

ESTADÍSTICA MATEMÁTICA

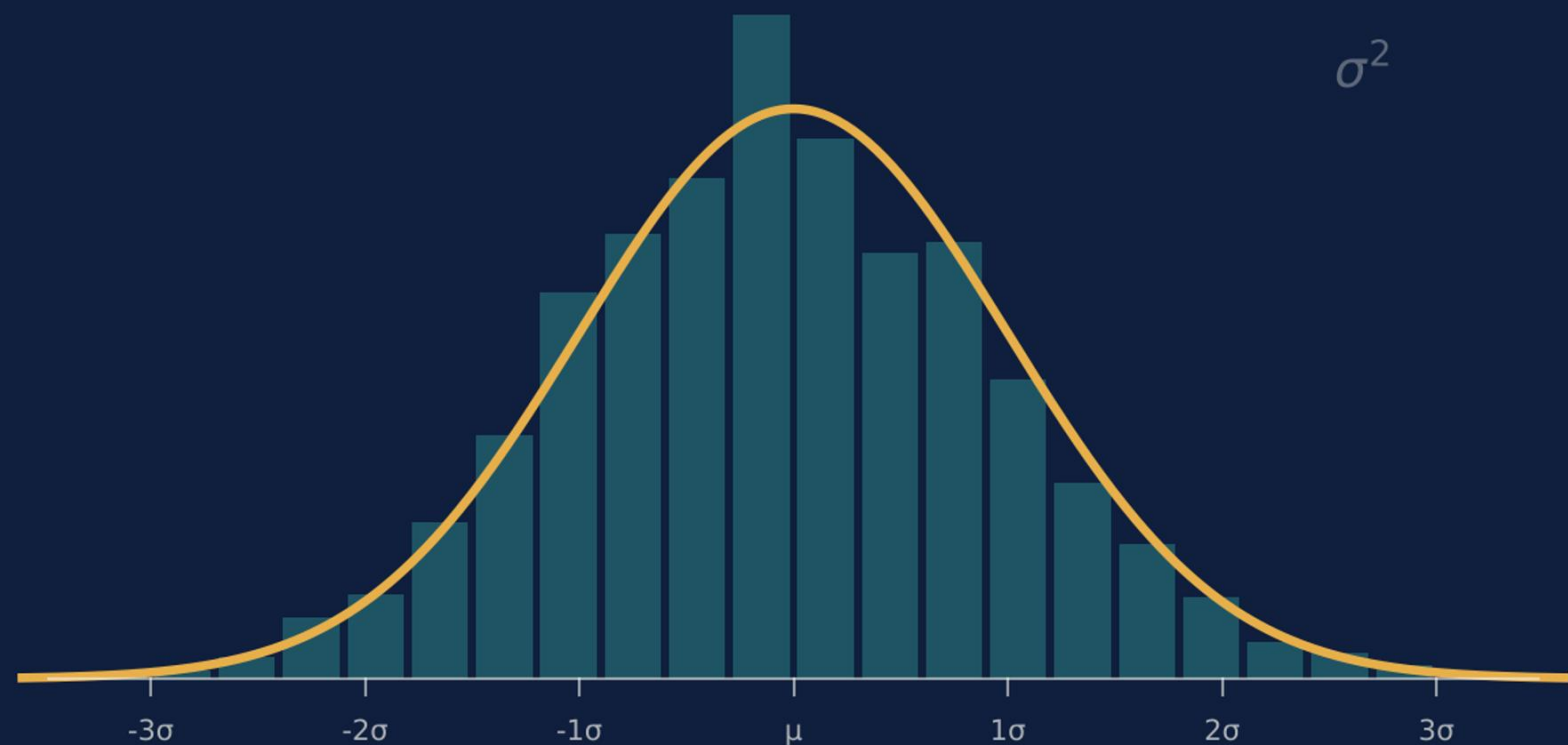
de 0 a 100

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

Del primer dato al razonamiento bayesiano:
teoría, ejemplos resueltos y ejercicios

$$P(A|B)$$

$$\sigma^2$$



Marlon Falcón Hernández

Implemento BIM 2.0 · Valencia

Edición 2026

Prólogo

Este libro te lleva de no saber nada de estadística (nivel 0) a dominar la inferencia moderna (nivel 100). Cada capítulo sube un peldaño y se apoya en el anterior, así que conviene leerlo en orden.

La estadística responde a una pregunta práctica: ¿qué podemos afirmar sobre el mundo a partir de datos incompletos? Un presupuesto de obra, un ensayo clínico o una campaña de ventas se deciden con muestras, nunca con la población entera. Aquí aprenderás a resumir esos datos, a medir la incertidumbre y a decidir con un riesgo de error conocido.

Cómo usar este libro

Nivel	Capítulos	Qué sabrás hacer
0-25	1-3	Clasificar datos, construir tablas y calcular media, mediana, varianza y cuartiles
25-50	4-6	Calcular probabilidades, aplicar Bayes y manejar las distribuciones clave
50-75	7-9	Entender el teorema central del límite, construir intervalos y contrastar hipótesis
75-100	10-11	Ajustar regresiones, usar métodos bayesianos, no paramétricos y multivariantes

Cada capítulo tiene tres partes: teoría, ejemplos resueltos y ejercicios. Las soluciones están en los apéndices. Las fórmulas aparecen en recuadros; no hace falta memorizarlas, sino entender cuándo usarlas.

Cada capítulo incluye además una sección **En Python** con código listo para ejecutar. Necesitas Python 3.10 o superior y estas librerías:

```
pip install numpy pandas scipy matplotlib statsmodels scikit-learn
```

Índice

1. Qué es la estadística: datos y variables
2. Estadística descriptiva: frecuencias y tendencia central
3. Dispersión, posición y forma
4. Probabilidad
5. Variables aleatorias discretas
6. Distribuciones continuas

- 7. Muestreo y teorema central del límite
- 8. Estimación
- 9. Contraste de hipótesis
- 10. Correlación y regresión
- 11. Temas avanzados: Bayes, no paramétrica, remuestreo y multivariante
- 12. Apéndices: formulario, soluciones y glosario

Capítulo 1 · Qué es la estadística (nivel 0)

La estadística es la ciencia de recoger, resumir e interpretar datos para tomar decisiones con incertidumbre. Se divide en dos ramas: la **descriptiva**, que resume lo observado, y la **inferencial**, que generaliza de una muestra a una población.

1.1 Población, muestra y parámetro

- **Población:** el conjunto completo que nos interesa (todas las viviendas de Valencia).
- **Muestra:** el subconjunto que realmente observamos (300 viviendas encuestadas).
- **Parámetro:** un número que describe la población (precio medio real, μ). Suele ser desconocido.
- **Estadístico:** un número calculado con la muestra (precio medio muestral, $\bar{x}$). Lo usamos para estimar el parámetro.

Convención de notación: letras griegas (μ , σ , p) para parámetros; letras latinas ($\bar{x}$, s , $\hat{p}$) para estadísticos.

1.2 Tipos de variables

Tipo	Subtipo	Ejemplo	Operaciones válidas
Cualitativa	Nominal	Color, provincia	Contar, moda
Cualitativa	Ordinal	Nivel de satisfacción (bajo/medio/alto)	Ordenar, mediana
Cuantitativa	Discreta	Nº de habitaciones	Todas las aritméticas
Cuantitativa	Continua	Superficie en m ² , tiempo	Todas las aritméticas

La escala importa: no tiene sentido calcular la media de «provincia», pero sí su moda.

1.3 Cómo se obtienen los datos

- **Estudio observacional:** se mide sin intervenir. Detecta asociaciones, no causas.

- **Experimento:** se asigna un tratamiento al azar. Permite inferir causalidad.
- **Sesgo de selección:** ocurre cuando la muestra no representa a la población (encuestar solo por internet a mayores de 80 años).

Ejemplo resuelto 1.1

Una constructora anota, para 5 obras: tipo (reforma/obra nueva), nº de plantas y coste en miles de €. Tipo es cualitativa nominal; nº de plantas es cuantitativa discreta; coste es cuantitativa continua.

En Python

Carga los datos en un DataFrame y declara el tipo de cada variable; `describe()` aplica a cada una el resumen que le corresponde.

```
import pandas as pd

obras = pd.DataFrame({
    "tipo": ["reforma", "obra nueva", "reforma", "obra nueva", "reforma"],
    "plantas": [1, 4, 2, 3, 1],
    "coste_miles": [45.0, 820.5, 96.2, 610.0, 38.7],
})
obras["tipo"] = obras["tipo"].astype("category")
print(obras.dtypes)
print(obras.describe(include="all"))
```

Ejercicios 1

1. Clasifica: código postal, temperatura en °C, talla de camiseta (S/M/L), nº de hijos.
2. Una app pregunta a sus usuarios si están satisfechos con la app. ¿Qué sesgo puede haber?
3. Distingue parámetro y estadístico en: «El 42 % de 1.000 encuestados votará A».

Capítulo 2 · Frecuencias y tendencia central (nivel 10)

Resumir datos empieza por contarlos y termina buscando un valor «típico». Las tres medidas de centro —media, mediana y moda— responden a esa pregunta de formas distintas.

2.1 Tablas de frecuencias

Para n datos con valores distintos $x_1, \dots, x_k$:

- **Frecuencia absoluta** n_i : cuántas veces aparece x_i .

- **Frecuencia relativa** $f_i = n_i / n$. Las f_i suman 1.
- **Frecuencia acumulada** $N_i = n_1 + \dots + n_i$.

Ejemplo: nº de incidencias diarias en una obra durante 20 días.

Incidencias x_i	n_i	f_i	N_i
0	6	0,30	6
1	8	0,40	14
2	4	0,20	18
3	2	0,10	20

Con datos continuos se agrupan en **intervalos de clase** y se dibuja un **histograma**: el área de cada barra es proporcional a la frecuencia.

2.2 Media aritmética

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

En el ejemplo: $\bar{x} = 0 \cdot 0,30 + 1 \cdot 0,40 + 2 \cdot 0,20 + 3 \cdot 0,10 = 1,10$ incidencias/día.

Propiedades: la suma de desviaciones $\sum (x_i - \bar{x})$ vale 0; si $y = a + b \cdot x$, entonces $\bar{y} = a + b \cdot \bar{x}$.

Defecto: es sensible a valores extremos.

Variantes: **media ponderada** $\sum w_i x_i / \sum w_i$; **media geométrica** (para tasas de crecimiento); **media armónica** (para velocidades o ratios).

2.3 Mediana y moda

- **Mediana** Me : el valor central de los datos ordenados. Si n es impar, el dato $(n+1)/2$; si es par, la media de los datos $n/2$ y $n/2+1$. En el ejemplo, los datos 10 y 11 valen 1, así que $Me = 1$.
- **Moda** Mo : el valor más frecuente. En el ejemplo, $Mo = 1$.

Regla práctica: si la distribución tiene cola a la derecha (salarios, precios), $media > mediana$. Para datos asimétricos, la mediana representa mejor al «individuo típico».

Ejemplo resuelto 2.1

Salarios (miles €) de 7 empleados: 22, 24, 25, 26, 27, 29, 120. Media = $273/7 = 39,0$. Mediana = 26. El director (120) arrastra la media; la mediana describe mejor la plantilla.

En Python

Tabla de frecuencias y medidas de centro del capítulo, calculadas con pandas y SciPy.

```
import numpy as np
import pandas as pd
from scipy import stats

incidencias = np.repeat([0, 1, 2, 3], [6, 8, 4, 2])
tabla = pd.Series(incidencias).value_counts().sort_index().to_frame("n")
tabla["f"] = tabla["n"] / tabla["n"].sum()
tabla["N"] = tabla["n"].cumsum()
print(tabla)

salarios = np.array([22, 24, 25, 26, 27, 29, 120])
print("Media:", salarios.mean().round(1))
print("Mediana:", np.median(salarios))
print("Moda:", stats.mode(incidencias, keepdims=False).mode)
print("Media geométrica:", stats.gmean([1.1, 0.8, 1.3]) - 1)
print("Media armónica:", stats.hmean([60, 40]))
```

Salida esperada: Media 39,0; mediana 26; moda 1; media geométrica $\approx 0,046$; media armónica 48.

Ejercicios 2

1. Calcula media, mediana y moda de: 4, 7, 7, 8, 10, 12, 15, 7.
2. Un coche va a 60 km/h a la ida y 40 km/h a la vuelta. ¿Cuál es la velocidad media del trayecto?
3. Una inversión crece +10 %, -20 % y +30 % en tres años. Calcula la tasa media anual.

Capítulo 3 · Dispersión, posición y forma (nivel 20)

Dos obras con el mismo coste medio pueden tener riesgos muy distintos: el centro no basta, hay que medir cuánto varían los datos.

3.1 Rango y rango intercuartílico

- **Rango** = $\text{máx} - \text{mín}$. Simple, pero depende solo de dos datos.
- **Cuartiles** Q_1 , $Q_2 = \text{Me}$, Q_3 : dividen los datos ordenados en cuatro partes con un 25 % cada una. Los **percentiles** P_k generalizan la idea (P_{90} deja el 90 % por debajo).
- **Rango intercuartílico** $\text{RIC} = Q_3 - Q_1$. Es robusto frente a extremos.
- **Atípicos** (regla de Tukey): datos por debajo de $Q_1 - 1,5 \cdot \text{RIC}$ o por encima de $Q_3 + 1,5 \cdot \text{RIC}$. Es la base del **diagrama de caja**.

3.2 Varianza y desviación típica

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

Se divide entre $n - 1$ (corrección de Bessel) para que s^2 sea un estimador insesgado de σ^2 (lo demostraremos en el capítulo 8). Fórmula de cálculo rápido:

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

Propiedad: si $y = a + b \cdot x$, entonces $s_y = |b| \cdot s_x$. Sumar una constante no cambia la dispersión.

Coefficiente de variación $CV = s / \bar{x}$. Compara dispersión entre variables con distinta unidad o escala.

Puntuación típica $z = (x - \bar{x}) / s$: cuántas desviaciones típicas se aleja un dato de la media.

3.3 Forma: asimetría y curtosis

$$g_1 = \frac{\frac{1}{n} \sum (x_i - \bar{x})^3}{s^3} \quad g_2 = \frac{\frac{1}{n} \sum (x_i - \bar{x})^4}{s^4} - 3$$

- $g_1 > 0$: cola a la derecha; $g_1 < 0$: cola a la izquierda; $g_1 \approx 0$: simétrica.
- $g_2 > 0$ (leptocúrtica): colas más pesadas que la normal; $g_2 < 0$ (platicúrtica): colas más ligeras.

3.4 Desigualdad de Chebyshev

Para cualquier distribución, al menos $1 - 1/k^2$ de los datos está a menos de k desviaciones típicas de la media. Con $k = 2$, al menos el 75 %; con $k = 3$, al menos el 88,9 %. Si los datos son normales, la regla empírica afina: 68 % / 95 % / 99,7 %.

Ejemplo resuelto 3.1

Datos: 2, 4, 4, 5, 7, 8. $\bar{x} = 30/6 = 5$. Desviaciones al cuadrado: 9, 1, 1, 0, 4, 9 $\rightarrow$ suma 24. $s^2 = 24/5 = 4,8$; $s \approx 2,19$. $CV = 2,19/5 = 0,44$ (44 %).

En Python

Dispersión, cuartiles, atípicos de Tukey, forma y diagrama de caja.

```

import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

x = np.array([2, 4, 4, 5, 7, 8])
print("Varianza (n-1):", x.var(ddof=1))
print("Desv. típica:", x.std(ddof=1).round(2))
print("CV:", (x.std(ddof=1) / x.mean()).round(2))

y = np.array([3, 5, 6, 8, 9, 11, 13, 14, 40])
q1, q3 = np.percentile(y, [25, 75])
ric = q3 - q1
print("Q1, Q3, RIC:", q1, q3, ric)
print("Atípicos:", y[(y < q1 - 1.5 * ric) | (y > q3 + 1.5 * ric)])
print("Asimetría:", stats.skew(y).round(2), "Curtosis:",
stats.kurtosis(y).round(2))

plt.boxplot(y, orientation="horizontal")
plt.title("Diagrama de caja")
plt.savefig("caja.png")

```

Salida esperada: $s^2 = 4,8$; $s \approx 2,19$; RIC = 7; atípico: 40.

Ejercicios 3

1. Calcula Q_1 , Q_3 y RIC de: 3, 5, 6, 8, 9, 11, 13, 14, 40. ¿Es 40 atípico?
2. Las notas de un examen tienen $\bar{x} = 6,2$ y $s = 1,4$. Ana sacó 8,5. Calcula su puntuación z.
3. Una máquina A rellena botes con $\bar{x} = 500$ g, $s = 5$ g; otra B con $\bar{x} = 250$ g, $s = 4$ g. ¿Cuál es relativamente más precisa?

Capítulo 4 · Probabilidad (nivel 30)

La probabilidad es el lenguaje matemático de la incertidumbre y el puente entre la estadística descriptiva y la inferencial.

4.1 Espacio muestral y sucesos

- **Experimento aleatorio:** su resultado no se puede predecir (lanzar un dado).
- **Espacio muestral Ω :** conjunto de resultados posibles, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- **Suceso:** subconjunto de Ω . Operaciones: unión $A \cup B$ («A o B»), intersección $A \cap B$ («A y B»), complementario A^c («no A»).

4.2 Axiomas de Kolmogórov

1. $P(A) \geq 0$ para todo suceso A .
2. $P(\Omega) = 1$.
3. Si $A_1, A_2, \dots$ son incompatibles dos a dos, $P(\cup A_i) = \sum P(A_i)$.

Consecuencias: $P(A^c) = 1 - P(A)$; $P(\emptyset) = 0$; y la **regla de la suma**:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Con resultados equiprobables vale la **regla de Laplace**: $P(A) = \text{casos favorables} / \text{casos posibles}$.

4.3 Combinatoria

Situación	¿Importa el orden?	¿Se repite?	Fórmula
Permutaciones de n	Sí	No	$n!$
Variaciones de n tomados de k	Sí	No	$n! / (n-k)!$
Variaciones con repetición	Sí	Sí	n^k
Combinaciones de n tomados de k	No	No	$C(n,k) = n! / (k! (n-k)!)$

4.4 Probabilidad condicionada e independencia

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{para } P(B) > 0$$

A y B son **independientes** si $P(A \cap B) = P(A) \cdot P(B)$, es decir, si saber B no cambia la probabilidad de A . No confundir con incompatibles ($P(A \cap B) = 0$).

4.5 Probabilidad total y teorema de Bayes

Si $B_1, \dots, B_k$ forman una partición de Ω :

$$P(A) = \sum_i P(A \mid B_i) P(B_i) \quad \text{y} \quad P(B_j \mid A) = \frac{P(A \mid B_j) P(B_j)}{\sum_i P(A \mid B_i) P(B_i)}$$

Bayes «invierte» la condición: de $P(\text{síntoma} \mid \text{enfermedad})$ pasa a $P(\text{enfermedad} \mid \text{síntoma})$.

Ejemplo resuelto 4.1 (el test diagnóstico)

Una enfermedad afecta al 1 % de la población. Un test detecta al 95 % de los enfermos (sensibilidad) y da negativo al 90 % de los sanos (especificidad). Si alguien da positivo:

$$P(+) = 0,95 \cdot 0,01 + 0,10 \cdot 0,99 = 0,0095 + 0,099 = 0,1085.$$

$$P(\text{enfermo} \mid +) = 0,0095 / 0,1085 \approx 0,088.$$

Solo un 8,8 % de los positivos está enfermo: con una prevalencia baja, la mayoría de positivos son falsos. Ignorar la probabilidad a priori es la **falacia de la tasa base**.

En Python

Bayes, combinatoria y una simulación de Monte Carlo para comprobar una probabilidad.

```
import numpy as np
from math import comb

# Teorema de Bayes: el test diagnóstico
prev, sens, espec = 0.01, 0.95, 0.90
p_pos = sens * prev + (1 - espec) * (1 - prev)
print("P(enfermo | +) =", round(sens * prev / p_pos, 3))

# Combinatoria
print("Comités de 3 entre 10:", comb(10, 3))

# Simulación: P(suma de dos dados = 7)
rng = np.random.default_rng(42)
dados = rng.integers(1, 7, size=(100_000, 2))
print("P(suma=7) simulada:", (dados.sum(axis=1) == 7).mean())
```

Salida esperada: $P(\text{enfermo} \mid +) \approx 0,088$; 120 comités; $P(\text{suma}=7) \approx 0,167$.

Ejercicios 4

1. Se lanzan dos dados. Calcula $P(\text{suma} = 7)$ y $P(\text{al menos un } 6)$.
2. ¿De cuántas formas se elige un comité de 3 personas entre 10?
3. El 60 % de las piezas viene del proveedor A (2 % defectuosas) y el 40 % de B (5 % defectuosas). Una pieza es defectuosa: ¿qué probabilidad hay de que venga de B?

Capítulo 5 · Variables aleatorias discretas (nivel 40)

Una variable aleatoria X asigna un número a cada resultado de un experimento. Si toma valores aislados (0, 1, 2...), es discreta y se describe con su **función de probabilidad** $p(x) = P(X = x)$, donde $\sum p(x) = 1$.

5.1 Esperanza y varianza

$$E[X] = \mu = \sum_x x \cdot p(x) \quad \operatorname{Var}(X) = \sigma^2 = E[(X - \mu)^2] = E[X^2] - \mu^2$$

Propiedades clave (a, b constantes):

- $E[aX + b] = a \cdot E[X] + b$; $\operatorname{Var}(aX + b) = a^2 \cdot \operatorname{Var}(X)$.
- $E[X + Y] = E[X] + E[Y]$ siempre.
- $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2 \cdot \operatorname{Cov}(X, Y)$; si son independientes, $\operatorname{Cov} = 0$.

La **función de distribución** $F(x) = P(X \leq x)$ es creciente, va de 0 a 1 y vale para cualquier tipo de variable.

5.2 Distribuciones discretas fundamentales

Distribución	Modela	$p(x)$	$E[X]$	$\operatorname{Var}(X)$
Bernoulli(p)	Un ensayo éxito/fracaso	$p^x(1-p)^{1-x}, x \in \{0,1\}$	p	$p(1-p)$
Binomial(n, p)	Nº de éxitos en n ensayos independientes	$C(n,x) \cdot p^x(1-p)^{n-x}$	np	$np(1-p)$
Geométrica(p)	Nº de ensayos hasta el primer éxito	$(1-p)^{x-1} \cdot p, x \geq 1$	$1/p$	$(1-p)/p^2$
Poisson(λ)	Nº de sucesos raros en un intervalo	$e^{(-\lambda)} \cdot \lambda^x / x!$	λ	λ
Hipergeométrica(N, K, n)	Éxitos al extraer n sin reemplazo	$C(K,x)C(N-K,n-x)/C(N,n)$	nK/N	$nK/N \cdot (N-K)/N \cdot (N-n)/(N-1)$

Aproximación de Poisson a la binomial: si n es grande y p pequeño ($n \geq 50, np \leq 5$), $\operatorname{Bin}(n, p) \approx \operatorname{Poisson}(\lambda = np)$.

Ejemplo resuelto 5.1

Un 3 % de los ladrillos de un lote sale defectuoso. Se toman 20 al azar. $X \sim \operatorname{Bin}(20; 0,03)$.

$P(X = 0) = 0,97^{20} \approx 0,544$. $P(X \geq 2) = 1 - P(0) - P(1) = 1 - 0,544 - 20 \cdot 0,03 \cdot 0,97^{19} \approx 1 - 0,544 - 0,336 = 0,120$. $E[X] = 0,6$ ladrillos.

Ejemplo resuelto 5.2

Una centralita recibe de media 4 llamadas por minuto. $X \sim \text{Poisson}(4)$. $P(X = 2) = e^{-4} \cdot 4^2 / 2! = 0,0183 \cdot 8 \approx 0,147$.

En Python

Las distribuciones de `scipy.stats` exponen `pmf`, `cdf`, `sf` (cola superior), `mean` y `var`.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

X = stats.binom(n=20, p=0.03)
print("P(X=0):", X.pmf(0).round(3))
print("P(X>=2):", X.sf(1).round(3))
print("E[X], Var[X]:", X.mean(), X.var())

Y = stats.poisson(mu=4)
print("P(Y=2):", Y.pmf(2).round(3))

k = np.arange(0, 13)
plt.bar(k, Y.pmf(k))
plt.title("Poisson(λ=4)")
plt.savefig("poisson.png")
```

Salida esperada: $P(X=0) \approx 0,544$; $P(X \geq 2) \approx 0,120$; $P(Y=2) \approx 0,147$.

Ejercicios 5

1. Demuestra que $E[X] = np$ para la binomial usando que X es suma de n Bernoulli.
2. Un examen tipo test tiene 10 preguntas con 4 opciones. Si se responde al azar, ¿ $P(\text{aprobar con } \geq 5 \text{ aciertos})$?
3. En una autovía hay de media 0,5 accidentes al día. ¿Probabilidad de que en una semana no haya ninguno?

Capítulo 6 · Distribuciones continuas (nivel 50)

Una variable continua puede tomar cualquier valor de un intervalo, así que $P(X = x) = 0$ para cada punto. Se describe con una **función de densidad** $f(x) \geq 0$ cuya área total es 1; las probabilidades son áreas.

$$P(a \leq X \leq b) = \int_a^b f(x) dx \quad E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

6.1 Uniforme y exponencial

- **Uniforme $U(a, b)$:** $f(x) = 1/(b-a)$ en $[a, b]$. $E[X] = (a+b)/2$; $\text{Var} = (b-a)^2/12$.
- **Exponencial(λ):** $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Modela tiempos de espera entre sucesos de Poisson. $E[X] = 1/\lambda$; $\text{Var} = 1/\lambda^2$. Es la única continua **sin memoria**: $P(X > s+t \mid X > s) = P(X > t)$.

6.2 La distribución normal

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$X \sim N(\mu, \sigma^2)$ es simétrica en torno a μ , con forma de campana. Toda normal se **tipifica** a la normal estándar $Z \sim N(0, 1)$:

$$Z = \frac{X - \mu}{\sigma}$$

Valores de referencia de $\Phi(z) = P(Z \leq z)$:

z	1,00	1,28	1,645	1,96	2,33	2,576
$\Phi(z)$	0,8413	0,9000	0,9500	0,9750	0,9900	0,9950

Una suma de normales independientes es normal: si $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$, $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

6.3 Distribuciones derivadas de la normal

Son las herramientas de la inferencia de los capítulos 8 y 9.

Distribución	Construcción	Uso principal
Chi-cuadrado χ^2_k	Suma de k normales estándar al cuadrado	Varianzas, bondad de ajuste, independencia
t de Student t_k	$Z / \sqrt{(\chi^2_k / k)}$	Medias con σ desconocida
F de Snedecor $F_{(k_1, k_2)}$	$(\chi^2_{k_1}/k_1) / (\chi^2_{k_2}/k_2)$	Comparar varianzas, ANOVA

La t se parece a la normal con colas más gruesas; cuando $k \rightarrow \infty$, $t_k \rightarrow N(0, 1)$.

Ejemplo resuelto 6.1

La resistencia de un hormigón sigue $N(30; 2^2)$ MPa. ¿Qué proporción no alcanza 27 MPa?

$z = (27 - 30)/2 = -1,5$. $P(Z < -1,5) = 1 - \Phi(1,5) = 1 - 0,9332 = 0,0668$. Un 6,7 % de las probetas falla.

En Python

Probabilidades, cuantiles (`ppf`) y la comparación visual entre la normal y la t.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats

hormigon = stats.norm(loc=30, scale=2)
print("P(R < 27):", hormigon.cdf(27).round(4))
print("Percentil 5:", hormigon.ppf(0.05).round(2))

print("z al 97,5 %:", stats.norm.ppf(0.975).round(3))
print("t(15) al 97,5 %:", stats.t.ppf(0.975, df=15).round(3))

espera = stats.expon(scale=200) # media 200 h
print("P(no falla en 300 h):", espera.sf(300).round(3))

x = np.linspace(-4, 4, 400)
plt.plot(x, stats.norm.pdf(x), label="N(0,1)")
plt.plot(x, stats.t.pdf(x, df=3), label="t(3)")
plt.legend(); plt.savefig("normal_t.png")
```

Salida esperada: $P(R < 27) \approx 0,0668$; $z = 1,96$; $t_{15} = 2,131$; $P \approx 0,223$.

Ejercicios 6

1. $X \sim N(100, 15^2)$. Calcula $P(85 < X < 130)$ y el percentil 90.
2. El tiempo entre averías es Exponencial con media 200 h. ¿ $P(\text{no falla en 300 h})$?
3. Si $X \sim U(0, 10)$, calcula $P(X > 7)$ y la varianza.

Capítulo 7 · Muestreo y teorema central del límite (nivel 60)

Este capítulo es la bisagra del libro: explica por qué una muestra de 1.000 personas basta para hablar de 47 millones. La clave es que los estadísticos muestrales son variables aleatorias con una distribución conocida.

7.1 Tipos de muestreo

- **Aleatorio simple:** todos los individuos tienen la misma probabilidad de salir. Es el supuesto de casi todas las fórmulas.

- **Estratificado:** se divide la población en grupos homogéneos (provincias) y se muestrea en cada uno. Reduce la varianza.
- **Por conglomerados:** se eligen grupos enteros al azar (obras completas). Más barato, menos preciso.
- **Sistemático:** uno de cada k individuos de una lista.

7.2 Distribución muestral de la media

Si $X_1, \dots, X_n$ son independientes con media μ y varianza σ^2 :

$$E[\bar{X}] = \mu \quad \text{y} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad \text{y} \quad \text{EE}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

El **error estándar** (EE) mide cuánto varía $\bar{x}$ de una muestra a otra. Decrece con $\sqrt{n}$: para reducir el error a la mitad hay que cuadruplicar la muestra.

7.3 Ley de los grandes números

Cuando $n \rightarrow \infty$, $\bar{X}$ converge en probabilidad a μ . Es la justificación de que las frecuencias relativas se estabilicen en torno a la probabilidad (lanzar una moneda 10.000 veces da casi un 50 % de caras).

7.4 Teorema central del límite (TCL)

Sea cual sea la distribución de las X_i (con varianza finita), para n suficientemente grande:

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\text{d}} N(0,1)$$

En la práctica, $n \geq 30$ suele bastar; si la población es muy asimétrica, se necesita más. El TCL explica por qué la normal aparece en todas partes: muchos fenómenos son suma de pequeños efectos independientes.

Caso particular para proporciones: si $np \geq 5$ y $n(1-p) \geq 5$,

$$\hat{p} \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

7.5 Distribución de la varianza muestral

Si la población es normal: $(n-1)S^2/\sigma^2 \sim \chi^2_{(n-1)}$, y además $\bar{X}$ y S^2 son independientes. De aquí sale el estadístico t:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Ejemplo resuelto 7.1

El gasto mensual en materiales por obra tiene $\mu = 12.000$ € y $\sigma = 3.000$ €, con distribución asimétrica. Se toman 36 obras. $EE = 3.000/6 = 500$ €. Por el TCL, $P(\bar{x} > 13.000) \approx P(Z > 2) = 0,0228$.

En Python

Simulación del TCL: aunque la población sea muy asimétrica (exponencial), la media de 36 datos sale casi normal y su error estándar coincide con $\sigma/\sqrt{n}$.

```
import numpy as np
import matplotlib.pyplot as plt

rng = np.random.default_rng(0)
# Población muy asimétrica: exponencial con media 12.000 €
n, repeticiones = 36, 10_000
medias = rng.exponential(scale=12_000, size=(repeticiones, n)).mean(axis=1)

print("Media de las medias:", medias.mean().round(0))
print("Error estándar empírico:", medias.std().round(0))
print("Error estándar teórico:", round(12_000 / np.sqrt(n)))

plt.hist(medias, bins=50, density=True)
plt.title("TCL: distribución de  $\bar{x}$  con  $n = 36$ ")
plt.savefig("tcl.png")
```

Salida esperada: Error estándar empírico ≈ 2.030 frente al teórico 2.000.

Ejercicios 7

1. ¿Qué tamaño de muestra hace falta para que el EE de la media sea $1/5$ de σ ?
2. El 40 % de los clientes prefiere el plan A. En una muestra de 200, ¿ $P(\hat{p} > 0,45)$?
3. Explica con tus palabras la diferencia entre la ley de los grandes números y el TCL.

Capítulo 8 · Estimación (nivel 70)

Estimar es usar la muestra para dar un valor (estimación puntual) o un rango (intervalo de confianza) del parámetro desconocido, junto con una medida de su precisión.

8.1 Propiedades de un buen estimador

Un estimador $\hat{\theta}$ de θ es:

- **Ins sesgado** si $E[\hat{\theta}] = \theta$. Sesgo = $E[\hat{\theta}] - \theta$.

- **Eficiente** si tiene la menor varianza entre los insesgados. La cota de Cramér-Rao fija ese mínimo: $\text{Var}(\hat{\theta}) \geq 1/(n \cdot I(\theta))$, con $I(\theta)$ la información de Fisher.
- **Consistente** si $\hat{\theta} \rightarrow \theta$ cuando $n \rightarrow \infty$.

El criterio global es el **error cuadrático medio**: $\text{ECM} = \text{Var}(\hat{\theta}) + \text{Sesgo}^2$. A veces un estimador algo sesgado tiene menor ECM.

Demostración de la corrección de Bessel: $E[\sum (X_i - \bar{X})^2] = \sum E[X_i^2] - n E[\bar{X}^2] = n(\sigma^2 + \mu^2) - n(\sigma^2/n + \mu^2) = (n-1)\sigma^2$. Por eso se divide entre $n-1$.

8.2 Método de máxima verosimilitud (MV)

La **verosimilitud** $L(\theta) = \prod f(x_i; \theta)$ mide lo compatibles que son los datos con cada valor de θ . El estimador MV la maximiza; en la práctica se maximiza el logaritmo:

$$\ell(\theta) = \sum_{i=1}^n \ln f(x_i; \theta), \quad \frac{d\ell}{d\theta} \Big|_{\hat{\theta}} = 0$$

Ejemplo (Poisson): $\ell(\lambda) = \sum x_i \ln \lambda - n\lambda - \sum \ln x_i!$. Derivando: $\sum x_i / \lambda - n = 0 \rightarrow \hat{\lambda} = \bar{x}$.

Los estimadores MV son consistentes, asintóticamente normales y asintóticamente eficientes. También existe el **método de los momentos**: igualar momentos muestrales y teóricos.

8.3 Intervalos de confianza

Un intervalo de confianza al 95 % es un procedimiento que, repetido en muchas muestras, captura el parámetro verdadero el 95 % de las veces. No significa «hay un 95 % de probabilidad de que μ esté en este intervalo concreto».

Parámetro	Supuestos	Intervalo
Media μ	σ conocida o n grande	$\bar{x} \pm z_{(\alpha/2)} \cdot \sigma/\sqrt{n}$
Media μ	Normal, σ desconocida	$\bar{x} \pm t_{(n-1, \alpha/2)} \cdot s/\sqrt{n}$
Proporción p	$n\hat{p} \geq 5$ y $n(1-\hat{p}) \geq 5$	$\hat{p} \pm z_{(\alpha/2)} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$
Varianza σ^2	Normal	$[(n-1)s^2/\chi^2_{(\alpha/2)}, (n-1)s^2/\chi^2_{(1-\alpha/2)}]$
Diferencia $\mu_1 - \mu_2$	Normales, varianzas iguales	$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2)} \cdot s_p \cdot \sqrt{1/n_1 + 1/n_2}$

Tamaño muestral para estimar una media con margen de error E : $n = (z_{(\alpha/2)} \cdot \sigma / E)^2$. Para una proporción, en el peor caso $p = 0,5$: $n = z^2 \cdot 0,25 / E^2$. Con 95 % y $E = 3\%$, $n \approx 1.068$.

Ejemplo resuelto 8.1

Se miden 16 vigas: $\bar{x} = 412$ kN, $s = 20$ kN. IC 95 % para μ con $t_{15}; 0,025 = 2,131$:

$$412 \pm 2,131 \cdot 20 / 4 = 412 \pm 10,7 \rightarrow [401,3; 422,7] \text{ kN.}$$

En Python

Intervalos de confianza, ajuste por máxima verosimilitud y tamaño muestral.

```
import numpy as np
from scipy import stats

# IC 95 % para la media (vigas, ejemplo 8.1)
n, media, s = 16, 412, 20
ic = stats.t.interval(0.95, df=n - 1, loc=media, scale=s / np.sqrt(n))
print("IC 95 %  $\mu$ :", np.round(ic, 1))

# IC 99 % para una proporción (320 de 500)
p = 320 / 500
ic_p = stats.norm.interval(0.99, loc=p, scale=np.sqrt(p * (1 - p) / 500))
print("IC 99 % p:", np.round(ic_p, 3))

# Máxima verosimilitud numérica: exponencial
rng = np.random.default_rng(1)
datos = rng.exponential(scale=5, size=200)
loc, escala = stats.expon.fit(datos, floc=0)
print(" $\lambda$  MV =", round(1 / escala, 3), "(1/ $\bar{x}$  =", round(1 / datos.mean(), 3),
      ")")

# Tamaño muestral
print("n =", int(np.ceil((1.96 * 10 / 2) ** 2)))
```

Salida esperada: IC μ : [401,3; 422,7]; IC p: [0,585; 0,695]; n = 97.

Ejercicios 8

1. Halla el estimador MV del parámetro λ de una exponencial.
2. En una encuesta a 500 clientes, 320 están satisfechos. Construye un IC 99 % para la proporción.
3. ¿Cuántas observaciones hacen falta para estimar μ con margen ± 2 , si $\sigma = 10$ y confianza 95 %?

Capítulo 9 · Contraste de hipótesis (nivel 80)

Un contraste decide si los datos aportan evidencia suficiente contra una afirmación inicial. Funciona como un juicio: se presume la inocencia (H_0) y solo se condena (se rechaza H_0) si las pruebas son muy improbables bajo esa presunción.

9.1 El procedimiento en cinco pasos

1. Plantear H_0 (hipótesis nula, «no hay efecto») y H_1 (alternativa), unilateral o bilateral.
2. Fijar el **nivel de significación** α (habitualmente 0,05).
3. Elegir el estadístico de contraste y su distribución bajo H_0 .
4. Calcular el **p-valor**: la probabilidad, si H_0 fuera cierta, de obtener un resultado al menos tan extremo como el observado.
5. Decidir: si $p \leq \alpha$ se rechaza H_0 ; si no, no se rechaza (que no es lo mismo que «aceptarla»).

9.2 Errores y potencia

	H_0 cierta	H_0 falsa
No rechazar H_0	Decisión correcta ($1 - \alpha$)	Error tipo II (β)
Rechazar H_0	Error tipo I (α)	Decisión correcta: potencia ($1 - \beta$)

La potencia aumenta con el tamaño muestral, con el tamaño del efecto y con α . Antes de un estudio se calcula n para alcanzar una potencia del 80 % o más.

9.3 Tests más usados

Pregunta	Test	Estadístico	Distribución bajo H_0
$\mu = \mu_0$? (σ desconocida)	t de una muestra	$t = (\bar{x} - \mu_0)/(s/\sqrt{n})$	$t_{(n-1)}$
$p = p_0$?	z de proporción	$z = (\hat{p} - p_0)/\sqrt{p_0(1-p_0)/n}$	$N(0,1)$
$\mu_1 = \mu_2$? independientes	t de dos muestras (Welch)	$(\bar{x}_1 - \bar{x}_2)/\sqrt{(s_1^2/n_1 + s_2^2/n_2)}$	t aprox.
¿Antes = después?	t pareada	t sobre las diferencias d_i	$t_{(n-1)}$

Pregunta	Test	Estadístico	Distribución bajo H_0
¿Ajusta a un modelo?	χ^2 de bondad de ajuste	$\Sigma(O - E)^2/E$	$\chi^2_{(k-1-m)}$
¿Dos variables independientes?	χ^2 de independencia	$\Sigma(O - E)^2/E$ en tabla $r \times c$	$\chi^2_{((r-1)(c-1))}$
¿Igualdad de k medias?	ANOVA de un factor	$F = CM_{\text{entre}} / CM_{\text{dentro}}$	$F_{(k-1, n-k)}$

En el χ^2 , $E = (\text{total fila} \times \text{total columna})/n$, y cada E debe ser ≥ 5 .

9.4 Comparaciones múltiples y buenas prácticas

- Con 20 tests a $\alpha = 0,05$, se espera un falso positivo aunque no haya ningún efecto. La **corrección de Bonferroni** usa α/m .
- Un p-valor pequeño no mide la importancia del efecto: informa también del **tamaño del efecto** (d de Cohen = $(\bar{x}_1 - \bar{x}_2)/s$) y del intervalo de confianza.
- Decidir el análisis antes de ver los datos evita el *p-hacking*.

Ejemplo resuelto 9.1

Un proveedor garantiza cemento con resistencia media de 42,5 MPa. Una muestra de 25 sacos da $\bar{x} = 41,6$ y $s = 2,0$. ¿Hay evidencia de que la media es menor? ($\alpha = 0,05$)

$H_0: \mu = 42,5$; $H_1: \mu < 42,5$. $t = (41,6 - 42,5)/(2/5) = -2,25$. Valor crítico $t_{24;0,05} = -1,711$. Como $-2,25 < -1,711$ ($p \approx 0,017$), se rechaza H_0 : hay evidencia de incumplimiento.

En Python

Contraste t a partir de estadísticos resumidos, test binomial exacto, χ^2 de independencia y ANOVA.

```
import numpy as np
from scipy import stats

rng = np.random.default_rng(7)
# Ejemplo 9.1: ¿la resistencia media es menor que 42,5 MPa?
n, media, s, mu0 = 25, 41.6, 2.0, 42.5
t = (media - mu0) / (s / np.sqrt(n))
p = stats.t.cdf(t, df=n - 1) # unilateral izquierda
print(f"t = {t:.2f}, p = {p:.4f}")
# Con los datos brutos en un array usarías:
# stats.ttest_1samp(datos, popmean=42.5, alternative="less")
```

```
# ¿Moneda trucada? 62 caras en 100
print("p binomial exacto:", stats.binomtest(62, 100, 0.5).pvalue.round(4))

#  $\chi^2$  de independencia
tabla = np.array([[30, 20], [15, 35]])
chi2, p, gl, esperados = stats.chi2_contingency(tabla)
print(f" $\chi^2$  = {chi2:.2f}, gl = {gl}, p = {p:.4f}")

# ANOVA de un factor (tres métodos de curado)
a, b, c = (rng.normal(m, 2, 5) for m in (30, 32, 35))
print("ANOVA:", stats.f_oneway(a, b, c))
```

Salida esperada: $t = -2,25$, $p \approx 0,017$; binomial exacto $p \approx 0,021$.

Ejercicios 9

1. Una moneda sale cara 62 veces en 100 lanzamientos. ¿Está trucada? ($\alpha = 0,05$)
2. Explica por qué «no rechazar H_0 » no demuestra que H_0 sea cierta.
3. Tres métodos de curado dan resistencias medias de 30, 32 y 35 MPa (5 probetas cada uno). ¿Qué test usarías y qué supuestos necesita?

Capítulo 10 · Correlación y regresión (nivel 90)

La regresión modela cómo cambia una variable respuesta Y cuando cambian una o varias variables explicativas X . Es la herramienta estadística más usada en ingeniería, economía y ciencia de datos.

10.1 Covarianza y correlación

$$s_{xy} = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y}) \quad r = \frac{s_{xy}}{s_x s_y}$$

- $r \in [-1, 1]$. $r = \pm 1$: relación lineal perfecta; $r \approx 0$: no hay relación **lineal** (puede haberla curva).
- Correlación no implica causalidad: helados y ahogamientos correlacionan porque ambos dependen del calor (variable de confusión).
- Test de significación: $t = r\sqrt{(n-2)/\sqrt{1-r^2}} \sim t_{(n-2)}$ bajo $H_0: \rho = 0$.
- **Correlación de Spearman**: r calculado sobre los rangos; robusta y válida para relaciones monótonas.

10.2 Regresión lineal simple

Modelo: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, con $\varepsilon_i \sim N(0, \sigma^2)$ independientes. Los **mínimos cuadrados** minimizan $\sum (y_i - \hat{y}_i)^2$:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} = r \frac{s_y}{s_x} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\hat{\beta}_1$ es el cambio medio en Y por cada unidad más de X. La recta siempre pasa por $(\bar{x}, \bar{y})$.

Coeficiente de determinación: $R^2 = SCR/SCT = 1 - SCE/SCT$, la proporción de variabilidad de Y explicada por X. En regresión simple, $R^2 = r^2$.

Inferencia sobre β_1 : $t = \hat{\beta}_1 / EE(\hat{\beta}_1) \sim t_{(n-2)}$, con $EE(\hat{\beta}_1) = s_e / \sqrt{\sum (x_i - \bar{x})^2}$ y $s_e^2 = SCE/(n-2)$.

10.3 Diagnóstico de residuos

Los supuestos se comprueban con los residuos $e_i = y_i - \hat{y}_i$:

1. **Linealidad:** el gráfico residuos vs. ajustados no muestra curvatura.
2. **Homocedasticidad:** la dispersión de residuos es constante (sin forma de embudo).
3. **Normalidad:** el gráfico Q-Q de residuos sigue una recta.
4. **Independencia:** sin patrones temporales (test de Durbin-Watson).

Atención a los **puntos influyentes** (distancia de Cook $> 4/n$): un solo dato puede mover la recta.

10.4 Regresión lineal múltiple

En forma matricial, $Y = X\beta + \varepsilon$, y la solución de mínimos cuadrados es:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- Cada β_j se interpreta «manteniendo constantes las demás variables».
- **R^2 ajustado** $= 1 - (1 - R^2)(n - 1)/(n - p - 1)$ penaliza añadir variables inútiles.
- **Multicolinealidad:** si las X están muy correlacionadas, los $\hat{\beta}$ se vuelven inestables. Se detecta con el VIF (> 10 es preocupante).
- Las variables cualitativas entran como **variables ficticias** (0/1).
- Si Y es binaria (sí/no), se usa **regresión logística**: $\log(p/(1-p)) = X\beta$.

Ejemplo resuelto 10.1

Superficie (x , m²) y coste (y , miles €) de 5 reformas: (50, 30), (70, 41), (80, 45), (100, 58), (120, 66). $\bar{x} = 84$, $\bar{y} = 48$. $S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = 1.530$; $S_{xx} = 2.920$.

$\hat{\beta}_1 = 1.530/2.920 \approx 0,524$ miles €/m²; $\hat{\beta}_0 = 48 - 0,524 \cdot 84 \approx 4,0$. Recta: $\hat{y} = 4,0 + 0,524x$. Para 90 m², $\hat{y} \approx 51,1$ mil €. Aquí $r \approx 0,997$: el ajuste es casi perfecto.

En Python

Regresión por mínimos cuadrados con `statsmodels`: el resumen da $\hat{\beta}$, errores estándar, t , p -valores y R^2 .

```
import numpy as np
import matplotlib.pyplot as plt
import statsmodels.api as sm

superficie = np.array([50, 70, 80, 100, 120])
coste = np.array([30, 41, 45, 58, 66])

X = sm.add_constant(superficie)
modelo = sm.OLS(coste, X).fit()
print(modelo.summary())
print("Predicción 90 m²:", modelo.predict([[1, 90]]).round(1))
print("r =", np.corrcoef(superficie, coste)[0, 1].round(3))

plt.scatter(superficie, coste)
plt.plot(superficie, modelo.fittedvalues)
plt.xlabel("m²"); plt.ylabel("miles €")
plt.savefig("regresion.png")
```

Salida esperada: $\hat{\beta}_0 \approx 3,99$; $\hat{\beta}_1 \approx 0,524$; $R^2 \approx 0,995$; predicción 51,1.

Ejercicios 10

1. Calcula R^2 del ejemplo 10.1 y su interpretación.
2. Un modelo predice el precio de una vivienda con $\hat{\beta} = 2.300$ €/m² (EE = 400). ¿Es significativo al 5 %?
3. ¿Por qué no se debe usar la recta del ejemplo 10.1 para predecir el coste de una reforma de 400 m²?

Capítulo 11 · Temas avanzados (nivel 100)

El último peldaño amplía la caja de herramientas: otra filosofía de inferencia (Bayes), métodos que no exigen normalidad y técnicas para muchas variables a la vez.

11.1 Estadística bayesiana

En el enfoque bayesiano el parámetro θ es una variable aleatoria que expresa nuestra incertidumbre. Se combina el conocimiento previo con los datos:

$$\underbrace{\pi(\theta \mid x)}_{\text{a posteriori}} \propto \underbrace{L(x \mid \theta)}_{\text{verosimilitud}} \underbrace{\pi(\theta)}_{\text{a priori}}$$

Prioris conjugadas: cuando la posterior es de la misma familia que la priori, el cálculo es cerrado.

Verosimilitud	Priori	Posterior
Binomial (k éxitos en n)	Beta(a, b)	Beta(a + k, b + n - k)
Poisson ($\sum x_i$ en n periodos)	Gamma(α, β)	Gamma($\alpha + \sum x_i, \beta + n$)
Normal (σ conocida)	Normal(μ_0, τ_0^2)	Normal con media ponderada por precisiones

Ejemplo: con priori Beta(1, 1) (uniforme) y 7 éxitos en 10 ensayos, la posterior es Beta(8, 4), de media $8/12 \approx 0,667$. Un **intervalo de credibilidad** del 95 % sí admite la lectura «hay un 95 % de probabilidad de que θ esté aquí».

Cuando no hay conjugación se usan métodos **MCMC** (Metropolis-Hastings, muestreo de Gibbs, Hamiltonian Monte Carlo) para simular la posterior. La comparación de modelos se hace con el **factor de Bayes** o criterios como WAIC.

11.2 Métodos no paramétricos

Se usan cuando no se cumple la normalidad o los datos son ordinales. Trabajan con rangos.

Situación	Test paramétrico	Alternativa no paramétrica
Dos grupos independientes	t de Student	U de Mann-Whitney
Dos medidas pareadas	t pareada	Rangos con signo de Wilcoxon
k grupos independientes	ANOVA	Kruskal-Wallis
k medidas repetidas	ANOVA de medidas repetidas	Friedman

Situación	Test paramétrico	Alternativa no paramétrica
¿Sigue una distribución dada?	—	Kolmogórov-Smirnov, Shapiro-Wilk (normalidad)

11.3 Remuestreo: bootstrap y permutaciones

- **Bootstrap:** se extraen B muestras (p. ej. 10.000) con reemplazo de los datos originales, se calcula el estadístico en cada una y su distribución empírica da el error estándar e intervalos (método de percentiles). Funciona para medianas, ratios o cualquier estadístico complicado.
- **Test de permutaciones:** se barajan las etiquetas de grupo muchas veces; el p-valor es la proporción de permutaciones con un estadístico tan extremo como el observado. No necesita supuestos distribucionales.

11.4 Estadística multivariante

- **Análisis de componentes principales (ACP):** busca direcciones ortogonales de máxima varianza (autovectores de la matriz de covarianzas). Reduce p variables a unas pocas componentes que conservan la mayor parte de la información.
- **Análisis de conglomerados (clustering):** agrupa observaciones similares. K-medias minimiza la suma de distancias al centroide; el jerárquico produce un dendrograma.
- **Análisis discriminante:** clasifica observaciones en grupos conocidos a partir de varias variables.
- **Normal multivariante** $N_p(\mu, \Sigma)$: generaliza la campana a p dimensiones; la distancia de Mahalanobis $(x - \mu)^T \Sigma^{-1} (x - \mu)$ mide lo atípico de un punto.

11.5 Hacia dónde seguir

Con estas bases puedes abordar series temporales (ARIMA), modelos lineales generalizados, modelos mixtos, análisis de supervivencia y aprendizaje automático estadístico. Todos se apoyan en los mismos pilares: verosimilitud, distribuciones muestrales e incertidumbre cuantificada.

En Python

Posterior conjugada, bootstrap, Mann-Whitney y componentes principales.

```
import numpy as np
from scipy import stats
from sklearn.decomposition import PCA
```

```

rng = np.random.default_rng(3)

# Bayes conjugado: priori Beta(1,1), 7 éxitos en 10
post = stats.beta(1 + 7, 1 + 3)
print("Media posterior:", post.mean().round(3), "IC 95 %:",
      np.round(post.interval(0.95), 3))

# Bootstrap de la mediana
datos = rng.lognormal(mean=3, sigma=0.6, size=60)
boot = stats.bootstrap((datos,), np.median, n_resamples=10_000,
                       method="percentile", random_state=rng)
print("IC 95 % mediana:", np.round(boot.confidence_interval, 2))

# No paramétrico: Mann-Whitney
g1, g2 = rng.exponential(10, 30), rng.exponential(14, 30)
print(stats.mannwhitneyu(g1, g2))

# ACP sobre 5 variables correlacionadas
base = rng.normal(size=(200, 2))
X = np.column_stack([base, base @ rng.normal(size=(2, 3)) +
                    rng.normal(scale=0.2, size=(200, 3))])
acp = PCA().fit(X)
print("Varianza explicada:", acp.explained_variance_ratio_.round(3))

```

Salida esperada: Media posterior 0,667, IC [0,39; 0,89]; el ACP concentra $\approx 98\%$ de la varianza en 2 componentes.

Ejercicios 11

1. Priori Beta(2, 2) y 3 éxitos en 12 ensayos. Escribe la posterior y su media.
2. Describe paso a paso cómo construirías un IC bootstrap al 95 % para la mediana.
3. Tras un ACP, las dos primeras componentes explican el 85 % de la varianza. ¿Qué significa?

Apéndices

A. Formulario esencial

Concepto	Fórmula
Media	$\bar{x} = \sum x_i / n$
Varianza muestral	$s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$

Concepto	Fórmula
Puntuación típica	$z = (x - \mu) / \sigma$
Prob. condicionada	$P(A B) = P(A \cap B) / P(B)$
Bayes	$P(B A) = P(A B)P(B) / P(A)$
Binomial	$P(X=k) = C(n,k)p^k(1-p)^{n-k}$
Poisson	$P(X=k) = e^{(-\lambda)}\lambda^k / k!$
Error estándar de la media	$\sigma / \sqrt{n}$
IC para μ (σ desconocida)	$\bar{x} \pm t \cdot s / \sqrt{n}$
IC para p	$\hat{p} \pm z \cdot \sqrt{(\hat{p}(1-\hat{p})/n)}$
Tamaño muestral (media)	$n = (z\sigma/E)^2$
Pendiente de regresión	$\hat{\beta}_1 = s_{xy} / s_x^2$
Correlación	$r = s_{xy} / (s_x s_y)$

B. Soluciones de los ejercicios

Capítulo 1. (1) Código postal: cualitativa nominal (aunque use números); temperatura: cuantitativa continua; talla: cualitativa ordinal; nº de hijos: cuantitativa discreta. (2) Sesgo de autoselección: responden los usuarios activos y quienes tienen opiniones extremas. (3) 42 % de los 1.000 es un estadístico ($\hat{p}$); el porcentaje real en el electorado es el parámetro (p).

Capítulo 2. (1) $\bar{x} = 8,75$; $Me = 7,5$; $Mo = 7$. (2) Media armónica: $2/(1/60 + 1/40) = 48$ km/h, no 50. (3) Media geométrica: $(1,1 \cdot 0,8 \cdot 1,3)^{1/3} - 1 \approx 4,6$ % anual.

Capítulo 3. (1) $Q_1 = 6$, $Q_3 = 13$, $RIC = 7$; límite superior $13 + 10,5 = 23,5$, así que 40 es atípico. (2) $z = (8,5 - 6,2)/1,4 \approx 1,64$. (3) $CV_A = 1$ %, $CV_B = 1,6$ %: A es relativamente más precisa.

Capítulo 4. (1) $P(\text{suma } 7) = 6/36 = 1/6$; $P(\text{al menos un } 6) = 1 - (5/6)^2 = 11/36 \approx 0,306$. (2) $C(10,3) = 120$. (3) $P(B|D) = 0,4 \cdot 0,05 / (0,6 \cdot 0,02 + 0,4 \cdot 0,05) = 0,020/0,032 = 0,625$.

Capítulo 5. (1) $X = \sum Y_i$ con $Y_i \sim \text{Bernoulli}(p)$; $E[X] = \sum E[Y_i] = np$. (2) $X \sim \text{Bin}(10; 0,25)$: $P(X \geq 5) \approx 0,078$. (3) En 7 días $\lambda = 3,5$: $P(0) = e^{(-3,5)} \approx 0,030$.

Capítulo 6. (1) $P(-1 < Z < 2) \approx 0,819$; $P_{90} = 100 + 1,2816 \cdot 15 \approx 119,2$. (2) $\lambda = 1/200$: $P(X > 300) = e^{(-1,5)} \approx 0,223$. (3) $P(X > 7) = 0,3$; $\text{Var} = 100/12 \approx 8,33$.

Capítulo 7. (1) $\sigma/\sqrt{n} = \sigma/5 \Rightarrow n = 25$. (2) $EE = \sqrt{(0,4 \cdot 0,6/200)} \approx 0,0346$; $z \approx 1,44$; $P \approx 0,074$. (3) La LGN dice que $\bar{x}$ se acerca a μ ; el TCL describe la forma (normal) de las fluctuaciones de $\bar{x}$ alrededor de μ .

Capítulo 8. (1) $\ell(\lambda) = n \ln \lambda - \lambda \sum x_i \Rightarrow \hat{\lambda} = 1/\bar{x}$. (2) $\hat{p} = 0,64$; $0,64 \pm 2,576 \cdot 0,0215 \rightarrow [0,585; 0,695]$. (3) $n = (1,96 \cdot 10/2)^2 \approx 96,04 \rightarrow 97$ observaciones.

Capítulo 9. (1) $z = (0,62 - 0,5)/0,05 = 2,4$; p-valor bilateral $\approx 0,016 < 0,05$: hay evidencia de que está trucada. (2) Porque la falta de evidencia puede deberse a poca potencia (muestra pequeña), no a que no exista el efecto. (3) ANOVA de un factor; supone normalidad en cada grupo, varianzas iguales e independencia (si fallan: Kruskal-Wallis).

Capítulo 10. (1) $R^2 = r^2 \approx 0,995$: la superficie explica el 99,5 % de la variación del coste. (2) $t = 2.300/400 = 5,75$, muy superior a 1,96: significativo. (3) Sería una extrapolación: el modelo solo se ajustó entre 50 y 120 m² y nada garantiza la linealidad fuera de ese rango.

Capítulo 11. (1) $\text{Beta}(2+3, 2+9) = \text{Beta}(5, 11)$; media $5/16 = 0,3125$. (2) Remuestrear n datos con reemplazo B veces, calcular la mediana en cada réplica, ordenar y tomar los percentiles 2,5 y 97,5. (3) Dos variables sintéticas resumen el 85 % de la variabilidad total; se puede trabajar en 2D perdiendo solo el 15 %.

C. Glosario

Término	Definición breve
Atípico	Dato muy alejado del resto (regla 1,5·RIC)
Error estándar	Desviación típica de un estimador
Estadístico	Función de la muestra que no depende de parámetros desconocidos
Hipótesis nula	Afirmación de «no efecto» que se contrasta
Intervalo de confianza	Rango que captura el parámetro con una frecuencia fijada
Nivel de significación (α)	Probabilidad máxima aceptada de error tipo I
p-valor	Probabilidad bajo H_0 de un resultado tan extremo o más
Potencia	Probabilidad de rechazar H_0 cuando es falsa
Sesgo	Diferencia entre el valor esperado del estimador y el parámetro
Verosimilitud	Probabilidad de los datos vista como función del parámetro

Aprende estadística desde cero hasta el nivel universitario avanzado.

Este libro recorre la estadística como una escalera de 100 peldaños. Empieza contando datos y calculando medias, y termina estimando parámetros por máxima verosimilitud, contrastando hipótesis y actualizando creencias con el teorema de Bayes.

Cada capítulo incluye:

- Teoría explicada en lenguaje claro y con notación rigurosa
- Ejemplos resueltos paso a paso
- Ejercicios con soluciones al final del libro

Ideal para estudiantes, profesionales de datos, ingenieros y cualquier persona que quiera decidir mejor con números.

