ALGORITMOS NUMERICOS PARA LA CONSTRUCCION DE TABLAS DE PROBABILIDAD

Jorge Mauricio Galbiati Riesco

En todas las áreas de actividad se debe tomar decisiones y en la inmensa mayoría de los casos este proceso de toma de decisiones se desarrolla en un escenario de incertidumbre. Esto es válido, por citar algunos casos, en la administración, la agricultura, las ciencias naturales, la sociología, el derecho, la economía, la política, la ingeniería, la investigación histórica, la medicina, la meteorología, etc. La estadística provee de modelos para la incertidumbre, que permiten optimizar la toma de decisiones, minimizando los riesgos. Estos modelos están constituidos por las llamadas *variables aleatorias*, que tienen asociadas *distribuciones de probabilidad*.

Este manual contiene *tablas de distribuciones de probabilidad* asociadas a las variables aleatorias **binomial**, **poisson**, **normal**, **ji-cuadrado**, **t de Student** y **F de Snedecor**, y fórmulas para la **uniforme**, la **exponencial**, la **gama**, y la **beta**. La *función de distribución* asigna a cada valor (llamado *cuantil*) la probabilidad acumulada hasta él, obtenida por suma de los valores de su *función de probabilidad* en los casos discretos, o por integración de su *función de densidad* en los casos continuos.

En estas páginas introductorias quiero explicar, en lineas generales, cómo se construyeron las tablas a objeto de que esta información le sea útil al lector que se vea enfrentado a tareas similares. El cálculo numérico tiene involucrado una serie de dificultades propias del hecho de trabajar con magnitudes numéricas extremas. Por una parte, las dificultades radican en el hecho que se trabaja con números extremadamente grandes, y con números extremadamente pequeños, que se combinan multiplicativamente, para dar resultados cuyas magnitudes son moderadas. Un pequeño error de redondeo en estos factores extremos, puede causar un error grave en el resultado final.

Por otra parte, el elevado número de operaciones puede hacer que los cálculos sean exageradamente lentos. A continuación se explicará, en cada caso, qué problemas surgieron, y cómo se resolvieron.
Para el cálculo de las distribuciones de probabilidad discretas, en este caso la **binomial** y la **poisson**, se utilizaron las conocidas fórmulas de las funciones de probabilidad, obteniéndose las distribuciones por suma acumulativa de los términos. En el caso de las discretas, cuando los parámetros son grandes, surgen situaciones en que se debe multiplicar números muy grandes por números muy pequeños. Dado que la capacidad de almacenaje del computador, por muy vasta que sea, es limitada, si ésta es sobrepasada por un número muy grande (o muy pequeño), se producirá un truncamiento, por lo que, el multiplicarlo por un número muy pequeño (o muy grande), conducirá a un resultado erróneo. Estudiemos cada caso individualmente:

La variable aleatoria **binomial** tiene su función de probabilidad

$$p(x) \;=\; \frac{n!}{x!(n-x)!}\; p^x\; (1-p)^{n-x} \qquad si\; x = 0, 1, 2, ..., n$$

Si el parámetro n es grande, los valores de los factoriales pueden ser enormes. Si se intenta calcularlos, es seguro que se sobrepasará la capacidad del sistema computacional. Por otro lado, la parte de las potencias de la fórmula binomial puede llegar a ser pequeñísima, y nos enfrentamos a situaciones problemáticas como las descritas anteriormente.
Sin embargo, si escribimos esta expresión como $p(x) = A(x) \cdot B(x)$ en que

$$A(x) = \frac{n(n-1)...(n-x+1)}{x!}[p(1-p)]^x$$

y

$$B(x) = (1-p)^{n-2x}$$

es fácil ver que ambas expresiones se pueden obtener, en forma recursiva

para n y p fijos, según la siguiente fórmula:

$$A(x) = A(x-1) \cdot \frac{n-x+1}{x} \cdot p(1-p)$$

y

$$B(x) = B(x+1) \cdot (1-p)$$

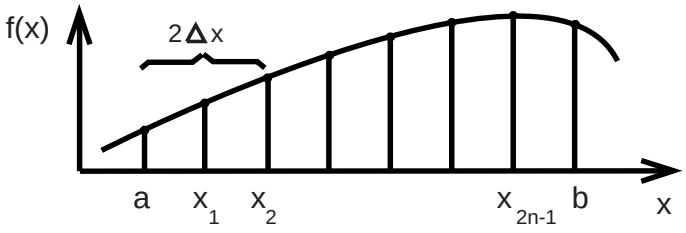
De esta forma y partiendo de $A(0) = 1$ y $B(n/2) = 1$, se pueden generar, recursivamente, todas las probabilidades binomiales, sin caer en el cálculo de expresiones excesivamente grandes o excesivamente pequeñas. La variable aleatoria **Poisson** tiene como función de probabilidad

$$p(x) = \frac{e^{-\lambda}\ \lambda^x}{x!} \qquad si\ x = 0, 1, 2, ...$$

La siguiente fórmula recursiva permite calcular los términos de esta función fácilmente, para un λ fijo, partiendo de $p(0) = e^{-\lambda}$:

$$p(x) = p(x-1) \cdot \frac{\lambda}{x}$$

Las distribuciones de probabilidad continuas contenidas en este manual son la **normal**, la **ji-cuadrado**, la **t de Student** y la **F de Snedecor**. El cálculo de los valores de las funciones de distribución se hizo por integración numérica, mediante la regla de Simpson. Esta consiste en aproximar el área bajo la curva de la función de densidad respectiva, por segmentos de lados verticales, en que la parte superior es un trazo parabólico que aproxima la curva, como muestra la figura:



Esto da como resultado la siguiente fórmula de cálculo: $\int_a^b f(x)dx \approx \frac{\Delta x}{3}\big[f(x_0) +$

$4f(x_1) + 2f(x_2) + 4f(x_3) + ...$

$... + 2f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n})\big]$ en que $\Delta x = \frac{b-a}{2n}, \quad x_0 = a$ y $x_{2n} = b$.
$f(x)$ es la función de densidad respectiva y n el número de segmentos. El ancho de cada segmento es $2\Delta x$. Para estas tablas se utilizaron segmentos de ancho 0.001, para asegurar la precisión requerida. Veremos cómo se abordó el cálculo de cada una de las distribuciones:

En el caso de la variable aleatoria **normal** La tabla entrega la función de distribución de la **normal standard**, cuya función de densidad es

$$f(x) \;=\; \frac{1}{\sqrt{(2\pi)}\sigma}\; exp\Big[-\frac{(x-\mu)^2}{2\sigma^2}\Big] \qquad para\ x \in (-\infty, +\infty)$$

La integración numérica, en el caso de la normal, no ofrece mayor dificultad. Se aprovecha la simetría de esta distribución para reducir el intervalo de integración.

Las restantes densidades tienen un término constante que involucra el cálculo de la función Gama, que se define como

$$\Gamma(t) = \int_0^\infty y^{t-1}e^{-y}dy$$

En general, no es integrable por métodos analíticos y para evaluarla tendría que recurrirse a procedimientos de integración numérica, lo que elevaría enormemente el volumen de cálculos necesarios. Sin embargo, las siguientes propiedades permiten obviar el cálculo directo de la función Gama: 1. $\Gamma(n) = (n-1)!$ *si n es entero no negativo*
2. $\Gamma(x) = (x-1)\cdot\Gamma(x-1) = (x-1)\cdot(x-2)\cdot\Gamma(x-2) = ...$
3. $\Gamma(1/2) = \sqrt{\pi}$

Gracias a estas propiedades, el cálculo del término constante de las funciones de densidad puede hacerse sin tener que evaluar la función gama. Veamos cómo se aplicaron en cada caso: La variable aleatoria **ji-cuadrado** tiene función de densidad

$$f(x) \;=\; \frac{1}{2^{k/2}\Gamma(k/2)}x^{k/2-1}e^{-x/2} \qquad si \quad x > 0$$

k es el parámetro *grados de libertad* de la ji-cuadrado. Es un entero no negativo, por lo tanto $k/2$ es entero o múltiplo entero de $1/2$. Por la propiedad 2 de la función gama, se tiene que

$$\Gamma(k/2) = (k/2-1)\Gamma(k/2-1) = (k/2-1)(k/2-2)\Gamma(k/2-2) = ...$$

Entonces partiendo de $\Gamma(1/2) = \sqrt{\pi}$ y de $\Gamma(1) = 1$, se pueden evaluar los términos $\Gamma(k/2)$ en forma recursiva, para todos los valores de k.

Lo anterior no funciona para la **ji-cuadrado** con un grado de libertad, puesto que resulta una integral impropia, que no puede calcularse mediante la regla de Simpson. En este caso recurrí al hecho que esta variable aleatoria (con un grado de libertad) es igual a una **normal estándar** elevada al cuadrado.
La función de densidad de la variable aleatoria **T de Student** es

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\big(k/2\big)} \cdot \frac{1}{\sqrt{(k\pi)}} \cdot \frac{1}{\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}}} \qquad para \quad x \in (-\infty, +\infty)$$

El valor k corresponde al parámetro *grados de libertad*. Como k es entero, $k/2$ o $(k+1)/2$ es entero, el otro término es múltiplo entero de $1/2$. Entonces después de algunos cálculos, en que se consideran por separado los casos en que k es par y k es impar, el cuociente de las funciones gama

$$G(k) = \Gamma\Big(\frac{k+1}{2}\Big)\Big/\Gamma(k/2)$$

puede llegar a expresarse en forma recursiva como

$$G(k) = \frac{k-1}{k-2} \cdot G(k-1)$$

válida desde $x = 3$ en adelante . Partiendo de $G(1) = \frac{1}{\sqrt{\pi}}$ y de $G(2) = \frac{\sqrt{\pi}}{2}$ se pueden generar todos los términos, para todos los valores de k. Por último, la variable aleatoria **F de Snedecor** tiene función de densidad

$$f(x) \;=\; \frac{\Gamma\left(\frac{n+d}{2}\right)}{\Gamma\left(\frac{n}{2}\right) \cdot \Gamma\left(\frac{d}{2}\right)} \cdot \left(n/d\right)^{n/2} \cdot \frac{x^{n/2-1}}{\left(1+\frac{n}{d}x\right)^{\frac{n+d}{2}}} \qquad si \quad x > 0$$

Los números enteros n y d son los parámetros de la distribución, y se denominan, respectivamente, *grados de libertad del numerador* y *grados de libertad del denominador*. Como en el caso de la t de student, la parte del cuociente de funciones gama

$$H(n,d) = \frac{\Gamma\left(\frac{n+d}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{d}{2}\right)}$$

puede desarrollarse utilizando las propiedades de la función gama, llegándose a las expresiones recursivas

$$H(n,d) = \frac{n-1}{n-2} \cdot H(n-2,d)$$

y

$$H(n,d) = \frac{n+d-2}{d-2} \cdot H(n,d-2)$$

Partiendo de $H(1,1) = 1/\pi$, $H(1,2) = H(2,1) = 1/2$ y de $H(2,2) = 1$, se pueden generar los valores de $H(n,d)$ para todo n y d. Con la **F** con un grado de libertad en el numerador ocurre lo mismo que con la **ji-cuadrado** con un grado de libertad, que resulta una integral impropia. En este caso usé el hecho que esta variable aleatoria es igual a una **t de student** con los grados de libertad del denominador de la **F**, elevada al cuadrado.

Hay un elemento que diferencia fundamentalmente la construcción las tablas de distribución Binomial, Poisson y Normal de las tablas de distribución Ji-cuadrado, T de Student y F de Snedecor.

En las tres primeras se dieron valores de los cuantiles, y se calcularon las probabilidades acumuladas hasta cada cuantil. Mientras que en las últimas tres tablas se fijaron valores de probabilidad acumulada, y se calcularon los cuantiles correspondientes. Esto se hizo aplicando la regla de Simpson, y verificando en cada paso, si se había alcanzado la probabilidad dada. En el fondo, para la Ji-cuadrado, la T de Student y la F de Snedecor se calcularon las inversas de las funciones de distribución. La razón de esto es que estas tablas se usan fundamentalmente para hacer estimaciones por intervalos de confianza y para hacer pruebas de hipótesis. En estos casos se fijan las probabilidades y se debe buscar sus cuantiles correspondientes.

Para validar las tablas, se rehicieron los cálculos utilizando el proceso inverso: Si se introdujeron las cuantilas y se obtuvieron la probabilidades acumuladas, en el proceso de validacón se introdujeron las probabilidades acumuladas, se calcularon los cuantiles, y se verificó que coincidieran con los valores introducidos originalmente. Por el contrario, si se introdujeron probabilidades acumuladas, se validó introduciendo los cuantiles obtenidas, y se verificó que las probabilidades resultantes coincidieran con las introducidas originalmente.

Los programas para el cálculo de estas tablas los hice utilizando el lenguaje de programación general Visual Basic 6.0, en ambiente Windows. Utilicé el formateador de texto LaTeX para escribir los textos y las tablas.