Inferencia Bayesiana

Una Herramienta para el Análisis de Datos

Notas del curso, UNAM Agosto-Diciembre 2025

Autor: Imelda Trejo

25 de agosto de 2025

Resumen

Estas notas introducen los fundamentos de la inferencia bayesiana para estudiantes de maestría y doctorado en matemáticas del Centro de Ciencias Matemáticas, UNAM Morelia. Cubrimos desde los conceptos filosóficos fundamentales hasta las bases matemáticas rigurosas, incluyendo espacios muestrales, distribuciones de probabilidad, el teorema de Bayes y el concepto de intercambiabilidad.

Índice

1. Contexto Histórico

El término de inferencia Bayesiana es en honor a Thomas Bayes.

Figura 1: Esta es probablemente la únia imagen que se tiene de Thomas Bayes, imagen obtenida de https://en.wikipedia.org/wiki/Thomas_Bayes

Thomas Bayes (1701-1761) fue un matemático y estadístico inglés que formuló el **teorema de Bayes**, una regla fundamental en estadística y aprendizaje automático moderno. Su trabajo dio origen a toda un área de la estadística: la **estadística bayesiana**, que se centra en determinar la probabilidad de un evento basándose en el conocimiento previo de condiciones que puedan estar relacionadas con dicho evento. Esto es, para dos eventos A y B con, $\mathbb{P}(B) > 0$:

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

Aquí, el evento A no lo conocemos con precisión, pero creemos que este evento tiene una distribución $\mathbb{P}(A)$, que también se le conoce como *distribución a priori*. Luego, la probabilidad de A se actualiza al conocer la relación entre los eventos A y B. Siendo, $\mathbb{P}(A \mid B)$ la *distribución posterior*, distribución que resulta al incorporar los nuevos conocimientos sobre el evento A.

En nuestro contexto estimaremos el parámetro θ dado datos:

$$\mathbb{P}(\theta \mid Datos) = \frac{\mathbb{P}(Datos \mid \theta)\pi(\theta)}{\mathbb{P}(Datos)}$$

El teorema de Bayes proporciona un marco matemático para actualizar nuestras creencias (falta de conocimientos) sobre cantidades inciertas a medida que observamos nuevos datos—un proceso que refleja cómo progresa el conocimiento científico.

2. Introducción a la Estadística

2.1. ¿Qué es la estadística?

La estadística es la ciencia que **recopila, organiza, analiza e interpreta datos** para entender fenómenos y ayudar en la toma de decisiones. Podemos dividirla en dos ramas principales:

- **Estadística descriptiva**: Resume y describe los datos mediante tablas, gráficas y medidas como media, mediana y varianza.
- **Estadística inferencial**: Utiliza datos muestrales para hacer **inferencias** sobre una población o proceso subyacente.

2.2. Una Distinción Fundamental: Interpretación vs. Explicación

> *"La estadística debe considerarse una interpretación de los fenómenos naturales, más que una explicación."*

Esta afirmación destaca un concepto crucial:

- La **estadística no explica por qué ocurre algo**; en cambio, **resume lo que observamos** y **cuantifica la incertidumbre** en nuestras observaciones. La forma de cuantificar la incertidumbre es con probabilidades.
- La **estadística** nos dice: *"Dadas estas observaciones, así es como se comportan los datos, y esta es la probabilidad de que este patrón sea real."*
- La **explicación** (el "por qué") típicamente proviene de una **teoría científica o modelo causal**.

Ejemplo 2.1. Patrones de influenza estacional.

- **Fenómeno**: "Los casos de gripe aumentan durante el invierno."

- **Análisis estadístico**: Analiza datos históricos, calcula aumentos promedio, estima probabilidades de brotes.
- **Explicación científica**: "El virus sobrevive mejor a temperaturas frías, y las personas pasan más tiempo en espacios cerrados en proximidad."

2.3. Por qué importa esta distinción

Confundir interpretación estadística con explicación causal lleva a:

1. Creer que correlación implica causalidad
2. Aplicar incorrectamente resultados estadísticos fuera de su contexto científico
3. Ignorar que la estadística **depende de la calidad y representatividad de los datos**

Ejemplo 2.2 (Ejemplo clásico). Las ventas de helado y los incidentes de ahogamiento muestran correlación positiva.

- **Observación estadística**: Cuando aumentan las ventas de helado, también aumentan los ahogamientos.
- **Inferencia causal incorrecta**: El helado causa ahogamientos.
- **Explicación real**: El calor es la causa común–en verano, más gente compra helado y más gente va a nadar, resultando en más ahogamientos.

2.4. Incertidumbre en Estadística

La **incertidumbre** representa **lo que no sabemos con certeza** sobre un fenómeno o valor de parámetro. En modelado matemático y análisis de datos, la incertidumbre surge de:

1. **Errores de medición** en la recolección de datos
2. **Variabilidad natural** en los fenómenos
3. **Observaciones limitadas**—observamos muestras, no poblaciones completas
4. **Limitaciones muestrales**—muestras pequeñas o no representativas

5. **Inadecuación del modelo**—los modelos son simplificaciones de la realidad

Ruido en los datos es otra expresión usada para expresar incertidumbre en las mediciones de los datos.

2.4.1. Cuantificación de la Incertidumbre

La estadística **cuantifica** esta incertidumbre mediante diferentes enfoques:

- **Enfoque frecuentista**: Intervalos de confianza, valores p
- **Enfoque bayesiano**: Distribuciones posteriores, intervalos creíbles

2.5. El Paradigma Bayesiano

Principio bayesiano fundamental: Cualquier cantidad que no conocemos con certeza (ya sea un parámetro, resultado futuro, o proceso no observado) se representa mediante una **distribución de probabilidad**.

Esta representación probabilística nos permite:

- Expresar grados de conocimiento sobre cantidades desconocidas
- Actualizar estos conocimientos conforme nuevos datos se vuelven disponibles
- Hacer predicciones con incertidumbre cuantificada

Usando la ditribución de probabilidad es fácil calcular predicciones con intervalos de credibilidad, ver Figura 2.

2.6. Introducción al análisis de datos

Tarea: Leer el capítulo 1 del libro [5]

A continuación se presenta un resumen del capítulo.

2.6.1. Conceptos fundamentales

El análisis de datos es una disciplina que nos permite extraer información significativa de conjuntos de observaciones. Para comprender esta disciplina, es esencial dominar algunos conceptos básicos que forman la base de todo estudio estadístico.

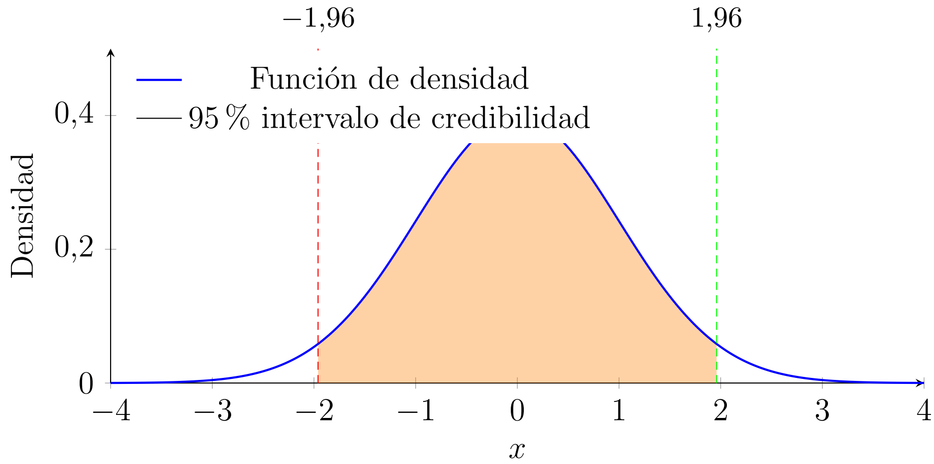

Figura 2: Distribucón Normal estándar: $\mu = 0$, $\sigma = 1$.

Definiciones básicas

Definición 2.1 (Población)**.** Una población es un conjunto de personas, objetos o eventos, de los cuales nos interesa estudiar algunas de sus características.

Definición 2.2 (Muestra)**.** Una muestra es cualquier subconjunto de una población.

Definición 2.3 (Variable)**.** Una variable es una característica de interés que posee cada elemento de una población y que podemos medir o registrar.

Una variable también puede considerarse como una pregunta que se le hace a cada elemento de la población, produciendo una respuesta en cada caso.

Ejemplo: En una población humana, podemos considerar la variable o pregunta: "¿Eres estudiante?" y obtener como respuesta "sí" o "no".

Definición 2.4 (Datos)**.** Mediante el término datos se entiende al conjunto de observaciones de una o varias variables de interés para todos los elementos de una muestra.

Generalmente, un conjunto de datos se organiza y almacena en una computadora en la forma de un arreglo bidimensional o tabla, donde las filas representan las observaciones (elementos de la muestra) y las columnas representan las variables.

Ejemplo de organización de datos Consideremos un estudio sobre estudiantes universitarios. La siguiente tabla muestra cómo se organizarían los datos:

ID	Edad	Sexo	Carrera	Promedio
1	20	Masculino	Ingeniería	8.5
2	19	Femenino	Medicina	9.2
3	21	Masculino	Derecho	7.8
4	22	Femenino	Psicología	8.9
5	20	Femenino	Ingeniería	8.1

Tabla 1: Ejemplo de organización de datos en forma tabular

2.6.2. Clasificación de variables

En el análisis de datos, es fundamental entender el tipo de variable con la que se trabaja, ya que esto determina qué métodos estadísticos son adecuados para su análisis. Las variables se clasifican principalmente en dos grandes categorías: cuantitativas y cualitativas.

Variables cuantitativas

Definición 1.5 (Variable cuantitativa): Una variable es cuantitativa si sus valores son números y representan una cantidad.

Las variables cuantitativas se subdividen en:

- **Discretas:** Toman valores enteros o contables. Ejemplos: número de hijos en una familia, número de automóviles vendidos, cantidad de libros en una biblioteca.
- **Continuas:** Pueden tomar cualquier valor dentro de un intervalo. Ejemplos: estatura, peso, longitud de un tornillo, temperatura.

Características de las variables cuantitativas:

- Permiten calcular promedios, varianzas y otras medidas estadísticas.
- Se pueden aplicar técnicas estadísticas (estimación de parámetros usando Poisson) paramétricas.
- Tienen sentido las operaciones aritméticas (suma, resta, multiplicación, división).
- Poseen un orden natural y distancias medibles entre valores.

Variables cualitativas (categóricas)

Definición 2.5 (Variable cualitativa)**.** Una variable es cualitativa si sus valores representan una cualidad, un atributo o una categoría. Se les llama también variables categóricas.

Las variables cualitativas se caracterizan por:

- Tomar valores en un conjunto limitado de categorías.
- Pueden o no tener un orden natural.
- Se utilizan para clasificar observaciones en grupos.

Las variables cualitativas se subdividen en:

- **Nominales:** No tienen un orden natural entre sus categorías.
- **Ordinales:** Sus categorías tienen un orden natural o jerárquico.
- **Binarias o dicotómicas:** Caso especial que toma exactamente dos valores.

Ejemplos de variables cualitativas:

- **Nominales:**
  - Grupo sanguíneo: A, B, AB, O
  - Estado civil: soltero, casado, divorciado, viudo
  - Color de ojos: azul, marrón, verde
  - Nacionalidad: mexicana, estadounidense, canadiense
- **Ordinales:**
  - Nivel educativo: primaria, secundaria, preparatoria, universidad
  - Calificación: excelente, bueno, regular, malo
  - Grupos de edad: joven, adulto, mayor
- **Binarias o dicotómicas:**
  - Sexo: masculino o femenino
  - Respuesta: sí o no
  - Presencia de enfermedad: presente o ausente
  - Estado de empleo: empleado o desempleado

Importancia de la clasificación

La correcta identificación del tipo de variable es crucial porque:

- Determina qué pruebas y medidas pueden aplicarse.
  - Variables numéricas: pruebas paramétricas, correlaciones, regresión, etc.
  - Variables categóricas: pruebas de proporciones, chi-cuadrada, etc.
- Define el tipo de representación gráfica.
  - Categóricas: diagramas de barras, diagramas de pastel.
  - Numéricas: histogramas, diagramas de caja (*boxplots*), funciones de densidad.
- **Interpretación de resultados:** Una media tiene sentido para variables numéricas, pero no para categóricas nominales.
- **Selección de medidas de tendencia central y dispersión:**
  - Numéricas: media, mediana, moda, desviación estándar, varianza, percentiles, cuartiles.
  - Categóricas: moda, frecuencias absolutas y relativas, porcentajes.

$$\text{Frecuencia absoluta} = \frac{\text{número de observaciones favorables}}{\text{Total de observaciones}}$$

2.6.3. La función `class()` en R

Devuelve el **tipo de dato** que R asigna a un objeto, indicando cómo R interpreta la variable y qué operaciones son posibles.

Tabla 2: Tipos de datos más comunes en R según `class()`

Tipo de dato	**Descripción**	**Ejemplo**
`numeric`	Números reales (decimales)	`3.14`
`integer`	Números enteros	`42L`
`character`	Texto	`"Hola"`
`logical`	Valores verdadero/falso	`TRUE`
`factor`	Variable categórica con niveles	`factor(c(‘‘A’’,‘‘B’’))`
`Date`	Fechas	`as.Date("2025-08-12")`

2.7. Ejercios de cómputo

Las bases de datos se encuentran en
https://github.com/imelit/Bayesian_inference_class_2025/tree/main
Actividades:

- ¿Qué tipo de variables son edad, sexo?
- ¿Cuántos tipos de enfermedades ajenas a las respiratorias se reportan (comorbidades)? Hacer una gráfica de barras con sus respectivas frecuencias.
- ¿Cuántos tipos de enfermedades respiratorias se reportaron? Realiza una gráfica de barras con sus respectivas frecuencias.
- Cuántos tipos de influenza se reportarón? Realiza una gráfica de barras para comparar estas.
- De entre los casos reportados por influenza. Realiza una gráfica de barras del total de casos estratificado por meses, partiendo de octubre a mayo, tempora de influenza, https://www.cdc.gov/flu/about/season.html.
- Visualiza la serie de tiempo de Influenza estratificado por grupos de edad (casos contra tiempo).
- ¿Qué relación hay entre comorbidades e infectados por influenza (que pasa con los decesos)?

Guia para el análisis:
https://www.cdc.gov/fluview/overview/fluview-interactive.html

3. Probabilidad y variables aleatorias

La probabilidad es el lenguaje y marco matemático que usamos para formular la inferencia estadística. Con la probabilidad, la inferencia estadística nos permite hacer afirmaciones o predicciones sobre una población más allá de los datos observados en una muestra[1]

[1] Para esta sección voy a usar el libro de clase [4], capítulo 2, y [3], capítulos 1 y 2.

3.1. Axiomas de probabilidad (axiomas de Kolmogórov)

Consideremos un experimento cuya espacio muestral es Ω. Para cada evento E del espacio muestral, suponga que el número $\mathbb{P}(E)$ es definido y satisface los siguientes axiomas.

Axioma 1

$$0 \leq \mathbb{P}(E) \leq 1$$

Axioma 2

$$\mathbb{P}(\Omega) = 1$$

Axioma 3 Para cualquier secuencia de eventos mutuamente exclusivos E_1, $E_2, \ldots$ (esto es para cual quiere $E_i \cap E_j =$ con $i \neq j$),

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

Observación 3.1. Establecimos que $\mathbb{P}(E)$ está bien definida; esto es

$$\mathbb{P} : \Omega \to [0, 1]$$

una función que asigna a cada evento E un número real entre 0 y 1. De manera formal, ésto tiene sentido en el contexto de espacios medibles, ver por ejemplo [1]. Sin embargo, en estas notas no entraremos en esos detalles.

Nos enfocaremos en los casos prácticos, donde casi todos los eventos con los que trabajamos son medibles. También supondremos que las variables aleatorias están bien definidas y no tendremos problemas "patológicos" de teoría de la medida, ver también la definición formal de variable aleatoria en [1].

3.2. Particiones

Pag. 14-16 del libro [4]

Definición 3.1 (Partición)**.** Una colección de eventos $\{H_1, H_2, \ldots, H_k\}$ forma una partición de H si:

1. $H_i \cap H_j = \emptyset$ para $i \neq j$ (mutuamente excluyentes)
2. $\bigcup_{i=1}^{k} H_i = H$ (exhaustivos)
3. $\mathbb{P}(H_i) > 0$ para todo i

Aquí $H = \Omega$ es el espacio muestral u conjunto de hipótesis a estudiar.

Ejemplos (libro pag 15):
Let H be someone's religious orientation. Partitions include

- {Protestant, Catholic, Jewish, other, none};
- {Christian, non-Christian};
- {atheist, monotheist, multitheist}.

Let H be someone's number of children. Partitions include

- {0, 1, 2, 3 or more};
- {0, 1, 2, 3, 4, 5, 6, ...}.

Let H be the relationship between smoking and hypertension in a given population. Partitions include

- {some relationship, no relationship};
- {negative correlation, zero correlation, positive correlation}.

3.3. Probabilidad condicional

Definición 3.2. Sean E y F dos eventos, si $\mathbb{P}(F) \geq 0$, se define

$$\mathbb{P}(E|F) := \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)},$$

y se lee, la probabilidad condicional de que ocurra E dado que F ya ocurrió.

3.4. Regla de Bayes

Supongamos que $\{H_1, H_2, \ldots, H_K\}$ es una partición de eventos del espacio muestral H, E es un evento de H y $\mathbb{P}$ es una función de probabilidad definida para todos los eventos de H. Los axiomas de probabilidad implican las siguientes identidades.

Regla de la probabilidad total (marginal)

$$\Pr(E) = \sum_{k=1}^{K} \Pr(E \cap H_k) = \sum_{k=1}^{K} \Pr(E \mid H_k) \Pr(H_k)$$

Regla de Bayes (forma extendida):

$$\mathbb{P}(H_j \mid E) = \frac{\mathbb{P}(E \mid H_j)\mathbb{P}(H_j)}{\mathbb{P}(E)} = \frac{\mathbb{P}(E \mid H_j)\mathbb{P}(H_j)}{\sum_{k=1}^{K} \mathbb{P}(E \mid H_k)\mathbb{P}(H_k)}$$

La **regla de la probabilidad total** es una herramienta muy útil de la probabilidad, ya que nos permite calcular la probabilidad de un evento E en situaciones donde su cálculo directo es complicado o imposible.

Ejemplo: Regla de la probabilidad total

Supongamos que una fábrica recibe piezas de tres proveedores distintos. Cada proveedor abastece un porcentaje diferente de la producción total, y además tiene distinta calidad en sus envíos. Queremos calcular la probabilidad total de que una pieza seleccionada al azar resulte defectuosa (esto es el evento E).

Los escenarios (hipótesis) son:

$$\begin{aligned} H_1 &: \text{ Proveedor 1 envía el } 40\,\% \text{ de las piezas,} \\ H_2 &: \text{ Proveedor 2 envía el } 35\,\%, \\ H_3 &: \text{ Proveedor 3 envía el } 25\,\%. \end{aligned}$$

También se conocen las probabilidades de defecto condicionadas al proveedor:

$$\mathbb{P}(E \mid H_1) = 0.01, \quad \mathbb{P}(E \mid H_2) = 0.03, \quad \mathbb{P}(E \mid H_3) = 0.02.$$

Directamente calcular $\mathbb{P}(E)$ es complicado porque intervienen varios proveedores. Aquí es donde la **ley de la probabilidad total** es esencial: descomponemos el cálculo en función de los escenarios H_i.

$$\mathbb{P}(E) = \mathbb{P}(E \mid H_1)\mathbb{P}(H_1) + \mathbb{P}(E \mid H_2)\mathbb{P}(H_2) + \mathbb{P}(E \mid H_3)\mathbb{P}(H_3).$$

$$\mathbb{P}(E) = (0.01)(0.40) + (0.03)(0.35) + (0.02)(0.25).$$

$$\mathbb{P}(E) = 0.0195.$$

Conclusión: La probabilidad de que una pieza elegida al azar sea defectuosa es de $1.95\,\%$.

En inferencia bayesiana, el conjunto

$$\{H_1, H_2, \ldots, H_K\}$$

suele referirse a **hipótesis mutuamente excluyentes** o a **escenarios de la naturaleza**. El evento E representa la **evidencia observada**, es decir, el resultado de un experimento, encuesta o estudio.

Cuando queremos **comparar hipótesis después de observar los datos**, una herramienta útil es el **cociente de probabilidades a posteriori**:

$$\frac{\mathbb{P}(H_i \mid E)}{\mathbb{P}(H_j \mid E)}.$$

Aplicando la **regla de Bayes**, se obtiene:

$$\frac{\mathbb{P}(H_i \mid E)}{\mathbb{P}(H_j \mid E)} = \frac{\mathbb{P}(E \mid H_i)\mathbb{P}(H_i)}{\mathbb{P}(E \mid H_j)\mathbb{P}(H_j)}.$$

Esta expresión se puede reorganizar como:

$$\frac{\mathbb{P}(H_i \mid E)}{\mathbb{P}(H_j \mid E)} = \underbrace{\frac{\mathbb{P}(E \mid H_i)}{\mathbb{P}(E \mid H_j)}}_{\text{Factor de Bayes}} \times \underbrace{\frac{\mathbb{P}(H_i)}{\mathbb{P}(H_j)}}_{\text{Creencias a priori}} .$$

Interpretación

- El **factor de Bayes** $\frac{\mathbb{P}(E|H_i)}{\mathbb{P}(E|H_j)}$ mide qué tan bien explica cada hipótesis la evidencia observada.
- El cociente $\frac{\mathbb{P}(H_i)}{\mathbb{P}(H_j)}$ refleja nuestras **creencias iniciales (a priori)**.
- En conjunto, el cociente a posteriori nos dice cómo deben ajustarse nuestras creencias una vez observada la evidencia.

Nota: la regla de Bayes no nos dice qué debemos creer después de ver los datos, sino cómo deben cambiar nuestras creencias previas al incorporar la información observada.

Ejemplo

Supongamos que queremos estimar el **nivel de apoyo a un candidato político**.

- Sea

$$H = \{\text{todas los porcentajes posibles de apoyo al candidato A}\}.$$

- Definimos dos hipótesis:

$$\begin{aligned} H_1 &= \{\text{más de la mitad de los votantes apoyan al candidato A}\}, \\ H_2 &= \{\text{ menos o igual a la mitad de los votantes apoyan al candidato A}\}. \end{aligned}$$

- Definimos la evidencia:

$$E = \{\text{en una encuesta, 54 de 100 personas dijeron apoyar al candidato A}\}.$$

El conjunto $\{H_1, H_2\}$ forma una **partición** de H. De interés está calcular:

$$\mathbb{P}(H_1 \mid E), \qquad \text{o bien} \qquad \frac{\mathbb{P}(H_1 \mid E)}{\mathbb{P}(H_2 \mid E)}.$$

Resolvermos este ejemplo más adelante.

3.5. Variables aleatorias y densidades de probabilidad

Definición 3.3 (Variable Aleatoria)**.** Una variable aleatoria X es una función medible $X : \Omega \to \mathbb{R}$.

Definición 3.4 (Variable aleatoria discreta)**.** Una variable aleatoria X es *discreta* si toma valores en un conjunto contable. En este caso, la **función de masa de probabilidad** (fmp) se define como

$$p(x) = \mathbb{P}(X = x), \quad \text{donde} \quad \sum_x p(x) = 1.$$

Definición 3.5 (Variable aleatoria continua)**.** Una variable aleatoria X es *continua* si existe una función $f(x) \geq 0$ tal que

$$\mathbb{P}(a < X \leq b) = \int_a^b f(x)\, dx \quad \text{para todo } a < b,$$

y además

$$\int_{-\infty}^{\infty} f(x)\, dx = 1.$$

A la función $f(x)$ se le llama **función de densidad de probabilidad** (fdp).

Observación 3.2 (Unificación de notación). En estas notas, por simplicidad, usaremos la misma letra $f(x)$ para referirnos tanto a la función de masa de probabilidad (en el caso discreto) como a la función de densidad de probabilidad (en el caso continuo). La diferencia estará dada por el contexto:

$$\mathbb{P}(X \in A) = \begin{cases} \displaystyle\sum_{x \in A} f(x), & \text{si } X \text{ es discreta,} \\ \displaystyle\int_A f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

3.6. Definiciones de esperanza y varianza

Definición 3.6 (Esperanza o media)**.** Sea X una variable aleatoria (discreta o continua) con función de desidad $f(x)$. La **esperanza** o **media** de X se define como

$$\mathbb{E}[X] = \mu = \begin{cases} \displaystyle\sum_{x} x\, f(x), & \text{si } X \text{ es discreta,} \\ \displaystyle\int_{-\infty}^{\infty} x\, f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Definición 3.7 (Varianza)**.** La **varianza** de una variable aleatoria X es un número que mide la dispersión de los valores de X alrededor de su media $\mu = \mathbb{E}[X]$. Se define como

$$\mathrm{Var}(X) = \mathbb{E}\big[(X-\mu)^2\big]\,.$$

En forma explícita:

$$\mathrm{Var}(X) = \begin{cases} \displaystyle\sum_{x} (x-\mu)^2\, f(x), & \text{si } X \text{ es discreta,} \\ \displaystyle\int_{-\infty}^{\infty} (x-\mu)^2\, f(x)\, dx, & \text{si } X \text{ es continua.} \end{cases}$$

Observación 3.3. Una fórmula equivalente y muy útil para la varianza es

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

3.7. Distribuciones Discretas Importantes

3.7.1. Distribución Bernoulli

Definición 3.8 (Distribución Bernoulli)**.**

$$X \sim \mathrm{Bernoulli}(\theta)$$

si $X = 0$ ó $X = 1$ con probabilidad de éxito $\mathbb{P}(X = 1) = \theta$.

fdp:

$$f(x) = \theta^x(1-\theta)^{1-x}$$

para $x \in \{0, 1\}$

Propiedades: $\mathbb{E}[X] = \theta$, $\text{Var}(X) = \theta(1-\theta)$

3.7.2. Distribución Binomial

Definición 3.9 (Distribución Binomial)**.**

$$X \sim \text{Binomial}(n, \theta)$$

representa el número de éxitos en n ensayos independientes Bernoulli(θ).

fdp:

$$f(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$$

para $x \in \{0, 1, \ldots, n\}$

Propiedades: $\mathbb{E}[X] = n\theta$, $\text{Var}(X) = n\theta(1-\theta)$

3.7.3. Distribución Poisson

Definición 3.10 (Distribución Poisson)**.**

$$X \sim \text{Poisson}(\lambda)$$

modela eventos raros que ocurren a tasa λ.

FMP:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

para $x \in \{0, 1, 2, \ldots\}$

Propiedades: $\mathbb{E}[X] = \text{Var}(X) = \lambda$

3.7.4. Distribución Binomial Negativa

Definición 3.11 (Distribución Binomial Negativa)**.**

$$X \sim \text{NegBin}(r, \theta)$$

modela el número de fallas antes de lograr r éxitos.

fdp:
$$f(x) = \binom{x+r-1}{x} \theta^r (1-\theta)^x$$
para $x \in \{0, 1, 2, \ldots\}$

Propiedades: $\mathbb{E}[X] = \frac{r(1-\theta)}{\theta}$, $\mathrm{Var}(X) = \frac{r(1-\theta)}{\theta^2} = E[X]\frac{1}{\theta}$.

3.8. Distribuciones Continuas Importantes

3.8.1. Distribución Beta

Definición 3.12 (Distribución Beta)**.**

$$X \sim \mathrm{Beta}(\alpha, \beta)$$

con soporte en $[0, 1]$.

FDP:
$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \leq x \leq 1$$

Propiedades: $\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}$, $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

3.8.2. Distribución Gamma

Definición 3.13 (Distribución Gamma)**.**

$$X \sim \Gamma(\alpha, \beta)$$

con forma $\alpha > 0$ y tasa $\beta > 0$.

FDP: $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ para $x > 0$

Propiedades: $\mathbb{E}[X] = \frac{\alpha}{\beta}$, $\mathrm{Var}(X) = \frac{\alpha}{\beta^2}$

3.8.3. Distribución Normal

Definición 3.14 (Distribución Normal)**.**

$$X \sim \mathrm{Normal}(\mu, \sigma^2)$$

con ubicación μ y escala σ.

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

Propiedades: $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$

3.8.4. Distribución Log-Normal

Definición 3.15 (Distribución Log-Normal)**.**

$$X \sim \mathrm{LogNormal}(\mu, \sigma^2)$$

si

$$\log(X) \sim \mathrm{Normal}(\mu, \sigma^2)$$

.

FDP: $f(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right)$ para $x > 0$
Propiedades: $\mathbb{E}[X] = \exp\left(\mu + \frac{\sigma^2}{2}\right)$, $\mathrm{Var}(X) = \exp(2\mu+\sigma^2)(\exp(\sigma^2)-1)$

distribución uniforme

distribución exponencial

3.9. Distribuciones conjunta y marginal

Definición 3.16 (Distribución conjunta)**.** Sean X y Y dos variables aleatorias (discretas o continuas). La **distribución conjunta** de (X, Y) se describe mediante una función $f(x, y)$ tal que:

$$\mathbb{P}\big((X,Y) \in A\big) = \begin{cases} \displaystyle\sum_{(x,y)\in A} f(x,y), & \text{si } (X,Y) \text{ son discretas,} \\ \displaystyle\iint_A f(x,y)\,dx\,dy, & \text{si } (X,Y) \text{ son continuas.} \end{cases}$$

En ambos casos, se cumple que

$$f(x,y) \geq 0, \qquad \sum_{x,y} f(x,y) = 1 \quad \text{o} \quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1.$$

Definición 3.17 (Distribuciones marginales)**.** Las **distribuciones marginales** de X y Y se obtienen a partir de la distribución conjunta $f(x,y)$:

$$f_X(x) = \begin{cases} \displaystyle\sum_y f(x,y), & \text{si } (X,Y) \text{ son discretas,} \\ \displaystyle\int_{-\infty}^{\infty} f(x,y)\,dy, & \text{si } (X,Y) \text{ son continuas,} \end{cases}$$

$$f_Y(y) = \begin{cases} \displaystyle\sum_x f(x,y), & \text{si } (X,Y) \text{ son discretas,} \\ \displaystyle\int_{-\infty}^{\infty} f(x,y)\,dx, & \text{si } (X,Y) \text{ son continuas.} \end{cases}$$

De este modo, $f_X(x)$ y $f_Y(y)$ describen las probabilidades o densidades de X y Y por separado.

3.10. Distribución condicional

Definición 3.18 (Distribución condicional)**.** Sean X y Y dos variables aleatorias con distribución conjunta $f(x,y)$ y marginales $f_X(x)$ y $f_Y(y)$.

La **distribución condicional** de X dado $Y = y$ se define como

$$f_{X|Y}(x \mid y) = \frac{f(x,y)}{f_Y(y)}, \quad \text{siempre que } f_Y(y) > 0.$$

De manera análoga, la distribución condicional de Y dado $X = x$ es

$$f_{Y|X}(y \mid x) = \frac{f(x,y)}{f_X(x)}, \quad \text{siempre que } f_X(x) > 0.$$

Observación 3.4. En notación unificada:

$$\mathbb{P}(X \in A \mid Y = y) = \begin{cases} \displaystyle\sum_{x \in A} f_{X|Y}(x \mid y), & \text{si } (X,Y) \text{ son discretas,} \\ \displaystyle\int_A f_{X|Y}(x \mid y)\,dx, & \text{si } (X,Y) \text{ son continuas.} \end{cases}$$

Observación 3.5 (Relación con la probabilidad condicional). La definición anterior generaliza la probabilidad condicional usual:

$$\mathbb{P}(X \in A \mid Y \in B) = \frac{\mathbb{P}\big((X,Y) \in A \times B\big)}{\mathbb{P}(Y \in B)},$$

siempre que $\mathbb{P}(Y \in B) > 0$.

De forma análoga, estas definiciones se extienden naturalmente al caso de la distribución conjunta de un número arbitrario de variables aleatorias

4. Inferencia Estadística

Dada una muestra aleatroia $\{Y_1, Y_2, \ldots, Y_n\}$ de la variable Y, que supondremos

$$Y \sim f(y \mid \theta)$$

nuestro objetivo es usar la inferencia estadística para *obtener estimaciones* del parámetro desconocido θ que caracteriza la distribución de Y.

4.1. Estimación por Máxima Verosimilitud

Supongamos que contamos con una función de densidad (o masa) de probabilidad

$$f(y \mid \theta),$$

y un conjunto de datos

$$Y_1 = y_1,\ Y_2 = y_2,\ \ldots,\ Y_n = y_n$$

para la variable aleatoria Y.

Definición 4.1 (The likelihood function)**.** Definimos la **función de verosimilitud** como

$$L(\theta) = \prod_{i=1}^{n} f(y_i \mid \theta).$$

Esta función depende solo de θ, ya que los valores y_i son observaciones fijas.

Definición 4.2 (Estimador de máxima verosimilitud)**.** El valor θ que maximiza $L(\theta)$:

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta \in \Theta} L(\theta)$$

se denomina **estimador de máxima verosimilitud** de θ.

Esta elección de θ por ser el máximo hace que los datos observados sean más plausibles bajo el modelo.

Ejemplo 1: estimar la probabilidad de éxito de una Distribución Bernoulli

Supongamos $Y_1, \ldots, Y_n \sim \text{Bernoulli}(\theta)$, con $\theta \in (0, 1)$.

- La verosimilitud es

$$L(\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}.$$

- La log-verosimilitud es

$$\ell(\theta) = \sum_{i=1}^{n} \big[y_i \log\theta + (1 - y_i)\log(1-\theta)\big].$$

- Derivando e igualando a cero:

$$\frac{d}{d\theta}\ell(\theta) = \frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{n - \sum_{i=1}^{n} y_i}{1-\theta} = 0.$$

- Resolviendo se obtiene el MLE:

$$\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

es decir, la proporción muestral de éxitos (frecuentista).

Ejemplo 2: Estimar la media de una Distribución Normal

Suponga que $Y_1, \ldots, Y_n$ es una muestra aleatoria i.i.d. con

$$Y_i \sim \mathcal{N}(\theta, \sigma^2), \quad \text{donde } \sigma^2 \text{ es conocida.}$$

Función de verosimilitud (en términos de θ):

$$\begin{aligned} L(\theta \mid y_{1:n}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}. \end{aligned}$$

Log-verosimilitud:

$$\ell(\theta \mid y_{1:n}) = \log L(\theta \mid y_{1:n}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2.$$

Derivada de la log-verosimilitud:

$$\frac{d\,\ell(\theta \mid y_{1:n})}{d\theta} = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} y_i - n\theta\right).$$

Igualando a cero:

$$\frac{d\,\ell(\theta \mid y_{1:n})}{d\theta} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} y_i - n\theta = 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{Y}.$$

(Opcional) Segundo derivada para verificar máximo:

$$\frac{d^2\,\ell(\theta \mid y_{1:n})}{d\theta^2} = -\frac{n}{\sigma^2} < 0 \quad \Rightarrow \quad \hat{\theta} \text{ es un máximo.}$$

Observación 4.1 (Sobre la existencia y unicidad del MLE). La unicidad del estimador de máxima verosimilitud (MLE) no está garantizada, y en ciertos casos pueden existir al menos dos valores de parámetro $\hat{\theta}_1 \neq \hat{\theta}_2$ tales que

$$L(\hat{\theta}_1) = L(\hat{\theta}_2) = \arg\max_{\theta\in\Theta} L(\theta).$$

En otras situaciones, el MLE puede no existir en absoluto en todo el espacio parametrico Θ, Ver ejmplos del Capítulo 2 del libro [2]

Observación 4.2 (Interpretación y propiedades del MLE).

- **Enfoque frecuentista:**

  El MLE trata los datos y_i como observaciones fijas y considera θ como un parámetro fijo desconocido.

- **Objetivo del MLE:** Busca el valor de θ que hace que los datos observados sean más plausibles, proporcionando así una estimación puntual del parámetro.

- **Propiedades deseables:** Bajo condiciones generales y para muestras grandes, el MLE tiene tres características importantes:

  - *Consistencia:* A medida que aumenta el tamaño de la muestra, el estimador $\hat{\theta}_{\text{MLE}}$ se aproxima al valor verdadero θ.
  - *Eficiencia:* Entre estimadores insesgados, el MLE tiende a tener la menor varianza posible, es decir, produce estimaciones más precisas.
  - *Normalidad asintótica:* Para muestras grandes, la distribución de $\hat{\theta}_{\text{MLE}}$ se aproxima a una distribución Normal centrada en θ, lo que permite construir intervalos de confianza y realizar pruebas de hipótesis.

Observación 4.3. En la práctica, supondremos que los datos han sido generados por una función de distribución, que consideramos como el *modelo* que describe los datos. La elección del modelo suele basarse en la experiencia y el conocimiento previo sobre la naturaleza de los datos: su forma, rango de valores y contexto en el que fueron observados.

4.2. Proceso de Aprendizaje Bayesiano

1. **Especificación de priori**: Codificar creencias iniciales sobre θ
2. **Observación de datos**: Recopilar datos y
3. **Evaluación de verosimilitud**: Evaluar $\mathbb{P}(y \mid \theta)$ para diferentes valores de θ
4. **Cómputo de posteriori**: Aplicar regla de Bayes para actualizar creencias
5. **Decisión/Predicción**: Usar posteriori para inferencia

4.3. Inferencia Estadística Bayesiana

En contextos estadísticos, la regla de Bayes se convierte en:

$$\boxed{\mathbb{P}(\theta \mid y) = \frac{L(y \mid \theta)\pi(\theta)}{m(y)}}$$

Donde:

- $\pi(\theta)$: **Distribución priori** (creencias antes de observar datos)
- $L(y \mid \theta)$: **Función de verosimilitud** (probabilidad de datos dado parámetro)
- $\pi(\theta \mid y)$: **Distribución posteriori** (creencias actualizadas tras observar datos)
- $m(y) = \int L(y \mid \theta)\pi(\theta)\, d\theta$: **Verosimilitud marginal** (constante normalizadora)

5. Intercambiabilidad

5.1. Definición e Intuición

Definición 5.1 (Intercambiabilidad)**.** Las variables aleatorias $X_1, X_2, \ldots, X_n$ son intercambiables si su distribución conjunta es invariante bajo permutaciones:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_{\pi(1)} = x_1, \ldots, X_{\pi(n)} = x_n)$$

para cualquier permutación π.

Intuición: No podemos distinguir entre las variables basándose en sus índices—son "simétricas" en nuestras creencias.

5.2. Teorema de De Finetti

Teorema 5.1 (De Finetti)**.** Si $X_1, X_2, \ldots$ son infinitamente intercambiables, entonces existe una medida aleatoria P tal que:

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int \prod_{i=1}^{n} P(A_i)\, d\mu(P)$$

Esto justifica el enfoque bayesiano: las observaciones intercambiables pueden modelarse como condicionalmente independientes dado un parámetro desconocido.

5.3. Implicaciones Prácticas

1. **Justifica modelos paramétricos**: La intercambiabilidad lleva naturalmente a asumir que las observaciones son i.i.d. dados los parámetros
2. **Elicitación de prioris**: Ayuda a determinar distribuciones prioris apropiadas
3. **Verificación de modelos**: Las violaciones sugieren inadecuación del modelo

Referencias

[1] Donald L Cohn. *Measure theory*, volume 1. Springer, 2013.

[2] Leonhard Held and Daniel Sabanés Bové. Likelihood and bayesian inference. *Statistics for Biology and Health. Springer, Berlin, Heidelberg*, 2020.

[3] Paul G Hoel. Introduction to statistical theory. 1971.

[4] Peter D. Hoff. *A First Course in Bayesian Statistical Methods*. Springer, 2009.

[5] Luis Rincón. *Una introducción a la estadística inferencial*. Universidad Nacional Autónoma de México, México, 2019.