DOCTRINA TECH

FUNDAMENTOS Y APLICACIONES DE LA ESTADÍSTICA DESCRÍPTIVA

Análisis e interpretación con Python y Google Colab

Borja Saltos, Thomás Ricardo
Tamami Pachala, Jorge Wilson
Salazar Guerrero, Rosario Jajaira

PRIMERA EDICIÓN 2025
EDITORIAL DOCTRINATECH

FUNDAMENTOS Y APLICACIONES DE LA ESTADÍSTICA DESCRIPTIVA

FUNDAMENTOS Y APLICACIONES DE LA ESTADÍSTICA DESCRIPTIVA

1era Edición 2025

Borja Saltos, Thomás Ricardo
Tamami Pachala, Jorge Wilson
Salazar Guerrero, Rosario Jajaira

Editorial: DOCTRINATECH S.A.S, Quito - Ecuador 2025

ISBN: 978-9942-7128-4-4

Área: PBT - Probabilidad y estadística

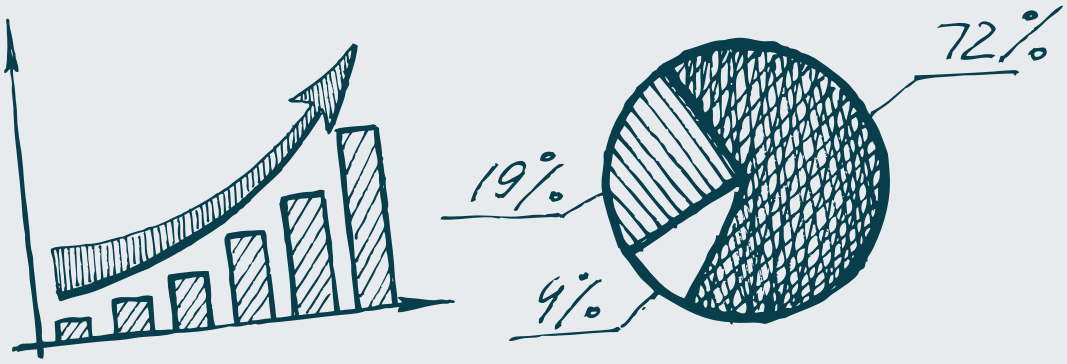
Materia: 519.5 - Matemáticas estadísticas

Tipo de Contenido: Libros universitarios

Páginas: 87

PROLOGO

0



Vivimos en una época definida por los datos, cada actividad humana desde la producción industrial hasta la investigación biomédica, genera enormes volúmenes de información que necesitan ser comprendidos, interpretados y transformados en conocimiento útil, para lo cual **la estadística** permite describirlos y analizarlos con precisión.

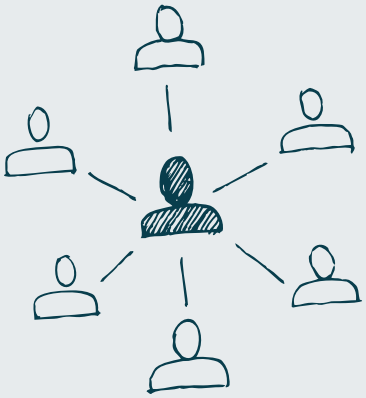
Este libro, surge con el objetivo de brindar una introducción clara y estructurada a los conceptos esenciales de la estadística descriptiva, combinando el rigor académico con un enfoque eminentemente práctico, está dirigido a estudiantes universitarios, docentes y profesionales de distintas áreas que deseen adquirir o reforzar competencias para manejar datos de forma sistemática y fundamentar sus decisiones en evidencias cuantitativas.

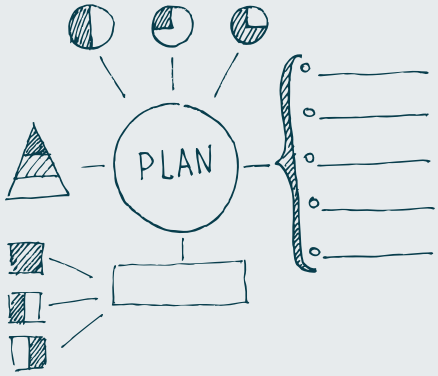
A lo largo de sus capítulos, el lector encontrará explicaciones detalladas sobre cómo recolectar, organizar, resumir y representar datos, se abordan los principales indicadores numéricos, como las medidas de tendencia central, dispersión y forma, así como las diversas técnicas gráficas que permiten visualizar las características más relevantes de los conjuntos de datos. Reconocer el papel que hoy juegan las herramientas tecnológicas es indispensable, por ello, este texto incorpora progresivamente el uso de Python, un lenguaje de programación ampliamente utilizado en análisis de datos, para mostrar cómo los procedimientos que tradicionalmente se realizaban con lápiz y papel pueden automatizarse.

Esta integración tiene un doble propósito: por un lado, demostrar la aplicabilidad inmediata de los conceptos teóricos en un entorno computacional; y por otro, reforzar la importancia de dominar los fundamentos, cabe enfatizar que este no es un manual de programación ni pretende formar programadores en Python, las rutinas computacionales se incluyen únicamente como un complemento ilustrativo, que respalda los procedimientos manuales.

FUNDAMENTOS

1





1.1 ¿Qué es la estadística? Definición y propósito

La estadística es una disciplina que forma parte esencial del pensamiento científico y del quehacer técnico en prácticamente todas las áreas del conocimiento, se ocupa del estudio sistemático de los datos: cómo recopilarlos, organizarlos, analizarlos, interpretarlos y presentarlos con el fin de comprender mejor los fenómenos que observamos y tomar decisiones fundamentadas.

Cuando se habla de estadística, se suele pensar inmediatamente en promedios, gráficos o tablas, si bien estos son componentes importantes, la estadística es mucho más que una colección de técnicas, es un proceso estructurado que permite transformar conjuntos de datos, a menudo desordenados y sin un patrón aparente, en conclusiones significativas, esto implica tanto describir lo que muestran los datos como proporcionar métodos para comprender la variabilidad inherente a casi todos los procesos observables.

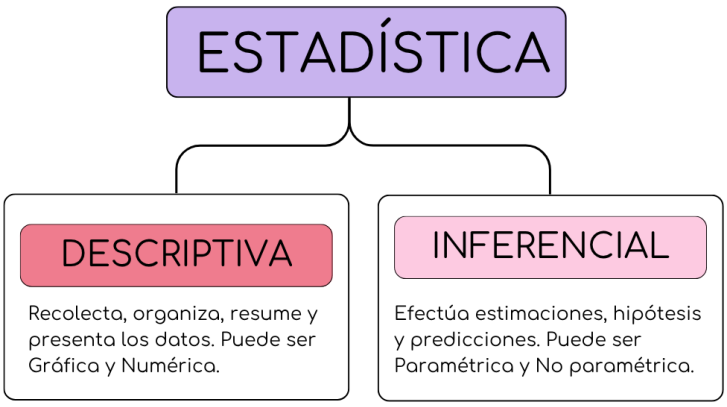
El estudio de la estadística parte del reconocimiento de que la mayoría de los fenómenos que interesan presentan fluctuaciones, ningún proceso es perfectamente constante, la estadística ofrece un lenguaje y un marco formal para describir esa variabilidad, resumirla y, en casos más avanzados, modelarla.

1.2 Estadística descriptiva y su delimitación

Dentro del vasto campo de la estadística, suele distinguirse dos grandes ramas, por un lado, la estadística descriptiva, cuyo objetivo es organizar y resumir la información contenida en un conjunto de datos concreto, sin pretender generalizar más allá de ese grupo observado, es la primera etapa del análisis estadístico y la más intuitiva, pues busca entender "qué muestran los datos" en términos claros y cuantificables.

Por otro lado, se encuentra la estadística inferencial, que utiliza la teoría de la probabilidad para extrapolar conclusiones desde una muestra a una población más amplia, estimar parámetros desconocidos y formular hipótesis que pueden ser contrastadas, si bien ambas ramas comparten herramientas y conceptos, este libro se centra exclusivamente en la estadística descriptiva, puesto que constituye el cimiento sobre el cual se edifican los análisis más complejos.

Ilustración 1



CLASIFICACIÓN DE LA ESTADÍSTICA

El estudio de la estadística no es un ejercicio abstracto, su impacto es tangible en casi todas las áreas del conocimiento y de la actividad humana por ejemplo en educación, sirve para evaluar el rendimiento de grupos de estudiantes y detectar brechas que requieren atención, en salud pública, permite monitorear la prevalencia de enfermedades y el efecto de intervenciones preventivas, en la industria, facilita el control de calidad, la optimización de procesos y la reducción de desperdicios, en economía, aporta herramientas para analizar mercados, prever tendencias y diseñar estrategias comerciales.

Así, la estadística se ha consolidado como el lenguaje de los datos, en un mundo caracterizado por la abundancia de información, ser capaz de comprender, resumir e interpretar datos de manera rigurosa se ha vuelto una habilidad fundamental para profesionales de todas las áreas.

1.3 El ciclo del análisis estadístico

Todo análisis estadístico comienza con la recopilación de datos relevantes, un paso que exige definir claramente el objetivo del estudio y los métodos de recolección adecuados dependiendo del problema, esto puede implicar diseñar encuestas, ejecutar experimentos controlados o extraer información de bases de datos existentes, una vez obtenidos los datos se estructuran en listas, tablas o conjuntos organizados que faciliten su manejo.

Una vez organizados, es necesario describir los datos para comprender sus características principales, este es el terreno de la estadística descriptiva que se ocupa de representar la información de forma concisa mediante tablas de frecuencia, gráficos ilustrativos como barras, histogramas, diagramas de caja y medidas numéricas que resumen aspectos clave como la posición central media, mediana, moda, la dispersión rango, varianza, desviación estándar y la forma de la distribución asimetría, curtosis.

Más allá de resumir, la estadística busca explicar y dar sentido a los datos, la variabilidad, es un elemento que revela la naturaleza cambiante de los fenómenos, analizar cómo se distribuyen los datos, qué tan concentrados están alrededor de un valor central o si muestran sesgos, permite interpretar lo que realmente sucede en el fenómeno estudiado.

El último paso es presentar los resultados de manera clara, honesta y útil para quienes deben usarlos, esto implica elaborar informes comprensibles, gráficos que resalten los patrones más importantes y argumentos sustentados en los análisis realizados, la estadística se convierte en un soporte imprescindible para la toma de decisiones informadas, ya sea en la gestión de empresas, en la política pública, en el diseño de procesos industriales o en el ámbito académico y científico.

1.4 Datos, población, muestra y variable

Dato

Un dato es la unidad básica de información que se obtiene al observar o medir una característica de interés, puede adoptar la forma de un número, un valor categórico o un resultado binario como “sí” o “no”. Los datos son los elementos que, al reunirse en conjunto, conforman la materia prima para el análisis estadístico. Por ejemplo:

- El peso de un estudiante (68 kg)
- El color de ojos de una persona “marrón”
- La respuesta “sí” a la pregunta “¿practica algún deporte?”

Los datos aislados suelen tener un significado limitado, es recién al agruparlos, organizarlos y analizarlos que revelan patrones o conclusiones relevantes.

Población

En estadística, el término población no se restringe a personas, sino que se usa para describir el conjunto completo de elementos o individuos que poseen una característica común y sobre los cuales se desea obtener información. Por ejemplo:

- Todos los estudiantes matriculados este semestre en una universidad forman la población para un estudio académico.
- Todos los focos fabricados en una planta durante un año constituyen la población para un análisis de calidad.

La población define el universo total de referencia para el estudio, es el grupo del que se buscan extraer conclusiones o realizar descripciones generales.

Muestra

Frecuentemente, resulta costoso, lento o incluso imposible examinar a cada elemento de la población, por ello se recurre a tomar una muestra, es decir, un subconjunto representativo de la población, seleccionado con métodos que aseguren que sus características reflejen razonablemente las del conjunto completo. Por ejemplo:

- Para conocer la satisfacción estudiantil en una universidad de 10.000 alumnos, puede encuestarse a una muestra de 500.
- Para verificar la resistencia de lotes de cemento, se prueban algunas bolsas extraídas aleatoriamente.

El análisis estadístico sobre la muestra permite luego describirla, en contextos de estadística inferencial, estimar características de toda la población, pero incluso en estadística descriptiva, el estudio suele centrarse en la muestra recolectada, que representa el objeto práctico del análisis.

Variable

Una variable es la característica o propiedad que se mide u observa en los elementos de la población o la muestra, es el aspecto concreto que genera los datos al ser observado en cada unidad. Las variables pueden adoptar diferentes valores según el individuo o elemento estudiado. Por ejemplo:

- El peso de estudiantes (medido en kg) es una variable cuantitativa, pues se expresa numéricamente.
- El color de cabello es una variable cualitativa o categórica, porque describe una cualidad o atributo sin representar magnitudes.
- La edad, la temperatura o la nota final en un curso son también variables cuantitativas.

Así, cuando se observa el peso de tres estudiantes y se obtienen los valores 65, 70 y 75 kg, estos son los datos, mientras que el peso es la variable que ha dado origen a tales datos.

1.5 Herramientas modernas

El estudio de la estadística ha experimentado una transformación radical gracias al desarrollo de la computación y de lenguajes de programación orientados al análisis de datos, hace apenas unas décadas, el cálculo manual y las hojas de papel cuadriculado eran los instrumentos principales para construir tablas de frecuencias, diagramas y obtener estadísticas básicas.

Hoy, en cambio, existen herramientas que permiten procesar volúmenes de datos impensables para los métodos tradicionales, garantizando rapidez, precisión y la posibilidad de reproducir los análisis cuantas veces sea necesario, entre los lenguajes que han impulsado esta revolución, Python ocupa un lugar destacado, originalmente concebido como un lenguaje de propósito general, se ha consolidado como la herramienta favorita para quienes trabajan en ciencia de datos, inteligencia artificial y análisis estadístico. Esto se debe a su sintaxis sencilla y legible, que facilita tanto el aprendizaje por parte de principiantes como la implementación de proyectos complejos por usuarios avanzados.

En el presente libro, Python se utilizará como soporte para automatizar procedimientos estadísticos que tradicionalmente se ejecutarían paso a paso con calculadora, permitiendo verificar resultados y dedicar más tiempo a la interpretación, por sí solo no bastaría para manejar datos de forma tabular o efectuar cálculos estadísticos complejos de manera eficiente. Para ello, se cuenta con un robusto ecosistema de bibliotecas especializadas que amplían enormemente sus capacidades, entre ellas, pandas se ha convertido en la columna vertebral del análisis de datos esta biblioteca facilita la creación y manipulación de estructuras llamadas Data Frames, similares a hojas de cálculo, que permiten almacenar datos con etiquetas de filas y columnas, filtrar registros, agrupar observaciones y aplicar funciones estadísticas sin complicaciones.

Ilustración 2

Herramientas Modernas

A través de pandas, operaciones que manualmente requerirían hojas extensas y cálculos repetitivos pueden realizarse en pocas líneas de código, garantizando resultados exactos y trazables, el soporte matemático fundamental lo provee numpy, una biblioteca que aporta estructuras eficientes para manejar vectores y matrices, junto con un amplio repertorio de funciones matemáticas. Gracias a numpy es posible calcular, por ejemplo, logaritmos y raíces de forma vectorizada, operaciones indispensables cuando se utiliza la regla de Sturges para determinar el número óptimo de clases en la construcción de tablas de frecuencia agrupadas.

El análisis estadístico no se limita a cálculos numéricos: buena parte de su poder reside en la capacidad de visualizar los datos, resaltar patrones y descubrir tendencias que difícilmente se percibirían en tablas, aquí es donde intervienen dos bibliotecas clave matplotlib y seaborn. El primero es un módulo esencial para la creación de gráficos en Python, que permite construir histogramas, diagramas de caja, polígonos de frecuencia, entre otros, personalizando cada aspecto visual del gráfico, seaborn por su parte, está construida sobre matplotlib y simplifica muchas de las tareas más frecuentes en el análisis estadístico, como representar distribuciones o comparar categorías, además de aplicar automáticamente estilos visuales que realzan la presentación.

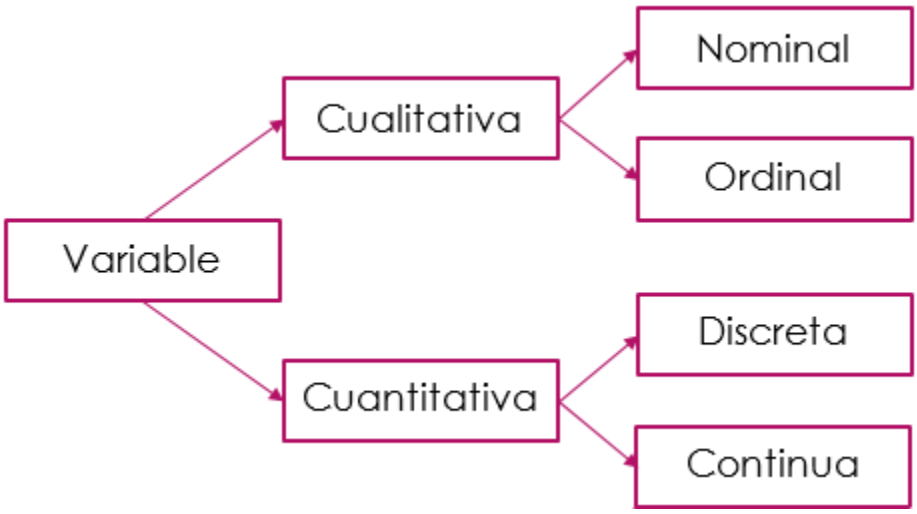
Para integrar estos recursos con la mayor comodidad posible, el libro se apoya en Google Colaboratory (Colab), un entorno que permite escribir y ejecutar código Python directamente desde el navegador, sin necesidad de instalar nada en el equipo local, al operar en la nube, Colab garantiza que los ejemplos funcionen de manera uniforme sin importar el sistema operativo, y ofrece además almacenamiento temporal y compatibilidad con Google Drive para guardar los avances. Su interfaz permite intercalar celdas de texto explicativo con código ejecutable, lo que la convierte en un espacio ideal para el aprendizaje progresivo, donde teoría, práctica y resultados pueden observarse en un mismo documento, a lo largo de los capítulos, cada concepto estadístico será primero desarrollado en su forma teórica y manual, con ejemplos trabajados paso a paso para consolidar la comprensión, posteriormente se mostrará cómo implementar el mismo proceso mediante Python en un notebook de Colab.

Las herramientas modernas no sustituyen el aprendizaje estadístico, sino que lo potencian, al manejar Python, Colab y sus bibliotecas asociadas, el lector ganará confianza no solo en los cálculos, sino también en la posibilidad de replicarlos sobre cualquier conjunto de datos, este enfoque permite consolidar la estadística como un saber aplicado, con el respaldo de tecnologías que hoy son estándar en la investigación, la industria y la toma de decisiones basada en evidencia.

1.6 Tipos de variables y escalas de medición

Al emprender el análisis estadístico de un fenómeno, es indispensable comprender primero qué tipo de variable se está estudiando y cómo se mide dicha característica, la naturaleza de la variable y su escala de medición determinan qué métodos de resumen, gráficos o cálculos son apropiados para describirla. No es lo mismo analizar la temperatura ambiental, que puede adoptar infinitos valores numéricos dentro de un rango, que clasificar el estado civil de un grupo de personas en categorías como soltero, casado o viudo.

Ilustración 3



CLASIFICACIÓN DE LAS VARIABLES ESTADÍSTICAS

Por esta razón, la estadística establece distinciones fundamentales entre variables cualitativas y cuantitativas, y describe los niveles o escalas de medición nominal, ordinal, de intervalo y de razón que permiten ubicar cada variable según el tipo de información que aporta y las operaciones matemáticas válidas para su tratamiento.

Cualitativas

Las variables cualitativas son aquellas que expresan atributos o cualidades, sin asociarse directamente a valores numéricos, su función es clasificar a los individuos u objetos en categorías o grupos con características comunes, se dividen principalmente en dos tipos:

- **Nominales:** Agrupan las observaciones en categorías sin que exista un orden natural entre ellas, el color de ojos azul, marrón, verde, el lugar de residencia urbano, rural, la marca de un producto Marca A, Marca B.
- **Ordinales:** También clasifican en categorías, pero en este caso sí existe un orden o jerarquía entre los valores posibles, aunque las diferencias entre ellos no sean cuantificables ni uniformes, por ejemplo, nivel de satisfacción bajo, medio, alto.

Para estas variables, la estadística descriptiva emplea principalmente tablas de frecuencia y gráficos como barras o sectores, que muestran cómo se distribuyen las observaciones entre las distintas categorías.

Cuantitativas

Las variables cuantitativas son aquellas que representan medidas numéricas, y permiten realizar operaciones aritméticas como sumas, restas o calcular promedios, se utilizan para describir características susceptibles de ser medidas y expresadas en números, estas variables se subdividen en:

- **Discretas:** Solo pueden tomar valores enteros, generalmente como resultado de un conteo, como número de hijos por familia, cantidad de autos vendidos en un mes.
- **Continuas:** Pueden adoptar cualquier valor dentro de un intervalo, debido a que resultan de una medición, teóricamente pueden presentar infinitos valores en un rango determinado por ejemplo la altura de las personas en metros o centímetros, el peso de productos en kg o g, el tiempo empleado en completar una tarea en minutos o segundos.

Las variables cuantitativas permiten aplicar un mayor repertorio de técnicas estadísticas, incluyendo cálculos de tendencia central, dispersión y gráficos como histogramas o diagramas de caja.

EJEMPLO 1.1

Para entender mejor estas diferencias, consideremos el análisis de un grupo de estudiantes universitarios:

- Si se registra el género masculino o femenino, se trata de una variable cualitativa nominal, porque clasifica sin jerarquía.
- Si se evalúa su nivel de satisfacción con la carrera bajo, medio, alto, se tiene una variable cualitativa ordinal, pues existe un orden implícito.
- Al medir la cantidad de cursos aprobados, se trabaja con una variable cuantitativa discreta, ya que son valores contables enteros.
- Si se mide la estatura en cm, corresponde a una variable cuantitativa continua, dado que puede tomar un número ilimitado de valores dentro de un rango.

Reconocer correctamente el tipo de variable antes de iniciar el análisis es esencial para elegir los métodos estadísticos adecuados y para interpretar los resultados de forma coherente.

Escalas

Una vez que se identifica el tipo general de variable cualitativa o cuantitativa, el siguiente paso es determinar su escala de medición, Esta clasificación es crucial porque indica el nivel de información que los datos aportan y en consecuencia, define qué operaciones matemáticas y procedimientos estadísticos son apropiados para su análisis. Existen cuatro escalas principales, ordenadas de menor a mayor nivel de sofisticación en cuanto a la cantidad y tipo de operaciones que permiten realizar: nominal, ordinal, intervalo y razón.

	DEFINICIÓN	VARIABLES	CARACTERÍSTICAS
NOMINAL	Sirve para clasificar un objeto sin establecer ningún orden.	Cualitativas nominales	Clasifica en categorías sin jerarquía, solo permite contar frecuencias.
ORDINAL	Clasifica los datos estableciendo un orden entre las categorías.	Cualitativas ordinales	Existe un orden, pero no se pueden medir diferencias exactas entre categorías.
INTERVALO	Medición numérica donde las diferencias son significativas, pero no existen ceros absolutos.	Cuantitativas continuas o discretas	Permite sumar y restar valores, no se pueden establecer razones (proporciones). Ej.: temperatura en °C.
RAZÓN	Medición numérica con diferencias iguales y ceros absolutos que indican ausencia total de la propiedad.	Cuantitativas continuas o discretas	Permite todas las operaciones aritméticas, incluyendo multiplicación y división. Ej.: peso, altura.

ESCALAS DE MEDICIÓN DE LOS DATOS

Escala nominal

La escala nominal es la forma más simple de medición, clasifica a los individuos u objetos en categorías mutuamente excluyentes, sin ningún orden inherente entre ellas. Solo permite identificar o distinguir grupos como:

- Tipo de sangre (A, B, AB, O).
- Estado civil (soltero, casado, divorciado, viudo).
- País de origen.

Para variables medidas en escala nominal, las únicas operaciones estadísticas válidas son las relacionadas con conteo y porcentaje, no tiene sentido calcular promedios o rangos en este contexto.

Escala ordinal

La escala ordinal también clasifica en categorías, pero añade un orden jerárquico entre ellas, no obstante, las distancias entre categorías no son necesariamente iguales ni cuantificables, ejemplos:

- Grado de satisfacción bajo, medio, alto.
- Nivel socioeconómico bajo, medio, alto.
- Clasificación de un torneo 1.er lugar, 2.do lugar, 3.er lugar.

Con datos ordinales se pueden establecer relaciones del tipo “mayor que” o “menor que”, pero no calcular con exactitud cuánto más o cuánto menos hay entre categorías, esto limita el uso de operaciones aritméticas y hace que las medianas y percentiles sean medidas más apropiadas que la media.

Escala de intervalo

La escala de intervalo corresponde a variables cuantitativas que, además de poseer un orden, presentan diferencias iguales y significativas entre valores, aunque carecen de un cero absoluto que indique la ausencia total de la característica medida, por ejemplos,

- Temperatura en grados Celsius o Fahrenheit.
- Fechas del calendario (por ejemplo, años).

En una escala de intervalo se pueden sumar y restar valores, así como calcular promedios, pero no es posible hacer comparaciones de tipo multiplicativo, por ejemplo, decir que 20 °C es el doble de calor que 10 °C no tiene sentido, porque el cero en la escala Celsius no representa la ausencia total de temperatura.

Escala de razón

La escala de razón es la más rica en información, cumple con todas las propiedades de las escalas anteriores clasificación, orden y distancias iguales y además posee un cero absoluto, que indica la ausencia total de la característica medida, esto permite realizar todas las operaciones aritméticas, incluidas multiplicaciones y divisiones, por ejemplo:

- Peso (0 kg significa ausencia total de masa).
- Altura (0 cm indica ausencia de longitud).
- Ingresos monetarios (0 indica falta total de ingreso).

Con variables en escala de razón es válido decir, por ejemplo, que una persona que gana $2000 tiene un ingreso el doble de quien gana $1000, o que un envase de 2 litros contiene el doble del volumen que uno de 1 litro.

Para ilustrar la diferencia entre estas escalas, consideremos:

Variable observada	Escala	Interpretación
Carrera universitaria	Nominal	Clasifica sin orden (Ingeniería, Derecho).
Nivel de satisfacción con el curso	Ordinal	Bajo, medio, alto (hay orden, sin distancias exactas).
Puntuación en examen	Intervalo	Diferencias iguales; sin cero absoluto (por ej. 0/20 no significa "ausencia de aprendizaje").
Tiempo empleado en resolver un test	Razón	Tiene cero absoluto; permite afirmar que 60 min es el doble de 30 min.

EJEMPLO 1.2

Supongamos que se realiza una encuesta a 100 estudiantes universitarios para conocer sus hábitos relacionados con el estudio y el bienestar. El cuestionario recoge la siguiente información:

- Carrera universitaria que cursa el estudiante.
- Número de horas semanales dedicadas al estudio.
- Nivel de satisfacción con su rendimiento académico (bajo, medio, alto).
- Número de cafés que consume al día.
- Peso corporal en kilogramos.
- Asiste o no a actividades deportivas (sí / no).

A partir de estas preguntas, analicemos cada variable para clasificarla según su tipo y escala.

Variable	Tipo	Subtipo	Escala	Comentario
Carrera universitaria	Cualitativa	Nominal	Nominal	Clasifica en categorías como Ingeniería, Derecho, Psicología, sin orden implícito.
Horas semanales de estudio	Cuantitativa	Continua	Razón	Puede tomar cualquier valor en el intervalo real (ej. 12,5 horas); tiene cero absoluto.
Nivel de satisfacción (bajo, medio, alto)	Cualitativa	Ordinal	Ordinal	Tiene un orden jerárquico, pero no diferencias iguales entre categorías.
Número de cafés por día	Cuantitativa	Discreta	Razón	Resulta de un conteo (0,1,2,3...); cero indica ausencia del hábito.
Peso corporal (kg)	Cuantitativa	Continua	Razón	Se mide en una escala con cero absoluto; admite todas las operaciones aritméticas.
Asistencia a deportes (sí / no)	Cualitativa	Nominal binaria	Nominal	Clasifica en dos categorías sin jerarquía.

Para el análisis estadístico en computador vamos a utilizar Python, especialmente cuando se trabaja con datos tabulares, utilizaremos su biblioteca panda, que es un paquete diseñado específicamente para manejar y analizar datos en forma de tablas estructuradas, parecidas a hojas de cálculo o tablas de bases de datos, donde se pueden realizar cálculos estadísticos, filtrar registros, agruparlos o crear resúmenes descriptivos. El objeto principal que pandas nos brinda para organizar y manejar datos es el Data Frame, que es básicamente una tabla con filas y columnas donde:

- Cada columna representa una variable como horas de estudio, peso.
- Cada fila representa un registro u observación como un estudiante con sus datos.

Esto hace que sea muy sencillo aplicar operaciones estadísticas a nivel de columnas como calcular promedios, contar frecuencias o ver la dispersión de los datos, para iniciar el análisis, primero necesitamos cargar datos en un Data Frame.

Qué instrucciones vamos a utilizar

```
import pandas as pd
```

Esta línea importa la biblioteca pandas y le da el alias pd, es una convención universal en Python, lo que facilita escribir pd.DataFrame en lugar de pandas.DataFrame.

```
df = pd.DataFrame(datos)
```

La función DataFrame de pandas sirve para crear una tabla estructurada a partir del diccionario datos, el resultado se almacena en la variable df, que será nuestro DataFrame. A partir de ese momento, df es un objeto pandas que nos permite:

- Acceder a columnas específicas.
- Obtener estadísticas descriptivas.
- Consultar los tipos de datos que ha reconocido automáticamente.

```
df.dtypes
```

Esta instrucción permite consultar el tipo de datos que pandas detectó para cada columna, por ejemplo, nos indicará si una columna fue interpretada como object (texto/categoría) o como float64 o int64 (números decimales o enteros). Esto es importante para el análisis estadístico, pues determina si podremos calcular medias, desviaciones o simplemente contar frecuencias.

```
df.describe()
```

Este método genera un resumen estadístico que incluye:

- count: cantidad de datos no nulos
- mean: media aritmética
- std: desviación estándar,
- min, max: valores mínimo y máximo,
- percentiles (o cuartiles), que describen la dispersión.

CÓDIGO 1.1

Crear un DataFrame con distintos tipos de variables.

```
import pandas as pd

datos = {‘Carrera': ['Ingeniería', 'Derecho', 'Psicología', 'Ingeniería', 'Derecho'], ‘Horas_estudio': [15.5, 10, 22, 12.5, 18], 'Nivel_satisfaccion': ['alto', 'medio', 'bajo', 'medio', 'alto'], ‘Cafes_dia': [2, 0, 3, 1, 4], ‘Peso_kg': [68.2, 74.5, 60.0, 80.3, 65.7], 'Asiste_deportes': ['sí', 'no', 'sí', 'no', 'sí']}

df = pd.DataFrame(datos)

print("DataFrame con distintas variables:")

print(df)

print("\nTipos de datos (según pandas):")

print(df.dtypes)

print("\nResumen estadístico de las variables numéricas:")

print(df.describe())
```

Salida

```
DataFrame con distintas variables:
    Carrera  Horas_estudio Nivel_satisfaccion  Cafes_dia  Peso_kg  \
0  Ingeniería         15.5             alto          2      68.2
1    Derecho         10.0            medio          0      74.5
2  Psicología         22.0             bajo          3      60.0
3  Ingeniería         12.5            medio          1      80.3
4    Derecho         18.0             alto          4      65.7
Asiste_deportes
0          sí
1          no
2          sí
3          no
4          sí
Tipos de datos (según pandas):
Carrera              object
Horas_estudio       float64
Nivel_satisfaccion   object
Cafes_dia             int64
Peso_kg             float64
Asiste_deportes      object
dtype: object
Resumen estadístico de las variables numéricas:
       Horas_estudio  Cafes_dia    Peso_kg
count       5.000000   5.000000   5.000000
mean       15.600000   2.000000  69.740000
std         4.682414   1.581139   7.869752
min        10.000000   0.000000  60.000000
25%        12.500000   1.000000  65.700000
50%        15.500000   2.000000  68.200000
75%        18.000000   3.000000  74.500000
max        22.000000   4.000000  80.300000
```

EJERCICIOS PROPUESTOS

Para cada uno de los siguientes ejemplos, indique:

- **Si la variable es cualitativa o cuantitativa.**
- **Si es cualitativa, especifique si es nominal u ordinal.**
- **Si es cuantitativa, indique si es discreta o continua.**
- **Señale la escala de medición: nominal, ordinal, intervalo o razón.**

1. Una encuesta a empleados registra el departamento donde trabajan (Producción, Ventas, Recursos Humanos, Contabilidad).
2. En un experimento físico se mide la temperatura en grados Celsius cada 5 minutos durante una hora.
3. Un hospital clasifica a los pacientes según el nivel de dolor reportado (ninguno, leve, moderado, severo).
4. Una empresa cuenta el número de productos defectuosos encontrados por semana en la línea de ensamblaje.
5. En un estudio nutricional, se anota el peso corporal de un grupo de adultos en kilogramos.
6. Se registra el código postal del domicilio de clientes de un banco.
7. En una encuesta sobre ocio, se pregunta cuántas películas vio en el cine una persona en el último mes.
8. Se evalúa el nivel académico alcanzado por un grupo de personas (primaria, secundaria, universidad, posgrado).
9. Una encuesta política anota si el encuestado planea votar en las próximas elecciones (sí / no).
10. Se mide el tiempo en segundos que tarda en completarse una operación industrial.

CÓDIGOS PROPUESTOS

Realiza los siguientes ejercicios en Google Colab:

1. Muestra solo la columna Carrera para observar los datos recogidos sobre la carrera universitaria.
2. Imprime el tipo de dato (dtype) de la columna Horas_estudio. ¿Corresponde a lo que esperarías para una variable continua?
3. Visualiza los valores de la columna Nivel_satisfaccion. ¿Puedes deducir por inspección si es ordinal o nominal, y por qué?
4. Cuenta cuántos estudiantes respondieron que asisten a actividades deportivas.
5. Calcula el número total de cafés consumidos entre cinco estudiantes del dataset (suma la columna Cafes_dia).
6. Muestra el peso promedio (Peso_kg) de los estudiantes, usando df.describe() o indexación sobre la columna.
7. Extrae y muestra solo las filas donde el número de cafés por día sea mayor o igual a 3.
8. Verifica si hay algún estudiante que no consume café y tampoco asiste a deportes. (Pista: busca filas donde Cafes_dia sea 0 y Asiste_deportes sea "no").
9. Muestra el resumen estadístico (describe()) únicamente para la columna Horas_estudio, sin imprimir el resto.
10. Imprime el tipo de dato de todas las columnas para confirmar si pandas reconoció adecuadamente las variables numéricas y las categóricas.

ORGANIZACIÓN Y PRESENTACIÓN DE DATOS

2

2.1 Tablas de Frecuencia

En estadística descriptiva, una de las herramientas más fundamentales para organizar y resumir datos es la tabla de frecuencia, ya que permiten representar de forma ordenada cómo se distribuyen los valores de una variable, mostrando no solo los datos individuales, sino también cuántas veces se repite cada uno, al construir una tabla de frecuencia se transforma un listado disperso de observaciones en una estructura comprensible, que facilita la identificación de patrones, concentraciones y posibles irregularidades en los datos.

Este recurso resulta indispensable tanto para variables cualitativas como cuantitativas, ya que proporciona una base clara para realizar posteriores análisis numéricos o gráficos, en esencia, las tablas de frecuencia constituyen el punto de partida para cualquier estudio que busque describir sistemáticamente un conjunto de datos, sirviendo como puente entre la recopilación inicial de información y su interpretación estadística.

2.2 Tablas para datos no agrupados

Cuando se dispone de datos que no presentan repeticiones excesivas ni requieren ser organizados en intervalos, es común emplear tablas de frecuencia para datos no agrupados.

Estas tablas muestran cada valor distinto de la variable y la cantidad de veces que aparece en el conjunto de datos su frecuencia, permitiendo un análisis directo de la distribución sin necesidad de agrupar, permiten resumir un conjunto de observaciones identificando cada valor observado y el número de veces que este se presenta. Este tipo de tabla es especialmente útil para:

- Variables cualitativas.
- Variables cuantitativas discretas con un rango moderado de valores.

No es recomendable para variables cuantitativas continuas con muchos valores distintos, donde se prefiere agrupar en intervalos.

Una tabla de frecuencia para datos no agrupados incluye típicamente:

- Valor (x): cada valor distinto de la variable observada
- Frecuencia absoluta (f): el número de veces que aparece ese valor en el conjunto de datos.
- Frecuencia relativa (fr): proporción que representa esa frecuencia respecto al total de datos, calculada como $fr = \frac{f}{n}$, donde n es el número total de observaciones.
- Frecuencia acumulada (F): suma progresiva de las frecuencias absolutas, que indica cuántas observaciones son menores o iguales a un determinado valor.

EJEMPLO 2.1

Supongamos que se realizó una pequeña encuesta a quince estudiantes universitarios, preguntándoles cuántas materias inscribieron este semestre, los datos recolectados fueron los siguientes:

4, 3, 5, 4, 4, 6, 3, 5, 4, 4, 5, 6, 5, 3, 4

Aquí, la variable “número de materias inscritas” es una variable cuantitativa discreta, porque se obtiene mediante conteo y toma valores enteros.

Construcción de la tabla de frecuencia simple

Paso 1: Identificar los valores distintos

Al observar los datos, los valores únicos que aparecen son:

3, 4, 5, 6

Paso 2: Contar cuántas veces aparece cada valor, frecuencia absoluta (f)

- 3 materias: aparece 3 veces
- 4 materias: aparece 6 veces
- 5 materias: aparece 4 veces
- 6 materias: aparece 2 veces

Paso 3: Calcular la frecuencia relativa (fr)

La frecuencia relativa se obtiene dividiendo la frecuencia absoluta entre el total de datos $n = 15$.

$$(fr) = \frac{f}{15}$$

Paso 4: Calcular la frecuencia acumulada (F)

Se suman progresivamente las frecuencias absolutas.

Tabla de frecuencias final

Materias inscritas (x)	Frecuencia absoluta (f)	Frecuencia relativa (fr)	Frecuencia acumulada (F)
3	3	0.20	3
4	6	0.40	9
5	4	0.27	13
6	2	0.13	15
Total	15	1.00	

Interpretación de la tabla

- La mayoría de los estudiantes (40%) inscribieron 4 materias, seguido por un 27% que inscribió 5 materias.
- La frecuencia acumulada indica, por ejemplo, que 13 de los 15 estudiantes (87%) inscribieron hasta 5 materias.

CÓDIGO 2.1

Partiendo del mismo ejemplo anterior se presenta la resolución del ejercicio en Python usando Google Colab, supongamos que se realizó una pequeña encuesta a quince estudiantes universitarios, preguntándoles cuántas materias inscribieron este semestre, los datos recolectados fueron los siguientes:

Qué instrucciones vamos a utilizar

pd.Series(...)

Esta instrucción convierte una lista simple de Python en una Serie de pandas, que es una estructura similar a una columna de una hoja de cálculo, las Series son muy útiles porque permiten aplicar de manera directa métodos estadísticos, contar valores, calcular medias o desviaciones, y mucho más, sin necesidad de escribir ciclos manuales.

value_counts()

Este método cuenta cuántas veces aparece cada valor distinto dentro de la Serie, es equivalente a construir manualmente la columna de frecuencia absoluta (f) en la tabla.

cumsum()

Este método obtiene la suma acumulativa, sirve para construir la frecuencia acumulada (F), que nos indica cuántos datos están contenidos hasta ese valor.

pd.DataFrame({...})

Aquí combinamos las tres Series (frecuencia, frecuencia_relativa y frecuencia_acumulada) en una sola tabla organizada (un DataFrame), donde cada columna corresponde a una de las frecuencias calculadas.

Codigo

```
import pandas as pd

# Crear la lista con los datos recolectados

materias = [4, 3, 5, 4, 4, 6, 3, 5, 4, 4, 5, 6, 5, 3, 4]

# Convertir la lista en una Serie de pandas

serie_materias = pd.Series(materias)

# Calcular la frecuencia absoluta (f)

frecuencia = serie_materias.value_counts().sort_index()

# Calcular la frecuencia relativa (fr)

frecuencia_relativa = frecuencia / len(serie_materias)

frecuencia_acumulada = frecuencia.cumsum()

# Construir la tabla final combinando todo en un DataFrame

tabla_frecuencia = pd.DataFrame({'f': frecuencia,'fr': frecuencia_relativa.round(2),

'F': frecuencia_acumulada})

# Mostrar la tabla de frecuencias

print("Tabla de frecuencias no agrupadas:")

print(tabla_frecuencia)
```

Salida

```
Tabla de frecuencias no agrupadas:
  f   fr  F
3 3  0.20  3
4 6  0.40  9
5 4  0.27  13
6 2  0.13  15
```

2.3 Tablas para datos agrupados: intervalos y marcas de clase

Cuando se trabaja con datos cuantitativos que tienen una gran variedad de valores distintos, como ocurre con frecuencia en variables continuas o discretas con amplio rango, las tablas de frecuencias simples se vuelven poco prácticas, en estos casos, es más apropiado agrupar los datos en intervalos, construyendo lo que se conoce como una tabla de frecuencias para datos agrupados. Este procedimiento permite resumir y organizar conjuntos voluminosos de datos, facilitando la identificación de tendencias generales y patrones que quedarían ocultos en una lista extensa de valores individuales.

Intervalos de clase

Un intervalo de clase es un rango que abarca una serie continua de valores de la variable, en lugar de mostrar cada valor por separado, se agrupan en estos intervalos, contabilizando cuántas observaciones caen dentro de cada uno. Por ejemplo, si se registran tiempos de respuesta en segundos de entre 10 y 50, se pueden construir intervalos como:

10 – 19 | 20 – 29 | 30 – 39 | 40 – 49

Cada intervalo se define por:

- Límite inferior: el menor valor que puede incluirse.
- Límite superior: el mayor valor permitido en el intervalo.

Con ello, la frecuencia absoluta f indica cuántos datos se encuentran dentro de cada rango.

Marcas de clase

Una vez construidos los intervalos, se introduce el concepto de marca de clase, que es simplemente el punto medio del intervalo. Se calcula como:

$$Marca\ de\ clase\ (x\_i) = \frac{Límite\ inferior + Límite\ superior}{2}$$

Las marcas de clase representan un valor promedio del intervalo y son muy útiles para cálculos posteriores, como estimar la media o la varianza en datos agrupados.

Regla de Sturges

La regla de Sturges es una fórmula empírica propuesta por el estadístico inglés Herbert Sturges en 1926, con el objetivo de determinar de manera aproximada cuántas clases o intervalos debe tener una tabla de frecuencias agrupadas, en función del tamaño de la muestra. Esta regla se convirtió en una referencia práctica muy utilizada en estadística descriptiva porque proporciona un número "razonable" de intervalos para resumir un conjunto de datos sin que la tabla resulte ni demasiado detallada ni demasiado simplificada.

Matemáticamente, la regla se expresa como:

$$k = 1 + 3.322 log_{10}(n)$$

donde:

- k es el número recomendado de clases o intervalos
- n es el número total de observaciones en el conjunto de datos
- log_{10} representa el logaritmo en base 10

Esta regla parte del supuesto de que los datos siguen aproximadamente una distribución normal, ajusta el número de clases según el tamaño de la muestra que indica que, a mayor número de datos, mayor cantidad de intervalos para representar bien la variabilidad. Es sencilla de aplicar, requiere solo el tamaño n de la muestra, y evita arbitrariedades al decidir el número de clases.

Proceso completo para construir la tabla de frecuencias

Paso 1. Determinar el rango (R)

El rango indica la amplitud total de los datos:

$$R = X_{max} - X_{min}$$

Donde:

- X_{max} es el valor máximo observado
- X_{min} es el valor mínimo observado

Paso 2. Calcular el número de clases (k)

Para definir cuántos intervalos o clases usar, se emplea la regla de Sturges.

Paso 3. Calcular el ancho del intervalo (c)

El ancho del intervalo se determina como:

$$c = \frac{R}{k}$$

Se recomienda redondear c a un número fácil de manejar (por ejemplo, 2, 5 o 10), según el contexto, para facilitar los cálculos y las interpretaciones.

Paso 4. Construir los intervalos

Partiendo del valor mínimo (o ligeramente menor para cubrir bien los datos), se suman incrementos de c para definir los intervalos sucesivos.

Ejemplo:

Si $X_{min} = 52, c = 5$

52–56,57–61,62–66

Paso 5. Contar frecuencias absolutas (f)

Para cada intervalo se cuenta cuántos datos caen dentro de ese rango, esto da la frecuencia absoluta (f).

Paso 6. Calcular la marca de clase (x_i)

Paso 7. Calcular la frecuencia relativa (fr)

Paso 8. Calcular la frecuencia acumulada (F)

EJEMPLO 2.2

Supongamos que tenemos un conjunto pequeño de 15 datos, correspondientes al tiempo (en minutos) que tardaron en ser atendidos 15 clientes:

12, 15, 17, 14, 20, 22, 13, 19, 25, 16, 23, 18, 21, 19, 24

Queremos organizar estos datos en una tabla de frecuencias con intervalos.

Paso 1. Determinar el rango (R)

Primero identificamos el valor máximo y el mínimo:

$Xmax = 25, Xmin = 12$

Luego calculamos el rango:

$R = Xmax - Xmin$

$R = 25 - 12$

$R = 13$

Paso 2. Calcular el número de clases con la regla de Sturges

Para $n = 15$, aplicamos la fórmula:

$k = 1 + 3.322 log_{10}(15)$

$k = 1 + 3.322x1.176$

$k = 1 + 3.91$

$k = 4.95 \approx 15$

Paso 3. Calcular el ancho del intervalo c

$c = \frac{R}{k}$

$c = \frac{13}{5}$

$c = 2.6$

Por comodidad, redondeamos a un número fácil de manejar, por ejemplo $c = 3$

Paso 4. Construir los intervalos

Comenzamos desde un punto un poco menor al valor mínimo para cubrir bien los datos (opcional, pero frecuente). Aquí empezaremos desde 12, así, los intervalos serían:

Intervalo	Límite inferior	Límite superior
1	12	14
2	15	17
3	18	20
4	21	23
5	24	26

Paso 5. Contar frecuencias absolutas (f)

Contamos cuántos datos caen en cada intervalo:

Intervalo	f
12 - 14	3
15 - 17	4
18 - 20	3
21 - 23	3
24 - 26	2

Paso 6. Calcular marcas de clase (x_i)

Intervalo	x_i
12 - 14	13
15 - 17	16
18 - 20	19
21 - 23	22
24 - 26	25

Paso 7. Calcular frecuencia relativa (fr) y acumulada (F)

$$fr = \frac{f}{15}$$

Intervalo	f	fr	F
12 - 14	3	0.20	3
15 - 17	4	0.27	7
18 - 20	3	0.20	10
21 - 23	3	0.20	13
24 - 26	2	0.13	15

Interpretación

- El intervalo con más clientes es el segundo (15-17 min), que agrupa el 27% de los casos.
- Con la frecuencia acumulada, vemos que el 67% de los clientes (10 de 15) fueron atendidos en hasta 20 minutos.

CÓDIGO 2.2

Partiendo del mismo ejemplo anterior se presenta la resolución del ejercicio en Python usando Google Colab, supongamos que tenemos un conjunto pequeño de 15 datos, correspondientes al tiempo (en minutos) que tardaron en ser atendidos 15 clientes:

Qué instrucciones vamos a utilizar

pd.DataFrame(...)

Permite crear un DataFrame, la estructura central de pandas para manejar datos en forma de tabla (con filas y columnas).

np.log10(n)

Esta función de numpy calcula el logaritmo decimal (base 10) del tamaño de la muestra n. Se usa para aplicar la regla de Sturges.

np.ceil(...)

También de numpy, ceil redondea un número hacia arriba al entero más próximo, se usa para redondear el ancho del intervalo.

pd.cut(...)

Esta función es una de las herramientas más potentes de pandas para agrupar datos en intervalos.

value_counts().sort_index()

value_counts() cuenta cuántas veces aparece cada categoría (en este caso, cada intervalo).

sort_index() reordena los intervalos según su orden lógico (por defecto, value_counts los ordena por frecuencia descendente).

cumsum()

Es un método de pandas que calcula la suma acumulada sobre una Serie, sirve para construir la frecuencia acumulada (F).

pd.DataFrame({...})

Finalmente, reunimos todas las columnas (marca, f, fr y F) en un solo DataFrame para formar la tabla final, así obtenemos una tabla ordenada y lista para análisis e interpretación.

Código

```
# Importar las librerías necesarias
import pandas as pd
import numpy as np
# Datos originales: tiempos de atención en minutos
datos = [12, 15, 17, 14, 20, 22, 13, 19, 25, 16, 23, 18, 21, 19, 24]
df = pd.DataFrame({'Tiempo': datos})
# Calcular número de clases con la regla de Sturges
n = len(df)
k = int(round(1 + 3.322 * np.log10(n)))
print(f"Número de clases (k) según Sturges: {k}")
```

```
# Calcular rango y ancho del intervalo
rango = df['Tiempo'].max() - df['Tiempo'].min()
c = np.ceil(rango / k)
print(f"Rango: {rango}, Ancho del intervalo (c): {c}")
# Construir los intervalos con pd.cut
limite_inferior = df['Tiempo'].min()
limite_superior = limite_inferior + k * c
bins = np.arange(limite_inferior, limite_superior + c, c)
df['Intervalo'] = pd.cut(df['Tiempo'], bins=bins, right=False)
# Calcular la frecuencia absoluta
frecuencia = df['Intervalo'].value_counts().sort_index()
# Calcular la frecuencia relativa
fr = frecuencia / n
# Calcular la frecuencia acumulada
F = frecuencia.cumsum()
# Calcular la marca de clase
marcas = [(interval.left + interval.right) / 2 for interval in frecuencia.index]
# Construir tabla final
tabla = pd.DataFrame({
   'x_i (marca)': marcas,
   'f': frecuencia,
   'fr': fr.round(2),
   'F': F
})
print("\nTabla de frecuencias agrupadas con marcas de clase:")
print(tabla)
```

Salida

```
Número de clases (k) según Sturges: 5
Rango: 13, Ancho del intervalo (c): 3.0
Tabla de frecuencias agrupadas con marcas de clase:
          x_i (marca)  f   fr   F
Intervalo
[12.0, 15.0)       13.5  3  0.20   3
[15.0, 18.0)       16.5  3  0.20   6
[18.0, 21.0)       19.5  4  0.27  10
[21.0, 24.0)       22.5  3  0.20  13
[24.0, 27.0)       25.5  2  0.13  15
```

EJERCICIOS PROPUESTOS

Tema: Tablas para datos no agrupados

Ejercicio 1

Se registra el número de mascotas que tienen 12 familias:

1, 0, 2, 1, 3, 2, 1, 0, 2, 3, 2, 1

Construir la tabla de frecuencias (f, fr, F).

Ejercicio 2

Horas semanales dedicadas a leer por 10 estudiantes:

3, 4, 2, 3, 5, 4, 2, 3, 4, 3

Elaborar la tabla de frecuencia simple y responder: ¿Qué porcentaje dedica exactamente 3 horas?

Ejercicio 3

Calificaciones (sobre 10) de un grupo de 8 estudiantes en un examen:

7, 8, 9, 7, 6, 8, 7, 9

Determinar la frecuencia relativa para cada calificación.

Ejercicio 4

Número de cafés consumidos en un día por 15 empleados:

1, 0, 2, 1, 3, 2, 1, 0, 2, 3, 2, 1, 0, 2, 3

Elaborar la tabla (f, fr, F) y calcular cuántos consumen hasta 2 cafés diarios.

Ejercicio 5

Encuesta sobre cantidad de veces que se fue al cine en el último mes (20 personas):

0, 1, 2, 1, 0, 3, 2, 1, 0, 2, 1, 3, 2, 1, 0, 2, 1, 0, 3, 1

Hacer la tabla de frecuencias.

Ejercicio 6

Número de hijos por 10 familias:

2, 3, 2, 1, 4, 3, 2, 1, 2, 3

Construir tabla y calcular el porcentaje de familias con más de 2 hijos.

Ejercicio 7

Número de libros leídos en el año por 12 estudiantes:

5, 7, 5, 6, 8, 5, 7, 6, 5, 8, 7, 6

Hacer tabla (f, fr, F) y señalar cuál es la frecuencia acumulada al llegar a 7 libros.

Ejercicio 8

Cantidad de veces que usaron transporte público en una semana (8 personas):

4, 5, 4, 3, 6, 5, 4, 5

Determinar la tabla y responder: ¿Qué proporción usó transporte al menos 5 veces?

Ejercicio 9

Veces que asistieron al gimnasio en el mes (12 personas):

0, 2, 1, 3, 2, 1, 0, 2, 1, 0, 2, 1

Elaborar tabla y calcular frecuencia relativa acumulada hasta 2 asistencias.

Ejercicio 10

Días de vacaciones tomados por 10 empleados:

10, 12, 10, 15, 12, 14, 10, 15, 12, 14

Crear la tabla (f, fr, F) e indicar qué porcentaje tomó 10 o 12 días.

CÓDIGOS PROPUESTOS

Tema: Tablas para datos no agrupados

Ejercicio 1

Datos: número de veces que un grupo de 10 personas consumió café en la semana:

2, 3, 2, 1, 3, 2, 1, 2, 3, 1

- Crea un DataFrame o Series con estos datos.
- Calcula la tabla de frecuencias no agrupadas (f, fr, F).
- ¿Cuál es la frecuencia relativa acumulada hasta 2 cafés?

Ejercicio 2

Datos: número de películas vistas en el último mes por 12 estudiantes:

1, 0, 2, 1, 1, 3, 2, 1, 0, 2, 1, 3

- Mostrar la tabla completa (f, fr, F).
- ¿Qué porcentaje no fue al cine ninguna vez?

Ejercicio 3

Datos: cantidad de materias inscritas por 15 alumnos:

4, 3, 5, 4, 4, 6, 3, 5, 4, 4, 5, 6, 5, 3, 4

- Construir la tabla de frecuencias no agrupadas.
- ¿Cuál es el valor modal (la cantidad de materias más frecuente)?

Ejercicio 4

Datos: número de hijos en 10 familias:

2, 3, 2, 1, 4, 3, 2, 1, 2, 3

- Elaborar la tabla (f, fr, F).
- ¿Qué proporción tiene hasta 2 hijos?

Ejercicio 5

Datos: calificaciones en un examen (sobre 10) de 8 estudiantes:

7, 8, 9, 7, 6, 8, 7, 9

- Crear la tabla y determinar qué porcentaje obtuvo 8 o más.

Ejercicio 6

Un técnico registró el tiempo (en segundos) que tarda en completarse un proceso automático, con los siguientes resultados en 20 mediciones:

32, 30, 28, 34, 33, 31, 29, 35, 34, 32, 31, 33, 32, 30, 29, 33, 31, 34, 32, 33

- Determinar el rango, el número de clases según Sturges y el ancho del intervalo.
- Construir los intervalos, marcas de clase y elaborar la tabla con f, fr, F.

Ejercicio 7

En una fábrica se midieron las longitudes (en cm) de 25 piezas producidas en un lote:

51, 53, 52, 54, 51, 52, 53, 52, 53, 51, 54, 52, 53, 54, 52, 53, 52, 54, 53, 52, 51, 53, 52, 54, 53

- Calcular el número de intervalos con la regla de Sturges.
- Construir la tabla agrupada con intervalos, marcas de clase, frecuencias y frecuencias acumuladas.

Ejercicio 8

Se registraron las temperaturas mínimas (°C) durante 30 días consecutivos en una ciudad:

18, 19, 21, 20, 19, 20, 22, 21, 20, 19,
21, 20, 19, 21, 20, 19, 18, 20, 21, 20,
19, 21, 20, 19, 20, 18, 20, 21, 20, 19

- Aplicar la regla de Sturges para obtener el número de clases.
- Determinar el ancho del intervalo.
- Elaborar la tabla con intervalos, marcas de clase, $f, fr\ y\ F$.

Ejercicio 9

Los pesos (en kg) de 15 atletas fueron los siguientes:

68, 72, 75, 70, 71, 69, 73, 74, 72, 70, 71, 69, 73, 75, 70

- Determinar rango, número de intervalos y ancho.
- Crear la tabla de frecuencias agrupadas con marcas de clase.

Ejercicio 10

Se midió la cantidad de litros consumidos de agua por día en 20 viviendas:

110, 120, 115, 125, 118, 122, 117, 119, 121, 116,
113, 124, 117, 115, 123, 118, 122, 114, 120, 119

- Hallar el número de clases usando Sturges.
- Calcular el ancho del intervalo.
- Construir la tabla final con intervalos, marcas de clase, f, fr y F.

2.3 Representación Gráfica de Datos

Gráficos de barras, sectores y pictogramas (variables cualitativas)

Cuando se trabaja con variables cualitativas, ya sean nominales (sin orden) u ordinales (con jerarquía), es fundamental utilizar representaciones gráficas que faciliten la comparación de las categorías. Las tablas de frecuencia resumen los

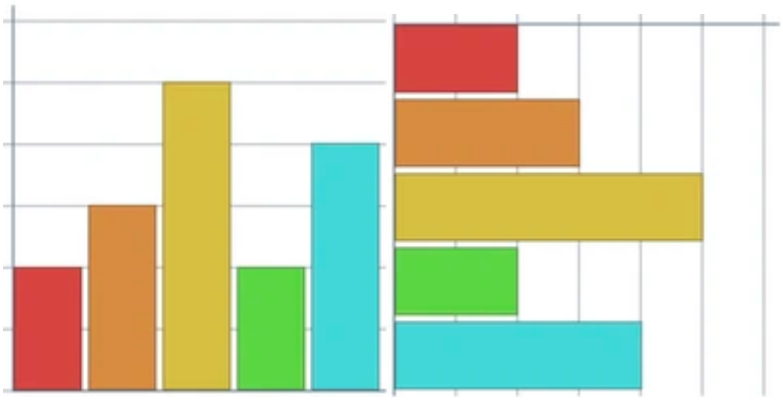
datos numéricamente, pero los gráficos hacen evidente la distribución de los datos de forma visual, ayudando a destacar categorías predominantes o menos frecuentes.

Entre los gráficos más empleados en estadística descriptiva para representar variables cualitativas se encuentran los gráficos de barras, los gráficos de sectores (circulares) y los pictogramas.

Gráfico de barras

El gráfico de barras es probablemente el más usado para datos cualitativos, cada categoría se representa mediante una barra rectangular cuya altura o longitud es proporcional a su frecuencia (absoluta o relativa).

Ilustración 5



GRÁFICOS DE BARRAS

- Puede orientarse vertical u horizontalmente.
- Es ideal tanto para variables nominales como ordinales.
- Se debe mantener la misma separación entre barras y no unirlas, pues representan clases distintas sin continuidad numérica (a diferencia de los histogramas).

Gráfico de sectores (o circular)

El gráfico de sectores, conocido popularmente como gráfico de pastel o pie chart, divide un círculo en sectores cuya amplitud es proporcional a la frecuencia o porcentaje de cada categoría, suele emplearse para mostrar la participación relativa de cada categoría respecto al total.

Es más adecuado para variables nominales, ya que el orden de las categorías alrededor del círculo no implica jerarquía.

Ventajas

- Ofrece una visión global de la composición del total.
- Destaca claramente cuál categoría tiene mayor proporción.

Ilustración 6

GRÁFICO DE SECTORES O CIRCULAR

Este tipo de grafico no resulta tan eficaz si hay muchas categorías con frecuencias similares o muy pequeñas, porque dificulta distinguir los sectores.

El pictograma

Es una forma visual que utiliza símbolos o imágenes repetidas proporcionalmente a la frecuencia, por ejemplo, si un ícono representa 5 casos, una categoría con frecuencia 15 se mostrará con 3 íconos.

Ilustración 7

PICTOGRAMA

- Es especialmente útil para presentaciones didácticas o divulgativas, ya que resulta muy intuitivo y atractivo.
- Debe usarse con cuidado para no distorsionar la percepción: los pictogramas deben ser siempre del mismo tamaño, y no usar áreas o volúmenes escalados arbitrariamente.

Elegir entre un gráfico de barras, de sectores o un pictograma depende del objetivo del análisis y del tipo de público, en todos los casos, estas representaciones cumplen con la función de traducir tablas de frecuencias cualitativas en imágenes comprensibles, facilitando la interpretación de datos categóricos y permitiendo observar rápidamente cuál o cuáles categorías predominan.

Histogramas, polígonos y ojivas (variables cuantitativas)

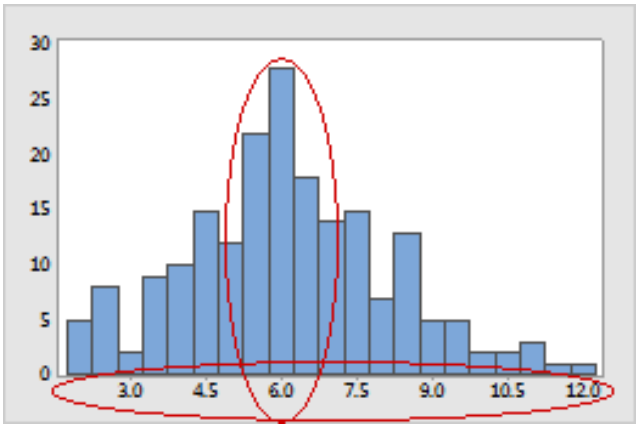
Cuando trabajamos con variables cuantitativas, especialmente continuas o discretas con muchos valores diferentes, las tablas de frecuencias agrupadas son fundamentales para resumir los datos, para visualizar la distribución de los datos y entender rápidamente su forma, concentración o dispersión, se utilizan representaciones gráficas específicas: el histograma, el polígono de frecuencia y la ojiva.

Estas representaciones son distintas de los gráficos usados para variables cualitativas (como barras y sectores), porque aquí se aprovecha la continuidad de la escala numérica, mostrando cómo se distribuyen los datos a lo largo del rango de valores.

Histograma

El histograma es la representación gráfica más común para variables cuantitativas agrupadas. Es similar a un gráfico de barras, pero con diferencias esenciales:

Ilustración 8



HISTOGRAMA

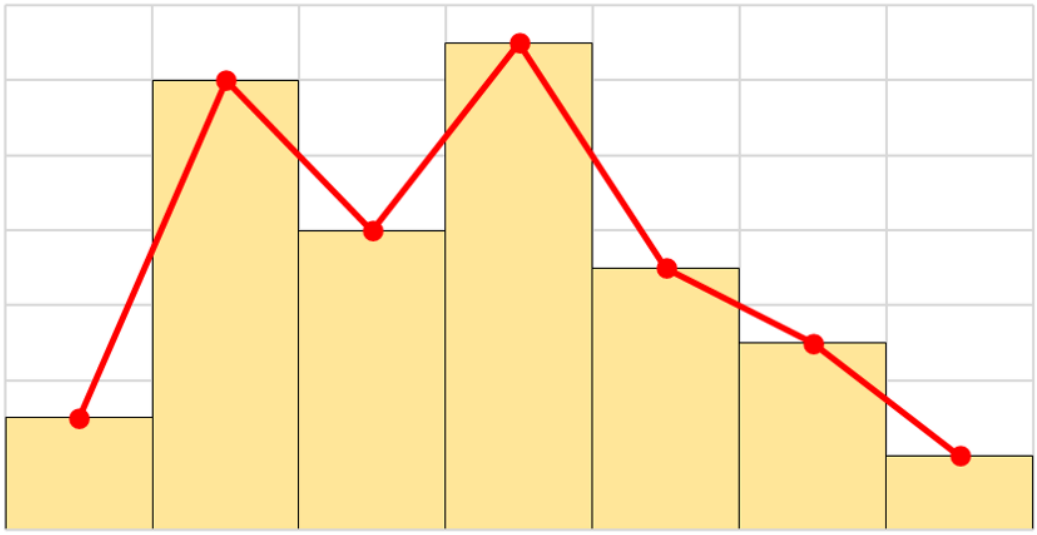
- Las barras del histograma están unidas entre sí, porque representan intervalos contiguos de una escala continua (no categorías aisladas).
- Cada barra corresponde a un intervalo de clase, y su altura indica la frecuencia absoluta o relativa.
- El eje horizontal representa los intervalos (por ejemplo, pesos, alturas, tiempos), mientras que el eje vertical muestra la frecuencia.
- En histogramas de frecuencias relativas o densidades, el área total de todas las barras suma 1 (o 100%).
- Permite apreciar la forma de la distribución: si es simétrica, sesgada a la derecha o izquierda, unimodal o bimodal.
- Ideal para detectar concentraciones o dispersiones.

Polígono de frecuencia

El polígono de frecuencia es un gráfico lineal que une mediante segmentos rectos los puntos que representan las marcas de clase frente a sus frecuencias, se construye ubicando un punto sobre cada marca de clase (eje x), a una altura igual a la frecuencia correspondiente (eje y). Luego, los puntos se conectan en orden, y es

habitual prolongar la línea hacia el eje horizontal agregando un intervalo ficticio antes y después con frecuencia cero, cerrando así el polígono.

Ilustración 9



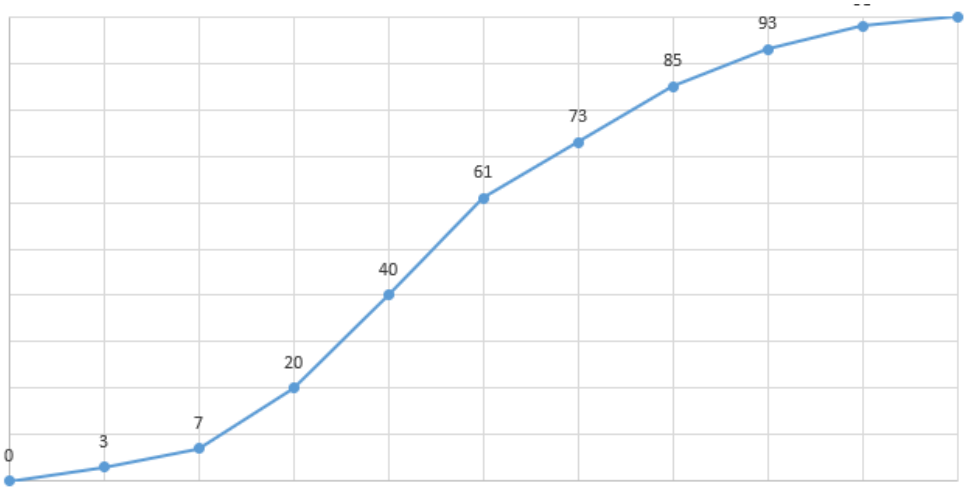
POLÍGONO DE FRECUENCIA

- El polígono de frecuencia facilita comparar varias distribuciones en el mismo gráfico, algo que resulta poco claro con histogramas superpuestos.
- Muestra con más fluidez los cambios de frecuencia a lo largo de los intervalos.

Ojiva (o gráfico de frecuencia acumulada)

La ojiva es una línea que representa la frecuencia acumulada (F) o la frecuencia relativa acumulada (Fr) frente a los límites superiores (o inferiores) de cada intervalo, se construye acumulando las frecuencias y graficándolas frente al límite correspondiente.

Ilustración 10



- Es muy útil para responder preguntas como "¿Qué porcentaje de datos está por debajo de cierto valor?"
- Permite ubicar percentiles, cuartiles o mediana de forma gráfica.
- Ideal para analizar cómo se acumulan los datos a lo largo del rango.

Los histogramas, polígonos y ojivas son herramientas gráficas esenciales en la estadística descriptiva para variables cuantitativas, cada uno aporta una perspectiva diferente: el histograma destaca la concentración, el polígono suaviza y facilita comparaciones, y la ojiva permite analizar la acumulación y localizar percentiles. Con su uso combinado, se obtiene un panorama completo del comportamiento de la variable estudiada.

4.3 Diagramas de caja (BOXPLOT)

El diagrama de caja y bigotes, conocido comúnmente por su nombre en inglés BOXPLOT, es un gráfico estadístico que resume visualmente la distribución de un conjunto de datos cuantitativos, destacando su tendencia central, dispersión y posibles valores atípicos. Se considera una herramienta fundamental en el análisis exploratorio de datos porque proporciona una descripción gráfica compacta de cinco estadísticas clave: el mínimo, el primer cuartil (Q1), la mediana (Q2), el tercer cuartil (Q3) y el máximo.

Ilustración 11

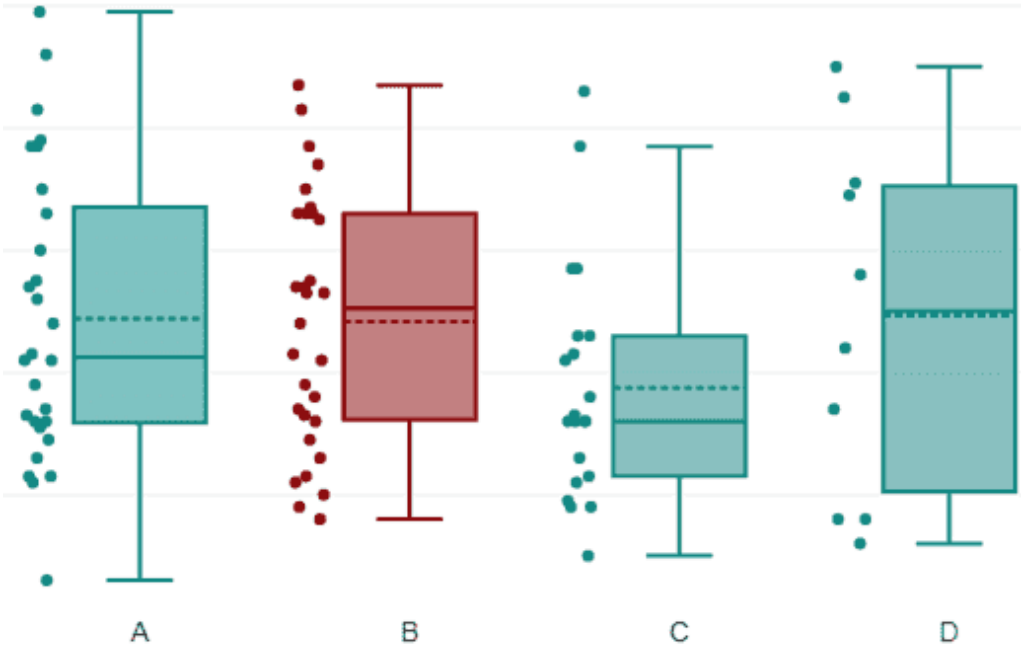


DIAGRAMA DE CAJAS (BOXPLOT)

El diagrama de cajas

- Resume de forma muy clara y gráfica la simetría o asimetría de los datos.
- Permite detectar rápidamente valores atípicos.
- Es excelente para comparar la distribución de la misma variable en diferentes grupos (por ejemplo, boxplots paralelos para comparar notas entre diferentes carreras).

¿Cómo se interpreta?

- Si la mediana está centrada en la caja y los bigotes tienen longitud similar, la distribución es aproximadamente simétrica.
- Si la mediana está más cerca de Q1, o un bigote es mucho más largo, indica sesgo (asimetría).
- Puntos fuera de los bigotes sugieren valores atípicos que pueden ser valores inusuales o errores en la recolección.

Supongamos que tenemos un conjunto de tiempos (en minutos) que tardaron estudiantes en terminar un examen, y construimos su boxplot.

- Si la mediana está cerca del centro de la caja y los bigotes son parecidos, podemos decir que los tiempos están bien distribuidos sin sesgo evidente.
- Si aparecen varios puntos fuera de los bigotes por el lado derecho, significa que hay algunos estudiantes que tardaron mucho más que la mayoría, indicando una cola hacia la derecha (distribución sesgada positivamente).

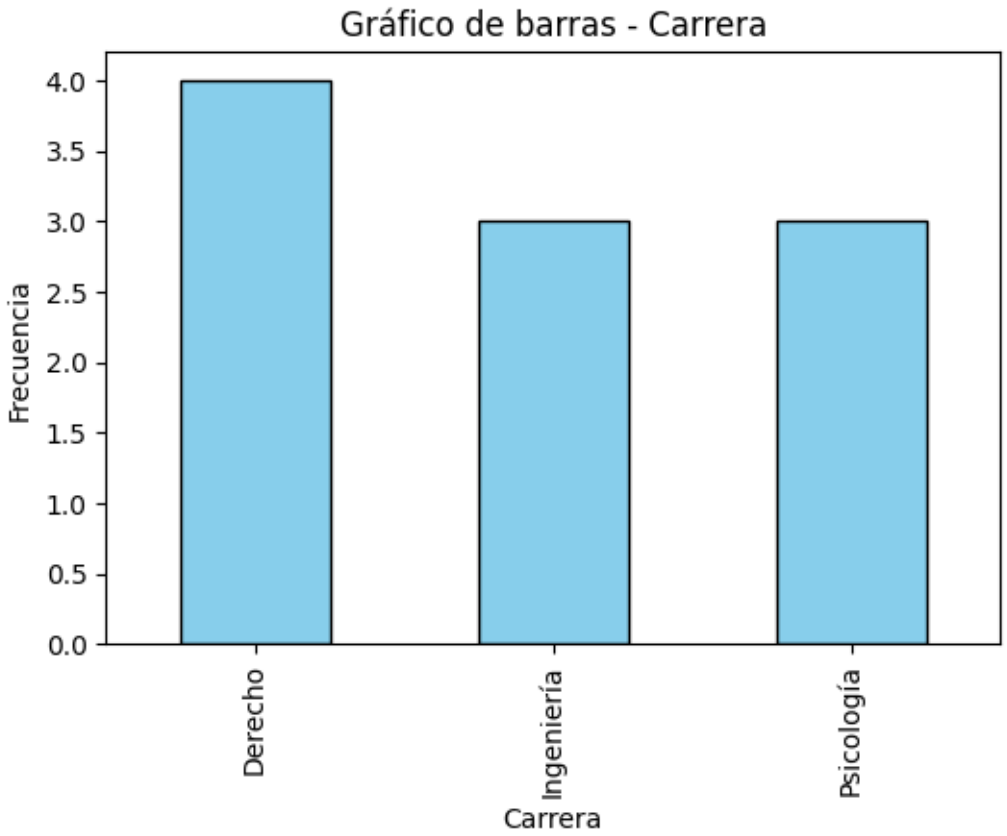
CÓDIGO 2.3

Grafico de barras en Google Colab.

Código

```
import pandas as pd
import matplotlib.pyplot as plt
# Datos cualitativos
carreras = ['Ingeniería', 'Derecho', 'Psicología', 'Ingeniería', 'Derecho',
        'Psicología', 'Derecho', 'Ingeniería', 'Psicología', 'Derecho']
df_cat = pd.DataFrame({'Carrera': carreras})
# Calcular frecuencias
frecuencia = df_cat['Carrera'].value_counts()
# Gráfico de barras
plt.figure(figsize=(6,4))
frecuencia.plot(kind='bar', color='skyblue', edgecolor='black')
plt.title('Gráfico de barras - Carrera')
plt.xlabel('Carrera')
plt.ylabel('Frecuencia')
plt.show()
```

Salida



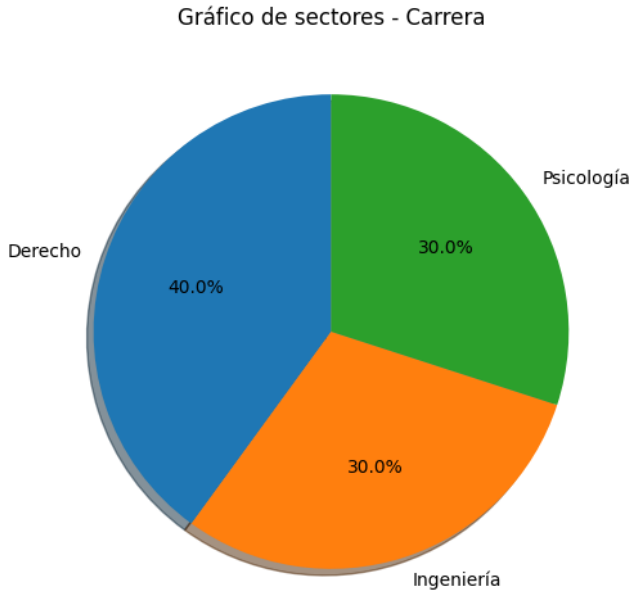
CÓDIGO 2.4

Gráfico de sectores en Google Colab.

Código

```
import pandas as pd
import matplotlib.pyplot as plt
# Datos cualitativos
carreras = ['Ingeniería', 'Derecho', 'Psicología', 'Ingeniería', 'Derecho',
        'Psicología', 'Derecho', 'Ingeniería', 'Psicología', 'Derecho']
df_cat = pd.DataFrame({'Carrera': carreras})
# Calcular frecuencias
frecuencia = df_cat['Carrera'].value_counts()
#Gráfico de sectores
plt.figure(figsize=(6,6))
frecuencia.plot(kind='pie', autopct='%1.1f%%', startangle=90, shadow=True)
plt.title('Gráfico de sectores - Carrera')
plt.ylabel('')
plt.show()
```

Salida



CÓDIGO 2.5

Pictograma en Google Colab

Código

```
import pandas as pd
import matplotlib.pyplot as plt
# Datos cualitativos
carreras = ['Ingeniería', 'Derecho', 'Psicología', 'Ingeniería', 'Derecho',
        'Psicología', 'Derecho', 'Ingeniería', 'Psicología', 'Derecho']
df_cat = pd.DataFrame({'Carrera': carreras})
# Calcular frecuencias
frecuencia = df_cat['Carrera'].value_counts()
#Pictograma con emojis
print("Pictograma simulado:")
for carrera, freq in frecuencia.items():
  print(f"{carrera}: {' 👤 ' * freq}")
```

Salida

Pictograma simulado:
Derecho: 👤 👤 👤 👤
Ingeniería: 👤 👤 👤
Psicología: 👤 👤 👤

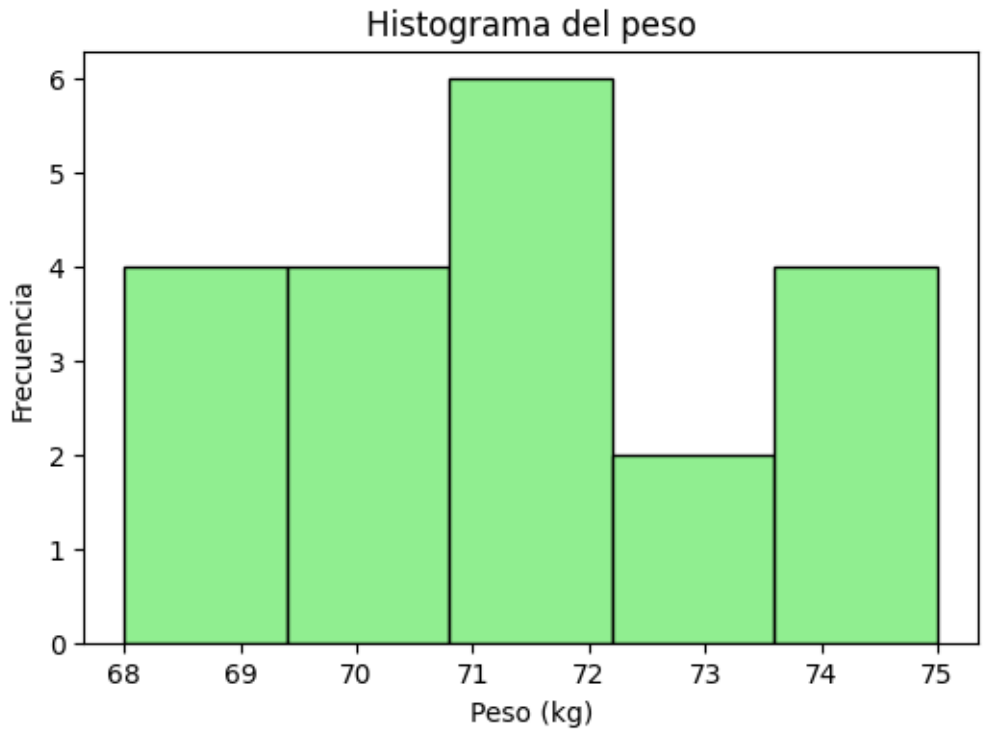
CÓDIGO 2.6

Histograma en Google Colab

Código

```
import numpy as np
# Datos cuantitativos
pesos = [68, 72, 75, 70, 71, 69, 73, 74, 72, 70, 71, 69, 73, 75, 70, 68, 72, 71, 74, 70]
# Histograma
plt.figure(figsize=(6,4))
plt.hist(pesos, bins=5, edgecolor='black', color='lightgreen')
plt.title('Histograma del peso')
plt.xlabel('Peso (kg)')
plt.ylabel('Frecuencia')
plt.show()
```

Salida



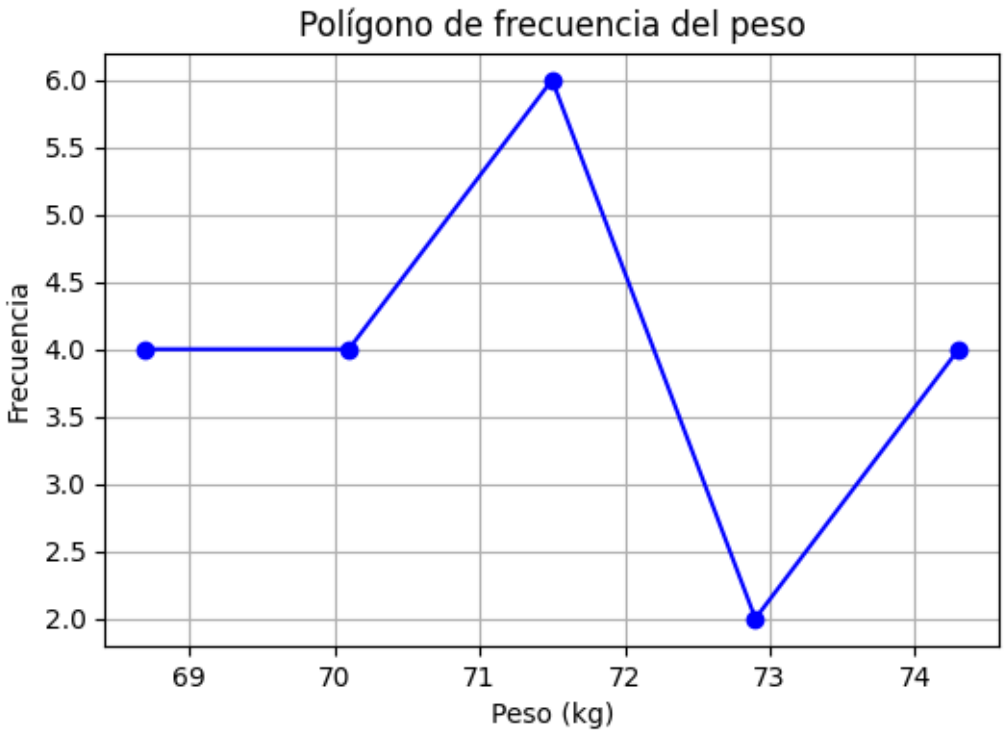
CÓDIGO 2.7

Polígono de frecuencia en Google Colab

Código

```
import numpy as np
# Datos cuantitativos
pesos = [68, 72, 75, 70, 71, 69, 73, 74, 72, 70, 71, 69, 73, 75, 70, 68, 72, 71, 74, 70]
# Calcular histograma para obtener centros y frecuencias
counts, bin_edges = np.histogram(pesos, bins=5)
bin_centers = 0.5 * (bin_edges[1:] + bin_edges[:-1])
# Polígono
plt.figure(figsize=(6,4))
plt.plot(bin_centers, counts, marker='o', linestyle='-', color='blue')
plt.title('Polígono de frecuencia del peso')
plt.xlabel('Peso (kg)')
plt.ylabel('Frecuencia')
plt.grid()
plt.show()
```

Salida



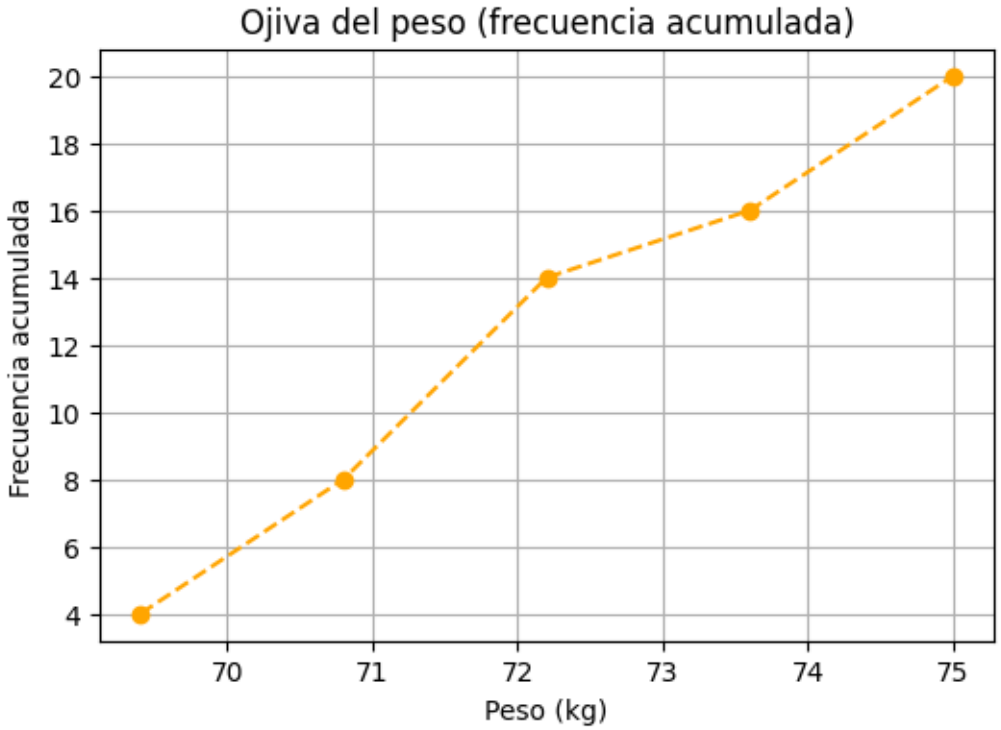
CÓDIGO 2.8

Ojiva o diagrama de frecuencia acumulada en Google Colab

Código

```
import numpy as np
# Datos cuantitativos
pesos = [68, 72, 75, 70, 71, 69, 73, 74, 72, 70, 71, 69, 73, 75, 70, 68, 72, 71, 74, 70]
# Frecuencia acumulada
cum_counts = np.cumsum(counts)
plt.figure(figsize=(6,4))
plt.plot(bin_edges[1:], cum_counts, marker='o', linestyle='--', color='orange')
plt.title('Ojiva del peso (frecuencia acumulada)')
plt.xlabel('Peso (kg)')
plt.ylabel('Frecuencia acumulada')
plt.grid()
plt.show()
```

Salida



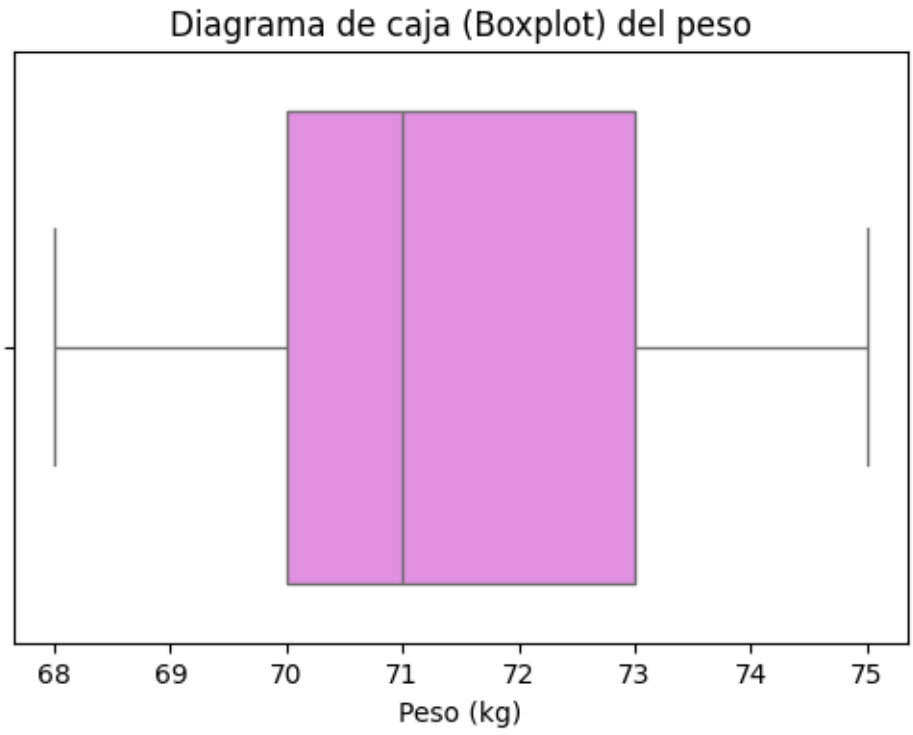
CÓDIGO 2.9

Diagrama de caja en Google Colab

Código

```
import numpy as np
import seaborn as sns
# Datos cuantitativos
pesos = [68, 72, 75, 70, 71, 69, 73, 74, 72, 70, 71, 69, 73, 75, 70, 68, 72, 71, 74, 70]
# Boxplot
plt.figure(figsize=(6,4))
sns.boxplot(x=pesos, color='violet')
plt.title('Diagrama de caja (Boxplot) del peso')
plt.xlabel('Peso (kg)')
plt.show()
```

Salida



CÓDIGOS PROPUESTOS

Tema: Gráficos Estadísticos

Ejercicio 1. Gráfico de barras (variable cualitativa)

Datos:

```
['Manzana', 'Banana', 'Manzana', 'Pera', 'Banana', 'Manzana', 'Pera',  'Banana', 'Pera', 'Banana', 'Manzana', 'Pera']
```

Construir un gráfico de barras que muestre la frecuencia de cada fruta.

Ejercicio 2. Gráfico de sectores (pie chart)

Datos:

```
['Netflix', 'Disney+', 'HBO', 'Netflix', 'HBO', 'Disney+', 'Netflix', 'Amazon', 'Disney+', 'Netflix', 'HBO', 'Amazon']
```

Hacer un gráfico de sectores que muestre la participación relativa de cada plataforma de streaming en las preferencias de 12 personas.

Ejercicio 3. Pictograma

Datos:

```
['Perro', 'Gato', 'Perro', 'Pez', 'Gato', 'Perro', 'Perro', 'Pez', 'Gato', 'Gato']
```

Imprimir un pictograma simulado usando un emoji o símbolo para representar cada mascota, repitiéndolo tantas veces como indica la frecuencia.

Ejercicio 4. Histograma

Datos:

```
[155, 160, 162, 158, 164, 167, 170, 162, 165, 159, 161, 168, 163, 169, 166, 160]
```

Construir un histograma que muestre la distribución de alturas (en cm) de 16 personas.

Ejercicio 5. Polígono de frecuencia

Datos:

[45, 48, 47, 49, 50, 51, 52, 48, 50, 47, 46, 49, 53, 51, 52, 48, 50, 47, 49, 51]

Construir un polígono de frecuencia que represente la distribución de puntuaciones en un test sobre 20 estudiantes.

Ejercicio 6. Ojiva

Datos:

[30, 32, 31, 35, 33, 36, 34, 32, 31, 34, 33, 35, 36, 34, 33, 32, 31, 35, 33, 34]

Elaborar una ojiva (frecuencia acumulada) para la cantidad de horas de estudio en un mes de 20 alumnos.

Ejercicio 7. Boxplot

Datos:

[78, 82, 85, 79, 81, 88, 90, 83, 87, 84, 82, 85, 89, 91, 86, 80, 83, 88, 85, 87]

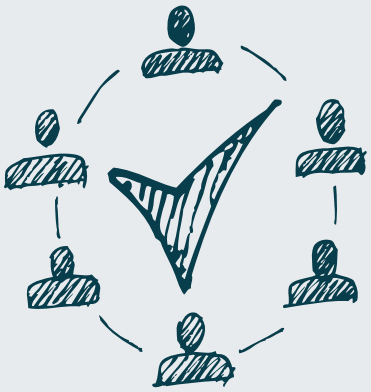
Crear un boxplot que muestre la distribución de las calificaciones finales (sobre 100) de 20 estudiantes.

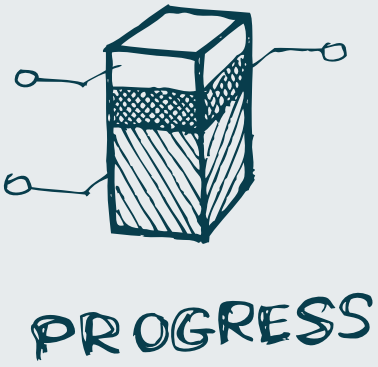
Los gráficos estadísticos —ya sean barras, sectores, pictogramas, histogramas, polígonos, ojivas o diagramas de caja— constituyen herramientas visuales esenciales para representar y comprender los datos, permitiendo identificar patrones, concentraciones, asimetrías y valores atípicos que difícilmente se apreciarían solo en tablas.

Elegir el tipo adecuado según la naturaleza de la variable (cualitativa o cuantitativa) y el objetivo del análisis es clave para comunicar de forma clara y efectiva los resultados, facilitando tanto la interpretación técnica como la divulgación a públicos no especializados.

MEDIDAS NUMÉRICAS

3





3.1 Medidas de Tendencia Central

Las medidas de tendencia central son parámetros estadísticos que permiten identificar un valor representativo alrededor del cual tienden a concentrarse los datos de un conjunto, se emplean para describir el comportamiento típico o promedio de una variable y son fundamentales para resumir grandes volúmenes de información en un solo valor que facilite su interpretación y comparación.

Entre estas medidas destacan la media aritmética, la mediana y la moda, cada una con propiedades particulares que las hacen más o menos adecuadas según la naturaleza de los datos y el objetivo del análisis, estudiarlas proporciona una primera aproximación al "centro" de los datos, sirviendo de base para análisis más avanzados sobre la dispersión o forma de la distribución.

Media aritmética simple

La media aritmética simple es la medida de tendencia central más utilizada en estadística, representa el promedio de un conjunto de datos, es decir, el valor obtenido al sumar todas las observaciones y dividir el total entre el número de datos. Matemáticamente, si se tienen n observaciones $x1, x2, x3, \dots, xn$, la media aritmética se calcula mediante la fórmula:

$$\bar{x} = \frac{x1, x2, x3, \dots, xn}{n} = \frac{\sum_{i=1}^{n} n}{n}$$

Donde:

- $\bar{x}$ representa cada uno de los valores del conjunto,
- x_i es el número total de datos
- n es el número total de datos

Características principales

- Sencillez: es fácil de calcular e interpretar, lo que la convierte en la medida más difundida.
- Uso general: se emplea para datos cuantitativos, tanto discretos como continuos.
- Sensibilidad: es afectada por valores extremos o atípicos, lo que puede desplazarla y hacer que no refleje adecuadamente el "centro" si existen datos muy alejados.

EJEMPLO 3.1

Supongamos que se registra el número de libros leídos en un mes por cinco estudiantes:

3, 5, 4, 6, 2

La media aritmética será:

$$\bar{x} = \frac{3+5+4+6+2}{4} = \frac{20}{5} = 4$$

Esto significa que, en promedio, los estudiantes leyeron 4 libros en el mes.

La media aritmética simple resume en un solo valor el comportamiento general del conjunto de datos, siempre debe analizarse complementada con medidas de dispersión para comprender cuán concentrados o dispersos están los datos respecto a ese promedio.

Media ponderada

La media ponderada es una extensión de la media aritmética que se utiliza cuando no todos los datos tienen la misma importancia o peso.

En lugar de considerar cada dato con el mismo nivel de influencia, la media ponderada asigna a cada valor un peso proporcional a su relevancia, frecuencia o impacto en el conjunto.

Definición formal

Si se tienen n datos $x1, x2, \dots, xn$, cada uno asociado a un peso $w1, w2, \dots, wn$, la media ponderada se calcula como:

$$\bar{x}_p = \frac{\sum_i^n w_i x_i}{\sum_i^n w_i}$$

Donde:

- $\bar{x}_p$ es la media ponderada
- x_i son los valores de los datos
- w_i son los pesos asociados a cada dato, que pueden representar frecuencia absoluta, importancia asignada o cualquier otro criterio.

Si todos los pesos son iguales, la media ponderada se reduce a la media aritmética simple, los pesos determinan cuánto "influye" cada dato en el promedio final: un valor con peso mayor afecta más el cálculo que uno con peso pequeño.

EJEMPLO 3.2

Un estudiante tiene tres calificaciones finales en un curso, todas sobre 10:

Actividad	Nota x_i	Ponderación w_i
Tarea	8.0	20% (=0.20)
Proyecto	9.0	30% (=0.30)
Examen final	7.0	50% (=0.50)

Fórmula general

$$\bar{x}_p = \frac{\sum_i^n w_i x_i}{\sum_i^n w_i}$$

Aplicación

Paso 1: Calcular el numerador

$$\bar{x}_p = \frac{(0.20x8.0) + (0.30x9.0) + (0.50x7.0)}{(0.20) + (0.30) + (0.50)}$$

$$\bar{x}_p = \frac{1.6 + 2.7 + 3.5}{1}$$

$$\bar{x}_p = 7.8$$

El promedio ponderado del estudiante, considerando el peso de cada actividad según la planificación del curso, es 7.8 sobre 10. Esto significa que el examen final, al representar un 50% del total, tiene el mayor impacto sobre la calificación definitiva.

CÓDIGO 3.1

Se muestra el mismo ejemplo anterior en Google Colab.

Código

```
import pandas as pd
# Datos del ejemplo
# Cada fila representa una actividad
datos = {
   'Actividad': ['Tarea', 'Proyecto', 'Examen final'],
   'Nota': [8.0, 9.0, 7.0],
   'Peso': [0.20, 0.30, 0.50] }
df = pd.DataFrame(datos)
# Calcular el producto
df['w_i * x_i'] = df['Nota'] * df['Peso']
# Sumar para hallar el promedio ponderado
numerador = df['w_i * x_i'].sum()
denominador = df['Peso'].sum()
media_ponderada = numerador / denominador
# Mostrar la tabla y el resultado
print("Tabla de cálculo de la media ponderada:")
print(df)
print(f"\nSuma de w_i * x_i: {numerador}")
print(f"Suma de w_i: {denominador}")
print(f"Media ponderada: {media_ponderada:.2f}")
```

Salida

```
Tabla de cálculo de la media ponderada:
    Actividad  Nota  Peso  w_i * x_i
0 Tarea   8.0   0.2      1.6
1 Proyecto   9.0   0.3      2.7
2  Examen final   7.0   0.5      3.5
Suma de w_i * x_i: 7.8
Suma de w_i: 1.0
Media ponderada: 7.80
```

Mediana y moda

La mediana y la moda complementan a la media en el análisis de los datos, la mediana muestra el punto central sin verse afectada por valores extremos, mientras que la moda indica el valor más frecuente, siendo particularmente útil para observar concentraciones o preferencias. Juntas ofrecen una visión más completa de la distribución.

Mediana

La mediana es el valor que ocupa la posición central en un conjunto de datos ordenados de menor a mayor, divide al conjunto en dos partes iguales: el 50% de los datos es menor o igual a la mediana y el otro 50% es mayor o igual.

- Si el número de datos (n) es impar, la mediana es el valor que queda exactamente en el medio.
- Si n es par, la mediana se obtiene promediando los dos valores centrales.

Es una medida de posición, no influenciada por valores extremos o atípicos, lo que la hace especialmente útil cuando la distribución está sesgada, es apropiada tanto para variables ordinales como cuantitativas.

EJEMPLO 3.3

Si es un conjunto ordenado:

```
3, 4, 5, 6, 8
```

$n = 5$ (impar), la mediana es el tercer valor:

$$\bar{x} = 5$$

Si el conjunto fuera:

3, 4, 5, 6

$n = 4$ (par), la mediana es el promedio de los dos valores centrales:

$$\bar{x} = \frac{4 + 5}{2} = 4.5$$

Moda

La moda es el valor que más se repite en el conjunto de datos. Es la medida de tendencia central que identifica el dato más frecuente.

- Puede existir más de una moda (distribución multimodal):
  - Si hay dos valores que se repiten con igual frecuencia, es bimodal.
  - Si más de dos, multimodal.
- Si todos los valores aparecen con la misma frecuencia, se dice que no hay moda.

Es la única medida de tendencia central que puede aplicarse directamente a variables cualitativas nominales, además de ordinales o cuantitativas, es muy útil para detectar concentraciones o preferencias dominantes.

EJEMPLO 3.4

Datos:

4, 5, 4, 6, 4, 7, 5

La moda es:

$$Mo = 4$$

porque se repite 3 veces.

Qué instrucciones vamos a utilizar

pd.Series(...)

Crea un objeto Serie de pandas, que es como una columna de datos, esto facilita el uso de métodos estadísticos incorporados.

median ()

Calcula la mediana, es decir, el valor central del conjunto de datos ordenado, pandas ordena internamente los datos y, según si hay un número par o impar de observaciones, devuelve el valor del medio o el promedio de los dos centrales.

mode()

Este método devuelve un objeto Serie que puede contener uno o más valores, dependiendo si hay una moda única, bimodal o multimodal.

CÓDIGO 3.2

Tenemos el número de libros leídos por 12 estudiantes:

3, 4, 5, 6, 4, 5, 4, 6, 7, 5, 4, 5

Código

```
import pandas as pd
datos = [3, 4, 5, 6, 4, 5, 4, 6, 7, 5, 4, 5]
serie = pd.Series(datos)
# Calcular mediana
mediana = serie.median()
# Calcular moda
moda = serie.mode()
# Mostrar resultados
print(f"Datos: {datos}")
print(f"Mediana: {mediana}")
# La moda puede devolver más de un valor (por si hay varias modas)
if len(moda) == 1:
  print(f"Moda: {moda.iloc[0]}")
else:
  print(f"Modas: {list(moda)}")
```

3.2 Medidas de Dispersión

Las medidas de dispersión o de variabilidad complementan a las medidas de tendencia central al proporcionar información sobre cómo se distribuyen o dispersan los datos alrededor de un valor central. Mientras la media, la mediana o la moda indican dónde tiende a concentrarse el conjunto, las medidas de dispersión describen qué tan alejados o próximos están los datos entre sí.

Evaluar la dispersión es esencial para entender la homogeneidad o heterogeneidad del fenómeno estudiado, identificar riesgos o variabilidades significativas y dar contexto a los promedios. Por ejemplo, dos grupos pueden tener la misma media, pero distribuciones muy diferentes, lo que puede alterar completamente la interpretación si no se considera la variabilidad.

Rango

El rango es la medida de dispersión más sencilla, calculada como la diferencia entre el valor máximo y el valor mínimo del conjunto de datos:

$$R = x_{max} - x_{min}$$

Mide la amplitud total del conjunto, indicando el intervalo dentro del cual se encuentran todos los datos. El rango sólo toma en cuenta los extremos, por lo que es muy sensible a valores atípicos y no refleja cómo están distribuidos los demás datos.

Varianza y desviación estándar

La varianza mide el promedio de las desviaciones cuadráticas respecto a la media, para un conjunto de datos $x1, x2, \ldots, xn$ con media $\bar{x}$:

Para población:

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

Para muestra

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

La varianza se expresa en unidades cuadradas, lo que dificulta su interpretación directa.

Desviación estándar

Por eso se utiliza la desviación estándar, que es la raíz cuadrada positiva de la varianza:

$$\sigma = \sqrt{\sigma^2},\ s = \sqrt{s^2}$$

Así se expresa en las mismas unidades que los datos originales, facilitando su interpretación como una medida del "promedio de desviación" respecto a la media.

Una desviación estándar pequeña indica que los datos están cercanos a la media, mientras que una grande revela mayor dispersión, es una de las medidas más importantes y ampliamente utilizadas en estadística descriptiva.

Coeficiente de variación (CV)

El coeficiente de variación (CV) relaciona la desviación estándar con la media, expresado generalmente en porcentaje. Se calcula así:

$$CV = \frac{\sigma}{\bar{x}} x100\%$$

O para muestras

$$CV = \frac{s}{\bar{x}} x100\%$$

El CV muestra el grado de dispersión relativo al tamaño del promedio, permitiendo comparar la variabilidad de conjuntos de datos que tienen diferentes unidades o escalas.

- Un CV bajo indica baja variabilidad relativa respecto a la media (más homogéneo).
- Un CV alto indica alta variabilidad relativa (más heterogéneo).

EJEMPLO 3.5

Consideremos el número de proyectos completados en un mes por cinco empleados:

3, 5, 4, 6, 2

El rango se calcula como:

$R = x_{max} - x_{min}$

$R = 6 - 2$

$R = 4$

Hallamos la media, necesaria para calcular varianza y desviación estándar:

$$\bar{x} = \frac{3 + 5 + 4 + 6 + 2}{5} = \frac{20}{5} = 4$$

Varianza

Paso 1: Tabla de desviaciones al cuadrado

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
3	-1	1
5	+1	1
4	0	0
6	+2	4
2	-2	4
		10

Paso 2: Cálculo de varianza

Como esto es una muestra $(n = 5)$:

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

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - x)^2}{n-1} = \frac{10}{5-1} = \frac{10}{4} = 25$$

Desviación estándar

$$s = \sqrt{s^2} = \sqrt{2.5} \approx 1.58$$

Coeficiente de variación

$$CV = \frac{s}{\bar{x}} x100\% = 39.5\%$$

Interpretación

- El rango indica que los empleados varían entre completar 2 y 6 proyectos.
- La desviación estándar muestra que, en promedio, se desvían aproximadamente 1.58 proyectos respecto a la media.
- El coeficiente de variación permite dimensionar esa variabilidad en términos relativos, mostrando que es casi un 40% del promedio, lo cual es importante para comparar con otros grupos o períodos.
- Esto significa que la variabilidad relativa respecto al promedio es del 39.5%, indicando una dispersión moderada alrededor de la media.

CÓDIGO 3.3

Qué instrucciones vamos a utilizar

pd.Series

Crea un objeto Serie de pandas, que es como una columna de datos. Esto facilita el uso directo de métodos estadísticos.

serie.max() y serie.min()

Calculan el máximo y mínimo del conjunto de datos, necesarios para hallar el rango.

- serie.max() devuelve el valor máximo.
- serie.min() devuelve el valor mínimo.

serie.var()

Calcula la varianza muestral (divide entre $n - 1$)

serie.std()

Calcula la desviación estándar muestral, es decir, la raíz cuadrada de la varianza.

serie.mean()

Calcula la media aritmética del conjunto.

Código

```
import pandas as pd
import numpy as np
# Datos: número de proyectos completados
datos = [3, 5, 4, 6, 2]
serie = pd.Series(datos)
# Calcular medidas de dispersión
# Rango
rango = serie.max() - serie.min()
# Varianza muestral (n-1)
varianza = serie.var()
# Desviación estándar muestral
desviacion = serie.std()
# Coeficiente de variación
media = serie.mean()
cv = (desviacion / media) * 100
# Mostrar resultados
print(f"Datos: {datos}")
print(f"Rango: {rango}")
print(f"Varianza: {varianza:.2f}")
print(f"Desviación estándar: {desviacion:.2f}")
print(f"Coeficiente de variación: {cv:.2f}%")
```

Salida

```
Datos: [3, 5, 4, 6, 2]
Rango: 4
Varianza: 2.50
Desviación estándar: 1.58
Coeficiente de variación: 39.53%
```

Las medidas de dispersión, en conjunto con las de tendencia central, proporcionan una descripción más completa de un conjunto de datos, permiten entender no sólo “dónde se concentran los datos”, sino cuán dispersos o compactos están alrededor de esa posición central, algo fundamental para el análisis y la toma de decisiones.

3.3 Medidas de Forma

Las medidas de forma son parámetros estadísticos que permiten describir cómo se distribuyen los datos respecto a la simetría y al "aplanamiento" o concentración alrededor de la media, complementan a las medidas de tendencia central y de dispersión, proporcionando una visión más completa del comportamiento del conjunto de datos.

Las dos medidas principales son:

- Asimetría (skewness), que indica el grado y dirección de la desviación de la distribución respecto a la simetría.
- Curtosis (kurtosis), que describe cuán concentrados o dispersos están los datos alrededor del promedio.

Asimetría

La asimetría mide el grado de desbalance o inclinación de la distribución respecto a su media.

- Si la distribución es simétrica, la asimetría es aproximadamente 0.
- Si está sesgada a la derecha (positiva), tiene una cola larga hacia valores mayores.
- Si está sesgada a la izquierda (negativa), tiene una cola hacia valores menores.

Ejemplos interpretativos

- Un ejemplo típico de asimetría positiva es el ingreso económico en una población: la mayoría gana alrededor del promedio, pero hay pocos con ingresos muy altos que alargan la cola hacia la derecha.

- Un caso de asimetría negativa sería la edad de jubilación si la mayoría se jubila alrededor de 65 años, pero con algunas personas que lo hacen antes, extendiendo la cola hacia la izquierda.

Curtosis

La curtosis mide el grado de apuntamiento o achatamiento de la distribución en torno a su media, comparado con una distribución normal.

- Una distribución con curtosis normal o mesocúrtica tiene un valor cercano a 0 (o 3 según la definición empleada), similar a la campana normal.
- Una distribución con alta curtosis (leptocúrtica) tiene valores concentrados cerca de la media y colas más pesadas (picos más altos y colas largas).
- Una distribución con baja curtosis (platicúrtica) es más plana, con menos datos concentrados alrededor de la media.

Tipos comunes

Tipo	Forma general	Significado
Mesocúrtica	Similar a normal	Curtosis ≈ 0
Leptocúrtica	Más alta y delgada	Datos muy concentrados en la media y colas más pesadas
Platicúrtica	Más achatada y ancha	Datos más dispersos, menor concentración en la media

El análisis conjunto de la asimetría y la curtosis ofrece un retrato más detallado de la forma de la distribución, permitiendo anticipar fenómenos atípicos, riesgos en variabilidad o posibles sesgos, son herramientas fundamentales en estudios exploratorios y en el diagnóstico de modelos estadísticos.

EJERCICIOS PROPUESTOS

Tema: Medidas de tendencia central, dispersión y forma.

Ejercicio 1

Calcular la media, mediana, moda, rango y la desviación estándar:

4, 6, 5, 7, 8

Ejercicio 2

Calcular la media y varianza muestral. ¿Cuál es el coeficiente de variación?

12, 15, 14, 11, 13, 12, 12

Ejercicio 3

Calcular la Media aritmética simple. Interpretar si la dispersión (calculando rango y s) es alta o baja respecto a la media.

20, 25, 30, 22, 28

Ejercicio 4

Identificar la moda, calcular la desviación estándar.

18, 21, 18, 19, 20, 18, 22

Ejercicio 5

Calcular la Media, mediana, moda y rango, ¿Cómo describirías la variabilidad de estos datos?

9, 11, 10, 12, 9, 8, 10

Ejercicio 6

¿Qué tipo de asimetría tiene?, estimar si presenta curtosis baja o alta comparada con una distribución normal.

3, 3, 4, 5, 6, 7, 8

Ejercicio 7

Calcular media, moda y varianza. Comentar sobre el coeficiente de variación.

10, 12, 10, 10, 14, 10, 10

Ejercicio 8

Calcular, media y mediana, ¿Hay indicios de asimetría?

7, 8, 10, 15, 18

Ejercicio 9

Calcular media y desviación estándar. ¿Qué efecto tiene el valor 80 en la media y la dispersión?

40, 42, 41, 43, 80

Ejercicio 10

Calcular media, mediana, desviación estándar y coeficiente de variación. Describir si el conjunto parece simétrico.

25, 30, 35, 40, 45, 50, 55

CÓDIGOS PROPUESTOS

Tema: Medidas de tendencia central, dispersión y forma, con Google Colab.

Ejercicio 1

Calcular media, mediana, moda, varianza y desviación estándar.

[5, 7, 8, 6, 9, 5, 7, 8]

Ejercicio 2

Hallar el coeficiente de variación y describir la homogeneidad.

[12, 14, 13, 11, 15, 12, 16]

Ejercicio 3

Calcular media, mediana y verificar asimetría con .skew().

[20, 22, 23, 24, 28, 29]

Ejercicio 4

Media, desviación estándar y curtosis (.kurt()).

[35, 36, 34, 33, 37, 32, 38]

Ejercicio 5

Tabla resumen con media, varianza, s y CV.

[8, 10, 9, 7, 6, 10, 11, 9]

Ejercicio 6

¿La distribución es simétrica? Calcular skewness.

[50, 55, 60, 65, 70, 75]

Ejercicio 7

Analizar cómo un valor extremo afecta media y desviación estándar.

[3, 3, 3, 3, 20]

Ejercicio 8

Calcular media, moda y coeficiente de variación.

[15, 16, 15, 17, 18, 16, 15]

Ejercicio 9

Media, desviación estándar y curtosis.

[100, 102, 98, 105, 95, 90]

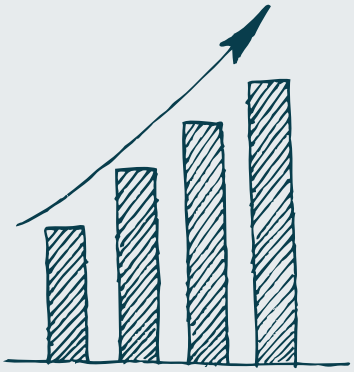
Ejercicio 10

Describir si hay asimetría o curtosis anormal comparado con normal.

[40, 42, 41, 43, 40, 42, 41]

CASOS PRÁCTICOS

4



4.1 Caso 1: Mejoramiento del tiempo promedio de atención en dos sucursales de una cadena de farmacias

La cadena de farmacias Ecuador tiene interés en comparar la eficiencia operativa de dos de sus sucursales (A y B), ubicadas en distintos barrios de la ciudad, para ello, durante una semana se registraron los tiempos de atención al cliente (en minutos) de 25 clientes en cada sucursal.

Los datos obtenidos fueron:

Sucursal A

7, 8, 9, 7, 6, 8, 9, 10, 8, 7, 7, 9, 8, 7, 6, 8, 9, 7, 8, 8, 9, 7, 6, 8, 7

Sucursal B

5, 6, 5, 7, 6, 6, 5, 7, 6, 6, 5, 6, 7, 6, 5, 6, 6, 7, 6, 5, 7, 6, 5, 6, 6

¿Qué se solicita?

Como responsable del análisis estadístico, debes calcular e interpretar, para cada sucursal:

- Media, mediana y moda.
- Varianza, desviación estándar y coeficiente de variación.
- Asimetría y curtosis.

Adicional para mejor presentación del informe se debe representar gráficamente la distribución de los tiempos, mediante:

- Histogramas comparativos.
- Diagramas de caja (boxplots) para visualizar diferencias en medianas, dispersión y posibles valores atípicos.

También deberemos comparar los resultados entre ambas sucursales, destacando:

- ¿Cuál atiende más rápido en promedio?
- ¿Cuál tiene mayor consistencia (CV)?
- ¿Se observa algún sesgo (asimetría) o mayor concentración/achatamiento (curtosis) en alguna de las sucursales?

Conclusión práctica: Formular recomendaciones basadas en los hallazgos. ¿Qué podría sugerirse a la sucursal menos eficiente para mejorar?

CÓDIGO 4.1

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Datos
sucursal_A = [7, 8, 9, 7, 6, 8, 9, 10, 8, 7, 7, 9, 8, 7, 6, 8, 9, 7, 8, 8, 9, 7, 6, 8, 7]
sucursal_B = [5, 6, 5, 7, 6, 6, 5, 7, 6, 6, 5, 6, 7, 6, 5, 6, 6, 7, 6, 5, 7, 6, 5, 6, 6]
#Crear DataFrame en formato largo
df = pd.DataFrame({'Tiempo': sucursal_A + sucursal_B,
  'Sucursal': ['A']*25 + ['B']*25})
# Función para calcular estadísticas
def resumen_estadistico(nombre, serie):
  media = serie.mean()
  mediana = serie.median()
  moda = list(serie.mode())
  varianza = serie.var()
  desviacion = serie.std()
  cv = (desviacion / media) * 100
  asimetria = serie.skew()
```

```
    curtosis = serie.kurt()
    print(f"=== {nombre} ===")
    print(f"Media: {media:.2f}")
    print(f"Mediana: {mediana}")
    print(f"Moda: {moda}")
    print(f"Varianza: {varianza:.2f}")
    print(f"Desviación estándar: {desviacion:.2f}")
    print(f"Coef. variación: {cv:.2f}%")
    print(f"Asimetría (skewness): {asimetria:.2f}")
    print(f"Curtosis: {curtosis:.2f}\n")

# Calcular y mostrar para cada sucursal
resumen_estadistico("Sucursal A", df[df['Sucursal']=='A']['Tiempo'])
resumen_estadistico("Sucursal B", df[df['Sucursal']=='B']['Tiempo'])

# Histogramas comparativos
plt.figure(figsize=(8,5))
sns.histplot(data=df, x='Tiempo', hue='Sucursal', bins=5, multiple='dodge', edgecolor='black')
plt.title("Histogramas comparativos de tiempo de atención")
plt.xlabel("Tiempo (min)")
plt.ylabel("Frecuencia")
plt.show()
# Boxplots comparativos sin advertencias
plt.figure(figsize=(8,3))
sns.boxplot(x='Sucursal', y='Tiempo', hue='Sucursal', data=df, palette=["skyblue", "lightgreen"], dodge=False)
plt.legend([],[], frameon=False)  # Oculta leyenda innecesaria
plt.title("Boxplots comparativos del tiempo de atención")
plt.ylabel("Tiempo (min)")
plt.show()
```

Salida

```
=== Sucursal A ===
Media: 7.72
Mediana: 8.0
Moda: [7, 8]
Varianza: 1.13
```

Desviación estándar: 1.06

Coef. variación: 13.75%

Asimetría (skewness): 0.16

Curtosis: -0.51

=== Sucursal B ===

Media: 5.92

Mediana: 6.0

Moda: [6]

Varianza: 0.49

Desviación estándar: 0.70

Coef. variación: 11.86%

Asimetría (skewness): 0.11

Curtosis: -0.82

Interpretación

- Sucursal A tiene un tiempo promedio de atención de 7.72 minutos, con una mediana muy cercana (8 min) y modas de 7 y 8 min, lo que indica que la mayoría de los tiempos se concentran en torno a esos valores.
- Sucursal B muestra un tiempo promedio menor, de 5.92 minutos, con mediana 6 min y moda claramente en 6 min, reflejando un proceso más rápido en promedio.

En términos de eficiencia del servicio, la Sucursal B atiende más rápido a sus clientes que la Sucursal A, con una diferencia promedio de casi 2 minutos menos por cliente, lo cual es relevante en contextos de alta demanda por otro lado:

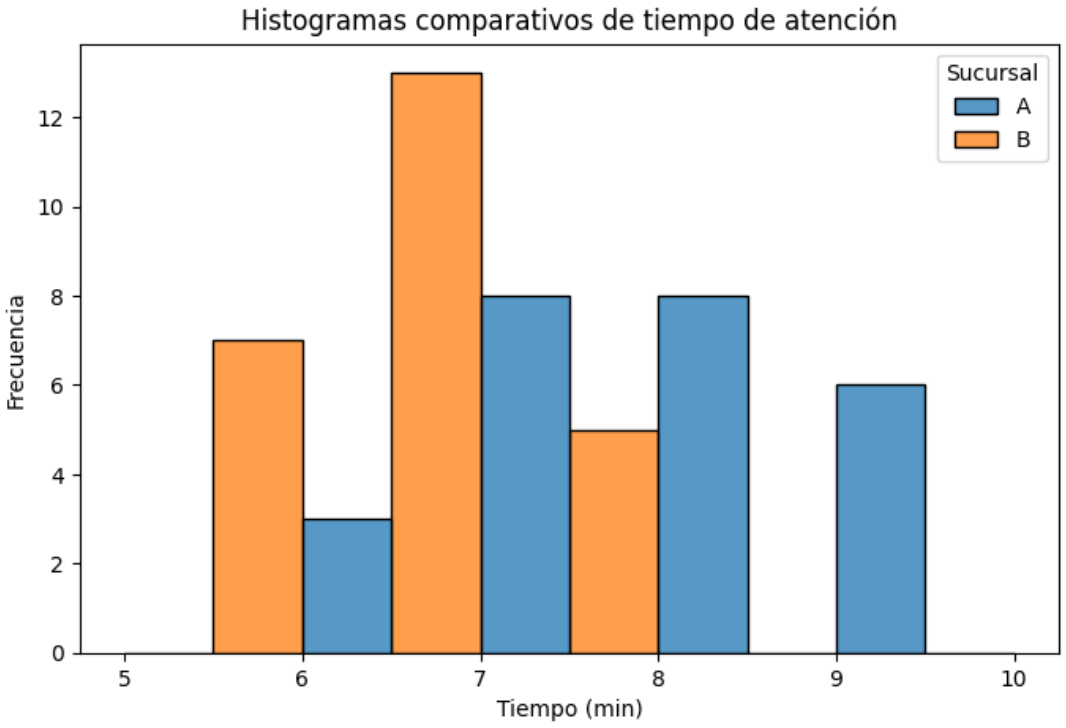
- La varianza y desviación estándar confirman que Sucursal A tiene una dispersión algo mayor (s=1.06s=1.06) frente a Sucursal B (s=0.70s=0.70).
- El coeficiente de variación (CV) muestra la variabilidad relativa al promedio, con: CV en A ≈ 13.75% CV en B ≈ 11.86%

Esto significa que, aunque ambos procesos son relativamente consistentes, Sucursal B no solo es más rápida sino también ligeramente más homogénea en sus tiempos, con menor variabilidad relativa. Ambas distribuciones presentan asimetrías ligeramente positivas (0.16 en A y 0.11 en B), lo que indica que, en ambos casos, existen colas un poco más extendidas hacia tiempos de atención mayores, estos valores son bajos, mostrando distribuciones casi simétricas.

La curtosis negativa en ambas sucursales (−0.51−0.51 en A y −0.82−0.82 en B) revela distribuciones ligeramente más planas (platicúrticas) que una normal. Esto indica menos concentración extrema en la media y colas algo menos pesadas.

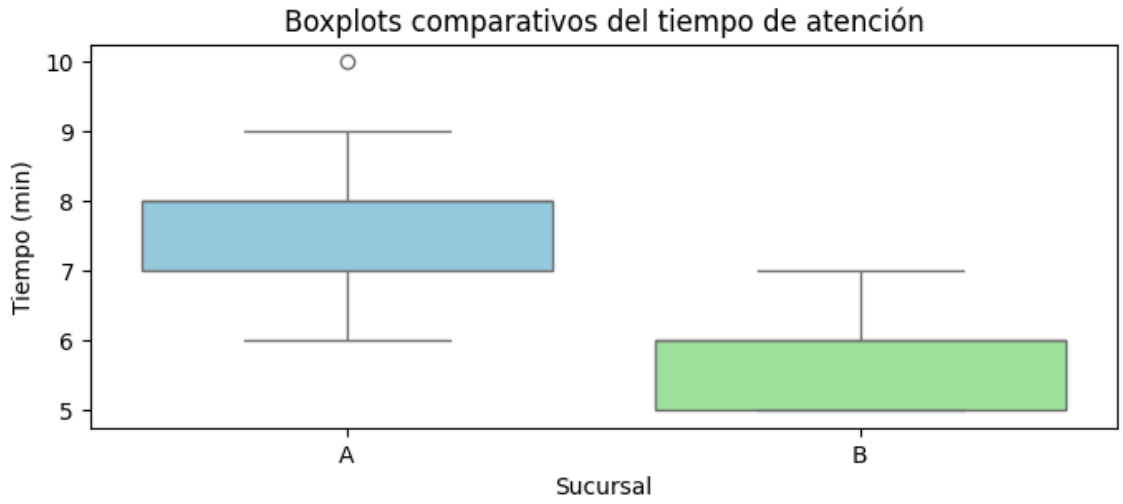
No hay evidencia de colas muy largas ni concentraciones atípicas, ambos procesos son razonablemente simétricos y con moderada dispersión, aunque B tiene distribución un poco más "extendida" en torno a su media.

Histogramas



Los histogramas comparativos muestran que Sucursal B concentra la mayoría de los tiempos en torno a 6 minutos, mientras que Sucursal A distribuye los tiempos entre 7 y 9 minutos, confirmando que el servicio en A es más lento y algo más disperso.

Diagramas de caja (boxplot)



Destacan que la mediana en A está más alta, y el rango intercuartílico (la "caja") ligeramente más amplio, confirmando la mayor variabilidad.

Conclusión

Sucursal B ofrece un servicio más eficiente, atendiendo en promedio casi 2 minutos menos por cliente y con un proceso ligeramente más homogéneo, lo que puede traducirse en menores tiempos de espera en fila y mayor satisfacción del cliente.

Recomendación

Se recomienda a la Sucursal A analizar sus procedimientos, identificar cuellos de botella (por ejemplo, tiempos de facturación o despacho de medicamentos) y adoptar buenas prácticas observadas en la Sucursal B, para lograr estandarizar y optimizar sus tiempos.

CÓDIGO 4.2

Caso2: Análisis del tiempo de permanencia de clientes en un restaurante.

Un restaurante quiere entender mejor cuánto tiempo permanecen sus clientes dentro del local, para optimizar la rotación de mesas y planificar el personal, para ello, durante dos días consecutivos registró el tiempo total de permanencia (en minutos) de 50 clientes seleccionados al azar.

Los datos fueron los siguientes:

42, 55, 48, 38, 61, 47, 53, 52, 45, 50, 58, 44, 49, 60, 39, 57, 46, 43, 51, 54, 48, 47, 59, 56, 62, 49, 46, 41, 55, 58, 50, 47, 52, 40, 60, 45, 51, 53, 48, 59, 57, 42, 56, 54, 43, 61, 44, 50, 47, 58

¿Qué se solicita?

Debemos realizar el siguiente análisis:

Tablas de frecuencia:

- Elaborar una tabla de frecuencia simple (no agrupada) mostrando cada valor único y su frecuencia.
- Construir una tabla de frecuencia por intervalos (usando regla de Sturges para determinar número de clases).

Medidas de tendencia central, dispersión y forma:

- Calcular media, mediana, moda.
- Hallar rango, varianza, desviación estándar y coeficiente de variación.
- Determinar asimetría y curtosis.

Gráficos:

- Histograma y polígono de frecuencia para la tabla agrupada.
- Ojiva (frecuencia acumulada).
- Boxplot para identificar concentración y posibles valores atípicos.

Responda:

- ¿Cuál es el tiempo típico de permanencia de los clientes?
- ¿El tiempo es homogéneo o hay alta variabilidad?
- ¿La distribución muestra colas largas (asimetría) o mayor concentración/aplanamiento (curtosis)?
- ¿Qué recomendaciones operativas se pueden hacer?

Código

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Datos
tiempos = [42, 55, 48, 38, 61, 47, 53, 52, 45, 50, 58, 44, 49, 60, 39, 57, 46, 43, 51,
54,48, 47, 59, 56, 62, 49, 46, 41, 55, 58, 50, 47, 52, 40, 60, 45, 51, 53, 48, 59, 57, 42,
56, 54, 43, 61, 44, 50, 47, 58]
serie = pd.Series(tiempos)
# Tabla de frecuencia simple (no agrupada)
tabla_simple = serie.value_counts().sort_index()
print("=== Tabla de frecuencia simple ===")
print(tabla_simple)
# Tabla de frecuencia agrupada por intervalos usando Sturges
n = len(serie)
k = int(np.ceil(1 + 3.322 * np.log10(n)))  # Regla de Sturges
minimo, maximo = serie.min(), serie.max()
rango = maximo - minimo
amplitud = np.ceil(rango / k)
# Construir intervalos
bins = np.arange(minimo, maximo + amplitud, amplitud)
serie_categorica = pd.cut(serie, bins=bins, right=False)
tabla_agrupada = serie_categorica.value_counts().sort_index()
print("\n=== Tabla de frecuencia agrupada ===")
print(tabla_agrupada)
# Medidas de tendencia central
media = serie.mean()
mediana = serie.median()
moda = list(serie.mode())
# Medidas de dispersión
varianza = serie.var()
desviacion = serie.std()
cv = (desviacion / media) * 100
# Medidas de forma
asimetria = serie.skew()
curtosis = serie.kurt()
# Mostrar resultados numéricos
print("\n=== Medidas de tendencia central, dispersión y forma ===")
print(f"Media: {media:.2f}")
```

```
print(f"Mediana: {mediana}")
print(f"Moda: {moda}")
print(f"Rango: {rango}")
print(f"Varianza: {varianza:.2f}")
print(f"Desviación estándar: {desviacion:.2f}")
print(f"Coef. variación: {cv:.2f}%")
print(f"Asimetría (skewness): {asimetria:.2f}")
print(f"Curtosis: {curtosis:.2f}")
# Histograma y polígono de frecuencia
plt.figure(figsize=(8,5))
counts, bin_edges, _ = plt.hist(serie, bins=bins, edgecolor='black', alpha=0.6, label='Histograma')
bin_centers = 0.5 * (bin_edges[1:] + bin_edges[:-1])
plt.plot(bin_centers, counts, marker='o', color='red', linestyle='-', label='Polígono de frecuencia')
plt.title("Histograma y polígono de frecuencia del tiempo de permanencia")
plt.xlabel("Tiempo (min)")
plt.ylabel("Frecuencia")
plt.legend()
plt.show()
# Ojiva (frecuencia acumulada)
frecuencia_acum = np.cumsum(counts)
plt.figure(figsize=(8,4))
plt.plot(bin_edges[1:], frecuencia_acum, marker='o', linestyle='--', color='blue')
plt.title("Ojiva (frecuencia acumulada)")
plt.xlabel("Tiempo (min)")
plt.ylabel("Frecuencia acumulada")
plt.grid()
plt.show()
# Boxplot
plt.figure(figsize=(8,2))
sns.boxplot(x=serie, color='lightgreen')
plt.title("Boxplot del tiempo de permanencia")
plt.xlabel("Tiempo (min)")
plt.show()
```

Salida

```
=== Tabla de frecuencia simple ===
38  1
39  1
40  1
41  1
42  2
43  2
44  2
45  2
46  2
47  4
48  3
49  2
50  3
51  2
52  2
53  2
54  2
55  2
56  2
57  2
58  3
59  2
60  2
61  2
62  1
Name: count, dtype: int64
=== Tabla de frecuencia agrupada ===
[38.0, 42.0)    4
[42.0, 46.0)    8
[46.0, 50.0)   11
[50.0, 54.0)    9
[54.0, 58.0)    8
[58.0, 62.0)    9
Name: count, dtype: int64
=== Medidas de tendencia central, dispersión y forma ===
Media: 50.60
Mediana: 50.0
Moda: [47]
Rango: 24
Varianza: 42.41
Desviación estándar: 6.51
Coef. variación: 12.87%
Asimetría (skewness): -0.02
Curtosis: -1.01
```

En la tabla de frecuencia simple, observamos que los tiempos más frecuentes son 47 min (4 ocurrencias), seguidos por 48, 50 y 58 minutos, cada uno con 3 ocurrencias, mostrando que el tiempo típico ronda entre 47 y 50 minutos. La tabla de frecuencia agrupada (según la regla de Sturges) divide los datos en 6 intervalos iguales, destacando:

- El intervalo [46.0, 50.0) concentra 11 clientes, siendo el más frecuente.
- Le siguen los intervalos [50.0, 54.0) y [58.0, 62.0), cada uno con 9 clientes, mostrando cierta dispersión hacia tiempos más largos.

Esto muestra que, aunque la mayor parte de los clientes permanece alrededor de 47 a 50 min, hay una distribución amplia que se extiende desde 38 hasta 62 minutos.

La media es de 50.6 min, mientras que la mediana es de 50 min, muy cercanas entre sí, la moda, sin embargo, se sitúa en 47 min, que es el tiempo más repetido individualmente.

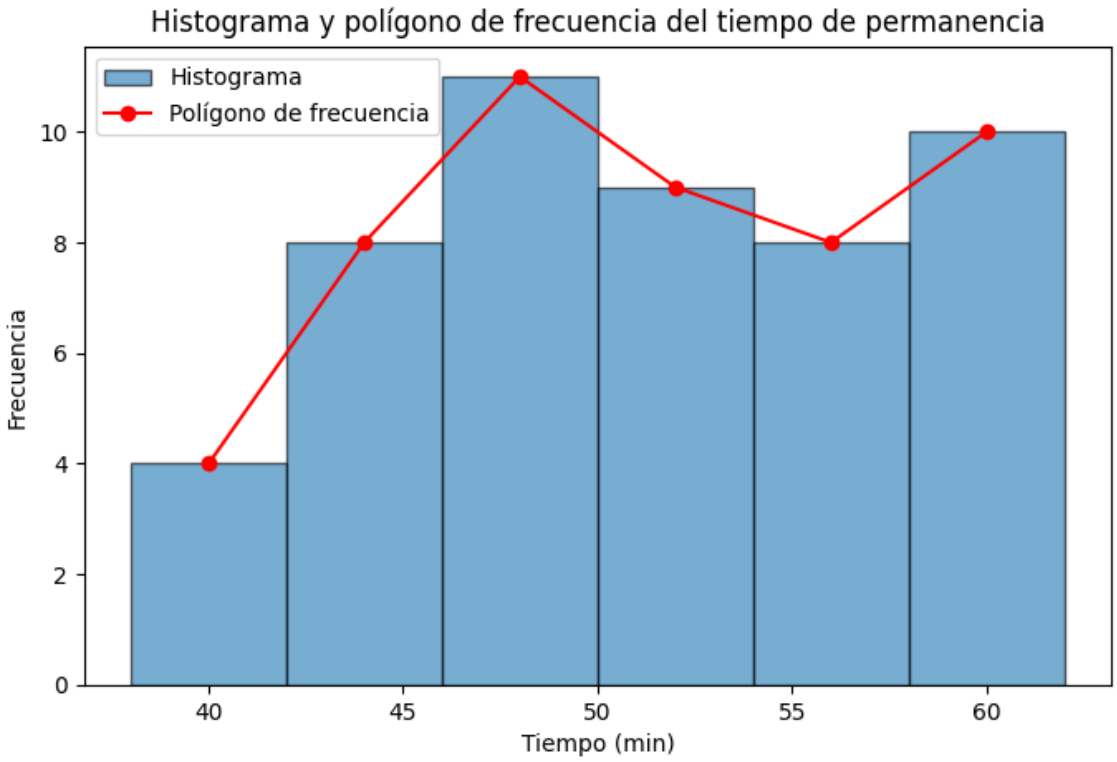
Esto indica una distribución ligeramente concentrada por debajo del promedio, sin grandes sesgos.

El rango es de 24 min (62 - 38), reflejando la amplitud total del fenómeno, la desviación estándar es de 6.51 min, que representa el desvío típico respecto al promedio. El coeficiente de variación (CV) es de 12.87%, indicando que, aunque hay una dispersión notable en minutos absolutos, es moderada respecto al promedio, es decir, la permanencia de los clientes es relativamente consistente en términos proporcionales.

La asimetría (skewness) es -0.02, prácticamente cero, mostrando que la distribución es casi simétrica, sin colas largas hacia ninguno de los extremos, la curtosis es -1.01, negativa, lo que indica que se trata de una distribución platicúrtica, es decir, más achatada que la normal, con menos concentración en el centro y colas más suaves.

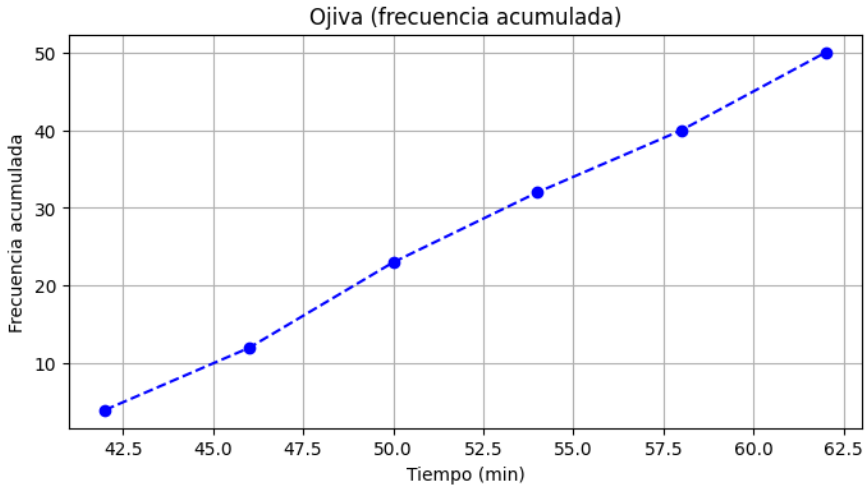
Esto sugiere que hay una dispersión algo mayor de clientes en torno a la media, sin mucha acumulación en un único valor.

Histograma



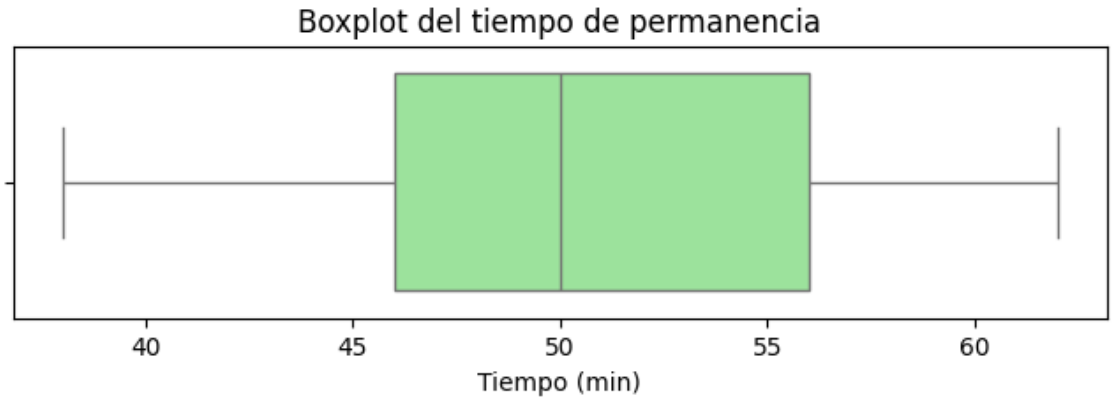
El histograma y el polígono de frecuencia muestran un pico alrededor de 47 a 50 min, pero con una distribución relativamente ancha.

Ojiva



La ojiva indica que aproximadamente el 50% de los clientes permanece en el restaurante hasta 50 min, confirmando la mediana.

Diagrama de cajas (boxplot)



El boxplot muestra una caja moderadamente amplia, sin valores atípicos evidentes, reforzando la idea de una dispersión razonable, con una mediana centrada.

Conclusión

- El tiempo típico de permanencia de un cliente en el restaurante es de aproximadamente 50 min.
- Existe una variabilidad moderada (CV ≈ 13%), por lo que la administración puede planificar la rotación de mesas con un tiempo estimado bastante confiable.
- Dado que la distribución no muestra asimetrías importantes ni colas largas, ni excesiva concentración, pueden mantenerse los turnos del personal y la planificación de cocina sin prever tiempos excepcionalmente largos o cortos.

Recomendación

Como recomendación, se podría diseñar estrategias para estimular rotaciones ligeramente más rápidas en las horas pico, asegurando mayor disponibilidad de mesas y reduciendo tiempos de espera.

ISBN: 978-9942-7128-4-4