Facultad de Ingeniería

Asignatura: Probabilidad y Estadística
Docente: Ing. Alberto Acosta López
Grupo: 020-84

Estudiantes:

Juan Felipe Wilches Gómez Código: 20231020137
Tomás Arévalo Montes Código: 20232020267
Daniel Vargas Arias Código: 20232020103

Bogotá, D.C.

ESTADÍSTICA DESCRIPTIVA

Índice

Resumen

El presente informe aborda los fundamentos de la estadística descriptiva, disciplina esencial para la recopilación, organización, análisis e interpretación de datos con el fin de facilitar la comprensión de fenómenos y apoyar la toma de decisiones informadas. Se estudian los principales tipos de variables y su clasificación, las medidas de tendencia central, de dispersión y de posición, así como las representaciones gráficas más utilizadas para describir conjuntos de datos. A través de ejemplos prácticos, se evidencia cómo la estadística descriptiva permite sintetizar grandes volúmenes de información, identificar patrones y comportamientos en los datos, y establecer bases sólidas para el análisis estadístico inferencial. En conjunto, este trabajo resalta la importancia de la estadística descriptiva como herramienta fundamental en los procesos científicos, tecnológicos y de gestión dentro del ámbito de la ingeniería.

1. Introducción

La estadística descriptiva es la disciplina que recolecta, organiza, resume y presenta datos con el propósito de facilitar su interpretación y la toma de decisiones (Carlos Pareja, 2025). Su origen se remonta a las prácticas censales de civilizaciones antiguas y al desarrollo de la *statistik* en el siglo XVIII (López, 2019). En el siglo XIX y XX, científicos como Quetelet, Pearson y Galton desarrollaron conceptos de tendencia central, correlación, regresión y diseño experimental que formalizaron la disciplina (El nacimiento de la estadística moderna. Francis Galton y Karl Pearson (2020) (Akal, 2020).

En el estado del arte actual, la estadística descriptiva se integra con la ciencia de datos y la inteligencia artificial: se automatiza el EDA (*exploratory data analysis*)(IBM, 2025), se aplican técnicas robustas para *big data* e IoT, y se enfatiza la ética y la interpretabilidad en la automatización de decisiones.

2. Desarrollo

2.1. Conceptos básicos

Población, muestra, variable y dato:

Población: conjunto completo de individuos u objetos de estudio. **Muestra:** subconjunto representativo de la población. **Variable:** característica que puede tomar

distintos valores. **Dato:** valor específico que toma una variable.

Tipos de variables

- **Cualitativas:**
  - Nominales: categorías sin orden (ej. tipo de sangre, género).
  - Ordinales: categorías con jerarquía (ej. nivel educativo).
  - Dicotómicas: dos categorías (sí/no).
- **Cuantitativas:**
  - Discretas: valores enteros (ej. número de hijos).
  - Continuas: valores reales (ej. peso, estatura).

Escalas de medición

- Nominal, ordinal, de intervalo y de razón, según la presencia de orden y cero absoluto.

2.2. Explicación analítica — Técnicas empleadas

Las principales técnicas empleadas incluyen:

- Cálculo de medidas para datos no agrupados: sumar y dividir por n (media), ordenar y hallar valor medio (mediana), contar frecuencias (moda).
- Cálculo de varianza y desviación estándar: promediar los cuadrados de las desviaciones respecto a la media (poblacional o muestral con n-1).
- Datos agrupados: usar marcas de clase (puntos medios) y frecuencias para estimar media y varianzas aproximadas.
- Percentiles y cuartiles: posición $k\cdot(n+1)/100$ (no agrupados) o interpolación dentro del intervalo para datos agrupados.
- Visualización: histogramas para distribuciones, boxplot para resumir mediana, IQR y posibles outliers.

3. Organización y presentación de datos

Objetivo: convertir la información cruda en una forma que facilite el análisis.

Técnicas y formatos

1. **Datos sin agrupar:** Simple lista de observaciones. Se ordenan para obtener percentiles y mediana.

   Usos: cálculos exactos de media, mediana, moda.

2. **Tabla de frecuencias:** Columnas típicas: valor / frecuencia absoluta / frecuencia acumulada / frecuencia relativa / frecuencia relativa acumulada.

   Permite ver concentración y construir gráficos.

3. **Datos agrupados:** Usar cuando el número de observaciones es grande o los valores son continuos.

   Se eligen clases (mismo ancho o variable), se cuenta frecuencia en cada clase y se usan marcas de clase (punto medio) para estimar medidas como la media y varianza.

4. **Visualizaciones:** Histograma (datos continuos o agrupados): forma de la distribución.

   Polígono de frecuencias o barras (discretos).

   Boxplot (diagrama de caja y bigotes): mediana, cuartiles, IQR, outliers.

   Diagramas de tallo y hoja: conserva datos originales y da idea de forma.

4. Medidas de tendencia central y dispersión

Conceptos y fórmulas (analíticamente)

Sea un conjunto de datos no agrupados $x_1, x_2, \ldots, x_n$.

Tendencia central

- **Media aritmética (muestral):**

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

- **Mediana:** valor central ordenado.

Si n impar:

$$\text{Mediana} = x_{(k)} \text{ con } k = \frac{n+1}{2}$$

Si n par:

$$\text{Mediana} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

- **Moda:** valor con mayor frecuencia.

Dispersión

- **Rango:**

$$R = x_{\text{máx}} - x_{\text{mín}}$$

- **Varianza poblacional:**

$$\sigma^2 = \frac{1}{N} \sum (x_i - \mu)^2$$

- **Varianza muestral (estimador insesgado):**

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

- **Desviación estándar:**

$$s = \sqrt{s^2} \quad (\text{o } \sigma \text{ si es poblacional})$$

- **Coeficiente de variación (CV):**

$$CV = \frac{s}{\bar{x}} \times 100\,\%$$

- **Desviación absoluta media (MAD):** Promedio de las desviaciones absolutas respecto a la media.

$$MAD = \frac{1}{n} \sum |x_i - \bar{x}|$$

5. Medidas de posición y forma

Las medidas de posición son indicadores estadísticos que permiten resumir datos en uno solo o dividir su distribución en intervalos del mismo tamaño, ayudan a

interpretar un conjunto de datos de forma rápida y sencilla, dando significado a un conjunto amplio de datos mediante una representación y repartición.

Se dividen en dos tipos, medidas de posición de tendencia central y tendencia no central, las medidas de posición central son los cuantiles, hacen divisiones iguales en la distribución de los datos, así, dan los valores superiores, medios e inferiores.

Medidas no centrales

- Cuartiles: $Q_1(25\,\%)$, $Q_2(50\,\%)$, $Q_3(75\,\%)$
- Percentiles: dividen en 100 partes.
- Rango intercuartílico: $IQR = Q_3 - Q_1$
- Sesgo: $Skew > 0$ (cola derecha), $Skew < 0$ (cola izquierda)
- Curtosis: mide el grado de concentración de la distribución (leptocúrtica si > 0)

Medidas centrales

- Media aritmética: promedio ponderado.
- Mediana: valor central de la distribución.
- Moda: valor más frecuente, útil en estudios de mercado.

6. Importancia y aplicaciones en ingeniería de sistemas

La estadística descriptiva es una herramienta fundamental en todas las ramas de la ingeniería, ya que proporciona métodos para organizar, resumir y presentar datos de procesos experimentos o sistemas.

En la ingeniería de sistemas es especialmente útil a la hora de realizar el front end de una aplicación de software ya que los datos numéricos y algebraicos no se pueden mostrar crudos a los usuarios o clientes, hablando en el caso de un software empresarial. Estos necesitan ser interpretados primero mediante la estadística descriptiva para poder darlos a entender al público.

Un ingeniero de sistemas debe garantizar que las aplicaciones y la infraestructura funcionen de manera eficiente. La estadística descriptiva es la base para lograrlo.

- **Caracterización de la carga:** Al utilizar métricas descriptivas (media, mediana, percentiles, desviación estándar), los ingenieros pueden analizar el tráfico de red, el tiempo de respuesta de los servidores o el uso de recursos (CPU, memoria). Por ejemplo, calcular el percentil 95 del tiempo de respuesta no solo ofrece un promedio, sino que describe el peor caso que experimenta la mayoría de los usuarios, lo cual es crucial para establecer acuerdos de nivel de servicio (SLA).

- **Identificación de cuellos de botella:** La visualización de datos mediante histogramas o diagramas de caja permite identificar rápidamente dónde se concentra la mayor latencia o variabilidad en los tiempos de procesamiento. Esta información es vital para enfocar los esfuerzos de optimización donde tendrán el mayor impacto en el rendimiento global del sistema.

7. Avances recientes y estado del arte

Los avances en la estadística descriptiva destacan por ser aplicaciones de esta en la tecnología contemporánea, siendo estos avances en áreas como la automatización, el Big Data y la integración con la inteligencia artificial.

La estadística descriptiva automatizada impulsada por los DS-Agents (Agentes de ciencias de datos) o la inteligencia artificial, véase el caso de DS-STAR, un agente de ciencia de datos desarrollado por Google Research que puede analizar automáticamente conjuntos de datos complejos, identificar variables, detectar distribuciones y generar un resumen de texto y gráfico, reduciendo el tiempo invertido en fases del *pipeline* de un proyecto como la fase de exploración de datos. Según Google Research (2025), "DS-STAR es un agente de ciencia de datos de última generación cuya versatilidad se demuestra por su capacidad para automatizar una variedad de tareas — desde análisis estadísticos hasta visualización y manipulación de datos — a través de diversos tipos de datos, culminando en un rendimiento destacado en el famoso benchmark DABStep." (Yoon & Nam, 2025).

Algoritmos avanzados son capaces de calcular límites estadísticos de un proceso y señalar automáticamente los valores atípicos o cambios bruscos en la distribución; también detectan anormalidades y ahorran tiempo y recursos en procesos tediosos.

Otros avances en la estadística descriptiva se presentan como herramientas para la narrativa de datos (*data storytelling*, del inglés), ayudando a crear coherentemente guías de usuario para comunicar los datos. Flourish, una plataforma desarrollada por Canva, se destaca por permitir a los usuarios "crear narrativas visuales interactivas y atractivas sin necesidad de conocimientos de programación" (Flourish, s. f.). Este software está especializado en generar gráficos, mapas de calor y visualizaciones dinámicas que facilitan la comprensión de conjuntos de datos complejos.

Asimismo, herramientas como Tableau, que posibilita la creación de *dashboards* interactivos enlazados con múltiples gráficas, representan ejemplos destacados de esta tendencia hacia la automatización y la accesibilidad en la comunicación de datos. Según Tom Perry, director sénior de datos, información e integración de Elsevier, "Tableau es un programa para cambiar la empresa, la cultura y la forma de analizar los datos y tomar decisiones. A medida que avanzamos en nuestro camino y le mostramos al personal lo fácil que era obtener información por su cuenta, tuvimos epifanías muy importantes. Ahora podemos dar un paso atrás y brindarle apoyo en su aprendizaje en lugar de tener que centralizar todo" lo que refleja la evolución de las plataformas analíticas hacia sistemas más inteligentes y autónomos, capaces de asistir al usuario en todo el proceso de análisis de datos, desde la exploración hasta la toma de decisiones (Tableu, 2021).

8. Análisis y resolución de ejemplos: comparación entre métodos manuales y digitales

Ejemplo 1 — Datos no agrupados (Tiempos de carga)

Para evaluar el rendimiento de una aplicación web durante el proceso de carga inicial, se registraron cinco mediciones correspondientes al tiempo (en segundos) que tarda la interfaz en mostrar completamente el contenido al usuario. Estos tiempos fueron obtenidos en condiciones controladas y utilizando el mismo dispositivo y conexión de red, con el fin de garantizar la comparabilidad de los resultados.

Los datos obtenidos fueron los siguientes: 2.1,2.3,1.9,2.5,2.2 segundos

A partir de estas observaciones, se solicita calcular manualmente y mediante herramientas digitales (Python) las principales medidas de estadística descriptiva: media, mediana, moda, varianza muestral y desviación estándar muestral. Posteriormente, se debe comparar ambos procedimientos para verificar la consistencia de

los resultados y analizar la variabilidad del tiempo de carga de la aplicación.

Datos ($n = 5$): 2.1, 2.3, 1.9, 2.5, 2.2 (segundos)

A) Procedimiento manual — paso a paso

1. Suma de observaciones: $2{,}1 + 2{,}3 + 1{,}9 + 2{,}5 + 2{,}2 = 11{,}0$
2. Media: $\bar{x} = \dfrac{11{,}0}{5} = 2{,}2$
3. Ordenar: 1.9, 2.1, 2.2, 2.3, 2.5 $\Rightarrow$ mediana = 2.2
4. Moda: no definida (ningún valor se repite)
5. Desviaciones y cuadrados:

$$(-0{,}1)^2 + (0{,}1)^2 + (-0{,}3)^2 + (0{,}3)^2 + (0{,}0)^2 = 0{,}20$$

6. Varianza: $s^2 = \dfrac{0{,}20}{5-1} = 0{,}05$
7. Desviación estándar: $s = \sqrt{0{,}05} \approx 0{,}2236$

Resultados manuales:

- Media = 2.2000 s
- Mediana = 2.2000 s
- Moda = No definida
- Varianza muestral = 0.0500 $\mathbf{s^2}$
- Desviación estándar = 0.2236 s

B) Solución digital en Python (VSCode)

[language=Python] Librerías necesarias import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns

1) Definir los datos datos = np.array([2.1, 2.3, 1.9, 2.5, 2.2])

2) Calcular medidas básicas mean = datos.mean() median = np.median(datos) mode = pd.Series(datos).mode() $\text{var}_sample = datos.var(ddof = 1) std_sample = datos.std(ddof = 1)$

3) Mostrar resultados print("Media:", mean) print("Mediana:", median) print("Moda(s):", mode.tolist()) print("Varianza muestral:", var$_sample)print("Desviaciónestándarmuestral : ", std_sample)$

4) Visualización rápida sns.histplot(datos, bins=5, kde=False) plt.title("Histograma: tiempos de carga") plt.xlabel("Tiempo (s)") plt.ylabel("Frecuencia") plt.show()

Explicación:

- `np.array` crea el vector de datos.
- `mean()`, `median()` y `mode()` calculan las medidas básicas.
- `var(ddof=1)` usa corrección de Bessel para varianza muestral.
- `sns.histplot` grafica la distribución.

C) Comparación manual vs digital

Medida	**Manual**	**Python (ejecución)**
Media	2.2000 s	2.2000 s
Mediana	2.2000 s	2.2000 s
Moda	No definida	Lista vacía o múltiples valores
Varianza muestral	0.0500 s^2	0.0500 s^2
Desviación estándar	0.2236 s	0.2236 s

Observación: Coincidencia completa entre cálculos manuales y digitales; la única diferencia está en la representación de la moda.

Ejemplo 2 — Datos agrupados (marcas y frecuencias)

Con el fin de analizar el comportamiento de los tiempos de procesamiento de un sistema informático bajo diferentes condiciones de carga, se recopilaron 30 observaciones que posteriormente fueron organizadas en intervalos para facilitar su interpretación. Cada intervalo representa un rango de tiempos (en segundos), junto con su marca de clase y la frecuencia correspondiente. La tabla de datos obtenida es la siguiente:

Tabla de datos

Intervalo	Marca (x_i)	Frecuencia (f_i)
2.0–2.4	2.2	8
2.4–2.8	2.6	15
2.8–3.2	3.0	5
3.2–3.6	3.4	2
Total		**30**

A partir de esta distribución de frecuencias, se solicita calcular la media agrupada, la varianza muestral y la desviación estándar muestral utilizando dos métodos:

Procedimiento manual, aplicando las fórmulas de estadística descriptiva para datos agrupados.

Procedimiento digital, empleando código en Python para verificar los resultados.

El objetivo es comparar ambos enfoques para confirmar la consistencia de los cálculos y analizar la variabilidad del sistema según los datos agrupados registrados.

Procedimiento manual

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}, \quad s^2 \approx \frac{\sum f_i (x_i - \bar{x})^2}{N-1}$$

$$\sum f_i x_i = 78{,}4 \quad \Rightarrow \quad \bar{x} = \frac{78{,}4}{30} = 2{,}6133$$

$$s^2 = \frac{3{,}3556}{29} = 0{,}1157, \quad s = 0{,}3402$$

Resultados manuales:

- Media agrupada = 2.6133
- Varianza = 0.1157
- Desviación estándar = 0.3402

Código en Python

import numpy as np

marcas = np.array([2.2, 2.6, 3.0, 3.4]) f = np.array([8, 15, 5, 2])

N = f.sum() $mean_agr = (marcas * f).sum()/Nnumerador = (f * (marcas - mean_agr) * *2).sum()var_agr = numerador/(N-1)std_agr = np.sqrt(var_agr)$

print("N:", N) print("Media agrupada:", mean$_a gr)print("Varianzaagrupada : ", var$_a gr)print("Desviacionestandar : ", std$_a gr)$

Comparación manual vs digital

Medida	**Manual**	**Python**
N	30	30
Media agrupada	2.6133	2.6133
Varianza	0.1157	0.1157
Desviación estándar	0.3402	0.3402

—

Ejemplo 3 — Percentiles, cuartiles y boxplot (Salarios)

Con el propósito de analizar la distribución salarial dentro de un equipo de trabajo de tamaño reducido, se recopilaron los salarios de diez empleados, expresados en millones de pesos. Este conjunto de datos permitirá estudiar la posición de los valores mediante medidas de orden, así como visualizar la dispersión y posibles valores atípicos mediante un boxplot.

Los salarios registrados fueron:

4.0,4.2,4.5,4.8,5.0,5.1,5.3,6.0,6.5,8.0

A partir de estos datos, se solicita:

Calcular manualmente los percentiles más relevantes, incluyendo:

Primer cuartil (Q1)

Mediana (Q2)

Tercer cuartil (Q3)

Rango intercuartílico (IQR)

Percentil 90 (P90)

Calcular los mismos valores mediante herramientas digitales, utilizando funciones de la biblioteca NumPy de Python para obtener percentiles y medidas de posición de manera automatizada.

Generar un boxplot que permita visualizar la distribución de los salarios, su dispersión y la posible presencia de valores extremos.

Finalmente, se comparan los resultados entre el procedimiento manual y el cálculo digital para evidenciar diferencias metodológicas —especialmente en los percentiles— y analizar cómo estas variaciones pueden influir en la interpretación de los datos.

Datos (n = 10, en millones)

4.0, 4.2, 4.5, 4.8, 5.0, 5.1, 5.3, 6.0, 6.5, 8.0

Resultados manuales

$$Q1 = 4{,}425, \quad \text{Mediana} = 5{,}05, \quad Q3 = 6{,}125, \quad IQR = 1{,}70, \quad P90 = 7{,}85$$

Código en Python

import numpy as np import matplotlib.pyplot as plt import seaborn as sns

salarios = np.array([4.0,4.2,4.5,4.8,5.0,5.1,5.3,6.0,6.5,8.0]) mediana = np.median(salarios) q1 = np.percentile(salarios, 25) q3 = np.percentile(salarios, 75) p90 = np.percentile(salarios, 90) iqr = q3 - q1

print("Mediana:", mediana) print("Q1:", q1) print("Q3:", q3) print(ÏQR:", iqr) print("P90:", p90)

sns.boxplot(x=salarios) plt.title("Boxplot de salarios (millones)") plt.xlabel("Salario (millones)") plt.show()

Comparación manual vs digital

Medida	Manual (k(n+1)/100)	Python (np.percentile)
Q1	4.425	4.58
Mediana	5.05	5.05
Q3	6.125	5.83
IQR	1.70	1.25
P90	7.85	6.65

—

Resumen comparativo (los 3 ejemplos)

Ejemplo	Medida	Manual	Python
1 (tiempos)	Media	2.2000	2.2000
1	s (muestral)	0.2236	0.2236
2 (agrupados)	Media agrupada	2.6133	2.6133
2	s agrupada	0.3402	0.3402
3 (salarios)	Mediana	5.05	5.05
3	Q1	4.425	4.58
3	Q3	6.125	5.83
3	P90	7.85	6.65

Conclusiones

La estadística descriptiva constituye la base fundamental del análisis de datos, ya que permite transformar información cruda en conocimiento comprensible mediante la recolección, organización, resumen y presentación de los datos. A lo largo del tiempo, ha evolucionado desde los antiguos censos y registros estatales hasta convertirse en una herramienta científica indispensable en todas las disciplinas, formalizándose gracias a los aportes de autores como Quetelet, Pearson y Galton, quienes consolidaron los conceptos de tendencia central, dispersión y correlación. En el ámbito de la ingeniería de sistemas, la estadística descriptiva es esencial para la gestión eficiente de datos y el diseño de soluciones informáticas, ya que permite analizar el rendimiento de los sistemas, detectar problemas y establecer métricas de desempeño. Los avances recientes integran esta rama con la inteligencia artificial, la automatización y el análisis de grandes volúmenes de datos (Big Data), mediante herramientas capaces de generar visualizaciones interactivas y análisis automáticos. En la actualidad, la estadística descriptiva cumple también una función ética y comunicativa al fomentar decisiones basadas en evidencia y en la interpretación responsable de la información. En síntesis, sigue siendo una herramienta científica vigente y en expansión, cuyo desarrollo continúa de la mano de la ciencia de datos y la inteligencia artificial, fortaleciendo las competencias analíticas necesarias para el ejercicio profesional de la ingeniería moderna.

Referencias

- Pareja, C., & Sevilla Arias, A. (2025, mayo 29). *Estadística descriptiva: Qué es, tipos y ejemplos*. Economipedia. Recuperado de https://economipedia.com/definiciones/estadistica-descriptiva.html
- López, J. F. (2019, noviembre 15). *Origen de la estadística—Qué es y su impacto histórico. Economipedia.*. Recuperado de https://economipedia.com/definiciones/origen-estadistica.html
- Akal, E. (2020, octubre 20). *DS-STAR: A state-of-the-art, versatile data science agent.*. Recuperado de https://www.nocierreslosojos.com/estadistica-moderna-histo
- IBM. (2025). *¿Qué es el análisis exploratorio de datos?*. Recuperado de https://www.ibm.com/es-es/think/topics/exploratory-data-analysis
- Yoon, J., & Nam, J. (2025, noviembre 6). *DS-STAR: A state-of-the-art versatile data science agent. Google Cloud.*. Recuperado de https://research.google/blog/ds-star-a-state-of-the-art-versatile-data-science-agent/
- Flourish. (s. f.). *Flourish*. Recuperado de https://flourish.studio/
- Tableau. (2021). *¿Qué es Tableau?*. Recuperado de https://www.tableau.com/es-es/why-tableau/what-is-tableau