ANÁLISE DOS DÍGITOS INDUSTRIAIS BASEADA NA LEI DE BENFORD E SUA APLICAÇÃO UTILIZANDO ROTINAS COMPUTACIONAIS

Raquel Cymrot*
Felipe Rigos da Rocha**
Dominique Santos Ferreira*

Resumo

O presente trabalho aborda a aplicação de uma ferramenta estatística utilizada na verificação da confiabilidade dos valores de amostras baseada na distribuição de Benford. Essa distribuição é composta das probabilidades de ocorrência de cada um dos números inteiros de 1 a 9 para o primeiro dígito significativo dos valores de uma amostra, formando uma curva logarítmica decrescente. O estudo se deu em dados relacionados com o setor industrial, em consonância com a atual tendência global da produção sustentável. O trabalho apresenta também duas rotinas computacionais para a implementação da análise dos dígitos em dois *softwares*: Excel e R.

Palavras-chave: Lei de Benford, confiabilidade, controle estatístico de processos.

* Universidade Presbiteriana Mackenzie (UPM).
** Whirlpool Latin America.

1 INTRODUÇÃO

Em processos industriais, é de grande importância a detecção de algum tipo de fraude, intencional ou não, na coleta de dados. No que diz respeito ao tratamento de dados provenientes da indústria, o termo fraude passa a ter significado mais amplo, englobando dados fora da distribuição causados por erros provenientes de processos computacionais, observações humanas ou manipulação de dados.

Em muitos processos, para mensuração dos dados, são usadas transmissões eletrônicas, manipulações computacionais e outros procedimentos de análises físicas e químicas. Tal situação aumenta a chance de haver uma alteração nos valores por conta de erros ocasionados durante o processo. Tais erros, por sua vez, podem alterar não somente a produtividade de uma empresa, mas também aspectos impactantes à natureza, como o desperdício de energia, o aumento de refugo e lixo industrial e a diminuição da eficácia de sistemas de filtragem. A detecção de tais erros pode levar a um redirecionamento de recursos para investimentos e aprimoramentos dos processos.

O presente trabalho propõe um estudo em uma área vital para a tomada de decisão derivada de análises estatísticas: a análise da confiabilidade de dados em processos industriais, utilizando um método estatístico de análise de dígitos baseado na Lei de Benford, uma distribuição anômala dos números inteiros de 1 a 9 ou de 0 a 9, objeto desta pesquisa. Uma vez comprovada a aderência de dados provenientes de certo fenômeno relacionado a um processo industrial à distribuição de Benford, outras amostras derivadas do mesmo fenômeno seguirão a mesma distribuição.

2 A LEI DE BENFORD

Em 1881, Simon Newcomb publicou artigo no qual relatou uma característica interessante encontrada em um livro de logaritmos – o desgaste das bordas diminuía com o decréscimo das páginas. Porém, somente em 1938, Frank Benford, físico da General Electric Company, publicou um artigo descrevendo o mesmo fenômeno que ele observou em mais de 20 mil dados analisados em diferentes amostras de diversas fontes, como distâncias de rios, estatísticas de beisebol, números de endereços, entre outras (HILL, 1996, 1998, 1999).

A distribuição de Benford se dá nos ditos dígitos significativos, isto é, nos dígitos à extrema esquerda dos valores, com exceção do zero, independentemente do número

de algarismos de cada valor da amostra. Benford notou empiricamente que, se analisado o primeiro dígito significativo dos valores de certas amostras, a probabilidade dos números naturais de 1 a 9 em certas distribuições não era de um para nove (11,11%), como esperado intuitivamente, mas que o número 1 tinha 30% de probabilidade de ocorrer, o número 2 tinha 17%, e assim por diante, formando uma curva logarítmica decrescente. Esse fenômeno é chamado em estatística de anomalia, e sua distribuição ficou conhecida como distribuição de Benford ou Lei de Benford, aplicada por meio da análise dos dígitos (HILL, 1996, 1998, 1999; NIGRINI, 1996).

Para se utilizar a Lei de Benford, os dados do fenômeno em estudo devem seguir algumas condições: os dados analisados devem descrever medidas de fenômenos similares, é necessário um grande número de observações por amostras (na literatura atual, não foi encontrado o uso da Lei de Benford em amostras com menos de 100 dados) (HALES et al., 2008), e os dados devem ser aleatórios e independentes, tendo como origem uma fonte natural, por conta de algum processo ou observação sem manipulação direta ou interferência humana. Tal condição resulta do fato de que não é possível aleatorizar uma série de dados de forma natural, sempre havendo tendências para certos valores (HILL, 1996, 1998, 1999). Para se obter uma aleatoriedade mais eficaz, além de o tamanho da amostra ser elevado, convém utilizar medidas ou observações de diferentes locais ou de diferentes períodos.

Hill (1996) apresentou o cálculo da função de probabilidade da distribuição de Benford, conforme Equação (1):

$$P(Di) = \log [1 + (1/Di)] \quad (1)$$

em que Di é o valor do primeiro dígito significativo, inteiro e não nulo; e P, a sua probabilidade de ocorrência.

A teoria se estendeu, e, de maneira geral, o primeiro dígito significativo pode ser formado por um ou mais algarismos. A função dessa generalização, também logarítmica, é apresentada na Equação (2):

$$P(Di \ldots Dk) = \log [1 + (1/(Di \ldots Dk)] \quad (2)$$

Encontrou-se também a função de probabilidade das distribuições dos dígitos subsequentes, como o segundo, terceiro e quarto dígitos significativos, mostradas nas equações 3, 4 e 5:

$$P(X = D2_i) = \sum_{k=1}^{9} \log_{10} (1+1/(10D1_k + D2_i)) \quad (3)$$

para $1 \leq D1_k \leq 9$; $0 \leq D2_i \leq 9$;

$$P(X = D3_j) = \sum_{k=1}^{9} \sum_{i=0}^{9} \log_{10} (1+1/(100D1_k + 10D2_i + D3_j)) \quad (4)$$

para $1 \leq D1_k \leq 9$; $0 \leq D2_i \leq 9$; $0 \leq D3_j \leq 9$;

$$P(X = D4_l) = \sum_{k=1}^{9} \sum_{i=0}^{9} \sum_{l=0}^{9} \log_{10} (1+1/(1000D1_k + 100D2_i + 10D3_j + D4_l)) \quad (5)$$

para $1 \leq D1_k \leq 9$; $0 \leq D2_i \leq 9$; $0 \leq D3_j \leq 9$; $0 \leq D4_l \leq 9$.

A Tabela 1 apresenta a distribuição do primeiro, segundo, terceiro e quarto dígitos, de acordo com a Lei de Benford.

TABELA 1

Distribuição de probabilidade dos primeiros quatro dígitos, conforme a Lei de Benford

Dígito	Probab. 1º dígito	Probab. 2º dígito	Probab. 3º dígito	Probab. 4º dígito
0	–	0,119679	0,101784	0,100176
1	0,301030	0,113890	0,101376	0,100137
2	0,176091	0,108821	0,100972	0,100098
3	0,124939	0,104330	0,100573	0,100059
4	0,096910	0,100308	0,100178	0,100019
5	0,079181	0,096677	0,099788	0,099980
6	0,066947	0,093375	0,099401	0,099941
7	0,057992	0,090352	0,099019	0,099902
8	0,051153	0,087570	0,098641	0,099863
9	0,045757	0,084997	0,098267	0,099824

Embora a teoria tenha sido proposta em 1938, a utilização prática da Lei de Benford somente tomou corpo a partir da década de 1980, principalmente no setor contábil. Sua aplicação na indústria ainda é nova, com poucos artigos científicos publicados no mundo (HÜRLIMANN, 2006).

A distribuição de Benford tem duas propriedades interessantes. A primeira é o fato de ser a única distribuição de dígitos significativos de dados reais que é invariante a mudanças de escala, ou seja, quando se multiplicam os valores de uma amostra por uma constante, a distribuição não se altera. Por exemplo, no estudo de dados provenientes de valores monetários que seguem a distribuição de Benford, a mudança de moeda não alterará as probabilidades dos dígitos significativos encontradas inicialmente. A segunda diz respeito à mudança de base dos dados. A mudança de base da função logarítmica não afeta a distribuição dos dígitos em relação à distribuição de Benford. O matemático Peter Schatte determinou que, com base na lei de Benford, o computador que minimiza o espaço de armazenagem (dentre todos os computadores

com base binária) é o de base 8. Tal característica está tendo uma grande repercussão na área da informática, fazendo com que vários pesquisadores explorem o uso de computadores com base logarítmica, possibilitando acelerar o processamento de dados (HILL, 1996, 1998, 1999).

Em alguns casos, observa-se que conjuntos de dados com pequena amplitude podem ser analisados mediante as funções descritas anteriormente. Foi levantada a questão de que certas distribuições seguem a Lei de Benford, porém não a partir dos primeiros dígitos (HALES et al., 2008). Essa abordagem é determinante para o presente trabalho pelo fato de as indústrias trabalharem com especificações em seus produtos e processos, resultando em valores dentro de uma amplitude predefinida. Isso faz com que alguns números nunca apareçam nos primeiros dígitos significativos.

3 O MÉTODO PARA A ANÁLISE DE DÍGITOS

A aplicação do método na busca de dados alterados (intencionalmente ou não) se dá por meio da comparação dos valores observados na amostra para o dígito em análise, em relação aos valores esperados para esse mesmo dígito, segundo a distribuição de Benford. Tal comparação se faz com a utilização de um teste de aderência.

O teste de aderência quiquadrado é realizado utilizando a estatística apresentada na Equação (6):

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} = - n \qquad (6)$$

com O_i igual à frequência observada na amostra, E_i igual à frequência esperada pela distribuição de Benford, k igual ao número total possível de dígitos (9 para o primeiro e 10 para os demais dígitos significativos) e n igual ao número de observações na amostra. O valor encontrado é comparado ao da distribuição quiquadrado, em um nível de significância estabelecido com (k – 1) graus de liberdade (SIEGEL; CASTELLAN JR., 2008).

Embora, em geral, o nível de significância (α) mais utilizado para amostras do tamanho usado na verificação da Lei de Benford seja de 5%, neste estudo se utilizou um nível de significância de 1%, conforme artigo de Hales et al. (2008), uma vez que se deseja minimizar a probabilidade de detectar falsos indícios de alterações. Caso a distribuição de algum dígito, em um fenômeno para o qual já é conhecido que seus valores seguem a lei de Benford, não siga tal distribuição, conclui-se que algum número (de 1 a 9 ou de 0 a 9) foi mais ou menos frequente do que deveria ser.

Quando $k \geq 2$, o teste quiquadrado não deverá ser utilizado se houver qualquer frequência esperada inferior a 1 ou se mais de 20% das frequências esperadas forem menores que 5 (SIEGEL; CASTELLAN JR., 2008). Nessas condições, as probabilidades apresentadas na Tabela 1 implicam que os tamanhos das amostras devem ser respectivamente maiores que 98, 56, 51 e 51 para a análise dos primeiros, segundos, terceiros e quartos dígitos. Vale salientar que, fixando o nível de significância do teste quiquadrado em 0,01, quanto menor for o tamanho da amostra, menor será o poder do teste, isto é, a probabilidade de o teste detectar a alteração nos dados quando de fato esta ocorreu.

Para todo o teste, foi calculado seu nível descritivo (valor-P) que é a probabilidade de obter estimativas mais desfavoráveis ou extremas do que a obtida na amostra (no caso, valores maiores que o quiquadrado observado), sendo rejeitadas as hipóteses com nível descritivo inferior a 0,01, o nível de significância adotado na pesquisa (MAGALHÃES; LIMA, 2010).

4 OS DADOS UTILIZADOS E SUAS RESPECTIVAS ANÁLISES

Neste trabalho, analisaram-se três conjuntos de dados em relação à distribuição de Benford. Dois conjuntos são derivados da produção industrial provenientes da área de qualidade, na qual é utilizada a ferramenta estatística controle estatístico de processos (CEP), e outro conjunto analisou um parâmetro de poluição.

Para a obtenção dos dados industriais, foram seguidos os procedimentos éticos necessários, sendo esta pesquisa aprovada pela Comissão de Ética em Pesquisa da instituição.

Os dois primeiros conjuntos de dados foram obtidos de processos industriais, mais especificamente do setor de qualidade, a fim de verificar a aderência de dois diferentes tipos de medida – dimensão e teor de elemento químico –, estudando a aplicação do método dos dígitos integrado ao CEP. Um dos conjuntos contém uma característica que demandou uma abordagem não encontrada em nenhuma referência pesquisada pelos autores.

O terceiro conjunto de dados se refere à área ambiental, os quais constam no *site* da Companhia de Tecnologia de Saneamento Ambiental (2012). Coletaram-se as partículas totais em suspensão (PTS – Hi-Vol) na estação de Cerqueira César, na cidade de São Paulo, no período de março de 2007 a março de 2012, sendo obtidos 276 valores (COMPANHIA DE TECNOLOGIA DE SANEAMENTO AMBIENTAL, 2012).

Em nenhum conjunto de dados estudados houve aderência do primeiro dígito à Lei de Benford; isso ocorreu porque, em conjunto de dados industriais, o primeiro dígito costuma estar preso a limites de especificação ou a limites naturais de ocorrência que abrangem uma amplitude limitada. A falta de análise do primeiro dígito não compromete a detecção de fraude, possível de ser detectada na análise dos demais dígitos.

4.1 Análises dos dados provenientes da área de qualidade de uma indústria

Da área de qualidade de uma indústria, foram estudados dois conjuntos de dados: teor de alumínio e comprimento do filtro.

Para o teor de alumínio, em porcentagem, em um produto químico industrial, tem-se uma especificação de 17,00% como limite inferior e de 22,00% como limite superior, com média de 19,50%. O primeiro dígito pode ter a ocorrência dos números 1 e 2, somente, e o segundo pode ser apenas os números 7, 8, 9 para o primeiro dígito igual a 1 e 0, 1 e 2 para o primeiro dígito igual a 2, impossibilitando a aderência à Lei de Benford.

A amostra analisada continha 147 medidas, todas com quatro dígitos (dois inteiros e dois decimais), e foram realizados, então, testes para o terceiro e quarto dígitos.

No teste do terceiro dígito, calculou-se um valor quiquadrado igual a 11,2723 (P = 0,2575), menor que o valor tabelado de 21,6660, não se rejeitando a hipótese de aderência à Lei de Benford. No teste do quarto dígito, calculou-se um valor quiquadrado igual a 26,1726 (P = 0,0019), maior que o valor tabelado de 21,6660, rejeitando-se, consequentemente, a hipótese de aderência à Lei de Benford.

Diante dessa observação, foi investigado junto à empresa um possível motivo para o resultado negativo. Descobriu-se que, no registro dos dados, o último dígito resulta de arredondamentos feitos pelos operadores, ou seja, há intervenção humana, o que altera a aleatoriedade.

O conjunto de dados de comprimento do filtro possui uma característica interessante que possibilitou a aplicação de uma abordagem não encontrada nas referências pesquisadas pelos autores.

As especificações inferior e superior para o comprimento desse componente são, respectivamente, iguais a 98,50 mm e 101,00 mm, com média igual a 99,75 mm. Nota-se que os valores variam de dois a três dígitos antes da vírgula (98 e 101). Dessa forma, o terceiro dígito dos valores abaixo de 100,00 mm tem condição de aderir à Lei de Benford, entretanto o terceiro dígito dos valores iguais a 100,00 mm ou acima desse valor pode ter como algarismos, de acordo com as especificações, apenas o zero e o 1, fazendo com que a condição de aleatoriedade se perca quando se considera a amostra como um todo.

Tal fato motivou a concepção de uma abordagem alternativa, considerando como terceiro dígito aquele imediatamente após a vírgula, para todos os valores da amostra, inclusive para os que têm três algarismos antes da vírgula. Dessa forma, o teste é feito em toda a amostra, sobre os primeiros valores livres de limites externos impostos.

No teste do terceiro dígito, calculou-se um valor quiquadrado igual a 16,1048, menor que o valor tabelado de 21,6660 (P = 0,0647), e, no teste do quarto dígito, calculou-se um valor quiquadrado igual a 14,5037, menor que o valor tabelado de 21,6660 (P = 0,1055). Segundo essa nova abordagem, houve aderência para o terceiro e quarto dígitos, confirmando assim a aleatoriedade para esses dígitos, não havendo indícios de interferência humana ou proveniente de componentes do processo.

4.2 Análise dos dados referentes à poluição

A amostra estudada continha 276 valores das partículas totais em suspensão (PTS – Hi-Vol), coletadas na estação de Cerqueira César, na cidade de São Paulo, entre março de 2007 e março de 2012. Para o primeiro dígito, o valor do quiquadrado observado foi igual a 74,2169, superior ao quiquadrado crítico igual a 20,0902 (P = 0,000), rejeitando-se a hipótese de aderência à distribuição de Benford. Tal rejeição ocorreu porque o primeiro dígito está vinculado às quantidades usuais de tais partículas na região estudada. Para o segundo dígito, o valor do quiquadrado observado foi igual a 13,4981, inferior ao quiquadrado crítico igual a 21,6660 (P = 0,023), logo não se rejeita a hipótese de aderência à distribuição de Benford, havendo, portanto, aleatoriedade.

5 IMPLEMENTAÇÃO DA ANÁLISE DOS DÍGITOS UTILIZANDO O EXCEL E O *SOFTWARE*

Programas computacionais proporcionam maior praticidade e eficácia em abordagens que requerem tempo e esforço. O objetivo desta seção é facilitar a utilização da Lei de Benford na indústria e indicar como devem ser realizados os cálculos necessários por meio de rotinas computacionais nos *softwares* Excel e R.

O uso de ferramentas estatísticas em planilhas eletrônicas é bem difundido, uma vez que os resultados obtidos fornecem informações que dão suporte a estratégias e decisões. A escolha do Excel justifica-se por sua consolidação no mercado em diversas aplicações, por deter sofisticados recursos, interface amigável e alta integração com os

demais aplicativos existentes, nos ambientes para o qual foi desenvolvido (ALMIRON et al., 2010; MIGLIOLI; OSTANEL; TACHIBANA, 2004).

Já o R é um sistema livre e colaborativo para cálculos estatísticos e construção de gráficos. Consiste em uma linguagem de programação, um ambiente de desenvolvimento, acesso a um conjunto de funções predefinidas e capacidade de compilar programas descritos em arquivos *script* (HORNIK, 2011; R DEVELOPMENT CORE TEAM, 2011). Seu intuito é integrar ferramentas gráficas, de cálculo e de análise de dados em um ambiente de desenvolvimento de programação (R DEVELOPMENT CORE TEAM, 2012; VANCE, 2009).

Segundo artigo publicado no *The New York Times* (VANCE, 2009), desde sua criação, o programa tem ganhado adeptos em todo o mundo. Sua estrutura de programação não é complicada quando comparada a outras linguagens, fazendo com que diversos estatísticos, engenheiros e cientistas explorem sua capacidade. Empresas como Google, Pfizer, Merck, Bank of America, entre outras, estão utilizando o *software* para diversos fins. O que torna o R uma plataforma tão aceita no meio científico e empresarial é a possibilidade de adaptar e modificar sua utilização para os mais diversos objetivos e a vasta gama de material já produzido (VANCE, 2009).

Para início da aplicação das rotinas, será suposto que os dados a serem utilizados já estejam em uma planilha do Excel, digitados com título em uma única coluna (suponha que seja a coluna A).

Cabe salientar que as rotinas funcionam para conjuntos cujos dados têm o mesmo número de dígitos significativos. Se tal situação não ocorrer, sugere-se utilizar as rotinas para teste de aleatoriedade apenas para os dígitos presentes em todo o conjunto de dados. A aleatoriedade dos demais dígitos poderá ser testada utilizando somente os dados que contemplarem tal dígito. Exceção deve ser feita para dados semelhantes aos do comprimento do filtro, nos quais há valores com um dígito a mais devido ao aumento do número de dígitos significativos do limite inferior para o limite superior de especificação. Nesse caso, devem-se analisar apenas os dígitos a partir das restrições impostas pela especificação, mas lembrando de realizar os testes de aderência para a posição de origem destes quando considerados os números com menos dígitos significativos.

5.1 Implementação no Excel

Para realizar a análise de dígitos no Excel, primeiramente se faz necessário que o primeiro dígito de todos os valores medidos seja diferente de zero. Se isso não ocorrer, deverão ser multiplicados todos os valores por 10^k, sendo k tal que o primeiro dígito da menor medição seja diferente de zero. A seguir, faz-se necessária a separação dos dígitos de cada medição. Para separar o dígito escolhido, utilizam-se as fórmulas apresentadas a seguir, as quais devem ser digitadas na célula à direita do primeiro dado e

arrastadas para as células abaixo. Como há o rótulo (título) na primeira célula, o primeiro valor do conjunto de dados deve estar digitado na célula "A2".

- Primeiro dígito:

=SE(ESQUERDA(A2;1)=0;DIREITA(ESQUERDA(A2;3);1);ESQUERDA(A2;1))

- Segundo dígito:

=SE(NÚM.CARACT($A2)<=1;0;SE(ÉERROS(PROCURAR(",";$A2));EXT.TEXTO(A2;2;1);SE(PROCURAR(",";$A2)<=2;EXT.TEXTO(A2;3;1);EXT.TEXTO(A2;2;1))))

- Terceiro dígito:

=SE(NÚM.CARACT($A2)<=2;0;SE(ÉERROS(PROCURAR(",";$A2));EXT.TEXTO(A2;3;1);SE(PROCURAR(",";$A2)<=3;SE(NÚM.CARACT(A2)<=3;0;EXT.TEXTO(A2;4;1));EXT.TEXTO(A2;3;1))))

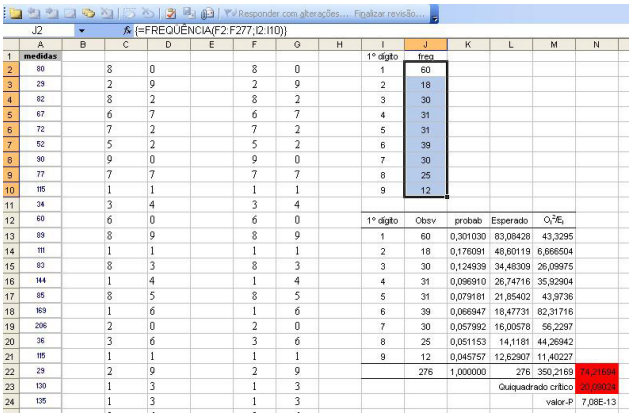
- Quarto dígito:

=SE(NÚM.CARACT($A2)<=3;0;SE(ÉERROS(PROCURAR(",";$A2));EXT.TEXTO(A2;4;1);SE(PROCURAR(",";$A2)<=4;SE(NÚM.CARACT(A2)<=4;0;EXT.TEXTO(A2;5;1));EXT.TEXTO(A2;4;1))))

A rotina descrita a seguir é ilustrada na Figura 1. Após a separação dos dígitos, o conjunto obtido deve ser copiado e colado como valores. Uma vez colado, esse novo conjunto deve ser todo selecionado, e, na caixa que aparecerá à esquerda, escolhe-se a opção "converter em número". Só então os dígitos estarão efetivamente separados, com seus respectivos valores numéricos.

A seguir, deve-se obter a frequência observada de cada dígito. Para o primeiro dígito, deve-se escrever em uma coluna os dígitos de 1 a 9. Na coluna do lado direito, devem-se selecionar as células vizinhas (no caso, as nove células). Então se faz necessário selecionar a função estatística frequência. Na caixa dessa função, devem-se completar a "Matriz_dados" com todos os primeiros dígitos que estão dispostos em uma coluna única e a "Matriz_bin" com os dígitos de 1 a 9 já digitados. A seguir, coloca-se o cursor no final da função e apertam-se simultaneamente as teclas "Control", "Shift" e "Enter". Dessa forma, será obtida a distribuição de frequências observada do primeiro dígito. Essa distribuição deve ser copiada e colada como valores em outras células. Ao seu lado, devem ser construídas as colunas com as probabilidades esperadas segundo a distribuição de Benford, com as frequências esperadas segundo a distribuição de Benford (probabilidades esperadas multiplicadas pelo número de medições) e com as frequências observadas ao quadrado divididas pelas frequências esperadas. A partir

dessa tabela, são finalizados os cálculos necessários para testar a aderência à distribuição de Benford. Para a análise dos dígitos posteriores, a única alteração é que os dígitos possíveis são agora de zero a 9.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	medidas								1º dígito	freq				
2	80		8	0		8	0		1	60				
3	29		2	9		2	9		2	18				
4	82		8	2		8	2		3	30				
5	67		6	7		6	7		4	31				
6	72		7	2		7	2		5	31				
7	52		5	2		5	2		6	39				
8	90		9	0		9	0		7	30				
9	77		7	7		7	7		8	25				
10	115		1	1		1	1		9	12				
11	34		3	4		3	4							
12	60		6	0		6	0		1º dígito	Obsv	probab	Esperado	O_i^2/E_i	
13	89		8	9		8	9		1	60	0,301030	83,08428	43,3295	
14	111		1	1		1	1		2	18	0,176091	48,60119	6,666504	
15	83		8	3		8	3		3	30	0,124939	34,48309	26,09975	
16	144		1	4		1	4		4	31	0,096910	26,74716	35,92904	
17	85		8	5		8	5		5	31	0,079181	21,85402	43,9736	
18	169		1	6		1	6		6	39	0,066947	18,47731	82,31716	
19	206		2	0		2	0		7	30	0,057992	16,00578	56,2297	
20	36		3	6		3	6		8	25	0,051153	14,1181	44,26942	
21	115		1	1		1	1		9	12	0,045757	12,62907	11,40227	
22	29		2	9		2	9			276	1,000000	276	350,2169	74,21691
23	130		1	3		1	3						Quiquadrado crítico	20,09024
24	135		1	3		1	3						valor-P	7,08E-13

Figura 1 Tela com exemplo de teste de aderência à Lei de Benford realizado.

5.2 Implementação no R

Foi realizada também uma rotina computacional com a utilização do *software* R. Os dados podem ser inseridos diretamente a partir do R Console ou importados de um arquivo do *software* Excel de extensão CSV (*comma separated values* – valores separados por vírgula).

Recomenda-se, em todo o processo de uso do R, a utilização de arquivos *scripts*. A rotina no R, desenvolvida neste trabalho, permite a realização da análise dos quatro primeiros dígitos. Se alguma das medições tiver menos de quatro dígitos significativos, deve-se somar 0,00001 em todas as medições para que as rotinas desenvolvidas funcionem adequadamente.

Para copiar os dados diretamente no R, estes não devem estar separados por dígito. Os dados devem ser copiados do Excel em um bloco de notas, devendo-se, a seguir, substituir as vírgulas por pontos utilizando os comandos editar e substituir. Nesse caso, utilizam-se, nesta ordem, as funções auxiliares contidas respectivamente nos apêndices C, D, E e F. Por último, executa-se a rotina contida no Apêndice A.

Para importar os dados do Excel, estes não devem estar separados por dígito. Importam-se os dados, lembrando-se de somar 0,00001, se necessário, a todos os valores, e se utilizam, nesta ordem, as funções auxiliares contidas respectivamente nos apêndices B, D, E e F. Por último, executa-se a rotina contida no Apêndice A.

A função *read.csv* é um dos modos de importar os dados do Excel para o R; ela gera um objeto *data frame*, sendo então necessário convertê-lo, pois a implementação realizada foi feita para uso de vetores. Consta no Apêndice B a conversão de um objeto *data frame* (o que se obtém quando se importam os dados do Excel com extensão csv para o R) para vetor. Essa conversão justifica-se pelo fato de o vetor ser, na maioria das vezes, mais facilmente manuseado do que um objeto *data frame*.

6 CONCLUSÃO

Quando, em um processo industrial, já se conhece previamente que o fenômeno tem por característica a aderência da distribuição de seus dígitos à distribuição de Benford, ao se constatar que a amostra proveniente desse fenômeno não adere a tal distribuição, surge uma oportunidade de investigar o motivo desse comportamento.

A amostra de teor de alumínio foi testada para o terceiro e quarto dígitos. A aderência do terceiro dígito comprovou a aleatoriedade e a não existência de interferências no processo. Já a não aderência do quarto dígito se deu por intervenção humana devido ao uso de aproximações, conforme informação obtida na indústria. Dessa forma, comprova-se a eficácia do uso da teoria da Lei de Benford na detecção de interferências nos dados.

A análise da amostra de comprimento de filtro apresentou uma nova abordagem quando os dados se mantêm entre os limites, inferior e superior, de uma especificação que tem diferentes números de dígitos significativos. Isso acarreta a necessidade de testar a aderência à Lei de Benford apenas nos dígitos livres de limites externos impostos. Novos trabalhos devem ser realizados para comprovar essa abordagem para outros conjuntos de dados com tais características.

Os gestores atuais vislumbram um desafio cada vez mais difícil que compõe a produção industrial em consonância com o meio ambiente: o alcance efetivo do desenvolvimento sustentável. Sabe-se, mais do que nunca, que os recursos naturais estão se esgotando e que o lixo industrial acarreta diversos prejuízos ambientais e sociais, e, nesse contexto, a Lei de Benford se mostra uma boa ferramenta na busca da sustentabilidade, uma vez que sinaliza possíveis erros de processo, bem como fraudes em dados divulgados.

Para os dados referentes à poluição, o primeiro dígito é afetado pelas características inerentes à variável estudada, estando vinculado às quantidades usuais das partículas na região, enquanto, para o segundo dígito, a aleatoriedade foi comprovada.

A análise dos dígitos se mostra como uma ferramenta poderosa no aprimoramento de processos industriais e no controle de poluentes emitidos pelas indústrias, sendo de baixo custo e fácil implementação com o uso dos aplicativos MS Excel ou R.

ANALYSIS OF INDUSTRIAL DIGITS BASED ON LAW OF BENFORD AND ITS APPLICATION USING COMPUTER ROUTINES

Abstract

This paper addresses the application of a statistical tool used for checking the reliability of sample values based on Benford distribution. This distribution is composed of the occurrence probabilities of each integer, from 1 to 9 for the first significant digit of the sample values, forming a decreasing logarithmic curve. This study focuses on data related to the industrial sector, in conformity with the current global trend of sustainable production. The paper also presents two computer routines developed for implementing the Analysis of Digits in the software: Excel and R.

Keywords: Benford's law, reliability, statistical process control.

REFERÊNCIAS

ALMIRON, M. G. et al. On the numerical accuracy of spreadsheets. *Journal of Statistical Software*, v. 34, n. 4, 2010. Disponível em: <http://www.jstatsoft.org/v34/i04/paper>. Acesso em: 10 mar. 2012.

COMPANHIA DE TECNOLOGIA DE SANEAMENTO AMBIENTAL. *Qualidade do ar*. São Paulo: Cetesb, 2012. Disponível em: <http://www.cetesb.sp.gov.br/ar/qualidade-do-ar/32-qualar>. Acesso em: 26 mar. 2012.

HALES, N. D. et al. Testing the accuracy of employee-reported data: an inexpensive alternative approach to traditional methods. *European Journal of Operational Research*, v. 189, p. 583--593, 2008.

HILL, T. P. A statistical derivation of the significant-digit law. *Statistic Science*, v. 10, p. 354-363, 1996. Disponível em: <http://www.math.gatech.edu/~hill/publications/cv.dir/stat-der.pdf>. Acesso em: 17 abr. 2008.

_____. The first digit phenomenon. *The American Scientist*, v. 86, n. 4, p. 358-376, 1998. Disponível em: <http://www.math.gatech.edu/~hill/publications/cv.dir/1st-dig.pdf>. Acesso em: 17 abr. 2008.

_____. The difficulty of faking data. *Chance*, v. 26, p. 8-13, 1999. Disponível em: <http://www.math.gatech.edu/~hill/publications/cv.dir/faking.pdf>. Acesso em: 17 abr. 2008.

HORNIK, K. *The R FAQ*, 2011. Disponível em: <http://CRAN.R-project.org/doc/FAQ/R-FAQ.html>. Acesso em: 16 set. 2011.

HÜRLIMANN, W. *Benford's law from 1881 to 2006*: a bibliography. Zürich, 2006. Disponível em: <http://arxiv.org/ftp/math/papers/0607/0607168.pdf>. Acesso em: 3 mar. 2008.

MAGALHÃES, M. N.; LIMA, A. C. P. *Noções de probabilidade e estatística*. 7. ed. São Paulo: Edusp, 2010.

MIGLIOLI, A. M.; OSTANEL, L. H.; TACHIBANA, W. Planilhas eletrônicas como ferramentas para apoio à decisão e geração de conhecimento na pequena empresa. In: ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO, 24., 2004, Florianópolis. *Anais...* Disponível em: <http://www.abepro.org.br/biblioteca/ENEGEP2004_Enegep0902_1706.pdf>. Acesso em: 4 out. 2010.

NEWCOMB, S. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, v. 4, n. 1, p. 39-40, 1881. Disponível em: <http://www.uvm.edu/~pdodds/files/papers/others/1881/newcomb1881a.pdf>. Acesso em: 26 mar. 2012.

NIGRINI, M. A taxpayer compliance application of Benford's law. *Journal of the American Taxation Association*, v. 1, p. 72-91, 1996.

R DEVELOPMENT CORE TEAM. *R*: *a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, 2011. Disponível em: <http://www.R-project.org/>. Acesso em: 13 dez. 2011.

_____. *An introduction to R*. Notes on R: a programming environment for data analysis and graphics. Version 2.13.1. Disponível em: <http://www.r-project.org/>. Acesso em: 1º fev. 2012.

SIEGEL, S.; CASTELLAN JR., N. J. *Estatística não-paramétrica para ciências do comportamento*. Métodos de pesquisa. 2. ed. Porto Alegre: Bookman, 2008.

VANCE, A. Data analysts captivated by R's power. *The New York Times*, 2009. Disponível em: <http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html?pagewanted=all>. Acesso em: 25 ago. 2011.

Contato

Raquel Cymrot
e-mail: raquel.cymrot@mackenzie.br

APÊNDICE A

Análise dos dígitos

```
#Separando os dígitos que formam os valores amostrados
Dados <- Digitos(Dados)

# Análise do primeiro dígito
Digito.1 <- Dados$d1
Freq1_Real <- quant(Digito.1, zero = F) # zero = F ou zero = FALSE #desconsidera os
valores nulos de Digito.1
Prob1_Benford <- pBenford(1:9, 1)
test1 <- chisq.test(Freq1_Real, p = Prob1_Benford)
test1

# Análise do segundo dígito
Digito.2 <- Dados$d2
Freq2_Real <- quant(Digito.2, zero = T)
Prob2_Benford <- pBenford(0:9, 2)
test2 <- chisq.test(Freq2_Real, p = Prob2_Benford)
test2

# Análise do terceiro dígito
Digito.3 <- Dados$d3
Freq3_Real <- quant(Digito.3, zero = T)
Prob3_Benford <- pBenford(0:9, 3)
test3 <- chisq.test(Freq3_Real, p = Prob3_Benford)
test3

# Análise do quarto dígito
Digito.4 <- Dados$d4
Freq4_Real <- quant(Digito.4, zero = T)
Prob4_Benford <- pBenford(0:9, 4)
test4 <- chisq.test(Freq4_Real, p = Prob4_Benford)
test4
```

APÊNDICE B

Para importação dos dados do Excel para o R

```
# O R entende este texto como comentário.

# No caso de seus dados serem importados do Excel,
# quando se importa um conjunto de dados do Excel para o R com a função
#read.csv, os valores são alocados em um data frame.
Dados <- read.csv("C:\\My Documents\\Benfords Law\\Dados.csv", sep = ";")
Dados # Observem como os dados estão dispostos.
is.data.frame(Dados) # A função retorna verdadeiro ou falso se o objeto
#é ou não um data frame. A resposta nesse caso é TRUE.

# Modificando o título das colunas dos dados
# Esse ajuste dos rótulos poderá ser necessário se o conteúdo do rótulo
#não estiver sendo reconhecido adequadamente pelo R.
# O comando abaixo modifica o nome atribuído à coluna de dados importada.
names(Dados) <- c("Dados").
Dados # Verificando novos nomes
Dados <- Dados$Dados # A partir de agora, a variável
#dados armazena somente os valores em um tipo vetor.

# Testando se o objeto é um vetor.
is.vector(Dados) # Resposta igual a TRUE.
```

APÊNDICE C

Para cópia dos dados diretamente no R

```
# No caso de seus dados serem inseridos diretamente no R,
# basta colar no R console os valores da amostra
#(sem o rótulo) após a execução dessa função.
Dados <- scan()
```

APÊNDICE D

Função auxiliar para separação dos valores dos dados em quatro dígitos

Utilize um arquivo *script* e copie o código abaixo.

```
# Exemplo de teste da função:
# x <- c(seq(0.11, 0.99, 0.1), rnorm(1e3)*1e-3, runif(1e3)*1e8)
# x1 <- Digitos(x)

Digitos <- function(x)
{
        y <- abs(x)

        for(i in 1:length(y))
        {
                while(y[i] < 999) y[i] <- y[i]*1e2
                # Enquanto os valores de 'y' forem menores
                # que 1000, será efetuada a multiplicação por dez.

                while(y[i] > 10000) y[i] <- y[i]/1e1
                # Enquanto os valores de 'y' forem maiores
                # que 9999, será efetuada a divisão por dez.
        }

        y <- trunc(y)

        d1 <- y %/% 1000

        d2 <- (y %% 1000) %/% 100

        d3 <- (y %% 100) %/% 10

        d4 <- y %% 10
```

```
        # Os vetores 'd1', 'd2', 'd3' e 'd4' são respectivamente os
    #valores dos dígitos que compõem o valor amostrado. Ex: 1234
    #onde d1 = 1, d2 = 2, d3 = 3, d4 = 4

    ans <- data.frame(x,y,d1,d2,d3,d4)
    # Os valores dos dados separados em dígitos são armazenados em
    # objeto do tipo 'data frame '.

    ans
}
```

APÊNDICE E

Função auxiliar das probabilidades dos dígitos escolhidos

Utilize um arquivo *script* e copie o código abaixo.

```
# Exemplo de teste da função pBenford(c(0:10, NA, NaN, Inf, -Inf), 1)
pBenford <- function(p, d)
{
    # d: Para qual dígito se quer a probabilidade?
    # 'd' igual a 1, 2, 3 ou 4.
    # p: Vetor de números para os quais se quer a probabilidade.

    Benford <- 0
    p <- p[!is.na(p) & !is.nan(p) & p<10 & p>=0]
    if(length(p) > 1)
    {
        if(d == 1)
        {
            for (j in 1:length(p))
            {
                if(p[j]!= 0)
                    Benford[j] <- log10(1 + p[j]^(-1))
                else
                    Benford[j] <- NA
            }
```

```
        }
        else
        {
            alfa <- 10^(d-1)
            m <- 1
            omega <- 0
            while((d-m) >= 1)
            {
                omega[m] <- 10^(d-m)
                m = m + 1
            }
            omega <- 9*sum(omega)
            Benf <- 0
            for (j in 1:length(p))
            {
                for (i in seq(alfa, omega, 10))
                    Benf <- log10(1 + (i + p[j])^(-1)) + Benf
                Benford[j] <- Benf
                Benf <- 0
            }
        }
    }
    else
        Benford <- NA
    ans <- Benford
    ans
}
```

APÊNDICE F

Função auxiliar para frequências dos dígitos significativos

Abra um novo arquivo *script* e copie o código abaixo.

```
# Exemplo de teste da função quant(c(1:10, NA, NaN, Inf, -Inf), F).
# Essa função retorna à frequência dos dígitos.
```

```
# Isto é, quantos dos dígitos são respectivamente iguais a 0, 1, 2,
# 3, 4, 5, 6, 7 e 9.
quant <- function(a, zero = FALSE)
{
# a: vetor de dados.
# zero: define se vamos ou não quantificar o número de zeros que compõem
# o vetor; se isso não for especificado, o vetor desconsiderará o número de zeros.
        q <- function(x, y)
        {
                # Desenvolveu-se uma função dentro de outra
                #função para facilitar os cálculos; a função
                #'q' incrementa a variável 'k' quando a
                #variável 'x' é igual à variável 'y'.
                k <- 0
                for (i in 1:length(x))
                {
                        if (x[i] == y)
                                k = k + 1
                }
                ans <- k
        }

        d <- 0
        zero <- zero

        a <- a[!is.na(a) & !is.nan(a)] #Retira valores perdidos
        #ou que não são contabilizados como 0/0.
        if(length(a) != 0)
        {
                if(zero == TRUE)#Considera os dígitos iguais a zero.
                #Essa parte da função contabiliza a frequência dos
                #dígitos e aloca o resultado no vetor 'd'.
                        for (j in 1:10)
                                d[j] <- q(a, j-1) #Chamada da função 'q'
                else #Quando NÃO se contam dígitos iguais a zero.
                        for (j in 1:9)
                                d[j] <- q(a, j) #O valor da função é
                                #armazenado no vetor 'd'.
                ans <- d
```

```
                    ans
                    #O vetor 'd' deterá a quantidade de cada número de zero
                    #a 9 presente no vetor 'a' se a variável lógica
                    #'zero' for igual a TRUE ou T.
                }
}
```