

APUNTES DE INTRODUCCIÓN

Estadística

lo esencial

Conceptos, fórmulas, métodos habituales y glosario de símbolos

Autor: David M.
davmacsolutions.com

Octubre de 2026

ÍNDICE

1. Ideas de partida	2
2. Tipos de variables	2
3. Estadística descriptiva	3
4. Representación gráfica	4
5. Probabilidad	4
6. Distribuciones de probabilidad	5
7. Muestreo y teorema central del límite	6
8. Intervalos de confianza	6
9. Contrastes de hipótesis	7
10. Elección de la prueba	8
11. Correlación y regresión	8
12. Errores frecuentes	9
13. Formulario	9
14. Glosario	11

1. IDEAS DE PARTIDA

La estadística tiene dos objetivos: *describir* los datos que tenemos y *generalizar* a partir de ellos, es decir, sacar conclusiones sobre un conjunto grande (la población) observando solo una parte (la muestra), y medir cuánta incertidumbre conllevan esas conclusiones.

Concepto	Qué es	Ejemplo
Población	El conjunto completo que interesa estudiar.	Todos los votantes del país.
Muestra	La parte de la población que se observa.	1.000 votantes encuestados.
Parámetro	Valor numérico de la población, normalmente desconocido. Se escribe con letras griegas (μ , σ , π).	Porcentaje real de apoyo a un partido.
Estadístico	Valor calculado en la muestra. Se escribe con letras latinas ($\bar{x}$, s , $\hat{p}$).	Porcentaje de apoyo en la encuesta.
Estadística descriptiva	Resume y representa los datos disponibles.	Medias, tablas, gráficos.
Estadística inferencial	Generaliza de la muestra a la población con un margen de error.	Intervalos de confianza, contrastes.

ATENCIÓN. Una muestra solo es útil si representa a la población. En un *muestreo aleatorio simple* cada individuo tiene la misma probabilidad de ser elegido. Si la muestra está sesgada (por ejemplo, una encuesta que solo responden los más interesados en el tema), aumentar su tamaño no corrige el problema.

2. TIPOS DE VARIABLES

El tipo de variable determina qué resúmenes, gráficos y pruebas son adecuados.

Tipo	Escala	Descripción	Ejemplos
Cualitativa	Nominal	Categorías sin orden.	Grupo sanguíneo, país.
	Ordinal	Categorías ordenadas; las distancias entre ellas no son comparables.	Nivel de estudios, escala de satisfacción.
Cuantitativa	Discreta	Valores aislados, normalmente recuentos.	Número de hijos, visitas diarias.
	Continua	Cualquier valor dentro de un intervalo.	Estatura, tiempo, temperatura.

3. ESTADÍSTICA DESCRIPTIVA

Para resumir una variable numérica se necesitan, como mínimo, tres cosas: dónde está el *centro*, cuánto se *dispersan* los datos y qué *forma* tiene la distribución.

3.1 Medidas de tendencia central

Media aritmética

$\bar{x} = \frac{1}{n} \sum x_i$. Adecuada para datos simétricos; muy sensible a valores extremos.

Mediana

Valor central de los datos ordenados: deja el 50 % por debajo y el 50 % por encima. Robusta frente a valores atípicos; preferible con datos asimétricos (salarios, precios de vivienda).

Moda

Valor más frecuente. La única medida de centro aplicable a variables nominales.

Ejemplo 3.1. Salarios en miles de euros: 20, 22, 24, 25, 200. La media es 58,2 y la mediana 24. El valor extremo arrastra la media; la mediana describe mucho mejor al trabajador típico.

3.2 Medidas de dispersión

Medida	Definición e interpretación
Rango	Máximo – mínimo. Sencillo, pero depende solo de dos observaciones.
Varianza muestral	$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$. Promedio de desviaciones al cuadrado; se expresa en unidades al cuadrado.
Desviación típica	$s = \sqrt{s^2}$. En las mismas unidades que los datos: es la distancia típica a la media.
Rango intercuartílico	$RIC = Q_3 - Q_1$. Amplitud del 50 % central de los datos; robusto.
Coefficiente de variación	$CV = s/\bar{x}$. Dispersión relativa; permite comparar variables con distinta escala.

Observación. En la varianza muestral se divide entre $n - 1$ y no entre n (corrección de Bessel). Dividir entre n subestimaría, en promedio, la varianza de la población.

3.3 Posición, forma y valores atípicos

El *percentil* k es el valor que deja por debajo el k % de las observaciones; los *cuartiles* Q_1 , Q_2 (mediana) y Q_3 son los percentiles 25, 50 y 75. La *puntuación típica*

$$z = \frac{x - \bar{x}}{s}$$

indica a cuántas desviaciones típicas de la media se encuentra un valor. En datos aproximadamente normales, $|z| > 2$ es poco habitual y $|z| > 3$ es muy raro. Por convenio se considera *atípico* todo valor menor que $Q_1 - 1,5 RIC$ o mayor que $Q_3 + 1,5 RIC$.

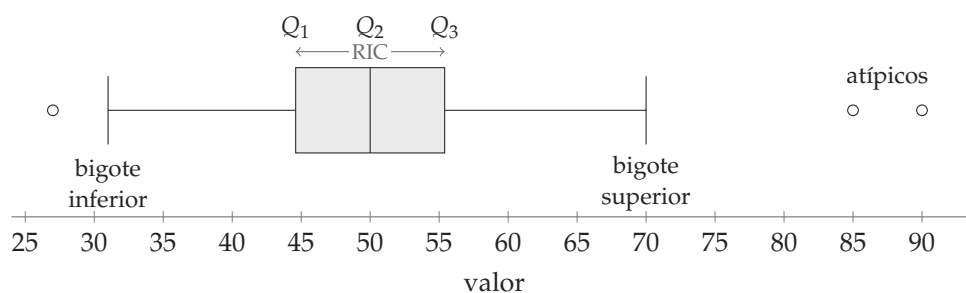


FIGURA 1. Diagrama de caja. Los bigotes llegan hasta el último dato no atípico.

La *asimetría* describe si una cola de la distribución es más larga que la otra. La media se desplaza hacia la cola; la mediana no.

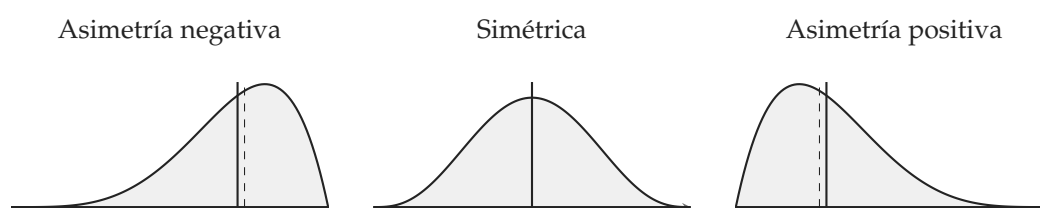


FIGURA 2. Línea continua: media. Línea discontinua: mediana (en la simétrica coinciden).

4. REPRESENTACIÓN GRÁFICA

Objetivo	Gráfico adecuado	A evitar
Distribución de una variable numérica	Histograma, diagrama de caja	Diagrama de sectores
Comparar una variable entre grupos	Cajas en paralelo; medias con intervalos	Barras sin medida de variabilidad
Frecuencias de una variable categórica	Diagrama de barras ordenado	Sectores con muchas categorías; 3D
Relación entre dos variables numéricas	Diagrama de dispersión	Dar solo el coeficiente r
Evolución temporal	Gráfico de líneas	Ejes truncados que exageran los cambios

Conviene representar los datos *antes* de calcular ningún resumen: un gráfico revela de inmediato valores atípicos, errores de registro, grupos diferenciados o relaciones no lineales que las cifras pueden ocultar.

5. PROBABILIDAD

La probabilidad de un suceso A , $P(A)$, es un número entre 0 (suceso imposible) y 1 (suceso seguro).

Regla	Expresión	Comentario
Complementario	$P(\bar{A}) = 1 - P(A)$	A menudo es más fácil calcular lo contrario.
Unión	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	Si son incompatibles, $P(A \cap B) = 0$.
Condicionada	$P(A B) = \frac{P(A \cap B)}{P(B)}$	Probabilidad de A sabiendo que ha ocurrido B.
Independencia	$P(A \cap B) = P(A) P(B)$	Conocer B no altera la probabilidad de A.
Teorema de Bayes	$P(A B) = \frac{P(B A) P(A)}{P(B)}$	Actualiza una probabilidad a la luz de nueva información.

Ejemplo 5.1 (falso positivo). Una enfermedad afecta al 1 % de la población. Una prueba detecta el 90 % de los casos (sensibilidad) y da positivo en el 9 % de los sanos. De cada 1.000 personas, 10 están enfermas y 9 de ellas dan positivo; de las 990 sanas, unas 89 dan también positivo. Por tanto, la probabilidad de estar enfermo tras un resultado positivo es

$$P(\text{enfermo} | +) \approx \frac{9}{9 + 89} \approx 0,09.$$

Confundir $P(A | B)$ con $P(B | A)$, o ignorar la frecuencia de base (prevalencia), es uno de los errores más comunes.

6. DISTRIBUCIONES DE PROBABILIDAD

Una distribución indica qué valores puede tomar una variable aleatoria y con qué probabilidad. La más importante es la *normal*, $N(\mu, \sigma)$, simétrica y en forma de campana.

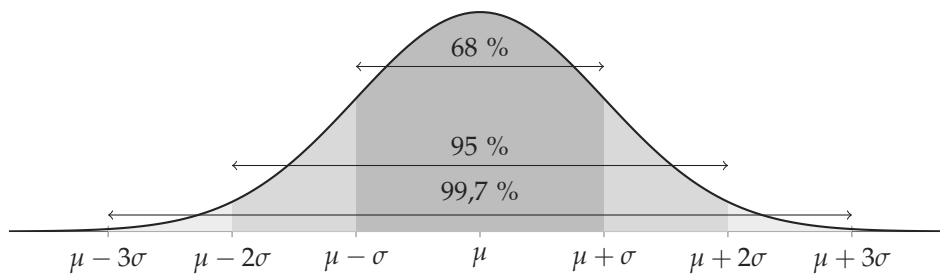


FIGURA 3. Regla empírica de la distribución normal.

Distribución	Tipo	Modela	Ejemplo
Normal $N(\mu, \sigma)$	continua	Fenómenos resultantes de muchos efectos pequeños.	Estaturas, errores de medida
Binomial $B(n, p)$	discreta	Número de éxitos en n ensayos independientes.	Caras en 10 lanzamientos
Poisson $\mathcal{P}(\lambda)$	discreta	Número de sucesos en un intervalo de tiempo o espacio.	Llamadas por hora
Uniforme	ambas	Todos los valores igualmente probables.	Un dado equilibrado
t de Student	continua	Como la normal, con colas más pesadas; inferencia con σ desconocida.	Intervalos para medias
Ji cuadrado χ^2	continua	Contrastes sobre frecuencias y varianzas.	Tablas de contingencia
F de Snedecor	continua	Cociente de varianzas.	ANOVA

7. MUESTREO Y TEOREMA CENTRAL DEL LÍMITE

Cada muestra produce una media distinta. La distribución de esas medias es la *distribución muestral*, y su desviación típica se llama *error estándar*:

$$EE(\bar{x}) = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}.$$

Definición (teorema central del límite). Si las observaciones son independientes y n es suficientemente grande (en la práctica, $n \geq 30$ suele bastar), la media muestral sigue aproximadamente una distribución normal, $\bar{x} \approx N(\mu, \sigma / \sqrt{n})$, aunque la variable original no sea normal.

Dos consecuencias prácticas:

- La precisión crece con $\sqrt{n}$: para reducir el error a la mitad hay que cuadruplicar la muestra.
- La desviación típica mide cuánto varían los *individuos*; el error estándar, cuánto variaría la *media* de una muestra a otra.

8. INTERVALOS DE CONFIANZA

Un intervalo de confianza ofrece un rango de valores plausibles para un parámetro. Su estructura general es

$$\text{estimación} \pm \text{valor crítico} \times \text{error estándar}.$$

Para una media, con confianza del 95 %:

$$\bar{x} \pm 1,96 \frac{s}{\sqrt{n}} \quad (\text{con } n \text{ pequeña se usa } t_{n-1} \text{ en lugar de } 1,96).$$

Nivel de confianza	90 %	95 %	99 %
Valor crítico $z_{\alpha/2}$	1,645	1,960	2,576

Ejemplo 8.1. En una muestra de 100 personas, el tiempo medio de sueño es $\bar{x} = 7,0$ h con $s = 1,2$ h. Entonces $EE = 1,2/\sqrt{100} = 0,12$ y el intervalo al 95 % es $7,0 \pm 1,96 \cdot 0,12 = [6,76; 7,24]$ horas.

ATENCIÓN. El 95 % se refiere al *procedimiento*: si se repitiera el estudio muchas veces, el 95 % de los intervalos así contruidos contendría el valor verdadero. No significa que haya un 95 % de probabilidad de que el parámetro esté en un intervalo concreto ya calculado. A mayor confianza o menor tamaño muestral, intervalo más ancho.

9. CONTRASTES DE HIPÓTESIS

Un contraste evalúa si los datos son compatibles con una afirmación de partida. El procedimiento es:

1. Formular la *hipótesis nula* H_0 (no hay efecto ni diferencia) y la *alternativa* H_1 .
2. Fijar el nivel de significación α , habitualmente 0,05.
3. Calcular el estadístico de contraste (z , t , χ^2 , F ...), que mide cuánto se alejan los datos de lo esperado bajo H_0 .
4. Obtener el *p-valor*: la probabilidad de observar un resultado igual de extremo o más, suponiendo que H_0 es cierta.
5. Decidir: si $p < \alpha$ se rechaza H_0 ; en caso contrario, no hay evidencia suficiente para rechazarla.

	H_0 verdadera	H_0 falsa
Se rechaza H_0	Error de tipo I (falso positivo), prob. α	Decisión correcta, prob. $1 - \beta$ (potencia)
No se rechaza H_0	Decisión correcta, prob. $1 - \alpha$	Error de tipo II (falso negativo), prob. β

TABLA 2. Tipos de error en un contraste.

La *potencia* aumenta con el tamaño muestral, con la magnitud del efecto y al disminuir la variabilidad. Suele exigirse una potencia mínima del 80 %.

ATENCIÓN. El *p-valor* *no* es la probabilidad de que H_0 sea cierta, ni mide la importancia del efecto: con muestras muy grandes, diferencias irrelevantes resultan significativas. Un $p \geq 0,05$ tampoco demuestra que no haya efecto. Por ello deben comunicarse siempre el *tamaño del efecto* y su intervalo de confianza.

10. ELECCIÓN DE LA PRUEBA

Pregunta	Datos	Prueba paramétrica	Alternativa no paramétrica
¿Difiere la media de un valor dado?	Un grupo, variable numérica	t para una muestra	Rangos con signo de Wilcoxon
¿Difieren dos grupos?	Dos grupos independientes	t de Student (o de Welch)	U de Mann-Whitney
¿Hay cambio antes y después?	Datos emparejados	t para datos pareados	Wilcoxon para datos pareados
¿Difieren tres o más grupos?	k grupos independientes	ANOVA de un factor	Kruskal-Wallis
¿Están asociadas dos variables categóricas?	Tabla de contingencia	χ^2 de independencia	Prueba exacta de Fisher
¿Se relacionan dos variables numéricas?	Pares (x, y)	Correlación de Pearson	Correlación de Spearman
¿Se puede predecir y a partir de x ?	Numéricas o mixtas	Regresión lineal	Regresión logística (y dicotómica)

Las pruebas paramétricas suponen normalidad aproximada (o muestras grandes) y, a menudo, varianzas similares. Las no paramétricas trabajan con rangos y son más robustas frente a asimetrías y valores atípicos. Cuando se realizan muchas pruebas a la vez conviene aplicar una corrección por comparaciones múltiples (Bonferroni, Holm).

11. CORRELACIÓN Y REGRESIÓN

El coeficiente de correlación de Pearson, $r \in [-1, 1]$, mide la intensidad de la relación *lineal* entre dos variables. Como orientación, $|r| < 0,3$ indica relación débil, entre 0,3 y 0,7 moderada y por encima de 0,7 fuerte, aunque el criterio depende de la disciplina.

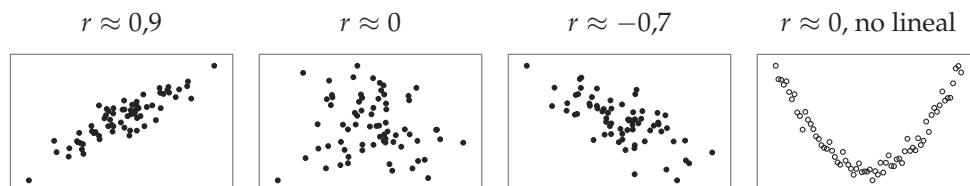


FIGURA 4. En el último panel la relación es evidente, pero no lineal, y r no la detecta.

La regresión lineal simple ajusta la recta $\hat{y} = b_0 + b_1x$ por mínimos cuadrados:

- b_1 (pendiente): variación media de y por cada unidad que aumenta x ;
- b_0 (ordenada en el origen): valor previsto de y cuando $x = 0$;
- R^2 (coeficiente de determinación): proporción de la variabilidad de y explicada por el modelo; en regresión simple, $R^2 = r^2$;

- residuos $e_i = y_i - \hat{y}_i$: su gráfico no debe mostrar ningún patrón.

ATENCIÓN. Correlación no implica causalidad. Las ventas de helados y los ahogamientos están correlacionados porque ambos aumentan con el calor, que actúa como *variable de confusión*. Establecer una relación causal requiere un experimento aleatorizado o un diseño observacional muy cuidadoso. Tampoco debe extrapolarse fuera del rango de x observado.

12. ERRORES FRECUENTES

Error	Descripción
Sesgo de selección	La muestra no representa a la población (voluntarios, supervivientes, no respuesta).
Correlación como causa	La asociación puede deberse a terceras variables o a causalidad inversa.
Media con datos asimétricos	En ingresos, precios o tiempos, la mediana describe mejor el caso típico.
Ignorar la frecuencia de base	Sin la prevalencia, las probabilidades condicionadas engañan.
Búsqueda de significación	Repetir análisis hasta obtener $p < 0,05$. Con 20 pruebas sin efecto real cabe esperar una «significativa».
Significativo = importante	La significación estadística no garantiza relevancia práctica.
Paradoja de Simpson	Una tendencia presente en cada subgrupo puede invertirse al agregar los datos.
Regresión a la media	Tras un valor extremo suele venir uno más moderado, sin intervención alguna.
Muestras pequeñas	Estimaciones inestables; los efectos grandes en estudios pequeños a menudo no se replican.
Gráficos engañosos	Ejes truncados, escalas distintas, efectos 3D, porcentajes sin el total de referencia.

13. FORMULARIO

Concepto	Fórmula
Media	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Varianza muestral	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
Desviación típica	$s = \sqrt{s^2}$
Coefficiente de variación	$CV = s/\bar{x}$
Puntuación típica	$z = (x - \mu)/\sigma$
Límites de atípicos	$Q_1 - 1,5 \text{ RIC}, \quad Q_3 + 1,5 \text{ RIC}$

Concepto	Fórmula
Error estándar de la media	$EE = s / \sqrt{n}$
Error estándar de una proporción	$EE = \sqrt{\hat{p}(1 - \hat{p}) / n}$
IC para la media	$\bar{x} \pm t_{n-1, \alpha/2} s / \sqrt{n}$
IC para una proporción	$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p}) / n}$
Estadístico t (una muestra)	$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$
Probabilidad condicionada	$P(A B) = P(A \cap B) / P(B)$
Teorema de Bayes	$P(A B) = \frac{P(B A)P(A)}{P(B)}$
Binomial	$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad E(X) = np, \text{ Var}(X) = np(1 - p)$
Poisson	$P(X = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad E(X) = \text{Var}(X) = \lambda$
Correlación de Pearson	$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$
Recta de regresión	$b_1 = r \frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1 \bar{x}$
Tamaño muestral (media)	$n = \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2, \quad E = \text{margen de error}$

14. GLOSARIO

14.1 Símbolos

Símbolo	Lectura	Significado
N	ene mayúscula	Tamaño de la población.
n	ene	Tamaño de la muestra.
x_i	equis sub i	Valor de la observación i -ésima.
Σ	sumatorio	Suma de todos los términos indicados.
μ	mu	Media de la población.
$\bar{x}$	equis barra	Media de la muestra.
σ	sigma	Desviación típica de la población.
σ^2	sigma cuadrado	Varianza de la población.
s, s^2	ese, ese cuadrado	Desviación típica y varianza de la muestra.
π o p	pi, pe	Proporción en la población (y probabilidad de éxito en la binomial).
$\hat{p}$	pe gorro	Proporción observada en la muestra.
$\theta, \hat{\theta}$	theta, theta gorro	Parámetro genérico y su estimador.
Me, Mo		Mediana y moda.
Q_1, Q_2, Q_3		Primer, segundo (mediana) y tercer cuartil.
P_k		Percentil k .
$P(A)$	pe de A	Probabilidad del suceso A .
$P(A B)$	pe de A dado B	Probabilidad de A condicionada a B .
$A \cap B, A \cup B$	intersección, unión	« A y B », « A o B ».
$\bar{A}$ o A^c	A complementario	El suceso «no A ».
X, Y		Variables aleatorias (mayúsculas); sus valores, en minúsculas.
$E(X)$	esperanza de X	Valor esperado (media teórica) de X .
$\text{Var}(X)$	varianza de X	Varianza teórica de X .
$\text{Cov}(X, Y)$	covarianza	Variación conjunta de dos variables.
$\sim$	se distribuye como	$X \sim N(0, 1)$: X sigue una normal estándar.
$\approx$	aproximadamente	Igualdad aproximada.
$N(\mu, \sigma)$	normal	Distribución normal de media μ y desviación σ (a veces se escribe con σ^2).
$B(n, p)$	binomial	Distribución binomial.
λ	lambda	Parámetro (media) de la distribución de Poisson.
z	zeta	Puntuación típica o estadístico de la normal estándar.
t	te	Estadístico de la t de Student.
χ^2	ji cuadrado	Estadístico y distribución ji cuadrado.
F	efe	Estadístico de Snedecor (cociente de varianzas).
ν o gl	nu, grados de libertad	Grados de libertad de una distribución.
H_0, H_1	hache cero, hache uno	Hipótesis nula y alternativa (también H_a).

Símbolo	Lectura	Significado
α	alfa	Nivel de significación; probabilidad de error de tipo I.
β	beta	Probabilidad de error de tipo II.
$1 - \beta$		Potencia del contraste.
p	p-valor	Probabilidad de un resultado igual o más extremo bajo H_0 .
$z_{\alpha/2}$		Valor crítico de la normal que deja $\alpha/2$ en cada cola.
r	erre	Coeficiente de correlación muestral de Pearson.
ρ	ro	Correlación poblacional; r_s o ρ también se usan para la de Spearman.
R^2	erre cuadrado	Coeficiente de determinación.
b_0, b_1		Ordenada en el origen y pendiente estimadas.
β_0, β_1		Coeficientes poblacionales de la regresión.
$\hat{y}$	y gorro	Valor predicho por el modelo.
e_i o ε_i	residuo, error	Diferencia entre valor observado y predicho.
$\binom{n}{k}$	n sobre k	Número combinatorio: formas de elegir k elementos de n .
$n!$	n factorial	$n \cdot (n - 1) \cdot \dots \cdot 2 \cdot 1$.

14.2 Siglas y abreviaturas

Muchos programas y artículos usan las siglas en inglés; se indican entre paréntesis.

Sigla	En inglés	Significado
DE, DT	SD	Desviación estándar o típica.
EE	SE, SEM	Error estándar (de la media).
RIC	IQR	Rango intercuartílico.
CV	CV	Coeficiente de variación.
IC	CI	Intervalo de confianza.
TCL	CLT	Teorema central del límite.
gl	df	Grados de libertad.
v.a.	r.v.	Variable aleatoria.
i.i.d.	i.i.d.	Independientes e idénticamente distribuidas.
MAS	SRS	Muestreo aleatorio simple.
ANOVA	ANOVA	Análisis de la varianza.
MCO	OLS	Mínimos cuadrados ordinarios.
EMV	MLE	Estimador de máxima verosimilitud.
ECM	MSE	Error cuadrático medio.
RECM	RMSE	Raíz del error cuadrático medio.
EAM	MAE	Error absoluto medio.
VP, FP	TP, FP	Verdaderos y falsos positivos.
VN, FN	TN, FN	Verdaderos y falsos negativos.
VPP, VPN	PPV, NPV	Valor predictivo positivo y negativo.
OR	OR	Razón de odds (de ventajas).
RR	RR	Riesgo relativo.

Sigla	En inglés	Significado
ROC, AUC	ROC, AUC	Curva ROC y área bajo la curva.
ECA	RCT	Ensayo controlado aleatorizado.
n.s.	n.s.	No significativo.

14.3 Términos

Estimador. Regla para calcular, a partir de la muestra, un valor aproximado de un parámetro.

Insesgado. Dicho de un estimador cuyo valor medio coincide con el parámetro que estima.

Robusto. Poco sensible a valores atípicos o a incumplimientos de los supuestos (la mediana lo es; la media, no).

Sesgo. Desviación sistemática respecto al valor verdadero, debida al diseño o al método.

Variable de confusión. Variable relacionada con la causa y el efecto que puede producir una asociación espuria.

Tamaño del efecto. Magnitud de una diferencia o relación, independiente del tamaño muestral (d de Cohen, r , OR...).

Significación estadística. Resultado con $p < \alpha$: poco compatible con H_0 .

Potencia. Probabilidad de detectar un efecto que realmente existe.

Prueba paramétrica. Supone una distribución concreta (normalmente la normal) para los datos.

Prueba no paramétrica. No supone una distribución concreta; suele basarse en rangos.

Homocedasticidad. Igualdad de varianzas entre grupos o a lo largo de la recta de regresión.

Valor atípico. Observación muy alejada del resto (*outlier*).

Prevalencia. Proporción de la población que presenta una característica en un momento dado.

Sensibilidad. Proporción de positivos reales que una prueba identifica como positivos.

Especificidad. Proporción de negativos reales que una prueba identifica como negativos.