Lectura fundamental

Estadística y probabilidad para ciencia de datos

Contenido

Palabras clave: probabilidad, variable aleatoria, distribución, inferencia, regresión, correlación, error, evaluación, predicción, modelo.

1. Introducción

En análisis de datos, muchas decisiones relevantes no se toman con certeza absoluta, sino bajo condiciones de variabilidad, incertidumbre y evidencia parcial. Una compra futura, la respuesta de un cliente, la demanda de un servicio o la probabilidad de incumplimiento no se conocen con exactitud anticipada; se estiman. Por eso, comprender la estadística no es un lujo teórico, sino una base para convertir datos dispersos en argumentos verificables y decisiones razonadas.

Quien trabaja con datos necesita distinguir entre una descripción superficial y una conclusión sustentada. No basta con observar una tabla o un gráfico; hace falta interpretar comportamientos, reconocer patrones, estimar márgenes de error y valorar qué tan sólido es un resultado. En ese tránsito, la probabilidad permite modelar incertidumbre, las distribuciones ofrecen marcos para entender comportamientos frecuentes y la inferencia posibilita pasar de una muestra a una afirmación razonable sobre una población.

En este recorrido también cobra especial importancia la relación entre variables. Muchas preguntas del campo profesional se formulan en esos términos: ¿a mayor inversión publicitaria aumentan las ventas?, ¿una variable explica otra?, ¿el patrón observado es estable o solo aparente?, ¿qué tanto error comete un modelo? Estas preguntas conectan con técnicas de regresión y con métricas que permiten valorar si una predicción o una explicación son suficientemente útiles para un caso aplicado.

2. Probabilidad y variables aleatorias

En ciencia de datos, uno de los retos más importantes es trabajar con situaciones donde no existe certeza absoluta. Por ejemplo, aunque se disponga de grandes volúmenes de datos, no es posible predecir con total seguridad si un cliente realizará una compra, si un usuario hará clic en un anuncio o cuántas ventas se generarán en un día determinado. Estas situaciones están marcadas por la **incertidumbre**, y es precisamente allí donde la probabilidad juega un papel fundamental.

La **probabilidad** es una herramienta matemática que permite cuantificar la incertidumbre. En términos simples, asigna un valor entre 0 y 1 a la posibilidad de que ocurra un evento, donde 0 indica que el evento es imposible y 1 que es completamente seguro. Este enfoque permite transformar situaciones inciertas en modelos analizables, lo cual es esencial en ciencia de datos. Desde esta perspectiva, la probabilidad no se limita a expresar una intuición sobre lo posible, sino que organiza la incertidumbre en términos cuantificables y comparables.

Para comprender mejor este concepto, es útil partir de la idea de un **experimento aleatorio**, es decir, un proceso cuyo resultado no puede predecirse con certeza antes de observarlo. Un ejemplo clásico es lanzar una moneda, pero en el contexto de la ciencia de datos, un experimento aleatorio puede ser el comportamiento de un usuario en una página *web*.

El conjunto de todos los posibles resultados de un experimento se denomina **espacio muestral**. Por ejemplo, si se analiza si un usuario hace clic en un anuncio, el espacio muestral podría ser: {clic, no clic}. A partir de este conjunto, se definen los **eventos**, que son subconjuntos del espacio muestral. En este caso, el evento de interés podría ser "el usuario hace clic".

2.1. Métodos de asignación de probabilidades

Una vez definidos los eventos, surge una pregunta fundamental: ¿cómo se asignan probabilidades en la práctica? Existen tres enfoques principales para hacerlo: el método clásico, el frecuentista y el subjetivo. Cada uno es útil en diferentes contextos dentro de la ciencia de datos.

2.1.1. Método clásico

El método clásico se basa en la idea de que todos los resultados posibles de un experimento son igualmente probables. En este caso, la probabilidad de un evento se calcula como la proporción entre los resultados favorables y el total de resultados posibles (Anderson, Sweeney y Williams, 2008). Por ejemplo, al lanzar un dado justo, la probabilidad de obtener un número par es:

$$P(par) = \frac{3}{6} = 0.5$$

Este método es útil cuando se conocen claramente todos los posibles resultados y estos tienen la misma probabilidad de ocurrir. Sin embargo, en ciencia de datos su aplicación es limitada, ya que muchos problemas reales no cumplen esta condición

2.1.2. Método frecuentista

Una de las formas más comunes de interpretar la probabilidad en ciencia de datos es desde el **enfoque frecuentista**. Según esta perspectiva, la probabilidad de un evento se estima a partir de su **frecuencia relativa** en un gran número de observaciones.

Por ejemplo, si de 100 usuarios, 30 hacen clic en un anuncio, se puede estimar que la probabilidad de clic es aproximadamente 0.30. Esta interpretación es especialmente útil porque conecta directamente con el análisis de datos reales.

Cuando se trabaja con datos observados de manera repetida, la frecuencia relativa ofrece una base empírica sólida para estimar probabilidades. Anderson *et al.*, (2008) señalan que este método resulta conveniente cuando existen datos suficientes que permiten estimar la proporción de veces que un resultado aparecerá al repetir muchas veces un experimento.

Este enfoque es ampliamente utilizado en ciencia de datos, ya que gran parte del análisis se basa en datos históricos, como registros de clientes, transacciones o comportamientos digitales.

2.1.3. Método subjetivo

No obstante, en muchos contextos no se dispone de datos suficientes o los fenómenos no son fácilmente repetibles. En estos casos, se recurre al método subjetivo, en el cual la probabilidad se asigna con base en la experiencia, el conocimiento previo y el juicio experto.

Por ejemplo, un analista puede estimar la probabilidad de éxito de un nuevo producto basándose en estudios de mercado, experiencia previa y tendencias del sector.

Aunque este enfoque incorpora cierto grado de subjetividad, sigue respetando los principios fundamentales de la probabilidad: los valores deben estar entre 0 y 1 y la suma de las probabilidades de todos los resultados posibles debe ser igual a 1 (Anderson *et al.*, 2008).

Esta distinción entre métodos es especialmente relevante en ciencia de datos, donde con frecuencia se combinan diferentes fuentes de información. Por ejemplo:

- Datos históricos → enfoque frecuentista
- Supuestos del modelo → enfoque clásico
- Juicio experto → enfoque subjetivo

En la práctica, el análisis de datos suele integrar estos enfoques para construir modelos más robustos y realistas.

2.1.4. Relación entre eventos

A medida que se profundiza en el análisis probabilístico, surge la necesidad de estudiar cómo se relacionan diferentes eventos. Por ejemplo, puede ser relevante conocer la probabilidad de que un usuario visite una página y, además, realice una compra. Para abordar este tipo de situaciones se utilizan reglas fundamentales como la regla de la suma y la regla del producto, que permiten combinar probabilidades de distintos eventos.

Sin embargo, uno de los conceptos más importantes en aplicaciones reales es la probabilidad condicional, que permite responder preguntas del tipo: "¿cuál es la probabilidad de que ocurra un evento dado que ya ocurrió otro?". Esta se define como:

$$P(A \mid B) = \frac{p(A \cap B}{P(B)}$$

Este concepto es especialmente relevante en ciencia de datos, ya que permite refinar el análisis en función de información adicional. Por ejemplo, no es lo mismo estimar la probabilidad de que un usuario realice una compra, que calcular dicha probabilidad sabiendo que el usuario ya visitó el sitio *web*.

Relacionado con lo anterior, se encuentra el concepto de **independencia de eventos**. Dos eventos son independientes si la ocurrencia de uno no afecta la probabilidad del otro, es decir, si $P(A \mid B) = P(A)$ (Anderson *et al.*, 2008).

2.2. Variables aleatorias

Una vez definido el experimento aleatorio, el análisis avanza hacia la variable aleatoria, es decir, hacia la representación numérica de los resultados posibles. Ese paso es fundamental porque convierte una situación incierta en un objeto susceptible de descripción matemática y análisis estadístico.

Las variables son características, propiedades o atributos que pueden tomar diferentes valores en un estudio o experimento. En términos prácticos, una **variable aleatoria** puede representar el número de compras realizadas en una hora, la cantidad de errores en una base de datos, el tiempo que tarda una consulta en responder o el monto de una transacción.

Una forma útil de diferenciar los tipos de variables aleatorias, como señalan Anderson *et al.* (2008), consiste en observar los valores que pueden asumir: si estos son separados o contables, se trata de una **variable discreta**; si pueden tomar cualquier valor dentro de un intervalo, se trata de una **variable continua**. Esta distinción no es menor, ya que condiciona el tipo de distribución, la forma de calcular probabilidades y la interpretación de los resultados.

Para describir el comportamiento de una variable aleatoria se utilizan funciones de probabilidad. En el caso particular de las **variables discretas**, que aparecen cuando se cuentan ocurrencias (como el número de usuarios que hacen clic en un anuncio, los *tickets* generados por un sistema o la cantidad de pacientes que llegan a un servicio en un periodo dado), se emplea la **función de masa de probabilidad** $f(x)$. Esta función asigna una probabilidad específica a cada valor posible de la variable, lo que permite, por ejemplo, construir tablas que indiquen la probabilidad de que un cliente compre 0, 1, 2 o más productos.

En el caso de las **variables continuas**, que describen magnitudes que se miden y no solo se cuentan (como el tiempo de permanencia de un usuario en una aplicación, la latencia de una consulta o el valor monetario exacto de una transacción), se utiliza la **función de densidad de probabilidad**. A diferencia de las variables discretas, en este contexto no se asignan probabilidades a valores puntuales, sino a intervalos de valores (Anderson *et al.*, 2008). Esta diferencia es fundamental, ya que implica que la probabilidad de que una variable continua tome exactamente un valor específico es igual a cero. Esto ocurre porque la función de densidad de probabilidad no proporciona directamente probabilidades para valores individuales, sino que estas se obtienen como áreas bajo la curva en un intervalo determinado. En la práctica, esto obliga a formular adecuadamente las preguntas analíticas: en lugar de preguntar "¿cuál es la probabilidad de que el tiempo sea exactamente 5 minutos?", resulta más apropiado plantear "¿cuál es la probabilidad de que el tiempo esté entre 4 y 6 minutos?". De esta manera, se ajusta el análisis a la naturaleza continua de la variable y se garantiza una interpretación correcta de los resultados.

2.3. Valor esperado y variabilidad

Una vez definida una variable aleatoria, es necesario resumir su comportamiento. Una de las medidas más importantes para ello es el **valor esperado** o media, que representa el resultado promedio que se esperaría en el largo plazo.

En estadística, el valor esperado o media de una variable aleatoria resume su localización central y representa, en términos sencillos, el resultado promedio que cabría anticipar en el largo plazo.

Anderson *et al.* (2008) lo definen como un promedio ponderado de los valores posibles de la variable, donde los pesos son precisamente las probabilidades asociadas a cada valor. Esta medida permite condensar la información de toda una distribución en un solo valor, facilitando su interpretación. Por ejemplo, si una variable representa las ventas diarias, el valor esperado permite estimar el número promedio de ventas, incluso si ese valor no coincide exactamente con ningún día específico.

Sin embargo, el valor esperado por sí solo no es suficiente para comprender completamente el comportamiento de una variable. Dos variables pueden tener la misma media, pero presentar niveles de variabilidad muy diferentes. Por esta razón, es necesario considerar medidas como la **varianza** y la **desviación estándar**, que indican qué tan dispersos están los valores respecto a la media.

La varianza puede interpretarse como un promedio ponderado de las desviaciones cuadráticas respecto a la media, utilizando nuevamente las probabilidades como pesos. Esta interpretación es especialmente relevante en ciencia de datos, ya que evita conclusiones simplistas. Un valor esperado atractivo no garantiza estabilidad: si la dispersión es alta, el fenómeno seguirá siendo impredecible y requerirá mayor cautela en la toma de decisiones.

3. Distribuciones fundamentales en Ciencia de datos

En la sección anterior se introdujeron las variables aleatorias como una forma de representar fenómenos inciertos mediante valores numéricos. Sin embargo, conocer únicamente los valores posibles que puede tomar una variable no es suficiente para comprender su comportamiento. Es necesario entender cómo se distribuyen esos valores, es decir, qué tan frecuentes son, cuáles son más probables y cuáles son poco comunes. Esta necesidad da origen al concepto de distribución de probabilidad.

Una **distribución de probabilidad** describe cómo se reparten los posibles valores de una variable aleatoria y con qué probabilidad (en el caso discreto) o densidad (en el caso continuo) se asocian. Esta idea es central porque, en lugar de estudiar datos como observaciones aisladas, permite identificar el patrón general de comportamiento de un fenómeno. En términos prácticos, una distribución actúa como un mapa que resume lo esperable, lo raro y lo extremo dentro de los datos.

Comprender las distribuciones es fundamental en ciencia de datos porque permite:

- Interpretar correctamente los datos observados
- Seleccionar modelos adecuados
- Identificar comportamientos atípicos
- Sustentar inferencias y predicciones

3.1. Distribuciones de probabilidad discretas

Las distribuciones discretas se utilizan cuando la variable aleatoria toma valores contables. Entre ellas, algunas tienen una relevancia especial en aplicaciones de ciencia de datos.

3.1.1. Distribución binomial

La **distribución binomial** es especialmente útil cuando un proceso puede describirse como una secuencia de ensayos independientes, cada uno con dos resultados posibles: éxito o fracaso. Este tipo de estructura aparece con frecuencia en problemas reales, como:

- Un usuario abre o no un correo

- Una transacción es aprobada o rechazada
- Un cliente responde o no a una oferta

Cuando se cumplen las condiciones del modelo (número fijo de ensayos, independencia y probabilidad constante de éxito), la distribución binomial permite calcular la probabilidad de obtener un determinado número de éxitos (Anderson *et al.*, 2008).

Por ejemplo, si se envía una campaña de correo a 100 usuarios y la probabilidad de apertura es del 30%, la distribución binomial permite estimar la probabilidad de que exactamente 40 usuarios abran el mensaje. Este tipo de análisis resulta clave en *marketing* digital y experimentación o *A/B testing*.

3.1.2. Distribución de Poisson

Otra distribución discreta de gran utilidad es la **distribución de Poisson**, que se emplea para modelar el número de veces que ocurre un evento en un intervalo de tiempo o espacio determinado.

Se aplica en situaciones como:

- Número de clientes que llegan a una tienda por hora
- Cantidad de fallas en un sistema
- Número de mensajes recibidos por minuto
- Incidentes en una red

Su valor práctico radica en que permite modelar procesos donde los eventos ocurren de manera relativamente independiente y dispersa en el tiempo o el espacio (Anderson *et al.*, 2008).

Por ejemplo, si un sistema recibe en promedio 5 solicitudes por minuto, la distribución de Poisson permite calcular la probabilidad de recibir exactamente 8 solicitudes en un minuto dado. Este tipo de modelado es común en sistemas de colas, monitoreo de servicios y análisis de operaciones.

3.2. Distribuciones continuas

Cuando las variables representan magnitudes medibles, como tiempos, distancias o valores monetarios, se utilizan **distribuciones continuas**. Entre ellas, la más importante en estadística y ciencia de datos es la distribución normal.

3.2.1. Distribución normal

La **distribución normal** (también conocida como distribución gaussiana) ocupa un lugar central debido a su amplia presencia en fenómenos reales y su importancia teórica. Se caracteriza por su forma de campana, donde los valores se concentran alrededor de la media y disminuyen progresivamente hacia los extremos.

Muchos fenómenos en la naturaleza y en sistemas sociales tienden a seguir aproximadamente una distribución normal, como:

- Altura de las personas
- Errores de medición
- Tiempos de respuesta en ciertos procesos

Una herramienta clave para interpretar esta distribución es la **regla empírica**, que indica que aproximadamente:

- El 68% de los valores se encuentra dentro de una desviación estándar de la media
- El 95% dentro de dos desviaciones estándar
- El 99.7% dentro de tres desviaciones estándar (Anderson et al., 2008)

Esta regla permite entender rápidamente qué valores son comunes y cuáles pueden considerarse atípicos.

En la práctica, las distribuciones de probabilidad permiten pasar de una colección de datos a una **representación estructurada del fenómeno**. Esto tiene múltiples aplicaciones:

- Modelar el comportamiento de usuarios
- Estimar probabilidades de eventos futuros
- Evaluar riesgos
- Detectar patrones y anomalías

Por ejemplo:

- Una distribución binomial puede modelar conversiones en una campaña
- Una distribución de Poisson puede describir llegadas de clientes
- Una distribución normal puede representar errores o variabilidad natural

Cada distribución ofrece una forma distinta de entender los datos, y la elección adecuada depende del tipo de variable y del fenómeno analizado.

Tabla 1. Distribuciones y usos frecuentes en análisis de datos

Distribución	Tipo de variable	Uso frecuente	Ejemplo aplicado
Binomial	Discreta	Conteo de éxitos en un número fijo de ensayos	Usuarios que hacen clic en una campaña
Poisson	Discreta	Conteo de eventos en un intervalo	Errores por hora en un sistema
Normal	Continua	Modelado de mediciones y base inferencial	Tiempo de respuesta de un servicio
Muestral de la media	Continua	Inferencia sobre promedios	Promedio de compra por cliente

Fuente: elaboración propia

Una ventaja pedagógica de estudiar distribuciones es que obligan a pensar la naturaleza del fenómeno antes de calcular. No toda variable se comporta como normal y no todo conteo debe modelarse con Poisson. Un uso mecánico de fórmulas conduce a errores de interpretación. En cambio, un análisis bien fundamentado comienza preguntando qué representa la variable, cómo se genera y qué supuestos resultan razonables en el contexto del problema.

Cómo mejorar...

Al enfrentarse a un conjunto de datos, describa primero la variable en lenguaje común y luego determine si está contando eventos o midiendo magnitudes. Esa distinción simple mejora la selección de la distribución y evita errores de modelado desde el inicio.

4. Inferencia estadística básica

En el análisis de datos rara vez se dispone de toda la información de interés. En la práctica, casi nunca es posible observar a toda la población, ya sea por limitaciones de tiempo, costo o acceso. Por ello, lo habitual es trabajar con una **muestra** y, a partir de ella, construir conclusiones sobre un conjunto mayor.

Este paso (de lo observado a lo no observado) no es trivial. Implica asumir cierto nivel de incertidumbre y requiere un enfoque riguroso que permita controlar el error. La **inferencia estadística** proporciona precisamente los métodos para realizar este proceso con criterios explícitos y fundamentos cuantitativos, permitiendo generalizar resultados de manera responsable (Diez, Cetinkaya-Rundel, & Barr, 2019).

4.1. Población, muestra y parámetros

Para entender la inferencia, es fundamental distinguir entre algunos conceptos clave. La **población** corresponde al conjunto total de elementos de interés, mientras que la **muestra** es un subconjunto de dicha población que se selecciona para su análisis.

A partir de la muestra se calculan medidas conocidas como **estadísticos** o **estimadores**, como el promedio o la proporción. Estos estadísticos se utilizan para aproximar características desconocidas de la población, llamadas **parámetros** (Anderson *et al.*, 2008).

Por ejemplo, el promedio de gasto calculado a partir de una muestra de clientes es un estadístico, mientras que el verdadero promedio de gasto de todos los clientes es un parámetro. El objetivo central de la inferencia estadística es estimar estos parámetros y evaluar qué tan confiables son dichas estimaciones.

4.2. Estimación

Uno de los primeros pasos en inferencia es la **estimación**, que consiste en aproximar el valor de un parámetro poblacional utilizando la información de la muestra.

4.2.1. Estimación puntual

La forma más sencilla es la **estimación puntual**, que utiliza un único valor para representar el parámetro. Por ejemplo, la media muestral puede emplearse como estimación de la media poblacional, o una proporción observada puede aproximar la proporción real.

Tabla 2. Parámetros y estimadores comunes

Parámetro poblacional		Estadístico o estimador puntual	
Descripción	**Fórmula**	**Descripción**	**Fórmula**
$\boldsymbol{\mu}$: media poblacional	$\mu = \frac{\sum x_i}{N}$	$\bar{x}$: media muestral	$\bar{x} = \frac{\sum x_i}{n-1}$
	$\boldsymbol{N}$: tamaño de la población		$\boldsymbol{n}$: tamaño de la muestra
$\boldsymbol{\sigma}$: desviación estándar poblacional	$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$	$\boldsymbol{s}$: desviación estándar muestral	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$
$\boldsymbol{\pi}$: proporción poblacional	$\pi = \frac{X}{N}$	$\bar{p}$: proporción muestral	$\bar{p} = \frac{x}{n}$

Fuente: elaboración propia

No obstante, esta aproximación tiene una limitación importante: **no informa sobre la incertidumbre asociada al valor estimado**. Un promedio por sí solo puede transmitir una falsa sensación de precisión, especialmente cuando se desconoce la variabilidad de los datos o el tamaño de la muestra.

4.2.2. Estimación por intervalo

Para abordar esta limitación, se utiliza la **estimación por intervalo**, que proporciona un rango de valores plausibles para el parámetro. Un **intervalo de confianza** no pretende ofrecer una certeza absoluta, sino expresar explícitamente la incertidumbre del proceso.

Por ejemplo, afirmar que el gasto promedio de los clientes se encuentra entre 45 y 55 unidades con un 95% de confianza es más informativo que reportar simplemente un valor puntual. Este enfoque obliga a interpretar los resultados con mayor prudencia y evita conclusiones apresuradas.

Para comprender mejor la estimación por intervalo, es útil tener en cuenta los siguientes conceptos:

- **Nivel de confianza:** el parámetro poblacional se da entre un rango de valores asociados con una probabilidad llamada Nivel de confianza (*1- α; coeficiente de confianza*). Indica la probabilidad de que el intervalo calculado contenga el valor verdadero del parámetro poblacional si se repitiera el muestreo muchas veces.
- **Intervalo de confianza:** rango de valores que probablemente contiene el valor verdadero de un parámetro de la población.
- **Precisión y margen de error:** el intervalo tiene un margen de error que depende del tamaño de la muestra, la variabilidad de los datos y el nivel de confianza elegido. Un intervalo más amplio aumenta la probabilidad de incluir el valor verdadero, pero es menos preciso.

Teniendo en cuenta lo anterior, para hacer estimaciones utilizando intervalos de confianza, se sigue la siguiente estructura:

$$Estimación\ puntual \pm margen\ de\ error$$

Así, para hacer la estimación por intervalo de la media poblacional a partir de una muestra de tamaño superior a 30, se utiliza la siguiente ecuación:

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Donde $\frac{s}{\sqrt{n}}$ representa el error estándar de la media.

$\bar{x}$:media muestral

s:desviación estándar

n:tamaño de la muestra

$z_{\frac{\alpha}{2}}$:valor crítico de la normal, que depende del nivel de confianza elegido. Por ejemplo, para un nivel de confianza del 90%, $z_{\frac{\alpha}{2}} = 1.645$.

Desde una perspectiva práctica, los intervalos de confianza permiten evaluar la precisión de una estimación: intervalos muy amplios sugieren alta incertidumbre o necesidad de más datos, mientras que intervalos más estrechos indican mayor precisión (Diez *et al.*, 2019).

4.3. Prueba de hipótesis

Además de estimar parámetros, en ciencia de datos es frecuente evaluar afirmaciones o contrastar si ciertos cambios han producido efectos. Para ello se utilizan las **pruebas de hipótesis**, que constituyen un marco formal para la toma de decisiones bajo incertidumbre.

Una prueba de hipótesis comienza con una afirmación tentativa sobre un parámetro poblacional. Como señalan Anderson *et al.* (2008), "a esta suposición tentativa se le llama **hipótesis nula** y se denota por H_0" (p. 369). Esta hipótesis representa una situación de referencia, como la ausencia de efecto o cambio.

Por su parte, la **hipótesis alternativa (H_1)** expresa la afirmación que compite con la hipótesis nula y orienta la decisión estadística. Por ejemplo:

- H_0: una nueva estrategia no cambia la tasa de conversión
- H_1: la estrategia sí cambia la tasa de conversión

A partir de los datos de la muestra, se evalúa si existe suficiente evidencia para rechazar la hipótesis nula.

4.3.1. El valor p y la toma de decisiones

El proceso de decisión se apoya en el **valor p (p-value)**, que mide la probabilidad de obtener resultados tan extremos como los observados, suponiendo que la hipótesis nula es verdadera.

Un valor p pequeño sugiere que los datos observados serían poco probables si la hipótesis nula fuera cierta, lo que constituye evidencia en su contra (Anderson *et al.*, 2008). En la práctica, este valor se compara con un nivel de significancia: α (por ejemplo, $\alpha = 0.05$) para decidir si se rechaza o no la hipótesis nula, cuando el valor $p < \alpha$, se rechaza la hipótesis nula.

Es importante enfatizar que esta decisión **no representa una verdad absoluta**, sino una conclusión condicionada por la muestra, el modelo utilizado y el nivel de significancia. La inferencia estadística no elimina la incertidumbre, sino que la hace explícita y gestionable.

4.3.2. Errores en la toma de decisiones

Dado que las decisiones se basan en información parcial, siempre existe la posibilidad de error. En este contexto, se distinguen dos tipos principales:

- **Error tipo I**: rechazar una hipótesis nula que en realidad es verdadera
- **Error tipo II**: no rechazar una hipótesis nula que es falsa

Esta distinción tiene implicaciones prácticas importantes. Por ejemplo, en un sistema de detección de fraude, clasificar una transacción legítima como fraudulenta (error tipo I) no tiene el mismo impacto que permitir una transacción fraudulenta (error tipo II). Por ello, la inferencia estadística no solo implica cálculos, sino también **juicio contextual**.

4.3.3. Nivel de significancia y error tipo I

En el contexto de las pruebas de hipótesis, el **nivel de significancia** y el **error tipo I** están directamente relacionados. De hecho, el nivel de significancia se define precisamente en términos de la probabilidad de cometer este tipo de error.

Recordemos que el **error tipo I** ocurre cuando se **rechaza la hipótesis nula (H_0) siendo en realidad verdadera**. Es decir, se concluye que existe un efecto, diferencia o cambio, cuando en realidad no lo hay.

Para controlar este riesgo, se introduce el **nivel de significancia**, que se denota generalmente como α. Este valor representa la **probabilidad máxima que se está dispuesto a aceptar de cometer un error tipo I**.

Si se trabaja con un nivel de significancia de $\alpha = 0.05$ esto significa que el analista acepta un **5% de riesgo** de rechazar incorrectamente una hipótesis nula verdadera. Dicho de en otras palabras, de cada 100 decisiones tomadas bajo estas condiciones, aproximadamente 5 podrían ser errores tipo I.

4.3.4. Inferencia en el contexto de ciencia de datos

La inferencia estadística adquiere especial valor cuando se integra con la toma de decisiones en contextos reales, como negocios, tecnología o gestión. En estos escenarios, no basta con obtener resultados numéricos; es necesario interpretarlos en función de su impacto práctico.

Por ejemplo:

- Un intervalo de confianza muy amplio puede indicar la necesidad de recolectar más datos
- Una hipótesis nula no rechazada puede sugerir que un cambio no tuvo efecto detectable
- Un valor p pequeño puede indicar una diferencia estadísticamente significativa, pero no necesariamente relevante en términos prácticos

Este último punto es especialmente importante: en ciencia de datos, **la significancia estadística no siempre implica relevancia práctica**. Por ello, el análisis debe complementarse con criterio y conocimiento del contexto.

Para reflexionar...

¿En cuántas ocasiones una decisión basada en datos se presenta como concluyente, cuando en realidad debería comunicarse como una conclusión provisional condicionada por la muestra, el método y el margen de error?

5. Relación entre variables y regresión simple

En el análisis de datos, una de las preguntas más frecuentes no es solo cómo se comporta una variable de manera individual, sino **cómo se relaciona con otras variables**. En muchos contextos, el interés está en comprender si dos cantidades tienden a cambiar juntas, en qué dirección lo hacen y con qué intensidad. Este tipo de análisis permite no solo describir fenómenos, sino también formular preguntas explicativas y construir modelos predictivos.

Por ejemplo, una organización puede preguntarse si existe relación entre el gasto en publicidad y las ventas, entre el tiempo de respuesta de un sistema y la satisfacción del usuario, o entre el número de visitas a una página y la probabilidad de compra. Identificar estas relaciones es un paso clave en la generación de conocimiento a partir de los datos.

Sin embargo, es importante establecer desde el inicio una distinción fundamental: **reconocer una asociación entre variables no implica necesariamente una relación causal**.

Esta advertencia es esencial en ciencia de datos, ya que evita interpretaciones erróneas y decisiones mal fundamentadas. Durante el análisis se busca saber si dos cantidades tienden a cambiar juntas, en qué dirección lo hacen y con qué intensidad. Esa exploración permite formular preguntas explicativas y predictivas. Sin embargo, reconocer una asociación no equivale automáticamente a comprender una causa, y por eso se requiere un tratamiento cuidadoso de la relación observada.

5.1. Asociación entre variables

El primer paso para estudiar la relación entre variables es analizar su **asociación**, es decir, observar si existe algún patrón conjunto en su comportamiento.

Cuando dos variables tienden a aumentar o disminuir juntas, se dice que existe una **relación positiva**. En cambio, si una aumenta mientras la otra disminuye, se habla de una **relación negativa**. Si no se observa un patrón claro, se considera que no hay una relación evidente.

Una herramienta fundamental para resumir esta relación es el **coeficiente de correlación de Pearson (*r*)**, que mide la intensidad y dirección de una relación lineal entre dos variables.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Donde:

x_i, y_i: valores observados

$\bar{x}, \bar{y}$: medias de las variables

$\sum$: sumatoria sobre todas las observaciones

Este coeficiente toma valores entre -1 y $+1$:

- Valores cercanos a $+1$ indican una relación lineal positiva fuerte
- Valores cercanos a -1 indican una relación lineal negativa fuerte
- Valores cercanos a 0 sugieren ausencia de relación lineal clara

Por ejemplo, un coeficiente de correlación de 0.85 indicaría una fuerte relación positiva, mientras que un valor de -0.70 reflejaría una relación negativa considerable.

Aunque esta medida es muy útil desde el punto de vista descriptivo, es importante reiterar que **no constituye evidencia suficiente de causalidad**. Es decir, una alta correlación no implica que una variable cause cambios en la otra.

5.2. Regresión lineal simple

Mientras la correlación describe cómo se comportan conjuntamente dos variables, la **regresión lineal simple** permite dar un paso adicional: construir una ecuación que relacione una variable con otra y que permita realizar estimaciones.

La regresión lineal simple permite construir un modelo matemático que describe la relación entre dos variables:

- Una variable independiente (explicativa)
- Una variable dependiente (respuesta)

Este modelo se expresa mediante una ecuación lineal:

$$y = \beta_0 + \beta_1 x$$

Donde:

β_0: Intercepto (Valor de y cuando $x = 0$)

β_1: pendiente (cambio de y por cada unidad de cambio de x)

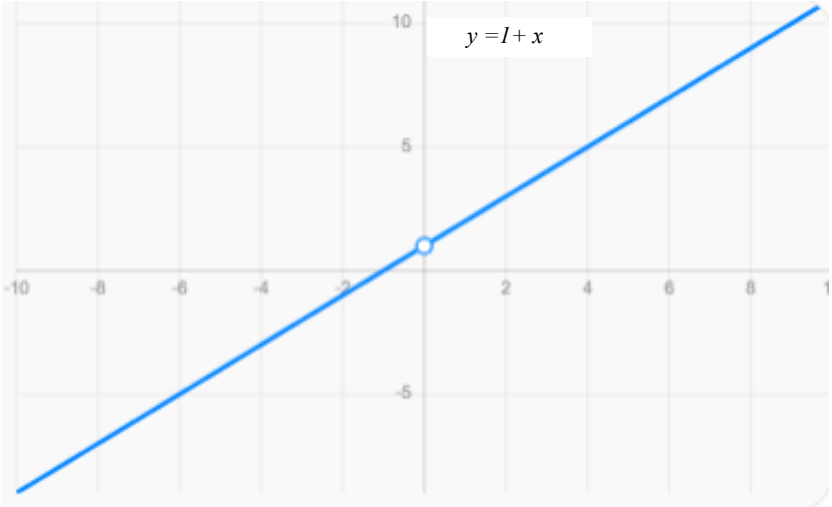

Figura 1. Representación de la recta
Fuente: elaboración propia

En términos prácticos, esta ecuación permite estimar cómo cambia una variable en función de otra. Por ejemplo, si se modela la relación entre inversión en publicidad (x) y ventas (y), la pendiente indica cuánto se espera que aumenten las ventas por cada unidad adicional invertida en publicidad.

Para construir esta ecuación, se utiliza el **método de mínimos cuadrados**, cuyo objetivo es encontrar la recta que mejor se ajusta a los datos observados.

Este método selecciona la recta que **minimiza la suma de los errores cuadrados**, donde cada error corresponde a la diferencia entre el valor observado y el valor estimado por el modelo (Anderson *et al.*, 2008).

Esta idea tiene una implicación importante: la recta ajustada **no pasa necesariamente por todos los puntos**, pero sí representa la tendencia general del conjunto de datos. En la práctica, esto permite:

- Resumir patrones
- Reducir la complejidad de los datos
- Generar predicciones aproximadas

5.2.1. Interpretación de la pendiente y los residuales

Una vez estimado el modelo de regresión, su utilidad no se limita a la ecuación obtenida. Es fundamental interpretar correctamente sus componentes, en particular la **pendiente** y los **residuales**, ya que estos permiten entender tanto la relación entre variables como la calidad del ajuste.

- La **pendiente (β_1)** indica cuánto cambia, en promedio, la variable dependiente cuando la variable independiente aumenta en una unidad, manteniendo el resto constante.
- Los **residuales** representan las diferencias entre los valores observados en los datos y los valores estimados por el modelo de regresión.

El análisis de los residuales permite evaluar qué tan bien el modelo captura el comportamiento de los datos. Si los errores son pequeños y no presentan patrones claros, el modelo puede considerarse adecuado. En cambio, patrones en los residuales pueden indicar que la relación no es realmente lineal o que faltan variables importantes.

Buen modelo:

- Los residuales se distribuyen de forma aleatoria
- No muestran patrones claros
- Tienen magnitud relativamente baja

Posibles problemas:

- **Patrón curvo.** Indica que la relación entre las variables **no es lineal**, pero el modelo sí lo es.
- **Tendencia.** Significa que hay una variable importante que no se incluyó en el modelo.
- **Dispersión desigual**. Si los residuales se abren o cierran (forma de embudo), significa que la variabilidad del error **no es constante**, es decir, el modelo predice mejor en algunos rangos que en otros.
- **Valores extremos**. Posibles *outliers* (observaciones que se desvían significativamente del patrón general de un conjunto de datos), lo cual quiere decir que existen observaciones que el modelo no logra explicar. Pueden ser errores en los datos, casos excepcionales o señales de otro fenómeno.

Como señalan Anderson *et al.* (2008), el análisis de los errores (residuales en la práctica) es fundamental para verificar si el modelo lineal es apropiado.

Tips

Los residuales muestran lo que el modelo **no logró explicar**.

Por eso, si presentan estructura o patrones, no son ruido, sino **señales de que el modelo está mal planteado o incompleto**.

5.2.2. Evaluación del modelo: coeficiente de determinación (R^2)

Para evaluar la calidad del ajuste del modelo, se utiliza el **coeficiente de determinación**, denotado como R^2. Este indicador mide la proporción de la variabilidad de la variable dependiente que es explicada por el modelo. Por ejemplo, un valor de $R^2 = 0.90$ indica que el 90% de la variación en la variable dependiente puede explicarse a partir de la variable independiente (Anderson *et al.*, 2008).

En regresión lineal simple:

$$R^2 = r^2$$

También puede expresarse como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Donde:

SST: suma total de cuadrados (variabilidad total)
SSR: variabilidad explicada por el modelo
SSE: variabilidad no explicada (errores)

Interpretación:

- $R^2 = 0.80 \rightarrow$ el 80% de la variabilidad es explicada por el modelo

- $R^2 = 0 \rightarrow$ el modelo no explica nada
- $R^2 = 1 \rightarrow$ el modelo explica perfectamente los datos

Como señalan Anderson *et al.* (2008), valores altos de R^2 indican mejor ajuste del modelo, aunque no garantizan que sea adecuado en todos los aspectos, "los valores grandes R^2 de implican que la recta de mínimos cuadrados se ajusta mejor a los datos" (s. p.). Esto permite evaluar de manera cuantitativa la capacidad explicativa del modelo.

Es importante notar que tanto el coeficiente de correlación como R^2 informan sobre la intensidad de la relación, aunque en escalas diferentes. Mientras la correlación indica dirección e intensidad, R^2 se enfoca en la proporción de variabilidad explicada.

5.2.3. Alcances y limitaciones de la regresión lineal

Aunque la regresión lineal simple es una herramienta poderosa, es importante reconocer sus limitaciones:

- Asume una relación lineal entre las variables
- No implica causalidad
- Puede verse afectada por valores atípicos
- No considera otras variables relevantes

Por esta razón, sus resultados deben interpretarse con criterio y en conjunto con el contexto del problema.

En el análisis aplicado, la regresión simple suele funcionar como una **puerta de entrada a modelos más complejos**. Antes de trabajar con modelos multivariados o algoritmos de aprendizaje automático, es fundamental comprender:

- Qué significa ajustar una recta
- Cómo interpretar una pendiente
- Qué representan los errores

- Cómo evaluar el ajuste de un modelo

Esta base conceptual permite interpretar correctamente modelos más avanzados y evita el uso mecánico de herramientas sin comprensión de fondo.

6. Métricas estadísticas para la evaluación de modelos

Una vez construido un modelo de regresión, es necesario evaluar su desempeño. No basta con obtener una ecuación o identificar una relación; es fundamental medir qué tan cerca están las predicciones de los valores reales.

Como se discutió anteriormente, los residuales representan el error de cada predicción. Sin embargo, para evaluar el modelo en su conjunto, es necesario resumir estos errores en indicadores globales. Las **métricas de evaluación** cumplen precisamente esta función.

Mientras que cada residual mide el error en una observación particular, las métricas de evaluación agregan esta información para ofrecer una visión general del desempeño del modelo.

Estas métricas permiten:

- Cuantificar el error promedio
- Comparar modelos
- Evaluar la precisión de las predicciones

6.1. Error Absoluto Medio (MAE)

El **Error Absoluto Medio (MAE)** calcula el promedio de los errores en valor absoluto:

$$MAE = \frac{1}{n}\sum | y_i - \hat{y}_i |$$

Donde,

y_i: valor observado.

$\hat{y}_i$:valor estimado (predicho)

n: número de observaciones

Esta métrica indica, en promedio, cuánto se equivoca el modelo en sus predicciones, manteniendo las mismas unidades de la variable analizada.

6.2. Error Cuadrático Medio (MSE)

El **Error Cuadrático Medio (MSE)** se basa en el promedio de los errores al cuadrado:

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$

Al elevar los errores al cuadrado, esta métrica penaliza más fuertemente las desviaciones grandes, lo cual es consistente con el método de mínimos cuadrados utilizado en la regresión.

6.3. Raíz del Error Cuadrático Medio (RMSE)

El **RMSE** es la raíz cuadrada del MSE:

$$RMSE = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2}$$

Esta transformación permite interpretar el error en las mismas unidades de la variable, facilitando su comprensión.

Cada una de estas métricas aporta una perspectiva distinta:

- **MAE**: error promedio directo
- **MSE**: mayor sensibilidad a errores grandes
- **RMSE**: equilibrio entre interpretación y penalización de errores

En la práctica, la elección depende del contexto y del tipo de problema.

6.4. Relación con el coeficiente de determinación

Como se abordó previamente, el coeficiente de determinación R^2 mide la proporción de la variabilidad de la variable dependiente que es explicada por el modelo.

A diferencia de las métricas basadas en error (MAE, MSE, RMSE), que evalúan la precisión de las predicciones, R^2 ofrece una medida del **ajuste global del modelo** (Anderson *et al.*, 2008).

Por esta razón, es recomendable utilizar estas métricas de manera complementaria, ya que cada una aporta información distinta sobre el desempeño del modelo.

6.5. Evaluación en contexto

En ciencia de datos, las métricas no deben interpretarse de forma aislada. Su significado depende del contexto del problema.

Por ejemplo:

- Un error de 10 unidades puede ser irrelevante en grandes volúmenes de ventas
- Pero puede ser crítico en aplicaciones médicas o financieras

Además, la evaluación debe complementarse con:

- análisis de residuales
- conocimiento del dominio
- impacto práctico de las decisiones

Las métricas de evaluación permiten cuantificar el desempeño de un modelo y determinar qué tan bien representa la realidad. A partir de los residuales, indicadores como MAE, MSE y RMSE resumen el error de predicción, mientras que R^2 permite evaluar su capacidad explicativa.

En conjunto, estas herramientas cierran el proceso analítico: no solo se construyen modelos, sino que se evalúan críticamente, lo que permite tomar decisiones más informadas y fundamentadas en datos.

Evaluar un modelo no depende de un único indicador. Es necesario combinar medidas de ajuste, error e interpretación para obtener una visión completa de su utilidad.

Tabla 3. Criterios para evaluar un modelo de regresión

Criterio	Qué indica	Pregunta útil
R^2 (coeficiente de determinación)	Proporción de variabilidad explicada por el modelo	¿Qué parte del comportamiento queda capturada por la relación lineal?
Residuales	Diferencia entre valores observados y estimados	¿Dónde y cómo se equivoca el modelo?
MAE (Error Absoluto Medio)	Magnitud promedio del error	¿Cuánto se equivoca el modelo en promedio?
MSE (Error Cuadrático Medio)	Error promedio penalizando más los errores grandes	¿Qué tanto afectan los errores grandes al modelo?
RMSE (Raíz del Error Cuadrático Medio)	Error promedio en las mismas unidades, sensible a grandes desviaciones	¿Qué tan grande es el error típico considerando desviaciones importantes?
Coherencia de la pendiente	Sentido e interpretación de la relación entre variables	¿La dirección del efecto tiene lógica en el contexto?

Fuente: elaboración propia

Una evaluación estadística rigurosa también exige distinguir entre buen ajuste y buena decisión. Un modelo puede explicar muy bien los datos históricos y aun así aportar poco a la acción si no es interpretable, si requiere información costosa o si sus errores tienen consecuencias críticas en contextos sensibles. Por eso, las métricas deben leerse siempre en relación con el problema real y con el uso previsto para el modelo.

Cómo mejorar...

Antes de presentar un modelo como bueno, traduzca sus métricas a lenguaje aplicado. En lugar de decir solo que el R^2es alto, explique qué parte del fenómeno logra representar y qué limitaciones siguen abiertas para la toma de decisiones.

7. Conclusiones

La estadística aplicada al análisis de datos comienza cuando se acepta que la incertidumbre no es un obstáculo accidental, sino una condición permanente de trabajo. La probabilidad permite representar esa incertidumbre, las variables aleatorias la vuelven operable y las distribuciones brindan modelos para describir comportamientos esperados. Con estas bases, el análisis deja de ser meramente descriptivo y adquiere capacidad explicativa, comparativa y predictiva.

La inferencia estadística enseña una lección decisiva: los datos observados no hablan por sí solos, sino a través de procedimientos que controlan el error y hacen explícitas las condiciones de la conclusión. Estimar, construir intervalos o probar hipótesis implica asumir responsabilidad interpretativa. Por ello, una lectura madura de resultados evita absolutismos y comunica hallazgos con precisión, prudencia y sentido aplicado.

El estudio de la relación entre variables y de la regresión lineal simple aporta una herramienta poderosa para comprender tendencias y generar predicciones iniciales. No obstante, su valor no radica en producir ecuaciones llamativas, sino en ofrecer modelos interpretables que puedan contrastarse con la realidad. Allí, métricas como el coeficiente de determinación y el análisis de residuales ayudan a reconocer tanto fortalezas como límites del ajuste.

Finalmente, evaluar modelos no es una fase periférica, sino el momento en que el análisis demuestra su calidad. Un modelo útil debe ser comprensible, razonablemente preciso y pertinente para la decisión que pretende apoyar. Cuando estos fundamentos se entienden de forma articulada, el trabajo con datos gana profundidad conceptual y también capacidad práctica para resolver problemas en contextos reales.

Referencias

Anderson, D., Sweeney, D. y Williams, T. (2008). *Estadística para administración y economía* (10ª edición). Cengage Learning.

Diez, D., Cetinkaya-Rundel, M. & Barr, C. (2019). *OpenIntro Statistics* (4ª edición). OpenIntro. https://www.openintro.org/go/?id=os4_for_screen_reader&referrer=/book/os/index.php

Información técnica

Autora: Ana María Palomino Marín

Asesor Pedagógico: Juan Sebastian Patarroyo Ocampo

Diseñador Gráfico: Andrés Felipe Figueroa Huertas