Novo método estatístico para análise da reprodutibilidade

José Alberto Martelli Filho*, Liliana Ávila Maltagliati**, Fábio Trevisan***,
Cássia Teresinha Lopes de Alcântara Gil****

Resumo

Este artigo propõe dois métodos estatísticos alternativos para avaliar a reprodutibilidade e o erro do método em trabalhos científicos que envolvam medidas quantitativas. Para a demonstração destes métodos foram utilizados dados reais de duas dissertações de mestrado apresentadas à Faculdade de Odontologia da Universidade Metodista de São Paulo. Os métodos utilizados foram propostos por Lin, Bland e Altman. Uma das vantagens destas análises em relação às tradicionalmente utilizadas, como o erro de Dahlberg, teste *t* pareado e coeficiente de correlação de Pearson, é que se pode utilizar um mínimo de 10 medidas em pares, livres de distribuição (não-normal), exigência que existe quando se utilizam testes paramétricos como teste *t* pareado e o coeficiente de correlação de Pearson.

Palavras-chave: Reprodutibilidade. Cefalometria. Estatística.

INTRODUÇÃO E REVISÃO DA LITERATURA

A pesquisa científica pode ser caracterizada como atividade intelectual intencional que visa responder às necessidades humanas, percebidas no indivíduo como sensação permanente de insatisfação. Pesquisar é o exercício intencional da pura atividade intelectual, visando melhorar as condições práticas de existência. Para que a pesquisa científica aconteça é necessário estar imbuído do espírito científico[6].

O espírito científico é uma atitude ou disposição subjetiva do pesquisador que busca soluções sérias, com métodos adequados, para o problema em questão. Traduz-se no senso de observação, no gosto pela precisão e pelas idéias claras, na imaginação ousada, mas regida pela necessidade da prova, na curiosidade que leva a aprofundar os problemas, na sagacidade e no poder de discernimento. O espírito científico deve manter sempre acesa a inquietação e a persistência em busca de novos achados e novos arranjos para o conhecimento[7].

Ainda que um primeiro método estatístico seja predominante nos estudos da verificação de erros casuais e sistemáticos entre duas séries de medidas para uma mesma grandeza, seja ela cefalométrica ou não - método proposto por Dahlberg (1940 apud HOUSTON[8], 1983), este trabalho vem propor uma nova maneira na busca pela eficiente

* Mestre em Odontologia, área de concentração Ortodontia, pela Universidade Metodista de São Paulo - UMESP.
** Professora Dra. Adjunta do Programa de Pós-Graduação em Odontologia, área de concentração Ortodontia da Universidade Metodista de São Paulo.
*** Mestre em Odontologia, área de concentração Ortodontia, pela Universidade Metodista de São Paulo - UMESP.
**** Professora Dra. do Programa de Pós-Graduação em Odontologia, área de concentração Ortodontia da Universidade Metodista de São Paulo.

análise da reprodutibilidade, descrevendo outro método estatístico para esta, comparando as metodologias e analisando as diferenças entre o método consagrado e o proposto.

Nos vários estudos radiográficos longitudinais existentes na literatura sobre reprodutibilidade, os autores têm usado o teste de Dahlberg (1940 apud HOUSTON[8], 1983) como forma de avaliação desta ocorrência. No entanto, Lin[10]; Bland e Altman[4] propõem métodos estatísticos de grande valor para avaliar a reprodutibilidade, de maneira que se analisa a acurácia para duas medidas repetidas de uma mesma grandeza (coeficiente de concordância[10]) e a variabilidade existente entre os dois momentos para medidas repetidas (limites de concordância[4]).

A marcação dos pontos cefalométricos é uma tarefa imprecisa e os erros associados devem ser quantificados e compreendidos[3]. Assim, os resultados de um estudo cefalométrico devem ser interpretados tendo-se em mente a incorporação de erros que neles existem. Battagel[3] considerou que nenhum método de avaliação de erros pode dar uma completa informação, porém, o de Dahlberg (1940 apud HOUSTON[8], 1983) constitui-se no meio matemático mais confiável de avaliação de erros de medições em associação com o coeficiente de confiabilidade descrito por Midtgärd et al. (1947 apud BATTAGEL[13], 1993).

Em um estudo comparativo entre análise cefalométrica pelo método manual e computadorizado, Brangeli et al.[5] utilizaram 50 telerradiografias em norma lateral de pacientes tratados na clínica de pós-graduação em Ortodontia da Faculdade de Odontologia de Bauru da Universidade de São Paulo e avaliaram a reprodutibilidade do método computadorizado e o erro entre dois examinadores, utilizando o teste de Dahlberg. Encontraram diferença estatisticamente significante em apenas uma entre as 16 grandezas utilizadas, para ambos os examinadores. Desta forma, concluíram que o método computadorizado indireto, empregando imagens digitais, quando comparado ao método manual, mostrara-se confiável e de boa reprodutibilidade. Porém, salientaram que a inclusão de erros ocorreu tanto na comparação entre os métodos como entre os examinadores, principalmente para as grandezas cefalométricas envolvendo os dentes e que o emprego de testes de reprodutibilidade e de erros de metodologia são imprescindíveis para a pesquisa científica.

Martelli[12] estudou o posicionamento da cabeça para a obtenção de telerradiografias, com duas metodologias. Utilizou 60 radiografias, repetidas com um intervalo médio de 30 dias. A reprodutibilidade foi testada com o coeficiente de concordância proposto por Lin[10], para a verificação da variabilidade entre os dois momentos, comparando os dois métodos de posicionamento do paciente. O dois métodos mostraram concordância de 0,823, em média. A utilização do coeficiente de Lin[10] permitiu apurar, numérica e graficamente, a concordância entre dois métodos distintos e a confiabilidade de um método, isoladamente, pela comparação de medidas repetidas.

Trevisan[14] também utilizou o método proposto por Lin[10], além do procedimento proposto por Bland e Altman[4], comparando, na verificação da reprodutibilidade da marcação dos pontos cefalométricos, os dois métodos com o já conhecido método proposto por Dahlberg (1940 apud HOUSTON[8], 1983). Os resultados obtidos sugeriram uma nova maneira de verificação da reprodutibilidade, não baseada em médias, mas na correspondência entre as medidas obtidas nos dois momentos distintos e na distribuição dos valores para a diferença entre as medidas em tempos diferentes.

Lin[10], Bland e Altman[4] detectaram deficiências nos métodos utilizados até então, para verificação da reprodutibilidade em medições repetidas. O método proposto por Lin[10] foi desenvolvido com a intenção de validar medidas de novos instrumentos, comparando estas novas medidas com outras geradas por métodos já consagrados (*gold-standards*), mas que poderia ser utilizado para a verificação da concordância entre dois pares de

medidas, de uma mesma amostra, em tempos diferentes. O valor do coeficiente de concordância (ρ c) pode variar entre -1 e 1, indicando concordância positiva máxima quando o valor é positivo (ou seja, a medida para determinada grandeza seria igual à segunda medida desta). A fórmula para o cálculo de ρ c é:

$$\rho_c = \frac{2\sigma_1\sigma_2.\rho}{2\sigma_1\sigma_2 + (\sigma_1-\sigma_2)^2 + (\mu_1-\mu_2)^2}$$

Onde σ_1 e σ_2 correspondem às variâncias da primeira e segunda série de medidas, ρ é o valor do coeficiente de Pearson para as duas séries, e μ_1 e μ_2 são as médias para as duas séries[9]. Apesar da complexidade da fórmula, os valores utilizados (variância e média, coeficiente de correlação de Pearson) estão disponíveis em programas de uso rotineiro (Microsoft Excel). Há um programa estatístico específico (Stata, Stata Corp., Estados Unidos da América) que provê um intervalo de confiança para a média do coeficiente de concordância, permitindo a análise da sua variabilidade. Este programa também fornece um gráfico para o coeficiente de reprodutibilidade (Fig. 1), facilitando a verificação visual da concordância.

Porém, apesar do gráfico e intervalo de confiança para o coeficiente de concordância, a acurácia não pode ser prontamente analisada, visto que o valor deste coeficiente é adimensional. Espera-se que, se duas séries de medidas (por exemplo, de uma mesma radiografia, analisadas para a verificação dos erros casuais e sistemáticos, que é um procedimento comum em estudos cefalométricos) apresentarem valores muito próximos uma da outra ou idênticos, quando se plotar uma série contra a outra, obter-se-á uma reta de perfeita concordância entre os pontos, com inclinação de 45°, conforme figura 2.

Problemas surgem quando as duas séries de medidas não possuem valores idênticos para os dois momentos, evento comum em estudos de reprodutibilidade. O coeficiente de correlação de Pearson mede a correlação entre as duas séries de medidas, mas não o quanto as medidas desviam da reta a 45°, portanto, impossível detectar a acurácia entre as duas séries de medidas feitas. Existe a premissa da distribuição normal das duas séries de medidas quando se utiliza o coeficiente de correlação de Pearson, fato este que não deveria ser subestimado em estudos sérios[1,10]. De acordo com Altman[2], se há uma mudança na escala das medidas (por exemplo, uma série de medidas feitas com metade do valor da segunda série),

	medida 1	medida 2
média	71,91	71,93
variância	16,45	20,57
Pearson	0,88	
pc	0,88	
intervalo 95%	0,803	-0,923

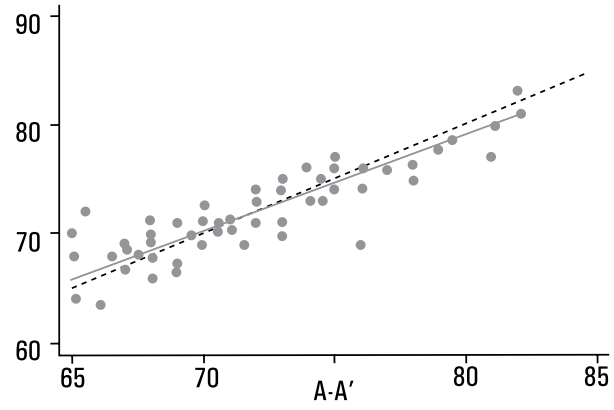

FIGURA 1 - Gráfico representativo do coeficiente de concordância entre pares de medidas (para perfeita concordância, as linhas cinza e tracejada devem se sobrepor).

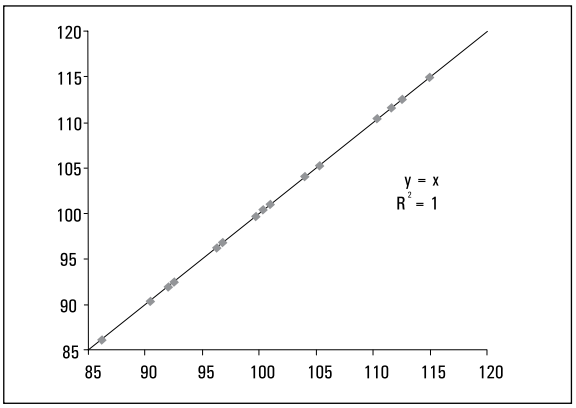

FIGURA 2 – Concordância perfeita entre as duas séries de medidas.

o coeficiente de correlação de Pearson não se altera, submetendo o pesquisador a erros grosseiros com relação à reprodutibilidade (Fig. 3A). Também há outra influência negativa no valor do coeficiente de Pearson, quando a variação na série de medidas é muito alta, originando valores maiores que em séries, com maior dispersão dos valores (Fig. 3B).

Outro procedimento muito utilizado para verificação da igualdade estatística entre as médias obtidas de duas séries medidas é o teste *t* pareado.

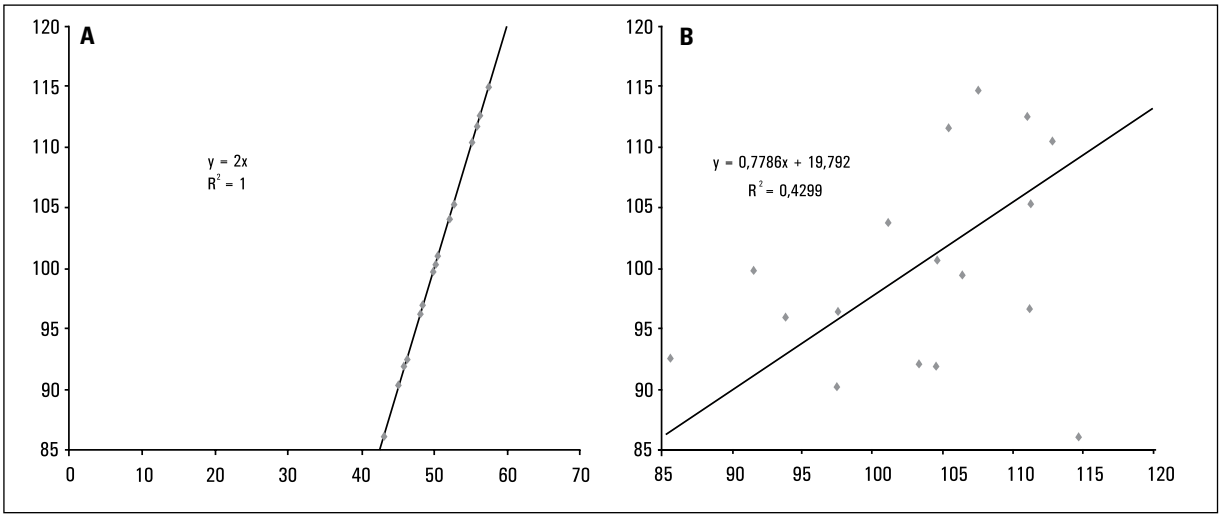

FIGURA 3 - **A)** Correlação entre duas séries de medidas, onde uma equivale à metade de outra. **B)** Correlação entre duas séries de medidas com elevada dispersão.

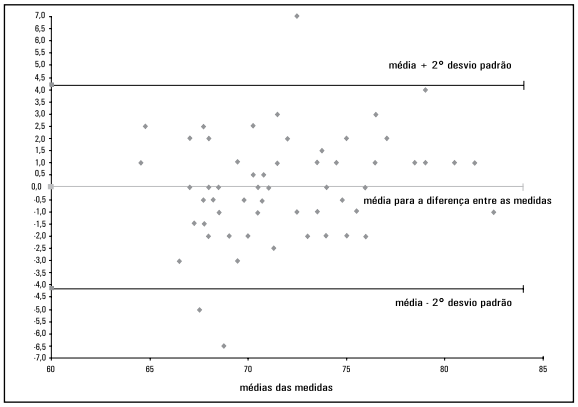

FIGURA 4 - Limites de concordância para os dados da tabela 1.

De novo, a premissa da distribuição normal das medidas se faz necessária, o que exigiria a repetição de uma quantidade enorme de medições, nem sempre possível no âmbito da Biologia. E, novamente, a dispersão dos valores (valor elevado para o desvio padrão) poderia influenciar negativamente o resultado deste teste, mascarando conclusões[2].

Como complemento ao coeficiente de concordância proposto por Lin[10], sugere-se a análise dos limites de concordância, conforme propõem Bland e Altman[4]. O procedimento é simples, bastando

plotar a diferença entre as duas medidas pela sua respectiva média (Fig. 4).

Os autores sugerem que, como provavelmente a diferença resultante tem distribuição normal (desde que o número de medidas repetidas não seja muito pequeno, fato que é respeitado nos estudos de reprodutibilidade), a variação existente para a diferença entre as medidas é igual à média

Tabela 1 - Dados do gráfico da figura 4.

Pacientes	Momento 1						Momento 2					
	A-A'	P-P'	ENA-ENA'	ENP-ENP'	VS-PP	VS-GoMe	A-A'	P-P'	ENA-ENA'	ENP-ENP'	VS-PP	VS-GoMe
1	71	71,5	30,5	38	99	117	73	72	28,5	39	100,5	115
2	66,5	67	40	49,5	101	115,5	69	69	39	49,5	101	116,5
3	69	63	40	38	91	114	68	66,5	33,5	36	92	111,5
4	71	68	42	40	88	109	70,5	67	43,5	41,5	88,5	112,5
5	70	73	51	52	91	111	68	70,5	52,5	52	91	109
6	77	70	52	47	87	122	81	77	47,5	47	89	120
7	68,5	71	41,5	44	91	123	68,5	71	40	42,5	92	124,5
8	64	58	51,5	46	85,5	114,5	65	60	50	46,5	86	115
9	69	59	34	42	98,5	116	70	61	34	42	99,5	115
53	74	70	43,5	46	93	110	76	71	44	46	92	108
54	81	81	41	43,5	92	112	82	84	38	42	93	111
55	74	73,5	36	41	96	117	75	74	37	42	96	118
56	63,5	63	46	45	90	113	66	64	46	45	90	115
57	71	69	50	46	88	115	68	62,5	51,5	48,5	86	116,5
58	75	72	35	39,5	95	115	73	68,5	37	41	93	117
58	70	56	44	41	87	128,5	65	50	46	42,5	86	132
60	76	74	38,5	38,5	90	117,5	75	72	39	37,5	89	118,5

As medidas A-A', P-P', ENA-ENA', ENA-ENP' estão expressas em milímetros.
As medidas VS-PP e VS-Go-Me estão expressas em graus.
Obs.: foram omitidos dados, para que a tabela pudesse ser colocada neste artigo.

Tabela 2 - Comparativo entre os métodos descritos.

Angulos e proporções	Limites de concordância[1]	Coeficiente de concordância[2]	Erro casual[3]	Média das diferenças	Desvio padrão	Erro padrão
nasolabial	8,05	0,958 - 0,996	1,63	1,18	2,04	0,57
nasolabial superior: nasolabial	9,87	0,672 - 0,965	1,81	-0,01	0,02	0,01
convexidade sem nariz	2,23	0,966 - 0,997	0,41	0,18	0,57	0,16
convexidade com nariz	3,55	0,930 - 0,994	0,63	-0,21	0,90	0,25
mentolabial	15,30	0,870 - 0,984	2,66	-0,24	3,90	1,08
terço inferior da face	2,12	0,987 - 0,999	0,40	-0,22	0,54	0,15
mentocervical	2,34	0,985 - 0,998	0,41	-0,07	0,60	0,17

1 - em graus ou milímetros[4].

2 - intervalo de confiança para 95%[10].

3 - em graus ou milímetros[8].

das diferenças somada ou subtraída de duas vezes o desvio padrão das diferenças. A figura 4 ilustra o exposto (as medidas que originaram o gráfico estão na tabela 1). A linha em cinza corresponde à diferença média entre as duas séries de medidas, e é claro que, se houver perfeita concordância entre as medidas, esta diferença será zero. Porém, como se trata da média das diferenças, alguém poderia questionar qual a variação existente para as diferenças par a par, o que fica fácil de se observar a partir da verificação dos pontos plotados (em cinza). As linhas superior e inferior (em preto) à linha da igualdade correspondem à variação existente nas observações efetuadas. Fica sob a responsabilidade do pesquisador aceitar ou não a variação observada.

DESCRIÇÃO DO MÉTODO E UTILIZAÇÃO

Para a demonstração dos novos métodos, foram utilizados dados reais das dissertações apresentadas como parte dos pré-requisitos para obtenção do título de Mestre em Odontologia, área de concentração Ortodontia, da Universidade Metodista de São Paulo. Omite-se, neste artigo, a metodologia utilizada para a obtenção dos valores, visto que são colocados única e somente para a ilustração dos métodos estatísticos.

Para o cálculo do coeficiente de concordância de Lin[10] e limites de concordância proposto por Bland e Altman[4] pode-se utilizar os recursos disponíveis em planilhas eletrônicas, como o Excel (Microsoft). Para o cálculo deste coeficiente para várias grandezas, a utilização de um programa estatístico específico deve ser planejada.

Como exemplo, um intervalo de confiança para o coeficiente de concordância entre 0,958 e 0,996 (valor próximo ao da máxima concordância) para o ângulo nasolabial não nos permite visualizar a variação existente na diferença entre as medidas realizadas em tempos diferentes (no total, 8°). Verifica-se, também, que para um intervalo de confiança para o coeficiente de concordância amplo (ângulo mentolabial 0,870 – 0,984), há limites de concordância mais amplos (15,30°, no total).

A vantagem do método proposto por Lin[10], em relação aos freqüentemente utilizados (erro de Dahlberg, teste *t* pareado e coeficiente de correlação de Pearson), é que se pode utilizar um mínimo de dez medidas em pares, que podem ser livres de distribuição (existe a premissa da distribuição normal dos valores das grandezas quando se utiliza prova paramétrica, como o teste *t* pareado e o coeficiente de correlação de Pearson, como comentado no texto anteriormente).

Para exemplificar a utilização e interpretação dos limites de concordância, utilizaremos a figura 4. Pode-se perceber que houve variação de aproximadamente 8° (de -4,0° no limite inferior até 4,0° no limite superior), no total, entre as medidas para a grandeza analisada (ângulo nasolabial). Se o pesquisador julgar conveniente, repete-se o procedimento da marcação dos pontos cefalométricos ou medidas diretas dos traçados, por exemplo, tomando o cuidado de verificar a metodologia utilizada. Pode chegar inclusive à conclusão que tal grandeza (ângulo ou medida linear) apresenta pouca confiabilidade por apresentar variação muito grande, apesar de meticulosidade na metodologia. A análise gráfica dos resultados é simples e intuitiva, porém, uma poderosa ferramenta para os estudos da reprodutibilidade.

DISCUSSÃO

Na tabela 2 pode-se visualizar o que foi exposto neste trabalho. O que se nota é que, apesar de alguns intervalos de confiança para o coeficiente de concordância serem muito pequenos (indicando elevada acurácia), os limites de concordância se mostram relativamente amplos (indicando elevada variabilidade), quando comparados aos valores correspondentes obtidos pelo método proposto por Dahlberg (1940 apud HOUSTON[8], 1983). Houston[8] recomendou maneiras para a detecção de erros sistemáticos, por meio de um teste de comparação entre médias (teste *t* pareado) e por meio de uma avaliação do erro aleatório conforme

proposto por Dahlberg (1940 apud HOUSTON[8], 1983). Recomendou, em primeiro lugar, que deveria haver um número suficiente de duplicações de medidas, sugerindo uma quantidade de 25 duplicações. Depois, sugeriu que o desvio padrão da diferença entre medidas duplicadas fosse dividido por dois (a justificativa seria que haveria erros aleatórios sistematicamente introduzidos em cada momento da medição, portanto, em dois momentos distintos haveria o dobro do erro). Portanto, para vários pares de medidas, Dahlberg (1940 apud HOUSTON[8], 1983) sugeriu que o erro entre medidas tomadas em tempos diferentes seria igual à raiz quadrada da somatória das diferenças entre os pares de medidas (d^2), dividida por duas vezes o número de pares de medidas (n), como na fórmula:

$$Se^2 = \frac{\Sigma d^2}{2n} \quad \text{ou} \quad Se = \sqrt{\frac{\Sigma d^2}{2n}}$$

Porém, ressaltou o autor, que esta fórmula somente seria aplicável quando não houvesse diferença entre as médias destes pares de medidas. Então, o procedimento proposto por Dahlberg (1940 apud HOUSTON[8], 1983) deveria ser utilizado somente sob estas condições. Por exemplo, para o ângulo nasolabial (Tab. 1), os limites de concordância foram 5,2° e -2,8°, para os limites superior e inferior (variação para 95% dos valores de 8,2°, no total, com média e desvio padrão para a diferença entre as medidas 0,03° ± 2,03°, enquanto o erro proposto por Dahlberg (1940 apud HOUSTON[8], 1983) foi de 1,63°. Esta discrepância entre valores ocorre porque a fórmula proposta por Dahlberg (1940 apud HOUSTON[8], 1983) se aproxima da fórmula do erro padrão, que leva em consideração a quantidade de medidas feitas. Para efeitos de comparação, se calculássemos o erro padrão para o ângulo nasolabial, ter-se-ia o quociente de 2,03° pela raiz quadrada de treze (número de medidas feitas em dois momentos, para os dados da tabela 2, tendo como resultado 0,55°, valor três vezes menor que o proposto por Dahlberg.

Ao pesquisador é necessário consciência de todas as possíveis implicações que resultados mal interpretados ou mal elaborados podem trazer: conclusões inconsistentes e duvidosas podem levar ao descrédito pesquisas que são muitas vezes realizadas com investimentos, financeiro e psicológico, elevados. O pesquisador deve conhecer a natureza da grandeza que está medindo (como ela se comporta biologicamente, se está distribuída normalmente, do ponto de vista estatístico); deve conhecer, também, qual a variação existente para cada ponto cefalométrico que vai fazer parte de seu estudo (nos sentidos horizontal e vertical[14,15]), e realizar um ensaio prévio para calibração de sua marcação de pontos cefalométricos. São esses os pré-requisitos fundamentais para que possa desenhar um projeto de seu estudo estatístico.

AGRADECIMENTOS

Ao Dr. Fábio Trevisan pela valorosa colaboração quanto à estatística.

Enviado em: Dezembro de 2003
Revisado e aceito: Abril de 2004

New statistical methods to evaluate reproducibility

Abstract
This article proposes two alternative statistical methods to evaluate reproducibility and method error present in scientific papers that deal with quantitative measurements. Real data was used to demonstrate theirs use. These methods were presented by Lin, Bland and Altman. One of the advantages of these analyses over the traditionally used (the Dahlberg error, paired t test, Pearson´s correlation coefficient) is that it is not required a normal distribution of the data, and only ten paired measurements should be enough. The method proposed by Dahlberg requires a minimum of 25 repeated measurements. The researcher must be aware of the implications that misinterpreted results or inconsistent conclusions may lead.

Key words: Reproducibility. Cephalometry. Statistics.

REFERÊNCIAS

1. ARANGO, G. H. **Bioestatística teórica e computacional**. 1. ed. São Paulo: Guanabara-Koogan, 2001.
2. ALTMAN, D. G. **Pratical statistics for medical research**. 1. ed. New York: Chapman & Hall, 1991.
3. BATTAGEL, J. M. A comparative assessment of cephalometric errors. **Eur J Orthod**, London, v. 15, no. 4, p. 305-314, 1993.
4. BLAND, J. M.; ALTMAN, D. G. Measuring agreement in method comparison studies. **Stat Methods Med Res**, London, v. 8, p. 135-160, 1999.
5. BRANGELI, L. Á. M.; HENRIQUES, J. F. C.; VASCONCELOS, M. H. F.; JANSON, G. Estudo comparativo da análise cefalométrica pelo manual e computadorizado. **Rev Assoc Paul Cir Dent**, São Paulo, v. 54, n. 3, 2000.
6. CARVALHO, A. M.; MORENO, E.; BONATO, F. R. O.; SILVA, I. P. **Aprendendo metodologia científica**: uma orientação para alunos de graduação. 1. ed. São Paulo: O Nome da Rosa, 2000.
7. CERVO, A. L.; BERVIAN, P. A. **Metodologia científica**. 4. ed. São Paulo: Makron Books, 1996.
8. HOUSTON, W. J. B. The analysis of errors in orthodontic measurements. **Am J Orthod**, St. Louis, v. 83, no. 5, p. 383-390, May 1983.
9. KRUMMENAUER, F.; DOLL, G. Statistical methods for the comparison of measurements derived from orthodontic imaging. **Eur J Orthod**, London, v. 22, p. 257-269, 2000.
10. LIN, L. I. A concordance correlation coefficient to evaluate reproducibility. **Biometrics**, Washington, D. C., v. 45, no.1, p. 255-268, Mar. 1989.
11. LOPES, P. A. **Probabilidades & estatística**. 1. ed. Rio de Janeiro: Reichmann & Affonso, 1999.
12. MARTELLI, J. A. F. **Estudo da reprodutibilidade na obtenção das telerradiografias em norma lateral pelo método da posição natural da cabeça**. 2003. 154 f. Dissertação (Mestrado em Ortodontia) – Faculdade de Odontologia da Universidade Metododista de São Paulo, São Bernardo do Campo, 2003.
13. BATTAGEL, J. M. A comparative assessment of cephalometric errors. **Eur J Orthod**, London, v. 15, no. 4, p. 305-314, 1993.
14. TREVISAN, F. **Análise fotogramétrica e subjetiva do perfil facial de jovens brasileiros, leucodermas, com oclusão normal**. 2003. 146 f. Dissertação (Mestrado em Ortodontia) – Faculdade de Odontologia da Universidade Metodista de São Paulo, São Bernardo do Campo, 2003.
15. TRPKOVA, B.; MAJOR, P. Cephalometric identification and reproducibility: a meta analisys. **Am J Orthod Dentofacial Orthop**, St. Louis, v.112, p. 165–170, Aug. 1997.

Endereço para correspondência
José Alberto Martelli Filho
Rua Sinharinha Frota, 1061 Centro
CEP: 19.990-060 Matão/SP
E-mail: martelli@process.com.br