Resultado de Elección

Simulación de elección dada una muestra de votantes

Practica Extra

Modelado y simulación

Samuel Salas Meza

ENES Unidad Morelia, UNAM

Introducción

Stanislaw Ulam, mejor conocido por su desarrollo de bombas nucleares, alguna vez se encontraba jugando un juego de solitario. Por afición a las matemáticas decidió encontrar la probabilidad de ganar el juego de manera sencilla: Número de juegos ganadores entre número total de juegos. A pesar de ser un brillante matemático, fallo porque los cálculos resultaron ser demasiado complicados. Entonces decidió hacer inferencias estadísticas jugando algunos juegos y contando cuántos ganaba. Pero para obtener un resultado fiable, pensó que tardaría demasiado tiempo. Fue entonces que le llamó a su amigo John Newman, el inventor de la computadora de programa almacenado, quien accedió a ayudar con la simulación que conocemos hoy como el método Monte Carlo. Ese método sería de gran ayuda más adelante en el desarrollo de bombas atómicas de hidrógeno.
El método Monte Carlo usa inferencias estadísticas para estimar un valor a partir de una población y una muestra bajo el concepto de que si elegimos una muestra aleatoria, obtendremos un comportamiento similar al de la población.
Dato curioso: Monte Carlo es un casino.

Objetivo

Para esta práctica se tiene el objetivo poner a prueba el método Monte Carlo para ver si podemos encontrar el ganador de una elección dada una muestra aleatoria.

Metodología

Construcción de población:

Para encontrar la efectividad del método he fabricado datos de una elección donde se registró, para cada voto:
- Edad (18-80)
- Religión (Cristiano, Hindú, Ateo)
- Educación máxima (Primaria, Secundaria, Preparatoria, Universidad, Postgrado)
- Voto (A,B)

Como hacer a todos los datos equiprobables daría un resultado que tendería a una elección cerrada, seleccioné la probabilidad de que cierto atributo rinda a cierta elección de votante de la siguiente manera:

Edad:

Para edad se construyó una función de probabilidad lineal de tal manera que votar por A sea inversamente proporcional con la edad. Los de 18 años tendrán un 65% de probabilidad de votar por A y los más viejos, de 80 tendrán un 35% de probabilidad de votar por A.
Entonces la ecuación será: y-65=-35/62*(x-18)

En cuanto a población, se tuvo una probabilidad de 30% de caer entre los 18 y 30 años, un 25% de tener entre 31 y 50 años, un 20% de tener entre 50 y 70 años y un 15 de tener entre 70 y 80 años. Una vez seleccionado el grupo, es equiprobable la edad actual.

Religión:

Los cristianos tendrán un 42% de probabilidad de votar por A.
Los Hindúes un 73% de votar por A.
Los Ateos tendrán un 58% de votar por A.

Una persona tiene un 70% de probabilidad de ser cristiano, un 20% de ser Hindú y un 10% de ser Ateo.

Educación:

Para la educación se tiene una relación similar a la edad. Esta vez entre más educada está la gente, menos votará por A.
Primaria: 58% de votar por A
Secundaria: 57% de votar por A
Preparatoria: 51% de votar por A
Universidad: 50% de votar por A
Postgrado: 44% de votar por A

La población que curso máximo hasta la primaria es de 30%, los de secundaria tienen 30%, los de preparatoria tienen 20%, los de universidad 15% y los de postgrado 5%.

Cálculo del voto:

Para calcular el voto de una persona se tomaron en cuenta todas sus probabilidades y se tomará el promedio de ellas. Y utilizando un método de aleatorieidad en Python se decidió por quien votó.

Después se tomaron 10 muestras aleatorias de 10, 50, 100, 500, 1000 personas para encontrar a partir de cuántas personas se podía encontrar una predicción del resultado de la elección fiable. Una predicción fiable es de 10/10 aciertos.

Predicciones de muestra:

Edad:

Para eliminar la mayor cantidad de variabilidad posible, hice un acomodo por grupos. Primero hice un recuento de cuántas personas votaron por qué candidato y las agrupe por edades. Como por aleatoriedad algunos grupos quedarían descompensados o sobrecompensados en representación, tomé el promedio de votantes por el candidato “A” de la edad que deseo encontrar y sus cuatro edades más cercanas. Así encontré la probabilidad por grupos de edades.

Religión y educación

Para estos rubros simplemente conté el número de votantes por a que pertenecían a cierta religión o grupo educativo y los dividí por votantes totales de la muestra con su respectiva característica.

Excepciones:

Cuando el número de votantes era 0 en la sección de edad, buscaba el valor significativo más cercano y lo adoptaba.
Cuando el valor era cero en religión o educación, simplemente tomaba el 50 por ciento.

Resultados

Al correr la simulación con los parámetros que especifiqué salieron los siguientes resultados (el programa está en ingles para subirlo a Github):

The simulation was successfully ran with a sample size of: 10
It got the winner right in 5 iterations
It got the winner wrong in 5 iterations
The simulation was successfully ran with a sample size of: 50
It got the winner right in 3 iterations
It got the winner wrong in 7 iterations
The simulation was successfully ran with a sample size of: 500
It got the winner right in 4 iterations
It got the winner wrong in 6 iterations
The simulation was successfully ran with a sample size of: 1000
It got the winner right in 2 iterations
It got the winner wrong in 8 iterations
The winner of the election was a
502350 people voted for a
497650 people voted for b

Claramente no fue exitosa en predecir el ganador de la elección. Pero analicemos con mayor detalle. Si hacemos un plot de las gráficas de probabilidad para edad, religión y educación obtenemos, para 1000 datos:

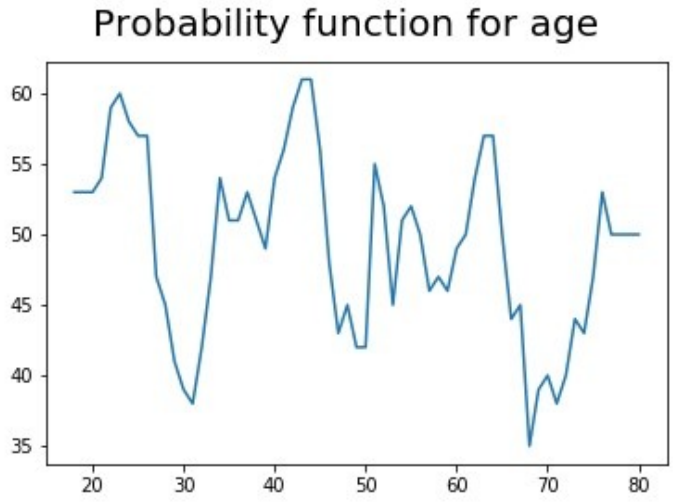

Figura 1: probabilidades por edad

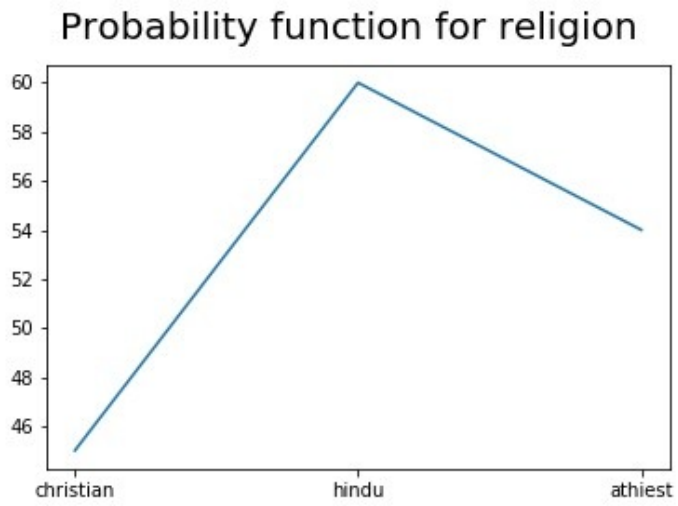

Figura 2: Probabilidades por religión

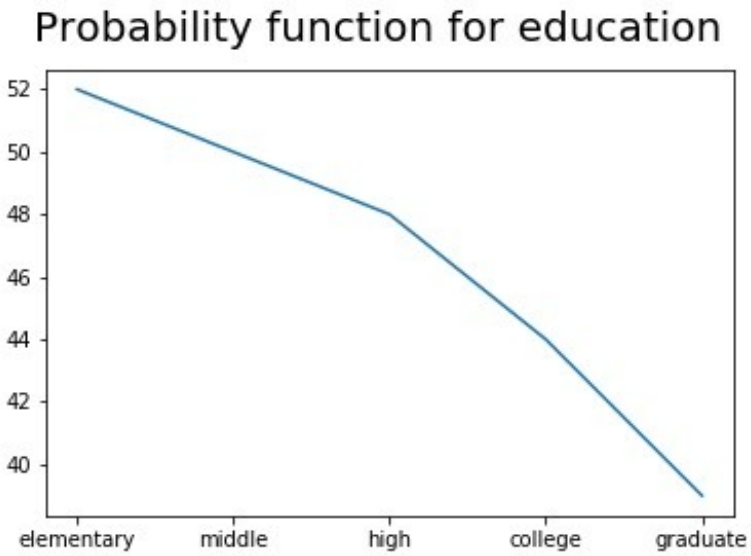

Figura 3: Probabilidades por grado educativo

Conclusiones

Aunque la gráfica de edades no muestra la tendencia esperada, las otras dos gráficas muestras valores cercanos a los que usamos en las probabilidades para definir a la población. Cabe mencionar que la población no tiene exactamente las proporciones que definimos para ella. De hecho el diccionario de predicción es una muestra aleatoria de una población creada con aleatoriedad. Puede ser esa una razón por la cual una gráfica tan demandante en datos como la de edades esté tan variable.

Ahora tocando el tema más importante: la predicción no fue acertada. Probablemente se eligieron valores que no mostraban una clara ventaja. Como resultado hubo una votación demasiado cercana para hacer una predicción fiable. Para poner a prueba esta hipótesis podemos cambiar el valor de probabilidad de los cristianos a 56 por ciento.

Obtenemos los siguientes resultados:
The simulation was successfully ran with a sample size of: 10
It got the winner right in 7 iterations
It got the winner wrong in 3 iterations
The simulation was successfully ran with a sample size of: 50
It got the winner right in 7 iterations
It got the winner wrong in 3 iterations
The simulation was successfully ran with a sample size of: 500
It got the winner right in 9 iterations
It got the winner wrong in 1 iterations
The simulation was successfully ran with a sample size of: 1000
It got the winner right in 10 iterations
It got the winner wrong in 0 iterations
The winner of the election was a
535541 people voted for a
464459 people voted for b

Ahora con una ventaja ligeramente más clara, los resultados fueron muchísimo más acertados. Aún cuando las gráficas no salieron del todo bien:

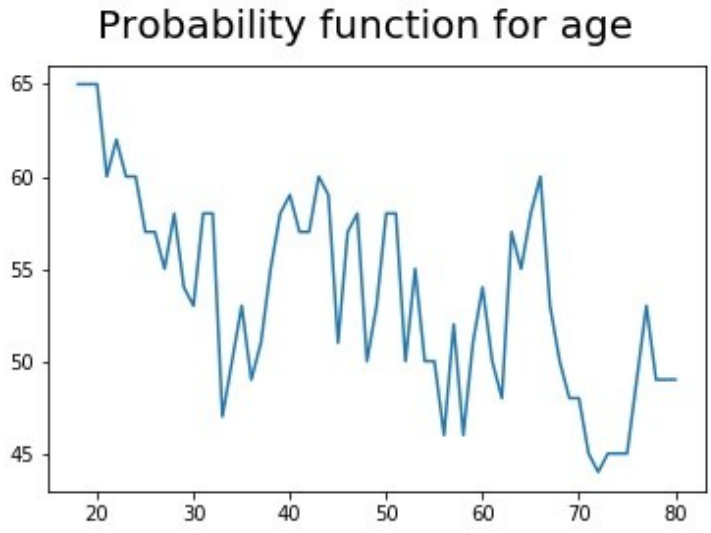

Figura 4: Probabilidad por edad

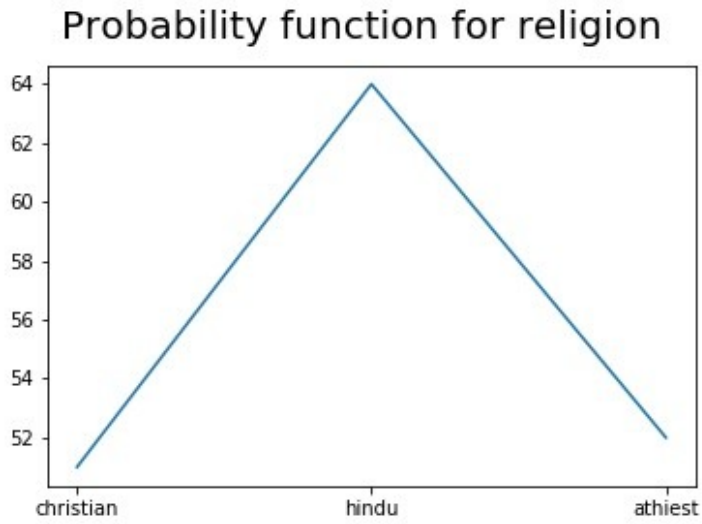

Figura 5: Probabilidad por religión

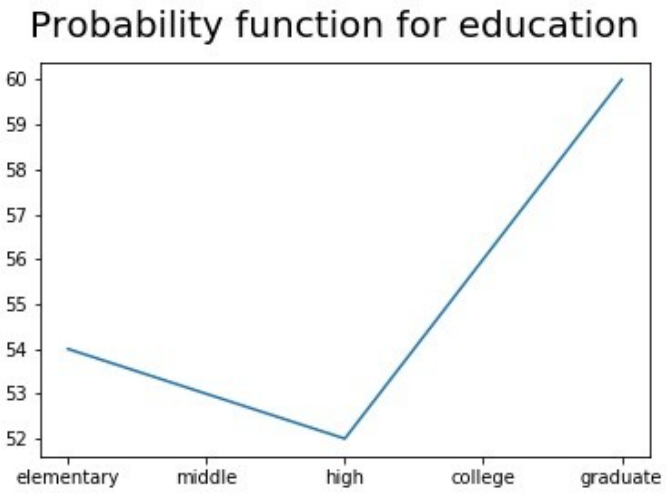

Figura 6: Probabilidad por educación

Esta vez la probabilidad por educación estuvo fuera de los valores esperados para universitarios y graduados de postgrado. Afortunadamente, el modelo multivariado pudo encontrar la respuesta correcta. Además los resultados fueron bastante cercanos. La simulación predijo 527908 votos para A y 472092 para b, que se acerca bastante a lo que en realidad pasó.

Esta herramienta no es perfecta, pero separar las probabilidades en muchas variables resulta ser un método efectivo y un buen ejemplo de regresión a la media en la ley de los grandes números.