Reporte de Proyecto Individual

CA0204. Herramientas para Ciencias de Datos I

Benjamín Gutiérrez Padua. C4F813

En el presente trabajo se planteó como objeto de investigación el estudio de la aleatoriedad en los resultados del Sorteo del Premio Mayor del "Gordo Navideño", aprovechando que nos encontramos en vísperas de la realización de dicho sorteo y articulando, de manera aplicada, los contenidos vistos en el curso de Probabilidad (CA0721) y en el curso de Herramientas para Ciencias de Datos I (CA0204), ambos cursados en el presente ciclo lectivo.

El objetivo central del trabajo ha consistido en integrar herramientas probabilísticas y de análisis de datos para responder a la siguiente pregunta de investigación:

> **¿Los resultados del sorteo extraordinario de Navidad "Gordo Navideño" de la Junta de Protección Social se comportan como realizaciones de un proceso aleatorio puro, o muestran indicios de patrones estadísticos recurrentes?**

En otras palabras, se busca someter a prueba la hipótesis de que los dígitos que componen el resultado ganador (serie y número) se comportan como variables aleatorias discretas, aproximadamente uniformes e independientes, tal como cabría esperar de un mecanismo de azar bien diseñado. Para ello se requiere contar con una herramienta probabilística que permita contrastar formalmente esta hipótesis a partir de los datos históricos.

En este contexto se identificó la prueba chi-cuadrado como una herramienta adecuada, tanto en su versión de bondad de ajuste (para evaluar si la distribución de los dígitos es compatible con una distribución uniforme discreta) como en su versión de independencia (para evaluar si, por ejemplo, los dígitos de decena y unidad del número de dos dígitos se comportan de manera independiente). En

ambos casos, el estadístico de prueba (X)exp(2) compara las frecuencias observadas con las frecuencias teóricas esperadas bajo la hipótesis nula.

De forma general, el procedimiento de aplicación de la prueba chi-cuadrado puede resumirse en los siguientes pasos:

1. **Formulación de hipótesis.**

   Se plantea una hipótesis nula H0 y una hipótesis alternativa H1.

   - En el caso de bondad de ajuste, H0suele afirmar que los datos siguen una distribución teórica dada (por ejemplo, uniforme discreta en {0,…,9}.
   - En el caso de independencia, H0 establece que dos variables categóricas son independientes entre sí.

2. **Cálculo del estadístico X2.**

   A partir de las frecuencias observadas y las frecuencias esperadas bajo H0, se calcula el estadístico chi-cuadrado, que mide el tamaño de la discrepancia entre lo observado y lo que se esperaría si H0 fuera cierta.

3. **Obtención del p-valor y decisión.**

   Bajo H0, el estadístico (X)exp(2) sigue (aproximadamente) una distribución chi-cuadrado con ciertos grados de libertad. El p-valor se interpreta como la probabilidad, suponiendo que H0 es verdadera, de obtener un valor del estadístico al menos tan extremo como el observado.

   - Si el p-valor es menor que el nivel de significancia α (por ejemplo, 0.05), se rechaza H0.
   - Si el p-valor es mayor que α, no se rechaza H0, y se concluye que los datos son compatibles con el modelo planteado (por ejemplo, uniforme o independiente).

En el contexto del presente proyecto, un p-valor grande indica que las discrepancias entre las frecuencias observadas y las esperadas no son lo suficientemente grandes como para considerar que se apartan de lo que produciría un mecanismo de azar puro; es decir, los datos resultan compatibles con la uniformidad y la independencia según el caso.

Cabe resaltar que para la obtención del data set se tuvo que solicitar por correo electrónico a la Junta de Protección Social (JPS) dado que los mismos no se encontraban en el apartado de “datos abiertos” de su plataforma, ya una vez obtenidos los datos se debió realizar un tratamiento de los mismos, no tanto por el control de valores ausentes, ni de outliers (que imputarlos hubiese introducido un sesgo de investigación) , si no porque el formato en el que se compartió se debió de transformar a un csv y luego a las variables de interés de estudio cambiarles en algunos casos su estructura para poder aplicarles adecuadamente funciones de R y sus librerías.

Respecto a este último aspecto se realizó uso de las librerías **readr**, **dplyr**, **tidyr** y **ggplot2** para la lectura de datos, su manipulación y reorganización respectivamente, y por supuesto para la generación de las gráficas para el componente del Análisis Exploratorio de Datos (EDA), dinámica por la que se optó representar las frecuencias absolutas para cada una de las variables aleatorias identidficas en el trabajo, a saber cada uno de los dígitos que corresponde al resultado del ganador del premio mayor.

Esta representación resultó ser sumamente beneficiosa pues permitió representar de manera intuitiva la distribución de las realizaciones materializadas para cada variable aleatoria durante todos los sorteos realizados, a su vez se aplicó por periodos temporales de 5 años (ventanas temporales) para reconocer si habían patrones particulares de ocurrencia de ciertos valores, de lo cual no se observó nada más allá de las caracterísitcas de la aleatoriedad per se del evento.

De este modo, realizadas las pruebas descritas se determinó que el Sorteo del Premio Mayor del Gordo Navideño en Costa Rica, obedece a patrones plausibles de aleatoriedad pura.