Comunidad Hispano

GRUPO DE USUARIOS DE R DE MADRID

Grupo R Madrid madRid

Fecha: Jueves 28 abril 2022
Lugar: Facultad de Estudios Estadísticos (sala S7-11) + Zoom
Hora: 7:00pm - 8:30pm
Acceso: Libre (plazas limitadas por protocolo Covid)
Cómo llegar: Buses desde Moncloa, 133, 83, 162 (cruzar la A6 por túnel) - F y G (dejan en Historia, cruzar A6 por debajo). Hay parking gratuito pero con número limitado de plazas.

Más detalles en: http://madrid.r-es.org

Presentaciones:

Juan M. Gutierrez
" Creación de bases de datos aleatorias".

Nuestros patrocinadores:

R consortium

UNED MÁSTER EN BIG DATA Y DATA SCIENCE

UNIVERSIDAD COMPLUTENSE MADRID
Facultad de Estudios Estadísticos (UCM)

Creación de bases de datos aleatorias

73 - Reunión Grupo de R: Jueves 28 de marzo de 2022

- Generación de bases de datos aleatoria para producción.
- Fabricar datos asintóticos manteniendo cierto rigor estadístico es una alternativa a la ausencia de bases de datos convencionales, principalmente porque a veces no se puede observar la población a estudiar.
- Crear una base de datos aleatoria es una buena alternativa para inferir esta población. Las investigaciones en salud son un buen ejemplo de su aplicación.
- En esta reunión veremos:
- Repaso a las técnicas de muestreo
- Generación de números aleatorios.
- Creación de una base de datos aleatoria desde 0
- Materiales, R y RStudio.

FUNDAMENTOS PARA LA CREACIÓN DE VARIABLES ALEATORIAS

MUESTREO

Tipos de Muestreo

TIPO	MÉTODO
Probabilístico Utilizan alguna forma de selección aleatoria de los puntos muestrales. Ventajas: son los únicos que permiten realizar inferencias, calcular el error de muestreo Desventajas: más complejo en su realización.	Aleatorio simple
	Sistemático
	Estratificado
	Por conglomerados
No probabilístico No involucran una selección aleatoria de los puntos muestrales. Ventajas: menos laborioso, más económico y de fácil realización. Desventajas: no permite realizar inferencias acerca de la población.	Juicio u opinión
	Por cuotas
	Accidental
	Por conveniencia
	Bola de nieve

Los métodos estadísticos y de aprendizaje automático se benefician cuando proporcionamos el volumen adecuado de datos sin embargo, podemos desarrollar buenos modelos incluso con conjuntos de datos razonablemente pequeños (estadísticamente).

El truco aquí es una técnica adecuada de muestreo de datos.
El muestreo no es más que un subconjunto de datos.

- Muestreo no probabilístico no es posible hacer inferencias sobre la población.
- Muestreo Aleatorio es posible hacer inferencias sobre la población

MUESTREO ALEATORIO SIMPLE Cada miembro de la población tiene la misma chance de ser seleccionado. Los individuos que constituirán la muestra son elegidos aleatoriamente mediante números aleatorios obtenidos por Tabla o programas en computadora. La población debe ser homogénea respecto a la variable de interés.

MUESTREO SISTEMÁTICO Se utiliza con frecuencia en lugar de un muestreo aleatorio. Cada individuo es seleccionado según la enésima posición de una lista de los miembros de la población, el primero de ellos se escoge al azar. Mientras que la lista no contiene ningún orden oculto, este método es tan bueno como el método de muestreo aleatorio. Su única ventaja con respecto a la técnica de muestreo aleatorio es la simplicidad

MUESTREO ESTRATIFICADO SIMPLE

Consiste en subdividir a la población en grupos homogéneos en función al estudio que se desea realizar. Es apropiado cuando la población ya está dividida en estratos y los estratos tienen diferente tamaño y es necesario tener ambos en cuenta. Refleja de forma más precisa las características de la población estratificada en comparación con otro tipo de muestras.

MUESTREO DE CONGLOMERADOS

La población está subdividida en subpoblaciones llamadas conglomerados. Los conglomerados deben presentar toda la variabilidad de la población. Además los conglomerados deben ser muy parecidos entre sí. La selección de los conglomerados que integran la muestra es al azar. Todos los elementos del conglomerado representan la población, de modo que conviene incluirlos a todos en la muestra. Una muestra de conglomerados, usualmente produce un mayor error muestral y es menos precisa en las estimaciones que una muestra aleatoria simple del mismo tamaño pero es menos costosa y mas rápida de muestrear.

MUESTREOS NO PROBABILÍSTICOS

- **Muestreo de juicio u opinión:** los elementos de la muestra son seleccionados mediante juicio personal.
- **Muestreo por cuotas:** se requiere conocer la población y/o los individuos más representativos. Se fijan cuotas que consisten en número de individuos con determinadas condiciones.
- **Muestreo accidental:** los individuos de la muestra se obtienen sin ningún plan, son elegidas producto de circunstancias casuales.
- **Muestreo incidental o de conveniencia:** se seleccionan directa e intencionalmente a los individuos de la población que formaran la muestra. Se usa en estudios exploratorios y en pruebas piloto.
- **Muestreo bola de nieve:** la premisa es que los elementos se relacionen entre sí. Se localizan algunos individuos de la población y estos conducen a otros que llevan a otros y así hasta tener una muestra de tamaño suficiente.

GENERACIÓN DE NÚMEROS Y VARIABLES DE FORMA ALEATORIA

- **Generación de números aleatorios:**
  - **GCLS Generadores Congruenciales Lineales**

- **Generación de variables aleatorias individuales:**
  - **Método de inversión para distribuciones continuas**
  - **Método de inversión para distribuciones discretas**
  - **Método de aceptación-rechazo**

- **Otros métodos para generar la N(0,1)**
  - **Método de Box-Müller**
  - **Método de Marsaglia**

- **Cálculo de integrales por simulación:**
  - **Método Hit or Miss**
  - **Método de la media muestral**
  - **Estimación de integrales no acotadas**

Generadores coherentes lineales simples (GCLS)

- **Cada número se calcula a partir del anterior usando una función lineal junto con una reducción modular:**

$$x_i = (ax_{i-1} + c) \mod m \quad 0 \le x_i < m \quad a > 0 \quad c \ge 0 \quad m > 0$$

a **= multiplicador**
c **= incremento**
m **= módulo**

- **De otra forma:** $x_i \equiv (ax_{i-1} + c) - m \cdot \mathrm{ENT}\left(\frac{ax_{i-1} + c}{m}\right)$

- **Si se desea una secuencia dentro del intervalo (0,1), entonces:**

$$u_i = \frac{x_i}{m} \longrightarrow u_i \equiv au_{i-1} \mod 1 \quad 0 < u_i < 1$$

GENERACIÓN DE VARIABLES ALEATORIAS INDIVIDUALES

Método de la inversión para distribuciones continuas

Sea X v.a. / $F(x) = P(X \le x) \wedge F'(x) \ge 0$

Entonces: ⇨

$$F^{-1}(z) = \min\{x / F(x) \ge z\} \quad / \quad 0 \le z \le 1$$
$$F^{-1}(0) = -\infty \quad \wedge \quad F^{-1}(1) = +\infty$$

Si la imagen de $F(x) \sim U(0,1)$, entonces $X=F^{-1}(U)$ tiene a F como función de distribución:

$$P(X \le x) = P\left[F^{-1}(U) \le x\right]$$

acumulando probabilidades:

$$P\left\{F\left[F^{-1}(U)\right] \le F(x)\right\} = P[U \le F(x)] = F(x)$$

Es decir, para generar un valor a partir de una distribución F:

1. se extrae un valor $u \in [0,1]$
2. se calcula $x = F^{-1}(u)$

En general sería el método preferible para la simulación de una variable continua (siempre que se disponga de la función cuantil). Está basado en los siguientes resultados:

Si $\overline{X}$ es una variable aleatoria con función de distribución F continua, y estrictamente monótona (invertible), entonces:

$$U = F(X) \sim \mathcal{U}(0,1)$$

➢ Caso de la distribución exponencial:

$$u = F(x) = 1 - e^{-\frac{x}{\mu}} \quad \forall x > 0 \quad \mu > 0$$

$$x = -\mu \ln(1-u) = F^{-1}(u)$$

La distribución exponencial exp(λ) de parámetro λ > 0 tiene como función de densidad $f_x = \lambda e^{-\lambda x}$ si x ≥ 0 y como función de distribución:

1. Generar $U \sim \mathcal{U}(0,1)$.
2. Devolver $X = -\dfrac{\ln(1-U)}{\lambda}$.

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

➢ Caso de la distribución logística:

$$u = F(x) = \frac{1}{1 + e^{-\frac{x-\alpha}{\beta}}} \quad -\infty < x < \infty \quad \beta > 0$$

$$x = \alpha + \beta ln\left(\frac{u}{1-u}\right)$$

Método de la tabla de búsqueda en distribuciones discretas

- **No se necesita conocer la función de distribución de partida.**
- **Sea X v.a. discreta:**

$$p(x_i) = P(X = x_i) \quad / \quad p(x_i) \geq 0 \quad \text{y} \quad \sum_{i=1}^{n} p(x_i) = 1$$

- **Supongamos que: $x_1 < x_2 < \cdots < x_n$. Entonces:**
  1. **se escoge un número aleatorio a partir de $U(0,1)$**
  2. **Regla de decisión:**

$$\begin{cases} X = x_1 & si \quad 0 \leq u < p(x_1) \\ X = x_j & si \quad \sum_{i=1}^{j-1} p(x_i) \leq u < \sum_{i=1}^{j} p(x_i) \end{cases}$$

Tabla de búsqueda conociendo la función de distribución

- Se puede acelerar el proceso si se conoce la relación entre $p(x_i)$ y $p(x_{i+1})$
- Ejemplo 1: $B(n,p)$

$$\text{Relación} = \frac{p(K=k+1)}{p(K=k)} = \frac{\binom{n}{k+1} p^{k+1}(1-p)^{n-k-1}}{\binom{n}{k} p^{k}(1-p)^{n-k}} = \frac{(n-k)p}{(k+1)(1-p)} = \frac{n-k}{k+1} \cdot C$$

- **P1: Definición de variables y valores iniciales:**

  $C = p/(1-p)$, $K = 0$, $PROB = P(K=0) = (1-p)^n$; $SA = PROB$

- **P2: Se genera un número aleatorio según $U(0,1)$**
- **P3: Comprobación: si $u < SA \rightarrow k^* = 0$, pero si $u \geq SA$, ir a P4**
- **P4: $PROB_{k+1} = PROB_k \cdot C \cdot (n-k)/(k+1)$**
- **P5: Se actualizan valores: $SA_{k+1} = SA_k + PROB_{k+1}$ y $K = k+1$**
- **P6: se vuelve a P3**

- **De forma más general: se desea generar los valores de $f(x)$ / $x \in [a, b]$. Entonces, se escoge otra función de densidad, $h(x)$, y una constante positiva c que envuelve $f(x)$ tal que:**

$$c \cdot h(x) \geq f(x) \quad siendo \quad c = \sup_{[a,b]} \left(\frac{f(x)}{h(x)} \right)$$

- **Los pasos del algoritmo son:**
  1. **se escoge un número aleatorio, tal que usando $h(x)$, se genere un valor, y**
  2. **se escoge un número aleatorio u distribuido según una $U(0,1)$**
  3. **Si** $u < \frac{f(y)}{c \cdot h(y)} \rightarrow$ **se acepta y como valor generado a partir de $f(x)$**

$g(y)$

Ejemplo: generación de una N(0,1)

$$F(z)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{x^2}{2}}\,dx \quad / \quad -\infty < x < \infty$$

- **Proceso en dos fases:**
  - **fase 1: se obtiene |z|**
  - **fase 2: se fija el signo al valor obtenido**

- **Elementos necesarios:**

$$f_{|z|}(x)=\frac{2}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \quad / \quad 0 \le x < \infty$$

$$h(x)=e^{-x} \quad / \quad 0 \le x < \infty$$

$$c=\sqrt{\frac{2e}{\pi}}=\frac{1}{eficiencia}=\frac{1}{0,76}$$

$$g(x)=exp\left[-\frac{(x-1)^2}{2}\right]$$

OBTENCIÓN DEL VALOR:

P1: se generan (u,v) a partir de una $U(0,1)$

P2: usando v y el método de la inversión para una función exponencial, se obtiene un valor de la densidad exponencial $x_1 = h(v) = -lnv$

P3: se calcula $g(x_1) = exp[-(x_1-1)^2/2]$

P4: se comprueba si $u < g(x_1)$:

si así ocurre, se acepta x como un valor positivo de $f_{|z|}(x)$

si no, se rechaza el par (u,v) y se vuelve a P1

OBTENCIÓN DEL SIGNO

P5: se genera w a partir de $U(0,1)$

P6: para los valores aceptados en P4:

si $w < 0,5 \rightarrow -x$

si $w \geq 0,5 \rightarrow +x$

Bibliografía de referencia

- **Gentle, J.E. (2003):** *Random Number Generations and Monte Carlo Methods. Second Edition.* Springer

- **Herzog, T.N. y Lord, G. (2002):** *Applications of Monte Carlo Methods to Finance and Insurance*. Actex Publications, Inc.

- **Robert, C.P. y Casella, G. (1999):** *Monte Carlo Statistical Methods*. Springer.

- **Rubinstein, R. (1981):** *Simulations and the Monte Carlo Method*. John Wiley & Sons.

Generación de una normal multivariante

$$f_X(x) = \frac{1}{(2\pi)^{-n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \quad \mathbf{X} \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- **Si Σ es definida positiva y simétrica, existe una única matriz triangular C tal que verifica que: Σ = CC´**

$$\mathbf{C} = \begin{bmatrix} c_{11} & 0 & \cdot & \cdot & 0 \\ c_{21} & c_{22} & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ c_{n1} & c_{n2} & \cdot & \cdot & c_{nn} \end{bmatrix} \Rightarrow \mathbf{X} = \mathbf{CZ} + \boldsymbol{\mu}$$

$\mathbf{Z} = [z_1 \quad z_2 \quad \cdot \quad \cdot \quad z_n]'$ ← **vector normal con media el vector nulo y Σ = I, siendo $z_i \sim N(0,1)$**

- **Problema: la obtención de C** ⇐ **Descomposición de Cholesky**

La descomposición de Cholesky

Es un proceso recurrente. Por ejemplo para un caso con tres variables:

Para x_1: $x_1 = c_{11} z_1 + \mu_1 \quad / \quad E(x_1) = c_{11} \cdot 0 + \mu_1 = \mu_1$

$$var(x_1) = c_{11}^2 = \sigma_{11} \Rightarrow c_{11} = \sqrt{\sigma_{11}}$$

Para x_2: $x_2 = c_{21} z_1 + c_{22} z_2 + \mu_2 \quad / \quad E(x_2) = \mu_2$

$$var(x_2) = \sigma_{22} = c_{21}^2 + c_{22}^2$$

$$cov(x_1, x_2) = \sigma_{12} = E\left[c_{11} z_1 (c_{21} z_1 + c_{22} z_2)\right] = c_{11} c_{21} E(z_1^2) = c_{11} c_{21}$$

Para x_3: $x_3 = c_{31} z_1 + c_{32} z_2 + c_{33} z_3 + \mu_3 \quad / \quad E(x_3) = \mu_3$

$$var(x_3) = \sigma_{33} = c_{31}^2 + c_{32}^2 + c_{33}^2$$

$$cov(x_1, x_3) = \sigma_{13} = E\left[c_{11} z_1 (c_{31} z_1 + c_{32} z_2 + c_{33} z_3)\right] = c_{11} c_{31} \Rightarrow c_{31} = \frac{\sigma_{31}}{c_{11}}$$

$$cov(x_2, x_3) = \sigma_{23} = E\left[(c_{21} z_1 + c_{22} z_2)(c_{31} z_1 + c_{32} z_2 + c_{33} z_3)\right] = c_{21} c_{31} + c_{22} c_{32}$$

- **Expresión general:**

$$c_{ij} = \frac{\sigma_{ij} - \sum_{k=1}^{j-1} c_{ik} c_{jk}}{\left(\sigma_{jj} - \sum_{k=1}^{j-1} c_{jk}^2 \right)^{1/2}}$$

siendo:

$$\sum_{k=1}^{0} c_{ik} c_{jk} = 0 \qquad \text{y} \qquad 1 \le j \le i \le n$$

- **Pasos del algoritmo:**
  - **P1: se genera** $Z = [z_1, z_2, \ldots, z_n]'$ / $z_i \sim N(0,1)$
  - **P2: se calculan los** c_{ij} **de C**
  - **P3: se obtiene X como** $X = CZ + \mu$

Bibliografía de referencia

- **Cario, M.C. y Nelson, B.L. (1997):** Modeling and generating random vectors with artibrary marginal distributions and correlation matrix. *Technique Report. Department of Industrial Engineering and Management Science, Northwestern University.*

- **Gentle, J.E. (2003):** *Random Number Generations and Monte Carlo Methods. Second Edition.* Springer

- **Herzog, T.N. y Lord, G. (2002):** *Applications of Monte Carlo Methods to Finance and Insurance*. Actex Publications, Inc.

- **Niavarani, M.R. y Smith, A.J.R. (2013):** Modeling and Generating Multi-Variate-Attribute Random Vectors Using a New Simulation Method Combined with NORTA Algorithm. *Journal of Uncertain Systems*. vol, 7, no 2, pp. 83-91

- **Robert, C.P. y Casella, G. (1999):** *Monte Carlo Statistical Methods*. Springer.

- **Rubinstein, R. (1981):** *Simulations and the Monte Carlo Method*. John Wiley & Sons.

BOOTSTRAPPING

Objetivo: evaluar el grado de exactitud que puede tener un estadístico calculado con una muestra (x_1,x_2,...,x_n) como indicativo del comportamiento global de una población

Técnica: regeneración de muestras un número elevado de veces para poder realizar inferencias.

Número de muestras posibles (Hall, 1992):

$$\binom{2n-1}{n} = \frac{(2n-1)!}{n!(n-1)!}$$

Punto de partida: $\mathbf{x} = (x_1, x_2, \ldots, x_n) \leftarrow$ **muestra aleatoria de tamaño** n**, con la cual se elabora un estadístico e(x) -Ejemplo: la media-**

Mediante algún mecanismo se escogen al azar números enteros comprendidos entre 1 y n**:**

$$j_1, j_2, \ldots, j_n \quad / \quad p(j_i) = \frac{1}{n} \quad \forall i = 1, \ldots, n$$

Nueva muestra:

$$x_1^* = x_{j_1}; x_2^* = x_{j_2}, \ldots, x_n^* = x_{j_n}$$

La muestra regenerada lo es con reemplazamiento $\Rightarrow j_i$ **se puede elegir varias veces**

Bibliografía de referencia

- **Efron, B. and Tibshirani, R. (1993):** *An Introduction to the Bootstrap*, Chapman and Hall.

- **England, P. and Verrall, R. (1999):** Analytic and bootstrap estimates of prediction errors in claim reserving. *Insurance: Mathematics and Economics*, 25: 281-293.

- **England, P. (2002):** Addendum to "Analytic and bootstrap estimates of prediction errors in claim reserving". *Insurance: Mathematics and Economics*, 31: 461-466

- **Hall, P. (1992):** *The Bootstrap and the Edgeworth Expansion,* Springer-Verlag.

- **Renshaw, A.E. (1994):** On the second moment properties and the implementation of certain GLIM based stochastic claims reserving models. *Actuarial Research Paper No. 65*, Department of Actuarial Science and Statistics. City University, London.

- **Renshaw, A.E. and Verrall, R.J. (1994):** A stochastic model underlying the chain-ladder technique. *Proceedings XXV ASTIN Colloquium*, Cannes.

Ahora un ejemplo de aplicación en R