Marco

aclarar que solo se trabajará uno de los 2 métodos del paper: **The Smallest Nonconforming Set of Records**

o teórico: Ley de Benford y representación logarítmica

Después de la introducción y la motivación del trabajo, el siguiente paso es entender con más detalle qué es la Ley de Benford desde el punto de vista probabilístico y por qué tiene sentido usarla como referencia de "normalidad" en los dígitos de los datos contables y financieros.

Representación logarítmica y primer dígito

Antes de entrar a la fórmula explícita de la Ley de Benford, es útil ver por qué el **primer dígito** de un número se puede leer directamente a partir de la **parte fraccionaria** de su logaritmo en base 10. Esta observación es la base de la interpretación logarítmica de la ley.

Todo número positivo x se puede escribir de forma única como

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

con $k \in \mathbb{Z}$ y $m \in [1, 10)$. El factor 10^k nos dice en qué "orden de magnitud" se encuentra el número (si está entre 1 y 10, entre 10 y 100, entre 100 y 1000, etc.), mientras que m es la "mantisa" normalizada en el intervalo $[1, 10)$. El punto importante es que el **primer dígito significativo** 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.

Podemos formalizar esto con logaritmos. Sea $x > 0$. Definimos

$$k = \lfloor \log_{10} x \rfloor,$$

de modo que se cumple

$$10^k \leq x < 10^{k+1}.$$

Luego definimos

$$m := \frac{x}{10^k},$$

con lo cual $m \in [1, 10)$ y, como antes, $x = m \cdot 10^k$.

Tomando 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 y una "parte fraccionaria" $\log_{10} m$. En notación de parte fraccionaria,

$$\{\log_{10} x\} = \log_{10} m,$$

donde $\{y\}$ denota la parte fraccionaria de y (es decir, y menos su parte entera). Esto muestra que **la parte fraccionaria del logaritmo de** x contiene exactamente la misma información que la mantisa m en $[1, 10)$.

Ahora fijemos un dígito $d \in \{1, 2, \ldots, 9\}$. Decir que el primer dígito de x es d equivale a decir que la mantisa m cae en el intervalo

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

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

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

Recordando que $\log_{10} m = \{\log_{10} x\}$, podemos reescribirlo como

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

Por tanto, tenemos la equivalencia

$$\text{"el primer dígito de } x \text{ es } d\text{"} \quad \Longleftrightarrow \quad \{\log_{10} x\} \in [\log_{10} d,\ \log_{10}(d + 1)).$$

En palabras: **el primer dígito de un número positivo** x está completamente determinado por la región del intervalo $[0, 1)$ donde cae la parte fraccionaria de $\log_{10} x$. Cada dígito d corresponde a un subintervalo $[\log_{10} d,\ \log_{10}(d + 1))$.

Como ejemplo concreto, si tomamos $x = 400$, tenemos

$$\log_{10}(400) = 2.60206\ldots$$

La parte entera es 2 y la parte fraccionaria es aproximadamente 0.60206. Por otro lado,

$$\log_{10}(4) \approx 0.60206, \qquad \log_{10}(5) \approx 0.69897,$$

de modo que

$$\{\log_{10}(400)\} \in [\log_{10} 4, \log_{10} 5).$$

Según la equivalencia anterior, esto indica que el primer dígito de 400 es 4, como es de esperar. Casos como $x = 200$ o $x = 3000$ caen exactamente en los extremos de los intervalos, pero adoptando la convención estándar de intervalos semiabiertos $[\log_{10} d, \log_{10}(d + 1))$ se asignan correctamente al dígito $d + 1$, coincidiendo con la lectura usual del primer dígito.

Uniformidad logarítmica e intuición multiplicativa

La construcción anterior es puramente geométrica. Para convertirla en un modelo probabilístico necesitamos hacer una hipótesis sobre cómo se distribuye la parte fraccionaria del logaritmo.

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

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

de modo que, por construcción, $Y \in [0, 1)$. Tal como vimos, el primer dígito de una realización x de X viene determinado por el subintervalo $[\log_{10} d,\ \log_{10}(d+1))$ en el que cae el valor de Y (con $d \in \{1, \dots, 9\}$).

La hipótesis central que conecta esta construcción con la Ley de Benford es que, para muchos fenómenos "naturales" y cuando los datos son suficientemente heterogéneos en escala, la variable Y **se comporta aproximadamente como una uniforme en** $[0, 1)$, es decir $Y \approx \text{Unif}(0, 1)$. Decir que $Y \sim \text{Unif}(0, 1)$ significa que, para cualquier subintervalo $[a, b) \subset [0, 1)$,

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

es decir, la probabilidad que cae en un subintervalo depende únicamente de la longitud del subintervalo (todos los puntos del segmento tienen la misma "densidad" de ocurrencia).

Bajo esta hipótesis de **uniformidad logarítmica**, la probabilidad de que el primer dígito de X sea d es exactamente la longitud del subintervalo correspondiente:

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

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

Esta hipótesis no es gratuita. Hill (1995) demuestra de forma rigurosa que, bajo condiciones muy generales, **mezclas** razonables de distribuciones y de escalas conducen a la Ley de Benford: si se combinan datos provenientes de orígenes distintos y que cubren varios órdenes de magnitud, la distribución límite de esta parte fraccionaria tiende a ser uniforme, y, en consecuencia, los primeros dígitos siguen la distribución logarítmica. De forma complementaria, Boyle (1994) muestra que listas de números obtenidas mediante productos, cocientes y potencias enteras de variables aleatorias también convergen a Benford. En ambos casos la intuición es fuertemente **multiplicativa**: en escala logarítmica, productos se convierten en sumas, y sumas largas de variables aleatorias son el terreno natural del Teorema del Límite Central.

La Ley de Benford como distribución discreta

Con la interpretación anterior, podemos presentar la Ley de Benford como una **distribución de probabilidad** sobre el primer dígito. 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 en el sentido del curso de Probabilidad.

A partir de aquí pueden definirse, con la notación habitual, su **esperanza**

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

y su **varianza**

$$\mathrm{Var}(D_1) = \mathbb{E}[D_1^2] - \left(\mathbb{E}[D_1]\right)^2, \qquad \mathbb{E}[D_1^2] = \sum_{d=1}^{9} d^2\, \mathbb{P}(D_1 = d).$$

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: 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.

Cuándo aplicar Benford (ejemplos y limitaciones)

No todos los conjuntos de datos deberían analizarse con Benford. La ley funciona bien cuando los datos:

- se extienden a través de varios órdenes de magnitud (por ejemplo: poblaciones, montos que van desde unidades hasta millones),
- no están forzados por topes, mínimos o reglas de diseño,
- y provienen de procesos que combinan muchas fuentes o factores multiplicativos (precios $\times$ cantidades, intereses compuestos, etc.).

En este tipo de contextos se ha observado que la distribución empírica del primer dígito se parece bastante a la distribución logarítmica. Ejemplos típicos donde Benford suele ajustarse son montos de ventas, cifras de ingresos, saldos contables acumulados, tamaños de poblaciones y volúmenes de transacciones bursátiles.

En cambio, la ley **no es adecuada** para datos que son códigos o números "asignados": números de factura, números de cuenta, cédulas, códigos postales, precios fijados "a propósito" (9,99; 19,99; 10 000 exactos), o conjuntos donde casi todos los valores son específicos de una empresa (por ejemplo, muchos códigos internos). En estos casos se rompe la hipótesis de mezcla/heterogeneidad y la uniformidad logarítmica deja de tener sentido.

Esta distinción es clave para el uso del modelo del paper: antes de aplicar la Ley de Benford y sus pruebas de conformidad, el auditor debe tener una buena razón para pensar que el dataset **a priori** debería seguir Benford. Sólo así tiene sentido interpretar desviaciones grandes como señales de posible manipulación o anomalía y alimentar el modelo de selección de muestras con esos resultados.

Ley de los Grandes Números y frecuencias relativas

Para pasar de la Ley de Benford (como modelo teórico) a algo que podamos **contrastar con datos**, necesitamos conectar las probabilidades teóricas con las **frecuencias observadas**. Aquí entra directamente la **Ley de los Grandes Números (LGN)**, aplicada no a promedios numéricos cualquiera, sino a **indicadores de eventos**.

Sea X la variable aleatoria que representa el dígito que estamos analizando (por ejemplo, el primer dígito significativo), con valores posibles en $\{1, 2, \dots, 9\}$ y probabilidades teóricas

$$\mathbb{P}(X = j) = p_j, \qquad j = 1, \dots, 9,$$

donde, en el caso de la Ley de Benford para el primer dígito, $p_j = \log_{10}(1 + \frac{1}{j})$.

Tomemos ahora una muestra de tamaño n de esta variable:

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

que modela, por ejemplo, los primeros dígitos de n registros contables. Para cada dígito fijo j, definimos una **variable indicadora**

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

Esta $I_k^{(j)}$ es una variable aleatoria discreta con

$$\mathbb{P}(I_k^{(j)} = 1) = p_j, \qquad \mathbb{P}(I_k^{(j)} = 0) = 1 - p_j,$$

y, por tanto,

$$\mathbb{E}[I_k^{(j)}] = p_j.$$

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

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

Aquí es donde usamos la LGN: la media muestral de variables i.i.d. converge a su esperanza. Aplicada a las indicadoras, nos da que, cuando n crece,

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

es decir,

$$\frac{O_j}{n} \longrightarrow p_j \quad \text{(casi seguramente o en probabilidad, según la versión).}$$

En palabras: **si el modelo teórico es correcto**, y tomamos suficientes observaciones, la fracción de veces que aparece cada dígito j debería aproximarse a la probabilidad teórica p_j. Este es el puente entre:

- lo que dice la teoría (distribución de X), y
- lo que se observa en los datos (frecuencias relativas O_j/n).

Cuando el modelo es Benford, esto significa que, para datos "limpios" y con tamaño de muestra grande, la distribución empírica de los primeros dígitos debería parecerse bastante a la distribución logarítmica de Benford.

Estadísticos: qué son y por qué aparecen aquí

En el curso de Probabilidad se trabaja sobre todo con variables aleatorias y sus distribuciones. En el contexto de contrastar modelo vs. datos aparece un concepto muy cercano, pero formulado "desde la muestra": el de **estadístico**.

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

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

Como cada X_k es una variable aleatoria, el valor de T también es aleatorio, y por lo tanto tiene una distribución de probabilidad asociada. Algunos ejemplos muy conocidos de estadísticos son:

- la **media muestral** $\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$,
- la **varianza muestral**,
- el **conteo** de cuántas veces ocurre cierto evento (por ejemplo, cuántas veces el primer dígito es 1).

En el estudio de la Ley de Benford, las cantidades O_j y las frecuencias relativas O_j/n son estadísticos construidos a partir de la muestra. A partir de ellos se definen otros estadísticos que resumen, en un solo número, qué tan lejos están los datos observados de lo que predice Benford. Los tres más importantes que usa el artículo son:

- el **estadístico chi–cuadrado** X^2, que suma desajustes al cuadrado,
- el **MAD** (Mean Absolute Deviation), que promedia desajustes absolutos,
- los Z–scores individuales para dígitos específicos.

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 todavía no se ha visto formalmente "pruebas de hipótesis", sí tenemos todas las herramientas de Probabilidad para entender de dónde salen las fórmulas de estos estadísticos y cómo se conectan con la LGN y el Teorema del Límite Central.

Teorema del Límite Central y el origen del estadístico χ^2

El siguiente paso es entender **cómo medir cuantitativamente** qué tan lejos están las frecuencias observadas de las probabilidades teóricas. Para eso no basta con la LGN; necesitamos una idea de cómo se comportan las **fluctuaciones** alrededor de los valores esperados. Ahí entra el **Teorema del Límite Central (TLC)**.

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

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

La diferencia

$$O_j - n\,p_j$$

mide cuánto se desvía la frecuencia observada del valor esperado teórico. Por construcción, esta diferencia es una **suma de muchas pequeñas desviaciones** (cada $I_k^{(j)} - p_j$ aporta un "error"); bajo hipótesis estándar (independencia y misma distribución), el TLC se puede aplicar a la suma

$$O_j = I_1^{(j)} + \cdots + I_n^{(j)}.$$

La versión clásica del 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** $N(0,1)$:

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

Es decir, cada desviación estandarizada (dígito por dígito) es aproximadamente normal. Si miráramos sólo un dígito j, podríamos trabajar con este Z_j y un test basado en la normal; de hecho, esto es la idea detrás de los Z–scores que aparecen en el artículo para dígitos particulares (como 0, 5, 9 o pares como 00 y 99). En este trabajo, sin embargo, esos Z se usan sobre todo como apoyo: el foco principal está en dos estadísticos colectivos más simples de interpretar: el χ^2 y el MAD.

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

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

Cada término puede reescribirse como

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

El factor $(1-p_j)$ está entre 0 y 1, por lo que no cambia la idea esencial: **cada término** es aproximadamente el cuadrado de una normal estándar (es decir, algo parecido a Z_j^2 con $Z_j \sim N(0,1)$). Por la definición que se ve en Probabilidad, la **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 heurística 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.). A diferencia del χ^2, el MAD no tiene detrás una distribución tan "clásica" en Probabilidad, pero sigue la misma lógica: **si las frecuencias observadas están cerca de las teóricas**, el MAD será pequeño; si hay desajustes importantes, el MAD crece.

- Los Z–scores individuales comparan la frecuencia de un dígito específico (por ejemplo, el dígito 0 en los centavos, o el par 99 en los dos últimos dígitos) contra lo que predeciría Benford, usando la aproximación normal del TLC:

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

En este proyecto no profundizamos en pruebas formales basadas en Z, pero sí es útil entender que se trata simplemente de **desviaciones estandarizadas**: valores Z muy grandes en valor absoluto indican que ese dígito concreto aparece mucho más (o mucho menos) de lo que diría el modelo.

En resumen, desde el punto de vista del curso de Probabilidad:

- cada uno de estos objetos (X^2, MAD, Z) es un **estadístico**, es decir, una función de la muestra;
- su definición usa de manera directa conceptos como variables indicadoras, esperanza, varianza y el Teorema del Límite Central;
- y todos ellos sirven, en el artículo, para cuantificar **qué tan lejos** están los datos observados de la distribución de Benford. En la siguiente parte del marco teórico se explicará cómo estos estadísticos se convierten en restricciones específicas dentro del modelo de programación matemática que busca el "smallest nonconforming set of records".

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 dígitos y probabilidades de referencia

En el paper se consideran varios **tests de dígitos**, que se indexan por T:

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

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:
  - para primeros dígitos, $e(i,T)$ viene de la Ley de Benford,
  - para últimos dígitos, $e(i,T)$ suele ser aproximadamente 1/10 o 1/100 (distribución casi uniforme).

Estas $e(i,T)$ juegan el papel de los p_j en la parte teórica: son las probabilidades que se esperan si los datos son "limpios".

Variables $h_t(i,T)$ y y_t: cómo se escribe el conteo en el modelo

Para poder incorporar todo esto en un modelo de programación matemática, el artículo introduce dos tipos de variables:

- Para cada registro t y cada categoría i en el test T, se define

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

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 $i \neq 4$.

- 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** (dataset "limpio") 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 clásico de Pearson para bondad de ajuste:

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

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

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

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

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

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

En palabras, el MAD es simplemente el promedio de las desviaciones absolutas entre las frecuencias relativas observadas y las probabilidades teóricas de Benford. No está ligado a una distribución tan clásica como la χ^2, pero es fácil de interpretar:

- si todas las $\hat{p}(i,T)$ son muy cercanas a $e(i,T)$, el MAD será muy pequeño;
- si hay dígitos con desajustes grandes, el MAD crecerá.

El paper lo incorpora mediante la ecuación (2):

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

Aquí $S_2^*(T)$ es un umbral elegido con base en criterios usados en auditoría digital (por ejemplo, valores recomendados por Nigrini y otros autores). 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.

En conjunto, las restricciones de tipo chi–cuadrado y MAD garantizan que, mirando el patrón de dígitos de forma global, el dataset que queda se parece razonablemente a lo que predice Benford.

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

Además de las medidas globales, el artículo introduce restricciones tipo Z–score para **dígitos específicos**. La idea es sencilla: 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.

Para un dígito (o par) i y test T, la frecuencia relativa en el dataset que queda es

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

y la probabilidad teórica es $e(i,T)$. Por modelo binomial, la desviación estándar teórica de $\hat{p}(i,T)$ es aproximadamente

$$\sqrt{\frac{e(i,T)\,[1-e(i,T)]}{N-k}}.$$

El estadístico

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

es entonces una desviación estandarizada: para $N-k$ grande, el TLC sugiere que $Z(i,T)$ es aproximadamente normal estándar si el modelo es correcto.

El artículo no desarrolla una prueba formal completa basada en estos Z, pero sí usa la idea de que valores de $|Z(i,T)|$ "demasiado grandes" señalan dígitos individuales sospechosos. Esto se traduce a 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.

En este proyecto no profundizamos en todas las sutilezas estadísticas de estos Z–scores; basta con entender que son estadísticos que miden, en unidades de desviación estándar, cuánto se aleja la frecuencia de un dígito concreto de su valor teórico, y que el modelo exige que estos desajustes individuales no sean excesivos.

Conexión final con el "Smallest Nonconforming Set of Records"

Resumiendo, la sección "The Smallest Nonconforming Set of Records" del paper se apoya completamente en los conceptos probabilísticos vistos en el curso:

1. Se modela el comportamiento de los dígitos mediante la **distribución discreta de Benford** (y, para últimos dígitos, una distribución casi uniforme).
2. La **LGN** justifica que las frecuencias relativas observadas deberían aproximarse a las probabilidades teóricas cuando los datos son limpios.

3. El **TLC** permite aproximar las fluctuaciones de los conteos por distribuciones normales y, a partir de ahí, justifica el uso del estadístico χ^2 y de los Z–scores.
4. Se definen **estadísticos de prueba**:

   - $X^2(T)$ (chi–cuadrado colectivo),
   - $\mathrm{MAD}(T)$ (desviación media absoluta),
   - $Z(i,T)$ (desviación estandarizada por dígito), todos ellos funciones de la muestra, es decir, estadísticos en el sentido básico de Probabilidad.

5. Cada uno de estos estadísticos se compara con un **valor crítico** fijado por el auditor ($S_1^*(T)$, $S_2^*(T)$, $Z^*(i,T)$). Estas comparaciones se convierten en restricciones en el modelo de programación matemática:

   - si el dataset que queda tras eliminar algunos registros cumple todas las restricciones, se considera **conforme** con Benford;
   - si no, el modelo debe eliminar más registros (poner más $y_t = 1$) hasta que lo sea.

El problema que resuelven Gomes da Silva y Carreira es, precisamente,

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

sujeto a todas estas restricciones de conformidad. Es decir, buscan el **conjunto más pequeño de registros no conformes** que, una vez retirado, deja un dataset cuyo patrón de dígitos pasa las pruebas de conformidad con la Ley de Benford. Esta formulación conecta directamente los conceptos teóricos de variables aleatorias discretas, LGN, TLC y distribución chi–cuadrado con una herramienta concreta para auditoría en grandes bases de datos financieros.