Variables aleatorias: El caso continuo

Gabriel Cavada Ch.[1]
[1]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*

Random variables: The continuous case

El tratamiento de una variable aleatoria continua es matemáticamente más complejo que el de una variable aleatoria discreta, debido a la completitud de los números reales, lo que significa que cualquier subconjunto de números reales tiene infinitos elementos y sus elementos no son enumerables; en consecuencia, si consideramos una variable de naturaleza continua, como por ejemplo el peso de una persona adulta, de sexo femenino y sana, que en una determinada población la podríamos situar entre los 40 y 70 kilógramos, es obvio asumir que una persona incluida en esta población, puede tener como peso cualquier valor comprendido en la dispersión dada. Si pensamos en el experimento consistente en "extraer una persona al azar y pesarla" y definimos el suceso A: "la persona pesa exactamente 48 kilógramos", entonces la P(A) = 0, ya que se trata de escoger un solo valor de un espacio muestral que tiene infinitos elementos. Este hecho nos lleva a renunciar a calcular la probabilidad de un evento como A, en cambio, nos podemos preguntar por la probabilidad del siguiente evento: B: "la persona escogida pesa entre 50 y 55 kilógramos", es claro que si contáramos con evidencia empírica, es decir con datos, que pudiésemos graficar, la estimación de la probabilidad del evento B, sería el área más oscura en la Figura 1.

Esta área es aproximadamente el 41% del total, es decir P(B) = 0,41; pues bien, si quisiéramos formalizar esta idea, tendríamos que conocer la función matemática en cuyo trazado se circunscribe el histograma mostrado en la Figura 1, es decir, deberíamos conocer la expresión matemática de la función mostrada en la Figura 2 y calcular el área destacada.

Figura 1.

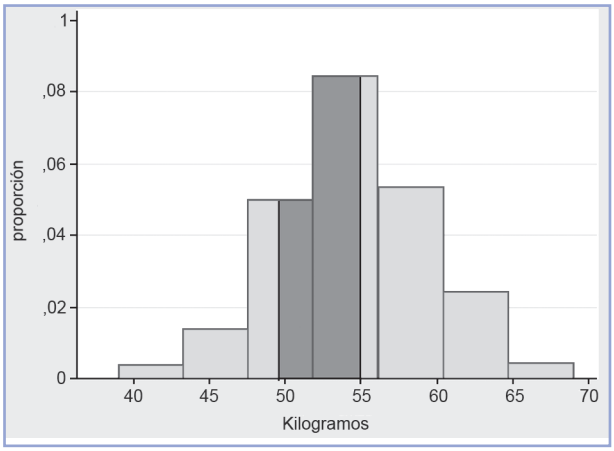
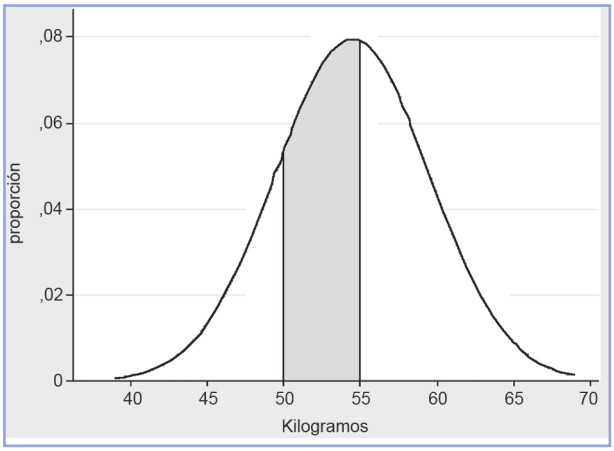


Figura 2.



La función a la que hacemos referencia recibe el nombre de "función densidad de probabilidades", que abreviaremos como fdp. Una función, f(x), es una fdp si cumple con dos condiciones:

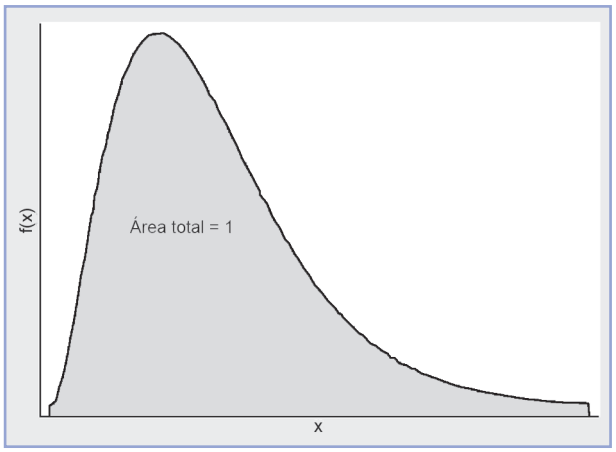
i. Es una función no negativa, es decir f (x) ≥ 0, para cualquier valor de x.

ii. El área total que ella encierra bajo su gráfico y el eje x es igual a 1. En símbolos: $\int_{-\infty}^{\infty} f(x)dx = 1$

Rincón de la Bioestadística

Gráficamente:

Figura 3.

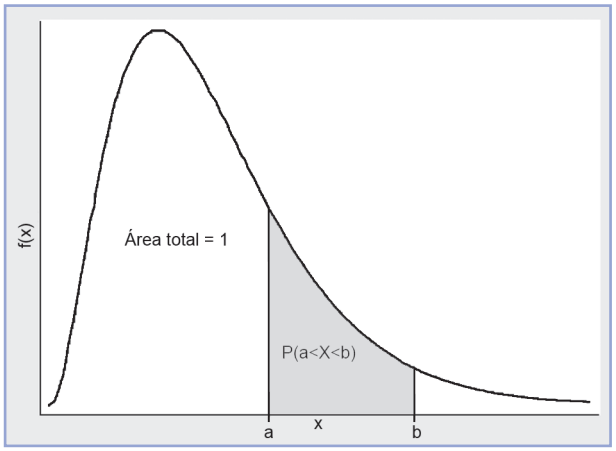


En estas condiciones, es posible calcular la probabilidad de que la variable X se encuentre entre los valores a y b; en símbolos:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

O sea, el área que encierra f(x) entre las verticales X = a y X = b:

Figura 4.

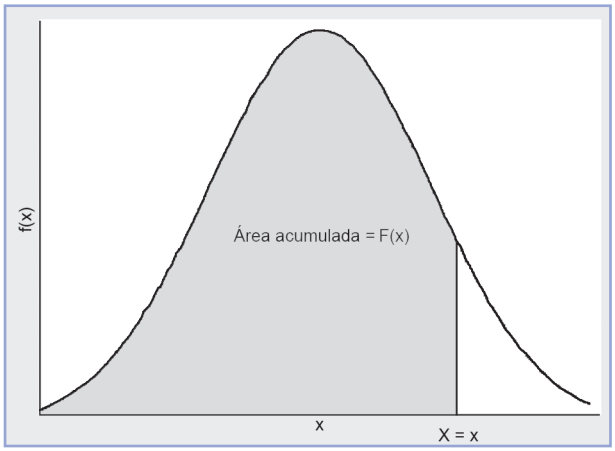


Se define la función de distribución de probabilidades (fdp) como:

$$F(X) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx$$

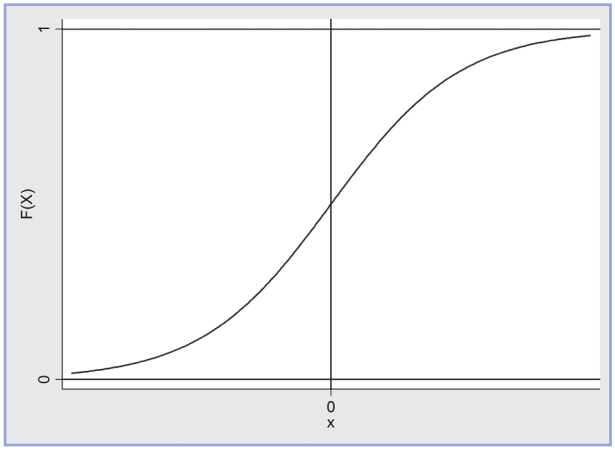
Es decir la fdp cuantifica el área acumulada bajo la fdp hasta el punto x, como muestra la siguiente Figura 5:

Figura 5.



Se observa que F (X) es acotada, ya que, $0 \leq F(X) \leq 1$ y que F(X) es siempre una función creciente, como muestra la Figura 6:

Figura 6.



Así:

$$P(a \leq X \leq b) = F(b) - F(a)$$

Notar que:

$$P(a \leq X \leq b) = P(a < X < b)$$

Ya que P (X = a) = P (X = b) = 0.

Caracterización de una variable aleatoria contínua

Dada una función densidad de probabilidades se define la Esperanza matemática o valor esperado de la variable X a la expresión:

$$E(X) = \mu = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

El momento de orden 2 está dado por la expresión:

$$E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx$$

Expresiones que permiten calcular la varianza de la variable X, a través de la expresión:

$$Var(X) = E[X^2] - (E[X])^2$$

Algunas distribuciones de probabilidad contínua

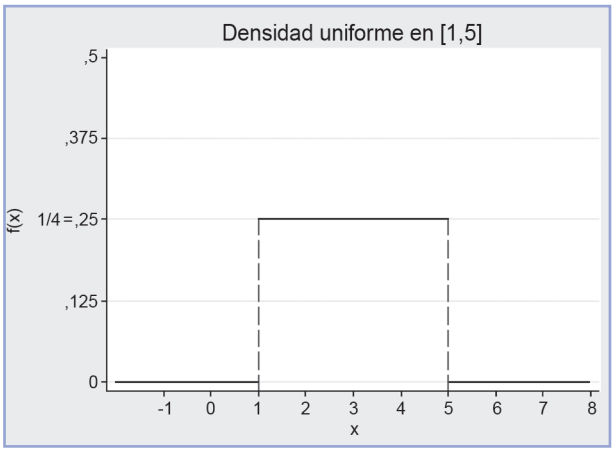
La distribución uniforme, X~U[a,b]

La variable X sigue una distribución uniforme en el intervalo [a,b], si su fdp es:

$$f(x) = \begin{cases} \dfrac{1}{b-a} \; si \; x \in [a, \, b] \\ 0 \;\; en \; otros \; casos \end{cases}$$

Cuyo gráfico, en el caso de la U[1,5] es:

Figura 7.

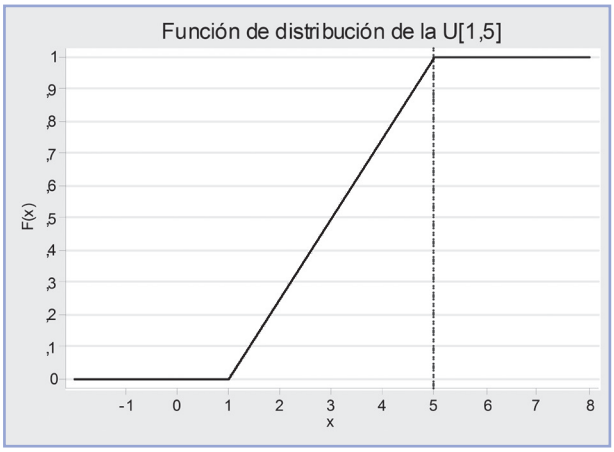


La función de distribución de probabilidades es:

$$F(x) = \begin{cases} 0 \; si \; x < a \\ \dfrac{x-a}{b-a} \; si \; a \le x \le b \\ 1 \; si \; x > b \end{cases}$$

El gráfico de la fdp, en el caso de la U[1,5] es:

Figura 8.



La esperanza y la varianza son:

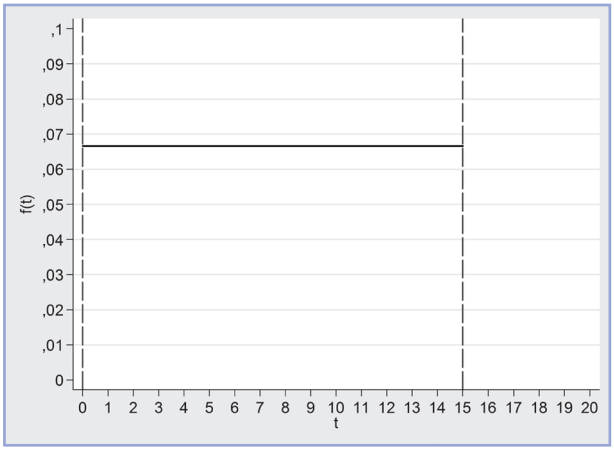
$$E(X) = \frac{a+b}{2}$$

$$Var(X) = \frac{(b-a)^2}{12}$$

Ejemplo: Una persona llega en forma aleatoria entre las 12:00 y las 12:15 horas a una determinada estación de Metro. Si un tren pasa exactamente a las 12:00 horas y los trenes tienen una frecuencia de 5 minutos. Calcular la probabilidad de que la persona espere más de 2 minutos un tren.

Si t es la variable tiempo de espera, en minutos, t~U[0,15], esta fdp la representamos gráficamente así:

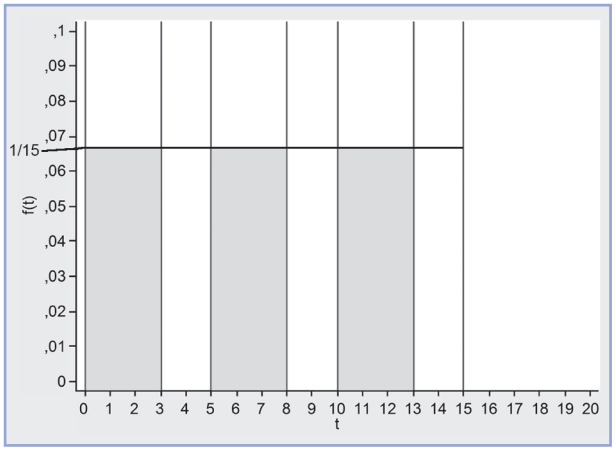
Figura 9.



Rincón de la Bioestadística

Luego, las áreas amarillas indican cuando ocurre el suceso de interés:

Figura 10.



En consecuencia, si A es el evento de esperar más de 2 minutos un tren:

$$P(A) = 3 \cdot \frac{3}{15} = \frac{9}{15} = 0{,}6$$

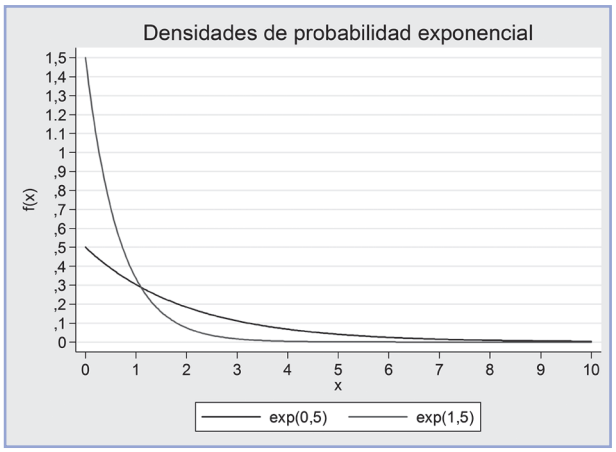
La distribución exponencial, X~exp(α)

Una importante distribución de probabilidades de una variable continua es la llamada distribución exponencial, que es la base para el análisis de sobrevida.

La variable X sigue una distribución exponencial de parámetro α si su fdp es:

$$f(x) = \begin{cases} a \cdot e^{-ax} & si \;\; x \geq 0 \\ 0 \;\; en\ otros\ casos \end{cases}$$

Figura 11.

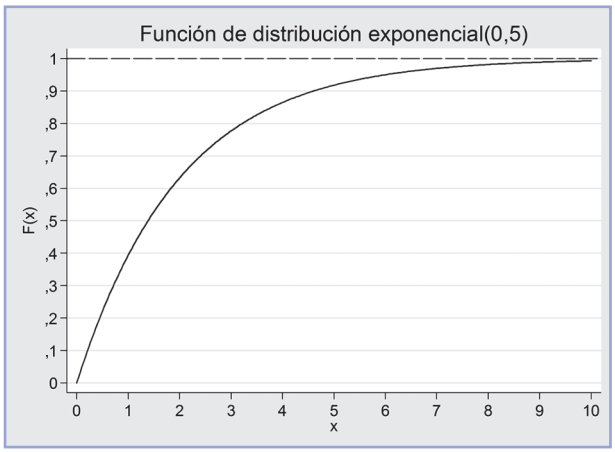


La función de distribución de probabilidades es:

$$F(x) = \begin{cases} 0, si \;\; x < 0 \\ 1 - e^{-ax}, si \;\; x \geq 0 \end{cases}$$

El gráfico de la fdp, en el caso de la exp(0,5) es:

Figura 12.



La esperanza y la varianza son:

$$E(X) = \frac{1}{\alpha}$$

$$Var(X) = \frac{1}{\alpha^2}$$

Ejemplo: Se sabe que el tiempo de duración de un marcapasos sigue una distribución exponencial. En base a registros de una serie de casos, se ha encontrado que en promedio, estos marcapasos han durado 60 meses. Calcular la probabilidad de que un marcapasos dure menos de 6 años:

Llamando x al tiempo de duración de un marcapasos, se tiene que E[x] = 60, con lo que α = 1/60, así la ftp para x es:

$$f(x) = \frac{1}{60} e^{-\frac{1}{60}x}, si \;\; x > 0$$

con lo que la ftp es

$$F(x) = 1 - e^{-\frac{1}{60}x}, si \;\; x > 0$$

Con lo que la probabilidad de que el marcapasos dure menos de 6 años es:

$$P(x \leq 72) = 1 - e^{-\frac{1}{60}72} = 0{,}6988 \approx 0{,}70$$

La distribución Normal o Gausstiana, X~N(μ,σ²)

Sea X una variable continua con recorrido R; diremos que X tiene una distribución normal (o gaussiana) si su fdp es:

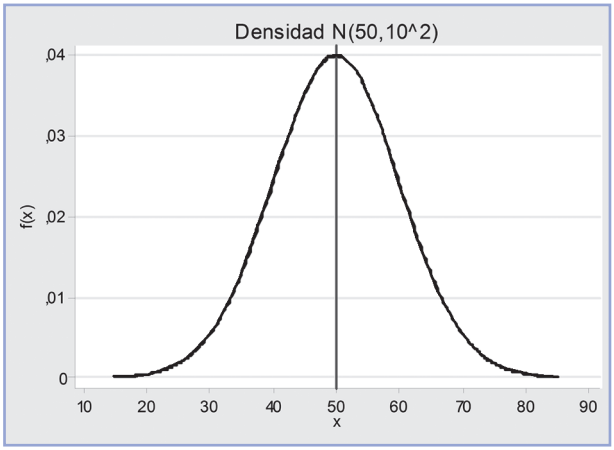
$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty,\mu\in R,\sigma\in R^+$$

El gráfico de la densidad $N(\mu,\sigma^2)$ es una curva tal que:

a) tiene máximo absoluto en $x = \mu$
b) es simétrica respecto a la vertical $x = \mu$
c) tiene puntos de inflexión en $x = \mu - \sigma$ y $x = \mu + \sigma$
d) se aproxima asintóticamente al eje de abscisas, lo que se refleja en la relación

$$f(\mu-3\sigma)=(\mu+3\sigma)=\frac{1}{100}f(\mu)$$

Figura 13.



La esperanza y la varianza son:

$$E(X)=\mu$$

$$Var(X)=\sigma^2$$

La distribución normal estándar o típica

Si z~N(0,1) se habla de la distribución normal estándar o típica, así:

$$F(Z)=\int_{-\infty}^{Z}\frac{1}{\sqrt{2\pi}}e^{-\frac{Z^2}{2}}dz=\phi(Z)$$

función que se encuentra tabulada.

Estandarización de variables aleatorias normales

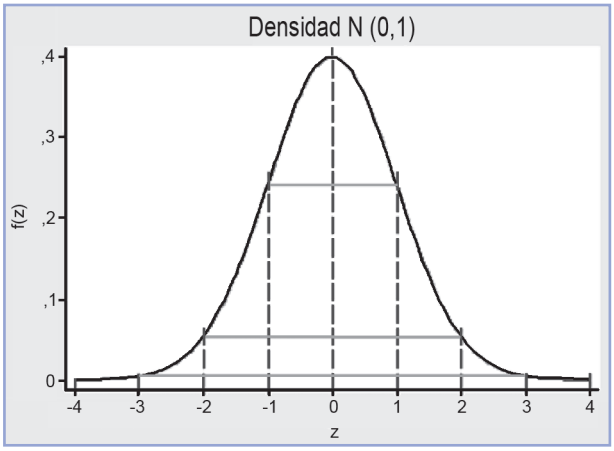
La estandarización de una variable aleatoria normal, es la transformación de dicha variable en una variable con distribución normal estándar, este proceso se obtiene usando el siguiente teorema:

Si $X\sim N(\mu,\sigma^2)$ entonces $z=\frac{X-\mu}{\sigma}\sim N(0,1)$

Al valor de z se le llama puntaje típico y representa la distancia de X a su promedio en unidades de desviación estándar.

En consecuencia, al estudiar la distribución normal estándar, se pueden generalizar algunas cosas de interés, como las probabilidades que se muestran en la Figura 14:

Figura 14.



Es decir, la probabilidad de encontrarse en torno al promedio en ± 1 desviación estándar es 68,3%, ± 2 desviaciones estándar es 95,5% y en ± 3 desviaciones estándar es 99,3%. Este resultado permite tener una respuesta aproximada a la interrogante si una colección de datos tiene una distribución normal.

Algunos percentiles clásicos de la normal estándar:

P_z	z
1,0%	-2,33
2,5%	-1,96
5,0%	-1,64
10,0%	-1,28
50,0%	0,00
90,0%	1,28
95,0%	1,64
97,5%	1,96
99,0%	2,33

Rincón de la Bioestadística

Propiedades de la distribución normal

Algunas propiedades de la distribución normal son las siguientes:

a) Si $X \sim N(\mu,\sigma^2) \rightarrow Y = a+b\cdot X \sim N(a+b\cdot\mu, b^2\sigma^2)$

b) Si $X \sim N(\mu_x, \sigma^2_x)$ e $Y \sim N(\mu_y, \sigma^2_y)$ y además X independiente de Y $\rightarrow X \pm Y \sim N(\mu_x \pm \mu_y, \sigma^2_x + \sigma^2_y)$

c) Si son independientes tales que $X_i \sim N(\mu_i, \sigma^2_i) \rightarrow \{X_i\}_{i=1}^{n}$

d) Si son independientes e idénticamente distribuidos $N(\mu,\sigma^2) \rightarrow \sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n}\mu_i, \sum_{i=1}^{n}\sigma^2_i\right)$ $N(n\mu, n\sigma^2)$

Tal vez uno de los resultados más notables de la estadística, sea el conocido **Teorema Central del Límite,** cuyo enunciado es:

Si $\sum_{i=1}^{n} X_i \sim$ son independientes e idénticamente distribuidas tales que E [X] = μ y Var [X]= $\sigma^2 \rightarrow \{X_i\}_{i=1}^{n}$ $N(n\mu, n\sigma^2)$ cuando $\sum_{i=1}^{n} X_i \sim$. Este teorema también se puede enunciar del siguiente modo:

Si $\{X_i\}_{i=1}^{n}$ son independientes e idénticamente distribuidas tales que E [X] = μ y Var [X]= $\sigma^2 \rightarrow z = \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ cuando $n \rightarrow \infty$. Este teorema es básico para construir la inferencia estadística paramétrica.

La distribución normal, es la distribución más usada y abusada por el usuario de la estadística. La forma de la clásica campana de Gauss, nos da la idea común de normalidad, esto es "ni mucho ni poco", sin embargo la normalidad estadística no siempre coincide con el concepto de normalidad clínica.

La Figura 15, muestra la diferencia distributiva del índice de masa corporal estandarizado (referencia CDC 2000) de un grupo de escolares chilenos respecto a la distribución teórica que es la normal estándar. Se puede deducir que este grupo de niños es aproximadamente 2 desviaciones estándar más obeso que la referencia:

Figura 15