Estendendo o MySQL com Funções para Detecção de Fraudes através da Distribuição de Benford

Adiel Mittmann[1], **Eros Comunello**[2], **Aldo von Wangenheim**[1]

[1]INCoD – Instituto Nacional para Convergência Digital
Universidade Federal de Santa Catarina – UFSC
Florianópolis – SC – Brasil

[2]4Vision Lab – Universidade do Vale do Itajaí – UNIVALI
Florianópolis – SC – Brasil

adiel@inf.ufsc.br, eros.com@univali.br, awangenh@inf.ufsc.br

Abstract. *Benford's law governs the frequency of digits in data sets according to their position within numbers. In particular, this law states the expected frequency for digits 1–9 in the first significant digit. It is possible to detect suspect data upon learning that they do not follow Benford's distribution, when it is known that they should. This article presents a MySQL extension (whose source code is available) that provides functions allowing the estimation of how well numeric data match Benford's distribution. Experiments were conducted on three data sets in order to show how these functions can be used to detect suspect behavior.*

Resumo. *A lei de Benford governa a frequência com que os algarismos aparecem em conjuntos de dados numéricos de acordo com a posição em que eles se encontram dentro dos números. Em particular, esta lei especifica a frequência esperada para os algarismos de 1 a 9 na posição do primeiro dígito significativo. Pode-se detectar dados suspeitos quando se descobre que eles não seguem a distribuição de Benford, sabendo-se que eles deveriam segui-la. Este artigo apresenta uma extensão ao MySQL (cujo código-fonte está disponível) com funções que permitem estimar quão bem dados numéricos seguem a distribuição de Benford. Experimentos foram conduzidos em três conjuntos de dados para mostrar como as funções podem ser usadas para detectar comportamento suspeito.*

1. Introdução

A lei de Benford afirma que, em determinados conjuntos de dados numéricos, os algarismos seguem uma distribuição específica conforme a sua posição dentro do número. A primeira formulação desta lei foi produzida por Newcomb [Newcomb 1881], ainda no século XIX, que observou:

> "Que os dez dígitos não ocorrem em igual frequência deve ser evidente a qualquer um que faça muito uso das tabelas de logaritmos e que perceba quão mais rápido ficam gastas as primeiras páginas do que as últimas. A primeira figura significativa é mais comumente 1 do que qualquer outro dígito, e a frequência diminui até o 9."

Tabela 1. Tabela de probabilidade de ocorrência dos algarismos para os primeiros dois dígitos significativos [Newcomb 1881].

Algarismo	1º	2º	Algarismo	1º	2º
0		0,1197	5	0,0792	0,0967
1	0,3010	0,1139	6	0,0669	0,0934
2	0,1761	0,1088	7	0,0580	0,0904
3	0,1249	0,1043	8	0,0512	0,0876
4	0,0969	0,1003	9	0,0458	0,0850

Além de discutir as possíveis razões para o fenômeno, Newcomb calculou a probabilidade da ocorrência dos algarismos nos primeiros dois dígitos significativos, reproduzida aqui na Tabela 1. Naturalmente, o primeiro dígito significativo de qualquer número nunca é zero, e portanto a probabilidade de sua ocorrência é nula neste caso. Newcomb concluiu ainda que "no caso da terceira figura a probabilidade vai ser quase a mesma para cada dígito, e para a quarta e restantes a diferença será quase inexistente."

Mais tarde, em 1938, Benford expandiu o trabalho de Newcomb, chamando o fenômeno de "lei dos números anômalos" [Benford 1938]. Em seu artigo, Benford torna explícitas as fórmulas que governam a distribuição dos dígitos. Em particular, a fórmula para o primeiro dígito é

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

em que $d \in D$, com $D = \{1, \ldots, 9\}$, é o algarismo em questão. Um gráfico de barras desta função pode ser visto na Figura 1.

A distribuição da Figura 1 pode ser utilizada na detecção de fraudes em dados numéricos [Bolton and Hand 2002, Durtschi et al. 2004, Tödter 2009]. Se é sabido que, em um determinado domínio, a distribuição do primeiro dígito significativo segue a distribuição de Benford, é possível afirmar que há fraude ao constatar-se que os números observados não seguem esta distribuição.

Informações numéricas das mais variadas naturezas são armazenadas em bancos de dados relacionais. A detecção de fraude nestes dados numéricos pode ser relevante em muitos domínios, especialmente o financeiro. É possível extrair os dados do banco de dados e processá-los em sistemas externos, porém todos os dados numéricos precisam ser exportados, já que não se pode fazer uma análise de qualidade sobre dados sumarizados, pois é preciso levar em consideração o primeiro dígito significativo de cada número envolvido. É mais conveniente, portanto, que essa análise seja feita dentro do próprio SGBD (sistema gerenciador de banco de dados). Desta forma, estatísticas para a detecção de fraude ficam disponíveis automaticamente para todos os clientes do SGBD, assegurando-se também uma bom desempenho computacional.

A linguagem SQL (linguagem de consulta estruturada), utilizada para fazer consultas ao banco de dados, provê uma série de funções que podem ser usadas para o cálculo de estatísticas, como é o caso das funções `SUM` para a soma de valores e `AVG` para a média. Estatísticas mais complexas, como o desvio padrão, também são suportadas por alguns SGBDs. As estatísticas utilizadas para a detecção de fraude através da distribuição de

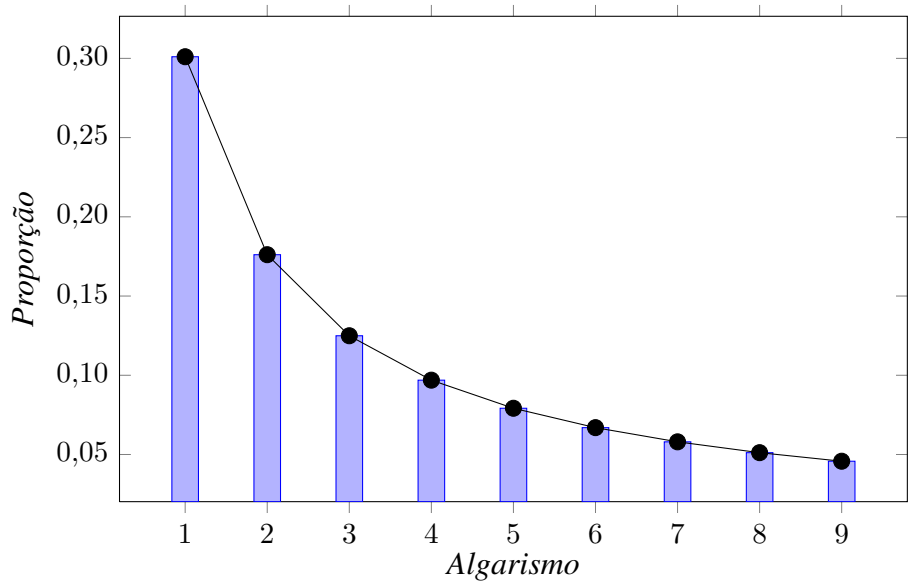

Figura 1. Distribuição do primeiro dígito significativo em conjuntos numéricos de acordo com a lei de Benford.

Benford, porém, são complexas e difíceis de ser implementadas em uma consulta simples sem utilizar tabelas temporárias.

Este artigo propõe uma extensão ao MySQL [Oracle Corporation 2012] que disponibiliza funções estatísticas capazes de auxiliar na detecção de fraudes em dados numéricos. O código-fonte da extensão pode ser acessado no endereço `https://github.com/adiel-mittmann/mysql-benford`. Mostra-se neste trabalho que dados numéricos reais tendem a seguir a distribuição de Benford, e são apresentados também exemplos de dados que não a seguem, por fraude ou outro motivo.

A Seção 2 introduz as estatísticas que podem ser utilizadas para detectar a conformidade de conjuntos numéricos à distribuição de Benford; a Seção 3 apresenta a extensão ao MySQL que implementa estas estatísticas; a Seção 4 mostra resultados práticos obtidos com a extensão; por fim, os comentários finais são feitos na Seção 5.

2. Estatísticas para a distribuição de Benford

Para que se possa afirmar que o primeiro digito significativo de um conjunto numérico não segue a distribuição de Benford (e que, portanto, há provável fraude em certos cenários), é preciso que sejam definidas estatísticas e seus valores críticos. Quando testes estatísticos são realizados para verificar conformidade com a distribuição de Benford, parte-se da hipótese nula, H_0, de que a os dados observados seguem a distribuição de Benford; ao rejeitar-se H_0, fica estatisticamente provado que os dados não seguem esta distribuição.

Uma estatística que pode ser utilizada para este fim é a χ^2, cuja fórmula é

$$\chi^2 = \sum_{d=1}^{9} \frac{[O_n(d) - E_n(d)]^2}{E_n(d)}, \tag{2}$$

em que $O_n(d)$ é o número observado de casos em que d é o primeiro dígito significativo

Tabela 2. Valores críticos propostos para a estatística m^* [Morrow 2010].

	$\alpha = 0,10$	$\alpha = 0,05$	$\alpha = 0,01$
m^*	0,851	0,967	1,212

e $E_n(d)$ é o número esperado de casos, ou seja,

$$E_n(d) = P(d)N, \tag{3}$$

em que N é o número total de observações. O valor da estatística pode então ser comparado com os valores críticos da distribuição χ^2 com $9 - 1 = 8$ graus de liberdade. Com um nível de significância de $\alpha = 0,01$, por exemplo, rejeita-se H_0 quando o valor de χ^2 é superior a $20,090$, provando-se então que os dados não seguem a distribuição de Benford.

Outra estatística, proposta por Leemis [Leemis et al. 2000] e expandida mais tarde por Morrow [Morrow 2010], é

$$m^* = \sqrt{N} \max_{d \in D} |O_f(d) - E_f(d)|, \tag{4}$$

em que $O_f(d)$ é a frequência observada de d como o primeiro dígito significativo e $E_f(d) = P(d)$ é a frequência esperada deste dígito. Valores críticos para os níveis de significância $0,10$, $0,05$ e $0,01$ foram calculados por Morrow [Morrow 2010] e constam da Tabela 2. Nota-se que tanto χ^2 quanto m^* dependem do número de observações feitas, de maneira que, quanto maior for N, mais se espera que o primeiro dígito significativo siga a distribuição de Benford.

Uma outra maneira de avaliar a discrepância entre dados observados e a distribuição de Benford é utilizar a medida original proposta por Leemis [Leemis et al. 2000], qual seja,

$$m = \max_{d \in D} |O_f(d) - E_f(d)|. \tag{5}$$

O valor de m é, portanto, a maior diferença encontrada entre a frequência observada e a esperada. É uma medida, contudo, que não depende de N, o que impede que seja utilizada em testes estatísticos. Sua utilidade surge quando se quer comparar dados que seguem uma distribuição similar, porém não idêntica, à de Benford. Em tais comparações, m fornece um índice de dissimilaridade.

3. Extensão ao MySQL

Uma extensão foi desenvolvida para prover ao MySQL a capacidade de avaliar se determinados conjuntos de números seguem ou não a distribuição de Benford. A maneira utilizada para realizar a extensão foi a criação de três funções definidas pelo usuário (*user-defined functions*, ou UDFs).

As UDFs foram escritas em C e são compiladas para formar um *plugin* para o MySQL. As três funções são agregadas, ou seja, não operam em um único número, mas sim num conjunto deles. As três funções são:

- **`BENF_CHISQ`**: realiza um teste χ^2 para os dados fornecidos, com $\alpha = 0,001$ e 8 graus de liberdade. A função retorna 1 quando a hipótese nula é rejeitada pelo teste

Listagem 1. Comandos para disponibilizar as funções propostas no MySQL.

```
CREATE AGGREGATE FUNCTION BENF_CHISQ RETURNS INTEGER SONAME 'benford.so';
CREATE AGGREGATE FUNCTION BENF_MSTAR RETURNS INTEGER SONAME 'benford.so';
CREATE AGGREGATE FUNCTION BENF_MAX   RETURNS REAL    SONAME 'benford.so';
```

Tabela 3. Sumário de estatísticas para os três conjuntos de dados.

Conjunto	N	χ^2	m^*	m
JOINVILLE	$14,196$	1	1	0,028
USDA	$409,288$	1	1	0,016
FINANCEIRO	$6,897$	1	1	0,040

Listagem 2. Consulta SQL que gera a Tabela 3.

```
SELECT COUNT(*), BENF_CHISQ(valor_pago), BENF_MSTAR(valor_pago),
       BENF_MAX(valor_pago) FROM empenhos;
SELECT COUNT(*), BENF_CHISQ(Nutr_val), BENF_MSTAR(nutr_val),
       BENF_MAX(nutr_val) FROM NUT_DATA;
SELECT COUNT(*), BENF_CHISQ(amount), BENF_MSTAR(amount),
       BENF_MAX(amount) FROM transactions;
```

χ^2, ou seja, há evidência estatística de que os dados não seguem a distribuição de Benford. Caso a hipótese nula não seja rejeitada, retorna 0.

- **BENF_MSTAR**: executa um teste estatístico com base em m^*, com $\alpha = 0,01$. O valor de retorno tem o mesmo significado que no caso de **BENF_CHISQ**: o valor 1 indica que, estatisticamente, os dados não seguem a distribuição de Benford, enquanto que 0 significa que não há evidência de que os dados não a seguem.
- **BENF_MAX**: o valor da medida m da Equação 5.

Após a geração da biblioteca dinâmica a partir do código-fonte C, as funções podem ser disponibilizadas no MySQL através do comando `CREATE FUNCTION` para UDFs. Os comandos necessários para as três funções utilizadas neste artigo são mostrados na Listagem 1.

Nota-se que as funções são `AGGREGATE` e que as duas primeiras retornam valores do tipo `INTEGER`, enquanto a terceira retorna `REAL`. As duas primeiras produzem valores booleanos, mas, como booleanos no MySQL nada mais são do que inteiros com zero correspondendo a falso, os valores destas funções são na prática inteiros. A última função, **BENF_MAX** retorna efetivamente um número real.

O código da Listagem 1 utiliza o valor `'benford.so'` para o parâmetro `SONAME` porque os experimentos deste artigo foram executados em ambiente Linux. Em outros sistemas, este nome pode variar; no Windows, por exemplo, o nome da biblioteca dinâmica seria `'benford.dll'`.

4. Experimentos

Foram realizados experimentos com três conjuntos de dados distintos, dois deles reais e um hipotético. Os conjuntos de dados, cujas propriedades estatísticas estão sumarizadas na Tabela 3, são os seguintes:

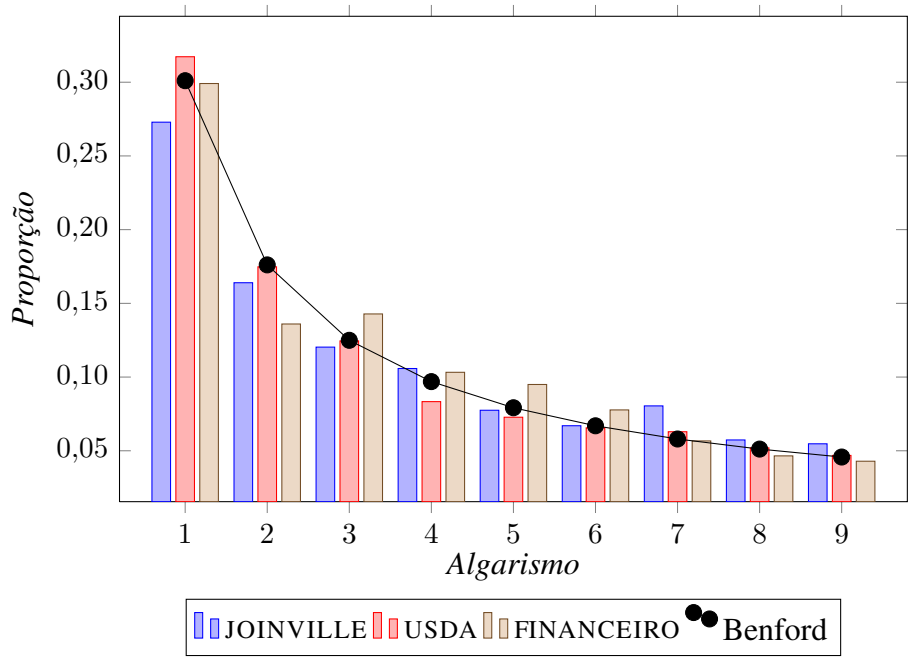

Figura 2. Distribuição do primeiro dígito significativo nos três conjuntos numéricos. Nota-se que todos seguem aproximadamente a distribuição de Benford.

- JOINVILLE: valores dos empenhos pagos pela prefeitura da cidade de Joinville – SC entre 01/01/2012 e 30/06/2012. Esta cidade foi escolhida por disponibilizar de forma acessível seus dados financeiros.
  Fonte: `http://www.joinville.sc.gov.br/`.
- USDA: banco de dados de valores nutricionais de alimentos do departamento de agricultura dos EUA (USDA).
  Fonte: `http://sourceforge.net/projects/usdanutr/`.
- FINANCEIRO: valores financeiros hipotéticos, extraídos de um banco de dados real porém com valores e datas alterados de forma a torná-los anônimos e utilizáveis nos experimentos deste artigo.

Todos os valores zero dos conjuntos foram excluídos, pois eles não contém nenhum dígito significativo e portanto não são analisáveis pela lei de Benford. Colunas contendo valores zero ou valores nulos podem ser passadas às UDFs descritas neste artigo, que os ignoram, mas preferiu-se aqui excluir estas linhas indesejáveis para que não fosse inflado artificialmente o número N de observações.

Percebe-se, através da Tabela 3, que, apesar de a função **`BENF_MAX`** deixar claro que as discrepâncias entre os dados observados e a distribuição de Benford são pequenas, elas são contudo estatisticamente significativas, já que as funções **`BENF_CHISQ`** e **`BENF_MSTAR`** rejeitam a hipótese nula de que os dados seguem esta distribuição. De fato, ao observar graficamente a distribuição do primeiro dígito nestes conjuntos de dados na Figura 2, pode-se ver que todos aproximam-se da distribuição de Benford, mas nenhum é exatamente igual.

Com frequência, ao analisarem-se grandes quantidades de observações, é mais prático utilizar a medida m, implementada pela função **`BENF_MAX`**, do que os testes estatísticos. Estes últimos são capazes de demonstrar que certos dados não seguem a distri-

buição de Benford, mas dizem pouco no que se refere ao quão perto os dados chegam de segui-la.

A Tabela 4 mostra a análise dos dados do conjunto JOINVILLE agrupados por mês. Apesar de a hipótese nula ser rejeitada com χ^2 e m^* para quase todos os meses, percebe-se que há uma boa uniformidade nos valores da medida m. O mês de maio apresenta o maior valor de m, mas ele não é suficientemente grande para levantar suspeitas.

A Tabela 5 mostra, no contexto do conjunto USDA, as estatísticas computadas para os cinco nutrientes com maior e menor adequação à distribuição de Benford. Esta tabela evidencia que, apesar da regularidade vista na Figura 2, os nutrientes mostram uma grande variação individual. Enquanto as cinco primeira linhas do resultado contêm nutrientes para os quais a hipótese nula é rejeitada com o teste m^*, os cinco nutrientes finais na lista exibem altos valores da estatística m. Valores neste patamares mostram, especialmente quando comparados aos valores dos outros nutrientes, que existe algo a ser investigado.

Tomando-se o "nutriente" água para investigação, obtém-se o gráfico de barras da Figura 3, que mostra a distribuição do primeiro dígito significativo da quantidade de água nos alimentos pesquisados. Percebe-se que, de fato, a distribuição obtida não segue, de maneira bastante evidente, a distribuição de Benford. Um exame mais minucioso do conjunto de dados USDA revela que os dados nutricionais são calculados com base em uma porção padronizada de 100 g. É intuitivo assumir que maior parte dos alimentos tem uma quantidade significativamente grande de água em sua composição; o que a Figura 3 revela, portanto, é que há muitos alimentos em que o percentual de água varia entre 50 e 80%.

Já os dados do conjunto FINANCEIRO podem ser agrupados por usuário, o que é mostrado na Tabela 6. Percebe-se que os lançamentos dos usuários tem o valor de m entre 2 e 4% — à exceção do usuário José, cujo m chega a 13%. A Figura 4 mostra graficamente a distribuição obtida com os lançamentos deste usuário. Este valor por si só não demonstra fraude, mas indica que uma investigação mais detalhada precisa ser levada a cabo. Se todos os usuários do sistema em questão tem exatamente o mesmo trabalho, é muito provável que José esteja agindo de má fé; por outro lado, é possível que José seja responsável por cadastrar valores que contenham alguma especificidade, como é o caso da água no conjunto USDA.

5. Conclusões

Vários conjuntos numéricos utilizados na prática seguem a distribuição de Benford. Os três conjuntos examinados neste artigo, JOINVILLE, USDA e FINANCEIRO aproximam-se desta distribuição quando todos os valores neles contidos são analisados graficamente. De um ponto de vista estritamente estatístico, contudo, pode-se demonstrar que eles não seguem a distribuição de Benford.

Por serem os testes estatísticos bastante exigentes, na prática pode-se usar a medida da máxima divergência entre o observado e o esperado. Assim a suspeita de fraude surge quando a divergência obtida é superior ao habitual. É o que foi observado no conjunto FINANCEIRO, em que a divergência máxima observada entre os funcionários não passa de 4%, à exceção de um em que a divergência é de 13%. Esta diferença significativa gera uma suspeita que precisa ser investigada.

Tabela 4. Estatísticas do conjunto de dados JOINVILLE, agrupadas por ano e mês. Somente meses com ao menos 1.000 pagamentos foram considerados.

Linha	Ano	Mês	Soma	N	χ^2	m^*	m
1	2012	1	226.670.564,58	2.716	1	0	0,014
2	2012	2	68.039.998,25	1.661	1	1	0,042
3	2012	3	41.725.962,08	1.787	1	1	0,041
4	2012	4	36.915.256,43	1.720	1	1	0,038
5	2012	5	36.045.489,58	1.439	1	1	0,069

Listagem 3. Consulta que gera a Tabela 4.

```
SELECT YEAR(data_emissao), MONTH(data_emissao), SUM(valor_pago), COUNT(*),
       BENF_CHISQ(valor_pago), BENF_MSTAR(valor_pago), BENF_MAX(valor_pago)
FROM empenhos
GROUP BY YEAR(data_emissao), MONTH(data_emissao) HAVING COUNT(*) >= 1000;
```

Tabela 5. Estatísticas para os nutrientes do conjunto de dados USDA, com valores ordenados pela estatística m. São exibidos os cinco menores e os cinco maiores valores. Apenas nutrientes com pelo menos 1.000 observações foram incluídos.

Primeiros 5					Últimos 5				
Linha	Nome	χ^2	m^*	m	*Linha*	Nome	χ^2	m^*	m
1	F18D0	0	0	0,0072	*75*	CYS_G	1	1	0,17
2	FASAT	0	0	0,0083	*76*	MG	1	1	0,18
3	VITC	0	0	0,010	*77*	ALA_G	1	1	0,21
4	VITA_IU	0	0	0,010	*78*	ARG_G	1	1	0,21
5	F18D2	1	0	0,011	*79*	WATER	1	1	0,23

Listagem 4. Consulta que gera a Tabela 5.

```
SELECT Tagname, BENF_CHISQ(Nutr_Val), BENF_MSTAR(Nutr_Val), BENF_MAX(Nutr_Val)
FROM NUT_DATA JOIN NUTR_DEF ON NUT_DATA.Nutr_No = NUTR_DEF.Nutr_No
GROUP BY NUT_DATA.Nutr_No HAVING COUNT(*) >= 1000 ORDER BY BENF_MAX(Nutr_Val);
```

Tabela 6. Estatísticas para o conjunto FINANCEIRO, com valores agrupados por usuário. O valor em negrito mostra comportamento anômalo.

Linha	Usuário	N	χ^2	m^*	m
1	Maria	1.345	1	0	0,026
2	José	1.357	1	1	**0,13**
3	Antônio	1.385	1	1	0,035
4	João	1.365	1	1	0,033
5	Francisco	1.445	1	1	0,040

Listagem 5. Consulta que gera a Tabela 6.

```
SELECT name, COUNT(*), BENF_CHISQ(amount), BENF_MSTAR(amount), BENF_MAX(amount)
FROM transactions JOIN users ON transactions.user_id = users.id
GROUP BY users.id;
```

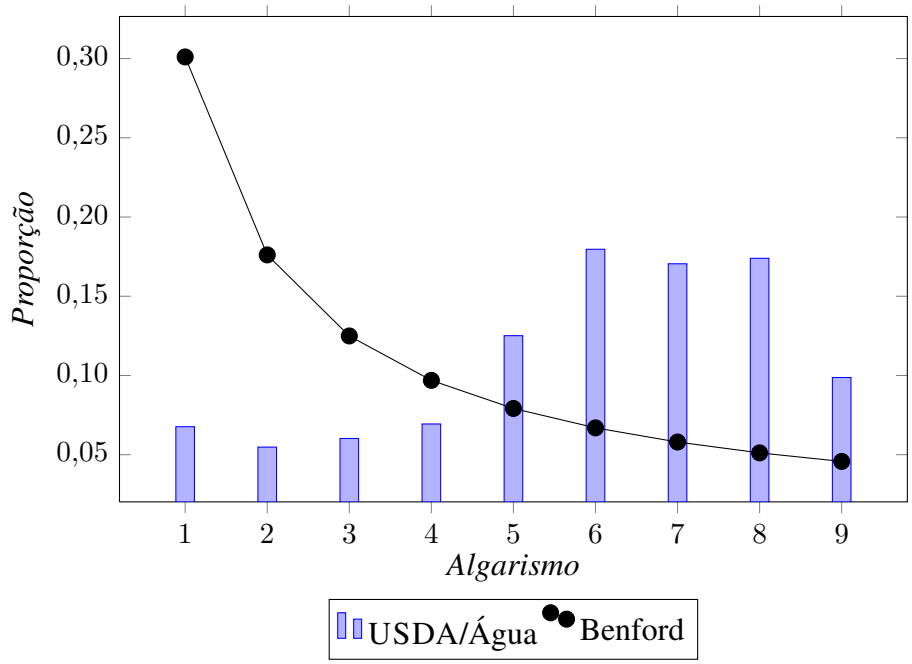

Figura 3. Distribuição do primeiro dígito significativo da quantidade de água em alimentos, segundo o conjunto de dados USDA.

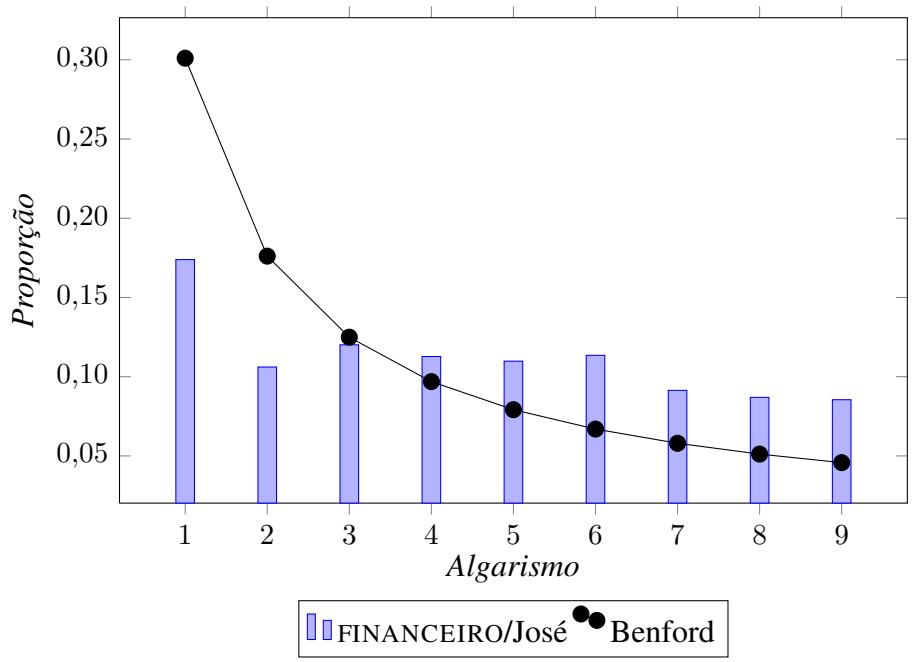

Figura 4. Distribuição do primeiro dígito significativo do usuário José, do conjunto de dados FINANCEIRO. Claramente os números lançados por este usuário não são compatíveis com a distribuição de Benford.

Existem casos em que naturalmente conjuntos numéricos não seguem a distribuição de Benford. Em geral, dados seguem esta distribuição quando os valores abrangem várias ordens de magnitude e não estão limitados a uma faixa específica. A água, no conjunto USDA, é um bom exemplo: a quantidade de água em 100 g de alimentos está limitada a 100 g e 83% dos valores estão acima de 10%. Fica claro que é preciso levar em consideração a natureza dos dados para julgar se a divergência observada é indicativo de um comportamento suspeito.

A extensão proposta neste artigo facilita a geração de relatórios em diversas aplicações. Um aplicativo que queira exibir ao usuário dados suspeitos pode fazê-lo com uma única consulta SQL, de forma simples e intuitiva, beneficiando-se da possibilidade de fazer agrupamentos e ordenação na própria consulta.

Referências

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572.

Bolton, R. J. and Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 17:235–249.

Durtschi, C., Hillison, W., and Pacini, C. (2004). The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting*, 5:17–34.

Leemis, L. M., Schmeiser, B. W., and Evans, D. L. (2000). Survival distributions satisfying benford's law. *The American Statistician*, 54:236–541.

Morrow, J. (2010). Benford's law, families of distributions and a test basis. `http://jmorrow.net/projects/benford/benfordMain.pdf`.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4:39–40.

Oracle Corporation (2012). *MySQL 5.5 Reference Manual*.

Tödter, K.-H. (2009). Benford's law as an indicator of fraud in economics. *German Economic Review*, 10:339–351.