Laboratorio de Telemática III

ETS Ingeniería de Telecomunicación de Barcelona

Práctica 3: Intervalo de confianza y régimen transitorio

Israel Martín, Alfonso Rojas, Francisco Barceló
Departamento de Ingeniería Telemática
Universidad Politécnica de Cataluña

Índice

1 Objetivos

1.1 Generales

El objetivo principal de la práctica es cuantificar los errores típicos aparecidos en la etapa de simulación. Para ello se validará un sistema M/M/N, cuya solución analítica es conocida, a partir del modelo G/G/N/M implementado en la práctica 1. De esta forma, el alumno podrá detectar las principales desviaciones entre los resultados teóricos y los producidos a través de simulación, así como proceder a su cuantificación.

1.2 Específicos

Los objetivos específicos planteados en la presente práctica son los siguientes:

- Validación de un sistema M/M/N a partir del modelo G/G/N/M
- Analizar los errores producidos por el régimen transitorio y la descarga de un sistema
- Analizar los errores de simulación mediante la validación de un sistema conocido. Cuantificar los mismos empleando intervalos de confianza.

2 La práctica

2.1 Régimen transitorio

Toda simulación parte de unas condiciones iniciales especificadas por el usuario. Normalmente, los modelos a simular parten de condiciones nulas. Por ejemplo, un sistema M/M/N partiría de una situación en la que ningún servidor está ocupado y no hay elementos en la cola.

De esta forma, al iniciar la simulación, el modelo alcanzará las condiciones de operación (régimen permanente) al cabo de un cierto tiempo. Ese espacio de tiempo hasta alcanzar el régimen permanente se conoce como régimen transitorio y es inherente a toda simulación. Las muestras pertenecientes al régimen transitorio distorsionan los resultados y por tanto deben ser eliminadas de todo estudio.

2.2 Intervalo de confianza sobre el valor medio

El cálculo del primer momento o valor medio de una variable aleatoria *X* implica normalmente el uso del estimador promedio

$$m_1 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{1}$$

De hecho la Ley de los Grandes Números nos dice que m_1 tiende al valor medio real μ de la variable aleatoria cuando *n* tiende a infinito. Fácilmente puede demostrarse que $E(m_1)=E(x_i)=\mu$.

Por otro lado sabemos que si los procesos que generaron las muestras x_i son independientes y además están igualmente distribuidas puede aplicarse el Teorema Central del Límite. En este caso m_1 se aproxima a una variable aleatoria normal con media μ y desviación típica $\sigma/\sqrt{n}$. Por conveniencia a la hora de utilizar las tablas de percentiles se utiliza una transformación de esta normal, es decir, conseguir el estadístico $A=(m_1-\mu)/(\sigma/\sqrt{n})$ que está distribuido como una N(0,1).

En la mayoría de aplicaciones prácticas (la simulación entre ellas), no se pueden conseguir secuencias infinitas y no es posible conocer el valor exacto de la media. Por ello interesa conocer el rango de valores sobre los que cabe esperar que se situará μ con elevada probabilidad, este rango se conoce como *el intervalo de confianza sobre el valor medio*. Para encontrarlo bastaría con establecer la probabilidad de acierto (1- α) o de error (α) y encontrar los límites del estadístico *A*:

$$P(c_1 \le \frac{m_1-\mu}{\sigma/\sqrt{n}} \le c_2) = P(m_1 - c_2 \cdot \sigma/\sqrt{n} \le \mu \le m_1 - c_1 \cdot \sigma/\sqrt{n}) = 1-\alpha \tag{2}$$

La normal es simétrica respecto la media por lo que $c_1=-c_2$ igual al valor que ofrece un percentil 1-α/2 en una N(0,1). El intervalo de confianza para la media μ obtenido es $m_1 \pm c_1 \cdot \sigma/\sqrt{n}$.

Para realizar este análisis se necesita conocer la desviación típica real de *X*. Es una premisa importante puesto que en general no resulta así. Por ello se puede utilizar una estimación del segundo momento (o varianza muestral):

$$S_n^2 = m_2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - m_1)^2 \tag{3}$$

Puede demostrarse que este estimador está sesgado, en particular, $E(m_2) = (n-1) \cdot \sigma/n$. Aplicando el factor de corrección *n/(n-1)*, denominada corrección de Bessel, se obtiene el estimador insesgado del segundo momento:

$$S_{n-1}^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} \left(x_i - m_1\right)^2 \quad (4)$$

Ambos estimadores tienden a la desviación típica real con infinitas muestras, la diferencia entre ellos viene dada cuando el número de muestras es pequeño. La primera aproximación al intervalo de confianza consistiría en sustituir la desviación típica en la expresión (2) por alguna de estas estimaciones. El problema es que no tendríamos en cuenta la distribución de m_2. Puede demostrarse que el estadístico $B = n \cdot S_n^2 / \sigma^2$ está distribuido como una chi-cuadrado de *(n-1)* grados de libertad. Podemos por tanto calcular la variación que puede sufrir la desviación típica real σ con elevada probabilidad tal como sigue:

$$P(s_1 \le \frac{n \cdot S_n^2}{\sigma^2} \le s_2) = 1 - \beta \quad (5)$$

En este caso la variable de chi-cuadrado no es simétrica, por lo que s_1 y s_2 no tienen el mismo valor absoluto como ocurría con c_1 y c_2 en la normal. Está aceptado utilizar s_1 tal que el percentil sea igual a $1-\beta/2$.

Observando los estadísticos *A* y *B* vemos que se podría hacer desaparecer la desviación típica real de la ecuación, dividendo ambas convenientemente: $A/\sqrt{B}$. Además puede demostrarse que *A* y *B* son independientes por lo que multiplicando por $\sqrt{(n-1)}$ se obtiene un estadístico conocido, la distribución de *t-Student* con *(n-1)* grados de libertad. Puede utilizarse el estimador insesgado incorporando la corrección de Bessel, por lo que definitivamente se obtendría el estadístico *T*:

$$T = \frac{m_1 - \mu}{\sqrt{S_n^2/(n-1)}} = \frac{m_1 - \mu}{\sqrt{S_{n-1}^2/n}} \quad (6)$$

, es decir, establecer el intervalo de confianza como:

$$P(m_1 - t_2 \cdot \sqrt{S_{n-1}^2/n} \le \mu \le m_1 - t_1 \cdot \sqrt{S_{n-1}^2/n}) = 1 - \alpha \quad (7)$$

La distribución de *t-Student* es simétrica, por lo que $t_1 = -t_2$. Según el Teorema Central del Límite, dicha variable también tiende a una variable normal de media 0 y desviación típica 1. La diferencia entre usar las distribuciones de *t-Student* o la *normal* puede considerarse despreciable cuando el número de muestras (i.e. grados de libertad) es superior a 50 (muchas fuentes recomiendan 100). Sin embargo, la distribución de *t-Student* se emplea en múltiples ocasiones, incluso si el número de muestras supera los 50 valores, puesto que proporciona valores más conservadores del intervalo de confianza.

2.3 Automatización de simulaciones

El cálculo del intervalo de confianza exige múltiples realizaciones para poder proceder a su cálculo (más de 50 en la mayoría de casos). Por lo tanto se debe optar por algún mecanismo que permita al usuario del simulador automatizar la obtención de dichas realizaciones. Para ello se pueden seguir dos estrategias sencillas, tal y como se detallan a continuación. En ellas se asume que se desean conseguir N realizaciones con M muestras cada una de ellas.

2.3.1 Batch means

Esta técnica se propuso inicialmente para sistemas en los que alcanzar el régimen permanente requiere de un tiempo considerable. La técnica de *batch means* consiste en llevar a cabo una única realización, con una duración en muestras igual a N * M. Una vez finalizada la realización, se dividen las muestras obtenidas en N secciones de M muestras (haciendo que cada sección se

comporte como una simulación autónoma) y se extraen los valores de interés de cada una de dichas secciones. Las ventajas de esta técnica son la sencillez y la minimización del impacto del transitorio. Sin embargo, puesto que las secciones proceden de una única realización, existe una alta correlación entre el final de una sección y el inicio de la siguiente. Por lo tanto, este mecanismo sólo es válido en simulaciones en las que el número de muestras M es alto.

2.3.2 Repeticiones independientes

Esta técnica consiste en llevar a cabo N realizaciones independientes de M muestras cada una de ellas. Cada una de dichas realizaciones deberá partir de una semilla distinta para ofrecer distintas casuísticas. La principal ventaja de esta técnica es que se asegura la independencia entre los valores obtenidos a partir de cada una de las simulaciones. Sin embargo, cada simulación presentará su propio transitorio (lo cual distorsiona los resultados y puede aumentar el tiempo de ejecución) y requiere de un proceso de automatización para llevar a cabo un número relativamente alto de repeticiones.

Para implementar repeticiones independientes, se deben seguir los siguientes pasos:

1. Utilizar el frontend de consola. Para ello, basta con generar el makefile empleando, además de las ya comentadas en prácticas anteriores, las siguientes opciones (mirar el manual de *Omnet* para más información):

   opp_makemake -u *Cmdenv*

2. Modificar el fichero omnetpp.ini de tal forma que incluya las siguientes directivas.

```
include seeds.ini

[General]
network=nombre_de_la_red_a_simular

[Cmdenv]
express-mode = yes
```

Figura 1: Directivas a incluir en el fichero omnetpp.ini

La directiva *include* se encarga de incluir el contenido de un fichero en el *omnetpp.ini* de forma dinámica, es decir, al lanzar la simulación. Se utilizará esta ventaja para generar un fichero *seeds.ini* que contenga la semilla a utilizar en cada caso.

3. Implementar un *script* con el que realizar la tanda de realizaciones. En el contenido de la práctica se puede encontrar un *script* llamado *mrun.sh* que se encarga de ello.
4. Llamar al *script*. En el caso de *mrun.sh*, la llamada seguirá el siguiente formato:

   sh mrun.sh número_de_realizaciones nombre_del_ejecutable_de_omnet directorio_salida [semilla_inicial]

   El *script* generará ficheros *omnetpp-índice.vec*, con índice igual a un valor numérico, con los resultados obtenidos para cada una de las semillas generadas, así como un fichero denominado *seeds-índice.ini* que contendrá la información de las semillas utilizadas para la ejecución número *índice*.

   Ej. ***sh mrun.sh*** *50 omnet_exec directorio_1 543*

   Esto ejecutaría 50 runs de *omnet_exec* (ejecutable para el modelo a simular), con una semilla

inicial igual a 543 y guarda los resultados en el *directorio_1*. Se tendrían que particularizar el fichero para que las claves generadas coincidieran en nombre con las utilizadas por vuestros generadores

5. Utilizar el *script fvector.sh* para escoger un vector de resultados y producir ficheros equivalentes a los omnetpp.vec que los contienen, pero sólo con las marcas de tiempo y los valores de la variable. De esta forma se consigue importar los datos a *Matlab* fácilmente. Este *script* recibe dos parámetros que son el número de vector a filtrar y la expresión que regula los ficheros afectados. Un posible ejemplo de ejecución sería:

   Ej. ***sh fvector.sh*** *0 '*.vec'* ← *Selecciona el vector 0 de los ficheros *.vec y produce archivos de nombre *.vec.vector_0 , con el contenido del vector*

2.4 Estudio previo

2.4.1 Estudio previo 1

Recupere las fórmulas del sistema *M/M/N* para calcular los valores medios de las variables indicadas en el Ejercicio 1. Calcule dichos valores medios para los parámetros indicados en el Ejercicio 2.

2.4.2 Estudio previo 2

¿Cómo se pueden recoger muestras de una variable en omnet? Es decir, ¿qué clase(s) tiene omnet disponibles para almacenar muestras de una variable? ¿qué métodos se deben escoger para hacerlo?. Consulte si es necesario el manual de omnet, así como el código de los distintos módulos proporcionados (especialmente *Server.cc*).

2.4.3 Estudio previo 3

¿Qué procedimiento hay que seguir para realizar repeticiones independientes en *Omnet*? Detallar los cambios a realizar en los fichero de configuración, la manera en la que hay que compilar el modelo y la forma en la que se utilizaran los distintos scripts disponibles.

2.4.4 Estudio previo 4

En los ejercicios se necesitará una función para calcular momentos, por lo que es necesario encontrar las probabilidades de estado de la variable particular. Para ello y conseguir una mayor generalidad deberá considerar estas probabilidades como porcentajes temporales. Esta función tendrá por nombre *time_weighted_mean* y recibirá por parámetro una matriz *W*, con el siguiente formato:

tiempo_1 *valor_1*
tiempo_2 *valor_2*
.....
tiempo_N *valor_N*

Tal y como se puede apreciar la primera columna de dicho fichero indica el tiempo en el que se cambia de valor y la segunda el valor al que se cambia. De esta forma, de acuerdo con la nomenclatura anterior, el *valor_1* se produciría durante un tiempo igual a *tiempo_2-tiempo_1*, etc. El objetivo de la función será calcular la media y la desviación típica de los valores (es decir la segunda columna) de la matriz *W* de acuerdo a su porcentaje de tiempo (primera columna), es decir, la implementación de las siguientes expresiones:

$$mean = \frac{1}{T_N} \sum_{i=1}^{N-1} tiempo(i) * valor(i) \tag{8}$$

$$std = \sqrt{\frac{1}{T_N} \sum_{i=1}^{N-1} tiempo(i) * valor(i)^2 - mean^2} \tag{9}$$

```
function [mean,std] = time_weighted_mean(W)
% W es una matriz Nx2 con los valores de la variable (segunda columna) y el tiempo de escritura de
%       los mismos (primera culomna)
```

```
% mean contiene el valor medio calculado
% std contiene la desviación típica calculada
%
% La funcion seguirá el siguiente algoritmo
%
% 1) Calcular la diferencia entre valores consecutivos de la primera columna
% (es decir W(i+1,1)-W(i,1)). Al vector que contiene esos valores se le llamará TD.
% 2) Calcular el producto entre componentes de TD y la segunda columna de W (es decir el
% producto del valor de la variable por el tiempo en que ha tenido validez dicho valor).
% Llamaremos a ese producto raw_avg.
% 3) Calcular el producto entre componentes de TD y la segunda columna de W al cuadrado
% Llamaremos a ese producto SM.
% 4) mean tomará por valor raw_avg / ultimo valor temporal de W
% 5) std tomará por valor la raiz cuadrada de: SM / ultimo valor temporal de W - mean*mean
```

2.4.5 Estudio previo 5

En el Ejercicio 4 se necesitará una función para calcular los intervalos de confianza. Dicha función deberá:

1.Cargar los valores generados por el modelo de la *M/M/1* para cada una de las simulaciones.

2.Obtener el valor medio del tiempo de espera en cola para cada simulación

De esta forma, construya en *Matlab* una función tal y como se especifica a continuación:

```
function result = confidence_interval(path)
% path es un string que hace referencia al directorio donde se encuentran los resultados del tiempo
% de espera en cola. Recordad que para generar dichos resultados podéis filtrar los .vec generados
% desde la consola con el comando fvector.sh numero_vector
% result es un vector que contiene las medias de los distintos ficheros analizados
%
% La funcion seguirá el siguiente algoritmo
%
% 1) Obtener la informacion de todos los  archivos situados en el directorio path (help dir)
% 2) Para cada archivo
% 2.1) Leer su contenido
% 2.2) Obtener el valor medio
% 2.3) Guardar el valor medio en el vector de salida.
```

2.5 Ejercicios

Durante la práctica, emplee el generador de *Twister de Mersenne* (generador por defecto en omnet) a menos que se indique lo contrario y con una semilla inicial igual al número que aparece en su *login* de acceso al sistema. Recuerde asimismo que los mensajes intercambiados por los distintos módulos son de tipo *Lt3Message* y que la red ya se implementó en la práctica 1,

2.5.1 Ejercicio 1 (Sesión 1)

Genere un fichero *omnetpp.ini* con el que simular una *M/M/N*. Utilice para ello el código proporcionado en esta práctica, el cual ha sido extraído de la práctica 1.

Simule un modelo *M/M/10000* y cárguelo con *6000* **Erlangs** y un tiempo de servicio exponencial de *6* segundos. Ejecute una simulación que tenga una duración de *50000* muestras y calcule los

valores medios de las siguientes variables:

- Tiempo entre llegadas
- Número de unidades en cola
- Número de servidores ocupados
- Tiempo de espera en cola
- Probabilidad de demora

Represente la evolución de la ocupación del sistema (número de servidores ocupados) a lo largo de la simulación e identifique las distintas anomalías.

Proponga mecanismos para minimizar el impacto del transitorio y de la descarga del sistema (analice el código del *Server* y del *TrafficGenerator* si es necesario). Aplique uno de ellos y compare el error relativo cometido si no se eliminaran dichas desviaciones de los resultados obtenidos (para cada una de las variables).

Nota: Tenga presente que, para el cálculo del valor medio de algunas de las variables indicadas, se deberán ponderar los valores obtenidos por el tiempo en el que dichos valores tienen efecto (por ejemplo para el cálculo del número medio de unidades en la cola). Puede comparar estos resultados con los obtenidos aplicando PASTA, dado el carácter Markoviano del sistema y sabiendo que el simulador anota únicamente los cambios de estado de las variables de interés (se consigue así que los ficheros de resultados tengan el menor tamaño posible).

2.5.2 Ejercicio 2 (Sesión 1 y 2)

Simule un sistema *M/M/1* con una utilización del *75%* con un tiempo entre llegadas de *10* segundos. Para ello realice una simulación con *20000* muestras.

Encuentre el valor medio y la varianza de las variables indicadas en el Ejercicio 1 así como la función de densidad de probabilidad (*fdp* estimada a partir del histograma) del tiempo y del número de elementos en la cola. Compare estas curvas con las que corresponden únicamente a las unidades

que se esperan.

Proporcione los resultados teniendo en cuenta únicamente las primeras *200* muestras y compárelos con los obtenidos a partir de las primeras *2000* y también con todas las muestras. Indique los errores relativos cometidos en simulación en comparación con los valores teóricos.

Tenga presente las siguientes consideraciones/ayudas

- Recomendación: Para dibujar la evolución de la ocupación de los servidores se puede utilizar la función plot(x,y) de *Matlab*
- Para seleccionar un vector en concreto dentro del fichero omnetpp.vec se puede emplear la siguiente sentencia:

  awk '{ if ($1 == num_vector) print $2,$3}' < omnetpp.vec > nombre_fichero_salida.txt

2.5.3 Ejercicio 3 (Sesión 2)

Si la utilización teórica del sistema del Ejercicio 1 y 2 es la misma, ¿por qué no aparece el régimen transitorio en el Ejercicio 2? ¿Cuál es el número máximo de elementos en la cola $N_{q,max}$ alcanzado por el sistema en el Ejercicio 2? Justifique el resultado.

Encuentre un ajuste analítico de la función de probabilidad del número de elementos en la cola de las unidades que se esperan. Establezca el mejor ajuste como aquél que supere el test de Pearson. Proporcione a partir de este ajuste la probabilidad de que el sistema tenga $N_{q,max}+1$ unidades en la cola y compárela con el valor teórico real.

2.5.4 Ejercicio 4 (Sesión 3)

Este ejercicio pretende realizar un análisis del intervalo de confianza de las variables del tiempo (W_q) y número medio de los paquetes en la cola (N_q) de una *M/G/1*. Considere los mismos parámetros que en el Ejercicio 2 y una distribución *3_Erlang* para el tiempo de servicio. Repase el funcionamiento de *mrun.sh* y establezca el valor de la semilla inicial con el número de su login (para *lt3usr45* el valor inicial sería *45*). Para los siguientes sub-apartados calcule el intervalo de confianza teniendo en cuenta las alternativas expuestas en el apartado 2.2, es decir:

- Estadístico *A* y una estimación sesgada del segundo momento
- Estadístico *A* y una estimación insesgada del segundo momento
- Estadístico *A* y estadístico *B* para la estimación insesgada del segundo momento
- Estadístico *T* (i.e. una distribución de *t-Student*)

a) Realice *40* simulaciones con un número de muestras inferior a *500* y calcule el intervalo de confianza para la media de las variables teniendo en cuenta:

- Sólo *5* simulaciones de las *40* realizadas
- Todas las simulaciones

Comente los resultados.

b) Realice *40* simulaciones con un número de muestras superior a *20000* y calcule el intervalo de confianza para la media de las variables teniendo en cuenta:

- Sólo *5* simulaciones de las *40* realizadas
- Todas las simulaciones

Comente y compare los resultados con los obtenidos en el apartado anterior a).

c) Repita el caso anterior b) con el peor generador congruencial de los empleados en la práctica 2. Valore el impacto del generador en el cálculo del intervalo de confianza.

d) Asumiendo una media y varianza estimadas según el las muestras obtenidas en a), estime el número de realizaciones necesario para conseguir un intervalo de confianza menor al *7%* del valor de la media de las variables. Utilice para ello únicamente el estadístico A y una estimación insesgada del segundo momento. Comente la validez de la estimación.

e) Calcule el valor $t = \frac{(m_1 - \mu_0)}{\sqrt{S^2_{n-1}/n}}$, empleando para ello la media y la desviación típica estimada para el *caso b* (*40* simulaciones de *20000* muestras). Asumiendo que este estadístico debería seguir una distribución de *t-Student*, ¿considera correcto que el valor del tiempo medio real de espera en la cola puede ser el valor medio teórico más *1 s*. Utilice el mismo razonamiento que para el Test de Pearson (práctica 2).

f) Realice *10* experimentos, cada uno con *5* simulaciones de *500* muestras, y calcule el intervalo de confianza promedio expresándolo como un porcentaje de error respecto el valor medio del mismo. ¿Cuántos intervalos aciertan con la media teórica esperada de las variables?

g) Al igual que hizo en el ejercicio 3, encuentre un ajuste que supere el test de Pearson para la función de probabilidad del número de elementos en la cola de las unidades que se esperan para cada una de las *5* simulaciones con menos de *500* muestras. Encuentre el intervalo de confianza para el valor *p* de la geométrica buscada y extrapole ahora el resultado de simulación para encontrar la probabilidad de que el sistema tenga $N_{q,max}+1$ unidades en la cola.

h) Encuentre el intervalo de confianza para la desviación estándar de estas variables considerando el escenario que crea más oportuno.

i) Considere ahora como variable a estudiar la desviación estándar del tiempo y número medio de los paquetes en la cola en lugar de su media. Encuentre los intervalos de confianza sobre la media aplicados a estas desviaciones estándar según el estadístico *T* y compárelos con los obtenidos en el apartado h). Realice una valoración crítica de este procedimiento.