Respondiendo algunas preguntas que me hicieron

Pablo L. De Nápoli

Departamento de Matemática
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Probabilidades y Estadística para Matemática
Segundo cuatrimestre de 2021

Parte I

Calculando el área abajo de la curva normal

Planteo del problema

Consideremos la densidad normal estándar

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

Para ver que esta función es efectivamente una densidad de probabilidad debemos ver que

$$\int_{-\infty}^{\infty} f(x)\, dx = 1 \Leftrightarrow \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi} \Leftrightarrow \int_{0}^{\infty} e^{-x^2}\, dx = \sqrt{\frac{\pi}{2}}$$

Pero el integrando no admite una primitiva explícita en términos de funciones elementales, ¿cómo la calculamos? Como vimos en la clase 8, haciendo el cambio de variable: $x = y^2/2 \Rightarrow y = \sqrt{2x}, dx = y\, dy$,

$$\Gamma(1/2) = \int_0^{\infty} x^{-1/2}\, e^{-x}\, dx = \int_0^{\infty} \left(\frac{y^2}{2}\right)^{-1/2} e^{-y^2/2}\, y\, dy = \sqrt{2} \int_0^{\infty} e^{-y^2/2} dy$$

Entonces nuestra afirmación es equivalente a ver que $\Gamma(1/2) = \sqrt{\pi}$.

Solución usando lo que vimos en la clase anterior

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}\, e^{-x}\, dx = \lim_{\substack{R\to+\infty \\ r\to 0+}} \int_r^R x^{\alpha-1}\, e^{-x}\, dx \;\; \alpha > 0 \text{ función Gama de Euler}$$

$$B(\alpha_1, \alpha_2) = \int_0^1 (1-u)^{\alpha_1-1}\, u^{\alpha_2-1}\, du \quad \alpha_1, \alpha_2 > 0 \text{ función Beta de Euler}$$

La clase pasada probamos que:

$$B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

Elegimos $\alpha_1 = \alpha_2 = 1/2$

$$B(1/2, 1/2) = \frac{\Gamma(1/2)\Gamma(1/2)}{\Gamma(1)} = \Gamma(1/2)^2$$

Luego

$$\Gamma(1/2) = \sqrt{\pi} \Leftrightarrow B(1/2, 1/2) = \pi$$

Solución usando lo que vimos en la clase anterior (2)

Veamos entonces que

$$B(1/2,1/2) = \int_0^1 (1-u)^{-1/2}\, u^{-1/2}\, du = \pi$$

.

Hacemos el cambio de variable

$$u = \text{sen}^2\,\theta, du = 2\text{sen}\theta\ \cos\theta\ d\theta$$

$$(1-u)^{-1/2} = \frac{1}{\sqrt{1-\text{sen}^2\,\theta}} = \frac{1}{\cos\ \theta}$$

$$B(1/2,1/2) = \int_0^{\pi/2} \frac{1}{\cos\ \theta} \cdot \frac{1}{\text{sin}\ \theta} \cdot 2\text{sen}\theta\ \cos\theta\ d\theta = \int_0^{\pi/2} 2\ d\theta = \pi$$

¿Cómo se calcula la distribución acumulada de una variable $N(0,1)$?

La función error se define por

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$$

Desarrollando el integrando en serie de Taylor e integrando término a término, se encuentra la serie de Taylor para la función error:

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n z^{2n+1}}{n!(2n+1)} = \frac{2}{\sqrt{\pi}} \left(z - \frac{z^3}{3} + \frac{z^5}{10} - \frac{z^7}{42} + \frac{z^9}{216} - \cdots \right)$$

Usando la simetría de la densidad normal y un cambio de escala, se ve que la función de distribución acumulada de una variable $N(0,1)$ es

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-t^2}{2}} dt = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{x}{\sqrt{2}} \right) \right]$$

Parte II

Convoluciones discretas

Convoluciones discretas

Consideramos dos variables discretas X e Y independientes con valores enteros. La distribuciones puntuales vienen dadas por las sucesiones

$$p_k = p(k) = P\{X = k\}, q_k = q(k) = P\{Y = k\} \quad k \in \mathbb{Z}$$

que podemos pensar como funciones $p, q : \mathbb{Z} \to \mathbb{R}$. ¿Cuál es la distribución de $X + Y$?
Imitando la definición del caso continuo que vimos la clase pasada, definimos la convolución discreta de p y q por

$$(p * q)(k) = \sum_{m,n \in Z : m+n=k} p(m) \cdot q(n) = \sum_{m \in \mathbb{Z}} p(m) \cdot q(k-m)$$

Si X e Y toman valores naturales con probabilidad 1 (en $\mathbb{N}_0$ la fórmula se simplifica

$$(p * q)(k) = \sum_{m=0}^{k} p(m) \cdot q(k-m)$$

Convoluciones discretas (2)

Como en el caso continuo podemos probar

Proposición

La distribución puntual de probabilidades de $X + Y$ viene dada por $p * q$ es decir

$$P\{X + Y = k\} = (p * q)(k)$$

Demostración.

$$\begin{aligned} P\{X + Y = k\} &= \sum_{m,n:m+n=k} P\{X = m, Y = n\} \\ &= \sum_{m,n:m+n=k} P\{X = m\} \cdot P\{Y = n\} \text{ por independencia} \\ &= \sum_{m,n:m+n=k} p(m) \cdot q(n) = (p * q)(k) \end{aligned}$$

Suma de variables independientes con distribución de Poisson

Por ejemplo, redemostremos el siguiente resultado que vimos en la clase 7

Proposición

Si $X \sim \mathcal{P}(\lambda_1)$, $Y \sim \mathcal{P}(\lambda_2)$ y son independientes, entonces $X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$.

Demostración.

$$(p * q)(k) = \sum_{m=0}^{k} p(m) \cdot q(k-m) = \sum_{m=0}^{k} \frac{\lambda_1^m}{m!} \cdot e^{-\lambda_1} \cdot \frac{\lambda_2^{k-m}}{(k-m)!} \cdot e^{-\lambda_2}$$

$$= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{m=0}^{k} \frac{k!}{m! \cdot (k-m)!} \cdot \lambda_1^m \cdot \lambda_2^{k-m}$$

$$= e^{-(\lambda_1+\lambda_2)} \frac{1}{k!} \sum_{m=0}^{k} \binom{k}{m} \cdot \lambda_1^m \cdot \lambda_2^{k-m} = e^{-(\lambda_1+\lambda_2)} \cdot \frac{(\lambda_1+\lambda_2)^k}{k!}$$

Parte III

Distribuciones empíricas

Distribuciones empíricas

Al aplicar las ideas que estuvimos viendo a problemas con datos reales, se suele presentar la siguiente situación: tenemos un conjunto de datos

$$x_1, x_2, \ldots, x_n$$

que son una muestra tomada de una población cuya distribución no conocemos.
En este caso, la distribución empírica es la distribución que observamos en la muestra.
Algunos de los valores pueden repetirse. Llammos

$$\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k$$

a los valores observados pero sin repetición (entonces $k \leq n$). Esta manera de considerar los mismos datos se llama datos agrupados.
La distribución empirica es una distribución discreta donde la probabilidad de cada valor $\hat{x}_j$ es la correspondiente recuencia relativa observada en la muestra.

$$P\{X = \hat{x}_j\} = f_j = \frac{\#\{i : x_i = \hat{x}_j\}}{n}$$

Un ejemplito para entender la notación

Consideramos un estudiante cuyas notas en la carrera fueron

$$x_1 = 7, x_2 = 8, x_3 = 10, x_4 = 7, x_5 = 10, x_6 = 9, x_7 = 10, x_8 = 10, x_9 = 5, x_{10} = 10$$

entonces $n = 10$, pero sus valores únicos son

$$\hat{x}_1 = 7, \hat{x}_2 = 8, \hat{x}_3 = 10, \hat{x}_4 = 9, \hat{x}_5 = 5$$

(luego $k = 5$)y sus correspondientes fecuencias son

$$f_1 = \frac{2}{10}, f_2 = \frac{1}{10}, f_3 = \frac{5}{10}, f_4 = \frac{1}{10}, f_5 = \frac{1}{5}$$

Distribuciones empíricas (2)

Dada una muestra, la función de distribución de su distribución empírica puede escribirse como

$$F_X(x) = P\{X \leq x\} = \sum_{\hat{x}_j \leq x} f_j = \frac{\#\{i : x_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$$

Su esperanza será

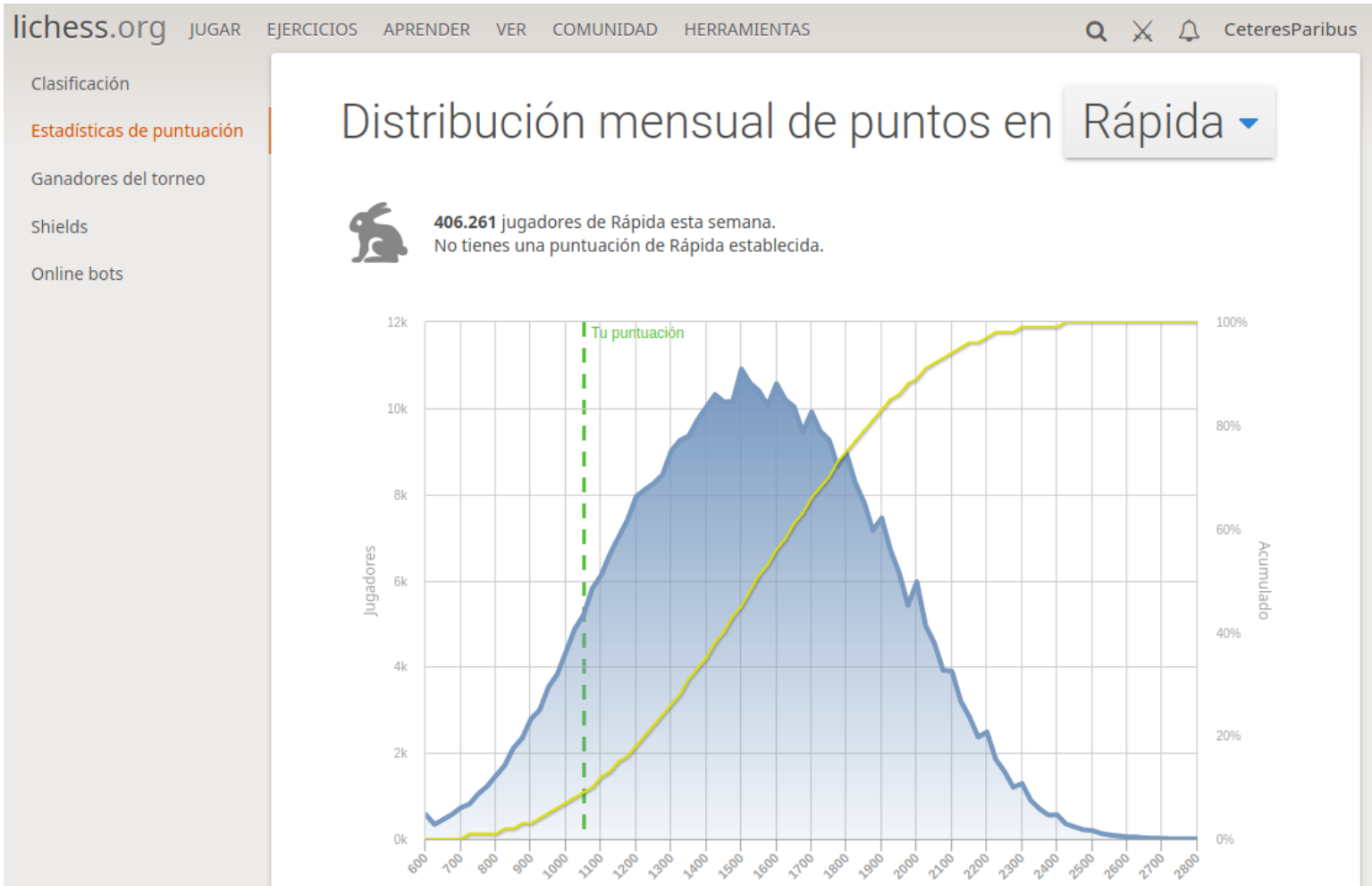
$$\overline{x} = \sum_{j=1}^{k} \hat{x}_j \cdot f_j = \frac{1}{n}\sum_{i=1}^{n} x_i$$

(es el promedio de los datos observados) y su varianza

$$\sigma^2 = \sum_{j=1}^{k} (\hat{x}_j - \overline{x})^2 \cdot f_j = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2$$

¡Vamos con un ejemplo real!

La página Lichess es una página para jugar al ajedrez en linea. En
`https://lichess.org/stat/rating/distribution/rapid`
me muestra la distribución de ELO (un parámetro usado en ajedrez para medir la habilidad relativa de los jugadores) en partidas rápidas esta semana.



¡Lo programamos!

Mirando el código fuente de la página, me bajé la lista de las frecuencias abosolutas observadas, y con ellas armé la distribución empírica.

Programita en Python, usando SciPy

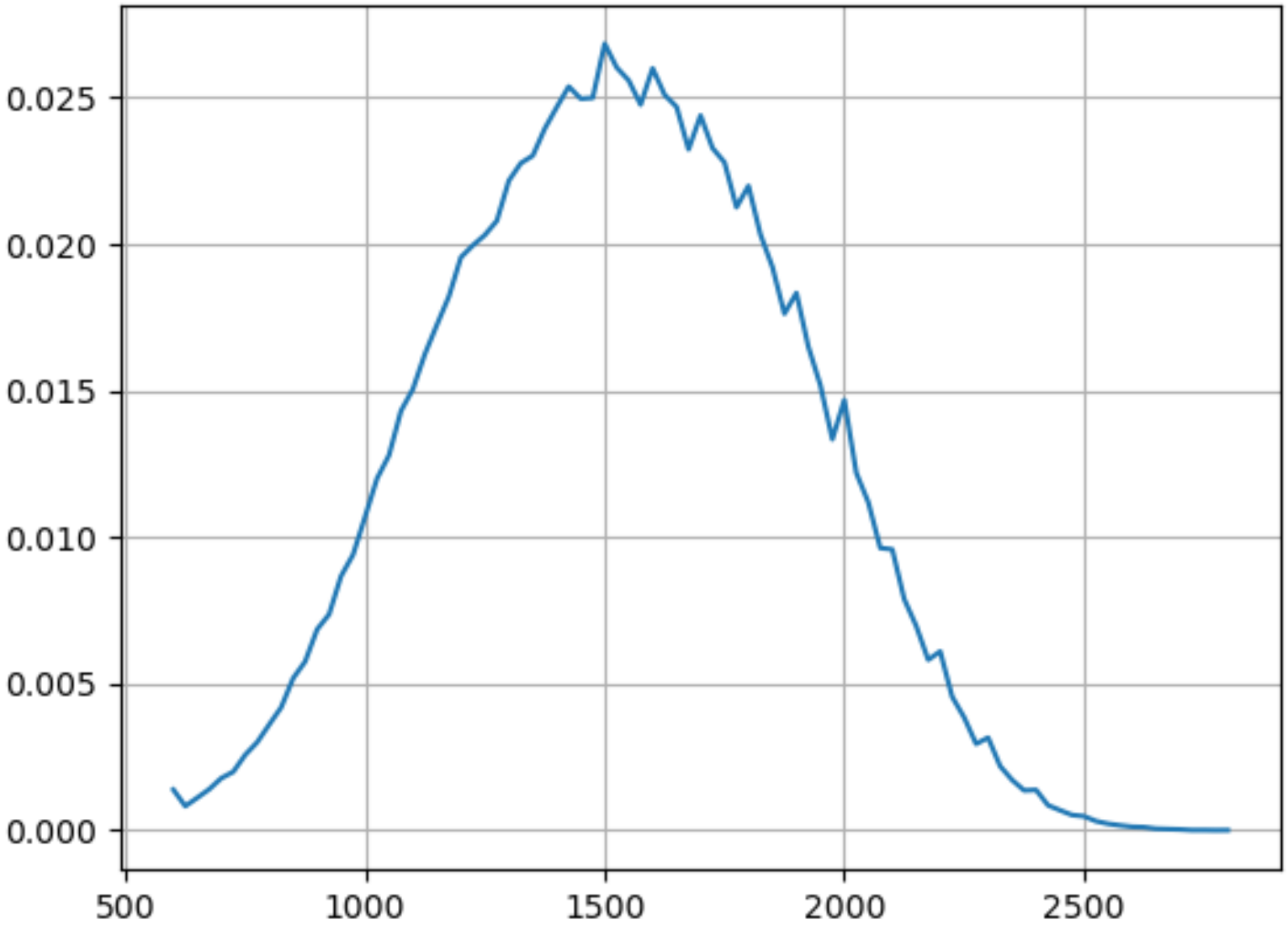
```
freq = [ 566, 337, 453, 572, 724, 813, 1048, 1223, 1467,
1706, 2110, 2341, 2787, 2996, 3526, 3823, 4347, ... ]

freq = np.array(freq)
delta_elo = 25
elo_final = 600 + len(freq) * delta_elo
elos = np.arange(600, elo_final, delta_elo)

total_jugadores = np.sum(freq)
freq_rel = freq / total_jugadores

data = scipy.stats.rv_discrete(values=(elos, freq_rel))
plt.plot(elos, freq_rel)
```

¡Veamos el gráfico!



Obteniendo los parámetros estadísticos de la distribución

Programita en Python, usando SciPy

```
print("Número total de jugadores=", total_jugadores)
print("media=", data.mean())
print("varianza=", data.var())
print("desviación estándar=", data.std())
print("mediana=", data.median())
q1 = data.ppf(0.25)
q3 = data.ppf(0.75)
iqr = q3 - q1
print("iqr=", iqr)
max_f = np.amax(freq_rel)
print("máxima frecuencia relativa=", max_f)
donde = np.where(freq_rel == max_f)
print("moda=", elos[donde])
```

Obteniendo los parámetros estadísticos de la distribución

Salida del programa

```
Número total de jugadores= 406261
media= 1531.5219157142824
varianza= 127980.78275326267
desviación estándar= 357.74401847307337
mediana= 1525.0
iqr= 525.0
máxima frecuencia relativa= 0.026844811586640115
moda= [1500]
```

Parte IV

Teoría de la predicción: Interpretación del coeficiente de correlación

Recordamos algunas definiciones

Dadas dos variables X e Y la covariancia entre ellas se define como

$$\text{Cov}(X, Y) = E[(X - \mu_X) \cdot (Y - \mu_Y)]$$

donde $\mu_X = E[X]$ y $\mu_Y = E[Y]$, y el coeficiente de correlación como

$$\rho = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}}$$

La desigualdad de Cauchy-Schwarz nos dice que $0 \leq |\rho| \leq 1$, o sea $-1 \leq \rho \leq 1$.

Me preguntaron en las clases anteriores sobre el significado de estos conceptos, así que hoy intentaremos aclararlo.

Como les dije ρ mide en qué medida Y se puede pensar como una función lineal de X (y viceversa). Hoy intentaremos precisar este concepto. Vamos a seguir esencialmente el capítulo 8 del apunte de Victor Yohai.

Contexto abstracto en el que vamos a trabajar

Consideramos un espacio de probabilidad $(\Omega, \mathcal{E}, P)$. Consideramos el espacio vectorial de las variables aleatorias con segundo momento finito

$$L^2(\Omega) = \{\text{variables aleatorias } X : \Omega \to \overline{\mathbb{R}} : E(X^2) < \infty\}$$

Recordamos que si $X \in L^2(\Omega)$,

$$E(|X|) \leq E(X^2)^{1/2} \text{ por la desigualdad de Jensen}$$

y

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

Por lo que las variables aleatorias en L^2 tienen esperanza y varianza finitas. $L^2(\Omega)$ es un espacio normado con la norma

$$\|X\| = E(X^2)^{1/2}$$

que proviene del producto interno

$$\langle X, Y \rangle = E(X \cdot Y)$$

Es un espacio con producto interno o espacio pre-Hilbert.

Contexto abstracto en el que vamos a trabajar (2)

Para que $L^2(\Omega)$ sea realmente un espacio vectorial normado, hay considerar iguales a las variables aleatorias X e Y tales que

$$P\{X = Y\} = 1$$

Con esta convención,

$$\|X\| = 0 \Rightarrow X = 0$$

Planteo del problema

Consideramos dentro de L^2 un subespacio S. Queremos aproximar una variable aleatoria Y por un elemento del subespacio $\widehat{Y} \in S$.
Particularmente, vamos a usar dos subespacios:

$$S_1 = \text{variables aleatorias constantes} = \langle 1 \rangle$$

y dada una variable aleatoria X vamos a considerar

$$S_2 = \{\alpha X + \beta : \alpha, \beta \in \mathbb{R}\} = \langle 1, X \rangle$$

La idea es que queremos usar $\widehat{Y}$ para predecir el valor de Y, por eso en la teoría de probabilidades se lo llama un predictor de Y.

¿Cuál es la mejor manera de elegir $\widehat{Y}$? Eso depende de cómo midamos el error en la aproximación. Vamos a usar el criterio del error cuadrático medio. Queremos minimizar

$$\text{ECM}(Y, \widehat{Y}) = E(|Y - \widehat{Y}|^2) = \|Y - \widehat{Y}\|^2$$

Un lema de álgebra lineal

Lema

Sea V un espacio con producto interno y $S \subset V$ un subespacio. Consideramos $x_0 \in V$. Entonces $s_0 \in S$ es el elemento de S que minimiza la distancia a x_0

$$d(x, s) = \|x - s\| \quad x \in S$$

si y sólo si s_0 es la proyección ortogonal de x_0 sobre s es decir:

$$x_0 - s_0 \in S^{\perp}$$

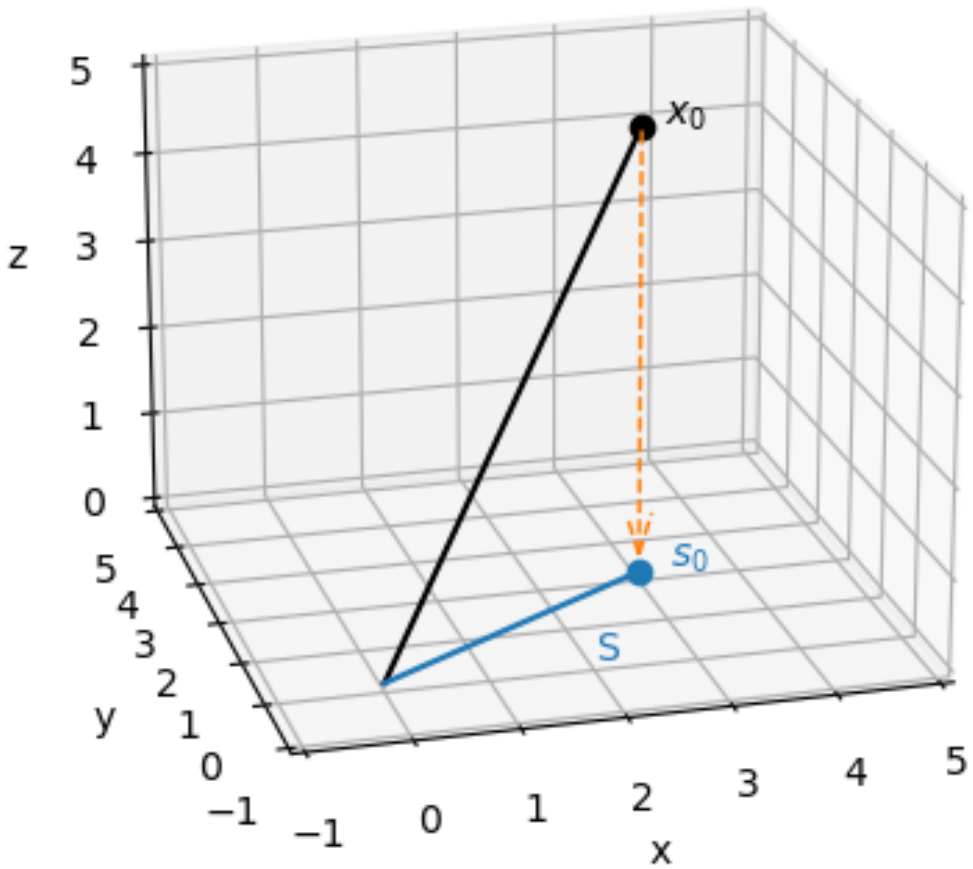
o sea

$$\langle x_0 - s_0, s \rangle = 0 \text{ para todo } s \in S \tag{1}$$

Nuestro $V = L^2(\Omega)$ es un espacio de dimensión infinita, pero este lema funciona exactamente igual que en dimensión finita (y con la misma prueba).

Si S fuera de diemnsión finita, es suficiente verificar la condición de ortogonalidad (**??**) para s en una base de S.

Ilustración gráfica de la proyección ortogonal



Predicción por variables aleatorias constantes

Apliquémoslo al primero de nuestros ejemplos

$$S_1 = \text{variables aleatorias constantes} = \langle 1 \rangle$$

(subespacio de dimensión 1).
Dada $Y \in L^2$, la condición para que $\widehat{Y} \in S_0$ sea el predictor constante que minimiza el error medio cuadrático es según el lema (con $S = S_0$, $x_0 = Y$,$s_0 = \widehat{Y}$):

$$\langle Y - \widehat{Y}, 1 \rangle = 0$$

o sea:

$$E[(Y - \widehat{Y}) \cdot 1] = 0$$

Como $\widehat{Y}$ es constante, esto nos dice que el mejor predictor de Y es:

$$\widehat{Y_0} = E[Y]$$

y entonces el error medio cuarático en esta aproximación será

$$\text{EMC}_1 = \min_{\widehat{Y} \in S_1} \|Y - \widehat{Y}\|^2 = \|Y - \widehat{Y}_0\|^2 = E[(Y - \widehat{Y_0})^2] = \text{Var}(Y)$$

Predicción por funciones lineales de X

Ahora dada otra variable aleatoria X, consideramos

$$S_2 = \{\alpha X + \beta : \alpha, \beta \in \mathbb{R}\} = \langle 1, X \rangle$$

(subespacio de dimensión 2). Según el lema, las condiciones de ortogonalidad que debe verificar el predictor óptimo son:

$$\langle Y - \widehat{Y}, 1 \rangle = 0$$

$$\langle Y - \widehat{Y}, X \rangle = 0$$

o sea:

$$E[(Y - \widehat{Y}) \cdot 1] = 0$$

$$E[(Y - \widehat{Y}) \cdot X] = 0$$

Entonces los coeficientes α, β para el predictor óptimo deben satisfacer que

$$E[(Y - \alpha X - \beta) \cdot 1] = 0$$

$$E[(Y - \alpha X - \beta) \cdot X] = 0$$

Predicción por funciones lineales de X (2)

La primera condición dice que

$$E[Y] - \alpha E[X] - \beta = 0 \tag{2}$$

También multiplicándola por $E[X]$ obteneos que

$$E[(Y - \alpha X - \beta) \cdot E[X]] = 0$$

y entonces restándola de la segunda condición

$$E[(Y - \alpha X - \beta) \cdot (X - E(X))] = 0$$

Reemplazando el valor de β dado por (**??**),

$$E[(Y - \alpha X - E(Y) - \alpha E(X)) \cdot (X - E(X))] = 0$$

por lo tanto

$$E[(Y - E(Y)) - \alpha(X - E(X))] \cdot (X - E(X))] = 0$$

Entonces distribuyendo la esperanza, obtenemos

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E[X]) \cdot (Y - E(Y))] \\ &= \alpha \, E[(X - E(X))^2] \\ &= \alpha \text{Var}(X) \end{aligned}$$

Predicción por funciones lineales de X (3)

En resumen, hemos demostrado

Teorema

Sea $\widehat{Y}_0$ el predictor de menor error cuadrático medio en S_2. Vine dado por $\widehat{Y}_0 = \alpha X + \beta$ donde α y β se determinan por las ecuaciones:

$$\alpha = \frac{Cov(X, Y)}{Var(X)}$$

$$\beta = E(Y) - \alpha E[X]$$

Cálculo del error cuadrático medio

Calculemos el error cuadrático medio óptimo al aproximar Y por una función lineal de X.

$$\text{EMC}_2 = \min_{\widehat{Y} \in S_2} \|Y - \widehat{Y}\|$$

Primero usamos que en el predictor óptimo $\beta = E(Y) - \alpha E[X]$

$$\begin{aligned} ECM_2 &= \|Y - \widehat{Y}_0\|^2 = E[(Y - \widehat{Y}_0)^2] = E[(Y - \alpha X - \beta)^2] \\ &= E[(Y - \alpha X - (E[Y] - \alpha E[X]))^2] \\ &= E[((Y - E(Y)) - \alpha(X - E(X)))^2] \\ &= E[(Y - E(Y))^2] + \alpha^2 E[(X - E(X))^2] - 2\alpha E[(Y - E(Y)) \cdot (X - E(X))] \end{aligned}$$

Y usando que $\alpha = \frac{Cov(X,Y)}{\text{Var}(X)}$ en el predictor óptimo,

$$\begin{aligned} ECM_2 &= \text{Var}(Y) + \alpha^2 \text{Var}(X) - 2\alpha \text{Cov}(X, Y) \\ &= \text{Var}(Y) + \frac{\text{Cov}^2(X, Y)}{\text{Var}(X)} - 2\frac{\text{Cov}^2(X, Y)}{\text{Var}(X)} \\ &= \text{Var}(Y) - \frac{\text{Cov}^2(X, Y)}{\text{Var}(X)} \end{aligned}$$

Mejora en el error medio cuadático

Queremos comparar cuánto mejoró el error medio cuadrático al usar como predictor de Y una función lineal de X comparado con usar una variable aleatoria constante. Para ello consideramos el cociente

$$\frac{EMC_1 - EMC_2}{ECM_1} = \frac{\text{Var}(Y) + \frac{\text{Cov}^2(X,Y)}{\text{Var}(X)} - \text{Var}(Y)}{\text{Var}(Y)}$$

$$= \frac{\text{Cov}^2(X,Y)}{\text{Var}(X)\text{Var}(Y)} = \rho^2(X,Y)$$

Esto permite interpretar el coeficiente $\rho^2(X,Y)$ como el decrecimiento relativo del error cuadrático medio cuando se usa un predictor lineal basado en X en vez de un predictor constante. Por lo tanto $\rho^2(X,Y)$ mide la utilidad de la variable X para predecir a Y por una función lineal.

Algunas observaciones

Notamos que como $S_1 \subset S_2$, $\text{ECM}_2 \leq \text{ECM}_1$. Esto nos dice nuevamente que $|\rho(X, Y)| \leq 1$, o sea nos proporciona otra prueba de la desigualdad de Cauchy-Schwarz.

¿Qué significaría $|\rho| = 1$?. Según la fórmula anterior, esto implica que

$$\text{ECM}_1 - \text{ECM}_2 = \text{ECM}_1 \Rightarrow \text{ECM}_2 = 0$$

o sea que Y es una función lineal de X.

Notamos también que como para el predictor óptimo

$$\alpha = \frac{Cov(X, Y)}{\text{Var}(X)}$$

el signo de $\rho(X, Y)$ coincide con el signo de α. [Si $\rho(X, Y) > 0$ el predictor óptimo será una función lineal creciente de X, mientras que si $\rho < 0$ será una función lineal decreciente]