Resumen

En el presente artículo se realizan dos modelos matemáticos programables que pueden asistir en el análisis de auditoría, mediante la selección de las muestras de escrutinio más prometedoras, estas muestras más potencialmente significativas son escogidas de manera optimizada teniendo como referencia la distribución natural (fidedigna) de los dígitos, la cual se comprende como la ley de Benford, conocida también como la ley de los dígitos significativos.

A propósito de esta ley, en el presente trabajo escrito se dedica un apartado en realizar la reconstrucción de identidad de manera formal, para esto mismo se hace uso de las propiedades de logaritmos y la acotación de x-valores en un intervalo, derivando formalmente la propiedad, suponiendo así un motivación matemática del análisis de la ley de Benford. En simultáneo se señalan las limitaciones de esta ley, la cual se satisface para conjuntos de datos no tan restrictivos y que satisfagan la propiedad heterogénea, se extrapola la distribución de Benford a la probabilidad para el análisis de conjuntos de datos aplicada en variables indicadoras de eventos, justificando mediante la ley de los grandes números (LGN) la convergencia a la media teórica (convergencia casi segura o en probabilidad, según versión de LGN) .

Análogamente se describen los estadísticos utilizados en el modelo, a saber: La prueba chi-cuadrado que suma los desajustes al cuadrado, prueba de Mean Absolute Deviation (MAD) que mide la discrepancia respecto a la muestra observada y Z-score para dìgitos individuales que mide cuántas desviaciones estándar un punto de datos está por encima (o debajo) de la media verdadera. Dado que si la distribución de Benford proporciona probabilidades teóricas, es necesario conocer qué tanto se desvían los conjuntos de datos de estudio de está misma.

A su vez se explora la construcción de estos estadísticos mediante el criterio del Teorema del Límite Central (TLC), donde se puede comprender la distribución chi-cuadrado como la suma N variables aleatorias independientes donde N_i para cada $i = 1, \dots, n$ siguen una distribución Normal$(\mu = 0,\ \sigma^2 = 1)$, MAD se aborda desde su definición como medida absoluta del desajuste respecto a la cantidad de categorías de la desviación entre la frecuencia relativa observada y la probabilidad teórica de Benford, mientras que los Z-score individuales comparan la frecuencia de un dígito observada (individual o par) con los esperados teóricos de Benford, en síntesis se construye todo un marco teórico-metodológico que persigue rescatar las desviaciones más significativas con el objetivo de prevenir la pérdida información relevante en las muestras a auditar.

Dado que el filtrado de las muestras de auditoría se propone realizar en dos estados, a saber mediante la identificación del subconjunto más pequeño de aquellos resultados más no conformes (es decir atípicos) de un conjunto de datos lo cual se obtiene mediante mediante la asignación de ciertos parámetros por parte del auditor, y la selección de los K registros menos conformes, no obstante el trabajo se centrará en explorar el primer método supracitado.

Resumidamente se puede definir la heurística general del proceso de la identificación del subconjunto de datos más pequeño no conforme, en donde teniendo de referencia las probabilidades teóricas de Benford y las muestras observadas se construyen los estadísticos de prueba previamente mencionados, y mediante el establecimiento de unas bandas de desviación razonables se obtiene la depuración del conjunto de datos fidedigno, que ha de satisfacer comportarse en distribución como la ley natural de los dígitos, es decir, se persigue con esta técnica el filtrado de el conjunto de datos hasta obtener uno que puede comprenderse como depurado y libre de alteraciones ya sea por error o fraudes, para esto mismo se establece una cota S(T), donde T es un experimento, la cual el auditor estipula de referencia con el objetivo de tener una banda de tolerancia respecto a qué tanto soporta la desviación observada de los datos respecto a las teóricas de Benford.

Paralelamente con el propósito de someter a prueba la fiabilidad del modelo y a modo de ejemplificación de la aplicación del mismo, daSilva y Carreira seleccionaron de manera aleatoria 30 conjuntos de datos de cinco mil registros de cuatro dígitos cada uno, con la finalidad de medir en diferentes niveles de contaminación (manipulación de los conjuntos de datos) el grado de efectividad del modelo, pruebas para las que resumidamente se concluyó para un nivel de contaminación de un 10% del total del conjunto de datos, el modelo de la detección del conjunto de datos no conforme más pequeño satisface una efectividad prudente para su implementación en contextos reales, como limitantes se tiene que si las contaminaciones son muy escazas (umbral del 2%) la prueba para la construcción del conjunto de datos depurados se ve ensuciada debido a las lígeras desviaciones producidas por el modelo.

Introducción

Finalidad del trabajo y Explicación de los Modelos Planteados.

El trabajo confeccionado por Gomes da Silva y Carreira surge con la finalidad de proponer un criterio técnico-racional eficiente para la selección de muestras en pruebas de auditoría, focal-

izado principalmente en instituciones financieras, la necesidad de la investigación de este tema surge debido a que los auditores requieren gestionar de forma eficiente el tiempo invertido para el estudio de los conjuntos de datos, más aún se requiere que el auditor sea capaz de realizar un filtrado efectivo del conjunto de datos pues el examinar transacción por transacción es una meta poco realista.

Este objetivo es alcanzado mediante la segmentación en dos bandas de los conjuntos de datos. Primeramente, se identifican las muestras de auditoría observando las propiedades del subconjunto de datos remanente después de retirar la muestra a auditar; esta muestra a auditar proviene del conjunto más pequeño de los datos que no siguen la conformidad deseada en los registros.

Con la finalidad de determinar qué registro t ha de ser examinado, es decir, removido del conjunto inicial de los primeros N registros, se define una variable aleatoria indicatriz:

$$y_t := \begin{cases} 1, & \text{si el registro } t \text{ es removido,} \\ 0, & \text{en otro caso,} \end{cases} \qquad t \in \{1, 2, 3, \dots, N\}.$$

Con esta variable aleatoria se introduce el tamaño del conjunto de conformidad dado por:

$$N - K, \quad \text{donde} \quad K := \sum_{t=1}^{N} y_t.$$

La idea tras esto, tal como señalan da Silva & Carreira (2013), es identificar en qué datos se encuentra la contaminación y removerla, dado que se persigue la finalidad de recuperar el conjunto de datos original. Una vez hecha esta separación, se toman los K registros más no conformes, dado un tamaño de muestra fijado por el auditor.

Explicación de la Ley de Benford

La particularidad del modelo proviene del filtrado para elegir qué conjuntos de datos están fuera de los parámetros de conformidad siguiendo la Ley de Benford. Esta ley, explicada de modo intuitivo, se puede ejemplificar de la siguiente manera:

Suponer que se toma una muestra de datos fidedignos y heterogéneos, es decir, muy diversos, como por ejemplo: el registro de natalidad por año de un país, el número de distritos por cantón en Costa Rica, o facturas de compras en un supermercado.

De forma intuitiva podría suponerse que cada dígito $\{1, 2, \dots, 9\}$ tiene un peso equiprobable

de ocurrencia, es decir:

$$P(X = i) = \frac{1}{9} \approx 0.111.$$

No obstante, esto no es cierto. En realidad, el dígito 1 ocurre con mucha mayor frecuencia que los demás; el dígito 2 ocurre más que los restantes siete, y así sucesivamente hasta el 9. Los pesos de Benford típicos son aproximadamente: 30% para el 1, 18% para el 2, y valores decrecientes hasta el 9.

La probabilidad teórica de que el primer dígito sea i, donde $i = 1, 2, \dots, 9$, está dada por:

$$P(\text{primer dígito } = i) = \log_{10}\left(1 + \frac{1}{i}\right).$$

Esta expresión tiene una justificación intuitiva debido a que muchos datos económicos se generan mediante productos o cocientes de variables aleatorias, como sucede con el interés compuesto.

Para hacer más natural esta explicación, considérese el caso en que el crecimiento poblacional de un país está modelado con una distribución exponencial $\text{Exp}(\lambda = 1.1)$. En los primeros años, pasar de 100 a 1000 personas tarda más tiempo que pasar de 1000 a 2000, por el crecimiento progresivo de la exponencial. Esto implica más observaciones con primer dígito 1, justificando intuitivamente la ley.

Justificación

El trabajo de investigación *Selecting Audit Samples Using Benford's Law* involucra amplios temas probabilísticos aplicados en la industria financiera. En particular, se vincula fuertemente con:

- la distribución de probabilidad de los dígitos de los datos,
- el uso de la distribución natural de Benford como referencia,
- la identificación de desviaciones respecto a esta distribución.

Aquellos conjuntos que se alejan significativamente de Benford se consideran eventos poco probables bajo el modelo, y por ende, deben ser auditados.

Simultáneamente, el trabajo integra herramientas probabilísticas y estadísticas para pruebas de conformidad, evaluando qué tan lejos se alejan los datos observados del comportamiento esperado. Esto incluye conceptos tales como:

- variables aleatorias,
- funciones indicadoras,
- funciones objetivo basadas en probabilidad,
- restricciones construidas mediante desviaciones estadísticas.

Todos estos conceptos se articulan con el propósito de resolver un problema real: optimizar la selección de muestras de auditoría en datos financieros potencialmente afectados por errores o fraudes.

Objetivos

El objetivo del trabajo de investigación abordado consisten en proporcionar una herramienta técnica eficiente para la recolección de muestras a auditar teniendo como referencia los parámetros de conformidad previamente definidos por el auditor, para este propósito se establecen cierto criterios para realizar el filtrado del conjunto de datos, bajo los principios de la ley de Benford, por esto mismo el trabajo persigue contribuir al análisis dígital mediante modelos matemáticos en la prevención del fraude, delimitando de manera más concreta se puede afirmar:

Objetivo general:

- Proporcionar una metodología técnica para la reconstrucción de un conjunto de datos fidedignos mediante la depuración de las muestras más anómalas que no siguen la ley de Benford para la detección de errores y/o fraudes.

Objetivos específicos:

- Establecer de forma justificada la razón de la toma de referencia de la distribución natural de Benford como un estándar de las muestras a auditar, siempre que satisfagan sus hipótesis.
- Definir los estadísticos de prueba más adecuados, a saber: prueba chi cuadrado, MAD y Z-score para la detección de muestras anómalas del conjunto de datos observado.

- Escrutar la efectividad del modelo mediante la ejecución de experimentos con el propósito de medir el rango de rendimiento óptimo del modelo propuesto.

Marco Teórico

Ley de Benford

Después de la breve introducción y motivación a la Ley de Benford, el siguiente paso es entender de manera formal el cómo se desenvuelve esta, así como por qué tiene sentido usarla como referencia.

Representación logarítmica

De esta manera, se va a detallar primero por qué el primer dígito de cualquier número se puede leer directamente de la parte decimal de su logaritmo en base 10, para posteriormente detallar una propiedad del comportamiento de esta parte decimal cuando miramos específicamente V.As. De esta forma tenemos que:

Todo número positivo x se puede escribir de forma única como

$$x = m \cdot 10^k,$$

con $k \in \mathbb{Z}$ y $m \in [1, 10)$. Note que el primer dígito de x viene únicamente de m (no de k): si m está entre 1 y 2, el primer dígito es 1; si está entre 2 y 3, el primer dígito es 2; y así sucesivamente.

Ahora si tomamos logaritmos en base 10 obtenemos

$$\log_{10} x = \log_{10}(m \cdot 10^k) = \log_{10} m + \log_{10} 10^k = \log_{10} m + k.$$

Es decir, $\log_{10} x$ se descompone en una parte entera k, que a su vez es justamente el factor al que elevamos 10 antes, y una parte decimal $\log_{10} m$.

Ahora fijemos un dígito $d \in \{1, 2, \dots, 9\}$. Decir que el primer dígito de x es d equivale se puede ver como:

$$d \leq m < d + 1,$$

pues $m \in [1, 10)$. Como la función $\log_{10}$ es estrictamente creciente, al aplicar logaritmo a la

desigualdad obtenemos

$$\log_{10} d \leq \log_{10} m < \log_{10}(d+1).$$

Por tanto, tenemos la equivalencia

$$\text{“el primer dígito de } x \text{ es } d\text{”} \iff \log_{10} m \ \in [\log_{10} d,\ \log_{10}(d+1)).$$

Resumiendo: x está completamente determinado por la región del intervalo $[0,1)$ donde cae la parte fraccionaria de $\log_{10} x$.

Conexión con la probabilidad

Lo construido anteriormente es meramente algebraico, aún necesitamos saber cómo conectarlo con variables aleatorias, pues estas a las que les queremos estudiar el/los primer/primeros dígitos.

Si X es una variable aleatoria positiva que modela un monto observado, definimos

$$Y := \log_{10} X - \lfloor \log_{10} X \rfloor,$$

Y representa la parte decimal del logaritmo de X, y por la equivalencia anterior.

$$P(\text{primer digito de X} = d) = P\big(Y \in [\log_{10} d,\ \log_{10}(d+1))\big)$$

Lo que conecta principalmente esto con la Ley de Benford es que, para muchos fenómenos “naturales”, y si la variable X se genera de manera no sesgada, a partir de procesos multiplicativos (dígase acumulaciones de interés por el tiempo, ventas en donde se tenga artículos $\times$ precio, etc, entonces $Y \approx \text{Unif}(0,1)$. I.e para cualquier subintervalo $[a,b) \subset [0,1)$,

$$\mathbb{P}(a \leq Y < b) = b - a,$$

Así como $d \in \{0,1,...,9\} \implies \log_{10} d \ \wedge\ \log_{10} d + 1 \in [0,1)$

Por tanto:

$$\mathbb{P}(\text{primer dígito de X} = d) = \mathbb{P}(Y \in [\log_{10} d,\, \log_{10}(d{+}1))) = \log_{10}(d{+}1) {-} \log_{10}(d) = \log_{10}\left(1 + \frac{1}{d}\right).$$

Así se obtiene la fórmula clásica de Benford para el primer dígito.

Este resultado que $Y \approx \text{Unif}(0, 1)$ es mostrado en un trabajo como el de Hill (1995) o el de Boyle (1994) este último mostrando justamente que esto es cierto para procesos multiplicativos.

Distribución

Formalmente para presentar la Ley de Benford como una distribución definimos la variable aleatoria discreta

$$D_1 := \text{“primer dígito significativo de } X\text{”},$$

con soporte $\{1, 2, \dots, 9\}$. La función de masa de probabilidad (pmf) propuesta por Benford es

$$\mathbb{P}(D_1 = d) = \log_{10}\left(1 + \frac{1}{d}\right), \qquad d = 1, \dots, 9.$$

Esta pmf cumple $\mathbb{P}(D_1 = d) \geq 0$ para todo d y

$$\sum_{d=1}^{9} \mathbb{P}(D_1 = d) = 1,$$

por lo que D_1 es una variable aleatoria discreta válida.

De forma análoga se pueden definir variables para el segundo dígito D_2 (soporte $\{0, \dots, 9\}$) y para los dos primeros dígitos (D_1, D_2) (soporte $\{10, \dots, 99\}$), cuyas probabilidades se calculan de formas similares. En la práctica del artículo de Gomes da Silva y Carreira (2013), estas distribuciones teóricas sirven como referencia, pues se comparan las frecuencias observadas en los datos (conteos de dígitos) con las probabilidades que predice Benford para evaluar conformidad en distintos tests de dígitos.

Es importante recalcar que La Ley de Benford no aplica para cualquier conjunto de datos, y eso es algo que a veces se pasa por alto. Funciona bien cuando los valores abarcan varios órdenes de magnitud (de unidades a miles o millones), no están limitados por topes, mínimos artificiales

o reglas de diseño, y provienen de procesos variados o multiplicativos, como precios×cantidades, acumulaciones financieras, crecimiento exponencial, etc.

En este tipo de situaciones es común que la distribución del primer dígito se parezca bastante a la distribución logarítmica que describe Benford. Ejemplos típicos incluyen montos de ventas, ingresos, saldos acumulados, tamaños de poblaciones o volúmenes de transacciones financieras.

Por el contrario, la ley no es adecuada para datos que son esencialmente códigos o números asignados de forma arbitraria: números de factura, cuentas bancarias, cédulas, códigos postales, precios que se fijan "a propósito" (9,99; 19,99; 10 000 exactos), o listas dominadas por identificadores internos de una empresa. En estos casos no existe la mezcla de fuentes ni la variabilidad necesaria para que el mecanismo detrás de Benford tenga sentido.

Ley de los Grandes Números

Tomemos ahora una muestra de tamaño n de una VA que modela diferentes registros contables, por ejemplo, los montos de las facturas:

$$X_1, X_2, \dots, X_n,$$

Para cada dígito fijo j, definimos una variable indicadora

$$I_k^{(j)} := \begin{cases} 1, & \text{si } X_k = j, \\ 0, & \text{en otro caso,} \end{cases} \qquad k = 1, \dots, n.$$

Si contamos cuántas veces aparece el dígito j en la muestra, definimos

$$O_j := \sum_{k=1}^{n} I_k^{(j)},$$

de modo que O_j es la frecuencia absoluta del dígito j y la frecuencia relativa es

$$\frac{O_j}{n} = \frac{1}{n} \sum_{k=1}^{n} I_k^{(j)}.$$

Por LGN cuando conforme n crece:

$$\frac{O_j}{n} = \frac{1}{n}\sum_{k=1}^{n} I_k^{(j)} \longrightarrow \mathbb{E}[I_1^{(j)}] = p_j,$$

es decir,

$$\frac{O_j}{n} \longrightarrow p_j$$

En palabras simples: si el modelo teórico es el correcto y contamos con suficientes datos, entonces la proporción de veces que aparece cada dígito j debería acercarse a su probabilidad teórica p_j. Ese es el puente entre:

- la teoría, que nos dice cómo deberían distribuirse los dígitos, y
- los datos observados, donde vemos las frecuencias relativas O_j/n.

Estadísticos

Dada una muestra aleatoria $X_1, \dots, X_n$ (por ejemplo, los primeros dígitos de n transacciones), se llama **estadístico** a cualquier función

$$T = T(X_1, \dots, X_n).$$

La idea es que Benford da probabilidades teóricas p_j, la muestra produce frecuencias relativas O_j/n, y los estadísticos miden "distancia" entre ambos. Aunque en este curso no se vieron estadísticos, sí tenemos todas las herramientas de Probabilidad para entender de dónde salen.

Los más importantes que usa el trabajo son:

- el **estadístico chi–cuadrado** X^2.
- el **MAD** (Mean Absolute Deviation).
- los Z–scores individuales para dígitos específicos.

Origen del estadístico χ^2

El siguiente paso es entender cómo medir cuantitativamente qué tan lejos están las frecuencias observadas de las probabilidades teóricas. Ahí entra el Teorema del Límite Central (TLC). Por tanto se explicará un estadístico de los 3 usados, el cual corresponde a uno de lo más importantes, el X^2

Retomemos la notación anterior. Para un dígito fijo j, sabemos que

$$O_j = I_1^{(j)} + \cdots + I_n^{(j)}, \qquad \mathbb{E}[O_j] = n\,p_j.$$

El TLC nos dice que, para n grande, la variable aleatoria

$$\frac{O_j - np_j}{\sqrt{n\,p_j(1-p_j)}}$$

se comporta aproximadamente como una **normal estándar** $N(0,1)$:

$$\frac{O_j - np_j}{\sqrt{n\,p_j(1-p_j)}} \;\approx\; Z_j \sim N(0,1).$$

Para combinar la información de **todos** los dígitos a la vez, se construye el conocido estadístico

$$X^2 = \sum_j \frac{(O_j - E_j)^2}{E_j}, \qquad E_j := n\,p_j.$$

Cada término puede reescribirse como

$$\frac{(O_j - E_j)^2}{E_j} = \frac{(O_j - np_j)^2}{np_j} = \left(\frac{O_j - np_j}{\sqrt{np_j(1-p_j)}}\right)^2 \cdot (1-p_j).$$

El factor $(1-p_j)$ está entre 0 y 1, por lo que no afecta mucho, entonces cada término es aproximadamente el cuadrado de una normal estándar (algo como Z_j^2 con $Z_j \sim N(0,1)$). Sabemos que distribución chi–cuadrado es precisamente la distribución de la suma de cuadrados de normales estándar independientes:

$$\chi^2_k = Z_1^2 + \cdots + Z_k^2, \qquad Z_i \sim N(0,1).$$

Así, de manera intuitiva pero bastante sólida, podemos ver que el estadístico X^2 se comporta aproximadamente como una variable con distribución chi–cuadrado con un cierto número de grados de libertad.

MAD y Z–scores: otras medidas de desviación

Además del χ^2, el artículo utiliza otros dos tipos de estadísticos para medir discrepancias:

- El **MAD (Mean Absolute Deviation)** se define, para un test de dígitos dado, como

$$\text{MAD} = \frac{1}{n(T)} \sum_i \left|\hat{p}(i,T) - e(i,T)\right|,$$

  donde $\hat{p}(i,T)$ es la frecuencia relativa observada de la categoría i en el test T, $e(i,T)$ es la probabilidad teórica de Benford y $n(T)$ es el número de categorías posibles (9 para el primer dígito, 100 para dos dígitos, etc.). Estas notaciones serán más detalladas adelante.

- Los Z–scores individuales, que comparan la frecuencia de un dígito específico:

$$Z(i,T) = \frac{\hat{p}(i,T) - e(i,T)}{\sqrt{e(i,T)\,[1-e(i,T)]/(N-k)}}.$$

Metodología

Inicialmente, *Selecting Audit Samples Using Benford's Law* presenta dos aplicaciones principales para las auditorías utilizando los tres modelos estadísticos ya definidos. Para un análisis más compacto, se toma en cuenta la primer aplicación, denominada como *Smallest Nonconforming Set of Records*, la cual se modela al final de este apartado, pero de manera intuitiva, corresponde **al conjunto más pequeño de registros no conformes** obtenidos de un datset.

Estadísticos de prueba y pruebas de conformidad en el modelo

Hasta aquí hemos visto la distribución de Benford como modelo teórico y cómo la LGN y el TLC justifican que las frecuencias relativas en la muestra se acerquen a las probabilidades p_j. Falta explicar cómo esas ideas se convierten, en el artículo de Gomes da Silva y Carreira (2013), en **pruebas de conformidad** concretas y, luego, en restricciones dentro del modelo de programación matemática para el "smallest nonconforming set of records".

La lógica general es:

1. Benford entrega probabilidades teóricas $e(i,T)$ para cada dígito (o par de dígitos) i, en cada tipo de test T.

2. A partir de los datos (después de eliminar ciertos registros sospechosos) se calculan frecuencias relativas $\hat{p}(i, T)$.
3. Se construyen estadísticos de prueba (chi–cuadrado, MAD, Z–scores) que miden distancia entre $\hat{p}(i, T)$ y $e(i, T)$.
4. El modelo exige que esos estadísticos estén por debajo de ciertos valores críticos. Si no, elimina más registros (pone más $y_t = 1$) hasta que el subconjunto remanente "conforme" con Benford.

Tests de conformidad y notación

En el paper se consideran varios tests de dígitos, que se indexan por T, y se les denomina tests de conformidad.

- $T = 1$: primer dígito,
- $T = 2$: segundo dígito,
- $T = 3$: dos primeros dígitos,
- $T = 4$: último dígito,
- $T = 5$: dos últimos dígitos

Aunque es importante recalcar que estos test T son definidos por el auditor, y no son necesariamente siempre estos de arriba.

Para cada test T:

- el conjunto de categorías posibles (dígitos o pares de dígitos) se indexa por i;
- $n(T)$ es el número de categorías de ese test (por ejemplo, $n(1) = 9$ para primer dígito, $n(3) = 90$ para dos primeros dígitos);
- $e(i, T)$ es la probabilidad **teórica** asignada a la categoría i en el test T, básicamente la asignada por Benford.
-

$$h_t(i, T) = \begin{cases} 1, & \text{si el registro } t \text{ presenta el dígito (o par) } i \text{ en el test } T, \\ 0, & \text{en otro caso.} \end{cases}$$

Por ejemplo, si T es el primer dígito y el valor del registro t empieza con 4, entonces $h_t(4,T) = 1$ y $h_t(i,T) = 0$ si $i \neq 4$.

- Por último para decidir qué registros se retiran en la muestra a auditar, se define, para cada $t = 1, \ldots, N$,

$$y_t = \begin{cases} 1, & \text{si el registro } t \text{ se elimina (entra a la muestra a auditar),} \\ 0, & \text{si el registro } t \text{ se mantiene en el dataset “limpio”.} \end{cases}$$

Con esta notación:

- el tamaño de la muestra a auditar es

$$k = \sum_{t=1}^{N} y_t,$$

- el número de registros que se mantienen es $N - k$,
- el número de registros que se mantienen y que tienen el dígito i en el test T es

$$\sum_{t=1}^{N} h_t(i,T)\,(1 - y_t) = \sum_{t=1}^{N} h_t(i,T) - \sum_{t=1}^{N} h_t(i,T)\,y_t,$$

- por tanto, la frecuencia relativa observada de la categoría i en el test T, sobre el dataset que queda, es

$$\hat{p}(i,T) = \frac{\displaystyle\sum_{t=1}^{N} h_t(i,T) - \sum_{t=1}^{N} h_t(i,T)\,y_t}{N - k}.$$

Esta $\hat{p}(i,T)$ es exactamente el análogo de O_j/n que usamos en la parte teórica, sólo que adaptado para permitir que algunos registros hayan sido eliminados ($y_t = 1$).

Con esto ya se pueden escribir los estadísticos χ^2, MAD y Z directamente en términos de las variables del modelo.

Prueba 1: restricción tipo chi–cuadrado (ecuación (1))

Recordemos el estadístico:

$$X^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}, \qquad E_i = n\, p_i.$$

En el contexto del modelo, después de eliminar k registros el tamaño de la muestra es $N-k$, y el valor esperado para la categoría i si los datos siguieran $e(i,T)$ sería

$$E_i = (N-k)\, e(i,T).$$

El conteo observado es $(N-k)\,\hat{p}(i,T)$, de modo que

$$\frac{(O_i - E_i)^2}{E_i} = \frac{\big((N-k)\hat{p}(i,T) - (N-k)e(i,T)\big)^2}{(N-k)e(i,T)} = (N-k)\,\frac{\big(\hat{p}(i,T) - e(i,T)\big)^2}{e(i,T)}.$$

Sumando sobre todas las categorías i del test T obtenemos

$$X^2(T) = (N-k)\sum_i \frac{\big(\hat{p}(i,T) - e(i,T)\big)^2}{e(i,T)}.$$

El artículo traduce esto directamente a la ecuación (1), escribiendo $\hat{p}(i,T)$ en función de las $h_t(i,T)$ y de las y_t:

$$(N-k)\sum_i \left[\frac{\sum_{t=1}^{N} h_t(i,T) - \sum_{t=1}^{N} h_t(i,T)\, y_t}{N-k} - e(i,T)\right]^2 \frac{1}{e(i,T)} \;\le\; S_1^*(T), \qquad T \in X_{S_1}.$$

Interpretación probabilística:

- el término entre corchetes es "frecuencia relativa observada − probabilidad de Benford" para la categoría i en el test T;
- cada término aporta un "pedazo" de la suma chi–cuadrado;
- el factor $(N-k)$ recoge el efecto del tamaño efectivo de la muestra.

Bajo la hipótesis de que los datos siguen Benford, el TLC sugiere que este estadístico se distribuye aproximadamente como una χ^2 con ciertos grados de libertad. Por eso, el auditor puede fijar un **valor crítico** $S_1^*(T)$: si el lado izquierdo es demasiado grande, la desviación global de Benford es sospechosa. En el modelo, se exige que

$$X^2(T) \leq S_1^*(T)$$

para todos los tests T que se decidan usar. Si esta desigualdad no se cumple, el problema de optimización tendrá que incrementar algunos y_t (eliminar registros adicionales) hasta que la muestra remanente tenga un $X^2(T)$ aceptable.

Prueba 2: restricción tipo MAD (ecuación (2))

El segundo estadístico colectivo que usa el artículo es el **MAD (Mean Absolute Deviation)**. Para un test T dado, se define

$$\text{MAD}(T) = \frac{1}{n(T)} \sum_i \left|\hat{p}(i,T) - e(i,T)\right|,$$

donde $n(T)$ es el número de categorías para ese test.

El paper lo incorpora mediante la ecuación (2):

$$\frac{1}{n(T)} \sum_i \left| \frac{\sum_{t=1}^{N} h_t(i,T) - \sum_{t=1}^{N} h_t(i,T)\, y_t}{N-k} - e(i,T) \right| \leq S_2^*(T), \qquad T \in X_{S_2}.$$

De nuevo, el modelo exige que, para los tests T que se consideren relevantes, el MAD del dataset "limpio" esté por debajo del umbral. Si no, será necesario eliminar más registros.

Prueba 3: restricciones tipo Z–score (ecuación (3))

Además de las medidas globales, el artículo introduce restricciones tipo Z–score para dígitos específicos. Esto último con el fin de que, aunque el conjunto completo pueda parecer razonable, tal vez haya ciertos dígitos o pares de dígitos muy particulares (por ejemplo, 0, 5, 9, 00, 99) cuya frecuencia individual se vea claramente anómala.

De nuevo, el paper incorpora este mediante la ecuación (3):

$$\left| \frac{\frac{1}{N-k}\sum_{t=1}^{N} h_t(i,T)\,(1-y_t) - e(i,T)}{\sqrt{e(i,T)\,[1-e(i,T)]/(N-k)}} \right| \leq\ Z^*(i,T), \qquad (i,T) \in X_Z,$$

donde $Z^*(i,T)$ es un máximo tolerable fijado por el auditor.

Conexión con el "Smallest Nonconforming Set of Records"

Resumiendo, la sección "The Smallest Nonconforming Set of Records" del paper se apoya directamente en ideas básicas de probabilidad vistas en el curso:

1. Se modela el comportamiento de los dígitos usando la distribución discreta de Benford (y, para últimos dígitos, una distribución casi uniforme).
2. La Ley de los Grandes Números (LGN) respalda que, si los datos son "limpios", las frecuencias relativas observadas deben acercarse a las probabilidades teóricas.
3. El Teorema Central del Límite (TLC) permite aproximar las fluctuaciones de los conteos con distribuciones normales y, a partir de ahí, justifica el uso del estadístico X^2, MAD, y de los Z-scores.
4. Se construyen estadísticos de pruebaso.
5. Cada estadístico se compara con un valor crítico elegido por el auditor ($S_1^*(T)$, $S_2^*(T)$, $Z^*(i,T)$). Estas comparaciones se traducen en restricciones dentro del modelo de optimización:
    - si el dataset resultante después de remover algunos registros cumple todas las restricciones, se considera conforme con Benford;

- si no, el modelo debe eliminar más registros (asignar más y_t = 1) hasta que lo sea.

El objetivo que plantean Gomes da Silva y Carreira es, justamente:

$$\min \sum_{t=1}^{N} y_t$$

sujeto a todas las restricciones de conformidad. En otras palabras, buscan el conjunto más pequeño de registros no conformes que, una vez retirado, haga que el dataset pase las pruebas de Benford.

Parámetros Definidos y Protocolo Iterativo

El modelo requiere que el auditor defina ciertos parámetros antes de ejecutar el procedimiento, en particular:

- los **tests de conformidad** a utilizar (primer dígito, segundo dígito, dos primeros dígitos, últimos dígitos, etc.);
- los **estadísticos** asociados a cada test (χ^2, MAD o *Z*-score);
- los **valores críticos** correspondientes a los estadísticos seleccionados.

En la práctica, estas decisiones pueden generar incertidumbre para el auditor. Por ello, Gomes da Silva y Carreira (2013) proponen un **protocolo iterativo** que permite ajustar los valores críticos después de observar el tamaño de la muestra resultante.

Según este procedimiento, el auditor primero fija los tests y estadísticos a emplear, así como los valores críticos iniciales. A continuación, se resuelve el modelo y se obtiene la muestra de auditoría, formada por aquellos registros para los cuales:

$$y_t = 1.$$

Una vez obtenida la solución, el auditor evalúa si el tamaño de la muestra, denotado por z, es adecuado. En caso de no serlo, puede ajustarlo de la siguiente manera:

- Si desea **una muestra más grande**, puede **disminuir los valores críticos**, haciendo más estrictas las pruebas de conformidad.
- Si desea **una muestra más pequeña**, puede **incrementar los valores críticos**, haciendo las pruebas más permisivas.

El proceso continúa hasta que el auditor obtenga un tamaño z satisfactorio. Los registros seleccionados finalizan siendo aquellos con $y_t = 1$ en la solución correspondiente a los valores críticos aceptados.

Este protocolo asiste al auditor en la gestión del riesgo. Valores críticos elevados implican un mayor riesgo de pasar por alto registros fraudulentos (dado que la muestra será más pequeña), mientras que valores críticos bajos generan muestras más amplias, reduciendo la probabilidad de excluir registros relevantes para la auditoría.

Análisis de resultados y efectividad del modelo

Para analizar los resultados del modelo los autores generaron aleatoriamente una muestra de 30 datasets, con 5000 registros de cuatro (4) dígitos cada uno, generados del resultado de 1000 x 10^r (con r entre 0 y 1), y le aplicaron cuatro tipos típicos de manipulación de datos:

- Redondeo (se redondea el número al múltiplo de 100 más cercano)
- Barrera psicológica (reemplaza los últimos dos dígitos por 99 y le resta una unidad al segundo dígito)
- Invención de números (reemplaza el número por otro generado de la distribución uniforme)
- Uso repetido del mismo número (reemplaza los últimos registros por un único valor de los registros)

Estas simulaciones de datos contaminados se hicieron en cuatro grupos de los mismos 30 datasets cada uno, dejando en total 120 datasets, en los que el primer grupo presentaba un 2% de los datos contaminados; el segundo grupo presentaba un 10% de los datos contaminados; el tercer grupo presentaba un 20% de los datos contaminados; y, el último grupo presentaba un 40% de los datos contaminados. Finalmente el desempeño del modelo se analiza observando las siguientes dos métricas:

1. Porcentaje de registros contaminados dentro de la muestra auditada (cuál es el porcentaje de registros realmente contaminado de la muestra que señaló el modelo)
2. Porcentaje de registros contaminados detectados por el modelo (cuál es el porcentaje de contaminados que señaló el modelo respecto a los contaminados totales)

Efectividad según nivel de contaminación

Contaminación del 2%

- Tamaño promedio de la muestra auditada: 28 registros
- Registros contaminados dentro de la muestra auditada: 42%
- Porcentaje total de contaminaciones detectadas: 11%

Interpretación: La tasa de detección total es baja, esto se puede dar debido a que las desviaciones producidas son mínimas y muchas manipulación pasan inadvertidas. Esto evidencia que el poder de detección del Modelo 1 en contextos de fraude leve es limitado.

Contaminación del 10%

- Tamaño promedio de la muestra auditada: 327 registros
- Registros contaminados dentro de la muestra auditada: 77%
- Porcentaje total de contaminaciones detectadas: 50%

Interpretación: Ahora se observa una mayor contaminación, los patrones de desviación son más evidentes y el modelo puede aislar con mayor precisión los grupos manipulados. Aquí el Modelo 1 logra capturar la mitad de todas las manipulaciones y, además, 3 de cada 4 registros que marca son realmente fraudulentos.

Contaminación del 20%

- Tamaño promedio de la muestra auditada: 705 registros
- Registros contaminados dentro de la muestra auditada: 90%
- Porcentaje total de contaminaciones detectadas: 64%

Interpretación: La detección de datos manipulados aumenta cada vez más, casi todos los registros señalados por el modelo están realmente manipulados. Esto sugiere que el Modelo 1 es altamente efectivo en escenarios donde el fraude tiene un peso considerable dentro del dataset.

Contaminación del 40%

- Tamaño promedio de la muestra auditada: 1479 registros
- Registros contaminados dentro de la muestra auditada: 95%
- Porcentaje total de contaminaciones detectadas: 70%

Interpretación: El modelo prácticamente identifica solo registros alterados (95%). A nivel de contaminación total reconoce el 70% de las manipulaciones totales. Esta es la condición donde el modelo funciona mejor, confirmando que su fortaleza crece con la magnitud de la manipulación.

Variabilidad de resultados

En todos los niveles de contaminación la variación entre datasets disminuye conforme aumenta el nivel de manipulación. Esto indica que el Modelo es más estable y predecible en escenarios con más señales de fraude, pero funciona peor cuando la contaminación es baja.

Comparación con métodos alternativos

Los autores realizan una comparación con un método simple basado en Z-scores de dígitos, observando a nivel de la muestra auditada:

Contaminación 2%:

- Modelo 1: 28
- Método alternativo : 214

Contaminación 10%:

- Modelo 1: 327
- Método alternativo : 633

Contaminación 20%:

- Modelo 1: 705
- Método alternativo : 1078

Contaminación 40%:

- Modelo 1: 1479
- Método alternativo : 1851

Gracias a esto y al nivel de detección de los datos contaminados se observa que el modelo del paper genera muestras más pequeñas y es más eficaz a la hora de auditar valores manipulados.

Conclusión general del análisis

Se observa que el modelo tiene las siguientes ventajas:

- Mejora drásticamente con la magnitud del fraude.
- Optimiza el tamaño del audit sample, siendo más eficiente que métodos basados solo en Z-statistics.
- Detecta adecuadamente manipulaciones que afectan patrones de dígitos.

En conjunto, la evidencia muestra que el modelo es una buena herramienta para detectar manipulación en datos contables cuando las distorsiones son suficientemente significativas para alterar las distribuciones de Benford.

Conclusiones:

En conclusión, se observo como la probabilidad puede llegar a impactar y revoluciona ámbitos como lo es la auditoria, donde se le otorga un uso practico y vital. Ello se demuestra mediante el caso analizado pues a través de métodos centrados en la probabilidad, como lo son las funciones indicadoras, las propiedades para que un caso concreto se considere una variable aleatoria discreta, utilidad de TLC, las leyes de los grandes números y demás.

Con lo cual, se logro identificar y aproximar patrones numéricos esperados que detectaron desviaciones de los datos dando así indicadores de anomalías o irregularidades, con la finalidad de analizar conjuntos de datos para “auditar”. Ejemplificando como la probabilidad puede retomarse y ser clave en procesos de toma de decisiones y simplificación de acciones, como lo fue con la selección de muestras en auditoria de manera más eficiente.

Por último, se denotaron ciertas limitaciones en el área de la probabilidad, como lo es con la utilidad de la Ley de Benford para el análisis de datos pues se requieren condiciones específicas de los conjuntos de datos seleccionados, como lo es la cobertura de muchos ordenes de magnitud, es decir, que los valores tomados abarquen desde números pequeños hasta valores grandes; que el conjunto no tenga limitaciones como topes o mínimos y que la selección de datos provenga de muchos factores mezclados entre si como lo son ventas acumuladas, precios por cantidades, etc. Lo anterior se presenta pues Benford únicamente funciona de forma pertinente cuando los números representan procesos variados y multiplicativos, de aleatoriedad razonable. Pues si fueran datos con forma lógica, asignados por una regla o no varían lo suficiente, no se lograría una distribución logarítmica adecuada y las "desviaciones" vistas serian una "falsa alarma".

Referencias:

Boyle, J. 1994. An application of Fourier series to the most significant digit problem. American Mathematical Monthly (November): 879–886. Gomes da Silva, C., & Carreira, P. M. R. (2013). Selecting Audit Samples Using Benford's Law. Auditing: A Journal of Practice & Theory, 32(2), 53–65. Hill, T. 1995. A statistical derivation of the significant-digit law. Statistical Science 10 (4): 354–363. Newcomb, S. 1881. Note of the frequency of use of the different digits in natural numbers. American Journal of Mathematics 4: 39–40.