CALIDAD DE DATOS – TIPS PRÁCTICOS

Angel Monjarás

amonjaras@dataiq.com.mx

@QlikMaster

Qlik Dev Group

CALIDAD DE DATOS

Beneficios

¿Cómo se mide?

Técnicas útiles

BENEFICIOS DE LA CALIDAD DE DATOS

Eliminar información duplicada (ej. clientes)

Segmentación más precisa, eliminando “desconocidos”, “no definidos”, etc. (Marketing)

Mejorar la lealtad de los clientes (menos comunicaciones, más precisas)

Disminución de riesgos (menor margen de error)

Menores costos (no procesar ni almacenar datos basura)

MÉTRICAS DE LA CALIDAD DE DATOS

Exactitud

Unicidad

Integridad (Densidad)

Actualidad

Volatilidad (Caducidad)

MÉTRICAS DE LA CALIDAD DE DATOS

Exactitud	→	Cifras control, metodología, técnicas de comprobación
Unicidad	→	Cálculo
Integridad (Densidad)	→	Cálculo
Actualidad	→	Última actualización
Volatilidad (Caducidad)	→	Tiempo desde la última actualización

MÉTRICAS DE LA CALIDAD DE DATOS

Exactitud	→	Cifras control, metodología, técnicas de comprobación
Unicidad	→	Cálculo
Integridad (Densidad)	→	Cálculo
Actualidad	→	Última actualización
Volatilidad (Caducidad)	→	Tiempo desde la última actualización

MÉTRICAS CALCULADAS

Dentro de Qlik

Qlik Dev Group

MÉTRICAS CALCULADAS

▼ Vista previa

Añadir como dimensión

Añadir como medida

LAB

Densidad	93%
Ratio de subconjunto	100%
Contiene duplicados	verdadero
Total de valores distintos	366
Valores presentes distintos	366
Valores no nulos	12041
Etiquetas	$ascii $text

Vista previa de datos

COD BARRAS	No. Linea	COD ORACLE
736085445053	1	0736085445053AA
7501043161602	2	7501043161602AA
7501043161596	3	7501043161596AA
7503006916014	4	7503006916014AA
7501070635596	5	7501070635596AA
7501125106378	6	7501125106378AA
7501095402067	7	7501095402067AA

MÉTRICAS CALCULADAS

Densidad Promedio

69.05%

Unicidad Promedio

9.84%

Tabla	Campo	Llave	Renglones	Nulos	No Nulos	Únicos	Densidad	Unicidad
Totales			-	-	-	-	**69.05%**	**9.84%**
DATA2	CLAVE CLIENTE	NO	84046	6	84046	12413	100.00%	14.77%
FNDWRR	No. Linea	NO	12943	47	12932	12932	99.92%	99.64%
FNDWRR	NOMBRE COMERCIAL	NO	12943	155	12824	12224	99.08%	94.18%
FNDWRR	LABORATORIO	NO	12943	235	12744	674	98.46%	5.19%
FNDWRR	LAB	NO	12943	938	12041	366	93.03%	2.82%
FNDWRR	COD SIVEC	NO	12943	1048	11931	11189	92.18%	86.21%
DATA2	NÚMERO PEDIDO	NO	84046	6667	77385	38074	92.07%	45.30%
DATA2	TOTAL	NO	84046	6667	77385	24593	92.07%	29.26%
DATA2	NOMBRE CLIENTE	NO	84046	6667	77385	16866	92.07%	20.07%
DATA2	CLAVE MÉDICO	NO	84046	6667	77385	6237	92.07%	7.42%
DATA2	NOMBRE MÉDICO	NO	84046	6667	77385	5992	92.07%	7.13%
DATA2	PRODUCTO	NO	84046	6667	77385	3936	92.07%	4.68%
DATA2	PADECIMIENTO	NO	84046	6667	77385	615	92.07%	0.73%
DATA2	FECHA PEDIDO	NO	84046	6667	77385	264	92.07%	0.31%
DATA2	SUCURSAL SURTIDORA	NO	84046	6667	77385	85	92.07%	0.10%

FÓRMULAS

Renglones: =sum($Rows)

Llave: if(count(DISTINCT total<$Field> $Table)>1,'SI','NO')

No Nulos: $(=concat('if($Field=' & chr(39) & $Field & chr(39) & ',count({1}[' & $Field & '])', ',') & concat(right($Field&')',1)))

Nulos: $(=concat('if($Field=' & chr(39) & $Field & chr(39) & ',nullcount([' & $Field & '])', ',') & concat(right($Field&')',1)))

Únicos: =FieldValueCount($Field)

Densidad: =[No Nulos]/[Renglones]

Unicidad: =[Únicos]/([No Nulos]+[Nulos])

TÉCNICAS DE COMPROBACIÓN

Tablas simples

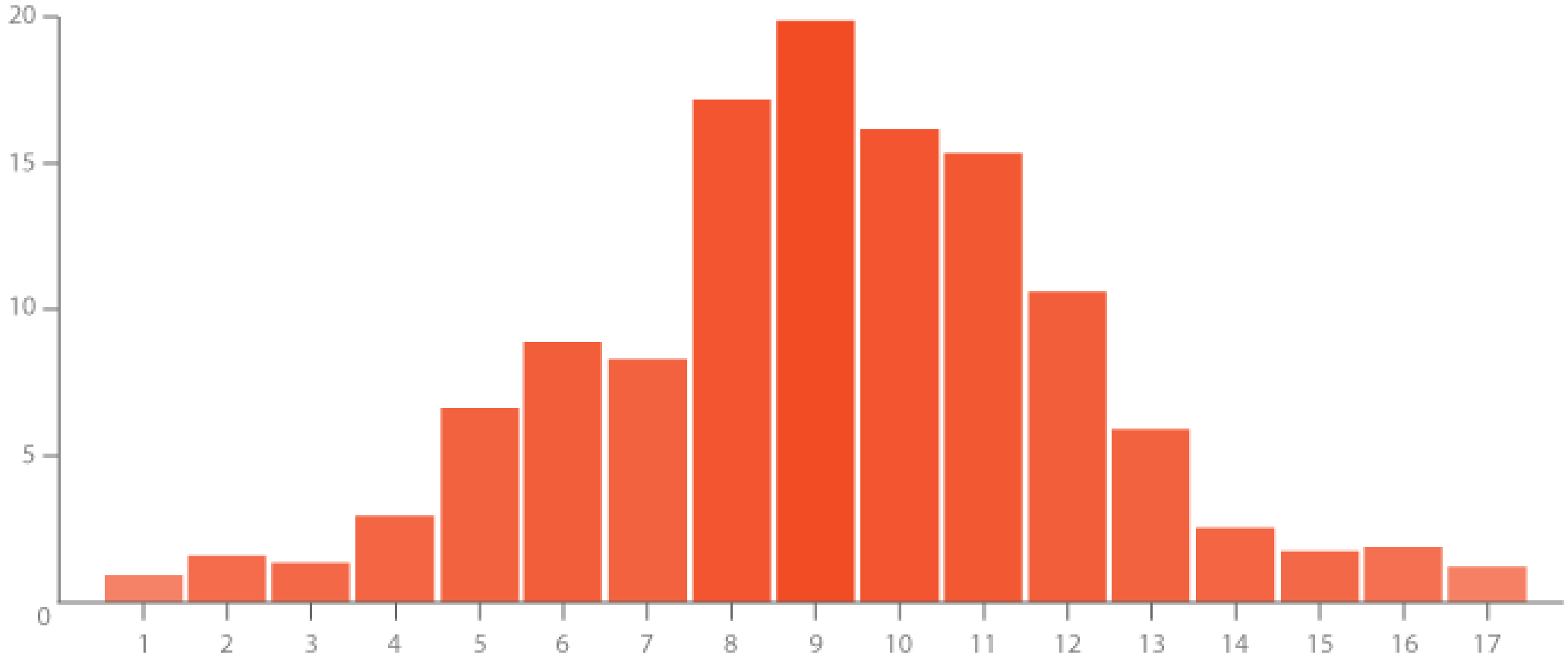
Qlik Dev Group

TABLAS SIMPLES

Sencillas y flexibles

Recomendadas para analizar datos desnormalizados

EJEMPLO…



TÉCNICAS DE COMPROBACIÓN

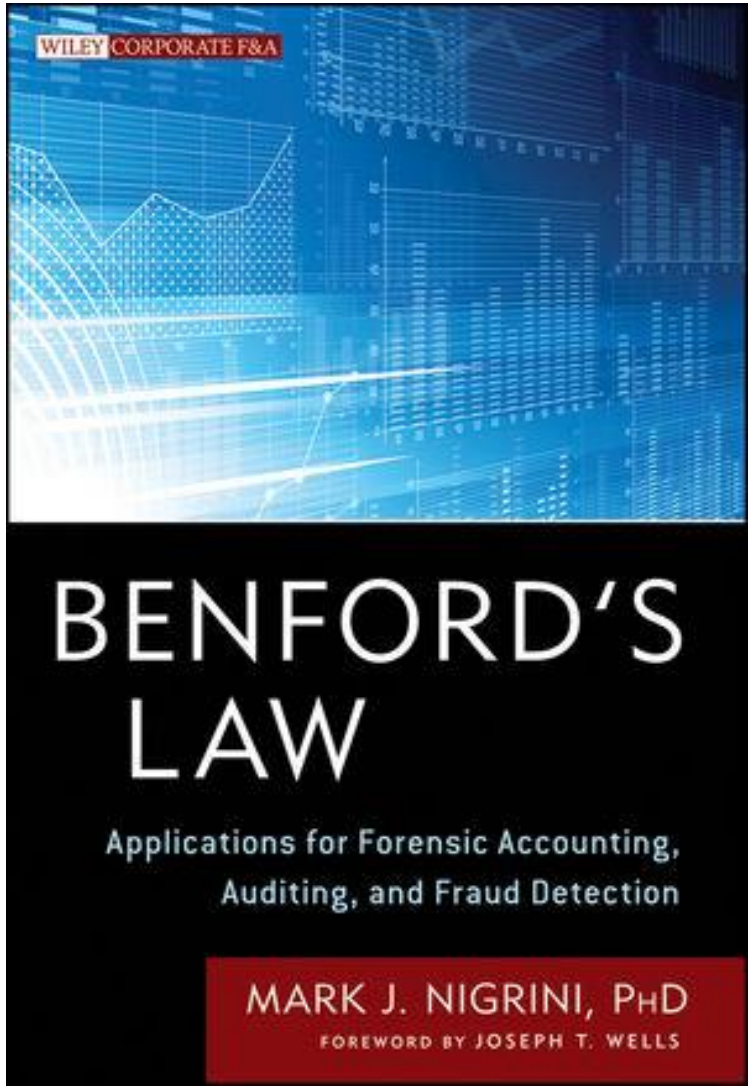
Histogramas

Qlik Dev Group

HISTOGRAMAS

Fáciles de entender y reconocer por los usuarios de negocio

Sirven para detectar o verificar patrones



TÉCNICAS DE COMPROBACIÓN

Ley de Benford

Qlik Dev Group

¿QUÉ ES LA LEY DE BENFORD?

También llamada “ley del primer dígito”

Es un método para predecir con gran exactitud los dígitos iniciales de una serie de números
no aleatorios

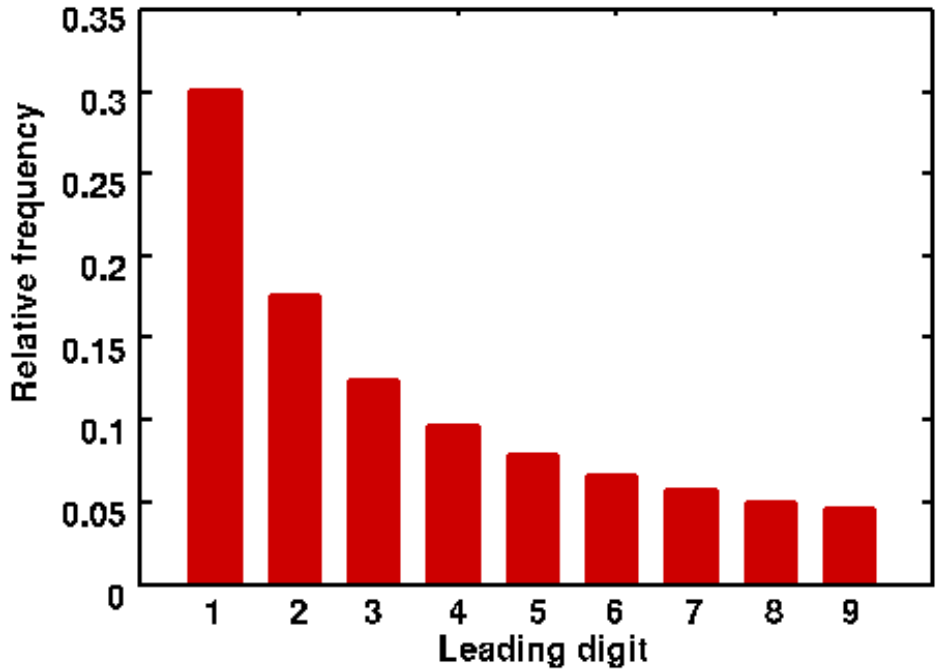
Wikipedia – definición simple:

- “…en los números que existen en la vida real, la primera cifra es 1 con mucha más frecuencia que el resto de los números.
- Además, según crece este primer dígito, más improbable es que se encuentre en la primera posición”

¿QUÉ ES LA LEY DE BENFORD?

- Fundamento: Los valores de las medidas del mundo real tienen una distribución logarítmica,
- La probabilidad de que un dígito (D) sea el primero en una cifra es = $\log_{10}(1+1/D)$

Ley de Benford	
Primer dítigo	**% que predice la Ley de Benford**
1	30.103
2	17.609
3	12.494
4	9.691
5	7.918
6	6.695
7	5.799
8	5.115
9	4.576



EJEMPLOS DE APLICACIÓN

Recibos de luz

Direcciones

Precios de acciones

Población

Tasas de mortandad

Longitud de ríos

Montos de facturación

Saldos de cuentas contables

Saldos de clientes

Valor de terrenos

¿A QUÉ TIPO DE DATOS APLICA?

Saldos totales que resulten de cualquier agregación:

Saldos contables, CxP, CxC, inventarios

Entre más etapas de cálculo se usen para obtener cada número de una serie, será más probable que el resultado se apegue a la predicción de la ley de Benford.

Números que resulten de una combinación matemática (ej.: Cantidad por precio)

Datos a nivel transacción (pagos, ventas, compras)

Cualquier cuenta o suma de los elementos de un conjunto de datos

¿CUÁNDO PUEDE QUE NO APLIQUE?

Números asignados (No. de cheque, No. de factura)

Números que tengan distribuciones distintas (distribución normal, distribución uniforme como números aleatorios, loterías o tirar un dado)

Números influenciados por el pensamiento humano (ej.: Precios con umbrales psicológicos como $99.99)

Saldos de cuentas creadas para propósitos específicos (ej. Registrar reembolsos de $100)

Conjuntos de datos con límites inferiores o superiores definidos (ej.: la estatura en metros casi siempre iniciará con 1 o 2)

“Efecto precio”. Por ejemplo, tickets de venta donde un producto con precio específico forma la mayor parte de las transacciones, o los totales de nómina por período

Cuando la muestra de datos sea muy pequeña

Números que no ocurren naturalmente (como números de teléfono o ip)

EJEMPLO CON QLIK SENSE

EXTRACCIÓN DEL PRIMER DÍGITO

En el script:

left (num ([TOTAL]) , 1) as [First Digit]

Si hay números menores a 1 o negativos:

left (purgechar ([MONEDA LOCAL] , '0.-‘) , 1) as [First Digit]

APLICACIONES PARA CALIDAD DE DATOS

Detección de pagos duplicados

Pagos sospechosos

Gastos sospechosos

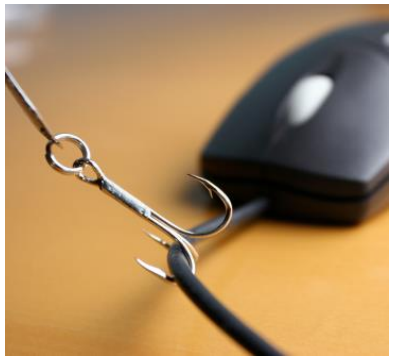
Detección de fraudes

Estimaciones influenciadas en presupuestos y provisiones

Números “inventados” en pronósticos (la distribución de los pronósticos debería ser igual a la de los números reales)

Errores sistémicos (ej.: Valores repetidos u omitidos debido a errores de lógica en el ETL)

DETECCIÓN DE FRAUDES



La ley de Benford es muy poderosa para detectar números “inventados” entre un conjunto de números calculados normalmente.

La metodología es aplicar la ley de Benford a cualquier serie de números calculados, y que se requiera una explicación cuando la serie no siga la ley.

La explicación puede deberse a alguna de las excepciones ya mencionadas, o a comportamiento anómalo y posiblemente malicioso.

Deben evaluarse los resultados para concluir si se trata de excepciones válidas o anomalías dignas de investigarse.

EJEMPLO: CUENTAS DE GASTOS

Un ejemplo clásico y sencillo:

Asumamos que los gastos mayores a $300 deben ser aprobados por la Gerencia General.

A menudo encontraremos muchos gastos justo por debajo de $300 para evitar la aprobación.

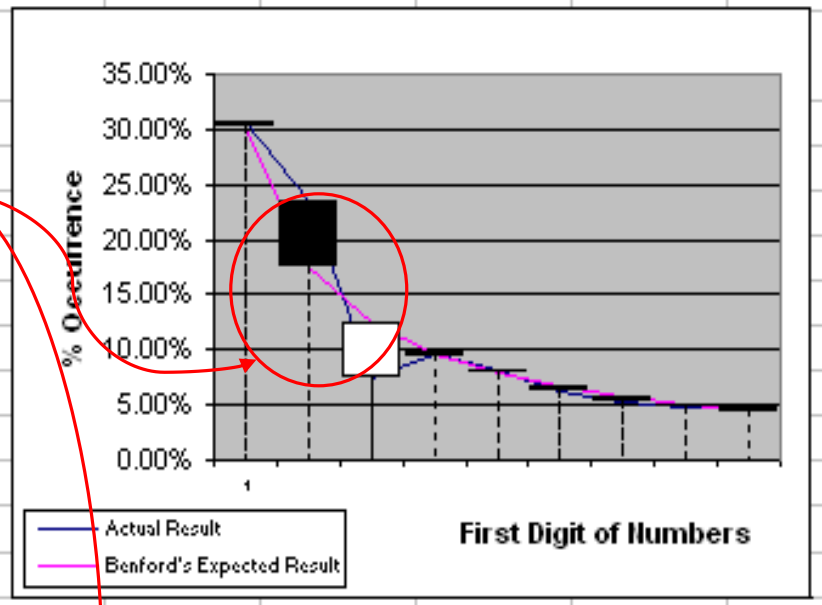
Esto se logra arreglando múltiples compras por debajo del límite, con o sin complicidad de los proveedores

La ley de Benford al rescate: Los primeros dígitos mostrarán anomalías (ej.: Preponderancia de 1s y 2s y menos 3s 4s y 5s de los esperados)

EJEMPLO: CUENTAS DE GASTOS

Anomalías en la distribución de las repeticiones de los primeros dígitos comparados con la distribución esperada de Benford*

First Digit	Total as a Percentage	Benford's
1	30.57%	30.10%
2	23.50%	17.61%
3	7.37%	12.49%
4	9.59%	9.69%
5	8.18%	7.92%
6	6.16%	6.70%
7	5.39%	5.80%
8	4.76%	5.12%
9	4.49%	4.58%
Grand Total	**100.00%**	**100.00%**

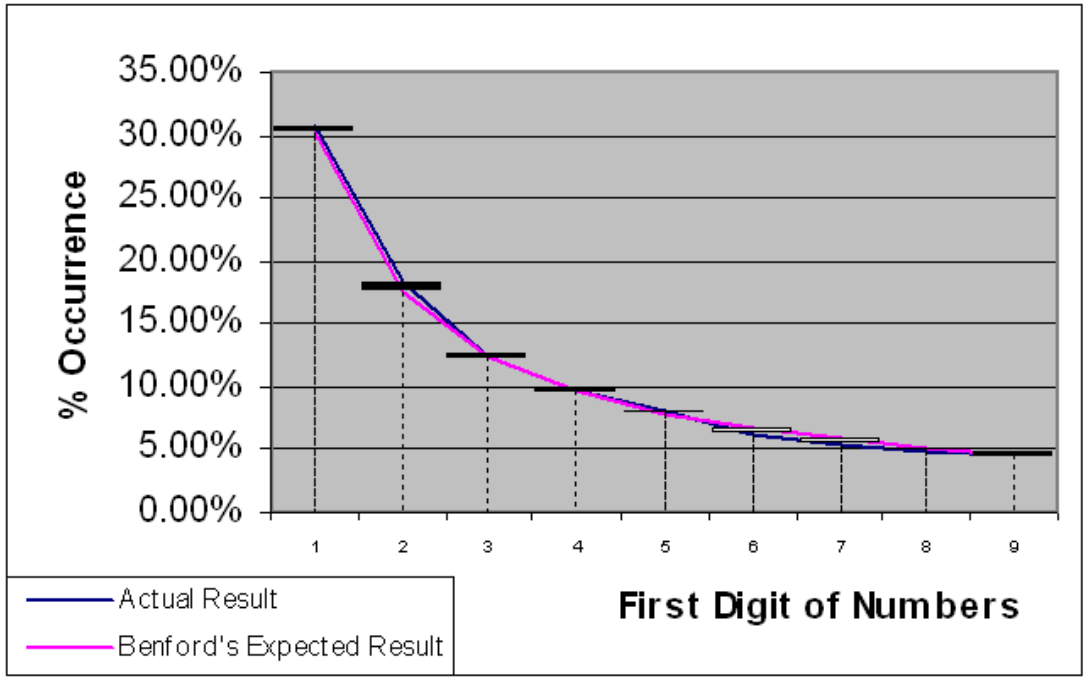


Transacciones sospechosas

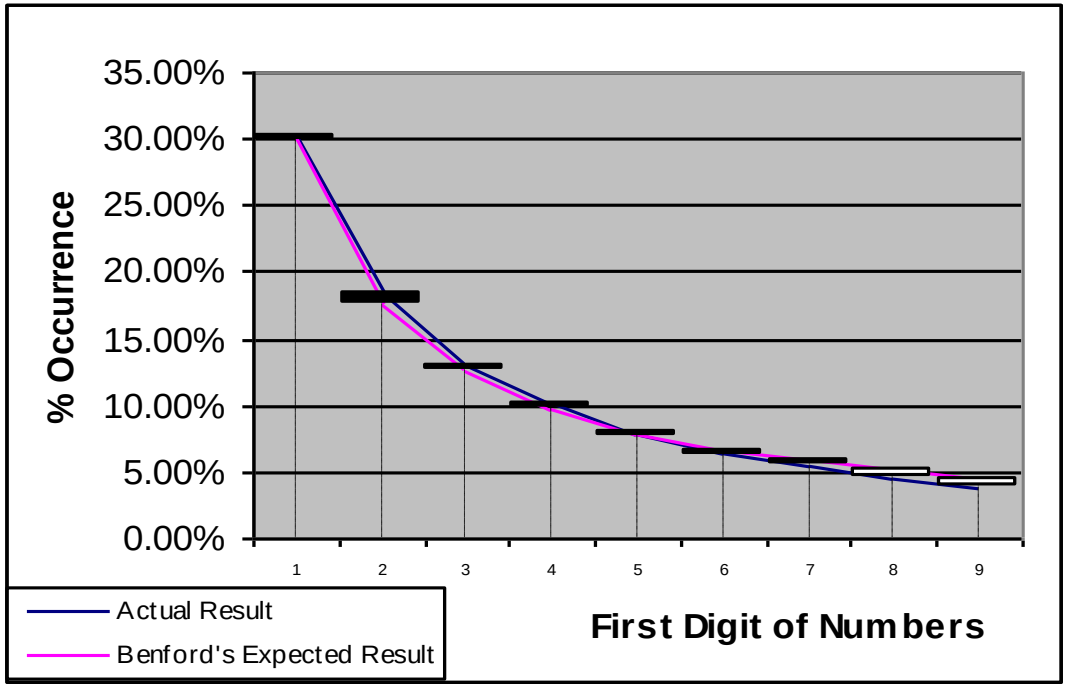
* Datos simulados para ilustrar el ejemplo

OTROS EJEMPLOS

6.5 millones de saldos de un DW

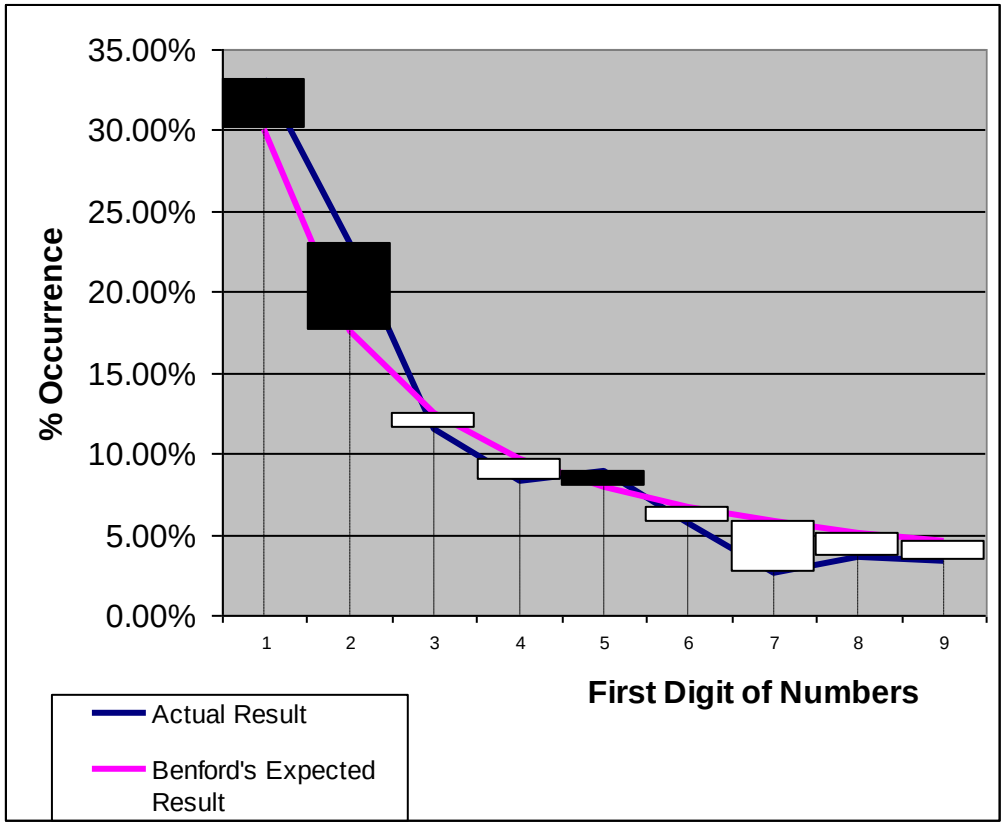


3.3 millones de transacciones de almacén



OTRO EJEMPLO: POCAS TRANSACCIONES

- Una tarjeta de crédito 976 transacciones en 12 meses “en crudo”



Correspondencia más o menos buena

Incluye muchos duplicados (pagos frecuentes, transferencias, montos “cerrados”, cargos bancarios, etc.)

Los comportamientos “conocidos” pueden filtrarse

Los datos (crudos o filtrados) pueden usarse como “huella digital” en un análisis comparativo con otros períodos.

CONCLUSIONES

La ley de Benford se aplica perfectamente a datos agregados.

Puede usarse para probar la calidad de datos.

Al encontra anomalías se puede revisar si la distribución “esperada” es realmente correcta.

Si los datos deberían ajustarse a la ley, hay que investigar más a fondo para saber las causas de las desviaciones (fraude, procesos erróneos, errores de ETL, repeticiones genuinas, etc.)

La ley de Benford es una técnica simple pero muy poderosa que puede agregarse a nuestras herramientas de Calidad de Datos.

¡MUCHAS GRACIAS!

Angel Monjarás

amonjaras@dataiq.com.mx

@QlikMaster

Qlik Dev Group