UNIVERSIDAD ICESI

Departamento de Economía

Facultad de Ciencias Administrativas y Económicas

Economics Lecture Notes

Introducción a la Ley de Benford para la detección de fraudes empleando Tableau Public

Julio Cesar Alonso Cifuentes

Icesi ECONOMICS LN No.8
Septiembre 2019

Introducción a la Ley de Benford para la detección de fraudes empleando Tableau Public

Julio Cesar Alonso Cifuentes

Icesi
ECONOMICS LN
No.8
Septiembre 2019

Universidad Icesi

Editor:
Carlos Giovanni González Espitia
Profesor tiempo completo, Universidad Icesi
cggonzalez@icesi.edu.co

Asistente editorial:
Laura María Otero López

Gestión Editorial
Departamento de Economía - Universidad Icesi

www.icesi.edu.co
Tel: 5552334. Fax: 5551441
Calle 18 # 122-135 Cali, Valle del Cauca, Colombia

Introducción a la Ley de Benford para la detección de fraudes empleando Tableau Public

Julio Cesar Alonso Cifuentes[*]

Contenido

Objetivos de Aprendizaje

Al finalizar la lectura de este documento se espera que el lector esté en capacidad de:

- Explicar en sus propias palabras que es la Ley de Benford y cómo puede ser empleada para la detección de fraudes.
- Por medio Tableau Public, determinar si un conjunto de datos cumple o no con la Ley de Benford

[*] Director del Centro de Investigación en Economía y Finanzas. Profesor Departamento de Economía, Universidad Icesi, Cali, Colombia. Jcalonso@icesi.edu.

1 Introducción

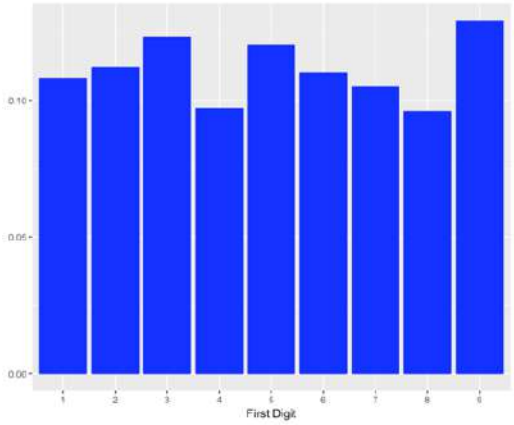
La detección de fraudes es en sí una disciplina a la cuál muchos contadores, estadísticos y economistas dedican su tiempo. Existen muchas herramientas para detectar fraudes o anomalías en un conjunto de datos que van desde modelos de regresión logística, hasta modelos de redes neuronales. Pero tal vez, una de las herramientas más sencillas y más usadas para encontrar fraudes en registros es la ley de Benford. La ley de Benford también es conocida como la ley de los Primeros Dígitos o el Fenómeno de los Dígitos Significativos.

Ésta establece un patrón que debe seguir el primer digito de una muestra grande de datos si estos son generados por un proceso "legitimo". Por ejemplo, la ley de Benford permite determinar si los encuestadores se han inventado respuestas al momento de recolectar la información de una encuesta, si un funcionario se dedicó a generar facturas sin soporte o si las declaraciones de aduana de un importador están siendo alteradas.

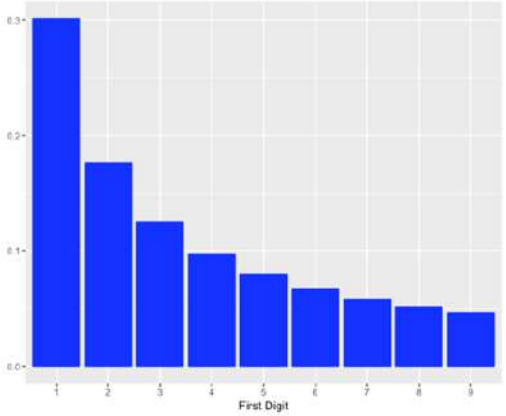
En términos sencillos, esta ley implica que el primer dígito de los números encontrados (registros) en bases de datos de fuentes diversas no muestran una distribución uniforme como intuitivamente se esperaría (Ver Figura 1, Panel a), sino que se organizan de tal manera que el dígito "1" es el más frecuente (30.1% de los registros), seguido de "2" (17.6%), "3" (12.5%), y así sucesivamente hasta "9" (4. %) (Ver Figura 1, Panel b). Es decir, se encuentra que aproximadamente el 30% de los registros deben iniciar por un "1", el 17. % por un "2" y así sucesivamente.

Figura 1. Posibles Distribuciones del primer dígito de los números de una base de datos grande.

Panel a: Distribución aproximadamente uniforme del primer dígito.



Panel b: Distribución "esperada" por la Ley de Benford para el primer.



Fuente: Cálculos Propios

La ley de Benford establece que la probabilidad de observar como primer digito a d (donde $d = 1,2,\cdots 9$) será:

$$P(d) = \log_{10}(d+1) - \log_{10}(d) = \log_{10}\left(\frac{d+1}{d}\right) = \log_{10}\left(1+\frac{1}{d}\right)$$

Esto implica la probabilidad de que ocurra cada uno de los 9 dígitos reportados en el Cuadro 1.

Cuadro 1. Probabilidad de ocurrencia del primer digito (P(d)) de acuerdo con la Ley de Benford

d	P(d)
1	30.1 %
2	17.6 %
3	12.5 %
4	9.7 %
5	7.9 %
6	6.7 %
7	5.8 %
8	5.1 %
9	4.6 %

Pero antes de entrar en el detalle de cómo hacer esto en Tableau Public, hablemos un poco de su historia. El primer registro sobre el tema se encuentra en un documento del astrónomo y matemático Simon Newcomb (Newcomb, 1881). La historia cuenta que Newcomb, mientras hojeaba páginas de un libro de tablas logarítmicas, notó que las páginas al principio del libro estaban más sucias que las páginas al final. Esto significaba que sus colegas, que compartían la biblioteca, preferían cantidades que comenzaban con el número uno en sus diversas disciplinas. El documentó esa regularidad en dos página publicadas en el American Journal of Mathematics en 1881 (Ver Newcomb (1881)), pero no presentó una demostración de por qué la regularidad debía cumplirse.

En 1938, el físico estadounidense Frank Benford revisó el fenómeno, y encontró esa misma regularidad y la llamó la "Ley de Números Anómalos" (Benford, 2012). Benford publicó en la revista científica titulada "Proceedings of the American Philosophical Society" un estudio en el que empleó 20,000 observaciones de datos recopilados de diversas fuentes para documentar la regularidad. Los datos empleados provenían de diferentes fuentes; desde caudales de ríos hasta pesos moleculares de compuestos químicos.

También consideró información de costos, números de direcciones, tamaños de población y constantes físicas. Todos ellos, en mayor o menor medida, siguieron una distribución decreciente de manera exponencial. Este físico no solo documentó la regularidad, sino que también presentó una justificación del porqué ocurre; pero no se presentó una demostración de ésta.

Posteriormente, en 1995, el fenómeno fue finalmente comprobado de manera estadística por el profesor de matemáticas de la Universidad de Gergia Tech Theodore P. Hill ((Hill, 1995)). El profesor Hill presentó una demostración estadística que da sustento al porqué se cumple esta regularidad. El encontró que, si se generan datos de manera legítima que se recolectan en bases de datos y se toman muestras aleatorias de estas, la distribución del primer dígito de la muestra convergerá a la distribución propuesta por Benford.

Se ha demostrado que este resultado se aplica a una amplia variedad de conjuntos de datos, incluidas facturas de electricidad, direcciones, precios de acciones, precios de viviendas, números de población, tasas de mortalidad, longitudes de ríos, etc. En general, se ha visto una serie de registros numéricos que siguen la ley de Benford cuando representa magnitudes de eventos o eventos, tales como poblaciones de ciudades, flujos de agua en ríos o tamaños de cuerpos celestes.

Sin embargo, se ha encontrado que, si los registros tienen límites mínimos o máximos preestablecidos, entonces esta ley no se cumple. Tampoco se cumple cuando los registros representan identificaciones, como números de identidad o de seguridad social, cuentas bancarias, números de teléfono.
Finalmente, los registros se ajustan mejor a esta ley si provienen de una distribución que tiene una media menor que la mediana, y los datos no se concentran alrededor de la media.

En la actualidad esta ley es empleada en diferentes áreas detectar patrones (o la falta de ellos) en conjuntos de datos. Esto puede llevar a aplicaciones importantes en la ciencia de los datos, como detectar anomalías o detectar fraudes.

Para terminar esta introducción, es importante mencionar que, así como existe una distribución para el primer dígito, también existe otra para el segundo, tercero y los demás dígitos. Esa distribución es diferente y por tanto la Ley de Benford estudiada no aplicaría para los dígitos a partir del segundo.

En lo que resta de este documento emplearemos la Ley de Benford en Tableau Public para detectar si existe algún tipo de anomalía o fraude en las facturas que se reportan en los pedidos que se presentan en la base de datos de facturas recibidas por una división de la empresa de servicios públicos West Coast. Los datos se encuentran en el archivo adjunto “datosBenford.txt” y provienen de Nigrini (2012).

Esta base de datos tiene los registros de 184,412 facturas por pagar por parte de la empresa de servicios públicos West Coast a sus proveedores. Los datos corresponden al 2010. Las variables disponibles son:

- "registro" consecutivo de la factura
- "VendorNum" número del proveedor
- "Date" fecha de emisión de la factura
- "Amount" valor de la factura.

Nuestra misión será detectar si existen o no anomalías en el valor facturado (la variable Amount).

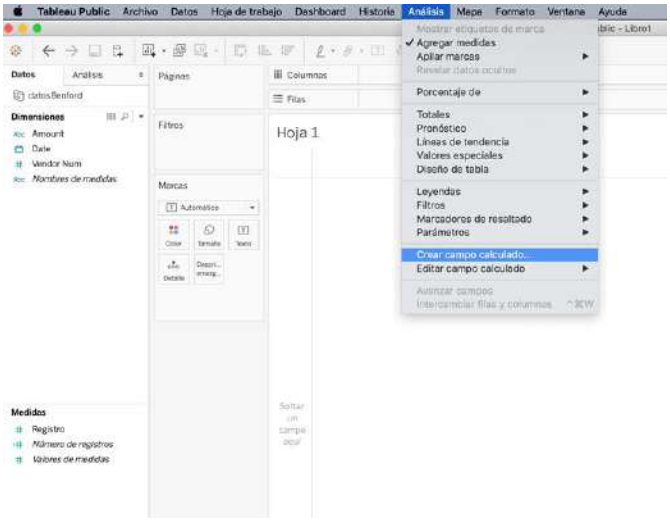
2 Aplicación en Tableau

Para responder la pregunta de si existen o no sospechas de anomalías o fraude en la información del valor facturado (la variable Amount) debemos graficar la distribución de los primeros dígitos de todas las facturas registradas (184,412) y compararla con la predicción de la Ley de Benford.

Para iniciar abramos Tableau Public y carguemos los datos del archivo adjunto "datosBenford.txt". Para esto, haga clic en el icono de Tableau y posteriormente en conectar a un "archivo de texto". Verá una ventana emergente en la que debe seleccionar la ubicación del archivo y haga clic en "abrir".

Concentrémonos en la variable del valor facturado (la variable Amount). Como queremos determinar si existe algún tipo de fraude, miremos cómo se comporta el primer dígito de estas ventas. Para esto tendremos que crear un campo calculado (**Calculated Field**). El primer paso será hacer clic en la Hoja 1 (parte inferior izquierda) e ir al menú de *Análisis* (Analysis) y escoger la opción de "*Crear campo calculado…*" (Create Calculated Field). (ver siguiente figura).

Gráfico 1:



Verá una ventana en la que podemos crear la variable calculada que deseamos.

En este caso tenemos que extraer el primer dígito de cada uno de los registros de las ventas. Es decir, debemos encontrar el dígito más a la izquierda de cada número. Llamemos a esta nueva variable "*Leftmost Integer*" y calculémosla empleando la siguiente fórmula:

LEFT (STR ([Amount]),1)

El nombre de la nueva variable debe colocarse en la primera ventana y la fórmula en el espacio más grande (ver siguiente figura). Noten que si no conocen el nombre de una función que se quiera aplicar, pueden hacer clic en el triángulo que se observa en el lado derecho de la ventana. Esto desplegará un menú de ayuda con las funciones disponibles. Haga clic en el botón de aceptar.

Gráfico 2:

Creemos un segundo campo que tenga la probabilidad de observar cada uno de los dígitos (d) de acuerdo con la ley de Benford ($P(d)$). Es decir, empleando la siguiente distribución de probabilidad:

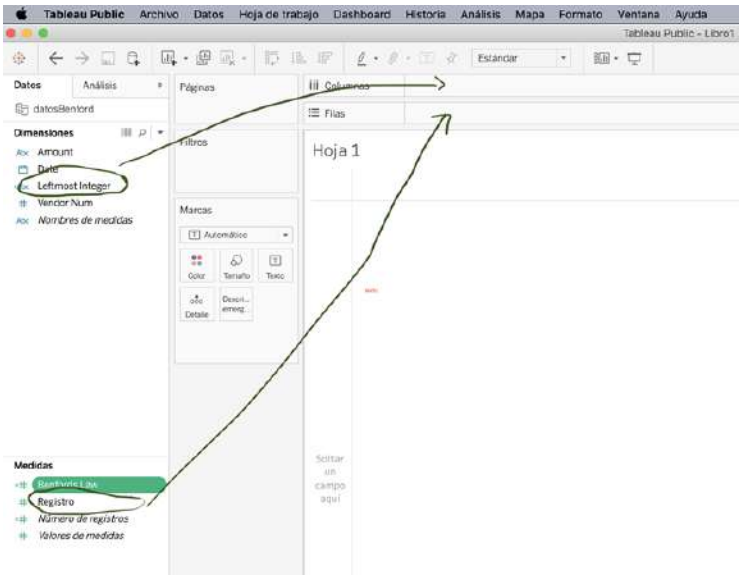
$$P(d) = \log_{10}(d+1) - \log_{10}(d)$$

El segundo campo se puede construir empleando la función:

```
LOG (INT ([Leftmost Integer]) +1)-LOG (INT ([Leftmost Integer]))
```

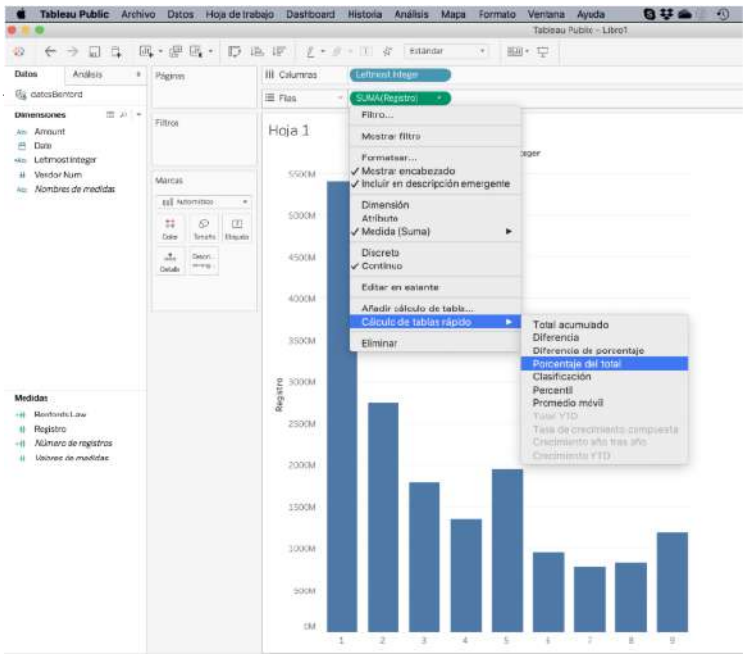
Y denominémoslo *"Benfords Law"*. A continuación, debemos graficar. Arrastre "Leftmost Integer" del **Área de dimensiones** del panel de datos al campo de columnas y el "Número de registros" (Number of Records) del **Área de medidas** al campo de filas (ver siguiente figura).

Gráfico 3:



Haga clic en “SUMA (Número de Reg…” que está en verde y seleccione “Cálculo de tablas rápido” y “Porcentaje del total” (ver siguiente figura).

Gráfico 4:

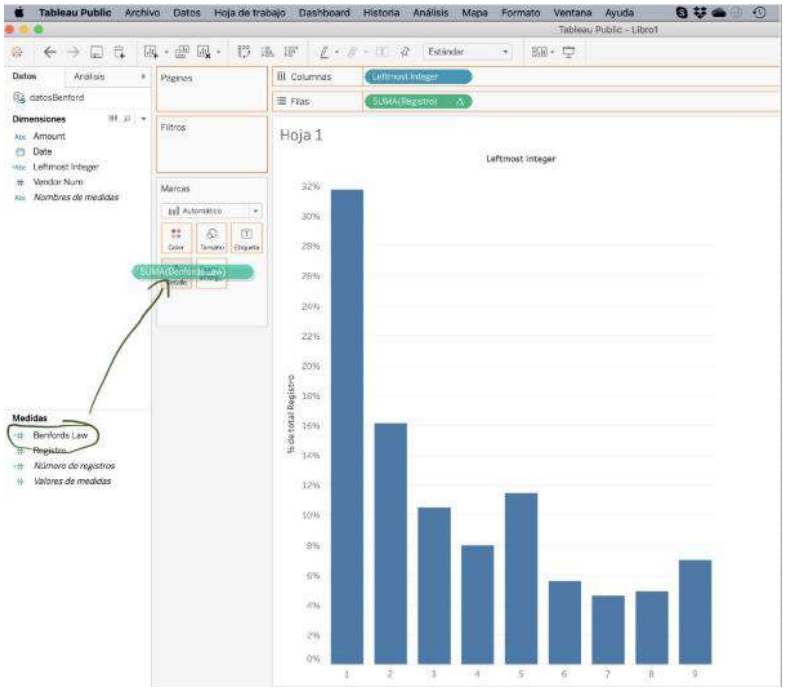


Después en la Hoja 1 tenemos la distribución de los primeros dígitos. Noten que las barras van disminuyendo de tamaño para los dígitos más grandes; es decir, se observa una mayor frecuencia en los dígitos menores y una menor en los dígitos mayores. Pero se presentan tres excepciones con los dígitos 5, 8 y 9. Esto sugiere que los datos en este caso no se podrían estar ajustando a la ley de Benford. Pero la decisión lo deberíamos tomar a "ojo".

El siguiente paso implicará sobreponer la distribución sugerida por la ley de Benford.

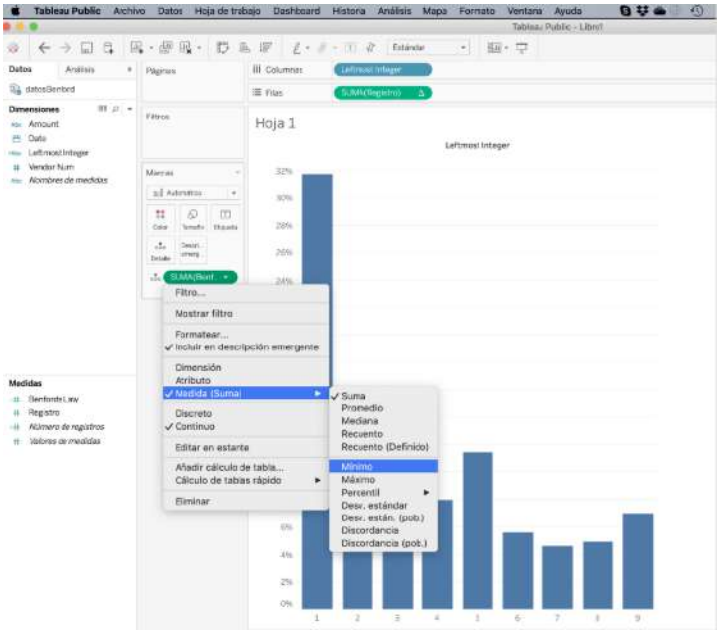
Arrastre "*Benfords Law*" del **Área de medidas** de la pestaña de datos a "Detalle" que se encuentra el **Tarjeta de Macas** (ver siguiente figura).

Gráfico 5



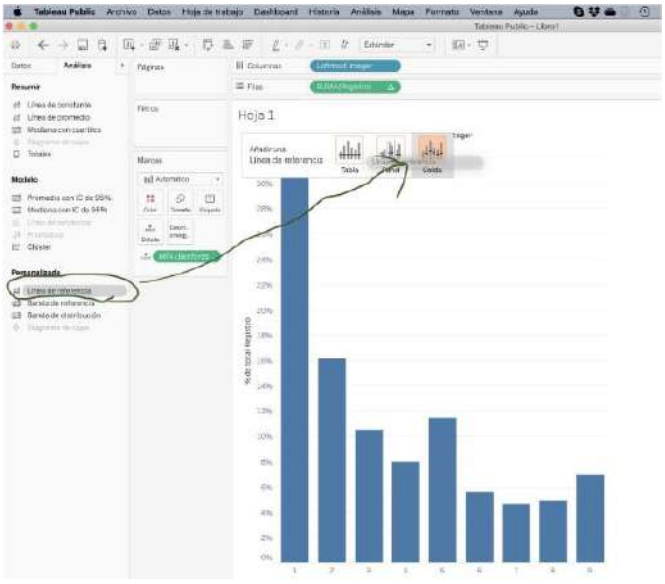
Posteriormente podemos hacer clic en "Benfords Law" de la **Tarjeta de Macas** (está en verde) y escoja "Medida" y posteriormente "Mínimo" (ver siguiente figura).

Gráfico 6:



Ahora pase de la pestaña de **Datos** a la de **Análisis** y arrastre "*Línea de referencia*" al área de vista (encima del gráfico). Observará que aparece una ventana, suelte la "*Línea de referencia*" en "*Celda*" (ver siguiente figura).

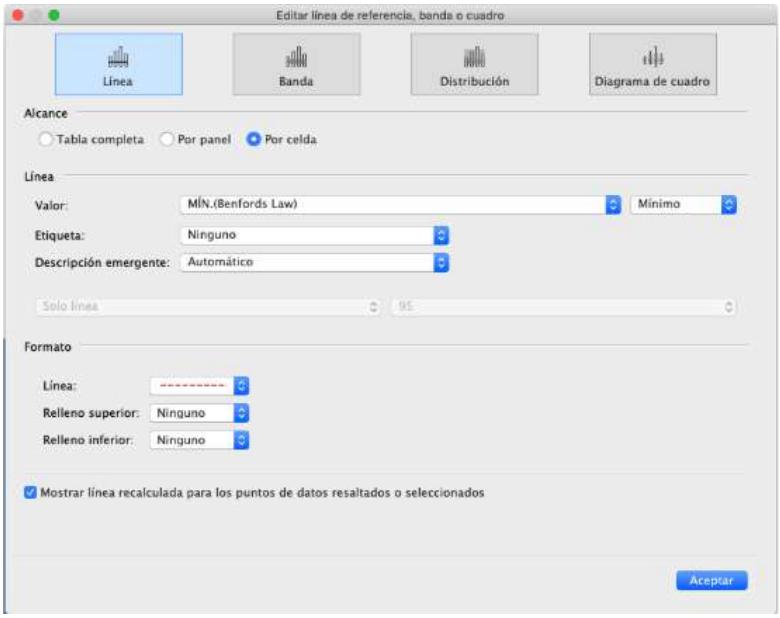
Gráfico 7:



A continuación, verá una ventana titulada “Editar línea de referencia, banda o cuadro”, en esa ventana haga clic en el campo de “*Valor*”. Esta ventana (ver siguiente figura) nos permite trazar una línea con la proporción que predice la ley de Benford.

En el campo de “Valor” escoja “*MIN (Benfords Law)*” y en seguida escoja “*Mínimo*” (ver siguiente figura).

Gráfico 8:

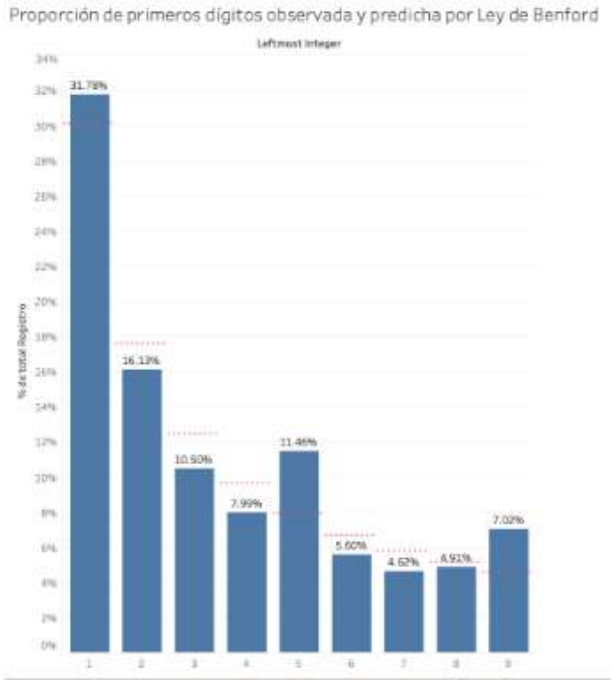


Ahora, terminaremos el gráfico. Escoja en el campo “Línea” la opción más delgada y la forma que desee, así como el color (ver figura anterior). Si desea ver los valores de las barras, haga clic en


.

La Figura 2 muestra nuestro resultado.

Figura 2. Histograma del primer dígito observado



Si bien los valores observados están muy cerca a los predichos por la ley de Benford (*pesperado=P*(*d*)) representados por la línea punteada roja, es importante hacer una prueba estadística que permita tomar la decisión.

Una forma de construir un intervalo de confianza del 99% para la proporción esperada es emplear el siguiente límite superior:

$$p_{esperado} + 2.58 \cdot \sqrt{\frac{p_{esperado}(1 - p_{esperado})}{n} + \frac{1}{2 \cdot n}}$$

Y el límite inferior es:

$$p_{esperado} + 2.58 \cdot \sqrt{\frac{p_{esperado}(1 - p_{esperado})}{n} + \frac{1}{2 \cdot n}}$$

donde *n* es el número de observaciones (184,412 en este caso). *Si se desea construir un intervalo con un nivel de confianza del 95%, entonces será necesario cambiar el número 2.58 por 1.96.*

Ahora creemos este intervalo. Para esto debemos construir un campo calculado con el límite superior y otro con el límite inferior. Para el límite superior usemos el nombre "upperBonford" y la fórmula:

```
[Benfords Law] +2.58 * SQRT ([Benfords Law] *(1- [Benfords
Law])/ 184412) + (1/ (2*184412))
```

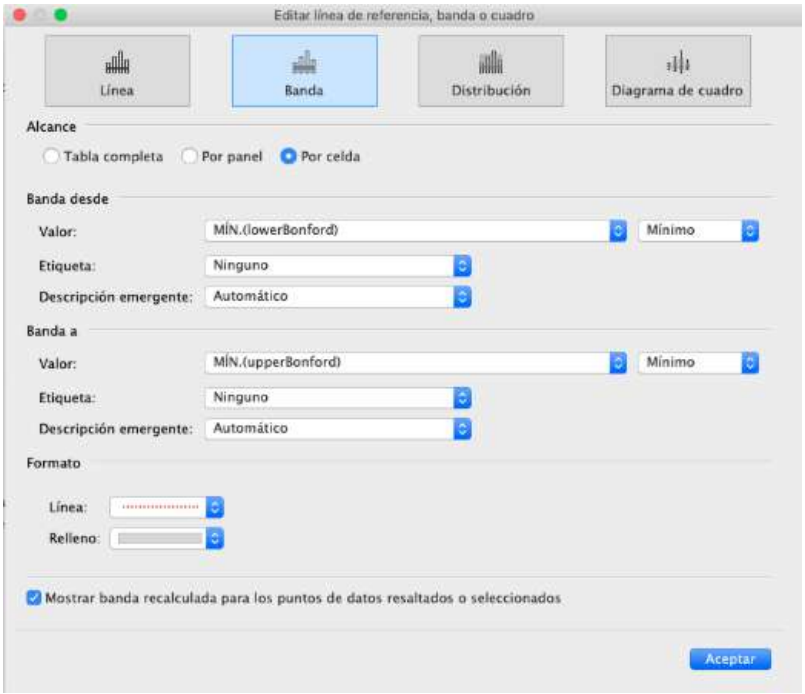
Y para el límite inferior, el nombre "lowerBonford" y la fórmula:

```
[Benfords Law] - 2.58*SQRT ([Benfords Law] *(1- [Benfords
Law])/ 184412) + (1/ (2*184412))
```

Finalmente grafiquemos estos intervalos de confianza. Arrastremos tanto "upperBonford" y "lowerBonford" del **Área de medidas** de la pestaña de datos a "Detalle" que se encuentra el **Tarjeta de Marcas**. Para ambos casos cambie (cómo lo hizo antes) la suma por el Mínimo. Esto se logra haciendo clic en cada una de estas medidas en la **Tarjeta de Marcas** (están en verde) y escoja "Medida" y posteriormente "Mínimo".

Ahora pase de la pestaña de **Datos** a la de **Análisis** y arrastre "*Banda de referencia*" al área de vista (encima del gráfico). Observará que aparece una ventana, suelte la "*Banda de referencia*" en "*Celda*". Veremos la siguiente ventana.

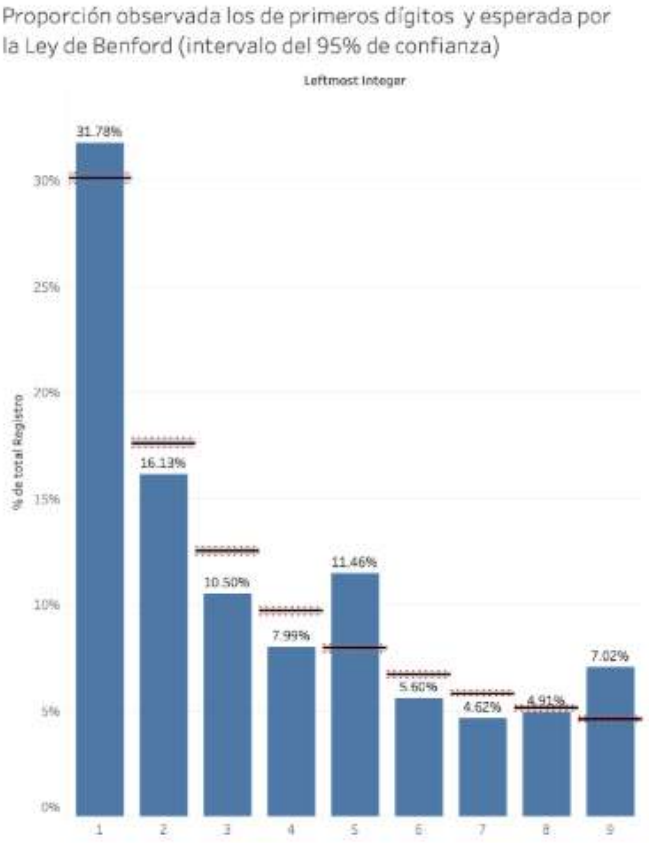
Gráfico 9:



Con este paso estamos construyen la parte inferior y superior del intervalo para la predicción de la proporción de la ley de Benford. En el campo “Valor” de la sección “Banda desde” escoja “MIN. (lowerBonfor)” al lado “Mínimo” y en la sección “Banda a” en “Valor” escoja “MIN. (upperBonford) (ver figura anterior).

En la siguiente figura observamos la figura 3.

Figura 3. Resultado final del gráfico realizado en Tableau Public



Noten que todos los intervalos no contienen el valor observado. Es decir, con un 99% de confianza es posible rechazar que la proporción observada es igual a la predicha por la ley de Benford. Es decir, hay indicios de anomalías o fraudes en el registro de la información.

3. Referencias

Benford, F. (2012). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, *78*(4), 551–572.

Hill, T. P. (1995). A Statistical Derivation of the Significant-Digit Law. *Statistical Science*, *10*(4), 354–363. https://doi.org/10.2307/2246134

Newcomb, S. (1881). Note on the Frequency of Use of the Different Digits in Natural Numbers. *American Journal of Mathematics*, *4*(1), 39–40. Retrieved from http://www.jstor.org/stable/2369148

Nigrini, M. J. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection (Vol. 586). John Wiley & Sons.*