Distribución gaussiana (detalle de su derivación)

Durante la última clase obtuvimos la forma funcional de la distribución de probabilidad gaussiana $p(x)$ a partir de 3 hipótesis sencillas. Con un poco de álgebra, llegamos a una expresión de la forma:

$$p(x) = A\, e^{-\frac{1}{2}kx^2}, \tag{1.0.1}$$

que permite determinar la probabilidad de hallar un valor de la variable entre x y $x + \Delta x$ calculando el producto $p(x)\Delta x$.

Notemos, no obstante, que para lograrlo nos queda todavía determinar las formas explícitas para los parámetros A y k. Dado que tenemos aún dos incógnitas que calcular, nos hacen falta dos ecuaciones (es decir, dos condiciones) para determinarlas.

La primera de tales condiciones surge del hecho de que la probabilidad total debe ser igual a 1. Matemáticamente, esta condición se escribe:

$$\int_{-\infty}^{\infty} p(x)\, dx = 1. \tag{1.0.2}$$

Por otro lado, podemos obtener una segunda condición a partir de exigir que la varianza calculada a partir de esa distribución de probabilidad $p(x)$ coincida con el valor de la varianza de la distribución de datos:

$$\int_{-\infty}^{\infty} (x - \bar{x})^2\, p(x)\, dx = \sigma_x^2. \tag{1.0.3}$$

En clase les mencioné que, utilizando la expresión funcional para la distribución gaussiana que obtuvimos (1.0.1) en estas dos ecuaciones es posible obtener las formas explícitas para A y k, que resultan ser:

$$A = \frac{1}{\sqrt{2\pi}\,\sigma_x}, \quad \text{y} \tag{1.0.4}$$

$$k = \frac{1}{\sigma_x^2}. \tag{1.0.5}$$

No obstante, en clase optamos por no mostrar explicitamente este resultado dado que se trata de un cálculo largo (y quizás tedioso) que, conceptualmente, no agrega mayor significado a nuestro resultado. El propósito de este documento es mostrar cómo realizar dicho cálculo, para quienes deseen saber cómo llevarlo a cabo.

Determinación del parámetro A

Según dijimos, para que la función $p(x)$ pueda considerarse una distribución de probabilidad, es necesario que el área total debajo de la curva que ella traza sea igual a 1. En otras palabras, debemos *imponer* dicha condición para hacer que A cumpla dicho requisito. Concretamente, tenemos

$$\int_{-\infty}^{\infty} p(x)\,dx = \int_{-\infty}^{\infty} A\, e^{-\frac{1}{2}kx^2}\,dx = 1,$$

lo que implica que debe valer:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}kx^2}\,dx = \frac{1}{A}.$$

Debido a la simetría que presenta el integrando (que es una función par de su argumento, i.e, que vale $p(x) = p(-x)$ para él), calcular la integral sobre los reales debe dar el mismo resultado que calcularla sobre los números reales positivos y multiplicarla por 2:

$$\int_{0}^{\infty} e^{-\frac{1}{2}kx^2}\,dx = \frac{1}{2A}.$$

Para simplificar aún más esta expresión, introducimos el cambio de variables $z = \sqrt{k/2}\,x$, que nos lleva a

$$\int_{0}^{\infty} e^{-z^2}\,dz = \frac{1}{2A}\sqrt{\frac{k}{2}}. \tag{1.1.6}$$

Toda la dificultad aquí yace en calcular la integral del lado izquierdo de la última igualdad. Para ello existen una gran variedad de métodos; para estas notas elijo aquel que permite lograrlo con las herramientas que provee cualquier curso de calculo básico. En particular, vamos a utilizar la técnica de diferenciación bajo el signo integral, que además les puede resultar útil para calcular integrales más allá de este curso. Para eso, consideremos la función

$$f(t) = \int_{0}^{\infty} \frac{e^{-t^2(1+x^2)}}{1+x^2}\,dx,$$

cuyo argumento t asumimos que es un número real. Notemos que, así definida, la función f no depende de la variable x. La razón por la que vamos a considerar esta función particular va a hacerse evidente más adelante. Observemos, en primer lugar, que si derivamos esta función (respecto de su único argumento t), obtenemos:

$$f'(t) = -2te^{-t^2} \int_0^\infty e^{-(tx)^2}\, dx.$$

Si hacemos la sustitución $z = tx$, resulta $dz = t\,dx$, de forma que

$$f'(t) = -2e^{-t^2} \underbrace{\int_0^\infty e^{-y^2}\, dy}_{\equiv G}.$$

Observen que la integral que aparece del lado derecho en esta expresión es justamente la que buscamos calcular (es el lado izquierdo de la ecuación 1.1.6), cuyo valor no depende de la variable t (ni de y). Dado que se trata de una constante, asignémosle un nombre simple: G. Podemos ahora integrar esta última expresión respecto de t en el intervalo [0,T]:

$$\int_0^T f'(t)\, dt = -2G \int_0^T e^{-t^2}\, dt.$$

De acuerdo al teorema fundamental del cálculo, la integral a la izquierda es igual a $f(T) - f(0)$, de lo se obtiene que

$$f(T) - f(0) = -2G \underbrace{\int_0^T e^{-t^2}\, dt}_{G} = -2G^2.$$

Si ahora consideramos el límite cuando $T \to \infty$ y despejamos G, llegamos a

$$G^2 = \frac{1}{2}\left[f(0) - \lim_{T\to\infty} f(T)\right],$$

y sólo nos resta evaluar los términos dentro del corchete del lado derecho. Comencemos por el primero. La función f evaluada en $t = 0$ se reduce, por definición, a $\int_0^\infty dx/(1+x^2)$; cuya primitiva es la arcotangente. Resulta entonces

$$f(0) = \int_0^\infty \frac{1}{1+x^2}\, dx = \arctan(\infty) - \arctan(0) = \frac{\pi}{2} - 0 = \frac{\pi}{2}.$$

Por otro lado, es fácil ver que el segundo sumando dentro del corchete tiende a cero. Para hacerlo evidente basta con considerar el límite cuando t tiende a infinito de la función $f(t)$ haciendo el cambio de variables $y = tz$, luego $dy = t\,dz$. En ese caso tenemos

$$\lim_{t\to\infty} f(t) = \lim_{t\to\infty} \int_0^t \frac{e^{-t^2(1+x^2)}}{1+x^2}\, dx = \lim_{t\to\infty} \left(te^{-t^2} \int_0^t \frac{e^{-y^2}}{t^2+x^2}\, dy \right) \leqslant \lim_{t\to\infty} \frac{e^{-t^2}}{t} G,$$

luego como G es una cantidad finita y t es positivo, sabemos que $\lim_{t\to\infty} f(t) = 0$.

Finalmente, hemos obtenido el resultado anticipado en clase [1.0.4],

$$A = \sqrt{\frac{k}{2\pi}}, \tag{1.1.7}$$

que es simplemente una relación entre los parámetros A y k que nos garantiza que la función $p(x)$ tiene una integral igual a 1 ó, en otras palabras, que $p(x)$ se encuentra *normalizada*. Con este resultado, nuestra distribución de probabilidad resulta:

$$p(x) = \sqrt{\frac{k}{2\pi}}\, e^{-\frac{1}{2}kx^2},$$

y aún nos queda pendiente determinar el valor del parámetro k que nos asegure que la integral de $(x-\bar{x})^2 p(x)$ sobre todos los valores posibles de la variable coincida con su varianza σ_x^2. Esto lo haremos en la siguiente sección.

Determinación del parámetro k

A fin de determinar el parámetro k (único parámetro restante), impongamos la segunda condición discutida en la sección inicial. Calculemos entonces la integral

$$\int_{-\infty}^{\infty} (x - \bar{x})^2\, p(x)\, dx,$$

e igualemos su resultado a la varianza de la variable x, σ_x^2.

El cuadrado del binomio en el integrando da lugar a 3 términos: uno con x^2, otro con $\bar{x}^2$ y un tercero de la forma $x\bar{x}$. Dado que en este caso el valor medio $\bar{x} = 0$ por construcción, la integral asociada al segundo término es nula. Por otro lado, puesto que $xp(x)$ es una función impar, la integral correspondiente al tercer término del binomio también se anula. Nos resta únicamente determinar la primera integral, que podemos reexpresar como

$$\int_{-\infty}^{\infty} x^2\, p(x)\, dx = 2\int_{0}^{\infty} x^2\, p(x)\, dx.$$

Sustituyendo por la forma funcional de $p(x)$, tenemos entonces la condición

$$2\sqrt{\frac{k}{2\pi}}\int_{0}^{\infty} x^2\, e^{-\frac{1}{2}kx^2}\, dx = \sigma_x^2.$$

La integral puede ahora evaluarse por partes usando $u = x$ y $dv = x\exp(-1/2kx^2)$, de lo que obtenemos

$$2\sqrt{\frac{k}{2\pi}}\left[\lim_{L\to\infty}\left(-\frac{x}{k}\, e^{-\frac{1}{2}kx^2}\right)\Bigg|_0^L + \frac{1}{k}\int_0^{\infty} e^{-\frac{1}{2}kx^2}\, dx\right] = \sigma_x^2.$$

El primer término entre corchetes es nulo, y del segundo conocemos su valor dado que hemos calculado la integral en la sección anterior, luego

$$2\sqrt{\frac{k}{2\pi}}\int_0^\infty x^2\, e^{-\frac{1}{2}kx^2}\, dx = \left(2\sqrt{\frac{k}{2\pi}}\right)\left(\frac{1}{2}\frac{1}{k}\sqrt{\frac{2\pi}{k}}\right)$$

de lo que resulta que la integral buscada vale k^{-1}. Por ende la relación entre el parámetro k y la varianza σ_x^2, impuesta por la condición (1.0.3), es

$$k = \frac{1}{\sigma_x^2},$$

coincidente con el resultado anticipado en clase, dado por la ecuación (1.0.5).

Distribución de Gauss completa

En función de lo calculado precedentemente, la expresión para la distribución de probabilidad $p(x)$ resulta:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x}\exp\left[-\frac{1}{2\sigma_x^2}x^2\right].$$

Ahora bien, esta distribución de probabilidad tiene media nula (según se observa de su paridad). No obstante resulta sencillo generalizar nuestro resultado para obtener una expresión que modele una distribución con media no nula: bastará con realizar un cambio de origen de coordenadas del tipo $x \to x - \bar{x}$. En ese caso nos queda:

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma_x}\exp\left[-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma_x}\right)^2\right].$$

El lector podrá comprobar fácilmente, si lo desea, que con esta forma para $p(x)$,fs: el valor medio de x coincide con $\bar{x}$, como resulta esperable.

Tenemos entonces la forma completa de la distribución de probabilidad de Gauss o *normal*, derivada a partir de 3 hipótesis sencillas.