

Parametrización de modelos

Ejercicio: Ajuste de datos a una distribución

1. Objetivo

El objetivo de la práctica consiste en que el alumno aplique los conocimientos adquiridos sobre parametrización de modelos a casos concretos relacionados con el análisis y la configuración de equipos informáticos.

Como segundo objetivo se pretende que el alumno tome consciencia del concepto de carga que soporta un sistema informático y cómo se representa y parametriza la carga cuando se debe usar con modelos analíticos. Para ello, se resolverá un problema sencillo consistente en parametrizar la carga que soporta un ecaminador (router) a partir de unas medidas realizadas en una red.

1.1 Conocimientos previos

Como requisitos previos a la realización de la práctica están:

- Conocimiento de conceptos básicos de estadística, tales como las distribuciones de probabilidad continuas y discretas y las técnicas de ajuste de distribuciones a datos.
- Conocimientos básicos sobre la utilización de la hoja de cálculo EXCEL para el análisis de datos.

2. La práctica

2.1 Enunciado de la práctica

Un ingeniero informático encargado del mantenimiento de una red de computadores ha monitorizado la llegada de paquetes a un encaminador (router). Durante un periodo de 90 segundos, han llegado 220 paquetes y se han medido los tiempos entre las llegadas sucesivas de los paquetes i e $i+1$, para $i=1,2,\dots,219$. En la hoja de cálculo que se adjunta con la práctica se dispone de los tiempos entre llegadas después de que hayan sido ordenados de forma creciente.

El ingeniero ha descompuesto el período de 90 segundos en 6 subperíodos sucesivos de 15 segundos cada uno, contando los paquetes que han llegado en cada subperíodo. Se observa que en todos los subperíodos llega, aproximadamente, el mismo número de paquetes, deduciéndose entonces que la cadencia de llegada de paquetes es prácticamente constante en el período. Además, los paquetes llegan de uno en uno, y no existen motivos para creer que el número de llegadas en intervalos diferentes no sean independientes. Basándose en estas consideraciones teóricas, el ingeniero presupone que los tiempos entre las llegadas de los paquetes siguen una distribución de tipo exponencial.

Se pretende que los alumnos de SIF asesoren a este ingeniero con objeto de parametrizar con exactitud el modelo (la distribución) de la carga (tráfico) que está soportando el encaminador.

2.2 Pasos de resolución

2.2.1 Selección de un tipo de distribución

En primer lugar se analizan los datos para seleccionar una familia de distribuciones. Para confirmar las informaciones aportadas inicialmente se **analizan los estadísticos muestrales**. Para calcularlos, partir de las observaciones almacenadas en una hoja de cálculo en A1:A219. En la barra de menús de la hoja seleccionar Herramientas y en el menú Herramientas, la opción Análisis de datos¹. En la ventana de análisis de datos seleccionar la función Estadística Descriptiva y generar los estadísticos en una hoja nueva (hoja 2), habiendo activado la opción Resumen de estadísticas.

En este caso se puede comprobar que la media muestral (0,399) supera a la mediana (0,27) y el coeficiente de asimetría es 1,47. Todo esto indica que la distribución no puede considerarse simétrica sino sesgada a la derecha. Además el coeficiente de variación (0,953), que debe calcular el usuario ya que Excel no lo calcula, está muy próximo a 1, que es el valor de la distribución exponencial teórica. Por todos estos motivos se puede deducir que los datos se ajustarán bien a distribuciones de la familia de las exponenciales y no muy bien a distribuciones de la familia de la normal.

Seguidamente se construyen sucesivos **histogramas** de los datos (seleccionando Herramientas→Análisis de datos→Histograma) para obtener una indicación de la distribución que se puede usar para modelar los datos. Se usan intervalos de 0,050 0,075 y 0,100 para construir las marcas de clases. Construir las marcas de clases en la hoja 1 usando tres columnas. Por ejemplo:

- En F2:F42 poner 0,000 0,050 0,100 0,150 ... 2,000
- En G2:G29 poner 0,000 0,075 0,150 0,225 ... 2,025
- En H2:H22 poner 0,000 0,100 0,200 0,300 ... 2,000

A partir de estos datos se realiza la construcción de los tres histogramas

Histograma 1:

Rango de entrada: A1:A219, Rango de clases: F2:F42 y seleccionar crear gráfico (hoja 3).

Como se puede comprobar el histograma presenta un perfil muy entrecortado. Probar con una agregación de los datos mayor, osea, con un rango de clase más ancho.

Histograma 2:

Rango de entrada: A1:A219, Rango de clases: G2:G29 y seleccionar crear gráfico (hoja 4).

Todavía se aprecia un perfil entrecortado, se agregan más los datos.

Histograma 3:

Rango de entrada: A1:A219, Rango de clases: H2:H22 y seleccionar crear gráfico (hoja 5).

Se puede aceptar la forma de este histograma que se asemeja mucho al de una distribución exponencial.

Ahora se puede realizar un **resumen de cuantiles** de las observaciones. Se recomienda realizarlo en la hoja 1, usando la función PERCENTIL de Excel. Al calcular los puntos medios

¹ Si esta opción de menú no aparece, se debe instalar el módulo correspondiente. Ir a Herramientas→Complementos y seleccionar Herramientas para análisis.

entre los cuartiles (0,323), los octiles (0,460) y los extremos (0,985) de las observaciones se ve que se “alejan” progresivamente de la mediana (0,270) hacia la derecha. Esto confirma que la distribución subyacente de las observaciones está sesgada a la derecha, como la exponencial.

Llegados a este punto de la práctica se puede presuponer (basándose en razones teóricas y en el análisis experimental realizado) que la distribución de los tiempos de llegada entre paquetes sigue una distribución exponencial.

2.2.2 Estimación de los parámetros de la distribución

Para estimar los parámetros de la distribución, se consulta una tabla de distribuciones en la que se indica la fórmula de los estimadores de máxima verosimilitud para los parámetros de la distribución.

En el caso de la distribución exponencial, el parámetro a calcular es la media de la distribución. El estimador de máxima verosimilitud de este parámetro es la media muestral (0,399).

2.2.3 Verificación del ajuste de la distribución teórica a las observaciones

Para comparar la calidad del ajuste de la distribución teórica seleccionada y parametrizada a las observaciones se pueden emplear métodos heurísticos o plantear diversos tests de hipótesis. Los métodos heurísticos consisten en comparar gráficamente

- las funciones de densidad de probabilidad,
- las funciones de distribución de probabilidad y
- los resúmenes de cuantiles

de la distribución teórica ajustada con los de las observaciones.

En primer lugar se **comparan las funciones de densidad** de probabilidad. Para ello se compara histograma de las observaciones con uno (o varios) obtenido directamente de distribución ajustada.

Para que manejar los datos de manera más cómoda, colocar en A1:A20 los valores iniciales de cada marca de clase, en B1:B20 un guión y en C1:C20 los valores finales de cada marca de clase. En D1:D20 se colocan el número de observaciones que hay en cada marca de clase. Todos estos valores se pueden copiar de la hoja 5 para pegarlos en la hoja 6. En E1:E20 se calculan las proporciones de observaciones en cada clase a partir de D1:D20. Esto es, se divide el número de observaciones en cada clase por el número total de observaciones. Por tanto en la columna E1:E20 tenemos la función de densidad muestral.

En la columna F1:F20 calculamos la función de distribución teórica. Para ello, se realiza la integral de la función de densidad de probabilidad en cada intervalo. La función de densidad teórica es $f(x)=2,506*EXP(-x / 0,399)$ y su integral es $p(j)=0,285*EXP(-2,51*\Delta*j)$ para $j=1,2,...,20$. Como se puede comprobar j es el número de clase y Δ el ancho de la clase. Esto se puede realizar muy bien en la hoja excel con la fórmula $0,2848*EXP(-2,51*\$cn)$.

Otra opción, en general bastante aproximada y que no precisa de la integración analítica de la función de distribución, consiste en calcular directamente el valor de la función de densidad teórica en el centro de cada una de las clases. Para ello utilizar la función DISTR.EXP de Excel colocando los resultados en G1:G20. Los parámetros que le hay que pasar a la función son:

- X: centro de cada clase.

Parametrización de modelos

Ajuste de datos a una distribución

- Lambda: inversa de la media.
- Acum: falso.

Se debe multiplicar el resultado de la función por 0,1, es decir, por el tamaño del intervalo.

Finalmente realizar un gráfico de columnas en la misma hoja 6 utilizando las columnas E, F y G como datos. Como se puede comprobar la aproximación es bastante aceptable.

Otra forma adicional de comprobar el ajuste de la distribución teórica a las observaciones es realizando una **comparación gráfica de los resúmenes de cuantiles** de las observaciones y de la distribución teórica ajustada. Se recomienda seguir este procedimiento. Primero se van a calcular los cuantiles de la distribución teórica. Copiar en una hoja nueva las observaciones en A1:A219 para trabajar con comodidad. En B2:B8 poner los rótulos de las variables a calcular: Inicial, Octil, Cuartil, Mediana, Cuartil, Octil, Final. En C2:C8 colocar los valores probabilidad acumulados por el cuantil correspondiente: $1/(2*n)$ 0,125 0,250 0,500 0,750 0,875 y $1-(1/2*n)$. En D2:D8 se calculan los valores de los cuantiles. Para ello será preciso invertir la función de distribución teórica, que es $F(x)=1-EXP(-x/0,399)$. Al invertirla, es decir, al obtener el cuantil inverso despejando la x en la ecuación anterior, se obtiene $x_q = -0,399*LN(1-q)$, tomando q los valores que tenemos en C2:C8. En la columna E2:E8 colocar los valores 1 1 2 2 2 1 1 para luego realizar un gráfico. Ahora se van a calcular los cuantiles de la distribución muestral. En la celda F2 se coloca el menor valor de las observaciones. En la columna F3:F7 se calculan los cuantiles muestrales usando la función PERCENTIL de Excel y en la celda F8 se coloca el mayor valor de las observaciones. En la columna G2:G8 colocar los valores 1 1 2 2 2 1 1 para luego realizar un gráfico.

Con las columnas D y E se realiza un gráfico de dispersión (resumen de cuantiles teórico) y justo debajo de este, usando las columnas F y G y con las mismas escalas se pone en la hoja otro gráfico de dispersión (resumen de cuantiles muestrales). Esta comparación visual permite verificar el excelente ajuste de los datos a la distribución exponencial ajustada.

Para verificar el ajuste de la distribución teórica a las observaciones de modo formal se realiza **un test de hipótesis Chi²**. Con la hoja Excel se puede proceder del siguiente modo:

Suponiendo que se ha de ajustar a una distribución exponencial de media 0,399 se eligen k=20 intervalos diferentes de modo que cada uno de ellos contenga la misma proporción p_j (o cantidad) de observaciones, con $p_j=1/k=0,05$ para $j=1,2,...,20$. Entonces $np_j=219*0,05=10,95$ lo que satisface las recomendaciones para la elección de los intervalos para realizar el test Chi² que son: iguales probabilidades p_j y $np_j \geq 5$.

Utilizar una hoja nueva.

Colocar en A1:A219 las observaciones ordenadas.

Colocar en B2:B21 los números de intervalo $j=1,2,...,20$.

En C2:C20 calcular los cuantiles que contienen $1/20, 2/20, 3/20, \dots, 19/20$ de las observaciones que se obtendrían de la distribución teórica. Para esto se aplica la fórmula obtenida anteriormente $x_q=-0,399*LN(1-q)$, siendo $q=1/20, 2/20, \dots, 19/20$.

Utilizando estos cuantiles como marcas de clase seleccionar Herramientas→Análisis de datos→Histograma. Las marcas de clase utilizadas por el histograma deben quedar ubicadas en D2:D21 y las frecuencias N_j en la columna E2:E21.

Ahora se va a calcular el estadístico Chi², que tiene la siguiente fórmula:

Parametrización de modelos
Ajuste de datos a una distribución

$$X^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j}$$

Para ello se calculan primero los valores que están dentro del sumatorio, usando las frecuencias de E2:E21. Estos valores se deben dejar en F2:F21. A continuación, suman estos valores en F22 para obtener el estadístico.

En F23 obtener el valor de la distribución con la función PRUEBA.CHI.INV de Excel. Si el nivel de error aceptable es del 10%, se coloca 0,1 como primer parámetro de esta función, y como segundo parámetro se colocan los grados de libertad, que son el número de intervalos utilizados menos una unidad, esto es, 19.

2.3 Presentación de resultados

Se presentará una memoria de la práctica desarrollando los contenidos que se detallan a continuación.

- Tabla con los estadísticos muestrales obtenidos a partir de los datos y un párrafo indicando las conclusiones que se pueden extraer de ellos.

ESTADÍSTICO	VALOR
Media	
Error típico	
Mediana	
Moda	
Desviación estándar	
Varianza de la muestra	
Curtosis	
Coefficiente de asimetría	
Rango	
Mínimo	
Máximo	
Suma	
Cuenta	
Coef. de variación	

Comentario:

- El histograma final de los datos, indicando claramente los intentos realizados para seleccionar el intervalo de clases más correcto.

Ajuste de datos a una distribución

- Un resumen de cuantiles que ayude a clarificar el tipo de distribución a seleccionar y las conclusiones que se deducen de él.

CUANTIL	VALOR	VALOR	PUNTO MEDIO
Mediana			
Cuartiles			
Octiles			
Extremos			

Conclusiones:

- La distribución seleccionada y sus parámetros.

- Mostrar la representatividad de la distribución teórica ajustada comparando el histograma obtenido de las muestras con el de la distribución teórica.

- Comprobar la representatividad realizando una comparación gráfica de los resúmenes de cuantiles de la distribución teórica y de las observaciones.

- Finalmente realizar la prueba χ^2 para verificar formalmente que se puede aceptar el ajuste de las observaciones por la distribución propuesta. ¿Se puede aceptar el ajuste? ¿Por qué?