¿Cuál es una ventaja del coeficiente de variación en comparación con otras medidas de dispersión?

A. No requiere conocimiento de la media.

B. Permite comparar la dispersión entre conjuntos de datos con diferentes unidades o medias.

C. Es más fácil de calcular que el rango.

D. Es menos sensible a valores atípicos que la mediana.

ANSWER: **B**

¿Qué indicaría una alta varianza en las calificaciones de un curso?

A. Que todas las calificaciones son iguales.

B. Que las calificaciones son muy consistentes.

C. Que hay una gran variabilidad en las calificaciones.

D. Que el curso es fácil.

ANSWER: **C**

¿Por qué es importante la visualización de la dispersión como los

box plots e histogramas?

A. Solo es relevante en el análisis de datos pequeños.

B. Ayuda a identificar tendencias y patrones en los datos.

C. Permite identificar valores atípicos y comparar la dispersión.

D. Muestra la correlación directa entre variables.

ANSWER: **C**

¿En qué tipo de datos es crucial utilizar pruebas no paramétricas según el texto?

A. Datos que siguen una distribución normal.

B. Datos cuantitativos continuos.

C. Datos no normales o categóricos.

D. Datos que no incluyen valores atípicos.

ANSWER: **C**

¿Qué medida de dispersión sería más adecuada para analizar la equidad en el acceso a servicios públicos entre diferentes regiones?

A. Media

B. Moda

C. Coeficiente de variación

D. Mediana

ANSWER: **C**

¿Cuál de las siguientes es una desventaja del rango como medida de dispersión?

A. Es difícil de calcular.

B. No considera los valores intermedios del conjunto de datos.

C. Siempre proporciona un valor muy alto.

D. Es más sensible a los valores atípicos que la media.

ANSWER: **B**

¿Qué proporciona el rango intercuartílico (IQR) que el rango no ofrece?

A. Una medida de la media

B. Una visión más detallada de donde se concentra la mayoría de las respuestas

C. La suma de los valores extremos

D. Una visión general rápida de la dispersión

ANSWER: **B**

Capítulo IV.

Distribuciones de Frecuencias y Tablas

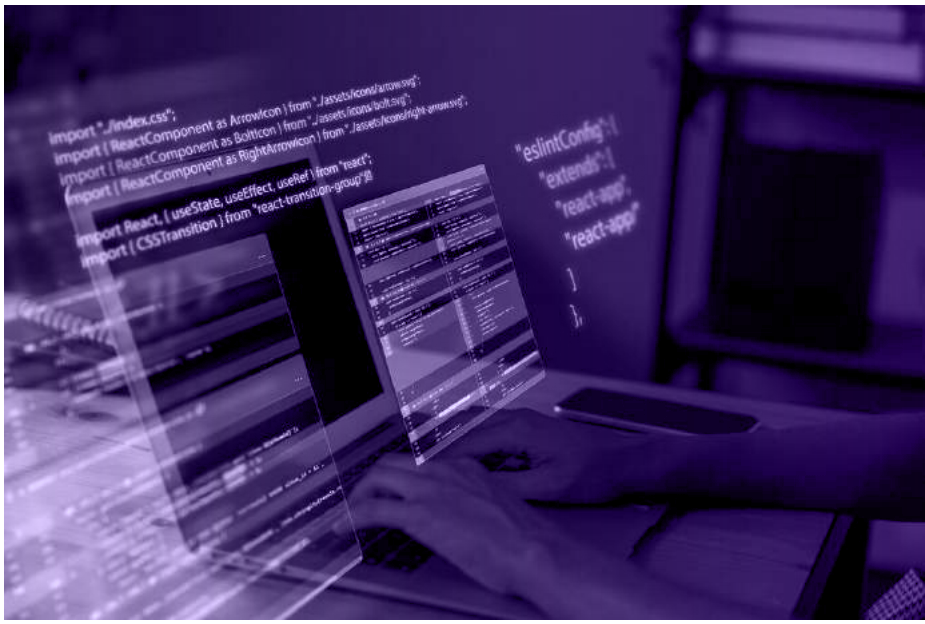
Capítulo IV.

Distribuciones de Frecuencias y Tablas

1. Teoría y construcción de distribuciones de frecuencia

Las distribuciones de frecuencia pueden ser discretas o continuas, dependiendo de la naturaleza de los datos. En una distribución de frecuencia discreta, los datos se presentan en categorías individuales y son contables. Por ejemplo, el número de veces que aparece cada color en un lote de canicas. En contraste, las distribuciones de frecuencia continua agrupan datos en rangos de valores, y se utilizan para datos que pueden tomar cualquier valor dentro de un intervalo, como las alturas de un grupo de personas.

Una distribución de frecuencia se construye determinando primero el número de intervalos (o "bins") que se utilizarán, lo que puede influir significativamente en la interpretación de los datos. Demasiados o muy pocos intervalos pueden distorsionar la representación visual de los datos, como los histogramas, que son una



herramienta común para visualizar distribuciones de frecuencia.

La elección del número de intervalos y su amplitud puede basarse en reglas como la Regla de Sturges o la Regla de Scott, las cuales proporcionan guías basadas en el tamaño de la muestra y la dispersión de los datos. Sin embargo, estas decisiones también deben considerar el contexto específico de los datos y los objetivos del análisis.

La construcción efectiva de una distribución de frecuencia implica varios pasos detallados:

- Determinación de los rangos de datos: Dependiendo de si los datos son discretos o continuos, se decide cómo agrupar los datos. En datos continuos, se deben definir rangos que son suficientemente amplios para mostrar tendencias y suficientemente estrechos para mantener la precisión.

- Conteo de frecuencias: Cada dato se cuenta en su respectivo rango o categoría. Este paso es fundamental porque las frecuencias resultantes son las que se analizan posteriormente.

- Representación visual: Los histogramas y los polígonos de frecuencia son dos de las representaciones visuales más comunes utilizadas para mostrar distribuciones de frecuencia. Estas herramientas ayudan a visualizar la forma de la distribución, lo

que puede indicar propiedades estadísticas como la simetría, la asimetría y la presencia de modas.

- Análisis e interpretación: Una vez construida la distribución, se analiza para interpretar qué indica sobre el conjunto de datos. Se buscan indicadores como el centro de los datos, la variabilidad, y si hay patrones o irregularidades que requieran investigación adicional.

Un ejemplo de la aplicación de estos conceptos se encuentra en el trabajo de Dewey (1992), quien discute diferentes métodos para formar distribuciones de frecuencia y compara su eficiencia y generalidad. Dewey señala que el método elegido puede depender significativamente del tamaño del conjunto de datos y de las características específicas de los datos.

2. Tablas de frecuencia en Python

Son particularmente útiles para comprender cómo se distribuyen las respuestas en encuestas, cuestionarios o evaluaciones, y para visualizar las preferencias o comportamientos de un grupo.

Aplicación en Ciencias Sociales y Educación

- Encuestas Educativas: Al analizar respuestas de estudiantes a

preguntas sobre, por ejemplo, métodos de enseñanza o satisfacción escolar.

- Estudios Demográficos: Para mostrar la distribución de características como edad, género, nivel educativo, etc.
- Investigación de Opinión Pública: Analizar la distribución de opiniones sobre temas sociales o políticos.

Supongamos que hemos realizado una encuesta sobre la satisfacción de los estudiantes con los servicios de biblioteca en una universidad. Los estudiantes calificaron su satisfacción en una escala de Likert de 1 a 5, donde 1 es 'Muy Insatisfecho' y 5 es 'Muy Satisfecho'.

Vamos a generar algunos datos simulados para esta encuesta y luego crearemos una tabla de frecuencia para analizar los resultados.

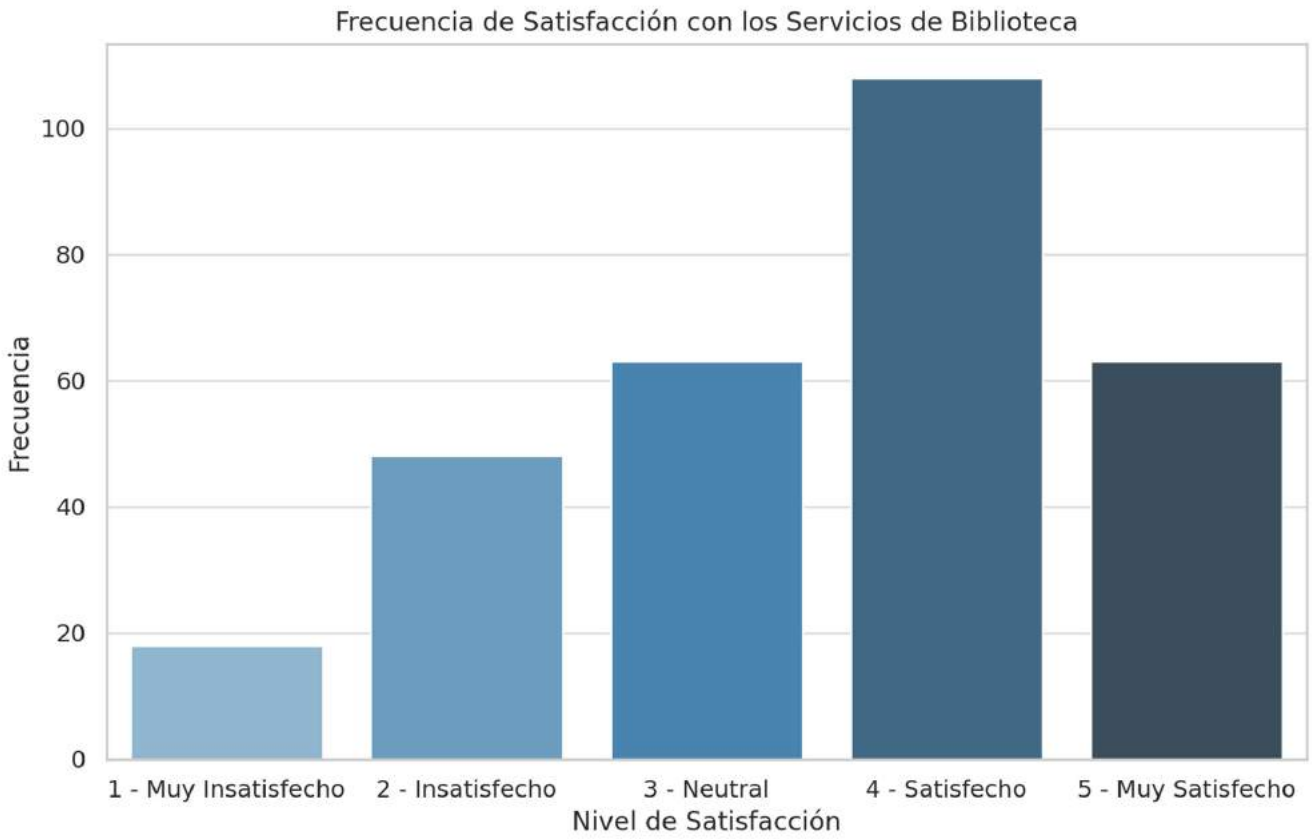
```
import numpy as np

import pandas as pd

# Simular respuestas de la encuesta

np.random.seed(42)

satisfaction_data = np.random.choice(['1 - Muy Insatisfecho',
'2 - Insatisfecho', '3 - Neutral', '4 - Satisfecho', '5 - Muy
Satisfecho'],

                                     300, p=[0.05, 0.15, 0.20, 0.40, 0.20])

# Crear DataFrame
```

```
df=pd.DataFrame(satisfaction_data,columns=['Satisfacción'])
# Crear tabla de frecuencia
frequency_table = df['Satisfacción'].value_counts().sort_index()
print(frequency_table)
```

Además de la tabla de frecuencia, podemos visualizar los resultados utilizando un gráfico de barras para una interpretación más intuitiva y visual de los datos.

```
import matplotlib.pyplot as plt
import seaborn as sns
# Configuración del estilo del gráfico
sns.set(style="whitegrid")
# Crear gráfico de barras
plt.figure(figsize=(10, 6))
sns.barplot(x=frequency_table.index, y=frequency_table.values, palette='Blues_d')
plt.title('Frecuencia de Satisfacción con los Servicios de Biblioteca')
```

```
plt.xlabel('Nivel de Satisfacción')
plt.ylabel('Frecuencia')
plt.show()
```



- 1 - Muy Insatisfecho: 18
- 2 - Insatisfecho: 48
- 3 - Neutral: 63
- 4 - Satisfecho: 108
- 5 - Muy Satisfecho: 63

La mayor frecuencia de respuestas está en el nivel "4 - Satisfecho",

indicando que la mayoría de los estudiantes están satisfechos con los servicios de biblioteca. La distribución muestra una tendencia positiva hacia la satisfacción, con una combinación significativa de respuestas en "5 - Muy Satisfecho".

Los niveles de insatisfacción ("1 - Muy Insatisfecho" y "2 - Insatisfecho") son relativamente bajos en comparación con los niveles más altos de satisfacción, pero todavía representan una porción considerable que podría ser objeto de investigación y mejora.

3. Histogramas y polígonos de frecuencia

Supongamos que una organización educativa ha realizado una prueba estandarizada a 1000 estudiantes de secundaria. La organización ha recopilado datos no solo de las puntuaciones, sino también del nivel socioeconómico (NSE) de cada estudiante, clasificado como bajo, medio o alto. Queremos explorar cómo las puntuaciones varían con respecto a los diferentes NSEs y analizar la equidad en la educación.

Objetivos:

- Generar histogramas para visualizar la distribución de puntuaciones para cada nivel socioeconómico.
- Crear polígonos de frecuencia para comparar visualmente estas distribuciones.
- Identificar posibles sesgos o tendencias en las puntuaciones en función del NSE.

Simularemos un conjunto de datos que incluye:

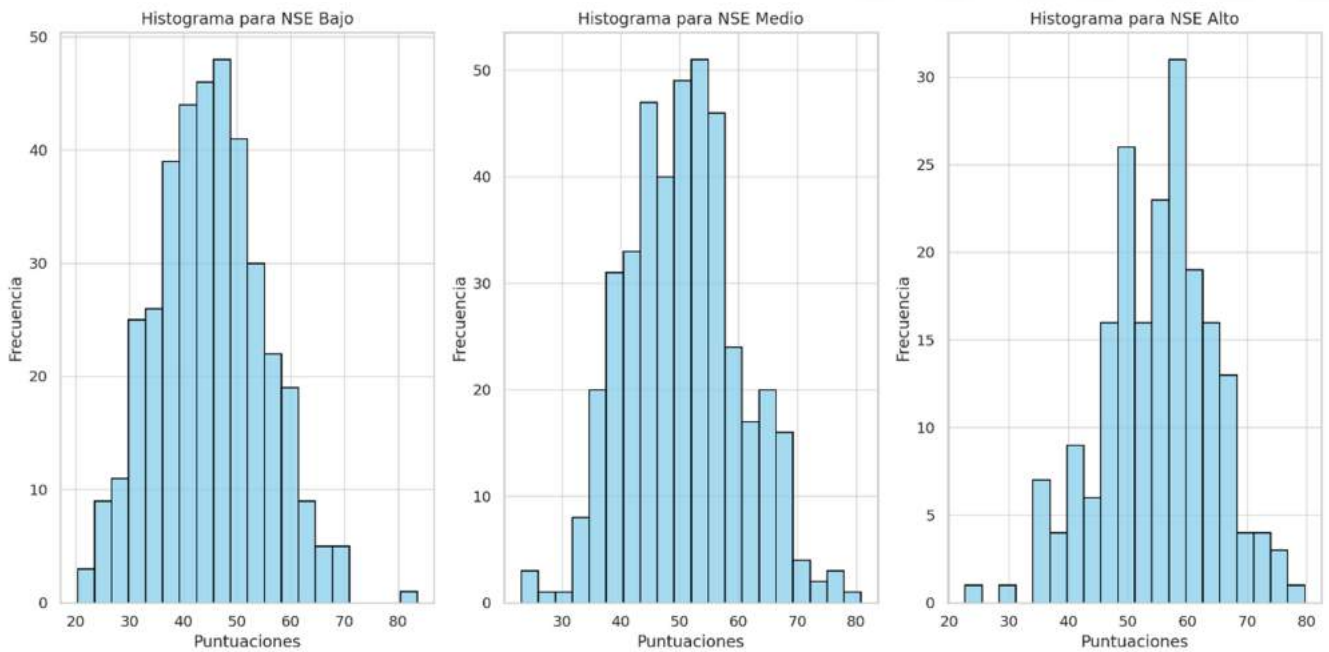
- puntuaciones: Un arreglo de puntuaciones obtenidas en la prueba.
- NSE: Un arreglo que indica el nivel socioeconómico de cada estudiante.

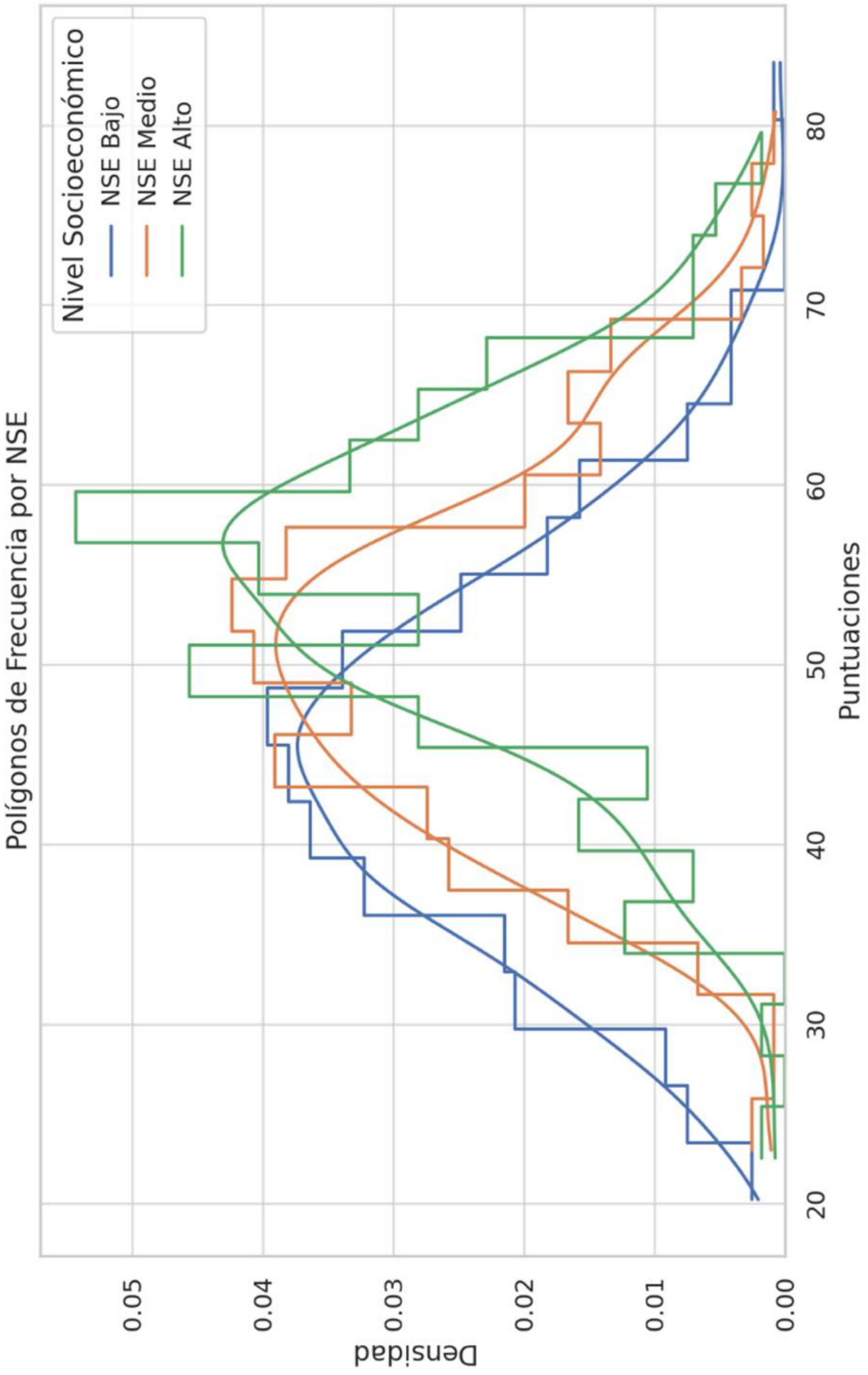
Vamos a simular los datos, crear histogramas y polígonos de frecuencia usando bibliotecas como numpy, pandas, matplotlib, y seaborn.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Configuración para la visualización
sns.set(style="whitegrid")
# Simulación de datos
np.random.seed(42)
tamaño_muestra = 1000
puntuaciones = np.random.normal(loc=50, scale=10, size=tamaño_muestra) # Puntuaciones centradas en 50 con desviación de 10
niveles_socioeconomicos = np.random.choice(['Bajo', 'Medio', 'Alto'], tamaño_muestra, p=[0.4, 0.4, 0.2])
# Creación del DataFrame
```

```
df = pd.DataFrame({
    'Puntuaciones': puntuaciones,
    'NSE': niveles_socioeconomicos
})
# Ajustar las puntuaciones para reflejar un posible sesgo socioeconómico
ajuste_nse = {'Bajo': -5, 'Medio': 0, 'Alto': 5}
df['Puntuaciones Ajustadas'] = df.apply(lambda x: x['Puntuaciones'] + ajuste_nse[x['NSE']], axis=1)
# Creación de los histogramas
plt.figure(figsize=(14, 7))
for i, nse in enumerate(['Bajo', 'Medio', 'Alto'], 1):
    plt.subplot(1, 3, i)
    subset = df[df['NSE'] == nse]
     sns.histplot(subset['Puntuaciones Ajustadas'], kde=False, bins=20, color='skyblue', edgecolor='black')
    plt.title(f'Histograma para NSE {nse}')
    plt.xlabel('Puntuaciones')
    plt.ylabel('Frecuencia')
plt.tight_layout()
plt.show()
```

```
# Creación de polígonos de frecuencia
plt.figure(figsize=(10, 6))
for nse in ['Bajo', 'Medio', 'Alto']:
    subset = df[df['NSE'] == nse]
    sns.histplot(subset['Puntuaciones Ajustadas'], kde=True, bins=20,stat="density",element="step",fill=False,label=f'NSE {nse}')
plt.title('Polígonos de Frecuencia por NSE')
plt.xlabel('Puntuaciones')
plt.ylabel('Densidad')
plt.legend(title='Nivel Socioeconómico')
plt.show()
```




Polígonos de Frecuencia por NSE
Nivel Socioeconómico
NSE Bajo
NSE Medio
NSE Alto
Densidad
0.05
0.04
0.03
0.02
0.01
0.00
20
30
40
50
60
70
80
Puntuaciones

- **Histogramas por NSE:** En los histogramas, cada gráfico representa la distribución de las puntuaciones ajustadas para cada nivel socioeconómico (Bajo, Medio, Alto). Las puntuaciones han sido ajustadas para reflejar un impacto hipotético del nivel socioeconómico sobre los resultados. Esto proporciona una visualización clara de cómo las puntuaciones se distribuyen dentro de cada grupo.

- **Polígonos de Frecuencia:** Este gráfico compara las densidades de las puntuaciones ajustadas para los tres niveles socioeconómicos. Los polígonos de frecuencia permiten una comparación visual directa entre los grupos, destacando diferencias en la forma y el rango de las distribuciones.

4. Distribuciones acumulativas

Supongamos que tenemos datos de una muestra de estudiantes que han participado en un examen estandarizado nacional. Los datos incluyen la puntuación del examen y el tipo de escuela a la que asisten los estudiantes.

Objetivos:

- Generar distribuciones acumulativas para las puntuaciones de los exámenes por tipo de escuela.

- Comparar estas distribuciones para identificar diferencias en el rendimiento académico.

- Evaluar las implicaciones de estas diferencias para la equidad en la educación.

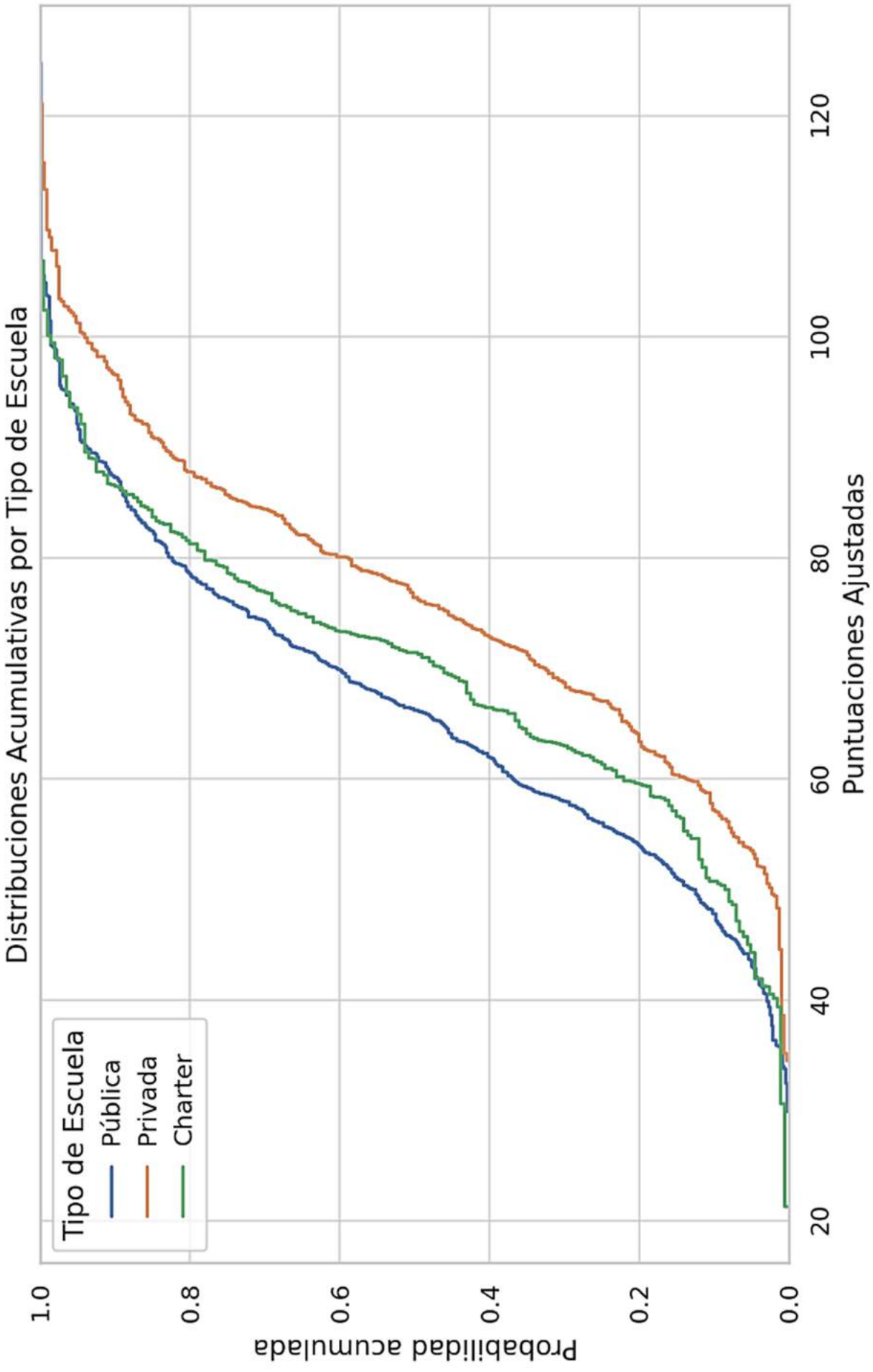
Simularemos un conjunto de datos que incluye:

- **puntuaciones:** Un arreglo de puntuaciones obtenidas en el examen.
- **tipo_escuela:** Un arreglo que indica el tipo de escuela que cada estudiante asiste (pública, privada, charter).

Usaremos Python para simular los datos y crear las distribuciones acumulativas utilizando numpy, pandas, matplotlib, y seaborn.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
# Configuración para la visualización
sns.set(style="whitegrid")
# Simulación de datos
np.random.seed(42)
tamaño_muestra = 1000
puntuaciones = np.random.normal(loc=70, scale=15, size=tamaño_muestra) # Puntuaciones centradas en 70 con desviación de 15
tipos_escuela = np.random.choice(['Pública', 'Privada', 'Charter'], tamaño_muestra, p=[0.5, 0.3, 0.2])
```

```
# Creación del DataFrame
df = pd.DataFrame({
    'Puntuaciones': puntuaciones,
    'Tipo de Escuela': tipos_escuela
})
# Ajustar las puntuaciones para reflejar un posible sesgo por tipo de escuela
ajuste_escuela = {'Pública': -3, 'Privada': 5, 'Charter': 0}
df['Puntuaciones Ajustadas'] = df.apply(lambda x: x['Puntuaciones'] + ajuste_escuela[x['Tipo de Escuela']], axis=1)
# Creación de las distribuciones acumulativas
plt.figure(figsize=(10, 6))
for tipo in ['Pública', 'Privada', 'Charter']:
    subset = df[df['Tipo de Escuela'] == tipo]
 sns.ecdfplot(subset['Puntuaciones Ajustadas'], label=f'{tipo}')
plt.title('Distribuciones Acumulativas por Tipo de Escuela')
plt.xlabel('Puntuaciones Ajustadas')
plt.ylabel('Probabilidad acumulada')
plt.legend(title='Tipo de Escuela')
plt.grid(True)
plt.show()
```


Distribuciones Acumulativas por Tipo de Escuela
Probabilidad acumulada
Puntuaciones Ajustadas
Tipo de Escuela
Pública
Privada
Charter
0.0
0.2
0.4
0.6
0.8
1.0
20
40
60
80
100
120

En el gráfico de distribuciones acumulativas por tipo de escuela, se puede observar cómo se distribuyen las puntuaciones ajustadas entre estudiantes de escuelas públicas, privadas y charter:

- **Escuelas Privadas:** Los estudiantes de escuelas privadas tienden a tener puntuaciones más altas, como se refleja en la curva que se desplaza hacia la derecha. Esto podría indicar un mejor rendimiento académico general, posiblemente debido a recursos más abundantes o métodos de enseñanza diferenciados.
- **Escuelas Públicas:** La curva para las escuelas públicas está ligeramente desplazada hacia la izquierda, indicando puntuaciones generalmente más bajas. Esto puede sugerir limitaciones en términos de recursos o apoyo académico en comparación con escuelas privadas.
- **Escuelas Charter:** Los estudiantes de las escuelas charter muestran una distribución intermedia, no alcanzando los picos de las escuelas privadas pero generalmente superando a las públicas en términos de puntuaciones más altas.

5. Análisis de frecuencias relativas y absolutas

- **Frecuencia Absoluta:** La frecuencia absoluta de un valor en un conjunto de datos es simplemente el recuento de cuántas veces aparece ese valor. Por ejemplo, si estás analizando las edades de un grupo de personas y tienes 30 personas en total, y 5 de ellas tienen 25 años, entonces la frecuencia absoluta de 25 años sería 5.

- **Frecuencia Relativa:** La frecuencia relativa de un valor en un conjunto de datos es el cociente entre la frecuencia absoluta de ese valor y el tamaño total del conjunto de datos. Se expresa generalmente como un porcentaje. Usando el ejemplo anterior, si 5 de 30 personas tienen 25 años, entonces la frecuencia relativa de 25 años sería 5/30=0.1667305=0.1667, o un 16.67%.

El análisis de frecuencias relativas y absolutas es útil para entender la distribución de los datos y puede ayudar a identificar patrones, tendencias o valores atípicos en un conjunto de datos. Por ejemplo, si estás analizando las edades de un grupo de personas y observas que la mayoría de las personas tienen edades entre 20 y 30 años, esto podría sugerir una tendencia demográfica en ese grupo específico.

6. Comprensión lectora

¿Qué tipo de datos se agrupan en categorías contables en las distribuciones de frecuencia?

A. Datos continuos

B. Datos discretos

C. Datos cualitativos

D. Datos binarios

ANSWER: **B**

¿Qué herramienta visual se utiliza comúnmente para representar distribuciones de frecuencia continua?

A. Diagramas de caja

B. Histogramas

C. Mapas de calor

D. Gráficos de línea

ANSWER: **B**

Al construir un histograma, ¿qué puede provocar una representación distorsionada de los datos?

A. Elegir un color de fondo neutro

B. Tener demasiados o muy pocos intervalos

C. Utilizar etiquetas claras en los ejes

D. Incluir todos los datos disponibles

ANSWER: **B**

¿Cuál es una regla para determinar el número de intervalos en un histograma?

A. Regla de Pearson

B. Regla de Sturges

C. Regla de Bayes

D. Regla de Newton

ANSWER: **B**

¿Qué indican los polígonos de frecuencia en una distribución de frecuencia?

A. La mediana de los datos

B. La correlación entre variables

C. La forma de la distribución

D. La varianza de los datos

ANSWER: **C**

¿Para qué tipo de análisis se utilizan principalmente las tablas de frecuencia en las ciencias sociales?

A. Para determinar la causalidad entre variables

B. Para describir cómo se distribuyen las respuestas en encuestas

C. Para calcular la regresión lineal

D. Para realizar pruebas de hipótesis complejas

ANSWER: **B**

¿Qué aspecto se debe considerar al analizar la satisfacción de los estudiantes con los servicios de una biblioteca usando una escala de Likert?

A. La frecuencia de las respuestas en cada nivel de satisfacción

B. La duración de cada encuesta completada

C. El color de la biblioteca

D. El número de libros disponibles

ANSWER: **A**

En el contexto de una encuesta, ¿qué mostraría una alta frecuencia en el nivel "4 - Satisfecho"?

A. Que ningún estudiante está satisfecho

B. Que la mayoría de los estudiantes están insatisfechos

C. Que la mayoría de los estudiantes están satisfechos

D. Que los servicios de la biblioteca son insuficientes

ANSWER: **C**

¿Qué metodología se utiliza para visualizar la distribución de puntuaciones ajustadas por nivel socioeconómico?

A. Mapas conceptuales

B. Histogramas y polígonos de frecuencia

C. Gráficos de dispersión

D. Diagramas Venn

ANSWER: **B**

¿Cuál es el propósito de ajustar las puntuaciones por nivel socioeconómico en un análisis de datos educativos?

A. Reducir la cantidad de datos a analizar

B. Compensar posibles sesgos inherentes a los niveles socioeconómicos

C. Simplificar los cálculos estadísticos

D. Incrementar la complejidad del modelo

ANSWER: **B**

¿Qué representa una distribución acumulativa en el contexto de las puntuaciones de un examen?

A. La probabilidad de obtener cualquier puntuación dada

B. El porcentaje de estudiantes que alcanza cada puntuación o menos

C. La diferencia entre las puntuaciones más altas y más bajas

D. La correlación entre puntuaciones y rendimiento académico

ANSWER: **B**

¿Cómo se interpreta un gráfico de distribuciones acumulativas en términos de rendimiento académico por tipo de escuela?

A. Muestra la dispersión de las puntuaciones entre los estudiantes

B. Indica qué tipo de escuela tiene el mayor número de estudiantes

C. Permite comparar el rendimiento académico entre tipos de escuela

D. Evalúa la efectividad de los métodos de enseñanza

ANSWER: **C**

En el análisis de frecuencias relativas, ¿qué información proporcio-

na este tipo de frecuencia?

A. El número total de observaciones para un valor específico

B. La proporción que un valor específico representa respecto al total

C. La diferencia entre la frecuencia más alta y más baja

D. La media de todas las frecuencias observadas

ANSWER: **B**

¿Cuál es una ventaja de visualizar los datos utilizando gráficos de barras junto con tablas de frecuencia?

A. Permite una interpretación más rápida y directa de los datos

B. Reduce el tamaño de los datos para su análisis

C. Incrementa la precisión de los cálculos matemáticos

D. Disminuye la variabilidad de los datos

ANSWER: **A**

Capítulo V.

Correlación y Covarianza

Capítulo V.

Correlación y Covarianza

1. Conceptos básicos de correlación y covarianza

Covarianza:

- La covarianza es una medida de cómo cambian dos variables juntas.
- Se calcula como el promedio de los productos de las desviaciones de cada variable respecto a su media.
- Una covarianza positiva indica que las dos variables tienden a cambiar en la misma dirección (es decir, cuando una aumenta, la otra también tiende a aumentar, y viceversa).
- Una covarianza negativa indica que las dos variables tienden a cambiar en direcciones opuestas (es decir, cuando una aumenta, la otra tiende a disminuir, y viceversa).

- Sin embargo, la covarianza no está estandarizada, lo que significa que su magnitud puede ser difícil de interpretar.

Correlación:

- La correlación es una medida estandarizada de la relación entre dos variables.
- Se calcula dividiendo la covarianza entre el producto de las desviaciones estándar de ambas variables.
- La correlación varía entre -1 y 1.
- Una correlación de 1 indica una correlación perfecta positiva, lo que significa que las dos variables tienen una relación lineal positiva perfecta.
- Una correlación de -1 indica una correlación perfecta negativa, lo que significa que las dos variables tienen una relación lineal negativa perfecta.
- Una correlación de 0 indica la ausencia de una relación lineal entre las dos variables.
- La correlación es una medida más fácil de interpretar que la covarianza porque está normalizada y su magnitud indica la fuerza y la dirección de la relación entre las variables.

2. Cálculo de la covarianza en Python

Los datos consisten en puntuaciones en matemáticas, ciencias y lengua, junto con el número de horas dedicadas al estudio por semana para cada materia. El objetivo es determinar cómo se rela-

cionan estas variables entre sí, específicamente buscando entender si los estudiantes que dedican más tiempo al estudio tienden a tener mejores puntuaciones y cómo estas relaciones varían entre las diferentes materias. Un paso a paso es:

- Preparación de Datos: Generaremos datos simulados para un conjunto de estudiantes.
- Cálculo de Covarianza: Usaremos numpy para calcular la matriz de covarianza entre las distintas variables.
- Interpretación: Analizaremos los resultados para entender las relaciones entre las variables.

```
import numpy as np

# Simulación de datos

np.random.seed(42)  # Semilla para reproducibilidad

num_estudiantes = 50

# Generar datos aleatorios que representan puntuaciones en matemáticas, ciencias, lengua y horas de estudio para cada una.

puntuaciones_matematicas = np.random.normal(75, 10, num_estudiantes)

puntuaciones_ciencias = np.random.normal(70, 12, num_estudiantes)

puntuaciones_lengua = np.random.normal(78, 8, num_estudiantes)
```

```
horas_estudio_matematicas = np.random.normal(5, 1.5,
num_estudiantes)

horas_estudio_ciencias = np.random.normal(4, 1.2, num_
estudiantes)

horas_estudio_lengua = np.random.normal(4.5, 1.3, num_
estudiantes)

# Juntar los datos en una matriz

datos = np.vstack((puntuaciones_matematicas,
puntuaciones_ciencias, puntuaciones_lengua,

                   horas_estudio_matematicas, horas_estudio_
ciencias, horas_estudio_lengua))

# Calcular la matriz de covarianza

matriz_covarianza = np.cov(datos)

# Imprimir la matriz de covarianza

print("Matriz de Covarianza:")

print(matriz_covarianza)
```

3. Interpretación del coeficiente de correlación de Pearson

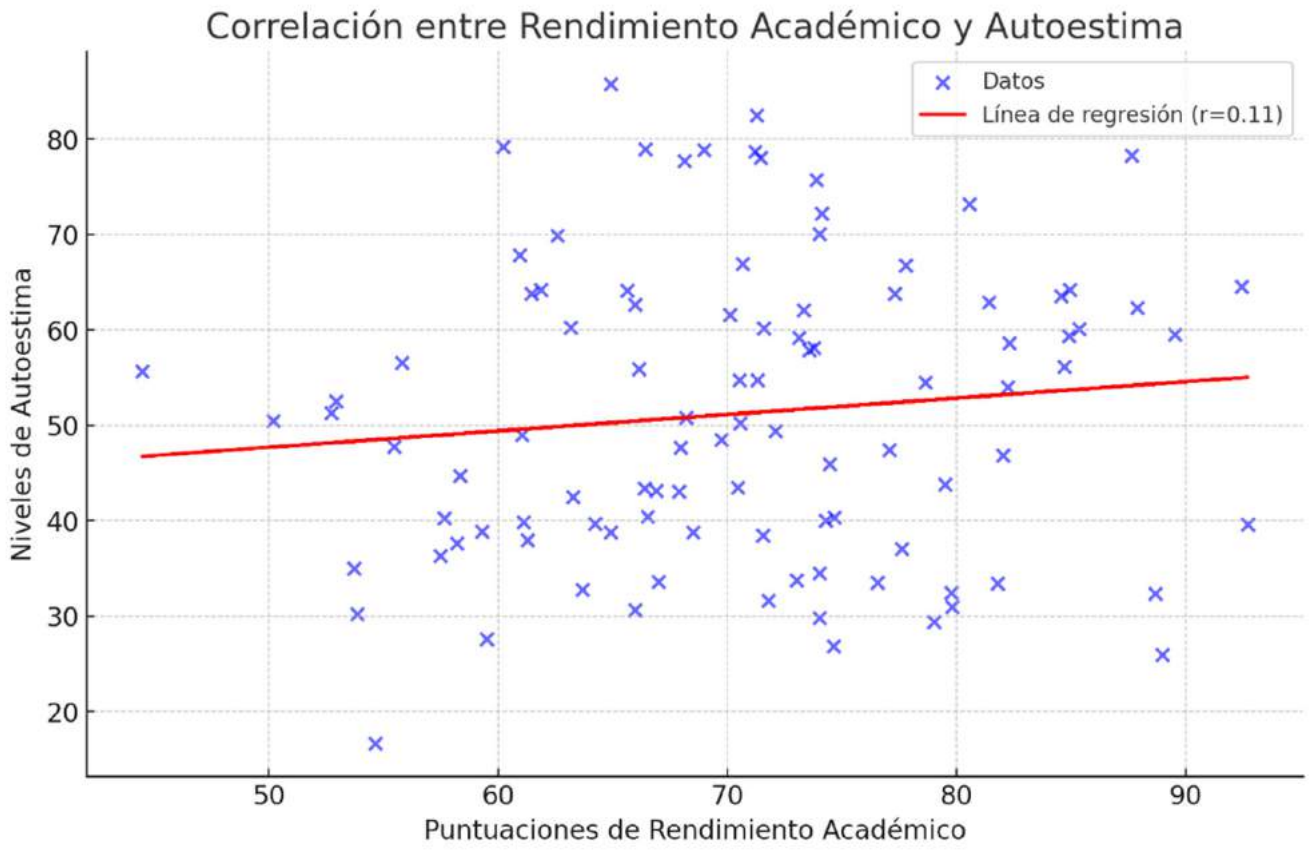
Los datos consistirán en las puntuaciones de rendimiento académico y los niveles de autoestima para un grupo de estudiantes de secundaria. Nuestra hipótesis es que una mayor autoestima está asociada con mejores resultados académicos.

- Preparación de Datos: Generaremos datos simulados para puntuaciones de rendimiento y autoestima.
- Cálculo del Coeficiente de Pearson: Usaremos la biblioteca scipy para calcular el coeficiente de Pearson.
- Interpretación: Analizaremos el coeficiente para entender la relación entre las variables.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import pearsonr, linregress
# Semilla para reproducibilidad
np.random.seed(0)
# Número de estudiantes
num_estudiantes = 100
# Datos simulados
puntuaciones_rendimiento = np.random.normal(70, 10, num_estudiantes)  # Media 70, desviación estándar 10
niveles_autoestima = np.random.normal(50, 15, num_estudiantes)  # Media 50, desviación estándar 15
# Cálculo del coeficiente de correlación de Pearson
coeficiente_pearson, p_value = pearsonr(puntuaciones_rendimiento, niveles_autoestima)
# Datos para la línea de regresión
```

```
slope, intercept, r_value, p_value_reg, std_err =
linregress(puntuaciones_rendimiento, niveles_autoestima)

# Crear un scatter plot

plt.figure(figsize=(10, 6))

plt.scatter(puntuaciones_rendimiento, niveles_autoestima,
color='blue', alpha=0.6, edgecolor='k', label='Datos')

plt.plot(puntuaciones_rendimiento, intercept + slope *
puntuaciones_rendimiento, color='red', label=f'Línea de
regresión (r={coeficiente_pearson:.2f})')

# Decoración del gráfico

plt.title('Correlación entre Rendimiento Académico y
Autoestima')

plt.xlabel('Puntuaciones de Rendimiento Académico')

plt.ylabel('Niveles de Autoestima')

plt.legend()

plt.grid(True)

# Mostrar el gráfico

plt.show()

# Imprimir resultados del coeficiente de Pearson

print(f"Coeficiente de correlación de Pearson: {coeficiente_
pearson}")

print(f"Valor p asociado a la correlación: {p_value}")
```

- Generación de Datos: Se crean conjuntos de datos para rendimiento y autoestima.
- Cálculo de Pearson: Se calcula el coeficiente de correlación de Pearson y su valor p.
- Regresión Lineal: Se calcula la línea de regresión para los datos.
- Gráfico Scatter Plot: Se genera un scatter plot con la línea de regresión. El coeficiente de correlación (r) se muestra en la leyenda para indicar la fuerza y la dirección de la relación lineal.
- Visualización y Resultados: El gráfico se muestra con todos los elementos decorativos necesarios, y se imprime el coeficiente y el valor p.



El coeficiente de correlación de Pearson resultó en 0.112, lo que indica una correlación lineal positiva muy débil entre las puntuaciones de rendimiento académico y los niveles de autoestima. El valor p asociado a esta correlación es 0.268, lo que sugiere que la correlación no es estadísticamente significativa en el nivel convencional del 0.05. Esto implica que, según nuestros datos simulados, no hay evidencia suficiente para afirmar que existe una relación fuerte y significativa entre el rendimiento académico y la autoestima de los estudiantes.

4. Correlación no implica causalidad: entendiendo la diferencia

"Correlación no implica causalidad" es un principio fundamental en estadística y ciencia que destaca la necesidad de ser cauteloso al interpretar relaciones entre variables. ¿Por qué?:

Correlación: La correlación describe la relación estadística entre dos variables. Si dos variables están correlacionadas, significa que hay una asociación entre ellas: cuando una variable cambia, la otra también tiende a cambiar de alguna manera. La correlación puede ser positiva (ambas variables cambian en la misma dirección), negativa (ambas variables cambian en direcciones opuestas) o nula (no hay una relación lineal entre las variables). Sin embargo, la correlación no proporciona información sobre la dirección de la relación ni sobre si una variable causa cambios en la otra.

Causalidad: La causalidad implica una relación de causa y efecto entre dos variables. Significa que un cambio en una variable cau-

sa un cambio en la otra variable. Establecer causalidad requiere evidencia adicional más allá de la correlación. Se necesitan experimentos controlados, estudios longitudinales u otras técnicas para determinar si existe una relación causal entre las variables. La causalidad implica una conexión más profunda y directa entre las variables que va más allá de una simple asociación estadística.

Es importante comprender que solo porque dos variables estén correlacionadas, no significa necesariamente que una cause la otra. Podría haber otros factores desconocidos o no considerados que influyan en la relación entre las variables. Por ejemplo, podría haber una correlación entre el consumo de helado y la tasa de ahogamientos en piscinas durante el verano, pero esto no significa que comer helado cause ahogamientos. En realidad, ambos pueden estar relacionados con la temperatura, que es el factor subyacente que influye en ambos.

5. Otros coeficientes de correlación: Spearman y Kendall

Coeficiente de Correlación de Spearman:

El coeficiente de correlación de Spearman se utiliza para medir la relación monotónica entre dos variables. A diferencia del coeficiente de correlación de Pearson, que asume que las variables están distribuidas normalmente y que la relación es lineal, el coeficiente de Spearman no hace tales suposiciones. En su lugar, se basa en el rango de los datos, lo que lo hace adecuado para datos ordinales o no paramétricos.

El coeficiente de Spearman se calcula rangificando los datos en lugar de utilizar los valores originales de las variables. Luego, se calcula el coeficiente de correlación de Pearson entre los rangos de las dos variables. El resultado es un valor que varía entre -1 y 1, donde 1 indica una correlación perfectamente positiva, -1 indica una correlación perfectamente negativa y 0 indica ausencia de correlación.

Coeficiente de Correlación de Kendall:

El coeficiente de correlación de Kendall es otro método para medir la concordancia entre dos conjuntos de datos. Al igual que el coeficiente de Spearman, el coeficiente de Kendall no asume ninguna distribución particular de los datos y es adecuado para datos ordinales o no paramétricos.

Este coeficiente mide la concordancia de los rangos entre las dos variables. Se calcula contando las concordancias y discordancias entre los pares de observaciones en las dos variables. El coeficiente de correlación de Kendall se interpreta de manera similar al de Spearman, donde 1 indica una correlación perfectamente positiva, -1 indica una correlación perfectamente negativa y 0 indica ausencia de correlación.

```
import numpy as np

import matplotlib.pyplot as plt

from scipy.stats import spearmanr, kendalltau

# Semilla para reproducibilidad

np.random.seed(0)

# Número de estudiantes

num_estudiantes = 100

# Datos simulados

puntuaciones_rendimiento = np.random.normal(70, 10,
num_estudiantes)  # Media 70, desviación estándar 10

niveles_autoestima = np.random.normal(50, 15, num_
estudiantes)  # Media 50, desviación estándar 15

# Cálculo del coeficiente de correlación de Spearman

coeficiente_spearman, p_value_spearman =
spearmanr(puntuaciones_rendimiento, niveles_autoestima)

# Cálculo del coeficiente de correlación de Kendall

coeficiente_kendall, p_value_kendall =
kendalltau(puntuaciones_rendimiento, niveles_autoestima)

# Imprimir resultados

print(f"Coeficiente de correlación de Spearman: {coeficiente_
spearman}")

print(f"Valor p asociado a Spearman: {p_value_spearman}")

print(f"Coeficiente de correlación de Kendall: {coeficiente_
kendall}")

print(f"Valor p asociado a Kendall: {p_value_kendall}")
```

6. Comprensión lectora

¿Qué indica una covarianza positiva entre dos variables?

A. Las variables tienden a cambiar en direcciones opuestas.

B. Las variables tienden a cambiar en la misma dirección.

C. Las variables no están relacionadas.

D. Las variables tienen una relación perfecta.

ANSWER: **B**

¿Qué medida utiliza la covarianza para estandarizar sus resultados y obtener un coeficiente de correlación?

A. La suma de cuadrados

B. El producto de las desviaciones estándar de las variables

C. La diferencia de las medias de las variables

D. La suma de las medias de las variables

ANSWER: **B**

¿Cuál es el rango de valores que puede tomar el coeficiente de correlación de Pearson?

A. 0 a 1

B. -1 a 1

C. -1 a 0

D. 0 a infinito

ANSWER: **B**

¿Qué significa un coeficiente de correlación de Pearson de -1?

A. Correlación positiva perfecta

B. Correlación negativa perfecta

C. No hay correlación

D. Correlación indefinida

ANSWER: **B**

¿Cómo se calcula la covarianza en Python usando numpy?

A. np.covariance()

B. np.var()

C. np.cov()

D. np.corrcoef()

ANSWER: **C**

¿Qué refleja un coeficiente de correlación de Pearson cercano a 0?

A. Una fuerte relación positiva

B. Una fuerte relación negativa

C. La ausencia de relación lineal

D. Una relación no lineal perfecta

ANSWER: **C**

¿Cuál es la principal diferencia entre la covarianza y la correlación?

A. La correlación es siempre negativa, mientras que la covarianza es positiva.

B. La covarianza está normalizada, mientras que la correlación no lo está.

C. La correlación está normalizada, mientras que la covarianza no lo está.

D. No hay diferencias; son términos intercambiables.

ANSWER: **C**

En el contexto de análisis estadístico, ¿qué significa "correlación no implica causalidad"?

A. Que la correlación entre dos variables nunca puede indicar una relación causal.

B. Que una correlación, aunque significativa, no puede establecer una relación causal sin evidencia adicional.

C. Que las correlaciones siempre indican causas directas.

D. Que la correlación y la causalidad son conceptualmente lo mismo.

ANSWER: **B**

¿Qué tipo de datos es apropiado para el coeficiente de correlación de Spearman?

A. Datos continuos normalmente distribuidos

B. Datos que siguen una distribución binomial

C. Datos ordinales o no paramétricos

D. Solo datos categóricos

ANSWER: **C**

¿Qué coeficiente de correlación es adecuado para medir la concordancia entre rangos?

A. Coeficiente de Pearson

B. Coeficiente de Spearman

C. Coeficiente de Kendall

D. Coeficiente de regresión

ANSWER: **C**

¿Qué indica un coeficiente de correlación de Spearman de 1?

A. Correlación negativa perfecta

B. Correlación positiva perfecta

C. No hay correlación

D. Datos insuficientes para determinar la correlación

ANSWER: **B**

¿Cuál es la principal utilidad del coeficiente de correlación de Ken-

dall en comparación con Pearson y Spearman?

A. Establece causalidad entre variables

B. Mide la concordancia de los rangos de datos

C. Calcula la varianza entre dos variables

D. Es menos sensible a datos atípicos

ANSWER: **B**

¿Cuál es el primer paso para calcular la covarianza de un conjunto de datos?

A. Sumar todas las variables

B. Normalizar los datos

C. Calcular la media de cada variable

D. Organizar los datos en una matriz

ANSWER: **C**

¿Qué debe considerarse al interpretar los valores del coeficiente de correlación de Pearson?

A. Que un valor más alto siempre indica una relación más fuerte

B. Que el coeficiente puede indicar tanto la fuerza como la dirección de la relación

C. Que solo los valores positivos son significativos

D. Que los valores cercanos a cero indican relaciones complejas

ANSWER: **B**

¿Cómo se interpretaría una correlación de Pearson de 0.85 entre las horas de estudio y las puntuaciones en matemáticas?

A. Indica una relación negativa fuerte entre las horas de estudio y las puntuaciones

B. Indica que no hay relación entre las horas de estudio y las puntuaciones

C. Sugiere una relación positiva fuerte entre las horas de estudio y las puntuaciones

D. Demuestra causalidad entre las horas de estudio y las puntuaciones

ANSWER: **C**

Referencias

Bibliográficas

Anghelache, C., Pârţachi, I., Anghel, M., Sacala, C., & Marinescu, A. (2016). Statistical-econometric Model for dispersion Analysis. Romanian Statistical Review Supplement, 64, 94-102.

Bhardwaj, A., & Sharma, K. (2013). Comparative Study of Various Measures of Dispersion. Journal of Advances in Mathematics, 1, 6-9. https://doi.org/10.24297/JAM.V1I1.6534

Bryant, W. C. (2019). Descriptive and Inferential Statistics. Companion Encyclopedia of Psychology. https://doi.org/10.4135/9781473920446.n5

Cooksey, R. (2020). Descriptive Statistics for Summarising Data. Illustrating Statistical Procedures: Finding Meaning in Quantitative Data, 61-139. https://doi.org/10.1007/978-981-15-2537-7_5

Dewey, M. (1992). Algorithms for frequency distributions: Effciency and generality comparisons. Statistics and Computing, 2, 213-220. https://doi.org/10.1007/BF01889681

Dierbach, C. (2014). Python as a first programming language. Journal of Computing Sciences in Colleges, 29, 153-154.

Dong, Y. (2023). Descriptive Statistics and Its Applications. Highlights in Science, Engineering and Technology. https://doi.org/10.54097/hset.v47i.8159

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374. https://doi.org/10.1098/rsta.2016.0360

Herschel, R., & Miori, V. M. (2017). Ethics & Big Data. Technology in Society, 49, 31-36. https://doi.org/10.1016/J.TECH-

SOC.2017.03.003

Johansson, R. (2018). Introduction to Computing with Python. Numerical Python. https://doi.org/10.1007/978-1-4842-0553-2_1

McKinney, W. (2010). Data Structures for Statistical Computing in Python. 56-61. https://doi.org/10.25080/MAJORA-92BF1922-00A

Nelli, F. (2015). The NumPy Library. 35-61. https://doi.org/10.1007/978-1-4842-0958-5_3

Nick, T. (2007). Descriptive statistics. Methods in molecular biology, 404, 33-52. https://doi.org/10.1007/978-1-59745-530-5_3

Samimi, H. (2013). Introduction to the Python programming language. Journal of Computing Sciences in Colleges, 29, 8-9.

ISBN: 978-9942-7221-6-4