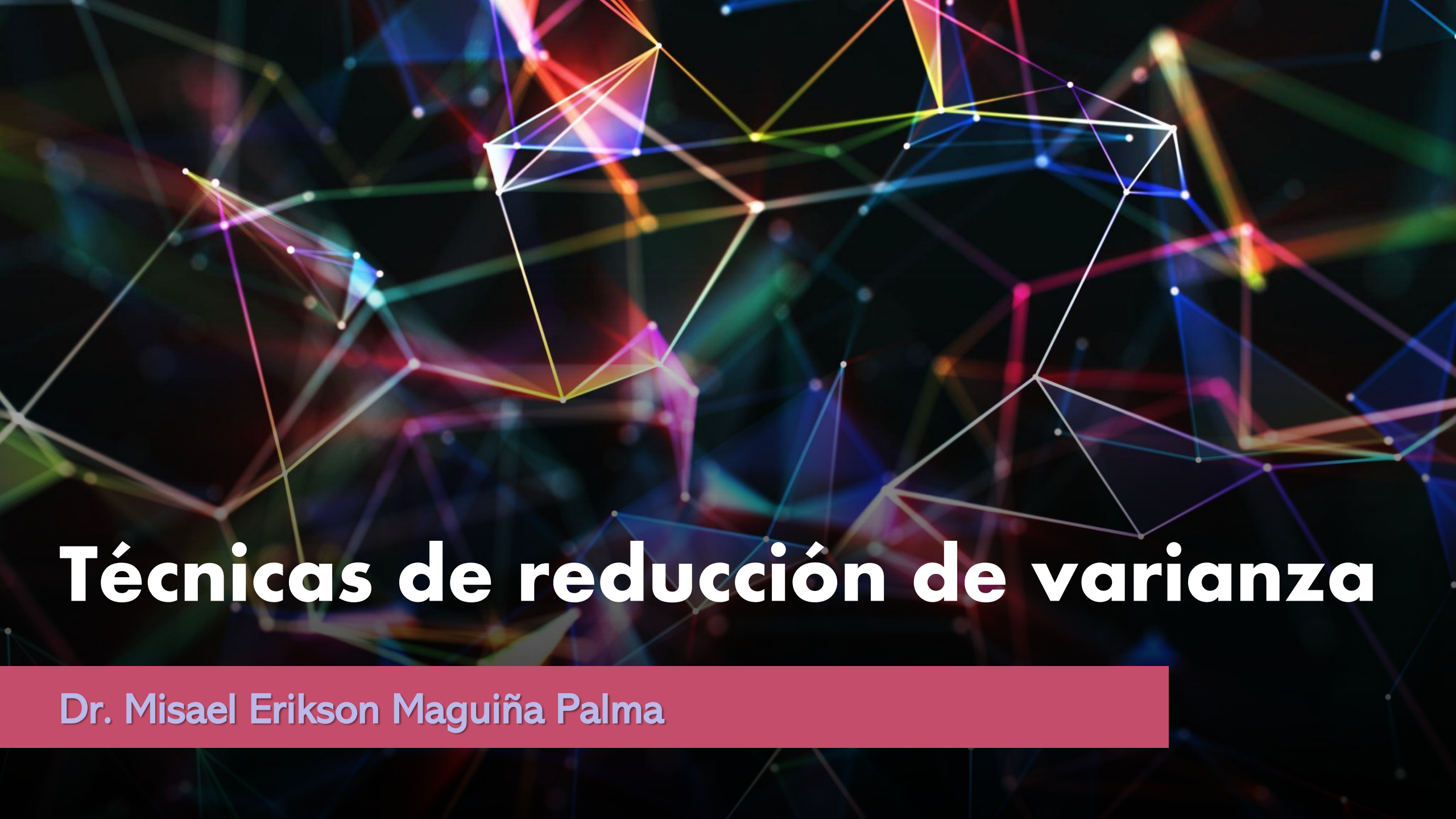


# Técnicas de reducción de varianza

Dr. Misael Erikson Maguiña Palma

# Reducción de la varianza

Éstas técnicas son aplicadas normalmente cuando se pretende ofrecer respuestas lo más precisas posibles (con menor costo computacional) y principalmente sobre cantidades medias.

- Supongamos que **estamos interesados en aproximar la media** de un estadístico mediante simulación y **no nos interesa aproximar su varianza**.

Existe un sinfín de técnicas encaminadas a reducir la varianza en un estudio de simulación (respecto a una aproximación estándar). Algunas de ellas son:

- Muestreo por importancia.
- Variables antitéticas.
- Muestreo estratificado.
- Variables de control.
- Números aleatorios comunes.

$$\mathrm{ECM} = E[(\bar{X} - \theta)^2] = \mathrm{Var}(\bar{X}) = \frac{\mathrm{Var}(X)}{n}$$

# Variables antitéticas

Supongamos que pretendemos aproximar

$$\theta = E(Z)$$

con $Var(Z) = \sigma^2$. Si generamos $n$ pares $(X_1, Y_1), ..., (X_n, Y_n)$ de $X \sim Y \sim Z$ con $Cov(X, Y) < 0$, el estimador combinado tiene menor varianza:

$$\begin{aligned} Var\left(\frac{\overline{X} + \overline{Y}}{2}\right) &= \frac{1}{4}\left(Var\left(\overline{X}\right) + Var\left(\overline{Y}\right) + 2Cov\left(\overline{X}, \overline{Y}\right)\right) \\ &= \frac{\sigma^2}{2n} + \frac{1}{2n}Cov(X, Y) \\ &= \frac{\sigma^2}{2n}\left(1 + \rho(X, Y)\right), \end{aligned}$$

que el equivalente a una muestra unidimensional independiente con el mismo número de observaciones $2n$ (con una reducción del $-100\rho(X, Y)\,\%$).

# Integración Monte Carlo

Para aproximar:

$$I = \int_0^1 h(x)\, dx$$

a partir de $x_1, x_2, \ldots, x_n$ $i.i.d.$ $\mathcal{U}(0,1)$. Podemos emplear:

$$I = E\left(\frac{h(U) + h(1-U)}{2}\right)$$

$$\approx \frac{1}{2n}\sum_{i=1}^{n}\left(h(x_i) + h(1-x_i)\right)$$

# Generación de variables antitéticas

Cuando se utiliza el método de inversión resulta sencillo obtener pares de variables con correlación negativa:

- $U \sim \mathcal{U}(0,1)$ para simular $X$.
- $1-U$ para simular la variable antitética $Y$.

En el caso general, si $X = h(U_1, \ldots, U_d)$ y $h$ es monótona puede verse (e.g. Ross, 1997) que $Y = h(1-U_1, \ldots, 1-U_d)$ está negativamente correlada con $X$.

Si $X \sim \mathcal{N}(\mu, \sigma)$ puede tomarse como variable antitética

$$Y = 2\mu - X$$

En general esto es válido para cualquier variable simétrica repecto a un parámetro $\mu$. (e.g. $X \sim \mathcal{U}(a,b)$ e $Y = a + b - X$).

# Ejercicio: Variables antitéticas en integración Monte Carlo

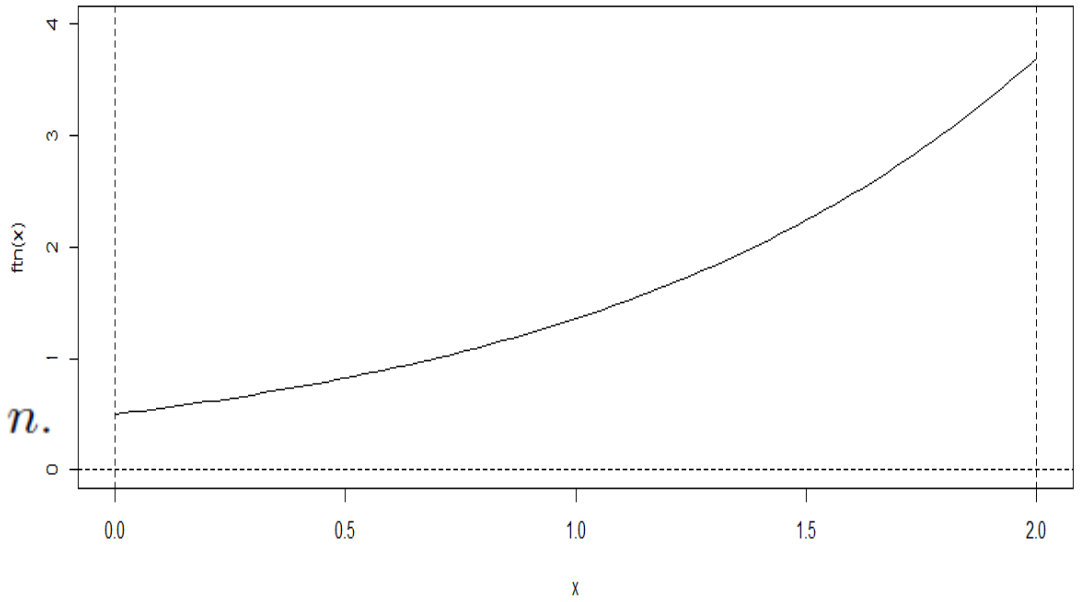
Crear una función que implemente la técnica de variables antitéticas para aproximar integrales del tipo:

$$I = \int_a^b h(x)\,dx$$

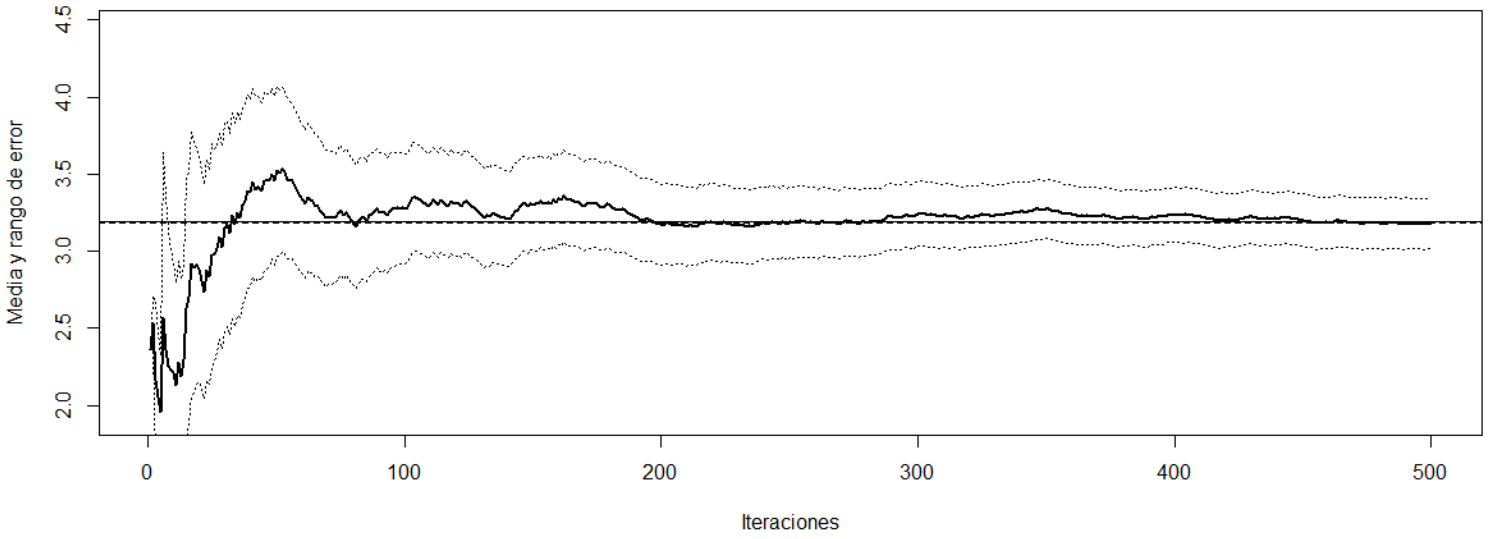
Emplearla para aproximar:

$$E\left(e^{\mathcal{U}(0,2)}\right) = \int_0^2 \frac{1}{2} e^x dx \approx 3.194$$

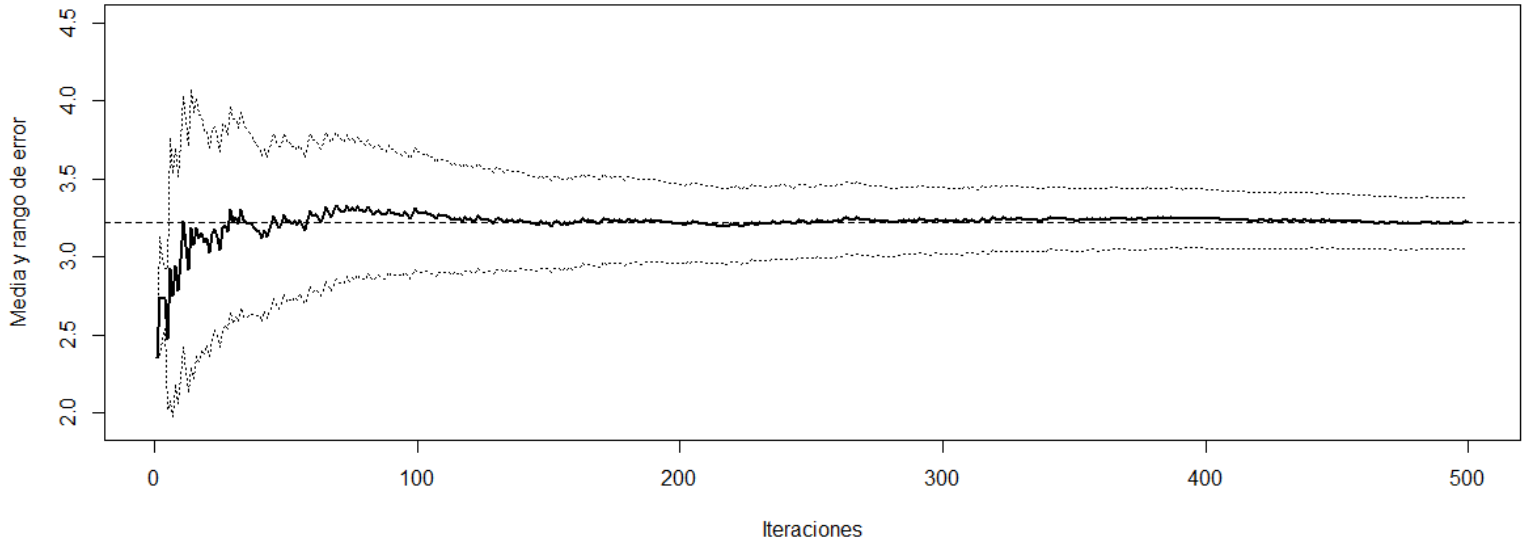
y representar gráficamente la aproximación en función de $n$.



# Aplicando Montecarlo




En este caso puede verse que la reducción teórica de la varianza es del 64.8%.



# Estratificación

Si se divide la población en estratos y se genera en cada uno un número de observaciones proporcional a su tamaño (a la probabilidad de cada uno) nos aseguramos de que se cubre el dominio de interés y se puede acelerar la convergencia.

- Por ejemplo, para generar una muestra de tamaño $n$ de una $\mathcal{U}(0,1)$, se pueden generar $l = \frac{n}{k}$ observaciones $(1 \leq k \leq n)$ de la forma:

$$U_{j_1}, \ldots, U_{j_l} \sim \mathcal{U}\left(\frac{(j-1)}{k}, \frac{j}{k}\right) \text{ para } j = 1, ..., k.$$

Si en el número de obsevaciones se tiene en cuenta la variabilidad en el estrato se puede obtener una reducción significativa de la varianza.

# Ejemplo: Muestreo estratificado de una exponencial

Supóngase el siguiente problema (absolutamente artificial pero ilustrativo para comprender esta técnica). Dada una muestra de tamaño 10 de una población con distribución:

$$x \sim exp\,(1)\ ,$$

se desea aproximar la media poblacional a partir de 10 simulaciones. Supongamos que para evitar que, por puro azar, exista alguna zona en la que la exponencial toma valores, no representada en la muestra simulada de 10 datos, se consideran tres estratos. Por ejemplo, el del 40% de valores menores, el siguiente 50% de valores (intermedios) y el 10% de valores más grandes para esta distribución.

# Ejemplo: Muestreo estratificado de una exponencial

El algoritmo de inversión (optimizado) para simular una exp(1) es:

1. Generar $U \sim U(0, 1)$.
2. Hacer $X = -\ln U$.

Dado que, en principio, simulando diez valores $U_1, U_2, \ldots, U_{10} \sim U(0,1)$, no hay nada que nos garantice que las proporciones de los estratos son las deseadas (aunque sí lo serán en media). Una forma de garantizar el que obtengamos 4, 5 y 1 valores, respectivamente, en cada uno de los tres estratos, consiste en simular:

- 4 valores de $U[0.6, 1)$ para el primer estrato,
- 5 valores de $U[0.1, 0.6)$ para el segundo y
- uno de $U[0, 0.1)$ para el tercero.

Otra forma de proceder consistiría en rechazar valores de U que caigan en uno de esos tres intervalos cuando el cupo de ese estrato esté ya lleno (lo cual no sería computacionalmente eficiente).

# Ejemplo: Muestreo estratificado de una exponencial

El algoritmo con la estratificación propuesta sería como sigue:

1. Para $i = 1, 2, \ldots, 10$:
2. Generar $Ui$:

   2a. Generar $U \sim U(0, 1)$.

   2b. Si $i \leq 4$ hacer $Ui = 0.4 \cdot U + 0.6$.

   2c. Si $4 < i \leq 9$ hacer $Ui = 0.5 \cdot U + 0.1$.

   2d. Si $i = 10$ hacer $Ui = 0.1 \cdot U$.
3. Devolver $Xi = -\ln Ui$.

No es difícil probar que:

- $Var(Xi) = 0.0214644$ si $i = 1, 2, 3, 4$,
- $Var(Xi) = 0.229504$ si $i = 5, 6, 7, 8, 9$ y
- $Var(X10) = 1$.

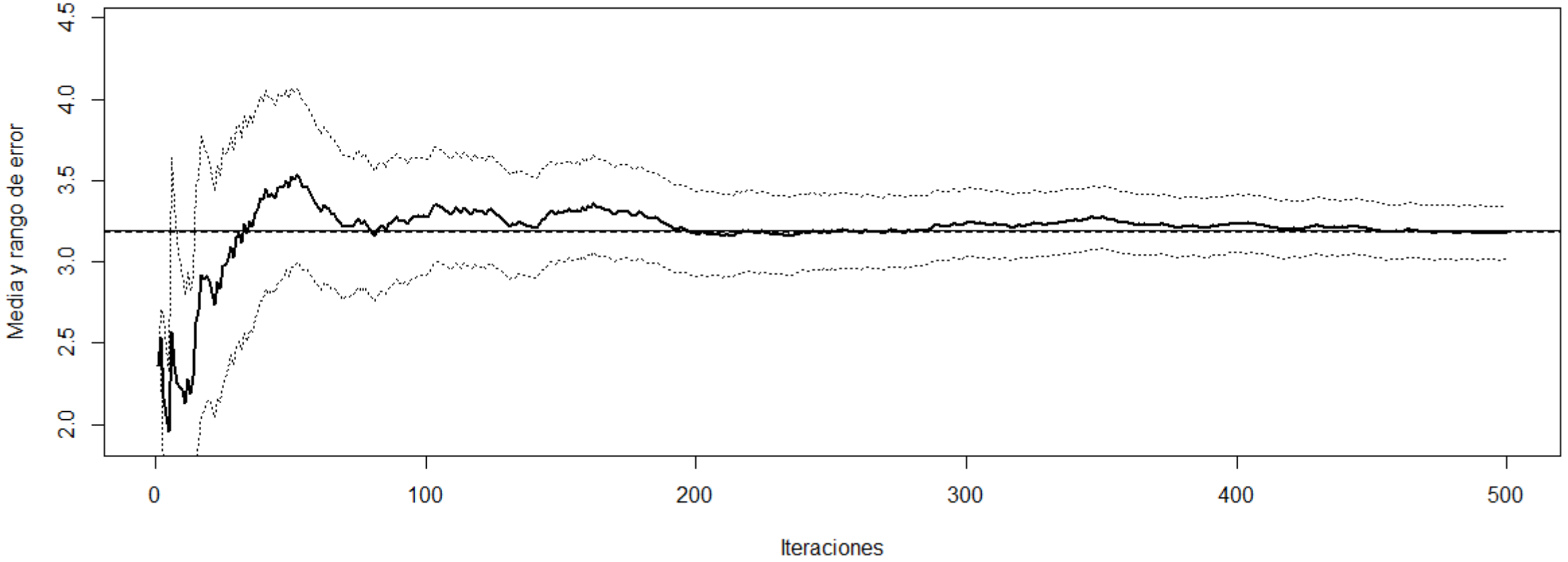
Como consecuencia:

$$Var(\overline{X}) = \frac{1}{10^2} \sum_{i=1}^{10} Var(X_i) = 0.022338$$

que es bastante menor que 1 (la varianza en el caso de muestreo aleatorio simple no estratificado).

# Ejercicio: Integración Monte Carlo con estratificación

Aproximar la integral anterior empleando la técnica de estratificación, considerando k subintervalos regularmente espaciados en el intervalo [0, 2]. ¿Como varía la reducción en la varianza dependiendo del valor de k?



# Variables de control

En este caso se trata de sacar partido tanto a una covarianza positiva como negativa. La idea básica es emplear una variable $Y$, con media conocida $\mu_Y$, para controlar la variable $X$ (con media desconocida), de forma que ambas variables estén "suficientemente" correlacionadas. La versión "controlada" de $X$ será:

$$X^* = X + \alpha\left(Y - \mu_Y\right)$$

con $E(X^*) = E(X) = \theta$. Puede verse que $Var(X^*) = Var(X) + \alpha^2 Var(Y) + 2\alpha Cov(X,Y)$ es mínima para:

$$\alpha^* = -\frac{Cov(X,Y)}{Var(Y)},$$

con $Var(X^*) = Var(X)\left(1 - \rho^2\left(X,Y\right)\right)$ (lo que supone una reducción del $100\rho^2\left(X,Y\right)\%$).

# Variables de control

En la práctica normalmente $\alpha^*$ no es conocida. Para estimarlo se puede realizar ajuste lineal de $X$ sobre $Y$ (a partir de los datos simulados $X_i$ e $Y_i$, $1 \leq i \leq n$):

- Si $\hat{x} = \hat{\beta}_0 + \hat{\beta}_1 y$ es la recta ajustada, con $\hat{\beta}_1 = \dfrac{S_{XY}}{S_Y^2}$ y $\hat{\beta}_0 = \overline{X} - \hat{\beta}_1 \overline{Y}$, la estimación sería:

$$\hat{\alpha}^* = -\hat{\beta}_1$$

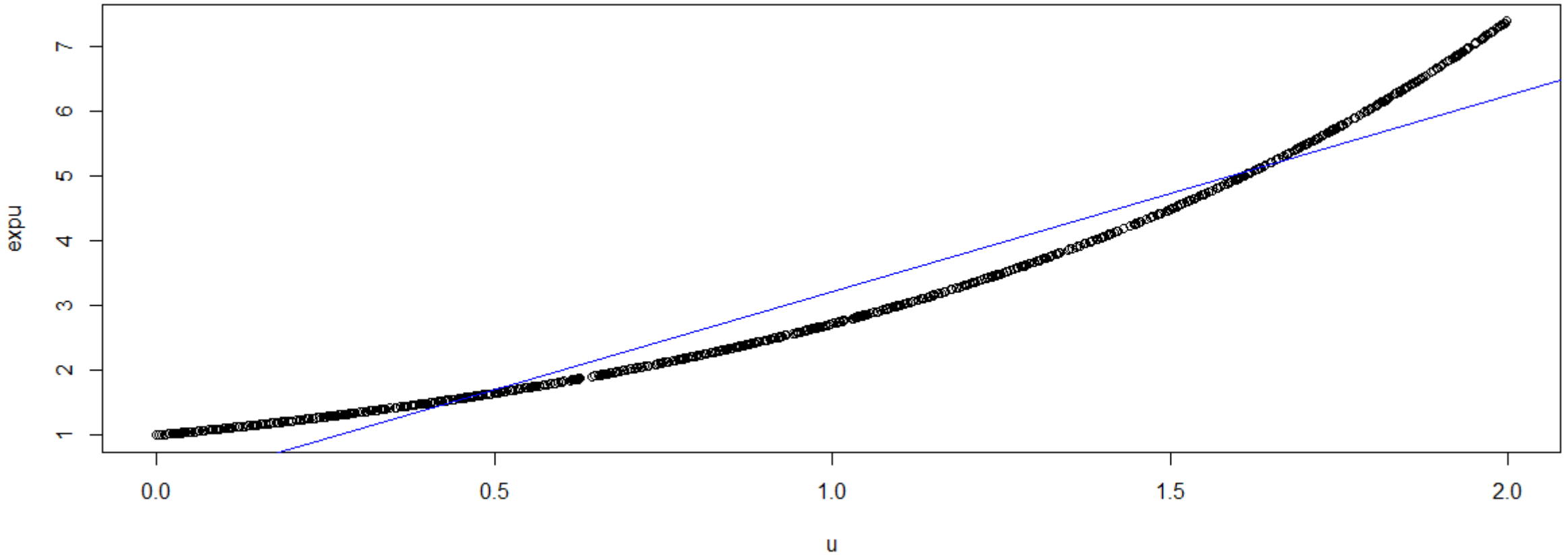
- Adicionalmente, para aproximar $\theta$:

$$\begin{aligned} \hat{\theta} = \overline{X}^* &= \overline{X} - \hat{\beta}_1 \left(\overline{Y} - \mu_Y\right) \\ &= \hat{\beta}_0 + \hat{\beta}_1 \mu_Y \end{aligned}$$

- Si $\mu_Y = 0 \Rightarrow \hat{\theta} = \overline{X}^* = \hat{\beta}_0$.

# Ejercicio: Integración Monte Carlo con variables de control.

Aproximar la integral anterior empleando la variable U~U(0, 2) para controlar la variable $e^U$. Se trata de calcular la media de exp(U(a, b)):



# Números aleatorios comunes

Se trataría de una técnica básica del diseño de experimentos: realizar comparaciones homogéneas (bloquear). Por ejemplo cuando se diseña un experimento para la comparación de la media de dos variables, se pueden emplear las denominadas muestras apareadas, en lugar de muestras independientes.

Supóngamos que estamos interesados en las diferencias entre dos estrategias (e.g. dos estimadores):

$$E(X) - E(Y) = E(X - Y).$$

Para ello se generan dos secuencias $X_1, X_2, \ldots, X_n$, e $Y_1, Y_2, \ldots, Y_n$ y se calcula:

$$\overline{X} - \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)$$

- Si las secuencias se generan de modo independiente:

$$Var\left(\overline{X} - \overline{Y}\right) = \frac{1}{n} \left(Var(X) + Var(Y)\right)$$

# Números aleatorios comunes

- Si se generar las secuencias empleando **la misma semilla**, los datos son dependientes:

$$Cov\left(X_i, Y_i\right) > 0$$

y tendríamos que:

$$\begin{aligned} Var\left(\overline{X} - \overline{Y}\right) &= \frac{1}{n^2}\sum_{i=1}^{N} Var\left(X_i - Y_i\right) = \frac{1}{n} Var\left(X_i - Y_i\right) \\ &= \frac{1}{n}\left(Var\left(X_i\right) + Var\left(Y_i\right) - 2Cov\left(X_i, Y_i\right)\right) \\ &\leq \frac{1}{n}\left(Var\left(X_i\right) + Var\left(Y_i\right)\right) \end{aligned}$$

En el capítulo de aplicaciones de la simulación se empleó esta técnica para comparar distribuciones de estimadores...

# Ejercicios fin de práctica

Aproximar mediante integración Monte Carlo (clásica) la media de una distribución exponencial de parámetro 1/2:

$$I = \int_0^\infty \frac{x}{2} e^{-\frac{x}{2}} dx$$

y representar gráficamente la aproximación en función de $n$. Comparar los resultados con los obtenidos empleando variables antitéticas, ¿se produce una reducción en la varianza?
*Nota:* . Puede ser recomendable emplear el método de inversión para generar las muestras (antitéticas) de la exponencial.