Métodos de reducción de varianza

Clase nro 12

Curso 2010

Métodos de reducción de varianza

- En la mayoría de las simulaciones, los experimentos tienen por objetivo obtener valores medios de los resultados que se muestrean (de sus distribuciones).
- El método utilizado es el de realizar varias replicaciones independientes.
- Cuanto mayor sea la varianza obtenida, mayor debe ser la cantidad de replicaciones a realizar.

Replicaciones

En el ***método de las replicaciones*** se ejecutan *n* corridas independientes, se estima el valor de la respuesta media, $\mu_x = E(X)$ y la Var(X).

La mejora obtenida al realizar *n* replicaciones, en lugar de una, surge de:

$$\overline{X} = \frac{\sum X_i}{n} \quad \text{y} \quad Var(\overline{X}) = \frac{1}{n^2} n Var(X) = \frac{Var(X)}{n}$$

la reducción obtenida está dada por

$$Var(X) - Var(\overline{X}) = (1 - \frac{1}{n}) Var(X)$$

S.E.D 2010

Tasas constantes

- El uso de tasas de arribo y/o estadías constantes (determinísticas) reduce considerablemente la varianza de las muestras obtenidas como resultados.
- Usarlos con precaución, todas las características estocásticas del sistema pueden perderse, al quizás obtener datos sobre o sub-evaluados.
- La incorporación de valores constantes puede ser útil como forma de establecer cotas para las medidas buscadas.

S.E.D 2010

Torrentes comunes

- Esta técnica se usa para determinar diferencias entre resultados, cuando los niveles de los factores son cambiados.
- Si X_i es el resultado de una corrida respecto a un determinado nivel e Y_i es el resultado respecto de otro nivel, entonces

$$\overline{X} - \overline{Y}$$

la diferencia entre las medias resultantes de *n* corridas para cada nivel, puede ser usada para estimar el valor esperado de la diferencia

$$E[\overline{X} - \overline{Y}]$$

Torrentes comunes

Para reducir el número necesario de corridas queremos minimizar la varianza de la estimación:

$$Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y}) - 2Cov(\overline{X}, \overline{Y})$$

Si $\overline{X}$ e $\overline{Y}$ son independientes, entonces $Cov(\overline{X}, \overline{Y}) = 0$.

Observar que los resultados X_i e Y_i de la corrida i están emparejados de tal modo, que se puede asumir que la diferencia entre ambos, se debe nada más que al cambio en el nivel del factor, entonces

$\overline{X}$ e $\overline{Y}$ no son independientes

Si la covarianza es grande, la varianza de la diferencia será mucho menor.

Torrentes comunes

- Esta condición es verdadera siempre que se usen diferentes torrentes de números para cada distribución de los experimentos (si no, se puede introducir correlación entre ellos). Solamente se cambian los niveles de los factores entre pares de experimentos.
- Por ejemplo en el sistema del hospital se utilizan distintos torrentes para los arribos, el tiempo de las estadías de los pacientes a operarse, a no operarse, tiempos de operación, tiempos pos-operatorios.

Torrentes comunes

- Ejemplo: cambiamos el factor número de camas; si lo aumentamos, entonces el largo de las colas y los tiempos de espera disminuirán. Las tasas de arribos y los tiempos de las actividades se mantendrán intocables entre ambos experimentos, así como las mismas entidades utilizarán los recursos en el mismo orden.

  Del mismo modo podemos razonar con el factor tiempo de apertura de la sala de operaciones.

Torrentes comunes

- Si cambiamos un solo factor, entonces las diferencias entre experimentos es bastante determinada y los análisis estadísticos entre resultados emparejados son más manejables.
- En simulaciones más complejas y cuando se cambia más de un nivel, entonces los efectos en los resultados pueden ser impredecibles.

Torrentes comunes

- Otro problema surge al intentar repetir corridas de modo de comparar con diferentes conjuntos de números pseudoaleatorios.
- Tener especial cuidado de no repetir ningún torrente de números ya que esto podría llegar a causar correlación entre los resultados de los experimentos.

Torrentes comunes

En resumen, el método de usar diferentes torrentes para cada distribución, y utilizar torrentes comunes en experimentos utilizando distintos niveles en uno (o más) factor(es), es un método efectivo para reducir la varianza en simulaciones de tipo comparativas. Observar que no facilita la interpretación del análisis estadístico y requiere de un gran número de torrentes de números pseudoaleatorios.

Método antitético

- Este método se basa en las llamadas variables antitéticas y en la hipótesis de que si un torrente de números pseudoaleatorios produce resultados de valores altos, entonces el torrente opuesto producirá resultados bajos, por lo tanto están correlacionados negativamente.

- *Variables antitéticas* se llaman a dos conjuntos muy especiales de torrentes (streams) de números pseudoaleatorios: u1, u2, u3, u4, u5 ... y su complementario (1-u1), (1-u2), (1-u3), (1-u4) ...

Método antitético

- Si una simulación se corre con dos torrentes de números antitéticos, realizando n pares de corridas, el promedio de los resultados deberá estar más cerca del valor esperado que el promedio de corridas usando torrentes de números independientes.
- Esto es cierto en modelos pequeños y simples. No siempre se cumple en sistemas complejos donde por ejemplo la salida de alguna actividad es luego la tasa de entrada de otra, etc...

Método antitético

- Es un método fácil de implementar y puede ser utilizado para testear problemas sencillos.

- Se ejecutan pares de corridas

$$(X_1^{(1)}, X_1^{(2)}) \ \ldots \ (X_n^{(1)}, X_n^{(2)})$$

- $X_j^{(1)}$ es el resultado de la corrida j usando el torrente **u**.
- $X_j^{(2)}$ es el resultado de la corrida j usando el torrente **(1-u)**.

Método antitético

- Por ser $\mathbf{X_j^{(1)}}$, y $\mathbf{X_j^{(2)}}$ muestras resultantes del modelo, entonces

$$\mathbf{E(X_j^{(1)}) = E(X_j^{(2)}) = \mu}$$

- El total de replicaciones es 2n.
- Los pares de resultados son independientes entre sí. (Cada par de muestras es independiente de los otros).
- Como $(X_j^{(1)})$, $(X_j^{(2)})$ están correlacionados negativamente entonces existe reducción de varianza, en modelos sencillos.

Método antitético

Considero k = 1,..., n corridas y los pares

$$\mathbf{(X_j^{(1)}, X_j^{(2)}) \quad j = 1,...,n.}$$

$$(*) \qquad X_j = \frac{X_j^1 + X_j^2}{2} \text{ son v.a. independientes entre sí.}$$

Considero $\overline{X}(n) = \dfrac{\sum_{j=1}^{n} X_j}{n}$ entonces

$$\text{Var}\left(\overline{X}(n)\right) = \frac{\sum_{j=1}^{n} Var(X_j)}{n^2} = \left[X_j \text{ son } i.i.d\right] = \frac{Var(X_j)}{n}$$

$$\text{por } (*) \quad \text{Var}\left(\overline{X}(n)\right) = \frac{Var\left(X_j^1\right) + Var\left(X_j^2\right) + 2\,Cov\left(X_j^1, X_j^2\right)}{2^2 n}$$

Método antitético

- ***Var*(X_j^1)** se estima con los valores obtenidos con ***u*** y ***n*** corridas.
- ***Var*(X_j^2)** se estima con los valores obtenidos con ***(1-u)*** y ***n*** corridas.

- La covarianza se estima con fórmula que veremos más adelante sobre las ***2n*** corridas.

Método antitético

- **La *Var(X)* se puede estimar a partir de las 2n corridas.**

$$\mathrm{Var}\left(\overline{X}(n)\right) = \frac{Var\left(X_j^1\right) + Var\left(X_j^2\right) + 2Cov\left(X_j^1, X_j^2\right)}{2^2 n}$$

$$Cov_{XY}(n) = \frac{\sum_{j=1}^{n}\left[X_j - \overline{X}(n)\right]\left[Y_j - \overline{Y}(n)\right]}{n-1}$$

Método antitético

- En Pascal_SIM las variables antitéticas se definen mediante la variable booleana antithetic; se introduce un segundo conjunto de torrentes antitéticas a las 32 existentes. Cuando la variable antithetic = "verdadera" entonces se resta a 1 el valor de la función rnd (ver pág 154).
- En EOSimulator, puede crearse un generador adicional para generar variables antitéticas.

Método de variables de control

- Este método trata (también) de aprovechar la correlación entre variables para obtener cierta reducción de la varianza.
- Sea X la v.a. que representa un resultado, supongamos que queremos estimar $E(X) = \mu$.
- Supongamos también que en la simulación hay otra variable que está correlacionada con X y que
  CONOCEMOS su $E(Y) = \nu$.
- Nota: Y = c en Davies y O'Keefe.

Método de variables de control

- Vamos a utilizar nuestro conocimiento sobre Y
  En el sentido de que va a acercar a X a su media μ (hacia abajo o hacia arriba), reduciendo su variabilidad de una corrida a otra.
- Por eso llamamos a Y *variable de control* de X , la utilizaremos para ajustar X, es decir para controlarla parcialmente. Para ello debemos cuantificar el porte de ese ajuste.

Ejemplo: Y puede ser la variable aleatoria de los arribos en una fila de espera.

S.E.D 2010

Variables de control

- Sea ***a*** (***k*** en Davies y O'Keefe) una constante a determinar que tiene el mismo signo que la correlación entre **X** e **Y**.
- **(Y-ν)** es la desviación de **Y** con respecto de su media **ν.**
- Calculamos el valor controlado de X

  $$X_c = X - a\,(Y-\nu)$$
- Si **X** e **Y** están correlacionados positivamente entonces ***a > 0,*** por lo que ajustaremos X, hacia abajo si Y > ν, y hacia arriba si Y < ν.
- Si X e Y están correlacionados negativamente haremos lo opuesto.

S.E.D 2010

Variables de control

- $\mathbf{X_c = X - a\,(Y-\nu)}$, $\boldsymbol{E(X) = \mu}$ y $\boldsymbol{E(Y) = \nu}$,

entonces para cualquier ***a,***

- $\boldsymbol{E(X_C) = \mu}$ y
- $\boldsymbol{Var(X_C) = Var(X) + a^2\, Var(Y) - 2a\, Cov(X,Y).}$

o sea que la dispersión de $\boldsymbol{X_C}$ es menor que la de $\boldsymbol{X}$
s.s.i $\boldsymbol{2a\, Cov(X,Y) > a^2\, Var(Y).}$

Variables de control

- Lo importante de este método es elegir el mejor valor de *a* (*k* en Davies y O'Keefe), de forma de minimizar el valor de $Var(X_C)$.
- Calculando la derivada de $Var(X_C)$ *en a,* obtenemos: $2aVar(Y) - 2Cov(X,Y)$, igualamos a 0 para obtener el mejor valor

$$a = Cov(X,Y) / Var(Y).$$

Variables de control

$Var(X_C) = Var(X) + a^2\, Var(Y) - 2a\, Cov(X,Y).$

$a = Cov(X,Y) / Var(Y).$

$$Var\ (Xc\ ) = \frac{Var\ (X\ )(1-\rho^{2})}{n}$$

$$\rho = \frac{Cov(X,Y)}{S(Y)\ S(X)} = \frac{Cov\ (X\ ,Y\ )}{\sqrt{Var\ (X\ )Var\ (Y\ )}}$$

Variables de control

- En la práctica esto no es tan fácil, ya que depende de la naturaleza de la variable ***Y***, muchas veces no conocemos ***Var(Y).***
- Entonces se puede utilizar un método alternativo para calcular "***a***" con los datos de la propia simulación.

Variables de control

- Supongamos que realizamos ***n*** corridas independientes para obtener las observaciones ***Y1, Y2, ... Yn*** de ***Y,*** así como ***X1, X2, ... Xn*** de ***X***.
- Sean ***X(n) , Y(n)*** los promedios obtenidos y $S_Y^2(n)$ la varianza estimada de ***Y*** en esas ***n*** observaciones.
- Estimamos la ***covarianza*** y la constante ***a*** como:

$$\bar{C}_{XY}(n)=\frac{\sum_{j=1}^{n}\left[X_j-\bar{X}(n)\right]\left[Y_j-\bar{Y}(n)\right]}{n-1}$$

$$a^*(n)=\frac{\bar{C}_{XY}(n)}{S_Y^2(n)} \qquad \bar{X}_c(n)=\bar{X}(n)-a^*(n)\left[\bar{Y}(n)-v\right]$$

Variables de control

- Qué variables usar como variables de control, no es muy fácil de determinar. En un simple M/M/1 podríamos estimar X como la demora en cola y usar como variable de control los tiempos de servicios (cuya tasa conocemos).

  En este caso se utiliza nada más que una sola variable de control.
- Pero la varianza de la estimación de una respuesta se puede ver mejorada utilizando más de una variable de control, se debe adaptar la fórmula a ellas (pág 156).
- En la elección de variables de control recordar que los tiempos de las actividades influyen en los largos de las colas y también en el uso de los recursos implicados.

Variables de control

Sin embargo las tasas de arribo solo influyen en las primeras actividades por eso es conveniente:

a. Identificar aquellos factores que puedan estar correlacionados con las respuestas.
b. Estimar la varianza y el promedio de cada factor, así como la covarianza entre ellos y la respuesta asociada.
c. Ajustar la estimación de la respuesta mediante la ponderación de las variables de control.
d. Este método es muy usado para realizar experimentos, ya que luego de haber calculado **a) y b)** se pueden realizar muchas corridas cambiando niveles de factores e hipótesis.

(pag. 157) $\overline{X}_c = \overline{X} + k(\overline{C} - \mu_c)$

S.E.D 2010

Resumen

- El ***uso de torrentes comunes*** de números es usable en simulaciones comparativas.
- El método de ***variables antitéticas*** es útil en simulaciones simples.
- El ***método de variables de control*** es útil en caso de realizar simulaciones para por ejemplo explorar el efecto del uso de distintos datos o políticas en una organización.

Cada uno de estos métodos se deben adecuar a la complejidad del sistema a simular. Por eso estudiar cuidadosamente cuáles son los beneficios obtenidos, teniendo en cuenta el trabajo adicional requerido para implementar el elegido.

S.E.D 2010