Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Simulación Estadística
Cali, abril de 2018

Generación de números aleatorios, modelos probabilisticos y pruebas de bondad de ajuste

Kevin Steven García[a], Cesar Saavedra[b]

1. Modelos probabilisticos

- Distribución Poisson

- **Pruebas de apoyo al riesgo sísmico tras la distribución Poisson:**
Este artículo trata de modelar el riesgo sísmico con una distribución Poisson y así lograr estimar la probabilidad de amenaza sísmica, esto con el fin de mitigar el riesgo de terremoto ya que estos no son predecibles. Primero debemos definir riesgo sísmico: en lugar de pérdidas económicas o bajas asociadas a terremotos, el riesgo sísmico se refiere a la tasa anual de un movimiento de tierra dado de excedencia.
A diferencia de la frecuencia de terremotos que se demostró siguiendo la distribución Poisson, se asume que el riesgo sísmico(la tasa anual de movimientos sísmicos del suelo) es el mismo tipo de variables aleatorias sin soporte tangible. En lugar de utilizar los algoritmos de probabilidad total empleados actualmente, este estudio aplicó la simulación Monte Carlo (MCS) para obtener la función de probabilidad de riesgo sísmico y luego la comparó con la distribución Poisson para ver si realmente está cerca de la predicción del modelo como se suponía. Sobre la base de un cálculo de referencia, el análisis muestra un buen acuerdo entre los dos, proporcionando por primera vez algunas pruebas de que el riesgo sísmico debería seguir la distribución Poisson. Para este articulo se utiliza una tasa media de 1.99 terremotos por año, es decir, nuestro parámetro para obtener las simulaciones va a ser $\lambda = 1.99$.
Para cumplir con el objetivo del articulo se generaron aleatoriamente valores correspondientes a 4 parámetros de terremoto, que son la magnitud, la distancia, la frecuencia, y el error del modelo y se evaluaron en una ecuación denominada ecuación de predicción de movimiento de tierra, que es un modelo de regresión que describe la correlación entre el movimiento del suelo, la magnitud y la distancia combinadas, obteniendo así valores al azar para las evaluaciones del riesgo sismico. Posteriormente, mediante la repetición de los cálculos para una serie de ensayos, se obtuvieron muchas muestras de peligro o riesgo sísmico, y luego se calculo el valor medio como las estimaciones de riesgo finales con el MCS. Para el método de Monte Carlo se sabe que es mucho más eficiente con un tamaño de muestra grande, en este articulo se tomaron 10000 repeticiones. Finalmente, con un número de muestras de peligro sísmico de simulación de Monte Carlo, los análisis estadísticos (pruebas de bondad de ajuste) atestiguan que su función de probabilidad debe seguir la distribución de Poisson, proporcionando evidencia tangible por primera vez en apoyo de esta presunción de ingeniería que se ha utilizado comúnmente en los análisis de terremotos.

- Distribución Logística

- **Aplicación de la regresión logística y de los algoritmos genéticos en el análisis del riesgo de crédito:**
Este articulo nos presenta el uso de la regresión logistica y algoritmos genéticos para la clasificación de buenos y malos pagadores en financiamientos bancarios y la identificación del mejor modelo en términos de calidad de ajuste. A partir de una muestra de 14.000 datos, suministrada por una gran institución financiera brasileña, se aplicaron las dos tecnicas. La regresión logistica presentó un mejor ajuste a los datos. Este estudio ilustró los procedimientos que una empresa puede adoptar para identificar el mejor modelo de concesión de créditos o prestamos, a partir del cual es posible dirigir la estrategia de la institución en el proceso de evaluación de solicitudes de prestamos bancarios.
Los modelos de crédito son utilizados por las instituciones financieras en general y especificamente por los bancos con algunas finalidades: reducir los costos del proceso de concesión de crédito; reducir el aumento de no recibir el valor concedido; mejorar el proceso de decisión de crédito y reducir el tiempo y el esfuerzo empleado en esa actividad. Los

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1628466. E-mail: cesar.saavedra@correounivalle.edu.co

modelos de acreditación son especificos para la aprobación en cada producto de crédito. Los productos de crédito pueden ser: crédito personal, cheque especial, financiaciones, entre otros. En este estudio, el producto en cuestión es el crédito personal.
El punto principal para la concesión de crédito es la evaluación del riesgo. Si el riesgo es mal evaluado, la empresa seguramente perdera dinero, ya sea por la aceptación de clientes malos que generarán perdidas al negocio, o ya sea por el rechazo de clientes buenos que generarían ganancias al negocio.
Para la realización de este estudio, se proporcionó información sobre la historia de los clientes que han contratado un crédito personal con una institución financiera, de todos estos clientes se seleccionaron al azar 7000 clientes buenos y 7000 clientes considerados malos. En el trabajo la muestra se divide en dos submuestras del mismo universo de interés: uno para la construcción del modelo, 8.000 datos (siendo 4.000 bueno y 4.000 malo) y otro para validación del modelo construido, 6.000 datos (siendo 3.000 bueno y 3.000 malo). La submuestra de construcción modelo se utiliza para la estimación de los parámetros del modelo y la submuestra de validación tiene como función para verificar la potencia de predicción de los modelos construidos.
Para la construcción del modelo se consideraron las siguientes variables explicativas (relacionadas con el perfil del cliente) y el uso y restricción (relacionados con restricciones de crédito y notas sobre otras operaciones de crédito al cliente): sexo, estado civil, teléfono en casa, teléfono Comercial, tiempo en el trabajo actual, salario del cliente, cantidad de parcelas a pagar, primera adquisición, tiempo en la residencia actual, importe de la parcela, valor total del préstamo, tipo de crédito, edad, CEP residencial, CEP comercial, código de profesión, nombre de la profesión, salario del cónyuge.72 categorías de las variables originales fueron obtenidas para ser utilizadas en la construcción de los modelos.
Las variables independientes enfocadas en este estudio fueron codificadas en forma de variables Dummies. Para cada variable el número de categorías (k) determinó el número de variables Dummies (k-1) incluidas en el procesamiento de la regresión logística. El procedimiento empleado para la inclusión de variables independientes en el modelo logístico fue el avance gradual. En los modelos de paso hacia adelante sólo se inicia con el término de la constante, excepto cuando se omite este parámetro en la especificación de modelado, y en cada etapa se introduce la variable con el nivel de significancia más bajo para la puntuación estadística, siempre que se trata de menos que un valor de remoción (corte), establecido en 0,05 en este trabajo.
entonces se establece que la probabilidad de que el cliente sea un buen pagador se da, según el modelo logístico, por: $p = \frac{e^Z}{1+e^Z}$
Considerando k-1 Dummies para cada variable de nivel k, se generaron 53 variables independientes, de las cuales 28 se incluyeron en el modelo. En este estudio, Z es la combinación lineal de las 28 variables independientes ponderadas por los coeficientes logísticos.
Finalmente, luego de obtener el modelo se evaluó su significancia y se concluyó que el modelo de regresión logística presentó resultados estadísticamente significativos, que refuerzan su idoneidad.

2. Simulación de variables aleatorias

2.1. Introducción

En esta sección se mostrarán los resultados de las simulaciones de ambas distribuciones (Poisson y Logistica) donde la Poisson se simuló por medio del método propio y la Logistic se simuló por medio de la transformada inversa, Se comparará gráficamente y con pruebas estadísticas como la kolmogorov y la χ^2, si los métodos de simulación de verdad simulan las variables aleatorias que queremos. Posteriormente, se concluirá acerca de los métodos de simulación usados teniendo en cuenta los resultados de la prueba y el contexto en el que se utiliza la distribución en los artículos seleccionados.

2.2. Descripción general de las distribuciones y las pruebas de bondad de ajuste utilizadas

2.2.1. Distribución Poisson

Esta distribución es una de las más importantes distribuciones de variable discreta. Sus principales aplicaciones hacen referencia a la modelización de situaciones en las que nos interesa determinar el número de hechos de cierto tipo que se pueden producir en un intervalo de tiempo o de espacio, bajo presupuestos de aleatoriedad.
Su función de densidad esta dada por:

$$f(x,\lambda) = \frac{e^{-\lambda} \cdot \lambda^x}{x!}; x \in \{0,1,2,3,...\}$$

En este caso, el artículo trata de modelar el riesgo sísmico en un año, definiendo riesgo sísmico como la tasa anual de un

movimiento de tierra dado de excedencia. Por lo anterior se plantea esta distribución ara modelar este fenómeno, ya que se trata de determinar el número de hechos de cierto tipo (en este caso de movimientos de tierra) que se producen en un intervalo de tiempo o de espacio (en este caso es acotado por un año).

2.2.2. Distribución Logística

2.2.3. Pruebas de bondad de ajuste

Las pruebas de bondad de ajuste tienen por objetivo determinar si los datos se ajustan a una determinada distribución de probabilidad, esta distribución puede estar completamente especificada o perteneciente a una clase paramétrica.Para la bondad de ajuste de los datos a una distribución fija dada, los contrastes básicos son el χ^2 de Pearson y el de Kolmogorov-smirnov.

Prueba χ^2 de Pearson:

Esta prueba plantea las hipotesis:

$$H_0 : f(x,\theta) = f_0(x,\theta)$$
$$H_1 : f(x,\theta) \neq f_0(x,\theta)$$

El estadístico de prueba es: $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$
Si n ∞, ($np_i > 5$), es la verdadera distribución (sin parámetros desconocidos), tiene una distrinución Chi-Cuadrado con k-m-1 grados de libertad. Donde k:Número de categorías mutuamente excluyentes, y m: Número de parámetros del modelo o distribución.
La regla de decisión esta dada por: Para un nivel de significancia de α se halla el valor $\chi^2_{1-\alpha,k-1}$ a partir de $P(\chi^2 > \chi^2_{1-\alpha,k-m-1}) = \alpha$
Si $\chi^2_{Obs} > \chi^2_{1-\alpha,k-1}$ se rechaza la hipótesis nula H_0

Prueba Kolmogorov-Smirnov:

Esta prueba, al igual que la anterior, sirve para probar si una muestra o un conjunto de datos tienen o siguen una distribución hipotetica.
Se plantean las siguientes hipótesis:

$$H_0 : F(x,\theta) = F_0(x,\theta)$$
$$H_1 : F(x,\theta) \neq F_0(x,\theta)$$

Estadístico de prueba: Sea $F_n(x)$ la función de distribución empirica basada en las muestra aleatoria $x_1, x_2, ..., x_n$, el estadistico de prueba para la hipotesis planteada es:

$$T = sup|F_n(x) - F_0(x)|$$

Región de rechazo:$R_c = T|T > W_{1-\alpha}$, donde $W_{1-\alpha}$ es calculada de la tabla de valores críticos.

2.3. Simulación de las observaciones:

2.3.1. Distribución Poisson:

Recordemos que el parámetro sobre el cuál se van a generar las observaciones es $\lambda = 1.99$ que es la tasa o el número esperado de movimientos sísmicos durante un año. La generación de estos números se hizo por el método de

2.3.2. Distribución Logística:

Con esta distribución, se buscó generar las probabilidades de ser un buen cliente dadas unas características (variables independientes del modelo) de cada una de las 14000 personas en cuestión, es decir, se generaron 14000 valores entre 0 y 1 que representan tales probabilidades.

2.4. Resultados

2.4.1. Distribucción Poisson:

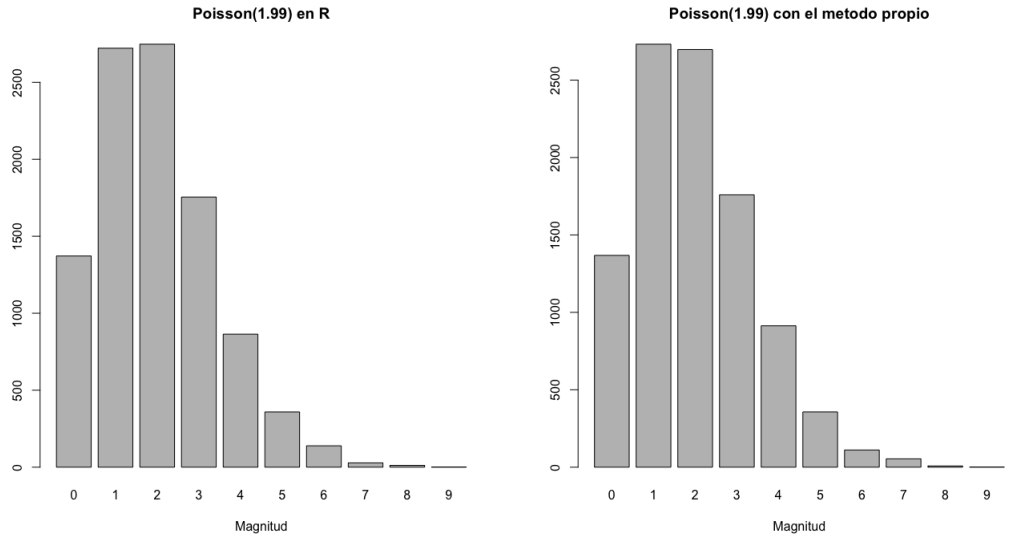

FIGURA 1: Comparación simulaciones en R y con el método propio de la distribución Poisson($\lambda = 1.99$)

Prueba de bondad de ajuste χ^2 de Pearson:

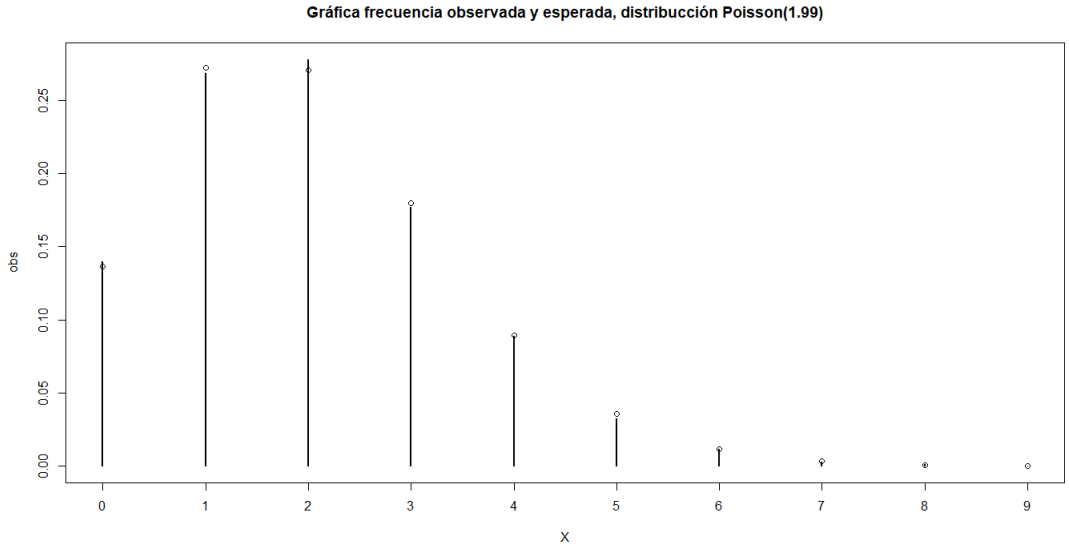

FIGURA 2: Frecuencia esperada vs observada de la distribución Poisson($\lambda = 1.99$)

El estadistico de prueba de la prueba χ^2 de Pearson nos dio $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.001783335$ y el $p - valor = P(\chi^2_{k-1} > \chi^2_{obs}) = P(\chi^2_9 > 0.00178335) = 1$. Por lo tanto, no rechazamos H_0 y concluimos que los datos generados por el método de

la transformada inversa se distribuyen Poisson($\lambda = 1.99$), esto quiere decir que el método de generación de datos utilizado realmente funciona.

2.4.2. Distribucion Logistica:

Prueba de bondad de ajuste Kolmogorov-Smirnov:

2.5. Conclusiones

2.6. Bibliografía

? ? ? ?