Modelado y Simulación (505103009)
Tema 6. Estimación paramétrica en simulación por eventos discretos

Javier Vales Alonso
Grado en Ingeniería Telemática

2020

Universidad Politécnica de Cartagena

Tabla de contenido

¿Cómo estudiar esta unidad?

1. Haga una primera lectura de la unidad. Concéntrese en ver las ideas generales y el uso del teorema central del límite en el contexto de simulación.
2. Haga una revisión más a fondo de los desarrollos matemáticos.
3. Implemente en MATLAB los métodos solicitados en la práctica.
4. En caso de dudas, puede consultar los libros de referencia, o contactar con el profesor.

Introducción

Objetivos

Los objetivos de este tema son:

- Describiremos **qué es un estimador**, y algunas de sus propiedades más importantes.
- Mostraremos cómo obtener **intervalos de confianza** para la estimación de medias a partir del **teorema central del límite**.
- Describiremos las peculiaridades que se presentan en el contexto de simulación.
- Explicaremos cómo implementar mecanismos para detener automáticamente la simulación al llegar a un nivel de confianza suficiente.

Introducción

Dado un sistema natural/artificial, éste puede estar caracterizado por unos **parámetros**, cuyos valores resultan de interés. Por ejemplo:

1. La altura media y su varianza en un conjunto de individuos.
2. El porcentaje de extranjeros en un conjunto de individuos.
3. El número de caras de un dado.
4. El tiempo esperado de respuesta $E\{T\}$ en un sistema G/G/k.

Si piensa con detenimiento **en estos ejemplos**, verá, que en todos los casos, **los parámetros desconocidos son de naturaleza determinista**, es decir, no están sujetos al azar.

Introducción (II)

Habitualmente no se puede calcular/observar directamente esos parámetros y la única alternativa es realizar un **muestreo** experimental:

- Escoger un conjunto pequeño del total de la población para los casos 1 y 2 y ver la altura o la nacionalidad de los individuos seleccionados.
- Realizar varias tiradas para el caso del dado.
- Ver el tiempo de respuesta de algunas tareas escogidas al azar.

El valor de cada muestra es **aleatorio**, ya que el azar determina que individuos/tareas escogemos o el valor de las tiradas de los dados.

Introducción (III)

A partir de las muestras, **aproximamos**:

1. La altura media o la varianza de una población, por la media o la varianza de la muestra.
2. El porcentaje de extranjeros de una población, por el porcentaje de extranjeros en la muestra.
3. El número de caras del dado, por la muestra de mayor valor.
4. El tiempo esperado de respuesta, por la media del tiempo de respuesta de las muestras.

La aproximación **es aleatoria** (depende de las muestras).
Intuitivamente, **a más muestras, mejor aproximación**.

Estimadores

Estimadores

Dado una muestra $\{X_1, X_2, \ldots, X_n\}$ de un proceso aleatorio, un **estimador** es una función que depende de dichas muestras, y de cuyo valor se intenta inferir algún parámetro o parámetros de la distribución del proceso aleatorio. Al ser las muestras variables aleatorias, **el estimador también es una variable aleatoria**.

Para un parámetro θ se denota una función estimadora del mismo como $\widehat{\theta}_n$ o simplemente $\widehat{\theta}$. Algunos estimadores habituales tienen una notación particular, como la **media muestral** $(\overline{X}_n)$ o la **cuasi-varianza muestral** (S^2_{n-1}).

Estimadores (II)

Propiedades fundamentales de los estimadores:

- **Insesgado**. Si $E\{\widehat{\theta}_n\} = \theta$
- **Consistente**. Si $\widehat{\theta}_n \xrightarrow{p} \theta$ cuando $n \to \infty$
- **Suficiente**. Si $f(x_1, \ldots, x_n|\theta) = u(x_1, \ldots, x_n)v(\widehat{\theta}_n, \theta)$
- **Robusto**.

Estimadores (III)

Ejemplo 1: Sea X una variable aleatoria con media $E\{X\} = \mu_X$, dados los siguientes estimadores de la media, indicar cuáles son insesgados y cuáles son consistentes.

1. $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (media muestral)
2. X_1
3. $X_1/2 + X_n/2$
4. $(\prod_{i=1}^{n} X_i)^{\frac{1}{n}}$
5. e^{π}

Estimadores (IV)

Ley (débil) de los grandes números (Kinchin). Dada una secuencia de muestras independientes idénticamente distribuidas (iid) $X_1, \ldots, X_n$, con media μ_X, se verifica que:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu_X \text{ cuando } n \to \infty$$

Estimadores (V)

Ejemplo 2: Otro estimador razonable para μ_X es el valor que minimice la suma de las diferencias observadas $|X_i - \widehat{\mu_X}|$, o, equivalentemente, de sus cuadrados $(X_i - \widehat{\mu_X})^2$. Es decir,

$$\widehat{\mu_X} = \arg\min_{\widehat{\mu_X}} \sum_{i=1}^{n} (X_i - \widehat{\mu_x})^2$$

Demuestre que este estimador también es la media muestral.

Estimadores (VI)

Ejemplo 3: Sea X una variable aleatoria discreta que toma valores uniformemente en el conjunto $\{1, \ldots, L\}$. Dados los siguientes estimadores de L, indicar cuáles son insesgados y cuáles son consistentes.

1. $\max\{X_1, \ldots, X_n\}$
2. X_1
3. $\frac{2}{n}\sum_{i=1}^{n} X_i$
4. e^{π}

Estimadores (VII)

Ejemplo 4: Sea X una variable aleatoria cuya varianza $E\{(X-\mu_X)^2\}$ denotaremos por σ_X^2. Demuestre que:

$$S_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$$

es un estimador insesgado de σ_X^2.

Intervalo de confianza de la media muestral

Intervalo de confianza de la media muestral

Dada una muestra aleatoria $\{X_1, \ldots, X_n\}$ cuyas componentes son iid, queremos calcular un intervalo de confianza $[\overline{X}_n - t, \overline{X}_n + t)$ para la media estadística (poblacional) μ_X con un nivel de significación α.

Es decir, encontrar el intervalo que verifique:

$$p(\overline{X}_n - t < \mu_X \leq \overline{X}_n + t) = 1 - \alpha$$

Llamaremos **calidad** al inverso del nivel de significación, es decir, a $1 - \alpha$, y **tolerancia** a la semi-longitud del intervalo, es decir, a t. A menudo, se trabaja con la **tolerancia relativa** $t^{\%} = \frac{t}{\overline{X}_n}$ para evitar efectos de escala.

Intervalo de confianza de la media muestral (II)

Teorema central del límite (TCL). Dado un conjunto $\{X_1, \ldots, X_n\}$ de muestras iid con media μ_X y varianza $\sigma_X^2 < \infty$, se verifica:

$$\overline{X}_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right) \text{ cuando } n \to \infty \tag{1}$$

Y, tipificando el resultado:

$$\frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{Z} \text{ cuando } n \to \infty \tag{2}$$

siendo $\mathcal{Z} = \mathcal{N}(0, 1)$.

Intervalo de confianza de la media muestral (III)

Sea Z una variable aleatoria con distribución $\mathcal{Z}$. Denotamos $z_{\alpha/2}$ como el punto que verifica:

$$p(Z > z_{\alpha/2}) = \frac{\alpha}{2}$$

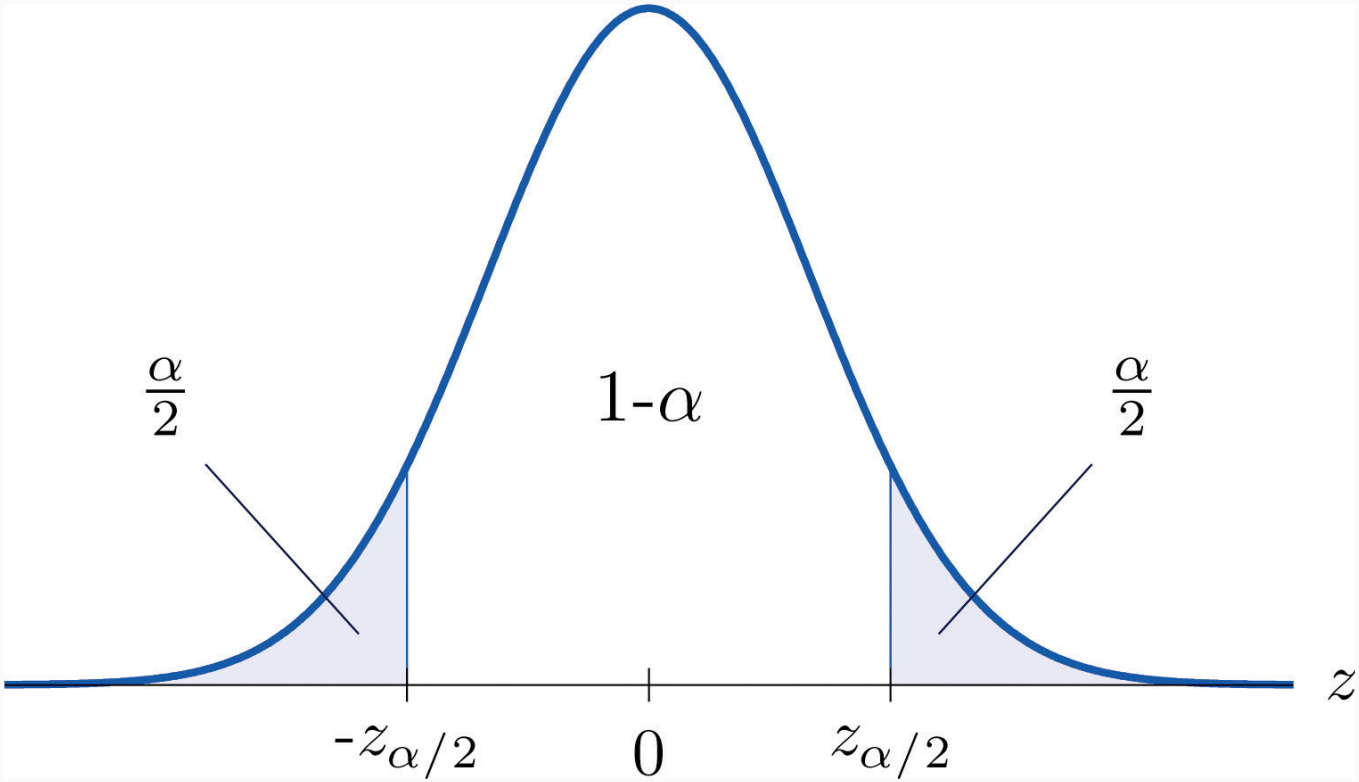
Y puesto que $\mathcal{Z}$ es simétrica se verifica también:

$$p(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}$$

Por tanto,

$$p(-z_{\alpha/2} < Z \leq z_{\alpha/2}) = 1 - \alpha$$

Intervalo de confianza de la media muestral (IV)



Función de densidad de la Gaussiana con distribución $\mathcal{Z}=\mathcal{N}(0,1)$.
Intervalo de mínima longitud con nivel de significación α: $(-z_{\alpha/2}, z_{\alpha/2}]$.

Intervalo de confianza de la media muestral (V)

Entonces, a partir de la expresión (2):

$$p(-z_{\alpha/2} < \frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$$

Al ser $\sigma_X \geq 0$ y multiplicar ambos lados de las inecuaciones por $\sigma_X/\sqrt{n}$, obtenemos:

$$p(-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} < \overline{X}_n - \mu_X \leq z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}) = 1 - \alpha$$

Restando $\overline{X}_n$ en cada lado de las inecuaciones:

$$p(-z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} - \overline{X}_n < -\mu_X \leq z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} - \overline{X}_n) = 1 - \alpha$$

Intervalo de confianza de la media muestral (VI)

Para centrar el intervalo en μ_X hay que invertir el signo de las inecuaciones (se recuerda que si se invierte $-a \leq b$ obtenemos $a \geq -b$). Por tanto, la inecuación a la izquierda pasa a la derecha y viceversa, y se obtiene:

$$p\left(\overline{X}_n - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}} \leq \mu_X < \overline{X}_n + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right) = 1 - \alpha$$

Concluimos entonces que $\mu_X \in \left[\overline{X}_n - z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{\sigma_X}{\sqrt{n}}\right)$ con nivel de significación α.

Intervalo de confianza de la media muestral (VII)

En la práctica, σ_X^2 es desconocida, por lo que se debe sustituir por su estimador insesgado S_{n-1}^2, resultando el intervalo:

$$\left[\overline{X}_n - z_{\alpha/2}\sqrt{\frac{S_{n-1}^2}{n}}, \overline{X}_n + z_{\alpha/2}\sqrt{\frac{S_{n-1}^2}{n}}\right) \tag{3}$$

Esto presenta un problema técnico, ya que al sustituir σ_X^2 por su estimador, la distribución de

$$\frac{\overline{X}_n - \mu_X}{\sqrt{S_{n-1}^2/n}}$$

ya no converge en distribución a $\mathcal{Z}$, sino a una *t*-de *Student* de $n-1$ grados de libertad.

Intervalo de confianza de la media muestral (VIII)

Si el número de muestras es pequeño, deberán usarse los valores $t_{\alpha/2,n-1}$ en vez de los $z_{\alpha/2}$ al calcular el intervalo de confianza.

No obstante, al aumentar n la t-de *Student* tiende en distribución a una $\mathcal{Z}$. En nuestro contexto podemos suponer cierta esta hipótesis, ya que un simulador puede generar un número tan grande de muestras como sea preciso con bajo coste, y usar el intervalo dado por la expresión (3).

Intervalo de confianza de la media muestral (IX)

Algoritmo 1: Cálculo de intervalo de confianza dada calidad

1 **Entradas**: $1-\alpha$, n, $\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n} X_i^2$

2 **Salida**: $[\overline{X}_n - t, \overline{X}_n + t)$

3 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

4 $S_{n-1}^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right]$

5 **Obtener** $z_{\alpha/2}$ mediante inspección en tabla $\mathcal{Z}$

6 $t = z_{\alpha/2}\sqrt{\frac{S_{n-1}^2}{n}}$

7 **Devolver** $[\overline{X}_n - t, \overline{X}_n + t)$

Intervalo de confianza de la media muestral (X)

Ejemplo 5: Calcule el intervalo de confianza para un nivel de significación del 5 % para un proceso del que se han obtenido los siguientes estadísticos:

$$n = 100, \quad \sum_{i=1}^{n} X_i = 1000, \quad \sum_{i=1}^{n} X_i^2 = 12475$$

Intervalo de confianza de la media muestral (XI)

Algoritmo 2: Cálculo de calidad dada la tolerancia relativa

1 **Entradas**: $t^{\%}$, n, $\sum_{i=1}^{n} X_i$, $\sum_{i=1}^{n} X_i^2$

2 **Salida**: $1 - \alpha$

3 $\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

4 $S_{n-1}^2 = \frac{1}{n-1} \left[\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2 \right]$

5 $t = t^{\%} \overline{X}_n$

6 $z_{\alpha/2} = \frac{t}{\sqrt{\frac{S_{n-1}^2}{n}}}$

7 **Obtener** $\alpha/2$ mediante inspección en tabla $\mathcal{Z}$

8 **Devolver** $1 - \alpha$

Intervalo de confianza de la media muestral (XII)

Ejemplo 6: Calcule la calidad para un intervalo centrado en la media con tolerancia relativa del 10 % para un proceso del que se han obtenido los siguientes estadísticos:

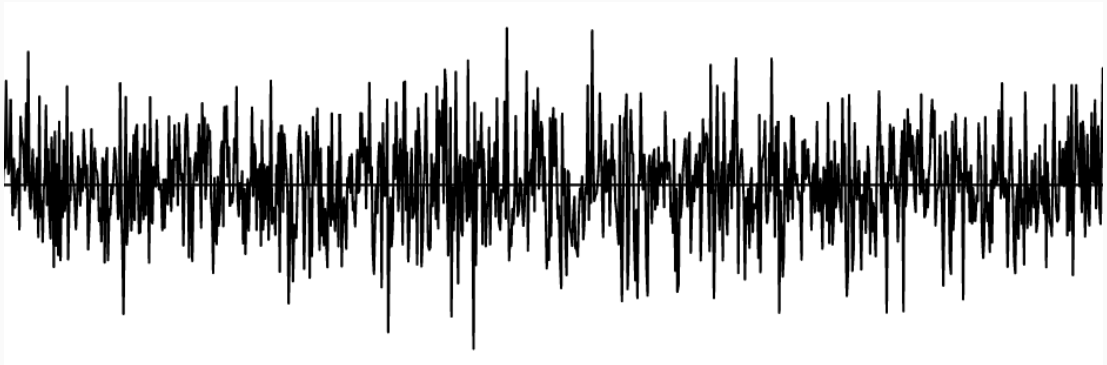
$$n = 100, \quad \sum_{i=1}^{n} X_i = 1000, \quad \sum_{i=1}^{n} X_i^2 = 12475$$

Estimación paramétrica en procesos estocásticos

Estimación paramétrica en procesos estocásticos

En el contexto de simulación por eventos discretos, **una realización del simulador es un proceso estocástico de tiempo discreto** X, del que se extrae una secuencia de muestras $\{X_1, \ldots, X_n\}$.

Para poder aplicar el **TCL las muestras deben ser iid y de energía (varianza) finita**, i.e., básicamente, una señal de ruido blanco limitada en ancho de banda:



Estimación paramétrica en procesos estocásticos (II)

En simulación **las muestras no son independientes entre sí y no podemos asumir que se verifique el TCL**. Por ejemplo, en un sistema de colas G/G/k el tiempo de salida de la tarea i-ésima estará fuertemente relacionado con las de las tareas cercanas.

A continuación estudiaremos cómo podemos adaptar el método de creación de intervalos de confianza en este contexto.

Estimación paramétrica en procesos estocásticos (III)

En general, asumiremos una serie de restricciones. En primer lugar, supondremos que los sistemas bajo estudio son **estacionarios**. Informalmente esto significa que no hay instantes de observación privilegiados.

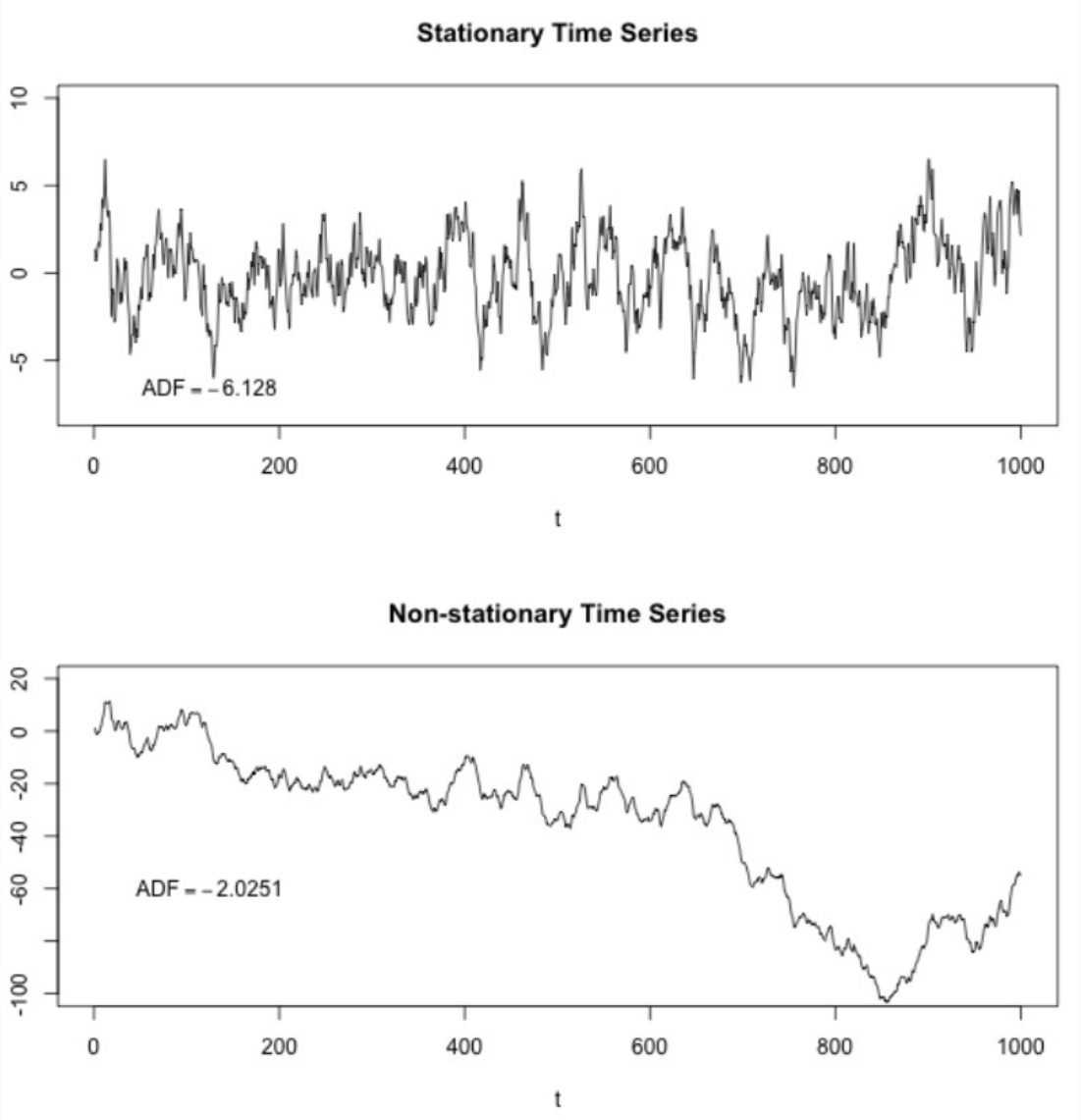
Analíticamente:

$$F_X(x_1, \ldots, x_n) = F_X(x_{1+i}, \ldots, x_{n+i})$$

para todo $n>0$ y todo $i>0$.

Es decir, las **distribuciones conjuntas de cualquier orden son invariantes en el tiempo**, y por tanto, también los serán sus momentos (esperanzas, varianzas, etc.).

Estimación paramétrica en procesos estocásticos (IV)



Estimación paramétrica en procesos estocásticos (V)

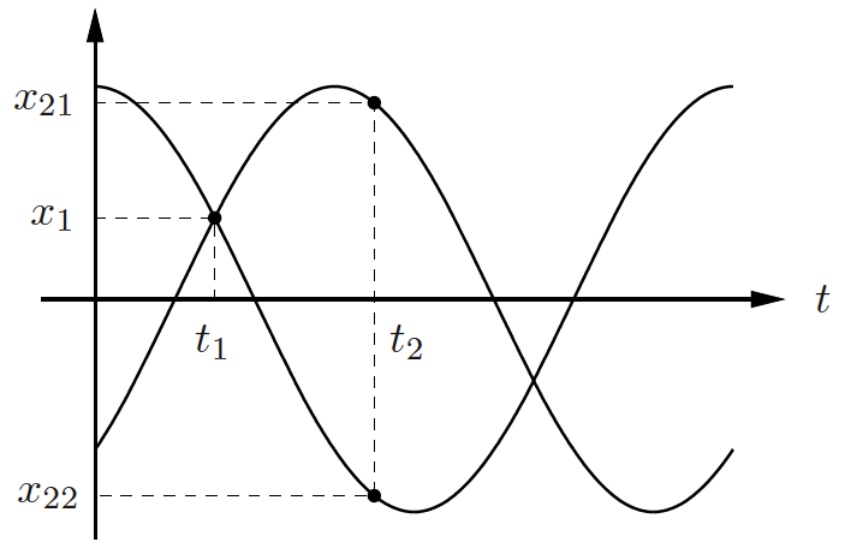
Ejemplo 7: Sea el proceso estocástico de tiempo continuo:

$$X(t) = \begin{cases} +\sin(t) & \text{con probabilidad } \frac{1}{4} \\ -\sin(t) & \text{con probabilidad } \frac{1}{4} \\ +\cos(t) & \text{con probabilidad } \frac{1}{4} \\ -\cos(t) & \text{con probabilidad } \frac{1}{4} \end{cases}$$

demostrar que no es estacionario.

Estimación paramétrica en procesos estocásticos (VI)

Ejemplo 8: Sea $X(t) = cos(\omega t + \Phi)$ con $\Phi \sim \mathcal{U}(0, 2\pi)$. Este proceso estocástico de tiempo continuo es estacionario. La figura siguiente muestra dos realizaciones del mismo:



Demostrar que la media $E\{X(t)\}$ es independiente del tiempo t.

Estimación paramétrica en procesos estocásticos (VII)

En segundo lugar, asumiremos que el sistema estacionario es **ergódico en media**. Esto es, que se cumple que la media estadística se coincide con la media temporal de la realización:

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{L^2} \mu_X = E\{X_i\}$$

Esta condición nos indica que es posible inferir la media estadística (μ_X) de la media temporal ($\overline{X}_n$) calculada a lo largo de **una** realización del proceso. **No todos los procesos estacionarios son ergódicos en media**, como veremos en el siguiente ejemplo.

Estimación paramétrica en procesos estocásticos (VIII)

Ejemplo 9:

Sea $U \sim \mathcal{U}(0,1)$. Se define el proceso estocástico $X_n = U$ para todo n. Demostrar que es estacionario pero no ergódico en media.

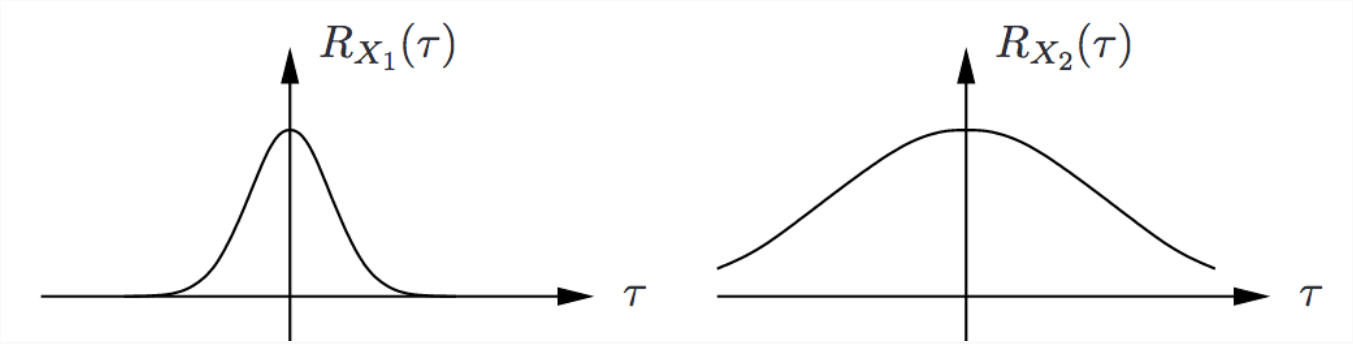
Estimación paramétrica en procesos estocásticos (IX)

Entonces, si el sistema estudiado verifica que es **estacionario y ergódico en media**, las muestras pueden considerarse identicamente distribuidas (ya que es estacionario) y la media temporal de una realización converge a la estadística.

Además, si consideramos muestras **suficientemente** alejadas en el tiempo, éstas pueden considerarse independientes, y estamos en condiciones de aplicar el TCL para el cálculo de intervalos de confianza, según lo indicado en la sección anterior.

Estimación paramétrica en procesos estocásticos (X)

La **autocorrelación de un proceso estocástico** de tiempo discreto se define como $R(i,j) = E\{X_i X_j\}$ y para un proceso estacionario se verifica que $R(i,j)$ sólo depende de $\tau = j - i$ al ser la distribución conjunta de X_i, X_j invariante en el tiempo.



La autocorrelación **mide el parecido del proceso tras un tiempo** τ. Puede existir correlación (dependencia) en diferentes escalas de tiempo (e.g., correlación corto plazo o a largo plazo).

Estimación paramétrica en procesos estocásticos (XI)

Existen versiones más relajadas de las condiciones anteriores. Para procesos son **estacionarios en sentido amplio**:

1. $EX_i)=\mu_X$ (independiente de i). Es decir, no hay instantes de observación privilegiados para observar la media estadística (todos poseen la misma media).
2. La autocorrelación $R_X(i,j) = E\{X_i, X_j\}$ sólo depende de la diferencia $j - i$.
3. $E\{(X_i)^2\} < \infty\}$ para todo i (varianza finita).

las distribuciones X_i no pueden considerarse idénticamente distribuidas, pero bajo ciertas condiciones técnicas (condición de Lindeberg) se **verifica el TCL sin ese requisito**.

Estimación paramétrica en procesos estocásticos (XII)

Ejemplo 10: Demostrar que el proceso estocástico definido en el ejemplo 7:

$$X(t) = \begin{cases} +\sin(t) & \text{con probabilidad } \frac{1}{4} \\ -\sin(t) & \text{con probabilidad } \frac{1}{4} \\ +\cos(t) & \text{con probabilidad } \frac{1}{4} \\ -\cos(t) & \text{con probabilidad } \frac{1}{4} \end{cases}$$

sí es estacionario en sentido amplio.

Notas de implementación

Notas de implementación

El objetivo fundamental desde el punto de vista de simulación es ofrecer **garantías sobre la bondad del resultado**.

Es decir, asegurar con una calidad alta que el resultado está en un intervalo de confianza suficientemente preciso. En virtud del TCL, para que el cálculo de la calidad sea correcto las muestras deben ser iid. Ello no siempre será posible.

Distinguiremos tres situaciones:

- **Sistema ergódico con correlación a corto plazo (CCP)**.
- **Sistema estacionario no ergódico o ergódico con correlación a largo plazo (CLP)**.
- **Sistema no estacionario**.

Sistema ergódico CCP

Los parámetros a buscar son medias estadísticas de interés (e.g., tiempo medio de respuesta en cola G/G/k, ratio de verificación de una condición, etc.).

En este caso, se hará **una única realización temporal** (en virtud de la ergocidad). Esta realización debe ejecutarse hasta que los intervalos de confianza para **todos los parámetros** de interés **alcancen una calidad mínima**.

El **cálculo de la calidad es costoso computacionalmente**, por lo que no debe repetirse hasta no recoger un número nuevo de muestras que sea significativo (e.g., calculando la calidad cada TEST muestras).

Sistema ergódico CCP (II)

Si el sistema posee un **régimen transitorio, las muestras recogidas en ese periodo deben descartarse**. Es decir, debemos garantizar que sólo se recogen muestras en régimen estacionario. Para ello, se eliminarán las H primeras muestras recogidas de la realización.

Asimismo, debe eliminarse la correlación entre muestras (para garantizar independencia), para ello se aplicará un **algoritmo de filtrado por bloques**: las muestras se recogerán en bloques de tamaño D y para cada bloque se calculará una única muestra que hará de representante de todo el bloque (e.g., la media del bloque, la primera muestra del bloque, la última muestra del bloque, etc.).

Sistema ergódico CCP (III)

Algoritmo 3: Simulación de procesos ergódicos CCP

Entradas: $(1-\alpha)_{min}$, $t^{\%}$, H, D, TEST, configuración del sistema
Salida: $[\overline{X}_n - t, \overline{X}_n + t)$
Inicialización (eventos de arranque, $n_t = 0$, $n_b = 0$)
repeat
    Dar paso de simulación;
    if *nueva muestra* **then**
        $n_t{=}n_t{+}1$;
        if $n_t > H$ **then**
            Añadir muestra al bloque y $n_b{=}n_b{+}1$;
            if $n_b == D$ **then**
                Calcular x representante del bloque y **Resetear** bloque ($n_b{=}0$);
                $n{=}n{+}1$, $\Sigma_X{=}\Sigma_X{+}x$, $\Sigma_{X^2}{=}\Sigma_{X^2}{+}x^2$;
                if *(n mod TEST == 0)* **then**
                    Calcular $(1-\alpha)$ con Alg. 2 con entradas $t^{\%}, n, \Sigma_X, \Sigma_{X^2}$
                end
            end
        end
    end
until $(1-\alpha) > (1-\alpha)_{min}$;

Sistema estacionario no ergódico o ergódico CLP

En este caso las muestras obtenidas a lo largo de una sola realización no son válidas para obtener la media estadística de interés, bien por ser un proceso no ergódico, bien por no poder seleccionar muestras independientes (ergódico CLP).

La única posibilidad es **hacer realizaciones independientes** (ver gestión de la semilla para este caso en Tema 4) y obtener **una muestra por realización**. Una vez **pasado el periodo transitorio, el instante de observación puede elegirse arbitrariamente** (en virtud de la estacionariedad).

Las muestras son iid y el intervalo de confianza se calcula con el Alg. 2. El **problema de este método es la gran cantidad de tiempo de computación que puede requerir**.

Sistema no estacionario

En este caso **los procesos son siempre transitorios**. El uso de simulación es **auxiliar**. No se medirán parámetros medios (en sentido estadístico), porque en este caso ya no existen, sino efectos transitorios. Por ejemplo, el tiempo promedio desde el arranque hasta que una cola G/G/k **inestable** tenga una ocupación determinada, o la probabilidad de que la cola G/G/k consiga volver a una ocupación al menos una vez tras llegar a ella por primera vez. Deberá ejecutarse **una realización independiente** para obtener cada muestra.

Como en el caso anterior, las muestras son iid y el intervalo de confianza se calcula con el Alg. 2.