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

GENERACION DE NUMEROS ALEATORIOS, MODELOS PROBABILISTICOS Y PRUEBAS DE BONDAD DE AJUSTE.

Kevin Steven García[a], Cesar Andres 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, buscando evaluar el riesgo de variables no predecibles, como los terremotos. Primero es necesario hacer un acercamiento a la definición de riesgo sísmico, el cual, en contraste a definiciones de índole económica o a bajas asociadas, 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 artículo se utiliza una tasa media de 1.99 terremotos por año, es decir, el parámetro para obtener las simulaciones va a ser $\lambda = 1.99$.
Para cumplir con el objetivo del artículo se generaron aleatoriamente valores correspondientes a 4 parámetros de terremoto, que son: la magnitud, la distancia, la frecuencia, y el error del modelo. Los valores generados se evaluaron en una ecuación denominada ?ecuación de predicción de movimiento de tierra?; esta ecuación 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 sísmico. Posteriormente, mediante la repetición de los cálculos para una serie de ensayos, se obtuvieron múltiples muestras de peligro o riesgo sísmico y luego se calculó el valor medio como las estimaciones de riesgo finales con el MCS. La eficiencia del método Monte Carlo se ve en gran medida aumentada en estudios donde el tamaño de la muestra es considerablemente grande, en el artículo 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 artículo presenta el uso de la regresión logística 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 técnicas señaladas con anterioridad, donde se obtuvo como conclusión importante que la regresión logística 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 préstamos bancarios.

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

Los modelos de crédito son utilizados por las instituciones financieras en general y específicamente 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 esta actividad. Los modelos de acreditación son específicos para la aprobación en cada producto de crédito y estos a su vez pueden ser de diferentes tipos como: crédito personal, cheque especial, financiaciones, entre otros. En el estudio presentado en el artículo, 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 perderá 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 del estudio se proporcionó información sobre la historia de los clientes que han contratado un crédito personal con una institución financiera, de todos estos clientes seleccionaron al azar 7000 clientes buenos y 7000 clientes considerados malos. En el artículo la muestra se divide en dos submuestras del mismo universo de interés: uno para la construcción del modelo con 8.000 datos (siendo 4.000 buenos y 4.000 malos) y otro para validación del modelo construido con 6.000 datos (siendo 3.000 buenos y 3.000 malos). La submuestra de construcción de modelo se utiliza para la estimación de los parámetros del modelo y la submuestra de validación fue usada 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. Como resultado 72 categorías de las variables originales fueron obtenidas para ser utilizadas en la construcción de los modelos.
Las variables independientes enfocadas en el 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 ocurre, 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 evidencian los resultados de las simulaciones de los artículos referenciados en el numeral 1 del presente documento, donde: la distribución Poisson fue simulada por medio del método propio y la distribución Logística se simuló por medio de la transformada inversa. Se comparó si los métodos de simulación de verdad simulaban las variables aleatorias deseadas, esto mediante método gráfico y haciendo uso de pruebas estadísticas de bondad de ajuste como la kolmogorov y la χ^2. Posteriormente, se concluye 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 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 está 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 y por tanto se plantea esta distribución para modelar el 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

La función de distribución de la logística se utiliza como modelo de crecimiento. Por ejemplo, con un nuevo producto a menudo encontramos que el crecimiento es inicialmente lento, entonces gana impulso, y finalmente se ralentiza cuando el mercado está saturado o se alcanza alguna forma de equilibrio. Esta distribución tiene uso en los modelos logit, ya que estos modelos trabajan con su función de distribución acumulada denominada función logistica, por la simpleza de esta para encontrar probabilidades. Su función de distribución acumulada está dada por:

$$F(x,a,b) = \frac{1}{1+e^{-(\frac{x-a}{b})}}$$

En este caso, el articulo trata de modelar las probabilidades de ser un buen cliente dadas 28 características o variables independientes (como sexo, estado civil, etc.), por medio del modelo logit.

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 puede pertenecer 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 hipotética. 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:

Es importante recordar que el parámetro sobre el cuál se van a generar las observaciones es $\lambda = 1.99$, el cual corresponde a la tasa o el número esperado de movimientos sísmicos durante un año. La generación de estos números se hizo por el método propio de la distribución a partir de la generación de distribuciones exponenciales y por medio del comando rpois del software estadístico R.

2.3.2. Distribución Logística:

El uso de esta distribución en la problemática del articulo es una función derivada de ella, es decir, no se usa directamente la función de densidad logística para describir el fenómeno, se usa su función de distribución acumulada, denominada también función logística. Por lo cuál, en este caso no fue posible extraer unos parámetros como tal del articulo para simular los datos, es decir, no se logro darle sentido real a los parámetros con los cuales se van a generar los datos. Teniendo en cuenta lo anterior, se generaron 14000 datos con distribución Logística($a = 0, b = 2$) para mostrar la efectividad del método generador aplicado y para mostrar los resultados de la prueba de bondad de ajuste.

2.4. Resultados

2.4.1. Distribucción Poisson:

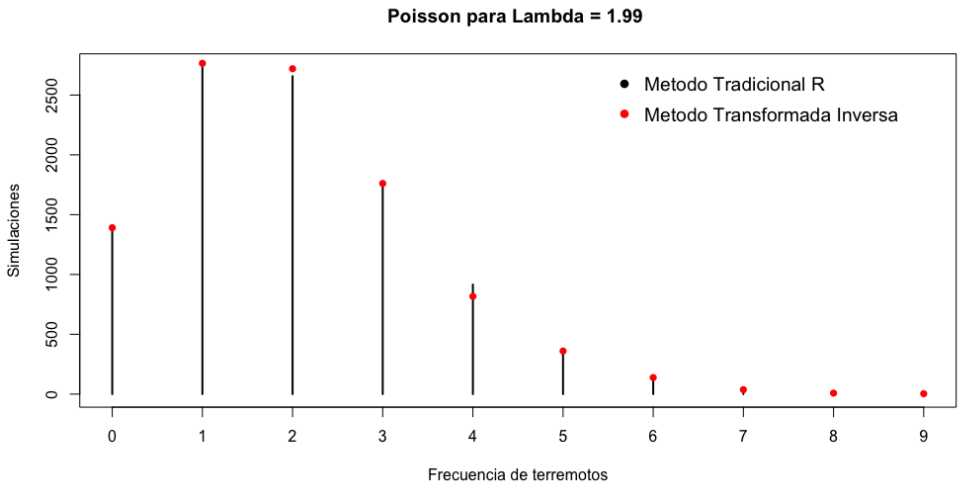

FIGURA 1: Comparación simulaciones en R y con el método propio de la distribución Poisson($\lambda = 1.99$)

En esta gráfica se puede observar como los valores obtenidos por el generador de R son casi identicos a los valores obtenidos por el generador propio de la distribución Poisson, por lo cuál se espera que la prueba de bondad de ajuste arroje que la distribución de los datos generados por el método propio de la Poisson realmente sigue una distribución Poisson($\lambda = 1.99$)

Prueba de bondad de ajuste χ^2 de Pearson:

observado	Esperado	$(observado - esperado)^2/Esperado$
0.1404	0.1367	0.000100
0.2733	0.2720	0.000006
0.2617	0.2706	0.000292
0.1761	0.1795	0.000064
0.0.933	0.0893	0.000179
0.0390	0.0355	0.000345
0.0125	0.0118	0.000041
0.0033	0.0033	0
0.0002	0.0008	0.00045
0.0002	0.0001	0.0001

Lo que nos permite obtener un valor total de 0.0015792 para un $\chi^2 = 0.0015792$

El estadístico de prueba de la prueba χ^2 de Pearson nos dio $\chi^2_{obs} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = 0.0015792$

El cuantil $\chi^2_{0.95,8} = 2.7326$ es mucho mayor que el estadistico de prueba obtenido $\chi^2_{obs} = 0.0015792$ por lo tanto no se rechaza H_0

El $p - valor = P(\chi^2_{k-m-1} > \chi^2_{obs}) = P(\chi^2_8 > 0.0015792) = 1$.

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:

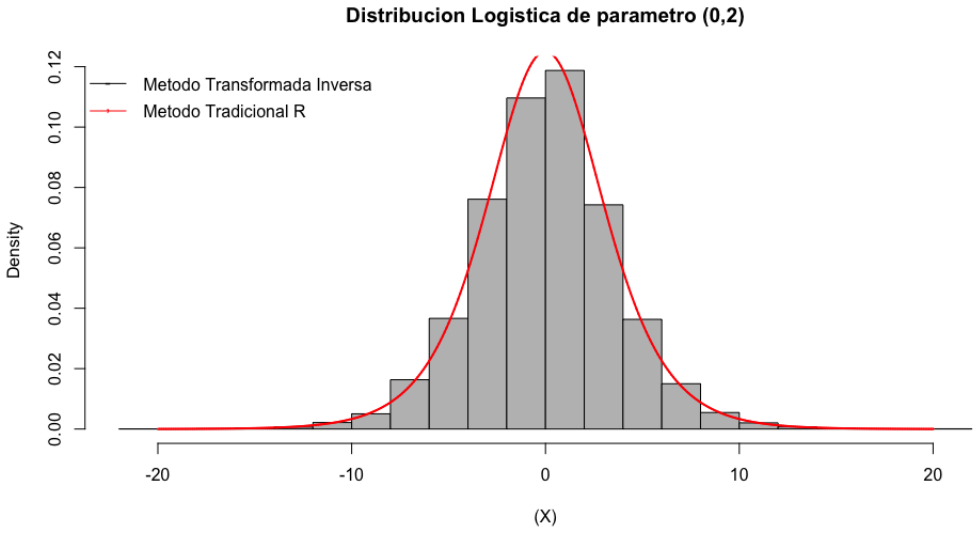

FIGURA 2: Comparación simulaciones en R y con el método de la transformada inversa para la distribución Logistica($a = 0, b = 2$)

en esta gráfica se puede ver claramente como el histograma de los datos generados por el método de la transformada inversa de la distribución Logistica muestra un comportamiento que sigue realmente la distribución esperada. La curva roja es la densidad teorica Logistica y vemos claramente se adapta al histograma generado. Por lo cuál, al igual que en la Poisson, esperamos que la prueba de bondad de ajuste no nos rechace H_0 concluyendo que realmente siguen una distribución Logistica.

Prueba de bondad de ajuste Kolmogorov-Smirnov:

```
> #Prueba de bondad de ajuste kolmogorov para la logistic:
> ks.test(X,'plogis',0,2)

        One-sample Kolmogorov-Smirnov test

data:  X
D = 0.004715, p-value = 0.9147
alternative hypothesis: two-sided
```

FIGURA 3: Resultados prueba de Kolmogorov para los valores generados con el método de la transformada inversa para la Logistic($a = 0, b = 2$)

En los resultados anteriores, se puede observar un p-valor de 0.9147, el cuál es demasiado grande. Lo cuál nos lleva a no rechazar H_0 y concluir que los datos generados por el método de la transformada inversa de la Logística, realmente siguen dicha distribución.

2.5. Conclusiones

- Se puede concluir que estas distribuciones tienen un uso muy importante en problemas sociales y económicos de gran tamaño. Por un lado, la distribución Poisson se utilizó para tratar de modelar y mitigar el riesgo sísmico (completar conclusión del articulo). Por otro lado, la distribución Logistic se usó para predecir u obtener la probabilidad por medio de un modelo logit, de que un cliente sea buen o mal pagador en cuanto a sus créditos personales dadas ciertas características, esto es muy importante en las empresas financieras, ya que con tan solo unas pocas preguntas personales al cliente, pueden obtener dicha probabilidad logrando disminuir considerablemente los riesgos crediticios, evitando así generar inmensas perdidas en la empresa.
- Con respecto a los métodos de simulación y pruebas de bondad de ajuste utilizadas, podemos ver que tanto el método utilizado para la distribución Poisson (método propio) como para la Logística (método de la transformada inversa) son muy acertados como generadores de variables aleatorias, ya que generan casi identicamente datos que siguen dichas distribuciones, y esto se ve apoyado con las pruebas de bondad de ajuste utilizadas, ya que ambas se nos rechazan por mucho, es decir, los datos generados muestran demasiada evidencia de que siguen las distribuciones hipotéticas.

2.6. Bibliografía

- **Statistical distributions, Catherine Forbes, Merran Evans, Nicholas Hastings, Brian Peacock, Fourth edition.**
- **http://www.redalyc.org/html/750/75040605/**
- **https://www.sciencedirect.com.bd.univalle.edu.co**