Modelos Estadísticos para la Regresión y la Clasificación

Práctico 2 - Probabilidad

Micaela Long

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

16 de agosto de 2024

Comentario Práctico 1
Ejercicio 3

Tanto la función median() de R como median() de Python calculan la mediana de esta forma:

- Si el largo de la lista es impar, entonces la mediana es el elemento que se encuentra en el lugar $\frac{n+1}{2}$ de la lista ordenada.
- Si el largo de la lista es par, entonces la mediana es el promedio de los elementos que se encuentran en los lugares $\frac{n}{2}$ y $\frac{n}{2}+1$ de la lista ordenada.

La mediana de [1,2,3,4] será $\frac{2+3}{2} = 2{,}5$.

Para nosotros **la mediana será siempre un elemento del conjunto**, por lo que utilizaremos la siguiente definición:

- Si el largo de la lista es impar, la mediana es el elemento que se encuentra en el lugar $\frac{n+1}{2}$ de la lista ordenada.
- Si el largo de la lista es par, la mediana es el elemento que se encuentra en el lugar $\frac{n}{2}$ de la lista ordenada.

La mediana de [1,2,3,4] será 2.

Para esto podemos definir una función que tenga como entrada una lista, y que: 1) si el largo de la lista es impar, devuelva el elemento que se encuentra en el lugar $\frac{n+1}{2}$ de la lista ordenada, o 2) si el largo de la lista es par, devuelva el elemento que se encuentra en el lugar $\frac{n}{2}$ de la lista ordenada.

Comentario Práctico 1
Ejercicio 3

Posible función en R:

```r
mediana <- function(muestra) {
  muestra <- sort(muestra)  # ordena la lista
  n <- length(muestra)  # calcula el largo de la lista

  if (n %% 2 == 0) {  # si el largo de la lista es par
    mediana <- muestra[n/2]
  } else {  # Si el largo de la lista es impar
    mediana <- muestra[(n %/% 2) + 1]
  }
  return(mediana)
}
```

Posible función en Python:

```python
def mediana(muestra):
    muestra.sort() # ordena la lista
    n = len(muestra) # calcula el largo de la lista
    if n % 2 == 0:  # si el largo de la lista es par
        mediana = muestra[n//2 - 1]
    else: # si el largo de la lista es impar
        mediana = muestra[n//2]
    return mediana
```

(ya están incorporadas en los laboratorios del práctico 1).

Práctico 2
Probabilidad

Material para hacer práctico 2 (disponible en EVA):

- Teóricos 12/8 y 14/8 (misma presentación)
- Cambio de variable - ejemplos
- Esperanza condicional - ejemplos

Ejercicio 1:

1. Si $X \sim U[0,1]$, halle la densidad de $Y = aX + b$ (analice la situación geométricamente).

2. Si X tiene densidad f_X y $Y = g(X)$ con g estrictamente monótona, pruebe que Y tiene densidad $f_Y(y) = \frac{1}{\left|\frac{dy}{dx}\right|} f_X(x)$ donde $x = g^{-1}(y)$.

3. Si X tiene densidad
$$f_X(x) = \begin{cases} e^{-x} & \text{si } x \geq 0 \\ 0 & \text{si no} \end{cases}$$
halle la densidad de $Y = \sqrt{X}$

Ejercicio 1
Parte 1

Tenemos $X \sim U[0,1]$, $Y = aX + b$. Queremos hallar la densidad $f_Y(y)$.
$X \sim U[0,1]$ entonces

$$f_X(x) = \begin{cases} 1 & \text{si } 0 \leq x \leq 1, \\ 0 & \text{en otro caso.} \end{cases}$$

luego

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(s)\, ds = \begin{cases} 0 & \text{si } -\infty \leq x \leq 0, \\ \int_0^x 1\, ds = x & \text{si } 0 \leq x \leq 1, \\ 1 & \text{si } x \geq 1. \end{cases}$$

Supongamos que $a > 0$. Usando $Y = aX + b$

$$\begin{aligned} F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(aX + b \leq y) &= \mathbb{P}\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right) \\ &= \frac{y-b}{a} \end{aligned}$$

si $\frac{y-b}{a} \in [0,1]$, o sea, si $y \in [b, b+a]$.

Tenemos la distribución, podemos calcular la densidad:

$$f_Y(y) = \frac{\partial F_Y(y)}{\partial y} = \frac{\partial \left(\frac{y-b}{a}\right)}{\partial y} = \frac{1}{a}$$

Esto es lo que estamos haciendo cuando usamos la fórmula para el **cambio de variable**:

$$f_Y(y) = f_X(x) \left. \frac{1}{\left|\frac{dy}{dx}\right|} \right|_{x=g^{-1}(y)} = f_X(g^{-1}(y)) \left| \frac{1}{g'(g^{-1}(y))} \right|$$

En este ejercicio (si $a > 0$):

- $g(x) = ax + b$
  - $g^{-1}(y) = \frac{y-b}{a}$
  - $g'(x) = a$
- $f_X(x) = 1$ si $x \in [0, 1]$

$$f_Y(y) = \overbrace{f_X(g^{-1}(y))}^{1} \left| \frac{1}{\underbrace{g'(g^{-1}(y))}_{a}} \right| = \frac{1}{|a|} = \frac{1}{a} \quad \text{en el intervalo } [b, b+a]$$

¿Qué está sucediendo geométricamente?

La transformación $Y = aX + b$ cambia el intervalo donde la densidad es no nula, y modifica la densidad.

¿Cuál es el rol de a y b?

- b traslada el intervalo
- $f_Y(y) = \frac{1}{|a|}$, entonces a comprime o ensancha el intervalo dependiendo de su magnitud:
  - $|a| < 1 \Rightarrow f_Y(y) > f_X(x) \Rightarrow$ comprime el intervalo
  - $|a| > 1 \Rightarrow f_Y(y) < f_X(x) \Rightarrow$ ensancha el intervalo
- Si además $a < 0$, invierte el intervalo.

Ejercicio 1
Parte 1

Para visualizarlo podemos hacer **simulaciones**:

- Imitaciones de un proceso real.
- Usamos una función generadora de números "aleatorios" para crear una muestra de datos que sigue la distribución uniforme, la normal estándar, etc.

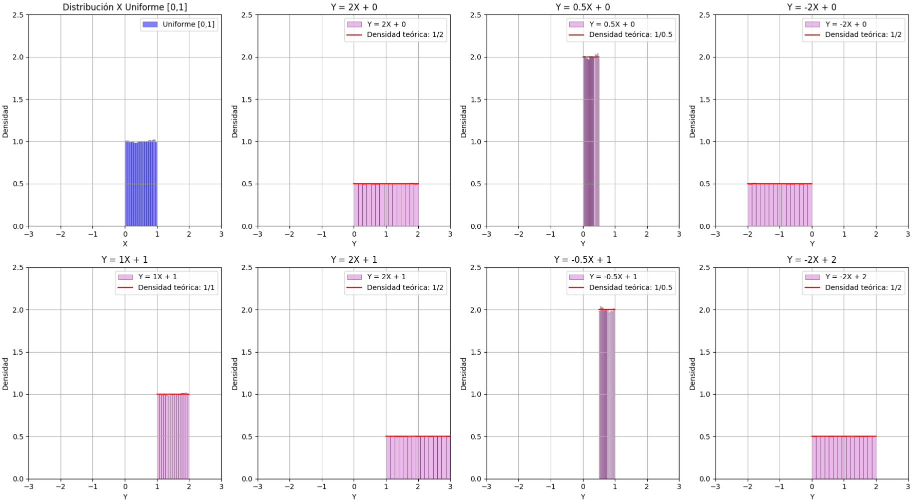
Veamos como hacerlo en Python (disponible en Laboratorios en R y Python/Laboratorios Python):

`https://colab.research.google.com/drive/16vHdPHTMmBFTyKbLbo0ZfZnXWmzBGnrB#scrollTo=YL2jqSehQa_T`

Disponible también en script de R (Laboratorios en R y Python/Laboratorios R)

Ejercicio 1
Parte 1

Simulación de uniformes:



Parte 2.

Sugerencia: Hacer lo mismo que hicimos en la parte 1, pero para g *estrictamente monótona, en lugar de* $g(X) = aX + b$.

Identificar en qué paso de la demostración usamos cada hipótesis!

Parte 3.

Sugerencia: Usar lo anterior.

Ejercicio 2: Densidades condicionales.

Sugerencias:

- *Hallar K que asegure que la función de densidad conjunta $f(x, y)$ es una densidad.*
- *Usar las definiciones vistas en el teórico.*
- *Hacer cuentas!*

Caso discreto:

$$\mathbb{E}(X \mid Y = y) = \sum_x x \cdot \mathbb{P}(X = x \mid Y = y) = \sum_x x \cdot \frac{\mathbb{P}(X = x, Y = y)}{\mathbb{P}(Y = y)}$$

Caso continuo:

$$\mathbb{E}(X \mid Y = y) = \int_{-\infty}^{+\infty} x \cdot f_{X|Y}(x \mid y)\, dx = \int_{-\infty}^{+\infty} x \cdot \frac{f(x, y)}{f_Y(y)}\, dx$$

Observaciones:

- $\mathbb{E}(X|Y) = g(Y)$ es una función de Y.
- Como Y es una variable aleatoria, $\mathbb{E}(X|Y)$ será una variable aleatoria que toma diferentes valores dependiendo de los valores que tome Y.
- $\mathbb{E}(X|Y = y)$ es un número: es el valor esperado de X cuando sabemos que Y ha tomado un valor particular y.

Ejemplo 1: Tirar un dado.

X variable aleatoria que toma valores [1,2,3,4,5,6] con probabilidad $1/6$.

$$\begin{aligned}\mathbb{E}(X) &= 1\cdot\mathbb{P}(X=1) + 2\cdot\mathbb{P}(X=2) + \cdots + 6\cdot\mathbb{P}(X=6)\\ &= \frac{1+2+3+4+5+6}{6} = 3{,}5\end{aligned}$$

Supongamos que Y es una variable aleatoria tal que

- Y toma valor 0 si el número que sale es par
- Y toma valor 1 si el número que sale es impar

Entonces

$$\begin{aligned}\mathbb{E}(X|Y=0) &= 2\cdot\mathbb{P}(X=2|Y=0) + 4\cdot\mathbb{P}(X=4|Y=0) + 6\cdot\mathbb{P}(X=6|Y=0)\\ &= 2\cdot\frac{\mathbb{P}(X=2,Y=0)}{\mathbb{P}(Y=0)} + 4\cdot\frac{\mathbb{P}(X=4,Y=0)}{\mathbb{P}(Y=0)} + 6\cdot\frac{\mathbb{P}(X=6,Y=0)}{\mathbb{P}(Y=0)}\\ &= 2\cdot\frac{1/6}{1/2} + 4\cdot\frac{1/6}{1/2} + 6\cdot\frac{1/6}{1/2}\\ &= 2\cdot\frac{1}{3} + 4\cdot\frac{1}{3} + 6\cdot\frac{1}{3}\\ &= 4\end{aligned}$$

De la misma forma tenemos:

$$\mathbb{E}(X|Y=1) = \frac{1+3+5}{3} = 3$$

Como Y toma solo los valores 0 y 1, tenemos

$$\mathbb{E}(X|Y) = \begin{cases} 4 \text{ si } Y = 0, \\ 3 \text{ si } Y = 1 \end{cases}$$

Escrito de otra forma...

$$\begin{aligned} \mathbb{E}(X|Y) &= 4(1-Y) + 3Y \\ &= 4 - Y \end{aligned}$$

es una función que depende de Y (toma sólo dos valores).

Observación: La variable Y nos aporta información adicional que podemos utilizar para **predecir** el valor de la variable X.

Ejemplo 2: Ingreso promedio en Uruguay.

Supongamos que estamos estudiando el ingreso promedio de las personas en Uruguay. Representamos el ingreso con una variable aleatoria X.

$\mathbb{E}(X)$ es el ingreso promedio en el país.

Supongamos que todos trabajamos.

Uruguay tiene 19 departamentos. Los representamos con la variable aleatoria Y que toma valores en $\{1, \cdots, 19\}$:

$$\mathbb{P}(Y = 1) = \mathbb{P}(\{\text{habitante sea de Artigas}\}) = \frac{\#\text{habitantes de Artigas}}{\#\text{habitantes de Uruguay}}$$

Puedo calcular el ingreso promedio de cada departamento, es decir:

$$\mathbb{E}(X|Y = y)$$

Esperanza condicional

Intuición:

- Si tuviera que **predecir** el ingreso de un habitante de Uruguay, utilizaría $\mathbb{E}(X)$.
- Si tuviera que **predecir** el ingreso de un habitante de Uruguay, sabiendo a qué departamento pertenece, sería más preciso utilizar $\mathbb{E}(X|Y)$.
- **Fórmula de la esperanza total (muy útil!):**

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y))$$

Ingreso promedio en Uruguay es igual al promedio de los ingresos promedio por departamento.

La fórmula de la esperanza total nos dice que podemos calcular la esperanza de X considerando cómo se comporta X dentro de cada "subgrupo"definido por Y, y luego promediando los resultados sobre todos los subgrupos, ponderando por la probabilidad de cada subgrupo.

Ejercicio 3: Descomposición de la varianza I.

Ejercicio 4: Descomposición de la varianza II.

Algunas fórmulas útiles.

- Descomposición de la varianza:

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X|Y)) + \mathrm{Var}(\mathbb{E}(X|Y))$$

- Esperanza de $g(X)$:
  Si X es absolutamente continua

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)\, dx$$

  Si X es discreta

$$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$$

Ejercicio 5: Concatenación de normales.
Para la parte 2 (simulaciones) hay código escrito en laboratorio (en R y Python).

Ejercicio 6: Esperanza condicional.

1. Si $\mathbb{E}(Y) = 2$, calcule $\mathbb{E}(X)$ sabiendo que $\mathbb{E}(X|Y) = -2Y + 3$.
2. Sea (X, Y) un vector aleatorio con función de densidad $f(x, y) = e^{-x-y}$. Calcule $\mathbb{E}(X|Y = y)$.
3. Sea X una variable aleatoria con $\mathbb{E}(X) = \mu$ e $Y = aX + b$ con a y b constantes. Calcule $\mathbb{E}(X|Y = y)$.
4. Si (X, Y) es un punto al azar en el triángulo $T = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0, x + y \leq 2\}$, calcule $P(Y > 1|X = x)$ y $\mathbb{E}(Y|X = x)$.

Ejercicio 6

Parte 1: Si $\mathbb{E}(Y) = 2$, calcule $\mathbb{E}(X)$ sabiendo que $\mathbb{E}(X|Y) = -2Y + 3$.

Usando la fórmula de la esperanza total:

$$\begin{aligned}\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(-2Y+3) &= -2\mathbb{E}(Y) + 3 \\ &= -4 + 3 \\ &= -1\end{aligned}$$

Parte 2: Sea (X, Y) un vector aleatorio con función de densidad $f(x, y) = e^{-x-y}$, $x \geq 0, y \geq 0$. Calcule $\mathbb{E}(X \mid Y = y)$.

Queremos calcular

$$\mathbb{E}(X \mid Y = y) = \int_0^\infty x \cdot f_{X|Y}(x \mid y)\, dx = \int_0^\infty x \cdot \frac{f(x,y)}{f_Y(y)}\, dx = \int_0^\infty x \cdot \frac{e^{-x-y}}{f_Y(y)}\, dx$$

Calculamos $f_Y(y)$

$$f_Y(y) = \int_0^\infty f(x,y)\, dx = \int_0^\infty e^{-x-y}\, dx = e^{-y}$$

y lo sustituimos en la fórmula de arriba

$$\mathbb{E}(X \mid Y = y) = \int_0^\infty x \cdot \frac{e^{-x-y}}{f_Y(y)}\, dx = \int_0^\infty x \cdot \frac{e^{-x-y}}{e^{-y}}\, dx = \int_0^\infty x \cdot e^{-x}\, dx = 1$$

Pensar **Partes 3 y 4**!
(Sugerencia: para la parte 4 hacer un dibujo.)

Ejercicio 7: Normal multivariada.
Sugerencia: Para la parte 2, recordar que en el caso de la distribución normal, la independencia y la no correlación son equivalentes.

Ejercicio 8: Combinación lineal de normales multivariadas.

Ejercicio 9: Distancia de Mahalanobis
En Python se llama igual que en R.

Viernes 23/8:

- Retomamos práctico 2 (sobre todo consultas).
- Práctico 3.