Simulación ITAM
29 de agosto de 2021 2021-II

Tarea 2.

La fecha de entrega es el lunes **10 de septiembre de 2021**.

Lecturas

- Casella y Robert, leer la sección 1.5 y el capítulo 2.
- Fast Generation of Discrete Random Variables Artículo de *Journal of Statistical Software, July 2004, Volume 11, Issue 3*
- Introduction to simulation using R Capítulo 13
- Marc C. Bove, et. al Effect of El Niño on U.S. Landfalling Hurricanes, Revisited

Problemas

1. **Problema de captura-recaptura**. Este tipo de problemas son aplicados en multitud de contextos. Un estadístico está interesado en el número N de peces que hay en un estanque. Para estimar su valor utiliza el siguiente procedimiento:

   a) Captura 250 peces, los marca y los regresa al estanque.

   b) Unos cuantos días después regresa y atrapa suficientes peces hasta que obtiene 50 peces marcados, en ese punto también tiene 124 peces no marcados, los que tuvo que cpturar para obtener los 50 peces (la muestra total es de 174 peces).

   - ¿Cuál es la estimación de N?
   - Hagan un programa que permita simular el proceso de obtener la primera y segunda muestra considerando como parámetros el tamaño N de la población de interés, el tamaño de la primera y segunda muestra y como datos a estimar son: de qué tamaño debe ser n_1 y n_2 para obtener una buena aproximación y ver cómo se afecta por el tamaño N.

2. Este problema es una versión simplificada de dos problemas comunes a los que se enfrentan las compañías de seguros: calcular la probabilidad de quebrar y estimar cuánto dinero podrán hacer. Los parámetros se dan en conjuntos, se pide simular combinaciones de los diferentes parámetros (probabilidad de reclamo, horizonte de tiempo).

Supongan que una compañía de seguros tiene activos por \$ 10^7 de pesos. Tienen $n = 1,000$ clientes que pagan individualmente una prima anual de \$5,500 al principio de cada año. Basándose en experiencia previa, se estima que la probabilidad de que un cliente haga un reclamo es $p \in \{0.01, 0.1, 0.15, 0.2\}$ por año, independientemente de reclamos previos de otros clientes. El tamaño X de los reclamos varía, y tiene la siguiente densidad con $\alpha = 5$ y $\beta = 125,000$:

$$f(x) = I(x \geq 0)\frac{\alpha\beta^\alpha}{(x+\beta)^{\alpha+1}}$$

(Tal X se dice que tiene una distribución *Pareto*, y en el mundo real no es un modelo infrecuente para el tamaño de los reclamos a las aseguradoras).

Suponemos las fortunas de la compañía aseguradora sobre un horizonte de $T \in \{5, 10, 20\}$ años. Sea $Z(t)$ los activos de la compañía al final del año t, así que $Z(0) = 10,000,000$, y $Z(t) = I_{Z(t-1)>0}\,\text{máx}\{Z(t-1) + \text{primas} - \text{reclamos}, 0\}$

Noten que si $Z(t)$ cae bajo 0 entonces la compañía quiebra y se queda ahí: si la compañía se va a bancarrota, deja de operar.

- Calcular la función de distribución F_X, $E[X]$, y $Var[X]$. Obtengan por simulación una muestra de X y su función de distribución, y comparen su estimado con la verdadera.
- Escriban una función para simular los activos de la compañía en el horizonte de tiempo T y estimen: (1) la probabilidad de que la compañía se vaya a bancarrota, y (2) Los activos esperados al final de cinco años.
- Toma de ganancias. Supongan ahora que la compañía toma ganancias al final de cada año. Esto es, si $Z(t) > 10,000,000$ entonces $Z(t) - 10,000,000$ se le paga a los accionistas. Si $Z(t) \leq 10,000,000$ entonces los accionistas no reciben nada ese año. Usando este nuevo esquema, estimar: (1) la probabilidad de irse a la quiebra, (2) los activos esperados después de T años, y (3) Las ganancias totales esperadas después de T años.

3. Proponer algoritmos (método y código, así como una corrida) para generar muestras de las siguientes densidades:

   (a) Cauchy

   $$f(x) = \frac{1}{\pi\beta\left[1 + \left(\frac{x-\gamma}{\beta}\right)^2\right]}$$

   donde $\gamma, x \in \Re$, $\beta > 0$.

   (b) Gumbel (o de valor extremo)

   $$f(x) = \frac{1}{\beta}\exp\left(-e^{-(x-\gamma)/\beta} - \frac{x-\gamma}{\beta}\right)$$

donde $\gamma, x \in \Re$, $\beta > 0$.

(c) Logística

$$f(x) = \frac{(1/\beta)e^{-(x-\gamma)/\beta}}{(1 + e^{-(x-\gamma)/\beta})^2}$$

donde $\gamma, x \in \Re$, $\beta > 0$.

(d) Pareto

$$f(x) = \frac{\alpha_2 c^{\alpha_2}}{x^{\alpha_2+1}}$$

donde $c > 0, \alpha_2 > 0, x > c$.

(e) Vasicek

$$f_{p,\rho}(x) = \sqrt{\frac{1-\rho}{\rho}} \exp\left(\frac{1}{2}\left\{\Phi^{-1}(x)^2 - \left(\frac{\sqrt{1-\rho}\Phi^{-1}(x) - \Phi^{-1}(p)}{\sqrt{\rho}}\right)^2\right\}\right)$$

donde $p, \rho \in (0, 1)$ y $\Phi(x)$ es la distribución normal estándar.

Para $\gamma = 0$ y $\beta = 1$ en cada inciso (a), (b) y (c), usen los algoritmos que obtuvieron para generar una muestra aleatoria de 5000 y obtengan $\bar{X}(n) = \sum_{i=1}^{n} X_i/n$ para $n = 50, 100, 150, \ldots, 5000$ para verificar empíricamente la ley fuerte de los grandes números. Hacer lo mismo para (d) con $c = 1$ y $\alpha_2 = 2$.

4. Grafiquen las siguientes densidades. Dar los algoritmos de transformación inversa, composición y aceptación-rechazo para cada una de las siguientes densidades. Discutir cuál algoritmo es preferible para cada densidad.

(a)

$$f(x) = \frac{3x^2}{2} I(x)_{[-1,1]}$$

(b) Para $0 < a < \frac{1}{2}$,

$$f(x) = \begin{cases} 0 \quad x \leq 0 \\ \frac{x}{a(1-a)} \quad 0 \leq x \leq a \\ \frac{1}{1-a} \quad a \leq x \leq 1-a \\ \frac{1-x}{a(1-a)} \quad 1-a \leq x \leq 1 \\ 0 \quad x \geq 1 \end{cases}$$

5. Considerando la transformación polar de Marsaglia para generar muestras de normales estándar, muestren que la probabilidad de aceptación de $S = V_1^2 + V_2^2$ en el paso 2 es $\pi/4$, y encuentren la distribución del número de rechazos de S antes de que ocurra una aceptación. ¿Cuál es el número esperado de ejecuciones del paso 1?

6. Obtengan una muestra de 1,000 números de la siguiente distribución discreta:

$$p(x) = \frac{2x}{k(k+1)}, x = 1, 2, \ldots, k$$

para $k = 100$.

7.
   - Desarrollen un algoritmo para generar una variable aleatoria binomial, usando la técnica de convolución (Hint: ¿cuál es la relación entre binomiales y Bernoullis?) Generar una muestra de 100,000 números. ¿Qué método es más eficiente, el de convoluciones o la función `rbinom` en R?
   - Ahora supongan que tienen Y como una suma de $n = 100$ variables Bernoulli con parámetro $p = 0.6$ correlacionadas, con correlación $\rho = 0.3$. Obtener por simulación la estimación de la densidad de Y.

8. Escribir una función para generar una mezcla de una distribución normal multivariada con dos componentes con medias μ_1 y μ_2 y matrices de covarianzas S_1 y S_2 respectivamente.

   a. Con el programa, generar una muestra de tamaño $n = 1000$ observaciones de una mezcla $50\,\%$ de una normal 4-dimensional con $\mu_1 = (0, 0, 0, 0)$ y $\mu_2 = (2, 3, 4, 5)$, y matrices de covarianzas $S_1 = S_2 = I_4$.
   - Obtener los histogramas de las 4 distribuciones marginales.

9. **Distribución de Wishart**. Suponer que $M = X'X$, donde X es una matriz $n \times d$ de una muestra aleatoria de una distribución $\mathcal{N}_d(\mu, \Sigma)$. Entonces M tiene una distribución Wishart con matriz de escala Σ y n grados de libertad, y se denota $W \sim W_d(\Sigma, n)$. Cuando $d = 1$, los elementos de X son una muestra aleatoria de una $\mathcal{N}(\mu, \sigma^2)$, por lo que $W_1(\sigma^2, n) \sim \sigma^2 \chi^2_{(n)}$.
   - Una forma de generar observaciones de una distribución Wishart, es generar muestras de multivariadas normales y calcular la matriz producto XX'. Programar este método. Noten que este método es muy costoso porque se tienen que generar nd valores aleatorios normales para determinar las $d(d+1)/2$ diferentes entradas de M.
   - Un método más eficiente se basa en la descomposición de Bartlett: sea $T = (T_{ij})$ una matriz triangular inferior de $d \times d$ con entradas independientes que satisfacen

     a) $T_{ij} \sim \mathcal{N}(0, 1)$ independientes, para $i > j$.

     b) $T_{ii} \sim \sqrt{\chi^2_{(n-i+1)}}$, $i = 1, \ldots, d$.

     Entonces la matriz $A = TT'$ tiene una distribución Wishart $W_d(I_d, n)$. Para generar variables $W_d(\Sigma, n)$, obtener la descomposición de Choleski $\Sigma = LL'$, donde L es triangular inferior. Entonces $LAL' \sim W_d(\Sigma, n)$. Implementar esta versión.

- Comparar en tiempo de ejecución de ambas versiones.

10. - Construyan un vector de 100 números crecientes y espaciados regularmente entre 0.1 y 20. Llámenlo `SIG2` . Ahora construyan otro vector de longitud 21 empezando en -1 y terminando en 1. Llámenlo `RHO`.
    - Para cada entrada σ^2 de `SIG2` y cada entrada de `RHO`:
        - Generar una muestra de tamaño $N = 500$ de una distribución bivariada normal $Z = (X, Y)$ donde $X \sim \mathcal{N}(0, 1)$ y $Y \sim \mathcal{N}(0, \sigma^2)$ y el coeficiente de correlación de X y Y es ρ. Z es una matriz de dimensiones 500×2.
        - Crear una matriz de 500×2, llámenlo `EXPZ`, con las exponenciales de las entradas de Z. ¿Qué distribución tienen estas variables transformadas?
        - Calculen el coeficiente de correlación, $\tilde{\rho}$ de las columnas de `EXPZ`. Grafiquen los puntos $(\sigma^2, \tilde{\rho})$ y comenten sobre lo que obtuvieron.