

Sesión 1: ¿Qué es la estadística y por qué simular?

Escuela de Verano – CIMAT

Marco Antonio Aquino López¹

¹CIMAT
Probabilidad y Estadística.

5 July 2025



Objetivos de la sesión

- Comprender la estadística como herramienta para modelar la incertidumbre.
- Introducir el concepto de variable aleatoria y su rol en la modelación del azar.
- Explorar la simulación como estrategia para estudiar fenómenos aleatorios.
- Estimar propiedades como media, varianza y distribución empírica.
- Usar Python para generar datos aleatorios y visualizarlos.

¿Por qué necesitamos la estadística?


- El mundo real está lleno de incertidumbre: clima, mercados, biología, decisiones humanas.
- La estadística nos ayuda a razonar en contextos donde el azar y la variabilidad están presentes.
- No basta con observar: debemos inferir, comparar, predecir, y validar.

Idea central: La estadística es nuestro lenguaje para entender lo incierto.

¿De dónde viene la estadística?

- La palabra “estadística” proviene del latín **status**, originalmente referida a la descripción del Estado.
- Siglos XVII–XVIII: recopilación de datos sobre población, nacimientos y recursos fiscales.

Bills of Mortality (1662)

<i>The Diseases and Casualties this Week.</i>			
			
A Botherie	4	Leppithum	18
Aged	45	Infants	22
Breeding	1	Kinglevil	4
Broken legges	1	Lethargy	1
Broke her skull by a fall in the street at St. Mary VVoolchurch	1	Livergrown	1
Children	38	Mangeome	1
Chilfoes	9	Pallie	1
Chilfomes	9	Plague	4237
Consumption	126	Purples	2
Convulsion	89	Quinine	5
Cough	1	Rising of the Lights	23
Droyn	53	Rupture	18
Feaver	248	Scurvy	1
Flox and Small-pox	11	Shingles	3
Fluk	1	Spotted Feaver	1
Frighted	2	Stilborn	166
Gout	1	Stone	4
Grief	3	Stopping of the Stomach	2
Gripping in the Guts	79	Serangury	17
Head-mould-shot	1	Suddenly	3
Jaundies	7	Surfie	2
		Teeth	74
		Thrush	111
		Tiffick	9
		Ulcer	1
		Vomiting	10
		Wende	4
		Wormes	20
Males 90		Males 2777	Plague-4237
Christened Females 81		Baried Females 2791	
In all 171		In all 5568	
Increased in the Burials this Week		249	
Parities clear of the Plague		27	Parities Infected 103
<i>For Advice of Bread set forth by Order of the Lord Mayor and Councill of Aldermen, A penny Wheaten Loaf to contain Nine Ounces and a half, and three half-penny White Loaves the like weight.</i>			

"By examining the weekly mortality data, I seek patterns that help us understand the city's health."

Nace la teoría de probabilidades

- Siglo XVIII: se formaliza la probabilidad como rama matemática.
- **Pierre-Simon Laplace** (1749–1827): introduce el *Teorema de Bayes* en forma general.
- **Carl Friedrich Gauss** (1777–1855): propone el método de mínimos cuadrados y la distribución normal.



Francis Galton

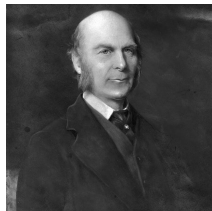


Karl Pearson

(Laplace y Gauss: pilares de la estadística matemática)

Del conteo a la inferencia moderna

- Siglo XIX: surge el concepto de *regresión hacia la media* (**Francis Galton**) y la **correlación** (**Karl Pearson**).
- La estadística comienza a abordar preguntas causales, no solo descriptivas.



Francis Galton



Karl Pearson

(Galton y Pearson: de la biología a la estadística)

Revolución del siglo XX

- **Ronald A. Fisher** establece los fundamentos del diseño experimental y la inferencia estadística.
- Se desarrollan dos paradigmas:
 - ▶ **Frecuentista**: Neyman, Pearson y Wald.
 - ▶ **Bayesiano**: Jeffreys, Savage y De Finetti.
- Hoy: la estadística es núcleo de disciplinas como epidemiología, inteligencia artificial y ciencias sociales.

"La estadística es la gramática de la ciencia." — Karl Pearson

¿Qué papel juegan los datos?

- No observamos fenómenos directamente: observamos **datos**.
- Los datos son muestras generadas por un **proceso aleatorio subyacente**.
- La estadística busca entender ese proceso.

¿Cómo lo hacemos? → Definiendo modelos con variables aleatorias.

¿Cómo modelamos el azar?

Para estudiar fenómenos aleatorios con rigor, usamos un modelo matemático basado en:

Espacio de probabilidad

$$(\Omega, \mathcal{F}, \mathbb{P})$$

- Ω : espacio muestral (todos los posibles resultados)
- \mathcal{F} : sigma-álgebra (eventos observables)
- \mathbb{P} : medida de probabilidad

Este modelo formaliza nuestra incertidumbre sobre el fenómeno que estamos estudiando.

¿Qué es una variable aleatoria?

Definición

Una variable aleatoria es una función:

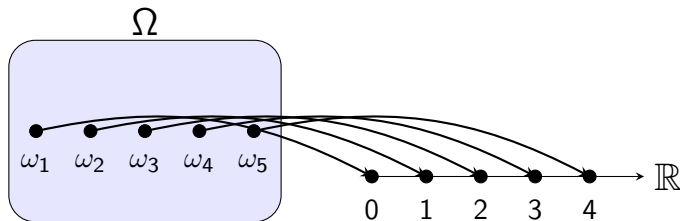
$$X : \Omega \rightarrow \mathbb{R}$$

que asocia un número real a cada resultado posible del experimento aleatorio.

- Es medible: nos permite asignar probabilidades a eventos del tipo $X \in B$
- Transforma el espacio abstracto Ω en una variable cuantitativa con la que podemos trabajar.

Visualizando el modelo

$$X : \Omega \rightarrow \mathbb{R}$$



Ejemplo: lanzamiento de una moneda

Modelo simple:

- $\Omega = \{\text{Cara}, \text{Cruz}\}$
- $\mathcal{F} = \mathcal{P}(\Omega)$
- $\mathbb{P}(\text{Cara}) = \mathbb{P}(\text{Cruz}) = 0.5$

Definimos:

$$X(\omega) = \begin{cases} 1, & \omega = \text{Cara} \\ 0, & \omega = \text{Cruz} \end{cases}$$

Entonces X es una variable aleatoria de Bernoulli.

¿Por qué usar esta estructura formal?

- Permite modelar fenómenos aleatorios con precisión matemática.
- Da lugar a cálculos de probabilidades, medias, varianzas, distribuciones.
- Es la base para construir modelos más sofisticados (regresión, procesos estocásticos, etc.).

Dato clave: Lo que observamos en el mundo real son realizaciones de variables aleatorias.

Probabilidad y Estadística: dos perspectivas

Probabilidad

Parte de un modelo (Ω, \mathbb{P}) y estudia sus propiedades.

Estadística

Parte de datos observados y busca inferir cómo es el modelo que los generó.

- Son dos caras de la misma moneda.
- Así como la física necesita matemáticas, la estadística necesita probabilidad.

¿Por qué simular?

Problema: A veces no podemos calcular exactamente la distribución, media o varianza de una variable aleatoria.

Solución: Usamos simulación computacional para generar datos artificiales que imitan el comportamiento de la variable.

- Se repite el experimento muchas veces.
- Se recolecta una muestra X_1, X_2, \dots, X_n .
- Se estima el comportamiento teórico de X .

¿Qué es una simulación?

Simular = Generar datos aleatorios controlados

- Podemos estudiar fenómenos complejos sin fórmulas explícitas.
- Es clave para:
 - ▶ Desarrollar intuición sobre probabilidades.
 - ▶ Visualizar fenómenos aleatorios.
 - ▶ Evaluar estimadores estadísticos.
- Forma la base de métodos como Monte Carlo, inferencia bayesiana, bootstrap y validación de modelos.

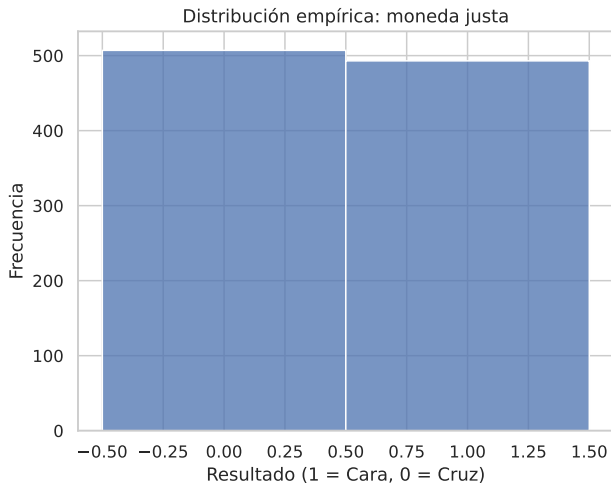
Ejemplo: simulando una moneda justa

Modelo: $X \sim \text{Bernoulli}(0.5)$

Podemos simular n lanzamientos de una moneda con Python: **Pregunta:** ¿Qué sucede si repetimos la simulación muchas veces?

Visualización: histograma de la moneda

Distribución empírica del experimento:



Ejemplo: simulando un dado justo

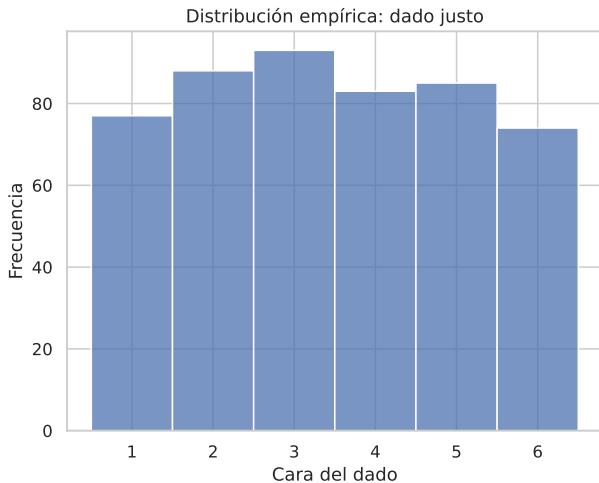
Modelo: $X \sim \text{Uniforme}\{1, 2, 3, 4, 5, 6\}$

Podemos simular n lanzamientos de un dado justo con Python:

Pregunta: ¿La frecuencia de cada cara se estabiliza si repetimos muchas veces el experimento?

Visualización: histograma de la dado

Distribución empírica del experimento:



Ventajas de simular

- Nos permite experimentar sin peligro y costos minimos costo.
- Podemos validar métodos cuando no conocemos la verdad.
- Es flexible: podemos simular muchos modelos distintos fácilmente.

Ejemplos:

- Simulación de epidemias, procesos genéticos o redes sociales.
- Métodos de bootstrap para estimar errores estándar.
- MCMC para inferencia bayesiana en modelos complejos.

Preguntas para discutir en clase

- ¿Qué tan “buena” es una simulación para aproximar la verdad?
- ¿Cómo cambia la media muestral con el tamaño de muestra?
- ¿Qué tan estable es la distribución empírica?
- ¿Qué pasaría si la moneda estuviera cargada?

Media y varianza de una variable aleatoria

Definiciones

Sea X una variable aleatoria discreta:

$$\mathbb{E}[X] = \sum_x x \cdot \mathbb{P}(X = x)$$

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Para una moneda justa: $\mathbb{E}[X] = 0.5$, $\text{Var}(X) = 0.25$
- En la práctica, usamos la muestra para estimar estas cantidades.

Estimación desde simulación

Dada una muestra simulada X_1, X_2, \dots, X_n , usamos:

- **Media muestral:**

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- **Varianza muestral:**

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Estas aproximan los valores esperados y dispersión de la variable aleatoria.

Código: estimar media y varianza

```
1 # Simular 1000 lanzamientos de una moneda justa
2 import numpy as np
3
4 n = 100
5 moneda = np.random.choice([0, 1], size=n)
6
7 media = np.mean(moneda)
8 varianza = np.var(moneda, ddof=1)
9
10 print(f"Media: {media:.3f}")
11 print(f"Varianza: {varianza:.3f}")
```

Nota: Puedes comparar con los valores teóricos 0.5 y 0.25.

Distribución empírica

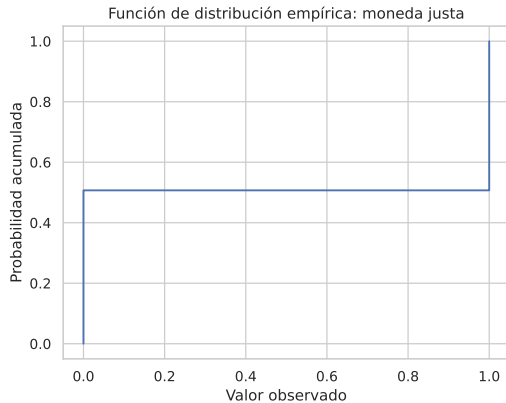
¿Qué es la distribución empírica?

Dada una muestra X_1, \dots, X_n , la función de distribución empírica (FDE) es:

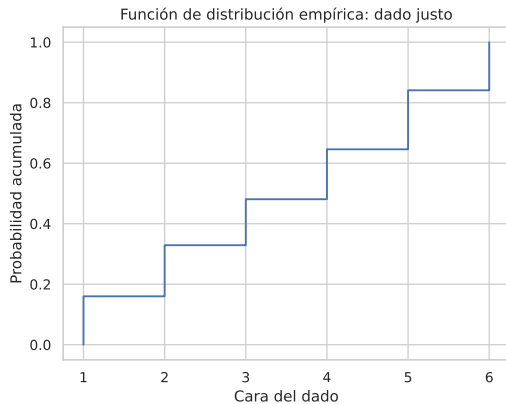
$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x)$$

- Representa la proporción acumulada de observaciones $\leq x$
- Es una estimación no paramétrica de la distribución verdadera
- Tiene saltos en los puntos de la muestra

función de distribución empírica



función de distribución empírica



Histograma vs FDE

- **Histograma:** muestra la frecuencia con que ocurren los valores.
- **FDE:** muestra cómo se acumulan los valores hacia la derecha.

Ambas herramientas son complementarias y ayudan a explorar los datos simulados.

¿Quieres que te genere el gráfico del histograma o de la FDE?

Ejemplo: lanzamiento de un dado justo

Modelo:

$$X \in \{1, 2, 3, 4, 5, 6\}, \quad \mathbb{P}(X = k) = \frac{1}{6}$$

Simulamos 500 lanzamientos y calculamos estadísticas:

```
1 n = 500 dado = np.random.randint(1, 7, size=n) media = np.mean(dado) varianza = np.var(dado, ddof=1)
2 print(f"Media: media:.3f") print(f"Varianza: varianza:.3f")
```

Histograma: dado justo

```
1 import matplotlib.pyplot as plt
2 import seaborn as sns
3
4 sns.histplot(dado, bins=np.arange(1,8)-0.5, discrete=True)
5 plt.title("Distribución empírica: dado justo")
6 plt.xlabel("Cara del dado")
7 plt.ylabel("Frecuencia")
8 plt.show()
```

Observación: Aunque el dado es justo, puede haber fluctuaciones por azar.

¿Y si el dado está cargado?

Supongamos que la cara 6 tiene el doble de probabilidad que las demás.

$$\mathbb{P}(X = 6) = 2p, \quad \mathbb{P}(X = k) = p, \text{ para } k = 1, \dots, 5$$

Como las probabilidades deben sumar 1:

$$5p + 2p = 1 \Rightarrow p = \frac{1}{7}$$

Usaremos simulación para estudiar cómo se ve este sesgo.

Simulación: dado cargado

```
1 caras = [1, 2, 3, 4, 5, 6]
2 pesos = [1, 1, 1, 1, 1, 2]
3 probabilidades = np.array(pesos) / sum(pesos)
4
5 dado_cargado = np.random.choice(caras, size=1000, p=probabilidades)
```

Pregunta: ¿Cómo cambia la distribución empírica comparada con el dado justo?

Histograma: dado cargado

```
1 sns.histplot(dado_cargado, bins=np.arange(1,8)-0.5, discrete=True)
2 plt.title("Distribución empírica: dado cargado")
3 plt.xlabel("Cara del dado")
4 plt.ylabel("Frecuencia")
5 plt.show()
```

¿Notas un sesgo hacia el 6? ¿Qué pasa si repites el experimento?

Comparación: dado justo vs cargado

- La distribución empírica del dado justo debe ser aproximadamente uniforme.
- En el dado cargado, la cara 6 aparece con más frecuencia.

Reflexión: ¿Cuántos datos necesitas para detectar el sesgo con claridad?

¿Qué pasa con la media cuando repetimos muchas veces?

Idea experimental: Si repetimos un experimento aleatorio muchas veces y calculamos la media acumulada...

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

- Al principio, puede variar mucho.
- Pero conforme aumenta n , parece estabilizarse cerca de un valor.
- ¿Ese valor refleja algo profundo del experimento?

Ejemplo: Moneda justa \rightarrow media observada tiende hacia 0.5 con más repeticiones.

Visualización: media acumulada

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 # Simular 10,000 lanzamientos
5 n = 10000
6 moneda = np.random.choice([0, 1], size=n)
7 medias = np.cumsum(moneda) / np.arange(1, n+1)
8
9 # Graficar
10 plt.plot(medias, label='Media acumulada')
11 plt.axhline(0.5, color='red', linestyle='--', label='Esperanza teórica')
12 plt.xlabel('Número de lanzamientos')
13 plt.ylabel('Media acumulada')
14 plt.legend()
15 plt.grid(True)
16 plt.title("Ley de los Grandes Números: moneda justa")
17 plt.show()
```

Interpretación de la gráfica

- Al inicio, la media muestral puede oscilar bastante.
- Conforme n crece, la media se estabiliza cerca de 0.5.
- Esta convergencia ilustra el comportamiento “promedio” de un fenómeno aleatorio.

¿Qué sucede si usamos una moneda cargada con $\mathbb{P}(X = 1) = 0.7$?

¿Qué aprendimos hoy?

- Modelamos el azar con variables aleatorias sobre espacios de probabilidad.
- Usamos simulaciones para estudiar fenómenos que no podemos resolver analíticamente.
- Estimamos media, varianza y distribución empírica.
- Visualizamos el comportamiento estadístico con gráficos y experimentos computacionales.

¡Todo esto es la base para construir inferencia estadística más adelante!

Preguntas abiertas para discusión

- ¿Qué limita la utilidad de una simulación?
- ¿En qué casos preferirías un modelo teórico versus uno simulado?
- ¿Cómo afectan el sesgo y la varianza a nuestras estimaciones?
- ¿Cómo puede ayudarnos una simulación a pensar estadísticamente?

Ejercicios para reflexionar (Sesión 1)

- 1 Simula 10 000 lanzamientos de una moneda con $p = 0.7$ y grafica la media acumulada. ¿Qué observas?
- 2 Modifica el código para probar con $p = 0.5$ y $p = 0.9$. ¿Cambia la velocidad de convergencia?
- 3 Justifica por qué la varianza de \bar{X}_n disminuye con n . ¿Qué papel juega esto en la estimación?

Sugerencia: Intenta modificar el código y discutir tus observaciones con tus compañeros.