Universidad de Sonora

División de Ciencias Exactas y Naturales

Maestría en Ciencias de Datos

"El saber de mis hijos
hará mi grandeza"

Proyecto final de Probabilidad y Estadística: Aplicación de la Ley de Benford para Análisis de Facturas

Alumno
Fernando Luna Ponce

Profesor:
Dra. Gudelia Figueroa Preciado

Hermosillo, Sonora, 6 de diciembre de 2022

Contenido

1. Antecedentes

La ley de Benford o ley del primer dígito asegura que, en el mundo real, el 1 aparece como primera cifra con mucha más frecuencia que el resto. Además, cuanto mayor es el dígito, menos probable es que se encuentre en primera posición. [1]

En 1881 el astrónomo y matemático Simon Newcomb descubrió que los libros donde consultaba tablas de logaritmos tenían más sucias las páginas con los números que empiezan por 1. Posteriormente el físico Frank Benford recuperó las investigaciones de Newcomb y saco a la luz esta propiedad, que nos a llegado con el nombre de LEY DE BENFORD. [2]

Newcomb incluso dedujo que la probabilidad de que un número empiezé por la cifra C, se puede calcular con la siguiente fórmula:

$$P = log(1 + 1/C) \tag{1}$$

En la siguiente tabla se muestra la distribución de cada número.

Clase	Probabilidad
1	32.13 %
2	17.60 %
3	12.49 %
4	9.69 %
5	7.91 %
6	6.69 %
7	5.79 %
8	5.11 %
9	4.57 %

Tabla 1: Distribución de probabilidades por cada dígito

¿Y qué tiene que ver todo esto con la investigación policial? Dado que la ley de Benford se cumple con cualquier colección de números aleatorios, también debería cumplirse con los importes de cobros y pagos de una empresa, así lo supuso el Dr.Mark J. Nigrini de la Universidad de Kansas, y empezó a llevar a cabo investigaciones usando este principio matemático, obteniendo rapidamente resultados.

Gracias a las investigaciones de todos estos científicos, la policía cuenta ahora con una nueva herramienta para detectar el fraude fiscal, con la que ya han conseguido detectar estafas en muchas empresas.

La prueba del primer dígito permite probar la razonabilidad de las cifras, la regla indica que si en esta prueba existen diferencias respecto a la tabla de frecuencias propuesta por la Ley de Benford, es probable que la información contenga riesgos de error o fraude por duplicaciones y anomalías. Esta prueba no intenta establecer muestras de auditoría. [3]

2. Objetivos

- Procesar nuestro conjunto de datos para obtener la información necesaria para comprobar la Ley de Benford.
- Realizar la prueba de bondad de ajuste χ^2 para verificar si nuestro conjunto de datos sigue una distribución de Benford.

3. Descripción de variables

Sea X una variable aleatoria discreta sobre [1,10), obtenida al tomar el primer dígito de la cantidad monetaria de cada elemento de una serie de facturas.

4. Descripción de la fuente de datos

Para efectuar el análisis se utiliza el conjunto de datos del Consejo de Transparencia y Protección de Datos de Andalucía. Esta institución es la que organiza el autogobierno de la Comunidad Autónoma de Andalucía, en España. Está integrada por el Parlamento de Andalucía, la presidencia de la Junta de Andalucía y el Consejo de Gobierno.

Este conjunto de datos se encuantra en formato de archivo separado por comas (csv) el cual contiene el listado de facturas por año desde el 2016 hasta el 2022, con un total de 2132 facturas. Para nuestro análisis solo es necesaria la columna de importe total, de esta solo se va extrae el primer dígito.

5. Aplicación de la Ley de Benford

Se estará analizando nuestro lista de facturas para ver si existen irregularidades, para lo cual se estará utilizando la prueba de Bondad de Ajuste χ^2 para determinar si se sigue o no una distribución de Benford.

Para realizar el procesamiento de nuestros datos se utilizará una libreta de jupyter. Lo primero a realizar es cargar nuestro archivo csv en un dataframe de pandas para realizar el procesamiento y poder obtener nuestra variable de interés. Después de generar nuestro dataframe nos quedamos solo con la columna de importe total, después extraemos el primer dígito de cada unos de los elementos. Posteriormente se utilizan algunas funciones de agregación para obtener cuantas veces aparace cada uno de los dígitos del 1 al 9 en nuestro conjunto de datos.

En la tabla siguiente podemos apreciar la frecuencia absoluta para cada uno de los dígitos, es decir la cantidad de veces que repite cada dígito en la primer posición de nuestros datos, también observamos la frecuencia relativa que resulta de dividir la frecuencia absoluta de cada dígito entre el total de elementos.

Digito	Frec. Absoluta	Frec. Relativa
1	684	0.320826
2	407	0.190901
3	260	0.121951
4	169	0.079268
5	114	0.053471
6	141	0.066135
7	123	0.057692
8	79	0.037054
9	154	0.072233

Tabla 2: Ocurrencia de cada dígito en nuestro conjunto de datos

Para poder hacer una comparación, necesitamos calcular la ocurrencia esperada de cada dígito, lo cual se obtiene multiplicando la probabilidad de ocurrencia de cada dígito por el total de elementos registrados, que son 2132.

Los resultados se observan en la siguiente tabla.

Digito	Ocurrencia esperada
1	641.795951
2	375.426564
3	266.369386
4	206.612148
5	168.814417
6	142.730555
7	123.638831
8	109.057178
9	97.554970

Tabla 3: Ocurrencia esperada de cada dígito

Una vez que obtuvimos la ocurrencia esperada y la frecuencia absoluta, procedemos a realizar un gráfica para comparar el valor esperado contra la frecuencia absoluta.

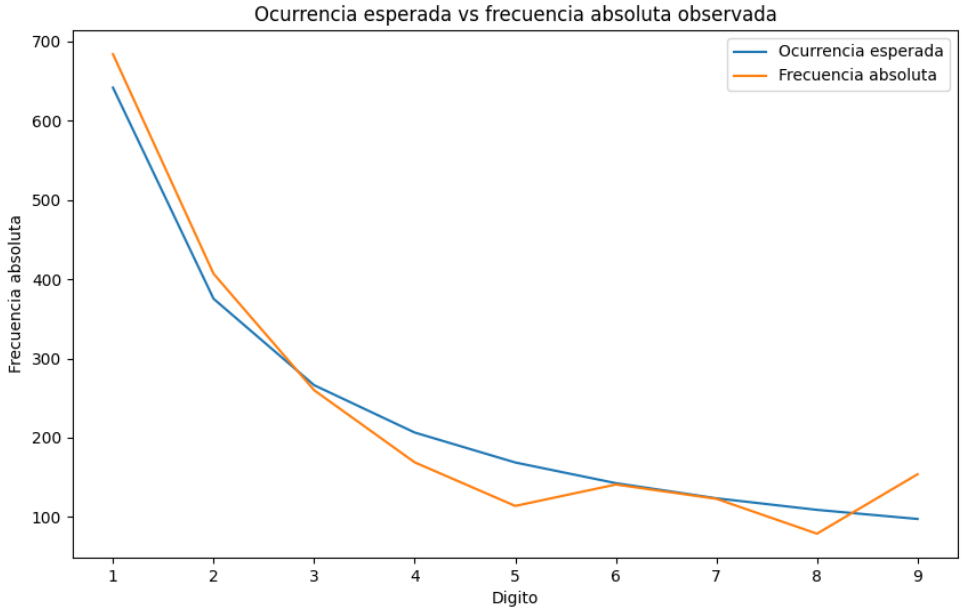

Figura 1: Comparación real vs esperado

Como se puede observar en la gráfica, se observan irregularidades en el valor real comparado con el valor esperado, por lo que se realizará la prueba de Bondad de Ajuste χ^2 para corroborar que efectivamente se trata de una anomalía.

5.1. Prueba de Bondad de Ajuste Chi-Cuadrada χ^2

El test de bondad de ajuste de Chi Cuadrado χ^2 permite comprobar si ciertos datos siguen una cierta distribución de probabilidad con un cierto error α. Para el caso de la Ley de Benford para el primer dígito, tendremos 8 grados de libertad y si buscamos hacer el test con un error del 5 %, entonces se aceptará que los datos siguen la Ley de Benford si $\chi^2 < 15{,}51$ y se rechazará en otro caso. [4]

Se realizó el cálculo en la libreta jupyter, los resultados arrojan una $\chi^2 = 71{,}19$ por lo que se estaría rechazando la hipótesis de que nuestros datos tienen una distribución de Benford.

6. Conclusiones

La ley de Benford debería de seguirse en el caso de las facturas, en nuestro conjunto de datos se observa que no se cumple, esto nos indicaría una posible alteración en las facturas, pero no es un indicio totalmente concluyente, por lo que sería necesario hacer más pruebas como la prueba del segundo dígito, los primeros dos dígitos o la prueba de los últimos dos dígitos. Encontrar hallazgos importantes en estas pruebas nos indican que posiblemente haya anomalías que se tienen que analizar mediante una auditoría.

Referencias

[1] Ley de benford: la fuerza de uno. https://www.ull.es/portal/cienciaull/ley-de-benford-la-fuerza-de-uno/. fecha consulta: 2021-12-06.

[2] La ley de benford. http://www.mat.ucm.es/cosasmdg/nuevos/Benford.html. fecha consulta: 2021-12-06.

[3] Ley de benford, técnica para el análisis de información. https://1library.co/article/ley-benford-t%C3%A9cnicas-an%C3%A1lisis-informaci%C3%B3n.zxv26koy. fecha consulta: 2021-12-06.

[4] La ley de benford. http://repositorio.cfe.edu.uy/bitstream/handle/123456789/247/Caputi%20Zunini,%20Maria%20,%20Ley.pdf?sequence=1. fecha consulta: 2021-12-06.