Curso:
Métodos de Monte Carlo
Unidad 5, Sesión 12: Tópicos adicionales sobre intervalos de confianza

Departamento de Investigación Operativa
Instituto de Computación, Facultad de Ingeniería
Universidad de la República, Montevideo, Uruguay

dictado semestre 1 - 2024

Contenido:

1. Intervalos de confianza simultáneos.

2. Estimación de cocientes.

3. Estimación secuencial.

Intervalos de confianza simultáneos

En el curso hemos discutido como generar intervalos de confianza para la estimación de un valor a partir de una corrida de Monte Carlo.

Sin embargo, en muchas ocasiones deseamos estimar varios parámetros al mismo tiempo, con los datos de la misma corrida. Por ejemplo, supongamos que queremos calcular λ_1 y λ_2, y que utilizando Monte Carlo calculamos dos intervalos de confianza $[I_1(S_1, n, \delta), I_2(S_1, n, \delta)]$ y $[I_1(S_2, n, \delta), I_2(S_2, n, \delta)]$.

Si analizamos las estimaciones por separado, podemos afirmar que
$\text{Prob}\left(\lambda_1 \in [I_1(S_1, n, \delta), I_2(S_1, n, \delta)]\right) \geq 1 - \delta$ y que
$\text{Prob}\left(\lambda_2 \in [I_1(S_2, n, \delta), I_2(S_2, n, \delta)]\right) \geq 1 - \delta$.

En cambio, si nos interesa poder hacer una afirmación simultánea sobre ambos estimadores, es decir calcular la probabilidad conjunta
$\text{Prob}\left(\lambda_1 \in [I_1(S_1, n, \delta), I_2(S_1, n, \delta)] \text{ y } \lambda_2 \in [I_1(S_2, n, \delta), I_2(S_2, n, \delta)]\right)$, la situación no es tan simple.

Si las estimaciones fueran independientes (obtenidas realizando dos corridas separadas, con distintas secuencias de números aleatorios), tendríamos que

Prob $(\lambda_1 \in [I_1(S_1, n, \delta), I_2(S_1, n, \delta)]$ y $\lambda_2 \in [I_1(S_2, n, \delta), I_2(S_2, n, \delta)]) =$
Prob $(\lambda_1 \in [I_1(S_1, n, \delta), I_2(S_1, n, \delta)])$ Prob $(\lambda_2 \in [I_1(S_2, n, \delta), I_2(S_2, n, \delta)]) \geq$
$(1 - \delta)^2 = 1 - 2\delta + \delta^2$ (notar que aún en este caso, el más simple, el nivel de confianza del intervalo conjunto cambia significativamente respecto al nivel de confianza de cada intervalo individual).

Esto implica un mayor gasto de recursos computacionales, por lo que en general vamos a utilizar una única corrida para calcular ambos parámetros. En este caso las estimaciones no son independientes (ya que son obtenidas con los mismos datos), y por lo tanto no podemos aplicar la fórmula previa directamente.

En su lugar, usaremos una fórmula propuesta por Bonferroni (nota biográfica:
`https://www.encyclopediaofmath.org/index.php/Bonferroni,_Carlo_Emilio` -

último acceso: 2023-03-02), que se basa en la desigualdad del mismo nombre (también conocida como desigualdad de Boole `http://mathworld.wolfram.com/BonferroniInequalities.html` -último acceso: 2023-03-02, y que deriva directamente del principio de inclusión-exclusión `http://mathworld.wolfram.com/Inclusion-ExclusionPrinciple.html` - último acceso: 2023-03-02).

Específicamente, si A_i, $i = 1, \ldots, L$ son L eventos distintos (no necesariamente independientes), la desigualdad de Bonferroni expresa que

$$\text{Prob}\left(\bigcap_{i=1}^{L} A_i\right) \geq 1 - \sum_{i=1}^{L}(1 - \text{Prob}\,(A_i)) = \sum_{i=1}^{L} \text{Prob}\,(A_i) - (L-1).$$

En nuestro caso, supongamos que queremos estimar L parámetros λ_i. Vamos a denotar A_i el evento $(\lambda_i \in [I_1(S_i, n, \delta), I_2(S_i, n, \delta)])$, por lo tanto $\text{Prob}\left(\bigcap_{i=1}^{L} A_i\right)$ corresponde a la probabilidad de que todos los parámetros de forma simultánea pertenezcan a los respectivos intervalos de confianza.

Dado que por construcción de los intervalos de confianza, tenemos que $\text{Prob}\,(A_i) \geq 1 - \delta$, por lo tanto $1 - \text{Prob}\,(A_i) \leq \delta$, y aplicando la desigualdad de Bonferroni, llegamos a

$$\text{Prob}\left(\bigcap_{i=1}^{L} A_i\right) \geq 1 - \sum_{i=1}^{L}(1 - \text{Prob}\,(A_i)) \geq 1 - L\delta.$$

Si bien esta fórmula es conservadora, resulta útil en la práctica siempre que el número de parámetros a estimar no crezca demasiado.

Es posible emplearla en cualquiera de las formas siguientes:

- Si ya tenemos predeterminado el nivel de confianza $1 - \delta$ para cada parámetro, podemos calcular el nivel de confianza $1 - L\delta$ que resultará para la hipótesis de pertenencia simultánea de todos los parámetros a los respectivos intervalos.

  En este caso, al aumentar el número de parámetros L, tenemos que

disminuye proporcionalmente la confianza con la que podemos afirmar se dará la pertenencia simultánea.

- Si en cambio tenemos predeterminado el nivel de confianza $1 - \delta'$ que deseamos obtener para la hipótesis de pertenencia simultánea de todos los parámetros a los respectivos intervalos, entonces podemos determinar para cada parámetro un intervalo de nivel $1 - \delta$ con $\delta = \delta'/L$, de esta forma obtendremos el nivel conjunto deseado.

  En este caso, al aumentar el número de parámetros L, tenemos que aumentará el nivel de confianza que exigiremos para cada intervalo individual; esto resultará en intervalos cada vez más anchos, disminuyendo la precisión de las estimaciones.

Las observaciones precedentes muestran las limitaciones del método; en ambos casos, es posible paliar parcialmente el problema aumentando el número de replicaciones para la corrida de Monte Carlo, n, en la medida de la disponibilidad de tiempo de cálculo.

Intervalos de confianza para un cociente

Sean $\mathcal{R}_1$ y $\mathcal{R}_2$ dos regiones en $\mathcal{J}^m$ con volúmenes positivos $\lambda(\mathcal{R}_1)$ y $\lambda(\mathcal{R}_2)$. Supongamos que deseamos estimar $\theta = \lambda(\mathcal{R}_1)/\lambda(\mathcal{R}_2)$. Se puede emplear de forma simple el método de Monte Carlo para calcular al mismo tiempo los estimadores $\bar{\lambda}_n(\mathcal{R}_1)$ y $\bar{\lambda}_n(\mathcal{R}_2)$ de $\lambda(\mathcal{R}_1)$ y $\lambda(\mathcal{R}_2)$ respectivamente, y emplear

$$\bar{\theta}_n = \left\{ \begin{array}{cc} \bar{\lambda}_n(\mathcal{R}_1)/\bar{\lambda}_n(\mathcal{R}_2) & \text{si } \lambda_n(\mathcal{R}_2) > 0 \\ 0 & \text{si } \lambda_n(\mathcal{R}_2) = 0 \end{array} \right.)$$

como estimador de θ.

Esa es la forma más simple del problema general de estimación de cocientes, que aparece en muchas situaciones concretas. Por ejemplo, si $\mathcal{R}_1 \subseteq \mathcal{R}_2$, entonces θ es la proporción del volumen de $\mathcal{R}_2$ que está en $\mathcal{R}_1$. En el caso de problemas de conteo, si $\mathcal{S}$ es un conjunto finito, $\mathcal{S}_\alpha$ el subconjunto de $\mathcal{S}$ con atributo α, $\mathcal{S}_\beta$ el subconjunto de $\mathcal{S}$ con atributo β, entonces en muchos casos es de interés calcular $\theta = |\mathcal{S}_\alpha \cap \mathcal{S}_\beta|/|\mathcal{S}_\alpha|$, la

proporción de elementos de $\mathcal{S}_\alpha$ que tienen también el atributo β (que es posible calcular sin necesidad de conocer $|\mathcal{S}|$).

Dado que $\bar{\theta}_n$ es el cociente de dos variables aleatorias (eventualmente correlacionadas, si se calculan en base a las mismas replicaciones), no es posible extrapolar sus propiedades directamente de las fórmulas que hemos utilizado en las secciones previas.

El siguiente teorema (que enunciamos aquí sin dar la demostración):

Teorema 1. *Sean X_i, Y_i, $i = 1, \ldots, n$ dos secuencias de variables aleatorias, cada secuencia i.i.d., con $\mu_X = E(X_i) \neq 0$, $\mu_Y = E(Y_i)$, $E\left(|X_i|^j\right) < \infty$ para $j \geq 1$, $\sigma_X^2 = Var(X_i)$, $\sigma_Y^2 = Var(Y_i) < \infty$, y $\sigma_{XY} = cov(X_i, Y_i)$. Definimos $\bar{X}_n = \sum_{i=1}^n X_i/n$, $\bar{Y}_n = \sum_{i=1}^n Y_i/n$, y $\theta = \mu_Y/\mu_X$. Entonces*

$$\lim_{n\to\infty} nE\left(\bar{Y}_n/\bar{X}_n - \theta\right) = \theta\left(\sigma_X^2/\mu_X^2 - \sigma_{XY}/(\mu_X\mu_Y)\right).$$

y

$$\lim_{n\to\infty} nE\left((\bar{Y}_n/\bar{X}_n - \theta)^2\right) = \theta^2\left(\sigma_X^2/\mu_X^2 - 2\sigma_{XY}/(\mu_X\mu_Y) + \sigma_Y^2/\mu_Y^2\right).$$

Este resultado es importante porque indica que el estimador $\bar{\theta}_N = \bar{Y}_n/\bar{X}_n$ no es insesgado, sino que tiene un error con un término dominante de orden $1/n$ (es asintóticamente insesgado, pero para n finito hay un error sistemático, más allá del introducido por la naturaleza aleatoria del muestreo).

Por lo tanto, para estimar θ, es preferible emplear el estimador

$$\tilde{\theta}_n = \bar{\theta}_n\left(1 + \left(\hat{\sigma}_X^2/\bar{X}_n^2 - \hat{\sigma}_{XY}/(\bar{X}_n\bar{Y}_n)\right)/n\right),$$

donde $\hat{\sigma}_{XY} = 1/(n-1)\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$ es un estimador de la covarianza entre X_i e Y_i (y los otros estimadores son los usuales).
Resultados de la literatura indican que $\lim_{n\to\infty} n\mathrm{E}\left(\tilde{\theta}_n - \theta\right) = 0$,

mostrando que este estimador corrige el término en orden $1/n$, por lo que posee un menor sesgo que el anterior. Además,

$$\lim_{n\to\infty} n\text{Var}\left(\tilde{\theta}_n\right) = \theta^2\left(\sigma_X^2/\mu_X^2 - 2\sigma_{XY}/(\mu_X\mu_Y) + \sigma_Y^2/\mu_Y^2\right),$$

por lo que la varianza del nuevo estimador tiene un comportamiento asintótico idéntico al de la anterior al menos hasta los términos de orden $1/n$.

En el caso especial en que X_i e Y_i son v.a. de Bernoulli tales que $Y_i \leq X_i$ (caso de especial interés en este curso), el sesgo remanente en $\bar{\theta}_n$ es exactamente $\text{E}\left(\bar{\theta}_n - \theta\right) = -\theta(1-\mu_X)^n$, por lo que tenemos una convergencia mejor que la esperable de los resultados genéricos precedentes.

Para calcular un intervalo de confianza, podemos utilizar una estimación de la varianza basada en el resultado discutido recientemente, insertándola en los intervalos de confianza estándar presentados en la Unidad 2

(sesiones 4 y 5), y recordando que es una fórmula aproximada ya que está descartando los términos de orden superior a $1/n$.

El estimador de la varianza Var $\left(\tilde{\theta}_n\right)$ sería entonces

$$V(\theta_n) = \bar{\theta}_n^2 \left(\hat{\sigma}_X^2/\bar{X}_n^2 - 2\hat{\sigma}_{XY}/(\bar{X}_n\bar{Y}_n) + \hat{\sigma}_Y^2/\bar{Y}_n^2\right)/n,$$

Estimación secuencial

Hasta el momento hemos siempre trabajado bajo la hipótesis de que el tamaño de muestra, n, se determina previamente a la realización de los experimentos, en base a consideraciones sobre el error deseado y el nivel de confianza a obtener (eventualmente empleando información obtenida en un experimento preliminar con n' muestras).

Sin embargo, puede ser de interés el intentar utilizar la propia información obtenida en el transcurso del muestreo, de manera de generar una condición de parada del muestreo cuando se detecta que se ha alcanzado la especificación de error y nivel de confianza deseados.

Esto permitiría un mejor uso de la capacidad de cómputo, de manera de no desaprovechar ciclos de cálculo cuando ya se ha obtenido la precisión especificada (y de no detener una simulación cuando todavía no se ha alcanzado).

Los métodos que intentan cumplir con este objetivo se conocen bajo el nombre de *métodos de estimación secuencial*. Si bien son

conceptualmente atractivos, en un contexto general su comportamiento incluye un error de aproximación, que sólo se desvanece cuando el error a obtener, ϵ, tiende a 0, esta limitación hace que no sean demasiado usados en la práctica.

Damos a continuación uno de estos métodos, expresado en el siguiente teorema (Chow y Robbins, 1965).

Teorema 2. *Sea β_i una secuencia de constantes positivas tales que $\lim_{k\to\infty} \beta_k = \beta = \Phi^{-1}(1-\delta/2)$.*

Si se realiza un muestreo Monte Carlo siguiendo este plan:

1. *entradas:* $\epsilon > 0$, $0 < \delta < 1$, $n_0 > 0$.

2. $N = n_0$

3. *Realizar n_0 replicaciones $X_1, \ldots, X_{n_0}$.*

4. *Repetir*

5. *N=N+1*

6. *Realizar replicación* N*-ésima,* X_N

7. $\bar{X}_N = (1/N) \sum_{i=1}^{N} X_i$

8. $T_N = \sum_{i=1}^{N} (X_i - X_N)^2$

9. *Hasta que* $(T_N + 1)/N \leq \epsilon^2 N/\beta_N^2$.

Entonces

$$\lim_{\epsilon \to 0} N/n_0(\epsilon, \delta) = 1 \text{ con probabilidad 1},$$

$$\lim_{\epsilon \to 0} \textit{Prob}\left(|\bar{X}_N - \mu| \leq \epsilon\right) = 1 - \delta,$$

$$\lim_{\epsilon \to 0} \textit{E}\left(N\right)/n_0(\epsilon, \delta) = 1,$$

donde

$$n_0(\epsilon, \delta) = \beta^2 \sigma^2 / \epsilon^2.$$

A partir de este teorema, podemos implementar el método, tomando un valor n_0 arbitrario para inicializar las estimaciones, y luego haciendo agregando una estimación por vez, hasta cumplir la condición de fin de bucle, momento en el que (aproximadamente) estaremos cumpliendo la especificación de error dada como entrada.

Preguntas para auto-estudio

- ¿Qué significa estimar intervalos de confianza simultáneos? ¿Qué dificultades hay para hacerlo con una única corrida? ¿Y que dificultades con múltiples corridas independientes?

- ¿Por qué no es conveniente usar el estimador más intuitivo $\bar{\theta}_N = \bar{Y}_n / \bar{X}_n$ para estimar el cociente de dos medidas o variables aleatorias? ¿Qué otro estimador puede usarse?

- ¿Qué significa realizar estimación secuencial?