Aproximación de probabilidades mediante la distribución normal de manera númerica

Universidad Autonoma de Coahuila
Facultad de Ciencias Fisico Matemáticas
Luis Eduardo Sánchez González

15 de junio de 2020

La distribución normal, también llamada distribución gaussiana o campana de Gauss. Está muy presente en multitud de fenómenos naturales debido al teorema del límite central: toda variable aleatoria que se pueda modelar como la suma de varias variables independientes e idénticamente distribuidas con expectativa y varianza finita, es aproximadamente normal.

1. Introducción

1.1. Variables aleatorias

Una **variable aleatoria** es una función que asocia un número real con cada elemento de un espacio muestral.

Se utiliza una letra mayúscula, digamos X, para denotar una variable aleatoria, y su correspondiente letra minúscula, x en este caso, para uno de sus valores.

Una variable aleatoria se llama **variable aleatoria discreta** si se puede contar su conjunto de resultados posibles. Cuando una variable aleatoria puede tomar valores en una escala continua, se le denomina **variable aleatoria continua.** En este caso nos concentraremos en las variables continuas.

En la mayoría de los problemas prácticos las variables aleatorias continuas representan datos medidos, como serían todos los posibles pesos, alturas, temperaturas, distancias o periodos de vida; en tanto que las variables aleatorias discretas representan datos por conteo, como el número de artículos defectuosos en una muestra de k artículos o el número de accidentes de carretera por año en una entidad específica.

1.2. Distribución continua de probabilidad

Una variable aleatoria continua tiene una probabilidad 0 de adoptar *exactamente* cualquiera de sus valores. En consecuencia, su distribución de probabilidad no se puede presentar en forma tabular.

Aunque la distribución de probabilidad de una variable aleatoria continua no se puede representar de forma tabular, sí es posible plantearla como una fórmula, la cual necesariamente será función de los valores numéricos de la variable aleatoria continua X, y como tal se representará mediante la notación funcional $f(x)$. Cuando se trata con variables continuas, a $f(x)$ por lo general se le llama **función de densidad de probabilidad**, o simplemente **función de densidad** de X.

La función $f(x)$ es una función de densidad de probabilidad para la variable aleatoria X, definida sobre el conjutno de los números reales, cumple

$$1) \qquad f(x) \geq 0, \ \forall x \in \mathbb{R}$$

$$2) \qquad \int_{-\infty}^{\infty} f(x)\, dx = 1$$

$$3) \qquad P(a < x < b) = \int_a^b f(x)\, dx$$

La distribución acumulada $F(x)$ de una variable aleatoria continua X con función de densidad $f(x)$ está dada por

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(t)\, dt, \quad \forall\ -\infty <\ t\ < \infty$$

Esto implica que

$$P(a < x < b) = F(b) - F(a)$$

Y por el teorema fundamental del cálculo podemos decir que

$$f(x) = \frac{dF(x)}{dx} = F'(x)$$

Los valores de $F(x)$ de función de distribución acumulada de variable aleatoria X cumple con:

1) Si $a < b \Rightarrow F(a) \leqslant F(b),\ \ \forall\, a, b\ \in \mathbb{R}$
2) $F(-\infty) = 0$
3) $F(\infty) = 1$

2. La distribución normal

La distribución de probabilidad continua más importante en todo el campo de la estadística es la distribución normal. Su gráfica, denominada curva normal, es la curva con forma de campana, la cual describe de manera aproximada muchos fenómenos que ocurren en la naturaleza, la industria y la investigación. Por ejemplo, las mediciones físicas en áreas como los experimentos meteorológicos, estudios de la precipitación pluvial y mediciones de partes fabricadas a menudo se explican más que adecuadamente con una distribución normal. Además, los errores en las mediciones científicas se aproximan muy bien mediante una distribución normal.

En 1733, Abraham DeMoivre desarrolló la ecuación matemática de la curva normal, la cual sentó las bases sobre las que descansa gran parte de la teoría de la estadística inductiva. La distribución normal a menudo se denomina distribución gaussiana en honor de Karl Friedrich Gauss (1777-1855), quien también derivó su ecuación a partir de un estudio de errores en mediciones repetidas de la misma cantidad.

Una variable aleatoria continua X que tiene la distribución en forma de campana de la se denomina variable aleatoria normal. La ecuación matemática para la distribución de probabilidad de la variable normal depende de los dos parámetros μ y σ, su media y su desviación estándar, respectivamente. Por ello, denotamos los valores de la densidad de X por $n(x; \mu, \sigma)$.

Distribución normal

La densidad de la variable aleatoria normal X, con media μ y varianza σ^2, es

$$n(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty$$

donde $\pi = 3{,}14159...$ y $e = 2{,}71828....$

Una vez que se especifican μ y σ, la curva normal queda determinada por completo.

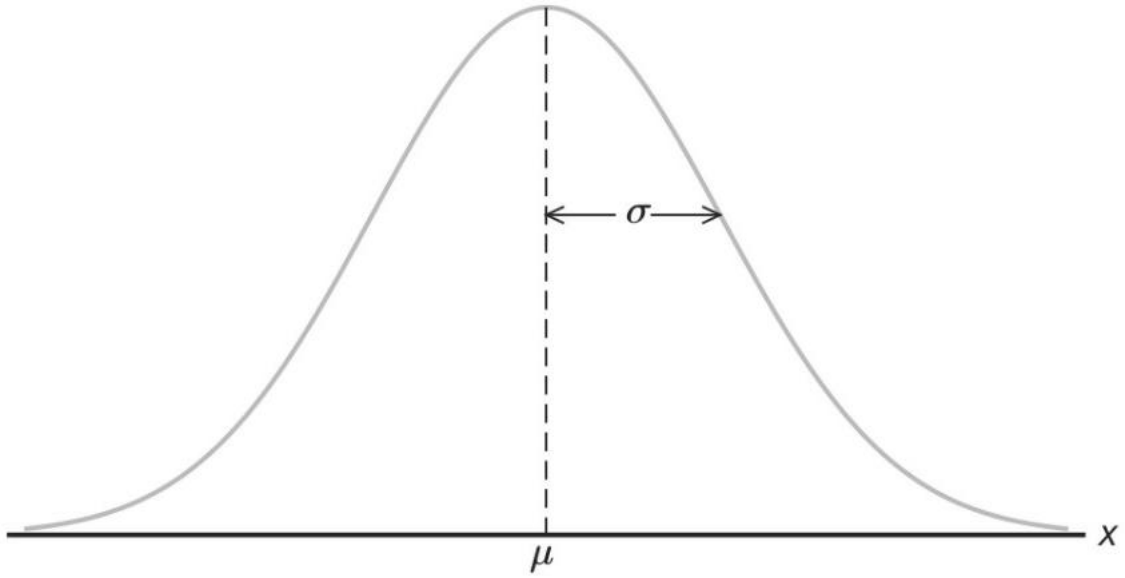

Figura 1: La curva normal.

3. Área bajo la curva

La curva de cualquier distribución continua de probabilidad o función de densidad se construye de manera que el área bajo la curva limitada por las dos ordenadas $x = x_1$ y $x = x_2$ sea igual a la probabilidad de que la variable aleatoria X tome un valor entre $x = x_1$ y $x = x_2$.

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma)\, dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx \tag{1}$$

es representada por el área de la región sombreada.

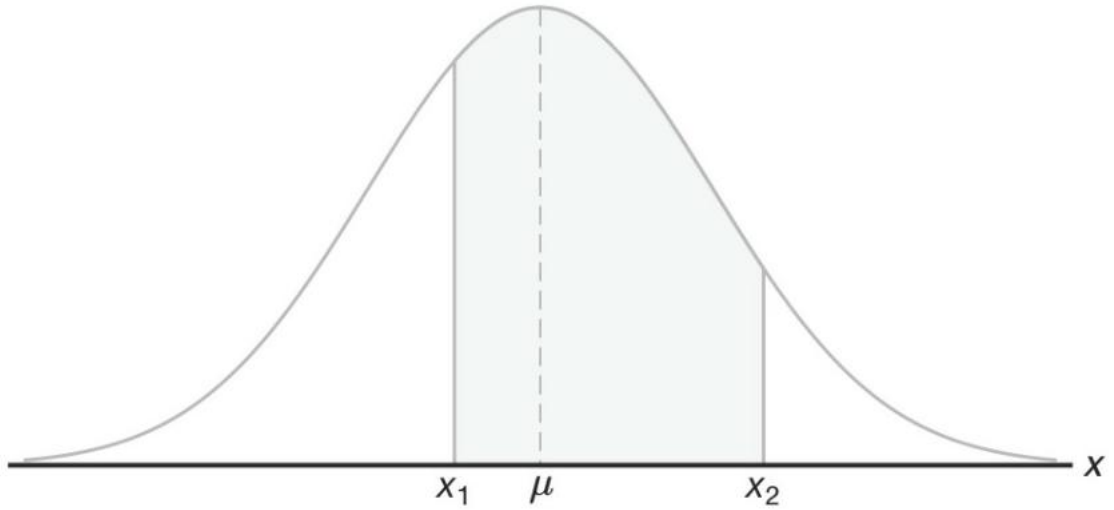

Figura 2: $P(x_1 < x < x_2) =$ área de la región sombreada.

Existen muchos tipos de programas estadísticos que sirven para calcular el área bajo la curva normal.

En este documento hablaremos acerca de un programa hecho en C++, en el cual calculamos la integral mediante un método númerico. Debemos mencionar que la dificultad que se enfrenta al resolver las integrales de funciones de densidad normal exige tabular las áreas de la curva normal para una referencia rápida. Sin embargo, sería inútil tratar de establecer tablas separadas para cada posible valor de μ y σ. Por fortuna, podemos transformar todas las observaciones de cualquier variable aleatoria normal X en un nuevo conjunto de observaciones de una variable aleatoria normal Z con media 0 y varianza 1. Esto se puede realizar mediante la transformación

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

Siempre que X tome un valor x, el valor correspondiente de Z es dado por $z = (x - \mu)/\sigma$. Por lo tanto, si X cae entre los valores $x = x_1$ y $x = x_2$, la variable aleatoria Z caerá entre los valores correspondientes $z_1 = (x_1 - \mu)/\sigma$ y $z_2 = (x_2 - \mu)/\sigma$. En conse- cuencia, podemos escribir

$$P(x_1 < x < x_2) = \int_{x_1}^{x_2} n(x; \mu, \sigma)\, dx = \frac{1}{\sqrt{2\pi\sigma}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{z_1}^{z_2} e^{-\frac{1}{2}z^2}\, dz = \int_{z_1}^{z_2} n(x; 0, 1)\, dz = P(z_1 < z < z_2)$$

La distribución de una variable aleatoria normal con media 0 y varianza 1 se llama **distribución normal estándar**.

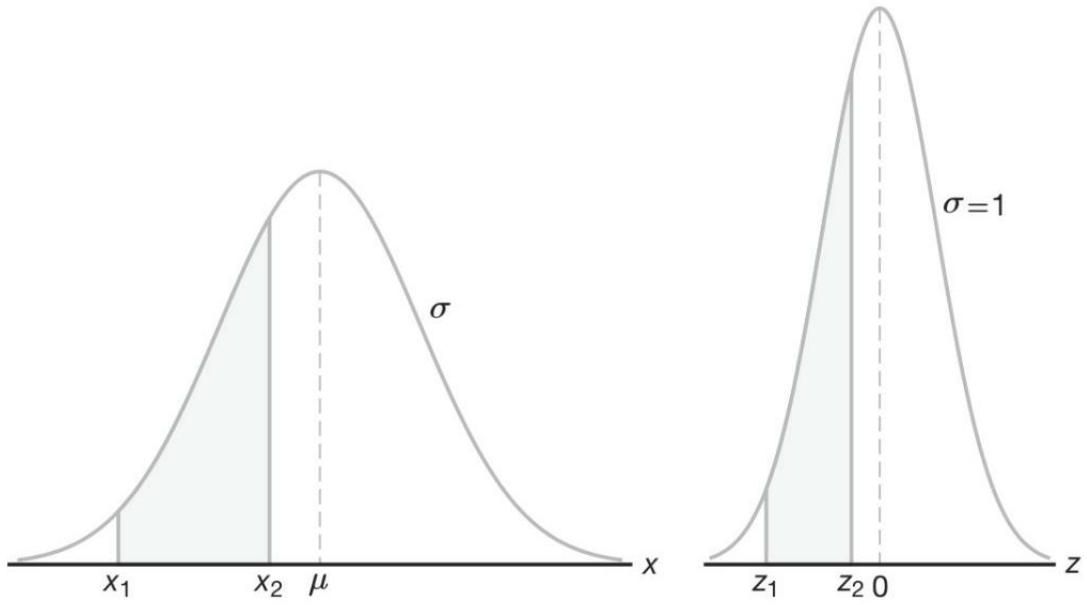

Figura 3: Distribución normal original y transformada.

Sabemos que una función de densidad de probabilidad de una variable aleatoria continua X se cumple que

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

Podemos demostrar que para la función asociada a la distribución de una variable aleatoria normal es una función de densidad de probabilidad, es decir

$$\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi\sigma}}\, dx = 1$$

Demostración:
Primero debemos tomar en cuenta la transformación hecha anteriormente, por lo tanto

$$\int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi\sigma}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz = 1$$

Definamos I de tal manera que

$$I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz$$

Ahora, podemos reescribir esta integral en función de una variable w

$$I = \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz = \int_{-\infty}^{\infty} e^{-\frac{1}{2}w^2}\, dw$$

Entonces multiplicando las dos integrales tenemos

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz\right)\left(\int_{-\infty}^{\infty} e^{-\frac{1}{2}w^2}\, dw\right)$$

Por el teorema de Fubini podemos reescribir y simplificar lo anterior de la forma siguiente

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2} e^{-\frac{1}{2}w^2}\, dzdw = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2+w^2)}\, dzdw$$

Al hacer un cambio de variables.

$$z = rcos\theta \qquad w = rsen\theta \qquad con \qquad 0 \leq r < \infty \qquad \wedge \qquad 0 \leq \theta \leq 2\pi$$

Lo que se conoce como coordenadas polares. Así:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2+w^2)}\, dzdw = \int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}r^2} r\, drd\theta = \int_0^{2\pi} \lim_{m\to\infty}\left[-e^{\frac{1}{2}r^2}\right]_0^m = \int_0^{2\pi} d\theta$$

Al tomar en cuenta la defínción de I, tenemos

$$I^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(z^2+w^2)}\, dzdw = \int_0^{2\pi} d\theta = 2\pi \qquad \Rightarrow \qquad I = \sqrt{2\pi}$$

Por lo tanto.

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}z^2}\, dz = \frac{1}{\sqrt{2\pi}} I = \frac{1}{\sqrt{2\pi}}\sqrt{2\pi} = 1$$

Que es lo que queriamos demostrar.

4. Aproximación normal a la binomial

Las probabilidades asociadas con experimentos binomiales se obtienen fácilmente a partir de la fórmula $b(x; n, p)$ de la distribución binomial o de alguna tabla cuando n es pequeña. Además, las probabilidades binomiales están disponibles en muchos paquetes de software. Sin embargo, resulta aleccionador conocer la relación entre la distribución binomial y la normal. La distribución normal a menudo es una buena aproximación a una distribución discreta cuando la última adquiere una forma de campana simétrica. Desde un punto de vista teórico, algunas distribuciones convergen a la normal a medida que sus parámetros se aproximan a ciertos límites. La distribución normal es una distribución de aproximación conveniente, ya que la función de distribución acumulativa se tabula con mucha facilidad. La distribución binomial se aproxima bien por medio de la normal en problemas prácticos cuando se trabaja con la función de distribución acumulativa.

Aproximación normal a la binomial

Si X es una variable aleatoria binomial con media $\mu = np$ y varianza $\sigma^2 = npq$, entonces la forma limitante de la distribución de

$$Z = \frac{X - np}{\sqrt{npq}}$$

conforme $n \to \infty$, es la distribución normal estándar $n(z; 0, 1)$.

Resulta que la distribución normal con $\mu = np$ y $\sigma^2 = np(1\breve{}p)$ no sólo ofrece una aproximación muy precisa a la distribución binomial cuando n es grande y p no está extremadamente cerca de 0 o de 1, sino que también brinda una aproximación bastante buena aun cuando n es pequeña y p está razonablemente cerca de 1/2.

5. Algoritmo para la distribución normal

Podemos realizar un pequeño algortimo que nos permita obtener probabilidades por medio de la distribución normal. Además de eso podemos hacer aproximaciones normal a la binomial solo modificando algunas condiciones.

Para esto necesitamos de alguna manera aproximar el área bajo la curva, es decir debemos aproximar la integral en un intervalo $[x_1, x_2]$. Sabemos que

$$P(x_1 < x < x_2) = \frac{1}{\sqrt{2\pi\sigma}} \int_{x_1}^{x_2} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\, dx$$

El resolver la integral parece algo complejo, por lo que se adopto utilizar un método númerico lo bastante eficiente para obtener buenas aproximaciones.

Además podemos implementar algo realemente sutil para obtener la aproximación normal a la distribución binomial, lo que se hará es corregir en un error que puede suceder normalmente. Sabemos que en una distribución normal σ no puede ser cero, por lo cual si $\mu = 0$ y $\sigma = 0$ el algoritmo lo tomará para hacer un "cambio de variable" y dar entrada a los parametros de la distribución binomial n, p. A partir de ahi se calcula μ y σ como vimos en la sección anterior. El algoritmo utilizado se muestra a continuación

Algoritmo 1: `Distribución normal`

Entrada: extremos x_1, x_2, parametros μ, σ
Resultado: aproximación $P(x_1 < x < x_2)$

Tome: $z_1 = \dfrac{x_1 - \mu}{\sigma}$ y $z_2 = \dfrac{x_2 - \mu}{\sigma}$;
si $\mu = 0$ & $\sigma = 0$ **entonces**
  Entrada: parametros distribución binomial, n, p
  $q = 1 - p$;
  $\mu = np,\ \sigma = \sqrt{npq}$;
  $x_1 = x_1 - 0{,}5,\ x_2 = x_2 + 0{,}5$;
  Tome: $z_1 = \dfrac{x_1 - \mu}{\sigma}$ y $z_2 = \dfrac{x_2 - \mu}{\sigma}$
fin
Método de Simpson(x_1, x_2) ;
return $P(x_1 < x < x_2)$;

En el algoritmo no se especifica como es el método de Simpson, para obtener más información acerca del método puede consultar el libro *Análisis númerico de Richard L. Burden y J. Douglas Faires*.

6. Implementación

El codigo consta de tres archivos para poder ejecutar el programa.

1. DistribucionNormal.cpp
2. Normal.cpp
3. Normal.h

Y dos archivos secundarios que nos ayudan en la graficación. El primer archivo contiene la función principal donde se manda llamar a la función Distribucion ubicada en el archivo *Normal.cpp*, en este archivo esta contenido todas las funciones que se utilizan en el programa, funciones que se define en el archivo de encabezado, *Normal.h*

Podemos observar las funciones definidas y sus respectivos parametros en el siguiente diagrama.

DistribucionNormal.cpp

```
class DistribucionNormal{

        public:

   //Evaluar en la funcion de densidad
                double Funcion(double z) ;
   //Resolver la integral
                double Integral(double h);
   //Obtener el error de la integral
                double Error(double epsilon);
   //Generar la grafica
                void grafica(double ti, double tf);
                void Distribucion();

        private:

   //Parametros de la distribucion
                double nu, sigma, pi;
   //Parametros de graficacion
                double t, ti, tf, f, dt, g;
   //Evaluacion en la funcion de densidad
                double F;
   //Parametros metodo de simpson
                double I,zo,z1,z2,z3,h3;
   //Condicion
                double zi, x;
   //Condiciones del problema
                double xo,x1, P;
   //Error en el metodo de Simpson
                double E, epsilon;
   //Parametros de la distribucion binomial
                double p, q, n;
                int opc;
};
```

Probemos el programa con dos sencillos problemas, uno correspondiente a la distribución normal y otro que corresponda a la aproximación normal a la binomial.

Problema 1

Cierta máquina fabrica resistencias eléctricas que tienen una resistencia media de 40 ohms y una desviación estándar de 2 ohms. Si se supone que la resistencia sigue una distribución normal y que se puede medir con cualquier grado de precisión, ¿qué porcentaje de resistencias tendrán una resistencia que exceda 43 ohms?

Terminal

```
$ c++ -o Normal DistribucionNormal.cpp && ./Normal

<<<<Probabilidad por medio de la distribución normal>>>>

Este programa determina la probabilidad de un suceso en base a la distribución
normal, todo esto se realiza mediante un método númerico. Además de eso
también este programa puede determinar una aproximación a la distribución
binomial por medio de la normal

Digite el intervalo de tiempo [xo,x1] en el cual quiere calcular la
probabilidad.

--------------------------------------------------------------------------------
 Opciones:
 * Si desea calcular la probabilidad de una variable aleatoria 'X' tal que
 P(X = x), entonces digite '0' para xo y x1.
 * Si desea calcular la probabilidad de una variable aleatoria 'X' tal que
 P(X < x), entonces digite '1000' para xo y para x1 digite el dato
 correspondiente.
 * Si desea calcular la probabilidad de una variable aleatoria 'X' tal que
 P(X > x), entonces digite '1000' para x1 y para xo digite el dato
  correspondiente.
--------------------------------------------------------------------------------

xo: 43
x1: 1000

INFO: Si desea realizar una aproximación de la distribución binomial a la
distribución normal, digite '0' para la media y la desviación estandar

Digite la media y la desviación estandar de la distribución.
Media: 40
Desviación: 2

Conforme a los datos que ingresaste, obtenemos que:

zo = 1.5     z1 = 4

La probabilidad en el intervalo [43,infinito] es:

P(x > 43) = 0.0693178
```

Problema 2

Un examen de opción múltiple tiene 200 preguntas, cada una con 4 respuestas posibles, de las que sólo una es la correcta. ¿Cuál es la probabilidad de que solamente adivinando se obtengan de 25 a 30 respuestas correctas para 80 de los 200 problemas sobre los que el estudiante no tiene conocimientos?

Terminal

```
$ c++ -o Normal DistribucionNormal.cpp && ./Normal

<<<<Probabilidad por medio de la distribución normal>>>>

Digite el intervalo de tiempo [xo,x1] en el cual quiere calcular la
probabilidad.

--------------------------------------------------------------------------------
 Opciones:
 * Si desea calcular la probabilidad de una variable aleatoria 'X' tal que
 P(X = x), entonces digite '0' para xo y x1.
 * Si desea calcular la probabilidad de una variable aleatoria 'X' tal que
 P(X < x), entonces digite '1000' para xo y para x1 digite el dato
 correspondiente.
 * Si desea calcular la probabilidad de una variable aleatoria 'X' tal que
 P(X > x), entonces digite '1000' para x1 y para xo digite el dato
  correspondiente.
--------------------------------------------------------------------------------

xo: 25
x1: 30

INFO: Si desea realizar una aproximación de la distribución binomial a la
distribuciónnormal, digite '0' para la media y la desviación estandar

Digite la media y la desviación estandar de la distribución.
Media: 0
Desviación: 0

ADVERTENCIA: Debe tomar en cuenta que para hacer una aproximación lo bastante
buena, se necesita que 'n' sea lo "suficientemente" grande.

Digite las condiciones para la distribución binomial.
n = 80
p = 0.25

Conforme a los datos que ingresaste, obtenemos que:

zo = 1.1619     z1 = 2.71109

La probabilidad en el intervalo [24.5,30.5] es:

P(24.5 < x < 30.5) = 0.118688
```