Métodos de reducción de varianza

1. Eficiencia de un estimador

La eficiencia de un estimador $\hat{\theta}(\boldsymbol{X})$ de un parámetro θ de un vector aleatorio $\boldsymbol{X}$ es

$$\mathsf{Eff}\big(\hat{\theta}(\boldsymbol{X})\big) \stackrel{\text{def}}{=} \frac{1}{\mathsf{MSE}\big(\hat{\theta}(\boldsymbol{X})\big)\mathsf{Coste}\big(\hat{\theta}(\boldsymbol{X})\big)}$$

donde

$$\mathsf{MSE}\big(\hat{\theta}(\boldsymbol{X})\big) \stackrel{\text{def}}{=} \mathbb{E}\Big[\big(\hat{\theta}(\boldsymbol{X}) - \theta\big)^2\Big] = \mathrm{Var}\big(\hat{\theta}(\boldsymbol{X})\big) + \mathsf{Sesgo}\big(\hat{\theta}(\boldsymbol{X}), \theta\big)^2$$
$$\mathsf{Coste}\big(\hat{\theta}(\boldsymbol{X})\big) \stackrel{\text{def}}{=} \text{tiempo esperado de cálculo de } \hat{\theta}(\boldsymbol{X})$$

Dados dos estimadores insesgados, $\hat{\theta}_1(\boldsymbol{X})$ y $\hat{\theta}_2(\boldsymbol{X})$, que requieren el mismo tiempo de cálculo, se tiene que

$$\mathsf{Eff}\big(\hat{\theta}_1(\boldsymbol{X})\big) > \mathsf{Eff}\big(\hat{\theta}_2(\boldsymbol{X})\big) \iff \mathrm{Var}\big(\hat{\theta}_1(\boldsymbol{X})\big) < \mathrm{Var}\big(\hat{\theta}_2(\boldsymbol{X})\big)$$

Para $\hat{\mu}_n(\boldsymbol{X})$ el estimador de Montecarlo de $\mu = \mathbb{E}\big[g(\boldsymbol{X})\big]$ se tiene que

$$\mathrm{Var}\big(\hat{\mu}_n(\boldsymbol{X})\big) = \frac{\sigma^2}{n} \qquad \mathsf{Sesgo}\big(\hat{\mu}_n(\boldsymbol{X}), \mu\big) = 0 \qquad \mathsf{Coste}\big(\hat{\mu}_n(\boldsymbol{X})\big) = cn$$

Por tanto, $\mathsf{Eff}\big(\hat{\mu}_n(\boldsymbol{X})\big) = \frac{1}{c\sigma^2}$. Es decir, al aumentar el tamaño de la muestra, la reducción en varianza se compensa exactamente con el incremento del tiempo de computación, por lo que no hay ganancia en eficiencia.

Para encontrar un estimador más eficiente se necesita, por tanto, reducir la varianza en un factor superior al incremento en tiempo de computación.

No hay una regla fija sobre cuán grande debe ser la mejora de la eficiencia proporcionada por un método de reducción de varianza para que valga la pena usarlo. En algunos escenarios, como la representación de gráficos por ordenador para películas animadas, donde miles de CPUs se mantienen ocupadas durante meses, una mejora del $10\,\%$ aporta un ahorro significativo. En otros escenarios, como un cálculo único, una ganancia de factor 60 que convierte un minuto en un segundo de espera puede que no justifique el coste de la programación de un método más complicado.

La ganancia de eficiencia necesaria para justificar el uso de un método es menor si el esfuerzo de programación puede amortizarse en muchas aplicaciones. El umbral es alto para un programa de un solo uso, más bajo por algo que añadimos a nuestra biblioteca personal, más bajo aún para código que se comparte y aún más bajo para código que se incluye en una biblioteca o herramienta de simulación de uso general.

2. Método del muestreo estratificado

Sean $\boldsymbol{X}$ un vector aleatorio (con d componentes) que sigue una distribución de probabilidad dada por la función de densidad f y $g\colon \mathbb{R}^d \to \mathbb{R}$ una función medible. Entonces,

$$\mu = \mathbb{E}\big[g(\boldsymbol{X})\big] = \int_D g(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$$

donde $D = \{\boldsymbol{x} \in \mathbb{R}^d \mid f(\boldsymbol{x}) > 0\}$.

Consideremos una partición de D en los estratos $D_1, \dots, D_J$ tal que para cada $j = 1, \dots, J$ se conoce $p_j = \mathbb{P}(\boldsymbol{X} \in D_j)$ y se sabe cómo generar valores aleatorios de la variable condicionada $\boldsymbol{X} \,|\, \boldsymbol{X} \in D_j$ (que sigue una distribución de probabilidad dada por la función de densidad $f_j(\boldsymbol{x}) = {}^1\!/_{p_j} f(\boldsymbol{x}) \mathbb{1}_{D_j}(\boldsymbol{x})$).

Para estimar el valor de μ mediante el método del muestreo estratificado, basta entonces calcular

$$\hat{\mu}_{\text{str},n} = \sum_{j=1}^{J} p_j \left(\frac{1}{n_j} \sum_{i=1}^{n_j} g(\boldsymbol{x}_{ij}) \right)$$

donde $\boldsymbol{x}_{1j}, \dots, \boldsymbol{x}_{n_j j} \overset{\text{i. i. d.}}{\sim} \boldsymbol{X} \,|\, \boldsymbol{X} \in D_j$ y $n_1 + \cdots + n_J = n$.

Se tiene que, en efecto, $\hat{\mu}_{\text{str},n}$ es un estimador insesgado de μ:

$$\begin{aligned}
\mathbb{E}\big[\hat{\mu}_{\text{str},n}\big] &= \sum_{j=1}^{J} p_j \mathbb{E}\left[\frac{1}{n_j} \sum_{i=1}^{n_j} g(\boldsymbol{x}_{ij}) \right] \\
&= \sum_{j=1}^{J} p_j \mathbb{E}\big[g(\boldsymbol{X}) \,|\, \boldsymbol{X} \in D_j\big] \\
&= \sum_{j=1}^{J} p_j \int_{\mathbb{R}^d} g(\boldsymbol{x}) f_j(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \sum_{j=1}^{J} \int_{D_j} g(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_D g(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \mathbb{E}[g(\boldsymbol{x})]
\end{aligned}$$

Denotemos $\mu_j = \mathbb{E}\big[g(\boldsymbol{X}) \,|\, \boldsymbol{X} \in D_j\big]$ y $\sigma_j^2 = \operatorname{Var}\big(g(\boldsymbol{X}) \,|\, \boldsymbol{X} \in D_j\big)$. Entonces la varianza del estimador viene dada por

$$\operatorname{Var}\big(\hat{\mu}_{\text{str},n}\big) = \sum_{j=1}^{J} p_j^2 \frac{\sigma_j^2}{n_j}$$

Esto proporciona una guía para elegir los estratos, ya que cuanto menor sean las σ_j, es decir, cuanto más constante sea la función g en cada estrato, menor será el error estándar $\sqrt{\operatorname{Var}\big(\hat{\mu}_{\text{str},n}\big)}$ de la estimación de μ por el método del muestreo estratificado.

Para construir un intervalo de confianza, podemos estimar el error estándar como sigue:

$$\hat{\mu}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} g(\boldsymbol{x}_{ij})$$

$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\big(g(\boldsymbol{x}_{ij}) - \hat{\mu}_j\big)^2$$

$$\widehat{\mathrm{Var}}\big(\hat{\mu}_{\mathrm{str},n}\big) = \sum_{j=1}^{J} p_j^2 \frac{\hat{\sigma}_j^2}{n_j}$$

Obsérvese que para que $\hat{\mu}_{\mathrm{str},n}$ esté bien definido es necesario que $n_j \geq 1$, para cada $j = 1, \ldots, J$, y para que $\widehat{\mathrm{Var}}\big(\hat{\mu}_{\mathrm{str},n}\big)$ esté bien definido es necesario, además, que $n_j \geq 2$, para cada $j = 1, \ldots, J$.

El estimador $\hat{\mu}_{\mathrm{str},n}$ depende del tamaño de cada uno de los estratos. Una manera simple de elegir esos tamaños es mediante una *asignación proporcional*, $n_j = p_j n$ (si, además, $p_j = 1/J$ para todo $j = 1, \ldots, J$, la asignación se dice que es *sistemática*). Entonces,

$$\hat{\mu}_{\mathrm{str,prop},n} = \sum_{j=1}^{J} p_j\left(\frac{1}{n_j}\sum_{i=1}^{n_j} g(\boldsymbol{x}_{ij})\right) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} g(\boldsymbol{x}_{ij})$$

y

$$\mathrm{Var}\big(\hat{\mu}_{\mathrm{str,prop},n}\big) = \sum_{j=1}^{J} p_j^2 \frac{\sigma_j^2}{n_j} = \frac{1}{n}\sum_{j=1}^{J} p_j \sigma_j^2$$

Por otra parte, por la ley de la varianza total se tiene que

$$\begin{aligned}\mathrm{Var}(g(\boldsymbol{X})) &= \mathbb{E}\Big[\mathrm{Var}\big(g(\boldsymbol{X}) \mid \boldsymbol{X} \in D_j\big)\Big] + \mathrm{Var}\Big(\mathbb{E}\big[g(\boldsymbol{X}) \mid \boldsymbol{X} \in D_j\big]\Big) \\ &= \sum_{j=1}^{J} p_j \sigma_j^2 + \sum_{j=1}^{J} p_j\big(\mu_j - \mu\big)^2\end{aligned}$$

Es decir, si denotamos $\sigma_{\mathrm{W}}^2 = \sum_{j=1}^{J} p_j \sigma_j^2$ a la varianza en los estratos y $\sigma_{\mathrm{B}}^2 = \sum_{j=1}^{J} p_j\big(\mu_j - \mu\big)^2$ a la varianza entre los estratos, entonces

$$\mathrm{Var}\big(\hat{\mu}_n\big) = \frac{1}{n}\big(\sigma_{\mathrm{W}}^2 + \sigma_{\mathrm{B}}^2\big) \qquad \text{y} \qquad \mathrm{Var}\big(\hat{\mu}_{\mathrm{str,prop},n}\big) = \frac{1}{n}\sigma_{\mathrm{W}}^2$$

Por lo tanto, siempre se tiene que $\mathrm{Var}\big(\hat{\mu}_{\mathrm{str,prop},n}\big) \leq \mathrm{Var}\big(\hat{\mu}_n\big)$. Además, cuanto mayor sea σ_{B}^2 con respecto a σ_{W}^2, mayor será la reducción de varianza conseguida mediante el método del muestreo estratificado con asignación proporcional.

La asignación proporcional no es necesariamente la mejor. Por ejemplo, dados dos estratos del mismo tamaño, pero con diferente varianza condicional, es más beneficioso generar menos valores en el estrato de menor varianza.

Asumiendo que el coste de generar cada valor es el mismo en cada estrato, la asignación óptima viene dada por

$$n_j = \frac{n p_j \sigma_j}{\sum_{k=1}^{J} p_k \sigma_k}, \quad j = 1, \ldots, J$$

siendo $\mathrm{Var}(\hat{\sigma}_{\text{str,ópt},n}) = \frac{1}{n}\left(\sum_{j=1}^{J} p_j \sigma_j\right)^2$.

La aplicación práctica de esta asignación óptima se ve dificultada por el hecho de que los σ_j no se conocen usualmente. Se puede usar un procedimiento en dos etapas:

1. En una primera etapa de tanteo se usa una asignación proporcional para generar valores $\boldsymbol{x}_{1j}, \ldots, \boldsymbol{x}_{n_j j} \overset{\text{i. i. d.}}{\sim} \boldsymbol{X} \,|\, \boldsymbol{X} \in D_j$ que se usan para estimar σ_j^2 mediante la cuasivarianza muestral:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} g(\boldsymbol{x}_{ij})$$
$$\hat{\sigma}_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(g(\boldsymbol{x}_{ij}) - \hat{\mu}_j\right)^2$$

2. En la etapa de producción se usan las estimaciones $\hat{\sigma}_j^2$ obtenidas para calcular el tamaño de muestra óptimo de cada estrato y realizar la estimación $\hat{\mu}_{\text{str,ópt},n}$.

La cantidad total de valores generados en la fase de tanteo debería ser pequeña, para que el coste no sea excesivo, pero lo suficientemente grande para obtener buenas estimaciones de σ_j. En caso contrario, podría ocurrir que la varianza de la estimación obtenida superara al de la estimación con asignación proporcional, e incluso al de la estimación por el método de Montecarlo directo.

3. Método del muestreo por importancia

Sean $\boldsymbol{X}$ un vector aleatorio con función de densidad *nominal* f_1 y $g\colon \mathbb{R}^d \to \mathbb{R}$ una función medible. Entonces

$$\mu = \mathbb{E}_{f_1}\left[g(\boldsymbol{X})\right] = \int_{\mathbb{R}^d} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x}$$

El método de muestreo por importancia consiste en considerar una función de densidad *instrumental* f_2 para $\boldsymbol{X}$, de tal manera que f_1 sea absolutamente continua con respecto a f_2 (es decir, $f_2(\boldsymbol{x}) \neq 0$ para todo $\boldsymbol{x}$ tal que $f_1(\boldsymbol{x}) \neq 0$). Si denotamos el soporte de una función f como $\text{sop}_f = \{\boldsymbol{x} \in \mathbb{R}^d \mid f(\boldsymbol{x}) \neq 0\}$ y definimos $L(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}$ como la *razón de verosimilitud*, se tiene que

$$\begin{aligned}
\mathbb{E}_{f_2}\left[g(\boldsymbol{x})L(\boldsymbol{x})\right] &= \int_{\mathbb{R}^d} g(\boldsymbol{x})L(\boldsymbol{x})f_2(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\text{sop}_{f_2}} g(\boldsymbol{x})L(\boldsymbol{x})f_2(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\text{sop}_{f_2}} g(\boldsymbol{x})f_1(\boldsymbol{x})\, d\boldsymbol{x}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{\text{sop}_{f_1}} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} + \int_{\text{sop}_{f_2} \setminus \text{sop}_{f_1}} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\text{sop}_{f_1}} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \int_{\mathbb{R}^d} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} \\
&= \mathbb{E}_{f_1}[g(\boldsymbol{x})] \\
&= \mu
\end{aligned}
$$

Se tiene entonces que el estimador

$$
\hat{\mu}_{\text{is},f_2,n}(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{x}_i) L(\boldsymbol{x}_i)
$$

donde $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n \overset{\text{i. i. d.}}{\sim} f_2$, es un estimador insesgado de μ.

Además,

$$
\begin{aligned}
\text{Var}_{f_2}\big(\hat{\mu}_{\text{is},f_2,n}(\boldsymbol{X})\big) &= \frac{1}{n} \text{Var}_{f_2}\big(g(\boldsymbol{X}) L(\boldsymbol{X})\big) \\
&= \frac{1}{n}\Big(\mathbb{E}_{f_2}\big[g(\boldsymbol{X})^2 L(\boldsymbol{X})^2\big] - \mathbb{E}_{f_2}\big[g(\boldsymbol{X}) L(\boldsymbol{X})\big]^2\Big) \\
&= \frac{1}{n}\left(\int_{\mathbb{R}^d} g(\boldsymbol{x})^2 L(\boldsymbol{x})^2 f_2(\boldsymbol{x})\, d\boldsymbol{x} - \mu^2\right) \\
&= \frac{1}{n}\left(\int_{\mathbb{R}^d} g(\boldsymbol{x})^2 L(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x} - \mu^2\right) \\
&= \frac{1}{n}\big(\mathbb{E}_{f_1}\big[g(\boldsymbol{X})^2 L(\boldsymbol{X})\big] - \mu^2\big)
\end{aligned}
$$

Como $\text{Var}_{f_1}\big(\hat{\mu}_n(\boldsymbol{X})\big) = \frac{1}{n}\big(\mathbb{E}_{f_1}\big[g(\boldsymbol{X})^2\big] - \mu^2\big)$, se tiene que $\text{Var}_{f_2}\big(\hat{\mu}_{\text{is},f_2,n}(\boldsymbol{X})\big) < \text{Var}_{f_1}(\hat{\mu}_n(\boldsymbol{X}))$ si y solo si $\mathbb{E}_{f_1}\big[g(\boldsymbol{X})^2 L(\boldsymbol{X})\big] < \mathbb{E}_{f_1}\big[g(\boldsymbol{X})^2\big]$.

Es posible determinar la densidad instrumental óptima: si $0 < \int_{\mathbb{R}^d} |g(\boldsymbol{x})| f_1(\boldsymbol{x})\, d\boldsymbol{x} < +\infty$, entonces la elección de f_2 que minimiza el valor de $\text{Var}_{f_2}\big(\hat{\mu}_{\text{is},f_2,n}(\boldsymbol{X})\big)$ es

$$
f_2^*(\boldsymbol{x}) = \frac{|g(\boldsymbol{x})| f_1(\boldsymbol{x})}{\int_{\mathbb{R}^d} |g(\boldsymbol{x})| f_1(\boldsymbol{x})\, d\boldsymbol{x}}
$$

y la mínima varianza es

$$
\left(\int_{\mathbb{R}^d} |g(\boldsymbol{x})| f_1(\boldsymbol{x})\, d\boldsymbol{x}\right)^2 - \left(\int_{\mathbb{R}^d} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x}\right)^2
$$

Este resultado de optimalidad es, en realidad, un resultado formal: por ejemplo, cuando g es siempre positiva o negativa se alcanzaría varianza 0, pero esta elección óptima requeriría conocer $\int_{\mathbb{R}^d} g(\boldsymbol{x}) f_1(\boldsymbol{x})\, d\boldsymbol{x}$, que es la integral que se quiere calcular.

No obstante, desde un punto de vista práctico el resultado sugiere probar con densidades instrumentales que sean lo más similares a $|g(\boldsymbol{x})| f_1(\boldsymbol{x})$ que sea posible. Es decir, las

tuplas $\boldsymbol{x}$ que deberían generarse con mayor probabilidad (las más «importantes») deberían ser aquellas con valores altos de $|g(\boldsymbol{x})|f_1(\boldsymbol{x})$. También hay que tener en cuenta que la varianza del método es menor cuanto menor sea $\mathbb{E}_{f_1}\left[g(\boldsymbol{X})^2 L(\boldsymbol{X})\right]$. Esto quiere decir que, puesto que el denominador de L es f_2, una densidad instrumental mal elegida puede dar lugar a una estimación de gran varianza, incluso infinita. Para evitarlo, los valores cercanos a cero de esta densidad instrumental deberían cancelarse con valores cercanos a cero de la función g o, en su defecto, la densidad instrumental f_2 debería acercarse a cero más lentamente que la densidad nominal f_1.