Avaliação de critérios estatísticos utilizados em programas interlaboratoriais para pesquisa de chumbo em sangue

Evaluation of statistical criteria applicable to interlaboratory comparisons for lead in blood

RIALA6/1070

Franca Durante de MAIO[2], Odair ZENEBON[1*], Paulo TIGLEA[1], Roberta I. S. OKURA[2], Alice M. SAKUMA[2]

* Endereço para correspondência: [1]Divisão de Bromatologia e Química, Instituto Adolfo Lutz, Av. Dr Arnaldo 355, CEP 01246 902, São Paulo, Brasil.

[2] Seção de Equipamentos Especializados, Instituto Adolfo Lutz, São Paulo, SP.

Recebido: 16/01/2006 – Aceito para publicação: 18/07/2006

RESUMO

A participação em Programas de comparação Interlaboratorial (PI) é uma ferramenta utilizada pelos laboratórios para garantir a confiabilidade de seus resultados analíticos. Em PIs a escolha do tratamento estatístico aplicado aos resultados pode afetar a estimativa do valor verdadeiro (valor designado) e a incerteza associada, o intervalo de aceitação e a avaliação de desempenho dos laboratórios. Empregando resultados dos laboratórios participantes do Programa de Ensaio de Proficiência para chumbo em sangue (PEP-Pb-s), coordenado pelo Instituto Adolfo Lutz, São Paulo, para diferentes níveis de concentração, foram comparados sete critérios estatísticos. O valor designado foi calculado usando a mediana, ou a média após exclusão dos dispersos, empregando os critérios de Hampel, Grubbs, Dixon e Z-score. O intervalo de aceitação foi calculado pelo Z-score empregando o desvio padrão ou utilizando as equações de Horwitz ou de Horwitz modificada por Thompson. Foi também utilizado o critério atualmente empregado no Programa, que considera $X_{designado} \pm 6$, para concentrações [Pb-s] ≤40 μg/100mL e $X_{designado} \pm 15\% X_{designado}$ para concentrações > 40 μg/100mL. Verificou-se que os valores designados obtidos pelos critérios estatísticos empregados não apresentaram diferenças significantes em um nível de significância de 5%. Porém, os intervalos de aceitação obtidos mostraram-se influenciados pelos critérios estatísticos empregados.

Palavras-Chave. programa interlaboratorial, ensaio de proficiência, técnicas estatísticas para ensaio de proficiência, chumbo em sangue.

ABSTRACT

The participation in Interlaboratory Comparison Programme (IP) is a tool used by laboratories to ensure reliability of analytical results. In IPs the selection of a statistical treatment applied to the results can affect the estimation of the "true" value (assigned value) and its uncertainty, the acceptance range and the performance assessment of the participants laboratories. Using the results of the participant laboratories in the Proficiency Test Programme for lead in blood, coordinated by Instituto Adolfo Lutz, São Paulo, for different levels of concentration, seven statistical techniques were compared. The assigned value was calculated using the median, or the average value after the elimination of the dispersed results, employing Hampel, Grubbs, Dixon and Z-score's criteria. The acceptance range was calculated according to Z-score, making use of the standard deviation or using Horwitz's equation or the Horwitz's one modified by Thompson. It was also used the current criterion applied to the Program, which considers $X_{designated} \pm 6$, for concentrations [Pb-s] ≤40 μg/100mL and $X_{designated} \pm 15\% X_{designated}$ for concentrations > 40 μg/100mL. It was verified that the assigned values obtained through the statistical techniques used did not present significant differences when the confidence level was 5%. However, the acceptance ranges obtained showed they were influenced by the statistical technique used.

Key Words. interlaboratorial programme, proficiency test, statistical criteria for proficiency test, lead in blood.

INTRODUÇÃO

A avaliação externa da qualidade é uma ferramenta utilizada pelos laboratórios a fim de garantir a confiabilidade de seus resultados analíticos e demonstrar competência técnica. A técnica estatística aplicada a um PI para a determinação do valor designado (a melhor estimativa do valor verdadeiro de uma concentração) e sua incerteza associada, e do intervalo de aceitação dos resultados pode afetar a avaliação do desempenho dos laboratórios participantes. Um Programa Interlaboratorial consiste de uma série de medições de uma ou mais propriedades, em uma amostra de determinado material, realizada de forma independente por um grupo de laboratórios, e pode visar diferentes objetivos, como compatibilização de resultados, validação de métodos de análise, certificação de materiais de referência e avaliação de desempenho de laboratórios[1] (programas de ensaios de proficiência). A participação em PIs é um dos requisitos da norma ISO/IEC 17025[2] para acreditação ou habilitação de um determinado ensaio junto aos órgãos reguladores nacionais e internacionais, na área de metrologia.

A escolha do tratamento estatístico para um PI depende, em primeiro lugar, dos objetivos deste e de suas características, como o grau de dispersão dos resultados apresentados (que é função da dificuldade analítica do ensaio), da faixa de concentração do analito em estudo, entre outros fatores inerentes à realização do ensaio. Deve ainda considerar as conseqüências de uma decisão incorreta na avaliação dos laboratórios.

Para PIs, são descritas na literatura diferentes técnicas para o cálculo do valor designado e das incertezas a ele associadas, inclusive critérios para exclusão de resultados que, dentro de uma margem definida de probabilidade, possam afetar estes valores[1,3,4,5]. Para a determinação do valor designado, o valor de consenso dos laboratórios participantes (média - após exclusão de valores dispersos –ou mediana dos resultados) é um dos critérios mais utilizados.

Outro critério estatístico que pode ser empregado na avaliação dos resultados de um PI é a estatística robusta, válida inclusive para resultados de ensaios que não apresentam comportamento segundo uma distribuição normal[6]. A vantagem da utilização da estatística robusta é que resultados dispersos não exercem grande influência na estimativa da posição central e da dispersão dos resultados não sendo, portanto, requerida aplicação de testes para verificação e exclusão de resultados dispersos ("outliers"). O valor designado é dado pela mediana (md) dos resultados. O valor de dispersão é calculado pela amplitude interquartílica normalizada[7].

De acordo com a Organização Mundial da Saúde, o principal objetivo da avaliação externa da qualidade é estabelecer a comparabilidade entre resultados de laboratórios, particularmente em investigações epidemiológicas, monitorização ambiental e biológica, pesquisas e outras atividades em Saúde Pública[8].

A concentração de chumbo sangüíneo é o indicador biológico mais utilizado para avaliar a exposição recente de um indivíduo a esse metal. Como o chumbo está presente no sangue em baixas concentrações, esta determinação apresenta dificuldades analíticas e, dada a sua complexidade, podem ocorrer erros aleatórios mesmo quando o ensaio é realizado por técnico treinado, com equipamento calibrado e com metodologia validada.

Neste trabalho, serão aplicados diferentes critérios estatísticos para avaliar os resultados de um programa interlaboratorial para chumbo em sangue, em três níveis de concentração, objetivando comparar os diversos tratamentos estatísticos para obtenção dos valores designados e dos intervalos de aceitação, avaliar o impacto à saúde de população exposta ao chumbo decorrente da adoção de diferentes critérios de aceitação dos resultados e avaliar os métodos mais adequados de tratamento estatístico para o PEP-Pb-s, na presença de resultados dispersos.

MATERIAL E MÉTODO

Foram selecionadas três amostras do Programa de Ensaio de Proficiência para Chumbo em Sangue coordenado pelo Instituto Adolfo Lutz, com participação de cerca de 30 laboratórios, com concentrações de chumbo sangüíneo variando de cerca de 16 a 70 μg/100mL, abrangendo níveis geralmente encontrados em indivíduos com exposição ambiental ou ocupacional.

Num PI, a maior dificuldade no tratamento estatístico dos resultados é a determinação do valor verdadeiro do item de ensaio que, no caso de ensaios quantitativos, refere-se à determinação da concentração verdadeira do analito na amostra de ensaio. Em geral, não se tem o valor verdadeiro da amostra de ensaio, calculando-se, então, o valor designado.

Neste caso os valores designados foram os valores de consenso dos laboratórios participantes, obtidos por meio da média ou da mediana dos resultados. No caso de cálculos do valor designado baseado na média, empregou-se primeiramente algum teste de detecção para exclusão de "outliers", de forma que esses não pudessem afetar fortemente as estimativas da média (valor designado) e do desvio padrão. Ao utilizar a mediana como valor designado, não houve necessidade de aplicar testes de detecção de valores dispersos, já que é uma estatística robusta. Para que o desvio padrão também fosse obtido de forma robusta, foi calculado a partir da equação de Horwitz, ou da equação de Horwitz modificada por Thompson[9], ambas baseadas na mediana dos resultados.

Foram comparados quatro testes de detecção de valores dispersos: Dixon, Grubbs, Z-score e Hampel[10,4], sendo este último baseado em estatística robusta. Após a aplicação dos testes de detecção de "outliers", foi possível obter as estimativas para os valores designados por meio da média e desvios padrão e, dessa forma, foram então estabelecidos os intervalos de aceitação.

Os critérios descritos para estabelecer o valor designado, o desvio padrão e o intervalo de aceitação estão combinados de diferentes maneiras nas Tabelas 1 e 2, totalizando sete tratamentos estatísticos (A, B, C, D, E, F e G), que foram avaliados conforme os objetivos e as necessidades de um Programa Interlaboratorial para Chumbo em sangue.

RESULTADOS E DISCUSSÃO

Observa-se pela Tabela 3 que, dentre todos os testes para exclusão de valores dispersos avaliados, o teste de Dixon é o que detecta a menor quantidade de dispersos. Necessita de valores tabelados, que dependem do tamanho da amostra e do nível de confiança adotado. Além disso, Dixon testa somente o menor ou maior valor presente na amostra, havendo necessidade, em alguns casos, da realização de vários passos para se detectar todos os "outliers".

Já o teste de Grubbs apresenta as mesmas dificuldades com relação a valores tabelados e a número de passos até a obtenção de todos os valores de interesse. No entanto, detecta, em geral, pelo menos o mesmo número de valores dispersos que o teste de Dixon e é menos trabalhoso em termos de cálculos matemáticos[4].

O Z-score utiliza o mesmo procedimento iterativo de eliminação dos valores dispersos do teste de Grubbs, com a vantagem de não depender de tabelas para sua realização, já que usa o índice z e avalia os valores de aceitação dentro do intervalo fixo de -2 a 2, com 95% de confiança. Por fixar um intervalo, esse teste torna-se mais rigoroso, sendo, dentre todos os testes aqui utilizados, aquele que mais detecta valores dispersos, levando a uma estimativa menor do desvio padrão.

O teste de Hampel é o único, dentre os avaliados, baseado em estatísticas robustas. Não depende de valores tabelados que variam com o tamanho da amostra e é preciso aplicá-lo somente uma vez, o que o torna menos trabalhoso em relação aos demais. Em comparação aos dois primeiros testes citados, detecta mais observações discrepantes.

Ao considerarmos as técnicas cujos valores designados foram obtidos a partir da mediana, nota-se que os desvios padrão, calculados por meio das equações de Horwitz e de Horwitz modificada por Thompson, são semelhantes. A equação de Horwitz modificada por Thompson leva em consideração a faixa de concentração do analito, porém, para as concentrações usualmente encontradas para o chumbo sangüíneo, μg/100mL, as duas equações são coincidentes. Os resultados obtidos para o valor designado e o intervalo de aceitação das três amostras escolhidas utilizando os diferentes critérios estatísticos encontram-se nos Tabelas 4 e 5.

Os valores designados obtidos pelos critérios estatísticos que empregam a média (A, B, C, D e E) não apresentaram diferenças estatisticamente significantes para as amostras X1, X2 e X3, cujos valores de p com 95% de confiança foram respectivamente: 0,988, 0,922 e 0,801. Usando as técnicas F e G, os valores designados obtidos pelo emprego das medianas mostraram-se semelhantes aos anteriormente obtidos pela

Tabela 1. Critérios estatísticos utilizados para o cálculo do valor designado e do intervalo de aceitação baseados na média.

Critério Estatístico	Teste de Detecção De outliers	Intervalo de aceitação
A	Hampel	$X_{designado} \pm 2s_A$
B	Z-score	$X_{designado} \pm 2s_B$
C	Grubbs	$X_{designado} \pm 2s_C$
D	Dixon	$X_{designado} \pm 2s_D$
E	Z-score	$X_{designado} \pm 6$, para concentrações ≤40 μg/100mL e $X_{designado} \pm 15\% X_{designado,}$ para concentrações > 40 μg/100mL

s_A, s_B, s_C e s_D são os desvios padrão obtidos a partir de cada um dos critérios estatísticos apresentados e calculados com os resultados válidos após a aplicação dos testes de detecção de valores dispersos.

Tabela 2. Critérios estatísticos para a determinação do valor designado e do intervalo de aceitação baseados na mediana.

Critério estatístico	Cálculo do desvio padrão	Intervalo de aceitação
F	$s_F = \frac{2^{(1-0,5\log C)} \times mediana}{100}$	$X_{designado} \pm 2s_F$
G	$s_G = 0,02C^{0,8495}$, $1,2.10^{-7} \leq C \leq 0,138$	$X_{designado} \pm 2s_G$

C = razão de concentração expressa em g/g

média.

Os resultados obtidos são concordantes com LINSINGER et al[4] que compararam diferentes critérios utilizados na eliminação dos dispersos e verificaram que a média, ao contrário do desvio padrão, não é muito afetada pelo critério utilizado. Isso significa que o valor designado não é muito influenciado pelo critério de exclusão selecionado, o mesmo não ocorrendo com os cálculos do intervalo de confiança e do critério de desempenho que, em geral, envolvem o valor do desvio padrão. Por esse motivo, o critério utilizado para eliminação de valores dispersos de aceitação deve estar claro no relatório do programa enviado aos participantes. Os intervalos indicados na tabela 5 são representados nas Figuras 1, 2 e 3. A tabela 6 compara a influência dos critérios sobre a percentagem de laboratórios com resultados considerados satisfatórios. Entre os aspectos a serem considerados na escolha de critérios estatísticos para o cálculo do intervalo de aceitação estão as suas possíveis conseqüências. No caso de análise de chumbo sangüíneo, se o laboratório, por erro na realização da análise, expressar um resultado menor que o real, um ou mais indivíduos podem deixar de ser afastados da fonte de exposição, sofrendo agravos à sua saúde, uma vez que o chumbo é um metal altamente tóxico e cumulativo. Se ocorrer o inverso, no caso de exposição ocupacional, poderão ocorrer perdas econômicas para a empresa ou para a Previdência Social, o que mostra a importância da exatidão de um resultado analítico. Neste caso, o critério estatístico empregado para avaliar o desempenho deve ser mais restritivo e considerar como

Tabela 3. Percentagem de valores dispersos detectados por teste e amostra (X1, X2 ou X3).

Teste de detecção de "outliers"	Percentagem de "outliers" detectados		
	X1 (n=27)	X2 (n=29)	X3 (n=27)
Hampel	7	10	15
Z-score	7	24	15
Grubbs	7	0	0
Dixon	4	0	0

Tabela 4. Valor designado para cada amostra, em µg/100mL, de acordo com a critério estatístico.

Critério estatístico	Valor designado para cada amostra		
	X1	X2	X3
A	16,08	41,31	65,54
B	16,08	42,11	65,54
C	16,08	42,25	67,44
D	16,42	42,25	67,44
E	16,08	42,11	65,54
F e G	16,38	42,00	67,15

Tabela 5. Intervalo de aceitação para cada amostra, em µg/100mL, de acordo com critério estatístico.

Critério estatístico	Intervalo de aceitação para cada amostra		
	X1	X2	X3
A	10,83-21,33	29,72-52,90	54,18-76,89
B	10,83-21,33	34,68-49,54	54,18-76,89
C	10,83-21,33	23,80-60,70	47,71-87,18
D	10,22-22,62	23,80-60,70	47,71-87,18
E	10,08-22,08	35,79-48,43	55,71-75,37
F e G	9,50-23,26	26,69-57,31	44,33-89,97

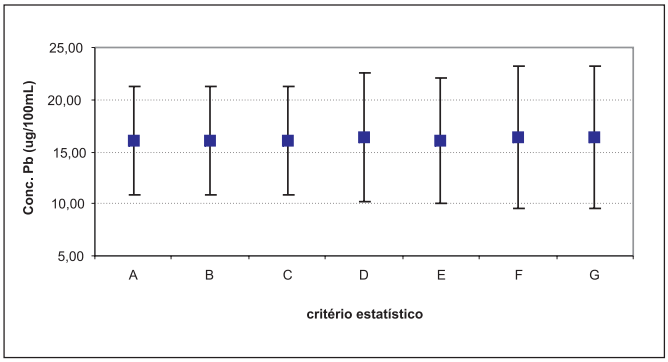

Figura 1.Intervalos de aceitação para a amostra X1.

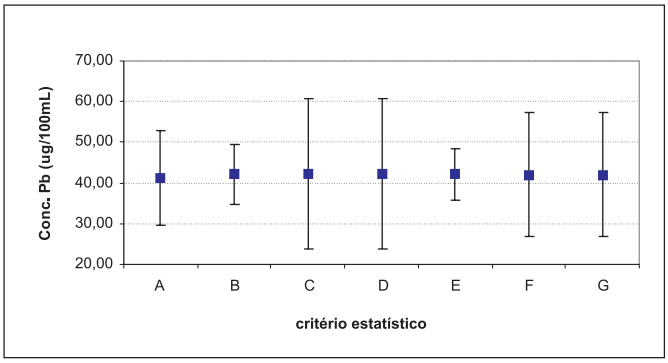

Figura 2.Intervalos de aceitação para a amostra X2.

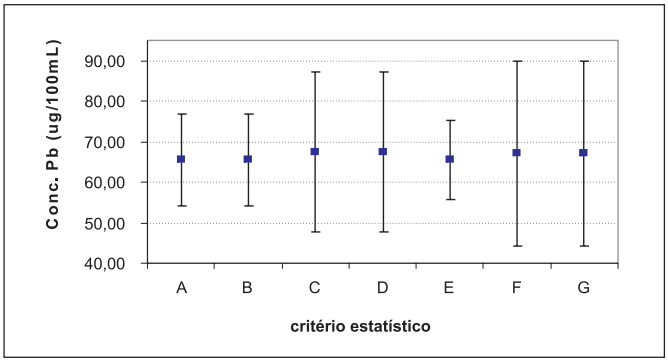

Figura 3. Intervalos de aceitação para a amostra X3.

Tabela 6. Percentagem de laboratórios com desempenho satisfatório por amostra, dependendo do critério estatístico utilizado.

Critério estatístico	**Percentagem de laboratórios com desempenho satisfatório**		
	X1 (n=27)	X2 (n=29)	X3 (n=27)
A	85	83	85
B	85	76	85
C	85	90	93
D	93	90	93
E	93	69	70
F e G	93	90	96

intervalo de aceitação uma faixa mais estreita de valores de concentração.

Analisando os resultados, verifica-se que para a amostra de baixa concentração (X1), os diferentes critérios empregados resultam em intervalos de aceitação semelhantes, o mesmo não ocorrendo para as demais concentrações.

Nas Figuras 2 e 3, tornam-se mais visíveis as diferenças obtidas no Intervalo de aceitação pelo emprego dos diferentes critérios. Verifica-se que os critérios C e D não se mostraram adequadas para o caso da determinação de chumbo sangüíneo, uma vez que, para a amostra X2, consideraram aceitáveis resultados de laboratórios com concentrações de Pb que variam desde o nível de não exposição para adultos até o Índice Biológico Máximo Permitido, de 60 μg/100 mL[11]. Para a amostra X3 o comportamento foi semelhante. Os critérios F e G apresentaram-se como equivalentes entre si para o cálculo dos intervalos de aceitação, os quais também se mostraram pouco restritivos, sendo que nas amostras de concentração alta os intervalos foram ainda superiores aos definidos pelas técnicas C e D. Para as amostras de concentrações média e alta, as técnicas A, B e E mostraram-se progressivamente mais restritivas.

No caso do critério estatística E, que é a empregada atualmente no Programa, o intervalo de aceitação independe da dispersão dos resultados do grupo, uma vez que não inclui esse parâmetro no seu cálculo. Este intervalo de aceitação também é o utilizado pelo PICC da Espanha e baseado no programa da United Kingdom External Quality Assessment Scheme for Lead in Blood (UKEQAS)[12]. O critério também leva em conta a dificuldade analítica, pois à medida que a concentração do chumbo sanguíneo diminui, o intervalo torna-se mais aberto. Assim, por exemplo, o valor 6 μg/100 mL, que é a semi-amplitude do intervalo de aceitação, representa 15% da concentração quando o valor designado é 40 μg/100 mL; já para um teor de 20 μg/100 mL, esse mesmo valor representa 30% da concentração.

CONCLUSÕES

A utilização de diferentes critérios estatísticos empregados para o tratamento de resultados de um programa interlaboratorial mostrou que o valor designado não foi significativamente influenciado pelo critério utilizado. No entanto, o intervalo de aceitação, que depende da estimativa do desvio padrão na maioria dos casos aqui analisados, variou dependendo do critério estatístico adotado.

Nem todos os critérios estatísticos comumente empregados para o tratamento de dados de um PI são adequados para o caso do Programa de Ensaio de Proficiência para chumbo em sangue, uma vez que o intervalo de aceitação deve ser mais restritivo e levar em consideração os agravos à saúde de indivíduos ou populações expostas que podem resultar de decisões incorretas.

O critério estatístico E, atualmente empregado no Programa, mostrou-se adequado aos três níveis de concentração e tem a particularidade de levar em consideração a dificuldade analítica do ensaio. Esse critério pode inclusive ser utilizado em programas interlaboratoriais cujos laboratórios participantes não estejam ainda completamente harmonizados.

REFERÊNCIAS

1. Associação Brasileira de Normas Técnicas [ABNT]. ABNT ISO / IEC Guia 43: ensaios de proficiência por comparações interlaboratoriais. Rio de Janeiro; 1999.
2. Associação Brasileira de Normas Técnicas [ABNT]. ABNT ISO / IEC Guia 17025: requisitos gerais para a competência de laboratórios de ensaios e calibração. Rio de Janeiro; 2005.
3. Davies PL. Statistical evaluation of interlaboratory tests. Fresenius Z Anal Chem 1988;331:513-9.
4. Linsinger TPJ, Kandler W, Krska R, Grasserbauer M. The influence of different evaluation techniques on the results of interlaboratory comparisons. Accred Qual Assur 1998;3:322-7.
5. Tholen DW. Statistical treatment of proficiency testing data. Accred Qual Assur 1998;3:362-6.
6. Analytical Methods Committee. Robust Statistics – How not to reject outliers. Part 1. Basic Concepts. Analyst 1989;114: 1693-7.
7. Chui QSH, Bispo JMA, Iamashita CO. Comparação de técnicas estatísticas aplicadas à avaliação de resultados de programas interlaboratoriais. Anais do III Congresso Internacional de Metrologia em Química, Curitiba, PR, de 30 de setembro a 02 de outubro de 2002.
8. World Health Organization. Regional Office for Europe. External quality assessment of health laboratories. Copenhagen; 1981.
9. Thompson M. Recent trends in inter-laboratory precision at ppb and sub-ppb concentrations in relation to fitness for purpose criteria in proficiency testing. Analyst 2000;125:385-6.
10. International Organization for Standardization [ISO]. ISO 5725-2:1994(E). Accuracy (trueness and precision) of measurement methods and results – Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. Genève; 1994.
11. Brasil. Norma Regulamentadora nº 7, Portaria nº 24, de 29.12.1994 da Secretaria de Segurança e Saúde no Trabalho. Diário Oficial [da] República Federativa do Brasil, Poder Executivo, Brasília, DF, 30 dez. 1994.
12. Bullock DG, Smith NJ, Whitehead TP. External quality assessment of assays of lead in blood. Clin Chem 1986;32(10):1884-9.