

Estadística Descriptiva

Sección 1: Fundamentos - Estadística, Probabilidad e Incertidumbre

1.1 El Ecosistema del Análisis de Datos: Estadística Descriptiva vs. Inferencial

El análisis de datos, en su núcleo, se divide en dos grandes ramas: la estadística descriptiva y la estadística inferencial. Comprender esta distinción es el primer paso para un análisis riguroso.

La **Estadística Descriptiva** se enfoca en la organización, resumen y presentación de los datos observados. Su objetivo es revelar patrones, tendencias y características inherentes a un conjunto de datos específico. Utiliza medidas como el promedio (media), la dispersión (diferencia entre valores) y la forma de la distribución (sesgo o asimetría).¹ En el contexto de la ciencia de datos, esta disciplina es la piedra angular del Análisis Exploratorio de Datos (EDA). No busca ir más allá de los datos que se tienen a mano; su propósito es "describir" lo que se ve.

La **Estadística Inferencial**, por otro lado, utiliza los resúmenes generados por la estadística descriptiva para hacer generalizaciones, predicciones o inferencias sobre una población más amplia de la cual se extrajo la muestra de datos. La estadística descriptiva es, por tanto, la materia prima indispensable para la inferencial. Un error en la descripción (por ejemplo, usar una media para caracterizar un conjunto de datos profundamente sesgado) inevitablemente conducirá a una inferencia errónea y, en consecuencia, a una toma de decisiones deficiente.

1.2 La Relación Fundamental: Probabilidad y Estadística ante la Incertidumbre

La estadística y la probabilidad están intrínsecamente ligadas; son dos caras de la misma moneda en la gestión de la incertidumbre. La incertidumbre es el estado natural en el que opera un científico de datos.

La **Probabilidad** es la rama de las matemáticas que modela la incertidumbre. Opera de manera deductiva: partiendo de un modelo conocido (ej. una distribución de probabilidad

teórica, como la normal o la exponencial ²⁾, predice la posibilidad de observar ciertos datos o eventos. Permite cuantificar la verosimilitud de resultados futuros.

La **Estadística** opera de manera inductiva: utiliza los datos observados (la muestra) para inferir las propiedades del modelo subyacente que los generó.

Esta simbiosis es fundamental en la ciencia de datos. Por ejemplo, al comparar el rendimiento de un modelo de *machine learning* (Modelo A) con uno nuevo (Modelo B), no es suficiente observar que el Modelo B tiene una métrica de error ligeramente mejor en un conjunto de validación.³ Esta diferencia podría ser producto del azar. Es aquí donde la estadística

inferencial, alimentada por la teoría de la probabilidad, se vuelve crucial. Se deben utilizar pruebas estadísticas para responder: ¿Es la mejora observada en el Modelo B

estadísticamente significativa, o es probable que haya ocurrido por casualidad?³

En esencia, la probabilidad y la estadística proporcionan el marco matemático para la toma de decisiones bajo incertidumbre, permitiendo cuantificar el riesgo de estar equivocados y movernos de una simple observación a una conclusión robusta.³

Sección 2: La Taxonomía de los Datos: Clasificación y Casuística Exhaustiva de las Variables

El error más fundamental en el análisis estadístico es aplicar una operación matemática a un tipo de variable que no la soporta. La elección de cualquier método (un gráfico, una medida de tendencia, un modelo de *machine learning*) está estrictamente supeditada y restringida por el tipo de variable. Un error en esta taxonomía invalida todo el análisis subsiguiente.

2.1 La Jerarquía Primaria: Datos Cualitativos (Categóricos) vs. Cuantitativos (Numéricos)

La primera distinción divide los datos en dos grandes familias ⁴⁾:

- **Datos Cualitativos (o Categóricos):** Describen una cualidad, característica o etiqueta.⁵ No son intrínsecamente numéricos, aunque a menudo se codifican numéricamente (ej. 1="Hombre", 2="Mujer").⁷ Responden a la pregunta "¿de qué tipo?".
- **Datos Cuantitativos (o Numéricos):** Describen una cantidad medible o contable. Siempre son números sobre los que las operaciones aritméticas tienen sentido.⁵ Responden a la pregunta "¿cuánto?" o "¿cuántos?".

2.2 La Jerarquía de Stevens (Escala de Medida): El Pilar del Rigor Estadístico

La clasificación primaria es útil, pero insuficiente. El pilar del rigor estadístico se encuentra en la jerarquía de las escalas de medida, propuesta por S.S. Stevens. Esta jerarquía es acumulativa: cada nivel hereda las propiedades del nivel anterior y añade una nueva restricción.⁸

2.2.1 Escala Nominal

- **Propiedades:** Identidad. Los valores son etiquetas o categorías mutuamente excluyentes.⁶ No existe un orden intrínseco o jerarquía.⁹
- **Operaciones Válidas:** Conteo (frecuencia), Moda. No se pueden promediar ni ordenar.
- **Ejemplos en Data Science:** `user_id`, `product_category`, `region`, `zip_code` (código postal)⁹, `gender`⁸, o variables *dummy* (0/1) que representan la pertenencia a una categoría.¹⁰
- **Error Común:** Calcular la "media de los códigos postales" de los clientes. El número resultante carece de todo significado estadístico, ya que el código postal es una etiqueta nominal.⁹

2.2.2 Escala Ordinal

- **Propiedades:** Identidad + **Orden**.⁷ Las categorías tienen una jerarquía o *ranking* claro.⁷
- **Limitación Clave:** Si bien se conoce el orden, las diferencias (intervalos) entre las categorías no son uniformes, medibles ni significativas.⁷ No se sabe si la "distancia" entre "Muy probable" y "Probable" es la misma que entre "Probable" y "Neutral".⁷
- **Operaciones Válidas:** Mediana, Percentiles, Frecuencias, Moda. La media aritmética está prohibida, ya que asume intervalos iguales.
- **Ejemplos en Data Science:** Resultados de encuestas de satisfacción (ej. "Muy Insatisfecho" a "Muy Satisfecho")⁷, nivel educativo (ej. "Licenciatura", "Maestría")¹⁰, clasificaciones de clientes (ej. "Bronce", "Plata", "Oro"), *rankings* de resultados.⁸

2.2.3 Escala de Intervalo

- **Propiedades:** Identidad + Orden + **Intervalos Iguales**.¹¹ La diferencia entre dos valores es constante y significativa. La diferencia entre 20°C y 10°C es la misma que entre 10°C y 0°C.⁷
- **Limitación Clave:** El **cero es arbitrario** y no representa la ausencia absoluta de la variable.⁷ 0°C no es "ausencia de calor", es simplemente un punto en la escala.
- **Operaciones Válidas:** Suma, Resta, Media, Mediana, Desviación Estándar.
- **Operaciones Prohibidas:** Multiplicación y División (Ratios). No se puede decir que

20°C es "el doble de caliente" que 10°C.

- **Ejemplos en Data Science:** Temperatura en Celsius o Fahrenheit ⁷, fechas (ej. marcas de tiempo datetime, donde el "año 0" es una convención) ¹², puntuaciones en tests estandarizados (ej. CI, SAT).¹³

2.2.4 Escala de Razón

- **Propiedades:** Identidad + Orden + Intervalos Iguales + **Cero Absoluto y Real**.¹¹ El cero representa la ausencia total de la variable.⁷
- **Operaciones Válidas:** Todas las operaciones aritméticas (Suma, Resta, Multiplicación, División, Ratios).¹¹ Se puede decir que un ingreso de 100,000€ es el doble de un ingreso de 50,000€.
- **Ejemplos en Data Science:** Ingresos, age (edad) ¹⁰, altura, peso ¹⁴, session_duration (duración de sesión en segundos), word_count (conteo de palabras) ¹⁵, velocidad.¹⁶ La mayoría de las variables cuantitativas en ciencia de datos pertenecen a esta escala.

La siguiente tabla sintetiza esta jerarquía fundamental, que dicta las operaciones válidas para cada tipo de variable.

Escala de Medida	Propiedades Clave	Cero	Operaciones Válidas (Tendencia)	Operaciones Válidas (Dispersión)	Ejemplo de Data Science
Nominal	Identidad, Categorías	N/A	Moda, Frecuencias	N/A (Solo conteo)	product_category (ej. 'Electrónica') ⁹
Ordinal	Identidad, Orden (Jerarquía)	N/A	Mediana, Moda, Percentiles	Rango Intercuartílico (IQR)	satisfaction_rating (ej. 'Muy Satisfecho') ⁷
Intervalo	Identidad, Orden, Intervalos Iguales	Arbitrario	Media, Mediana, Moda	Desviación Estándar, Varianza, IQR	temperature_celsius (Temperatura) ¹²
Razón	Identidad, Orden, Intervalos Iguales	Absoluto (Real)	Todas (Media, Mediana, Geométrica, etc.)	Todas (Desviación Estándar, CV, etc.)	revenue (Ingresos) ¹⁴

2.3 Casuísticas Avanzadas de Variables en Data Science

Para un análisis exhaustivo, la jerarquía de Stevens debe complementarse con casuísticas específicas que se encuentran habitualmente en la ciencia de datos.

2.3.1 Cuantitativas Discretas vs. Continuas

Esta es una subclasificación de las variables de Razón e Intervalo.¹⁷

- **Discretas:** Variables que se *cuentan*. Toman un número finito o contablemente infinito de valores, generalmente enteros.⁷ Ejemplos: "número de pantalones" ¹⁸, "número de hijos" ¹⁰, `page_views` (vistas de página).
- **Continuas:** Variables que se *miden*. Pueden tomar un número infinito de valores dentro de un rango determinado.⁷ Ejemplos: "peso" ¹⁸, "temperatura" ⁷, `session_duration`.

En la práctica de la ciencia de datos, esta distinción a menudo se vuelve borrosa. Variables como "ingresos" o "edad" son técnicamente discretas (se miden en céntimos o días), pero su *cardinalidad* (el número de valores únicos) es tan alta que se tratan como continuas para fines de modelado. El contexto y la cardinalidad de la variable importan más que la definición teórica pura.¹⁸

2.3.2 Variables Cíclicas

Esta es una casuística crucial en *feature engineering*. Se trata de variables donde los valores extremos son conceptualmente adyacentes.¹⁹

- **El Problema:** El ejemplo clásico es la hora del día. Para un modelo de regresión, el valor 23 (23:00) y el valor 0 (00:00) están muy alejados numéricamente, aunque en la realidad solo están separados por una unidad de tiempo.²⁰ Un modelo lineal no puede capturar la idea de que "la medianoche" es un continuo.
- **La Solución:** Estas variables no pueden usarse "en crudo". Deben ser transformadas para representar su naturaleza circular. La técnica estándar es descomponer la variable en dos nuevas características usando funciones trigonométricas: una componente de seno y una de coseno.²¹ Esto ubica los datos en un círculo de 2D, donde 23:59 y 00:00 están, de hecho, muy próximos.
- **Ejemplos:** Hora del día, día de la semana, mes del año ²¹, dirección del viento.¹⁹

2.3.3 Variables de "Log-Intervalo"

Estas son variables que operan inherentemente en una escala logarítmica, no lineal.¹⁹

- **El Problema:** En estas escalas, los intervalos no son aditivos, sino multiplicativos. Un cambio de pH de 7 a 6 (10 veces más ácido) no es comparable a un cambio de 5 a 4 (que también es 10 veces más ácido que 5, pero 100 veces más que 6).²³ Tratar estos datos con una media aritmética viola sus supuestos.

- **Ejemplos:** El pH (acidez) ¹³, los decibelios (dB) para el sonido ¹³ y la escala de Richter para terremotos.¹⁹ Para el análisis, estas variables a menudo se manejan en su escala logarítmica nativa o se transforman (usando la función exponencial) de nuevo a una escala lineal si el análisis lo requiere.

Sección 3: Análisis Univariante Exhaustivo: Medidas de Tendencia y Dispersión

Una vez establecida la taxonomía de la variable (Sección 2), el siguiente paso es seleccionar la herramienta estadística *correcta* para resumirla. El análisis univariante examina una variable a la vez, centrándose en dos propiedades clave: su "centro" (tendencia central) y su "propagación" (dispersión).

3.1 Medidas de Tendencia Central (MTC): Encontrando el "Centro"

La elección de una MTC depende críticamente de dos factores: la escala de medida de la variable (ver Tabla 1) y la presencia de valores atípicos (*outliers*) en la distribución.²⁴

3.1.1 Las Medidas Clásicas y su Robustez

- **Media Aritmética (Promedio):**
 - **Definición:** La suma de todos los valores dividida por el número de valores.¹
 - **Aplicabilidad:** Es la medida de tendencia central por defecto para datos de **Intervalo y Razón** que son razonablemente **simétricos**.²⁴ Es la base de innumerables técnicas inferenciales (como la prueba t y el ANOVA) debido a sus propiedades matemáticas manejables.²⁴
 - **Robustez: No es robusta.** Es extremadamente sensible a los valores atípicos.²⁴ Un solo *outlier* puede arrastrar la media y hacerla un mal representante del centro de los datos. Se dice que tiene un "punto de ruptura" del 0%, ya que un solo valor extremo puede distorsionarla infinitamente.²⁴
 - **Ejemplo de Falla:** Calcular el "ingreso medio" de un país. Un pequeño número de multimillonarios sesgará la media hacia arriba, dando una impresión falsa del ingreso de la persona típica.²⁷
- **Mediana (Percentil 50):**
 - **Definición:** El valor que ocupa la posición central en un conjunto de datos *ordenados*.²⁵ El 50% de los datos está por encima de ella y el 50% por debajo.²⁶
 - **Aplicabilidad:** Es la medida ideal para datos **Ordinales** ²⁴ y la mejor y más

robusta opción para datos de **Razón e Intervalo** que están **sesgados** o contienen *outliers*.²⁴

- **Robustez: Muy robusta.** Es insensible a los *outliers*, ya que su valor solo depende de la(s) observación(es) central(es), no de la magnitud de los valores extremos.²⁷ Tiene un "punto de ruptura" del 50%, lo que significa que hasta la mitad de los datos tendrían que estar contaminados para moverla.²⁴
- **Ejemplo de Uso:** En ciencia de datos, se debe usar la mediana por defecto al reportar sobre *salarios*, *precios de vivienda* o cualquier distribución de ingresos, ya que casi siempre están sesgados.²⁷

- **Moda:**

- **Definición:** El valor (o valores) que aparece con mayor frecuencia en el conjunto de datos.²⁴
- **Aplicabilidad:** Es la **única** medida de tendencia central válida para datos **Nominales** (ej. ¿cuál es el `product_category` más frecuente?).²⁴
- **Casuística:** También es útil en datos cuantitativos para identificar distribuciones *multimodales*. Por ejemplo, el análisis de las horas de uso de un servicio de streaming puede revelar dos modas: una por la mañana (antes del trabajo) y otra por la noche (después del trabajo). La media o mediana ocultaría esta estructura bimodal.

3.1.2 Medidas Avanzadas ("Exhaustividad")

Para un análisis riguroso, las medidas clásicas son a menudo insuficientes.

- **Media Ponderada:**

- **Concepto:** Una media aritmética donde no todas las observaciones contribuyen por igual. A cada observación se le asigna un "peso" (w_i) que refleja su importancia.³⁰ La fórmula es $\frac{\sum(x_i \times w_i)}{\sum(w_i)}$.
- **Aplicabilidad:** Se utiliza cuando los puntos de datos tienen diferentes niveles de importancia.³⁰
- **Ejemplos en Data Science:**
 1. **Métricas de Negocio:** Calcular la "calificación promedio de un producto" debe ponderarse por el "número de valoraciones" o "volumen de ventas" de cada producto. Un producto con 5 estrellas y 1000 ventas es más importante que uno con 5 estrellas y 1 venta.³⁰
 2. **Evaluación de Modelos:** Al promediar el *score* de un modelo en *k-folds* de validación cruzada, si los *folds* tienen tamaños diferentes, se debe usar una media ponderada por el tamaño de cada *fold* para obtener el rendimiento agregado correcto.
 3. **Finanzas:** Calcular el precio promedio ponderado de una cartera de acciones.³¹

- **Media Geométrica:**

- **Concepto:** La raíz n -ésima del producto de N números: $\sqrt[n]{x_1 \times x_2 \times \dots \times x_n}$.³³
- **Aplicabilidad:** Es la medida correcta para promediar valores que son *multiplicativos* por naturaleza, como tasas de crecimiento o ratios.³⁴
- **Ejemplo en Data Science:**
 1. **Crecimiento de Usuarios:** Si una *startup* crece un 100% el primer año (la base de usuarios se duplica: $\times 2.0$) y decrece un 50% el segundo año (se reduce a la mitad: $\times 0.5$).
 - **Error (Media Aritmética):** $(2.0 + 0.5) / 2 = 1.25$. Sugiere un crecimiento promedio del 25%.
 - **Correcto (Media Geométrica):** $\sqrt{2.0 \times 0.5} = \sqrt{1.0} = 1.0$. Sugiere un crecimiento promedio del 0%, lo cual es correcto, ya que el usuario vuelve al tamaño original.
 2. **Finanzas:** Es la medida estándar para calcular el retorno promedio de inversiones (tasas de interés compuesto).³⁴
- **Media Armónica:**
 - **Concepto:** El recíproco de la media aritmética de los recíprocos de los valores: $n / \sum(1/x_i)$.³⁵
 - **Aplicabilidad:** Es la medida correcta para promediar *tasas* o *ratios* cuando el numerador es constante.³⁵ El ejemplo clásico es promediar velocidades (ej. distancia/tiempo).
 - **Ejemplo de Uso (La Casuística Clave en Data Science): El F1-Score.**
 - El F1-Score, una métrica fundamental en la clasificación de *machine learning*, es la media armónica de la Precisión (P) y la Exhaustividad (Recall): $F_1 = 2 \times (P \times R) / (P + R)$.
 - ¿Por qué la media armónica? Porque penaliza severamente los valores bajos. Si un modelo tiene una Precisión excelente ($P=1.0$) pero una Exhaustividad terrible ($R=0.1$), una media aritmética sugeriría un rendimiento decente (0.55). Sin embargo, la media armónica ($F_1=0.18$) refleja correctamente que el modelo es deficiente porque una de sus métricas componentes ha colapsado. Se utiliza para encontrar un balance óptimo cuando ambas métricas (P y R) son igualmente importantes.
- **Media Truncada (o Recortada):**
 - **Concepto:** Una media calculada después de *descartar* un porcentaje fijo de los valores más bajos y más altos (ej. el 10% inferior y el 10% superior).³⁸
 - **Aplicabilidad:** Es un *compromiso* entre la media aritmética y la mediana.⁴⁰ Es más robusta a *outliers* que la media (porque los elimina), pero utiliza más información de los datos que la mediana (que solo usa el valor central).
 - **Ejemplo de Uso:** Se utiliza en competiciones deportivas (como el patinaje artístico) para calcular la puntuación final, eliminando las puntuaciones de los jueces de los extremos para mitigar el sesgo.⁴¹ En ciencia de datos, puede usarse para calcular un "promedio robusto" del tiempo de carga de una página web,

ignorando los tiempos extremadamente rápidos o lentos que pueden ser anomalías.

3.2 Medidas de Dispersión: Cuantificando la Variabilidad

La tendencia central por sí sola es incompleta; se necesita una medida de cuán "dispersos" están los datos alrededor de ese centro.¹

3.2.1 Medidas Absolutas (En unidades de datos)

- **Rango:** La diferencia entre el valor máximo y el mínimo.⁴² Es simple de calcular pero **no es robusto**, ya que depende exclusivamente de los dos *outliers* más extremos.⁴²
- **Varianza (\$s^2\$):** El promedio de las diferencias al cuadrado entre cada valor y la media.⁴³ Es matemáticamente fundamental, pero sus unidades están *al cuadrado* (ej. "dólares al cuadrado"), lo que dificulta su interpretación práctica.⁴⁴
- **Desviación Estándar (o Típica) (\$s\$):** La raíz cuadrada de la varianza.⁴³
 - **Aplicabilidad:** Es la medida de dispersión más común. Se interpreta en las **mismas unidades** que los datos originales.⁴⁵ Es el compañero natural de la **Media Aritmética**.
 - **Robustez:** Al igual que la media, **no es robusta**. Los *outliers* (que tienen grandes desviaciones de la media) se elevan al cuadrado en el cálculo de la varianza, por lo que la desviación estándar es muy sensible a ellos.²⁷
- **Rango Intercuartílico (IQR):**
 - **Definición:** La diferencia entre el percentil 75 (Q3) y el percentil 25 (Q1): $IQR = Q3 - Q1$.⁴⁶ Representa el rango en el que se encuentra el 50% central de los datos.
 - **Aplicabilidad:** Es el compañero natural de la **Mediana**. Es la medida de dispersión **robusta** estándar.²⁷
 - **Robustez:** Es robusta porque, al igual que la mediana, su cálculo ignora los valores en los extremos de la distribución (el 25% inferior y el 25% superior).⁴⁶ Es la base para la construcción de *Box Plots* (diagramas de caja y bigotes).
- **Desviación Absoluta Mediana (MAD):**
 - **Definición:** La *mediana* de las desviaciones absolutas con respecto a la *mediana* de los datos: $MAD = \text{mediana}(|x_i - \text{mediana}(x)|)$.⁴⁶
 - **Aplicabilidad:** Es la medida de dispersión **más robusta** disponible.⁴⁶ Es aún más resistente a los *outliers* que el IQR. Se utiliza en la detección de anomalías cuando se sospecha que los datos están fuertemente contaminados por valores atípicos.

3.2.2 Medidas Relativas (Adimensionales)

- **Coefficiente de Variación (CV):**

- **Definición:** Una medida de dispersión relativa, calculada como la desviación estándar dividida por la media (generalmente expresada como porcentaje): $CV = (s / \text{media}) \times 100\%$.⁴⁸
- **Aplicabilidad (La Casuística Clave):** Se utiliza cuando se necesita **comparar la variabilidad** de dos o más variables que tienen **medias muy diferentes** o que están medidas en **unidades diferentes**.⁴⁹ La desviación estándar no se puede comparar directamente en estos casos, pero el CV, al ser adimensional, sí.
- **Ejemplo en Data Science:** Se desea determinar qué métrica es más volátil: las ventas de un producto (Media = 1,000,000€, SD = 100,000€) o los clics en un anuncio (Media = 500 clics, SD = 100 clics).
 - *Análisis Erróneo:* Comparar las SD (100,000 vs 100) no tiene sentido.
 - *Análisis Correcto (CV):*
 - $CV(\text{Ventas}) = (100,000 / 1,000,000) = 0.10$ o 10%.
 - $CV(\text{Clics}) = (100 / 500) = 0.20$ o 20%.
 - *Conclusión:* Los clics son *relativamente* dos veces más volátiles (o dispersos) que las ventas.

Medida	Tipo	Robusta a Outliers?	Escala de Variable Aplicable	Caso de Uso Principal en Data Science
Media Aritmética	MTC	No	Intervalo, Razón	Datos simétricos; base para inferencia (ej. ANOVA). ²⁴
Mediana	MTC	Sí	Ordinal, Intervalo, Razón	Datos sesgados o con outliers (ej. ingresos, precios). ²⁷
Moda	MTC	Sí	Nominal, Ordinal, Intervalo, Razón	Datos categóricos; identificar multimodalidad. ²⁴
Media Ponderada	MTC	No	Razón	Promediar observaciones con diferente importancia (ej. métricas de negocio). ³⁰

Media Geométrica	MTC	No	Razón (positiva)	Promediar tasas de crecimiento o valores multiplicativos. ³⁵
Media Armónica	MTC	No	Razón (positiva)	Promediar tasas o ratios (ej. F1-Score: P y R). ³⁷
Media Truncada	MTC	Parcialmente	Intervalo, Razón	Compromiso entre robustez (mediana) y eficiencia (media). ³⁸
Rango	Dispersión	No	Ordinal, Intervalo, Razón	Cálculo rápido pero muy inestable; no recomendado. ⁴²
Varianza	Dispersión	No	Intervalo, Razón	Fundamental matemáticamente, pero difícil de interpretar. ⁴⁴
Desviación Estándar	Dispersión	No	Intervalo, Razón	Acompañante de la media; dispersión en unidades originales. ⁴⁵
IQR	Dispersión	Sí	Ordinal, Intervalo, Razón	Acompañante de la mediana; dispersión robusta (Box Plots). ⁴⁶
MAD	Dispersión	Muy Alta	Intervalo, Razón	Detección de anomalías; la medida de dispersión más robusta. ⁴⁶
Coef. de Variación (CV)	Dispersión (Relativa)	No	Razón	Comparar volatilidad de variables con diferentes escalas. ⁴⁹

Sección 4: Técnicas de Muestreo y sus Implicaciones en Data Science

El muestreo es el proceso de seleccionar un subconjunto (la muestra) de una población más grande para su análisis. Las conclusiones de cualquier modelo de *machine learning* o prueba estadística dependen enteramente de la calidad y el método de ese muestreo. La forma en que se selecciona la muestra determina si los hallazgos pueden ser generalizados.

4.1 El Dilema Central: Muestreo Probabilístico vs. No Probabilístico

- **Muestreo Probabilístico:** Cada miembro de la población objetivo tiene una probabilidad conocida y no nula de ser seleccionado.⁵¹ Este método se basa en la aleatorización y es la base de la inferencia estadística formal. Solo las muestras probabilísticas permiten generalizar los resultados a la población con un nivel de confianza medible.
- **Muestreo No Probabilístico:** La selección de la muestra se basa en la conveniencia, el juicio del investigador u otros criterios no aleatorios.⁵² Es más rápido, económico y, a menudo, la única opción viable. Sin embargo, introduce sesgos y no permite la generalización formal a la población.⁵²

4.2 Técnicas de Muestreo Probabilístico (El Ideal Científico)

- **Muestreo Aleatorio Simple:** Cada miembro de la población tiene exactamente la misma probabilidad de ser elegido.⁵¹
- **Muestreo Sistemático:** Se selecciona un punto de partida aleatorio y luego se elige a cada k-ésimo miembro de la población.
- **Muestreo Estratificado:** La población se divide primero en subgrupos homogéneos (estratos) que son mutuamente excluyentes (ej. por demografía, geografía).⁵¹ Luego, se realiza un muestreo aleatorio simple *dentro* de cada estrato.
 - **Aplicación Crítica en Data Science:** Esta técnica es **esencial** para manejar **datasets desbalanceados** en *machine learning*.⁵⁵ Si se está construyendo un modelo de detección de fraude donde solo el 1% de las transacciones son fraudulentas, un muestreo aleatorio simple podría resultar en una muestra de entrenamiento casi sin casos de fraude.⁵⁵ El muestreo estratificado (usando la variable "fraude" como estrato) *garantiza* que la muestra de entrenamiento y de prueba mantenga la proporción de clases (o una proporción diseñada, en caso de sobremuestreo/submuestreo), permitiendo que el modelo aprenda a identificar la clase minoritaria.⁵⁷

- **Muestreo por Conglomerados (Clusters):** La población se divide en grupos (conglomerados, ej. ciudades), y se seleccionan aleatoriamente *conglomerados enteros* para el análisis.

4.3 Técnicas de Muestreo No Probabilístico (La Realidad Práctica en DS)

- **Muestreo por Conveniencia:** El investigador selecciona la muestra que es más accesible (ej. estudiantes voluntarios, encuestas enviadas a una lista de correo).⁵² En la práctica, casi todos los datos recopilados de fuentes de Internet (tráfico de un sitio web, datos de redes sociales, usuarios de una aplicación) son muestras por conveniencia.
- **Muestreo por Cuotas:** Un intento de mejorar el muestreo por conveniencia.⁵³ El investigador establece cuotas para subgrupos (ej. 50% hombres, 50% mujeres) y luego llena esas cuotas por conveniencia.⁵³
- **Muestreo de Bola de Nieve:** Los participantes existentes reclutan a futuros participantes.⁵³ Es útil para estudiar poblaciones ocultas o difíciles de alcanzar (ej. usuarios de un software muy específico).⁵³

4.4 La Consecuencia del Muestreo: Validez Interna vs. Externa

El uso de técnicas de muestreo no probabilísticas, especialmente por conveniencia, introduce un **sesgo de muestreo** o **sesgo de selección**.⁶² La muestra no es representativa de la población de interés.⁶⁵ Esto tiene implicaciones directas en la validez del modelo:

- **Validez Interna:** Se refiere a si las conclusiones del estudio son válidas *para la muestra* que se estudió. Un A/B test puede tener una alta validez interna (ej. "La Versión B venció a la A en nuestro grupo de prueba").
- **Validez Externa:** Se refiere a la capacidad de *generalizar* los resultados del estudio a la población más amplia.⁶²

El problema central en ciencia de datos es que los modelos a menudo se entrenan en muestras por conveniencia (baja validez externa).⁶³ Un modelo de *churn* (abandono) entrenado con datos de usuarios de EE. UU. en 2023 puede funcionar perfectamente en su conjunto de prueba (alta validez interna), pero fallar estrepitosamente cuando se aplica a usuarios de Europa o Asia en 2024 (baja validez externa).

Esta limitación es crítica en el **A/B Testing**. Los participantes en un A/B test (ej. usuarios que visitaron el sitio durante una semana específica) son una muestra de conveniencia.⁶⁷ Si la Versión B gana, la conclusión rigurosa no es "B es mejor que A", sino "B fue mejor que A *para este grupo de usuarios en este período de tiempo*". La generalización a "todos los usuarios futuros" es un salto inferencial que conlleva riesgo.

Sección 5: Análisis Bivariante: La Matriz de Relación "Para Todo Tipo de Variables"

El análisis univariante describe los datos; el análisis bivariante busca relaciones entre ellos.⁶⁹ El objetivo es responder: "¿Cómo se relaciona la variable X con la variable Y?". La elección de la técnica depende estrictamente de la escala de medida (Sección 2) de las dos variables involucradas.⁷¹

5.1 Caso 1: Relación Cuantitativa (Q) - Cuantitativa (Q)

(Ej. Razón/Intervalo vs. Razón/Intervalo)

Se busca medir cómo covarían dos variables numéricas (ej. age vs. income).

- **Visualización: El Diagrama de Dispersión (*Scatter plot*)** es la herramienta principal. Permite una inspección visual de la *forma* (lineal, no lineal), la *dirección* (positiva, negativa) y la *fuerza* de la relación.⁷²
- **Medida (Lineal): Coeficiente de Correlación de Pearson (r)**
 - **Qué Mide:** La fuerza y la dirección de una **relación lineal** entre dos variables.⁷³ Varía de -1 (lineal negativa perfecta) a +1 (lineal positiva perfecta).
 - **Casuística:** Un r cercano a 0 **no** significa "ausencia de relación"; significa "ausencia de relación *lineal*".⁷² Una relación cuadrática perfecta (ej. en forma de U) puede tener un r de 0.
- **Medida (Monotónica): Coeficiente de Correlación de Spearman (ρ)**
 - **Qué Mide:** La fuerza y la dirección de una **relación monotónica**.⁷³ Una relación es monotónica si, a medida que X aumenta, Y consistentemente aumenta (o disminuye), pero no necesariamente a un ritmo constante (no lineal).
 - **Cómo Funciona:** Calcula la correlación de Pearson sobre los *rangos* (posiciones ordenadas) de los datos, no sobre los valores en sí.⁷³
 - **Cuándo Usar:**
 1. Cuando la relación visual en el *scatter plot* es claramente no lineal pero sí monotónica (ej. una curva creciente).
 2. Cuando una o ambas variables son de escala **Ordinal**.⁷³
 3. Cuando hay *outliers* significativos en los datos, ya que al operar sobre rangos, Spearman es robusto a ellos.

5.2 Caso 2: Relación Categórica (C) - Categórica (C)

(Ej. Nominal/Ordinal vs. Nominal/Ordinal)

Se busca medir la asociación entre dos variables categóricas (ej. `product_category` vs. `region`).

- **Visualización: Tabla de Contingencia** (o tabla cruzada) ⁷⁶ y **Mapas de Calor (Heatmaps)** o Gráficos de Barras Agrupadas/Apiladas.
- **Medida (Asociación): Prueba Chi-Cuadrado (χ^2)**
 - **Qué Mide:** Compara las frecuencias *observadas* en la tabla de contingencia con las frecuencias *esperadas* (teóricas) que se verían si las dos variables fueran completamente independientes.⁷⁷
 - **Interpretación:** Un valor p bajo (ej. $p < 0.05$) sugiere que existe una asociación estadísticamente significativa (es improbable que la distribución observada ocurra por azar).
 - **Limitación:** χ^2 solo indica si existe una asociación; no mide la *fuerza* ni la dirección de la misma.⁷⁷
- **Medida (Fuerza de Asociación): V de Cramer y Coeficiente Phi (ϕ)**
 - **Concepto:** Son coeficientes derivados de χ^2 que normalizan el valor para medir la *fuerza* de la asociación, escalándolo entre 0 (sin asociación) y 1 (asociación perfecta).⁷⁷
 - **Cuándo Usar:** El **Coeficiente Phi (ϕ)** se utiliza para tablas de 2×2 . La **V de Cramer** se utiliza para tablas más grandes ($N \times M$).⁷⁷

5.3 Caso 3: Relación Cuantitativa (Q) - Categórica (C)

(Ej. *Razón/Intervalo* vs. *Nominal/Ordinal*)

Se busca comparar la distribución de una variable cuantitativa entre diferentes grupos definidos por una variable categórica (ej. `income` vs. `education_level`).

- **Visualización: Box Plots (Diagramas de Caja) Comparativos.**⁷¹ Esta es la herramienta visual por excelencia, ya que permite comparar simultáneamente la mediana (centro), el IQR (dispersión) y los *outliers* de la variable (Q) para cada nivel de la variable (C).
- **Medida (Comparación de Medias):**
 - **Si la variable (C) tiene 2 niveles (Dicotómica): Prueba t de Student**
 - **Qué Mide:** Evalúa si la diferencia entre las *medias* de la variable (Q) en los dos grupos (definidos por C) es estadísticamente significativa o si podría deberse al azar.⁷⁹
 - **Si la variable (C) tiene >2 niveles: Análisis de Varianza (ANOVA)**
 - **Qué Mide:** Es una generalización de la prueba t. Compara la *varianza entre las medias de los grupos* con la *varianza dentro de los grupos* (usando el estadístico F).⁷⁹
 - **Interpretación:** Una prueba ANOVA significativa (valor p bajo) indica que *al menos una* de las medias de los grupos es significativamente diferente de las demás.⁷⁹
- **Medida (Correlación): Coeficiente de Correlación Biserial Puntual (r_{pb})**

- **Aplicabilidad:** Mide la fuerza de la relación entre una variable **continua (Q)** y una variable **genuinamente dicotómica (C)** (ej. test_score vs. passed/failed; income vs. bought/did_not_buy).⁸¹
- **Conexión Rigurosa:** La Correlación Biserial Puntual es matemáticamente equivalente a la Prueba t de Student.⁸³ De hecho, el cuadrado de r_{pb} se puede convertir directamente en el estadístico t . Probar la significancia de r_{pb} arroja el mismo valor p que una prueba t independiente.⁸³ Ofrecen dos perspectivas (asociación vs. diferencia de medias) sobre el mismo fenómeno.

	Variable 2: Categórica (C)(Nominal / Ordinal)	Variable 2: Cuantitativa (Q)(Intervalo / Razón)
Variable 1: Categórica (C) (Nominal / Ordinal)	C vs. C Visualización: Tabla de Contingencia, Mapa de Calor. Test (Asociación): Prueba Chi-Cuadrado (χ^2). ⁷⁷ Test (Fuerza): V de Cramer (para N x M), Phi (ϕ) (para 2x2). ⁷⁷	C vs. Q (Ver celda Q vs. C)
Variable 1: Cuantitativa (Q) (Intervalo / Razón)	Q vs. C Visualización: Box Plots Comparativos por grupo. ⁷⁸ Test (Diferencia de Medias): <ul style="list-style-type: none"> • Prueba t de Student (si C tiene 2 niveles).⁷⁹ • ANOVA (si C tiene >2 niveles).⁸⁰ Test (Asociación): Correlación Biserial Puntual (si C tiene 2 niveles). ⁸¹	Q vs. Q Visualización: Diagrama de Dispersión (Scatter Plot). ⁷² Test (Lineal): Correlación de Pearson (r). ⁷⁴ Test (Monotónica): Correlación de Spearman (ρ). ⁷³ Test (Ordinal-Q): Correlación de Spearman (ρ). ⁷⁵

Sección 6: Conclusión y Recursos de Aplicación (Datasets)

6.1 Resumen Narrativo

Este informe ha delineado un marco riguroso para la estadística descriptiva, diseñado específicamente para la práctica de la ciencia de datos. El flujo lógico es fundamental: la **taxonomía de las variables** (Sección 2) no es un ejercicio académico, sino el principio que dicta todas las acciones subsecuentes. Dicha taxonomía determina qué **medidas univariantes** (Sección 3) son válidas y robustas, y qué herramientas de **análisis bivariante** (Sección 5) deben seleccionarse de la matriz de relaciones.

Simultáneamente, la **teoría de muestreo** (Sección 4) envuelve todo el proceso, dictando los límites de nuestras conclusiones. La comprensión de que la mayoría de los datos de ciencia de datos son muestras de conveniencia limita la **validez externa** de los modelos, una restricción crítica que debe comunicarse en cualquier análisis profesional.

6.2 Anexo: Datasets Recomendados para la Enseñanza

Para que un curso de estadística descriptiva sea eficaz, la teoría debe anclarse en la práctica. Se recomiendan los siguientes datasets clásicos, disponibles públicamente (ej. en repositorios como Kaggle, UC Irvine Machine Learning Repository), ya que proporcionan ejemplos claros de todas las casuísticas discutidas.⁸⁴

1. Iris Dataset⁸⁵

- **Contexto:** Un dataset pequeño y limpio, ideal para empezar.
- **Ejemplos de Aplicación:**
 - **Q vs. Q (Sección 5.1):** Correlación de Pearson/Spearman entre `sepal_length` y `sepal_width`.
 - **Q vs. C (Sección 5.3):** ANOVA (o *Box Plots*) para comparar `sepal_length` (Q) entre las tres categorías de `species` (C).
 - **MTC (Sección 3.1):** Calcular la media y mediana de `petal_width`.

2. Titanic Dataset⁸⁸

- **Contexto:** Excelente para variables categóricas y datos faltantes.
- **Ejemplos de Aplicación:**
 - **C vs. C (Sección 5.2):** Tabla de contingencia y prueba Chi-Cuadrado para `survived` (C) vs. `gender` (C).
 - **Q vs. C (Sección 5.3):** Prueba t (o *Box Plots*) para comparar `age` (Q) entre

los grupos survived (C).

- **Escala Ordinal (Sección 2.2.2):** Analizar la variable Pclass (clase del pasajero) como ordinal.

3. Ames Housing / Boston Housing Dataset ⁸⁸

- **Contexto:** Datasets del mundo real para predecir precios de vivienda.
- **Ejemplos de Aplicación:**
 - **MTC y Dispersión (Sección 3):** La variable objetivo SalePrice ⁹¹ está fuertemente sesgada. Esto demuestra la necesidad de usar la **Mediana** y el **IQR** sobre la Media y la Desviación Estándar. También es un caso de uso clave para la transformación logarítmica (una casuística de la Sección 2.3.3).
 - **Matriz Bivalente (Sección 5):** Permite practicar todas las combinaciones: Q-Q (SalePrice vs. LotArea), C-C (Neighborhood vs. RoofStyle) y Q-C (SalePrice vs. Neighborhood).

Obras citadas

1. Estadística descriptiva e inferencial en el análisis de datos ..., fecha de acceso: noviembre 11, 2025, <https://www.cognodata.com/blog/estadistica-descriptiva-e-inferencial-analisis-datos/>
2. Fundamentos de estadística y probabilidad para Data Science - RPubS, fecha de acceso: noviembre 11, 2025, https://rpubs.com/jesusturpin/saa_intro_stats_03_1
3. PROBABILIDAD Y ESTADÍSTICA en la Ciencia de Datos - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=dZ9Vm7jqnTk>
4. Tipos de variables y ejemplos - Análisis Estadístico en R, fecha de acceso: noviembre 11, 2025, <https://learn-r-uah.netlify.app/resource/r-datatypes-example/>
5. 4 Tipos de Datos: Nominal, Ordinal, Discreto y Continuo | Great Learning, fecha de acceso: noviembre 11, 2025, <https://www.mygreatlearning.com/blog/tipos-de-datos/>
6. Scales of Measurement and Presentation of Statistical Data - PMC - PubMed Central - NIH, fecha de acceso: noviembre 11, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6206790/>
7. Tipos de datos en estadística: Tipos de datos nominales, ordinales ..., fecha de acceso: noviembre 11, 2025, <https://www.freecodecamp.org/espanol/news/tipos-de-datos-en-estadistica-tipos-de-datos-nominales-ordinales-de-intervalo-y-de-razon-explicados-con-ejemplos/>
8. Levels of Measurement | Nominal, Ordinal, Interval and Ratio - Scribbr, fecha de acceso: noviembre 11, 2025, <https://www.scribbr.com/statistics/levels-of-measurement/>
9. Tipos de variables - IBM, fecha de acceso: noviembre 11, 2025, <https://www.ibm.com/docs/es/spss-statistics/cd?topic=charts-variable-types>

10. Transformación de Variables Nominales y Ordinales en Numéricas en Python - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=KO7Tc0i4UEA>
11. ESCALAS DE MEDICIÓN - Dialnet, fecha de acceso: noviembre 11, 2025, <https://dialnet.unirioja.es/descarga/articulo/4942056.pdf>
12. fecha de acceso: noviembre 11, 2025, <https://mindthegraph.com/blog/es/interval-variables/>
13. 25 Interval Variable Examples (2025) - Helpful Professor, fecha de acceso: noviembre 11, 2025, <https://helpfulprofessor.com/interval-variable-examples/>
14. Types of Data and the Scales of Measurement | UNSW Online, fecha de acceso: noviembre 11, 2025, <https://studyonline.unsw.edu.au/blog/types-of-data>
15. Ratio Scales | Definition, Examples, & Data Analysis - Scribbr, fecha de acceso: noviembre 11, 2025, <https://www.scribbr.com/statistics/ratio-data/>
16. What is Ratio Scale? With 5 Examples - Voiceform, fecha de acceso: noviembre 11, 2025, <https://www.voiceform.com/blog-posts/ratio-scale>
17. Estadística. Tipos de variables. Escalas de medida - Evidencias en pediatría, fecha de acceso: noviembre 11, 2025, <https://evidenciasenpediatria.es/articulo/7307/estadistica-tipos-de-variables-escalas-de-medida>
18. Types of Statistical Variables | Quantitative Qualitative - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=nCszHELuwXk>
19. Levels of Measurement. How many types of variable are there? | by Nick Ward | Medium, fecha de acceso: noviembre 11, 2025, <https://medium.com/@nhward60/levels-of-measurement-7b0f9828e14c>
20. Encoding Cyclical Features for Deep Learning - Kaggle, fecha de acceso: noviembre 11, 2025, <https://www.kaggle.com/code/avanwyk/encoding-cyclical-features-for-deep-learning>
21. CyclicalFeatures — 1.7.0 - Feature-engine, fecha de acceso: noviembre 11, 2025, https://feature-engine.trainindata.com/en/1.7.x/user_guide/creation/CyclicalFeatures.html
22. Three Approaches to Encoding Time Information as Features for ML Models, fecha de acceso: noviembre 11, 2025, <https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-features-for-ml-models/>
23. Logarithmic Scale - GeeksforGeeks, fecha de acceso: noviembre 11, 2025, <https://www.geeksforgeeks.org/maths/logarithmic-scale/>
24. Medidas de tendencia central: Una visión general completa ..., fecha de acceso: noviembre 11, 2025, <https://www.datacamp.com/es/tutorial/central-tendency>
25. ¿Qué son las medidas de tendencia central y para qué sirven? - UNIR, fecha de acceso: noviembre 11, 2025, <https://colombia.unir.net/actualidad-unir/medidas-tendencia-central/>
26. Media, Mediana y Moda | Introduction to Statistics - JMP, fecha de acceso: noviembre 11, 2025, <https://www.jmp.com/es/statistics-knowledge-portal/measures-of-central-tende>

[ncy-and-variability/mean-median-and-mode](#)

27. Media y desviación estándar o mediana y rango intercuartil | Khan Academy en Español, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=iHRgWKYBiPQ>
28. Medidas de tendencia central y variabilidad | Introduction to Statistics - JMP, fecha de acceso: noviembre 11, 2025, <https://www.jmp.com/es/statistics-knowledge-portal/measures-of-central-tendency-and-variability>
29. ¿Qué es la media, la mediana y la moda? - QuestionPro, fecha de acceso: noviembre 11, 2025, <https://www.questionpro.com/blog/es/la-media-la-mediana-y-la-moda/>
30. Media ponderada: ¿Qué es y dónde se utiliza? (+Ejemplos) - Someka, fecha de acceso: noviembre 11, 2025, <https://www.someka.net/es/blog/media-ponderada/>
31. How to calculate the WEIGHTED AVERAGE (Solved Exercise) - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=DhlwGw4VC-s>
32. Media ponderada - Wikipedia, la enciclopedia libre, fecha de acceso: noviembre 11, 2025, https://es.wikipedia.org/wiki/Media_ponderada
33. Tipos de media: simple, ponderada, armónica, cuadrática y geométrica. Ejemplo 1, fecha de acceso: noviembre 11, 2025, https://www.youtube.com/watch?v=oXkq4WPS_Co
34. fecha de acceso: noviembre 11, 2025, <https://cursos.frogamesformacion.com/pages/blog/media-aritmetica-armonica-geometrica#:~:text=La%20media%20arm%C3%B3nica%20se%20calcula,o%20tasas%20de%20inter%C3%A9s%20compuesto.>
35. Media Geométrica y Media Armonica - Yeudy Maldonado - Prezi, fecha de acceso: noviembre 11, 2025, <https://prezi.com/p/gbuwgexmnbw2/media-geometrica-y-media-armonica/>
36. Media armónica - Universo Formulas, fecha de acceso: noviembre 11, 2025, <https://www.universoformulas.com/estadistica/descriptiva/media-armonica/>
37. ¿Cuáles son los casos de uso ideales para las medias geométrica y armónica? - Reddit, fecha de acceso: noviembre 11, 2025, https://www.reddit.com/r/AskStatistics/comments/1kvd89k/what_are_the_ideal_use_cases_for_geometric_and/?tl=es-419
38. Trimmed Mean / Truncated Mean: Definition, Examples - Statistics How To, fecha de acceso: noviembre 11, 2025, <https://www.statisticshowto.com/trimmed-mean/>
39. ¿Qué es la media truncada (acotada o recortada) y cómo calcularla con R? - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=lpH9xbSZ8Rk>
40. Estadística Básica con R: La media truncada o podada y la media winsorizada - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=dSugi3lp2Dc>
41. Medias recortadas (medias truncadas) - qué es, definición y preguntas frecuentes - Ikusmira, fecha de acceso: noviembre 11, 2025, <https://ikusmira.org/p/medias-recortadas-medias-truncadas/>
42. Medidas de Dispersión: Rango, varianza, desviación estándar y coef. de variación,

- fecha de acceso: noviembre 11, 2025,
<https://www.todoestadistica.com.mx/medidas/dispersion.html>
43. Medidas de tendencia central y dispersión - Medwave, fecha de acceso: noviembre 11, 2025, <https://www.medwave.cl/series/MBE04/4934.html>
44. Rango, desviación estándar y varianza - Prepa 8 - UNAM, fecha de acceso: noviembre 11, 2025,
http://prepa8.unam.mx/academia/colegios/matematicas/paginacolmate/applets/matematicas_IV/Applets_Geogebra/desestvar.html
45. Medidas de dispersión: rango, varianza y desviación estándar (video) - Khan Academy, fecha de acceso: noviembre 11, 2025,
<https://es.khanacademy.org/math/probability/data-distributions-a1/summarizing-spread-distributions/v/range-variance-and-standard-deviation-as-measures-of-dispersion>
46. Medidas de escala robustas - Wikipedia, la enciclopedia libre, fecha de acceso: noviembre 11, 2025, https://es.wikipedia.org/wiki/Medidas_de_escala_robustas
47. MAD (Desviación media absoluta), fecha de acceso: noviembre 11, 2025,
https://docs.oracle.com/cloud/help/es/pbcs_common/PFUSU/insights_metrics_MAD.htm
48. MEASURES OF DISPERSION (Range, Mean deviation, Variance, Standard deviation, Coefficient of varia... - YouTube, fecha de acceso: noviembre 11, 2025,
<https://www.youtube.com/watch?v=1sIBTCBG07s>
49. fecha de acceso: noviembre 11, 2025,
<https://www.jmp.com/es/statistics-knowledge-portal/measures-of-central-tendency-and-variability/standard-deviation#:~:text=%C2%BFCu%C3%A1l%20es%20la%20diferencia%20entre%20la%20desviaci%C3%B3n%20est%C3%A1ndar%20y%20el,datos%20en%20una%20escala%20com%C3%BAn.>
50. Desviación estándar | Introduction to Statistics | JMP, fecha de acceso: noviembre 11, 2025,
<https://www.jmp.com/es/statistics-knowledge-portal/measures-of-central-tendency-and-variability/standard-deviation#:~:text=%C2%BFCu%C3%A1l%20es%20la%20diferencia%20entre%20la%20desviaci%C3%B3n%20est%C3%A1ndar%20y%20el,datos%20en%20una%20escala%20com%C3%BAm.>
51. Tipos de muestreo - Repositorio CENTROGEO, fecha de acceso: noviembre 11, 2025,
<https://centrogeo.repositorioinstitucional.mx/jspui/bitstream/1012/163/1/19-Tipos%20de%20Muestreo%20-%20%20Diplomado%20en%20An%C3%A1lisis%20de%20Informaci%C3%B3n%20Geoespacial.pdf>
52. Muestreo no probabilístico: definición, tipos y ejemplos - QuestionPro, fecha de acceso: noviembre 11, 2025,
<https://www.questionpro.com/blog/es/muestreo-no-probabilistico/>
53. Aproximación a los distintos tipos de muestreo no probabilístico que ..., fecha de acceso: noviembre 11, 2025,
http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0864-21252021000300002
54. Muestreo estadístico, fecha de acceso: noviembre 11, 2025,

- <https://estudiosestadisticos.ucm.es/muestreo-estadistico>
55. Machine Learning with Imbalanced data - Kaggle, fecha de acceso: noviembre 11, 2025,
<https://www.kaggle.com/code/liamarguedas/machine-learning-with-imbalanced-data>
 56. Comparison of machine learning techniques to handle imbalanced COVID-19 CBC datasets, fecha de acceso: noviembre 11, 2025,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8372002/>
 57. How do I handle imbalanced datasets in classification problems? - Milvus, fecha de acceso: noviembre 11, 2025,
<https://milvus.io/ai-quick-reference/how-do-i-handle-imbalanced-datasets-in-classification-problems>
 58. Muestreo por Conveniencia: Qué es, ejemplos y características - Netquest, fecha de acceso: noviembre 11, 2025,
<https://www.netquest.com/blog/muestreo-por-conveniencia>
 59. Convenience sampling method: How and when to use - Qualtrics, fecha de acceso: noviembre 11, 2025,
<https://www.qualtrics.com/en-au/experience-management/research/convenience-sampling/>
 60. Muestreo por cuotas: método y ventajas | Qualtrics, fecha de acceso: noviembre 11, 2025,
<https://www.qualtrics.com/es-es/gestion-de-la-experiencia/investigacion/muestreo-por-cuotas/>
 61. Muestreo de bola de nieve: Qué es, ventajas y cómo realizarlo. - QuestionPro, fecha de acceso: noviembre 11, 2025,
<https://www.questionpro.com/blog/es/muestreo-de-bola-de-nieve/>
 62. What Is Convenience Sampling? | Definition & Examples - Scribbr, fecha de acceso: noviembre 11, 2025,
<https://www.scribbr.com/methodology/convenience-sampling/>
 63. Convenience Sampling: Definition, Method and Examples - Simply Psychology, fecha de acceso: noviembre 11, 2025,
<https://www.simplypsychology.org/convenience-sampling.html>
 64. Sampling bias - Wikipedia, fecha de acceso: noviembre 11, 2025,
https://en.wikipedia.org/wiki/Sampling_bias
 65. More than Just Convenient: The Scientific Merits of Homogeneous Convenience Samples - PMC - NIH, fecha de acceso: noviembre 11, 2025,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC5606225/>
 66. The Inconvenient Truth About Convenience and Purposive Samples - PMC - NIH, fecha de acceso: noviembre 11, 2025,
<https://pmc.ncbi.nlm.nih.gov/articles/PMC8295573/>
 67. What is A/B Testing in Data Science? - Caltech Bootcamps, fecha de acceso: noviembre 11, 2025,
<https://pg-p.ctme.caltech.edu/blog/data-science/what-is-a-b-testing>
 68. A/B Testing: A Complete Guide to Statistical Testing | by Francesco Casalegno - Medium, fecha de acceso: noviembre 11, 2025,

- <https://medium.com/data-science/a-b-testing-a-complete-guide-to-statistical-testing-e3f1db140499>
69. Técnicas estadísticas para identificar posibles relaciones bivariadas - SciELO Cuba, fecha de acceso: noviembre 11, 2025, http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1726-67182020000200008
 70. Análisis Bivariante: Guía y Técnicas | PDF - Scribd, fecha de acceso: noviembre 11, 2025, <https://es.scribd.com/document/618505351/Bivariantes>
 71. MEDB - GEOPY, fecha de acceso: noviembre 11, 2025, <https://web.bioucm.es/cont/geopy/medb/medb.html>
 72. Descriptiva bivalente - YouTube, fecha de acceso: noviembre 11, 2025, <https://www.youtube.com/watch?v=gpcFii1jFX4>
 73. fecha de acceso: noviembre 11, 2025, <https://support.minitab.com/es-mx/minitab/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/#:~:text=Por%20ejemplo%2C%20usted%20puede%20usar.de%20las%20capas%20de%20chocolate.&text=La%20correlaci%C3%B3n%20de%20Spearman%20eval%C3%BAa,dos%20variables%20continuas%20u%20ordinales>
 74. Una comparación de los métodos de correlación de Pearson y ..., fecha de acceso: noviembre 11, 2025, <https://support.minitab.com/es-mx/minitab/help-and-how-to/statistics/basic-statistics/supporting-topics/correlation-and-covariance/a-comparison-of-the-pearson-and-spearman-correlation-methods/>
 75. ¿Cómo decides qué coeficiente de correlación (Pearson o Spearman) es mejor usar para un conjunto de datos específico? : r/statistics - Reddit, fecha de acceso: noviembre 11, 2025, https://www.reddit.com/r/statistics/comments/76iw0w/how_do_you_decide_which_correlation_coefficient/?tl=es-419
 76. Análisis bivalente, fecha de acceso: noviembre 11, 2025, <https://openaccess.uoc.edu/bitstream/10609/148455/1/AnalisisBivariante.pdf>
 77. Diapositiva 1, fecha de acceso: noviembre 11, 2025, <https://www.uv.es/mperea/T5.ppt>
 78. Estadística descriptiva bivalente - YouTube, fecha de acceso: noviembre 11, 2025, https://www.youtube.com/watch?v=RegdZcnmr_I
 79. TEMA 3: RELACIÓN ENTRE UNA VARIABLE CUALITATIVA Y ..., fecha de acceso: noviembre 11, 2025, https://personales.unican.es/rasillad/docencia/G14/TEMA_3/relacion_entre_una_variable_cualitativa_otra_cuantitativa.html
 80. Sesión 5. Análisis de varianza (ANOVA) y correlación - Rodrigo Fernández Caba, fecha de acceso: noviembre 11, 2025, <https://rodrigofcaba.github.io/posts/Sesi%C3%B3n-V-An%C3%A1lisis-de-Varianza-y-Correlaci%C3%B3n/>
 81. Correlación punto-biserial - Calculadora estadística, fecha de acceso: noviembre 11, 2025, <https://numiqo.es/tutorial/point-biserial-correlation>

82. Correlación biserial puntual - Psicometría con R - Bosco Mendoza, fecha de acceso: noviembre 11, 2025,
<https://boscomendoza.com/correlacion-biserial-puntual-psicometria-con-r/>
83. [Q]Diferencia entre la correlación de Pearson y la correlación biserial puntual. - Reddit, fecha de acceso: noviembre 11, 2025,
https://www.reddit.com/r/statistics/comments/pzcfsp/qdifference_between_pearson_correlation_and_point/?tl=es-es
84. Datasets for Teaching and Learning - NC State University Libraries, fecha de acceso: noviembre 11, 2025,
<https://www.lib.ncsu.edu/formats/teaching-and-learning-datasets>
85. Datasets - UCI Machine Learning Repository, fecha de acceso: noviembre 11, 2025, <https://archive.ics.uci.edu/ml/datasets>
86. List of datasets for machine-learning research - Wikipedia, fecha de acceso: noviembre 11, 2025,
https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research
87. Find Open Datasets and Machine Learning Projects - Kaggle, fecha de acceso: noviembre 11, 2025, <https://www.kaggle.com/datasets>
88. Exploratory Data Analysis with Python Code Examples | by Sercan Gul, fecha de acceso: noviembre 11, 2025,
<https://sercangl.medium.com/exploratory-data-analysis-with-python-code-examples-3445fdeef702>
89. 5 Free Datasets to Kickstart Your Machine Learning Projects Today - MachineLearningMastery.com, fecha de acceso: noviembre 11, 2025,
<https://machinelearningmastery.com/5-free-datasets-to-kickstart-your-machine-learning-projects-today/>
90. Top Free Dataset Resources for Data Science Projects - GeeksforGeeks, fecha de acceso: noviembre 11, 2025,
<https://www.geeksforgeeks.org/data-science/top-free-dataset-resources-for-data-science-projects/>
91. Ames Housing Dataset - Kaggle, fecha de acceso: noviembre 11, 2025,
<https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset>