Marco Bayesiano para el análisis de datos, calibración de parámetros y modelamiento inverso

Variables Aleatorias

03

Universidad Industrial de Santander
U18 Fest

23–26 de septiembre de 2019

Introducción

- Cubrimos los elementos de la teoría de probabilidades
  - *Resultados, variables aleatorias, eventos, y probabilidades*
- El modelamiento probabilístico en la práctica se basa en variables aleatorias, eventos asociados a variables aleatorias, y sus probabilidades
- Usualmente, la definición de espacio de resultados Ω se puede tratar de manera implícita, i.e., no es necesario definir Ω explícitamente

Introducción

Ejemplo: Altura de una población

$$\Omega = \{\text{personas en la población}\}$$

- **Variable aleatoria**: Altura

$$A \colon \Omega \to \mathbb{R}$$

- **Eventos**:
  - Altura mayor o igual a cierto valor a: $P(A \geq a)$
  - Altura menor a cierto valor b: $P(A < b)$

Introducción

Ejemplo: Altura de una población

$$\Omega = \{\text{personas en la población}\}$$

- **Variable aleatoria**: Altura

$$A \colon \Omega \to \mathbb{R}$$

- **Eventos**:
  - Altura mayor o igual a cierto valor a: $P(A \geq a)$
  - Altura menor a cierto valor b: $P(A < b)$

Si el interés es en las *propiedades* de la "altura"...

- ...no es necesario pensar en Ω
- Sólo hace falta pensar en A y los valores que puede tomar

Funciones de masa y densidad de probabilidad

Asocian probabilidades a los valores de variables aleatorias

- **Variables discretas**:
  Función de *masa* de probabilidad (pmf)
- **Variables contínuas**:
  Función de *densidad* acumulada (cdf)
  Función de *densidad* de probabilidad (pdf)

Función de masa de probabilidad (pmf)

- **Variable aleatoria**

$$X\colon \Omega \to C$$

- **pmf** $p_X : C \to [0,1]$ definida como

$$p_X(c) = P(X = c) = P(\{\omega \in \Omega\colon X(\omega) = c\})$$

Función de masa de probabilidad (pmf)

Ejemplo: Lanzar dos dados. Suma $X\colon \Omega \to \mathbb{N}$

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

pmf:

- $p_X(1) = 0$
- $p_X(2) = 1/36$
- $p_X(3) = 2/36$
- $p_X(4) = 3/36$

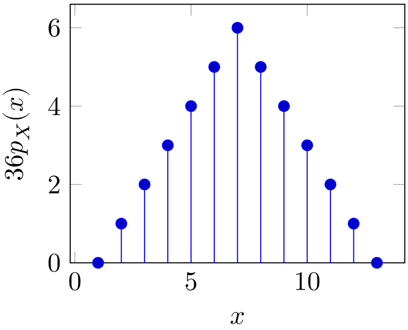
  $\dots$

Función de masa de probabilidad (pmf)

Ejemplo: Lanzar dos dados. Suma $X\colon \Omega \to \mathbb{N}$

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

pmf:



Función de masa de probabilidad (pmf)

- Para la variable aleatoria discreta $X\colon \Omega \to C$, se puede utilizar la pmf para calcular la probabilidad del evento del valor de X estar en cualquier subconjunto $B \subseteq C$:

$$P(X \in B) = \sum_{b \in B} p_X(b)$$

Función de masa de probabilidad (pmf)

- Para la variable aleatoria discreta $X\colon \Omega \to C$, se puede utilizar la pmf para calcular la probabilidad del evento del valor de X estar en cualquier subconjunto $B \subseteq C$:

$$P(X \in B) = \sum_{b \in B} p_X(b)$$

- Como consecuencia,

$$\sum_{b \in C} p_X(b) = 1$$

- **Ejemplo**: Lanzar dos dados

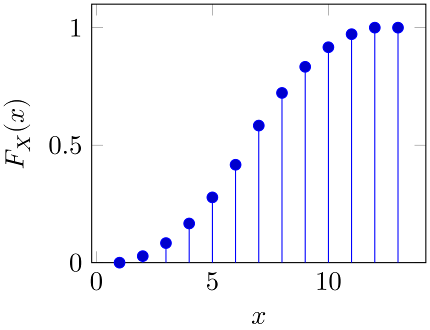
$$p_X(1) + p_X(2) + \cdots + p_X(11) + p_X(12) = 1$$

Función de densidad acumulada (cdf)

- **Variable aleatoria**

$$X\colon \Omega \to C$$

- **cdf** $F_X(c) := P(X \leq c) \in [0, 1]$
- **Ejemplo**: Lanzar dos dados



Función cuantil

Para la variable aleatoria $X\colon \Omega \to C$,

$$Q(p) := \inf\{x \in C\colon p \leq F_X(x)\}$$

- $Q(p)$ determina el menor valor x de X para el cual $P(X \leq x) \geq p$
- **Ejemplo**: Lanzar dos dados

p	0.1	0.2	0.5	0.7	0.9
$Q(p)$	4	5	7	8	10

Estadísticos

- **Media/Esperanza/Valor esperado**:

$$\mu \equiv \mathbb{E}[X] := \sum_{x \in C} x p_X(x)$$

El promedio de n muestras de la variable aleatoria en el límite $n \to \infty$, i.e.,

$$\frac{x_1 + x_2 + \cdots + x_n}{n} \to \mathbb{E}[X]$$

- **Varianza**:

$$\sigma^2 \equiv \mathbb{E}\left[(X - \mu)^2\right] = \sum_{x \in C} (x - \mu)^2 p_X(x)$$

Medida de desviación de la distribución alrededor de la media

- **Desviación estándar**: $\sigma = \sqrt{\sigma^2}$
- **Mediana**: $Q(0.5)$
- **Modo**: Valor mas probable, i.e, $\arg\max p_X(x)$

Muestrear variables aleatorias

- Podemos simular variables aleatorias utilizando *generadores de números pseudo-aleatorios* (PRNGs)
- PRNGs **no** generan números realmente aleatorios...
- ...pero generan una secuencia de números con propiedades aproximadamente iguales a las propiedades de secuencias de números aleatorios
- La secuencia de un PRNG es determinada por un valor inicial llamdo *semilla* (seed)
- Dada la misma semilla, un cierto PRNG genera la misma secuencia cada vez que es invocado

Variables aleatorias contínuas

- **Variable aleatoria**

$$X\colon \Omega \to C$$

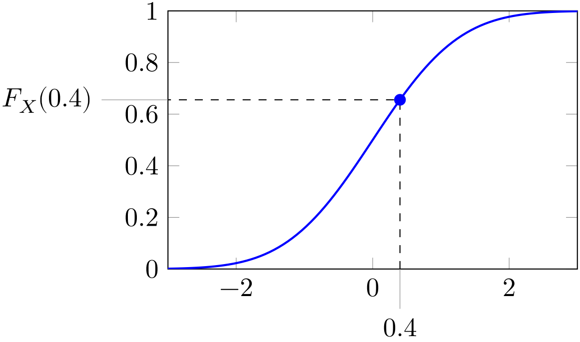
- **cdf** $F_X(c) = P(X \leq c) \in [0, 1]$

Variables aleatorias contínuas

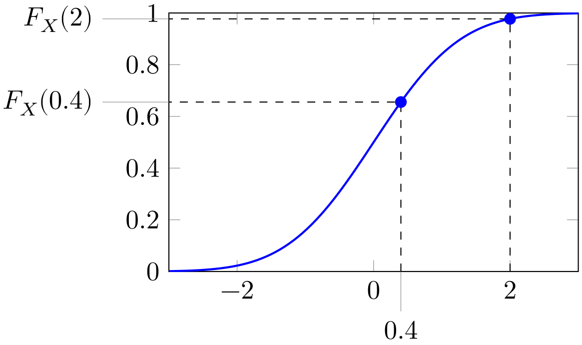
- **Variable aleatoria**

$$X\colon \Omega \to C$$

- **cdf** $F_X(c) = P(X \leq c) \in [0,1]$
- **Ejemplo**: Variable aleatoria normal estándar



Función de densidad cumulativa



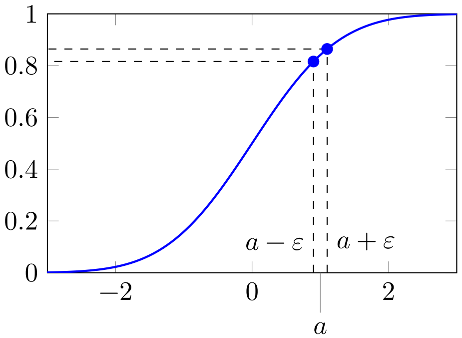
La cdf puede utilizarse para evaluar la probabilidad de varios escenarios:

- $P(X > 2) = 1 - P(X \leq 2) = 1 - F_X(2)$
- $P(0.4 \leq X \leq 2) = P(X \leq 2) - P(X \leq 0.4) = F_X(2) - F_X(0.4)$

Función de densidad de probabilidad (pdf)

- Para generalizar la pmf a variables aleatorias contínuas, hay que introducir el concepto de *densidad*
- La pdf (o *densidad* o *distribución*) $p_X(a)$ cuantifica la cantidad de probabilidad en un intervalo alrededor de a relativa a la longitud del intervalo

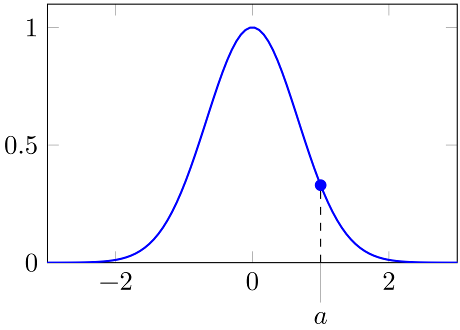
Ejemplo: Variable aleatoria normal estándar



Función de densidad de probabilidad (pdf)

- Para generalizar la pmf a variables aleatorias contínuas, hay que introducir el concepto de *densidad*
- La pdf (o *densidad* o *distribución*) $p_X(a)$ cuantifica la cantidad de probabilidad en un intervalo alrededor de a relativa a la longitud del intervalo

Ejemplo: Variable aleatoria normal estándar



Función de densidad de probabilidad (pdf)

- Si entendemos la probabilidad como la “masa” de una región del espacio de resultados (evento), la distribución $p_X(\cdot)$ indica la densidad en cada punto de esa región
- Para variables aleatorias *discretas*,

$$P(X \in B) = \sum_{b \in B} p_X(b)$$

- Para variables aleatorias *contínuas*,

$$P(X \in B) = \int_B p_X(x)\,\mathrm{d}x$$

- La integral $\int(\cdot)\mathrm{d}x$ generaliza la suma $\sum$ a variables aleatorias contínuas

Función de densidad de probabilidad (pdf)

- Para la variable aleatoria $X\colon \Omega \to C$, la pdf y la cdf están conectadas a través de la relación

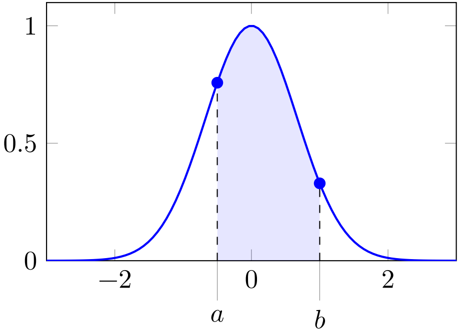
$$F_X(a) = P(X \leq a) = \int_{-\infty}^{a} p_X(x)\,\mathrm{d}x$$

- Como consecuencia de la definición,

$$P(X \in C) = \int_C p_X(x)\,\mathrm{d}x = 1$$

Área bajo la curva

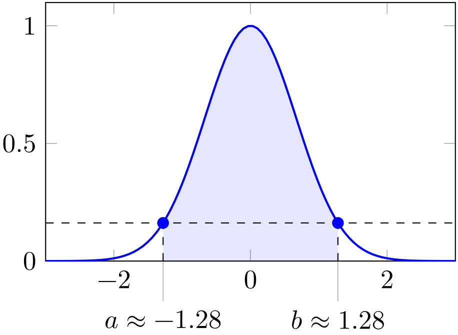
- La integral $\int_B p_X(x)\,\mathrm{d}x$ puede interpretarse como el área bajo la curva $y = p(x)$ en el intervalo B
- E.g., para el intervalo $B = (a, b)$,



- Dado que $\int_C p_X(x)\,\mathrm{d}x = 1$, eso indica que el área bajo la pdf es 1
- La cdf $F_X(a)$ corresponde al área bajo la curva entre $-\infty$ y a

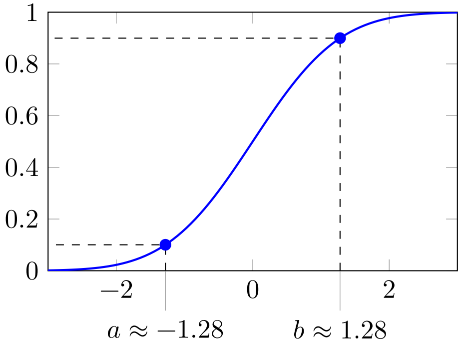
Ejemplo

- Variable aleatoria normal estándar $X \sim \mathcal{N}(0,1)$
- **Intervalo de mayor densidad (HDI)**: Para un valor $\alpha \in [0,1]$, hay un intervalo $B = (a,b)$ para el cual
  - $p_X(a) = p_X(b)$
  - $P(X \in B) = 1 - \alpha$
- Para $\alpha = 0.2$:



Ejemplo

- Variable aleatoria normal estándar $X \sim \mathcal{N}(0, 1)$
- **Intervalo de mayor densidad (HDI)**: Para un valor $\alpha \in [0, 1]$, hay un intervalo $B = (a, b)$ para el cual
  - $p_X(a) = p_X(b)$
  - $P(X \in B) = 1 - \alpha$
- Para $\alpha = 0.2$:



Ejemplo

- Para la variable aleatoria $X \sim \mathcal{N}(\mu, \sigma)$ y $\alpha = 0.05$,

$$\text{HDI} = (\mu - 1.96\sigma, \mu + 1.96\sigma)$$

- Éste intervalo se conoce comúnmente como el *intervalo de credibilidad de 95%*
- Compárese con la noción de *intervalo de confianza*, que visitaremos luego