II-011 - MEDIDA DE TENDÊNCIA CENTRAL MAIS ADEQUADA PARA DADOS DE MONITORAMENTO DE ESTAÇÕES DE TRATAMENTO DE ESGOTOS

Sílvia Corrêa Oliveira[(1)]
Doutora em Saneamento, Meio Ambiente e Recursos Hídricos da UFMG. Professora Adjunta do Departamento de Engenharia Sanitária e Ambiental da UFMG.
Marcos von Sperling[(1)]
Doutor em Engenharia Ambiental (Imperial College, Universidade de Londres – Inglaterra). Professor Titular do Departamento de Engenharia Sanitária e Ambiental da UFMG.

Endereço[(1)]**:** Departamento de Engenharia Sanitária e Ambiental – UFMG; Av. do Contorno, nº 842 – 7º andar – Centro – Belo Horizonte – MG – Brasil – CEP 30.110-060– Tel: (31) 3409-1935 – e-mail: silvia@desa.ufmg.br; marcos@desa.ufmg.br

RESUMO

No presente trabalho é efetuado um estudo para avaliar se a média aritmética é a medida de tendência central mais apropriada para dados de monitoramento de estações de tratamento de esgotos, utilizando dados típicos de concentrações efluentes de DBO de uma ETE operando em escala real. O estudo demonstra que para dados assimétricos e que apresentam um bom ajuste à distribuição lognormal, como é o caso de dados oriundos de ETEs, a média geométrica pode ser considerada mais apropriada. Mostra, ainda, que para assimetrias mais acentuadas (coeficientes de variação iguais ou maiores que 0,5), a mediana pode ser considerada uma boa medida alternativa de tendência central para tais dados, uma vez que apresenta sempre valores próximos aos obtidos para média geométrica. Os resultados obtidos servem de alerta, também, para a utilização indiscriminada de técnicas estatísticas clássicas para estimar parâmetros e testar hipóteses acerca de médias, uma vez que tais técnicas foram originalmente desenvolvidas para dados oriundos de populações normais. Quando a suposição de normalidade não é confirmada, os níveis de confiança podem ser distorcidos, os métodos estatísticos perdem potência e os resultados obtidos podem ser severamente imprecisos. Desta forma, quando usuários desrespeitam os pressupostos requeridos para utilização de testes paramétricos, os resultados poderão ser incorretos e não confiáveis.

PALAVRAS-CHAVE: Medida de tendência central, distribuição lognormal, efluentes, estações de tratamento de esgotos.

INTRODUÇÃO

As medidas de tendência central são utilizadas numa tentativa de descrever comportamentos "típicos" ou "médios" de um conjunto de dados aleatórios, sendo que as mais largamente utilizadas são a média aritmética ($\bar{X}$), a mediana (o valor do meio ou a média de dois valores do meio) e a moda (o valor mais frequentemente observado). Usualmente, a média aritmética tem sido utilizada, indiscriminadamente, para quantificar valores médios de dados de qualquer origem, porque, além de ser fácil de calcular, tem uma interpretação familiar e propriedades estatísticas que a tornam muito útil nas comparações entre populações e situações que envolvam inferências.

No entanto, para dados assimétricos, que não podem ser representados pela distribuição Normal, como é o caso de dados de concentrações afluentes e efluentes de estações de tratamento de esgotos (ETEs) e da maioria dos dados ambientais, esta medida de tendência central pode não ser a mais adequada. Estudos efetuados por diversos autores sobre a caracterização da distribuição de probabilidade dos dados efluentes de ETEs têm mostrado que a distribuição lognormal é mais representativa do comportamento da maioria dos constituintes (Niku *et al.* 1981, Berthouex e Hunter, 1981, Metcalf & Eddy, 2003, Oliveira, 2006). Para dados representados por distribuições assimétricas à direita, incluindo a lognormal, a média amostral $\bar{X}$ tem se mostrado inadequada devido ao impacto causado pela presença de valores elevados e atípicos ("outliers") que usualmente ocorrem. Estes altos valores, eventualmente presentes em pequenas quantidades, causam uma elevação na média aritmética, trazendo dúvidas se ela seria realmente a melhor medida de tendência central, já que a maior parte dos valores é menor que a média. Para tais observações assimétricas, a média geométrica

tem sido considerada por alguns pesquisadores uma substituta particularmente apropriada (McAlister, 1879, Niku *et al.*, 1981, Limpert *et al.*, 2001, Metcalf & Eddy, 2003).

O objetivo principal do presente trabalho é efetuar um estudo para avaliar se a média aritmética é a medida de tendência central mais apropriada para dados de monitoramento de estações de tratamento de esgotos, utilizando dados típicos de concentrações efluentes de DBO de uma ETE operando em escala real. São ilustradas, ainda, algumas propriedades das distribuições normal e lognormal, com o objetivo de alertar para as distorções que podem ser obtidas quando se aplicam propriedades da distribuição normal em dados assimétricos.

METODOLOGIA

A distribuição normal é a mais importante das distribuições contínuas, apresentando a forma de um sino e sendo utilizada para descrever o comportamento de variáveis aleatórias que flutuam de forma simétrica em torno de um valor central. Algumas das propriedades matemáticas do modelo normal fazem com que esta distribuição esteja na origem de toda a formulação teórica acerca da construção de intervalos de confiança, testes estatísticos de hipóteses, bem como da teoria de regressão e correlação (NAGHETTINI & PINTO, 2007). A partir da definição de apenas dois parâmetros, a média e o desvio padrão, é possível calcular a probabilidade de ocorrência de valores de interesse, como por exemplo, o percentual de valores que deverão estar acima ou abaixo de um determinado valor, ou entre dois valores definidos. A determinação destas probabilidades é realizada matematicamente através da integração da função de densidade de probabilidade (Equação 1) entre os pontos de interesse. No caso da distribuição normal, a integral não pode ser calculada exatamente, e a probabilidade entre dois pontos só pode ser obtida aproximadamente, por métodos numéricos. Esta tarefa é facilitada através do uso da distribuição normal padrão, que tem média $\mu = 0$ e desvio padrão $\sigma = 1$, largamente discutida em livros texto de estatística (SNEDECOR & COCHRAN, 1989, LEVINE *et al.*, 2000, BUSSAB & MORETTIN, 2002, NAGHETTINI & PINTO, 2007).

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} \qquad \text{para } -\infty < X < \infty \tag{1}$$

No caso da distribuição normal padrão, algumas dessas áreas calculadas são bastante difundidas, e estão representadas na Figura 1.

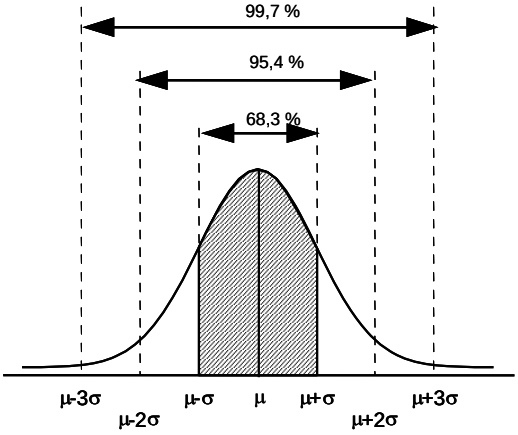

Figura 1: Probabilidades da distribuição normal

O exame da Figura 1 demonstra que 68,3% dos valores de *x* (área hachurada do gráfico) estão compreendidos entre os limites de 1 desvio padrão abaixo e acima da média (μ-σ, μ+σ). Do mesmo modo conclui-se que 95,4% da área corresponde ao intervalo (μ-2σ, μ+2σ), enquanto 99,7% está compreendida pela área situada entre os limites de μ-3σ e μ+3σ.

Uma relação correspondente pode ser observada para dados que seguem uma distribuição lognormal, considerando a média geométrica (μg) e desvio-padrão geométrico (σg). Neste caso, o intervalo $(\mu_g \times \sigma_g, \mu_g \div \sigma_g = \mu_g \overset{\times}{\div} \sigma_g)$ contém 68,3% dos dados e $(\mu_g \times \sigma_g^2, \mu_g \div \sigma_g^2 = \mu_g \overset{\times}{\div} \sigma_g^2)$ representa a faixa que compreende 95,4% dos valores. A Tabela 1 sumariza e compara algumas propriedades das distribuições normal e lognormal. As médias e os desvios padrão aritméticos e geométricos amostrais são representados por $\bar{x}$ e S e $\bar{x}g$ e Sg, respectivamente.

A média e o desvio padrão geométricos são representados pelas expressões:

$$\overline{X}g = \left(\prod_{i=1}^{n} X_i \right)^{1/n} \qquad (2)$$

$$Sg = \exp\left(\frac{\sum_{i=1}^{n} \left(\ln X_i - \ln \overline{X}g \right)^2}{n} \right)^{1/2} \qquad (3)$$

Onde o símbolo Π representa o produto de todos os valores amostrais.

Tabela 1: Relações entre as distribuições normal e lognormal

Propriedades	Distribuição	
	Normal	Lognormal
Média	$\bar{x}$, aritmética	$\bar{x}g$, geométrica
Desvio padrão	*S*, aditivo	*S*g, multiplicativo
Intervalos de confiança:		
68,3 %	$\bar{x} \pm S$	$\bar{x}g \overset{\times}{\div} Sg$
95,4 %	$\bar{x} \pm 2S$	$\bar{x}g \overset{\times}{\div} Sg^2$
99,7 %	$\bar{x} \pm 3S$	$\bar{x}g \overset{\times}{\div} Sg^3$

Nota: O símbolo $\overset{\times}{\div}$ = limite superior: $\bar{x}g$ x Sg;.limite inferior:. $\bar{x}g \div Sg$

Fonte: Adaptado de Limpert *et al*., 2001

Estas relações foram avaliadas a partir de um exemplo de aplicação com dados típicos de concentrações efluentes de DBO, obtidos em estações de tratamento de esgotos operando em escala real.

RESULTADOS E DISCUSSÃO

A Tabela 2 apresenta estatísticas descritivas referente à concentração de DBO efluente de uma ETE operando em escala real na região metropolitana de Belo Horizonte/MG. A Figura 1 mostra o histograma típico para o mesmo parâmetro, que evidencia o comportamento assimétrico à direita e o bom ajuste à função densidade de probabilidade (FDP) de uma distribuição lognormal (em vermelho no gráfico).

Tabela 2: Estatística descritiva referente à concentração de DBO efluente

Média	Desvio padrão	CV	Média Geométrica	D. Padrão Geométrico	Mediana	Valor Mín.	Valor Máx.
25	13	0,5	22	2	23	5	117

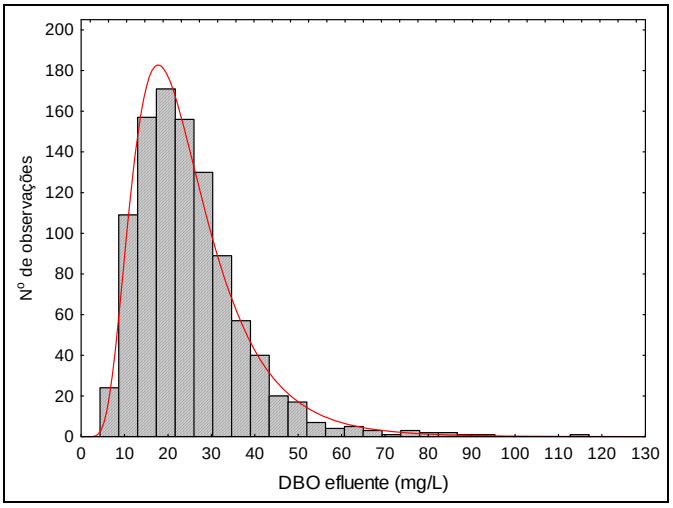

Figura 1: Histograma típico de concentrações efluentes de DBO

As propriedades das distribuições normal e lognormal, apresentadas na Tabela 1, foram testadas com os dados desta mesma ETE. Assim, foram calculados os percentuais de observações contidas nos intervalos característicos das duas distribuições, ou seja, o percentual de dados dentro da faixa $\overline{X} \pm S$ (média aritmética ± um desvio padrão), $\overline{X} \pm 2S$ e $\overline{X} \pm 3S$ e, ainda, o percentual contido em $\overline{X}g \,^{\times}\!\div\, Sg$ (média geométrica $^{\times}\!\div$ desvio padrão geométrico), $\overline{X}g \,^{\times}\!\div\, (Sg)^2$ e $\overline{X}g \,^{\times}\!\div\, (Sg)^3$. Os resultados obtidos são apresentados na Tabela 3.

Tabela 3: Percentual de observações contidas nos intervalos característicos das distribuições normal e lognormal, para as concentrações de DBO efluente de uma ETE na RMBH

Característica	Distribuição Normal			Distribuição Lognormal		
	$\bar{x} \pm S$	$\bar{x} \pm 2S$	$\bar{x} \pm 3S$	$\bar{x}g \,^{\times}\!\div\, Sg$	$\bar{x}g \,^{\times}\!\div\, Sg^2$	$\bar{x}g \,^{\times}\!\div\, Sg^3$
Valor teórico	68,3%	95,4%	99,7%	68,3%	95,4%	99,7%
Valor obtido	76,1%	96,4%	98,4%	68,2%	95,4%	99,6%

Fica evidente, pela observação da Tabela 3, a aplicabilidade das relações apresentadas na Tabela 1 para a distribuição lognormal, já que os percentuais esperados de dados contidos nos intervalos são muito próximos dos valores efetivamente observados. É importante ressaltar, também, as distorções obtidas quando se aplicam propriedades da distribuição normal em dados assimétricos, uma vez que os percentuais esperados de dados contidos nas faixas relativas a distribuições simétricas foram muito diferentes daqueles calculados.

Outra consideração importante a ser efetuada, em termos práticos, é a grande influência do coeficiente de variação (CV), ou seja, o desvio padrão dividido pela média aritmética, na forma da distribuição. Ott (1995) comenta que, para valores de CV menores que um sexto (0,16667), a função densidade de probabilidade da distribuição lognormal apresenta um comportamento muito próximo da distribuição normal. Por outro lado, elevados valores de CV são associados com expressivas assimetrias.

Valores de CVs típicos de estações de tratamento de esgotos, calculados por Oliveira (2006) em um estudo sobre 166 ETEs dos estados de Minas Gerais e São Paulo, compreendendo seis diferentes processos de tratamento de esgotos, foram utilizados para ilustrar tal comportamento. Neste estudo, foi observado que cerca de 60% dos dados de concentração efluente de DBO apresentaram coeficientes de variação acima de 0,5. Considerando as concentrações efluentes de coliformes termotolerantes, mais de 90% dos valores de CV foram maiores que 0,5, o que justifica a adoção usual da média geométrica como medida de tendência central deste parâmetro.

A função densidade de probabilidade da distribuição lognormal foi traçada para diferentes situações, utilizando a equação (1). Séries aleatórias de 1000 dados foram geradas na distribuição lognormal, considerando a mesma média aritmética apresentada pelos dados da ETE em estudo. Assim, no gráfico apresentado na Figura 2, para este mesmo valor de média aritmética, quatro valores diferentes de desvios padrão foram utilizados, gerando os seguintes coeficientes de variação: CV = 0,1, CV = 0,5, CV = 1,0 e CV = 2,0. Na Tabela 4 são apresentados os valores calculados das médias aritmética e geométrica e da mediana destes mesmos dados.

Como observado na Figura 2 e na Tabela 4, para dados que apresentam pequenos coeficientes de variação, a média aritmética poderá ser utilizada como uma medida adequada de tendência central e todas as suas propriedades serão aplicáveis. Mas, em presença de grandes assimetrias, a média geométrica será mais indicada e a utilização de técnicas estatísticas clássicas para estimar parâmetros e testar hipóteses acerca de médias não será adequada. Outra observação evidenciada nesta simulação é que a mediana pode ser considerada uma boa medida alternativa de tendência central para dados assimétricos, já que os valores calculados são geralmente muito próximos daqueles obtidos para a média geométrica.

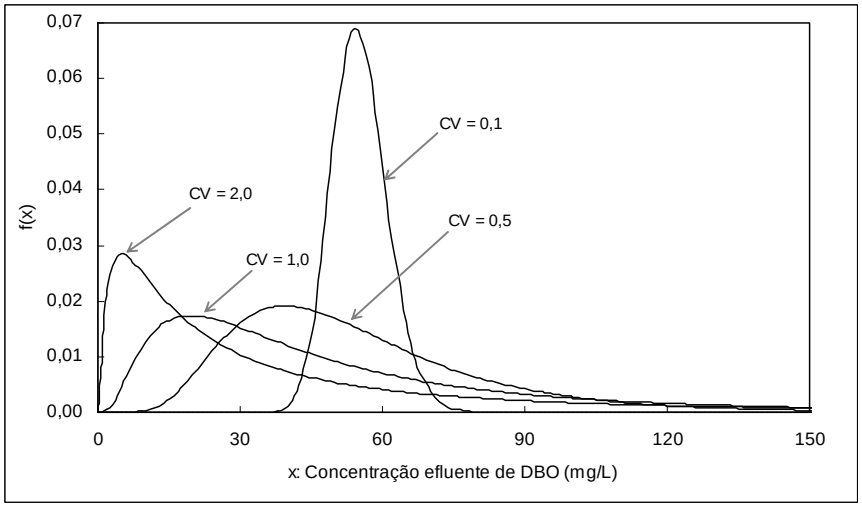

Figura 2: PDF para a distribuição lognormal, mostrando diferenças nas formas apresentadas para uma média aritmética constante e diferentes valores de CV

Tabela 4: Medidas de tendência central para dados que apresentam diferentes valores de CV

Parâmetro	Coeficientes de variação (CV)			
	0,1	0,5	1,0	2,0
Média aritmética	25	25	25	25
Média geométrica	25	22	18	11
Mediana	25	23	18	12

CONCLUSÕES

Dados oriundos de estações de tratamento de esgotos são frequentemente assimétricos e apresentam um bom ajuste à distribuição lognormal. Isto significa que a medida de tendência central mais apropriada não é a média aritmética, representativa de distribuições normais, mas a média geométrica. A mediana pode ser considerada uma boa medida alternativa de tendência central para tais dados, uma vez que apresenta geralmente valores similares aos obtidos para média geométrica. O conhecimento das propriedades da curva lognormal permite que, sendo calculados a média e o desvio padrão geométricos, se conheçam quais os valores mais freqüentes e quais os valores extremos esperados.

Deve-se notar que estes comentários dizem respeito à busca da melhor representação de uma medida de tendência central. No caso da utilização da média e do desvio padrão geométrico, o relatório técnico ou

trabalho científico deve deixar bem clara esta opção, uma vez que ela não tem sido a prática na grande maioria dos trabalhos redigidos, e muitas vezes os leitores têm maior dificuldade na sua compreensão. Neste sentido, a mediana é de compreensão mais fácil, e pode ser utilizada em várias aplicações. A média aritmética permanece importante quando se deseja dar importância aos valores mais elevados obtidos no monitoramento, que podem ter implicações importantes, por exemplo, em termos da avaliação do impacto do lançamento em um corpo receptor, ou no cálculo de cargas poluentes (carga = vazão x concentração).

Outra importante consideração a ser efetuada diz respeito aos testes estatísticos de comparações de médias, utilizados indiscriminadamente em conjuntos de dados de qualquer espécie, mas que foram originalmente desenvolvidos para dados oriundos de populações normais. Quando a suposição de normalidade não é confirmada, os níveis de confiança podem ser distorcidos, os métodos estatísticos perdem potência e os resultados obtidos podem ser severamente imprecisos. Desta forma, quando usuários desrespeitam os pressupostos requeridos para utilização de testes paramétricos, os resultados poderão ser incorretos e não confiáveis. Sendo assim, é muito importante o conhecimento da distribuição de probabilidade de dados de concentrações efluentes, para alimentação de modelos matemáticos de processos, seleção de métodos estatísticos adequados para avaliação de desempenho e interpretação de medidas de tendência central, com graus elevados de confiança, considerando sua variabilidade e confiabilidade.

REFERÊNCIAS BIBLIOGRÁFICAS

1. BERTHOUEX, P.M., HUNTER, W.G. Simple statistics for interpreting environmental data. *Journal of Water Pollution Control Federation*, v. 53, n. 2, p. 167-175, 1981.
2. BUSSAB, W. O.; MORETTIN, P. A. Estatística básica. São Paulo: Saraiva, 5 ª ed, 2002, 526 p.
3. LEVINE, D. M.; BERENSON, M. L.; STEPHAN, D. *Estatística: Teoria e aplicações: usando o Microsoft Excel em português*. Rio de Janeiro: LTC, 2000, 812 p.
4. LIMPERT, E.; STAHEL, W. A.; ABBT, M. Log-normal Distributions across the Sciences: Keys and Clues. *Bioscience*, v. 51, n. 5, pages 341–352, 2001.
5. McALISTER, D. The Law of the Geometric Mean. *Proceedings of the Royal Society of London*, v. 29, pp. 367-376, 1879.
6. METCALF & EDDY. *Wastewater engineering: treatment, and reuse*. New York: Metcalf & Eddy, Inc., 4 th. Ed., 2003, 1819p.
7. NAGHETTINI, M.; PINTO, E. J. A. *Hidrologia Estatística*, Boletim Técnico CPRM, 1ª ed., 2007, 552p.
8. NIKU, S., SCHROEDER, E.D., TCHOBANOGLOUS, G., SAMANIEGO F.J. Performance of activated sludge process: reliability, stability and variability. Environmental Protection Agency, EPA Grant Nº R805097-01, pp. 1 – 124, 1981b.
9. OLIVEIRA, S. M. A. C. *Análise de desempenho e confiabilidade de estações de tratamento de esgotos.* Tese (Doutorado em Saneamento, Meio Ambiente e Recursos Hídricos) - Escola de Engenharia, Universidade Federal de Minas Gerais, Belo Horizonte, 231 f. 2006.
10. SNEDECOR, G.W., COCHRAN, W.G. *Statistical Methods*. Ames: Iowa State University Press, 8th. ed., 1989. 503p.