Arquivos de Ciências do Esporte
Archives of Sport Sciences

Ponto de vista

doi: http://dx.doi.org/10.17648/aces.v6n2.2783

Procedimentos post hoc: orientação para praticantes de estatística em ciências da saúde

Post hoc procedures: a guide for statisticians practitioners in health science

Dahan da Cunha Nascimento[1,*]
Cristiane Rocha Silva[1]
Jonato Prestes[1]

Resumo

Procedimentos de comparações múltiplas ou post hoc são fundamentais em pesquisas no âmbito na saúde, esporte, educação e lazer. No entanto, a escolha do teste mais adequado por praticantes de estatística na área das ciências da saúde ainda gera dúvidas. O teste utilizado deve promover proteção contra o erro do tipo I, considerar o número de comparações ou grupos e deve ser conservador. No entanto, se o procedimento é muito conservador, diferenças genuínas podem ser negligenciadas ou aumentar o erro do tipo II. Com isso, o objetivo da presente revisão é apresentar um guia com mais de dez diferentes testes de comparações múltiplas e sugestões para o uso mais adequado desse tipo de estatística.

Palavras-chave: análise estatística, estatística e dados numéricos, protocolos.

Abstract

Procedures of pairwise multiple comparisons or post hoc tests are of great importance for health, sport, educational and recreational areas. However, doubts are commonly raised for the choice of the appropriate post hoc test by statisticians of health science. The pairwise multiple comparisons should give protection against type I error, must consider the number of comparisons or groups, and have adequate power. However, if the post hoc test is too conservative, there is a risk to miss an important effect or increase the type II error. Thus, the goal of the present review is to present a guide with more than ten pairwise multiple comparisons and suggestions for the adequate use of this type of statistics.

Key words: statistical analysis, statistics and numerical data, protocols.

Afiliação dos autores

[1]Programa de Pós-Graduação *Stricto Sensu*, Faculdade de Educação Física, Universidade Católica de Brasília, Brasília, Brasil.

**Autor correspondente*

Programa de Pós-Graduação em Educação Física, Universidade Católica de Brasília, Q.S. 07, Lote 01, EPTC, Bloco G, CEP: 71966-700, Distrito Federal, Brasília, Brasil.
e-mail: dahanc@hotmail.com.

Conflito de interesses

Os autores declararam não haver conflito de interesses.

Processo de arbitragem

Recebido: 27/03/2018
Aprovado: 13/08/2018

Introdução

O uso inadequado de métodos estatísticos é comum em pesquisas na área de ciências da saúde, até mesmo em artigos científicos publicados em jornais de grande impacto. Com isso, é preciso reconsiderar a prática da estatística e estimular disciplinas com o treinamento adequado em estatística em departamentos de pós-graduação, adaptar a estatística para a área de estudo do pesquisador/aluno e desenvolver ferramentas para promover educação e disseminação do conhecimento da estatística[1].

Dentre esses procedimentos estatísticos, os testes post hoc ou de comparações múltiplas na área de ciências da saúde ainda geram dúvidas para praticantes de estatística especialmente na área da Educação Física. O teste post hoc ideal é justamente aquele que protege contra a inflação de se realizar inferências falso-positivas. Além disso, muitos testes de post hoc já foram descritos e seu uso em situações ou circunstâncias específicas ainda deixam dúvidas[2,3].

Portanto o objetivo da presente revisão é apresentar testes de comparações múltiplas apropriados para detectar diferenças quando elas realmente existem em experimentos no âmbito da saúde e do esporte.

Delineamentos entre e dentre participantes

Durante a coleta dos dados podemos expor os participantes da pesquisa a diferentes manipulações de experimento, da forma entre participantes (também conhecido como delineamento independente) e dentre participantes (também conhecido como delineamento dependente). Para melhor entendermos como funciona o post hoc o exemplo citado abaixo será de fundamental importância.

Além disso, cabe ressaltar que somente testes paramétricos serão utilizados no presente artigo. No entanto, os autores ressaltam que a normalidade dos dados é uma das suposições frequentemente analisadas para determinar se procedimentos paramétricos (testes utilizados para dados que apresentam uma distribuição normal) ou não paramétricos (testes utilizados para dados que apresentam uma distribuição não normal)[4] serão utilizados.

Exemplo 1

Durante a graduação no curso de Educação Física, um aluno de iniciação científica (IC) procurou saber se o treinamento de força (TF) podia ser usado como uma forma de tratamento não farmacológico na hipertensão arterial sistêmica. Seguindo minuciosamente as orientações de seu professor ele fez uma revisão de literatura e encontrou evidências de que o TF é capaz de diminuir a pressão arterial (PA) em indivíduos hipertensos. Como ele queria se especializar em idosos e queria que esse tema fosse o seu projeto de trabalho de conclusão de curso. Decidiu estudar sobre o assunto e comparou a PA antes de realizar o TF (pré-treino PAS) e após 14 semanas de TF (pós-treino PAS) em idosas hipertensas.

Como eram as mesmas participantes (ver tabela 1) a comparação dentre participantes foi realizada (delineamento dependente). Por isso, como são dois momentos e os mesmos indivíduos analisados, o test t dependente é o mais indicado para esse tipo de delineamento.

Tabela 1
Dados da pressão arterial apresentados através da média e desvio padrão (DP)*.

	Momento 1	Momento 2	Momento 3
Participantes	Pré-treino PAS	Pós-treino PAS (14 semanas)	Pós-treino PAS (28 semanas)
1	160,00 mmHg	130,00 mmHg	122,00 mmHg
2	155,00 mmHg	142,00 mmHg	121,00 mmHg
3	148,00 mmHg	127,00 mmHg	124,00 mmHg
4	158,00 mmHg	125,00 mmHg	119,00 mmHg
5	162,00 mmHg	134,00 mmHg	123,00 mmHg
Média	156,60	131,60	121,80
Desvio Padrão	5,46	6,73	1,92
Valores de p do teste t	Momento 1 versus Momento 2 = 0,002 Momento 2 versus Momento 3 = 0,03 Momento 1 versus Momento 3 = 0,001		

PAS = pressão arterial sistólica; mmHg = milímetros por mercúrio; * = dados fictícios.

Vamos supor agora que ele queira avaliar em três momentos e resolva prolongar o seu projeto avaliando os efeitos após 28 semanas de treinamento. Neste momento, o test t não é o tratamento estatístico mais indicado para se aplicar aos dados, pois temos três momentos, outros tratamentos estatísticos como as análises de variância possuem melhor aderência a esse tipo de delineamento.

Por que não realizar vários testes t para comparar os momentos?

Quando realizamos uma pesquisa e aplicamos múltiplos testes t para a análise de nossos dados aumentamos as taxas de erros, ou melhor, inflacionamos um tipo de erro. Concluímos que houve diferenças entre os tratamentos quando na verdade não houve (erro tipo I). Ao realizarmos múltiplos testes inflacionamos o erro tipo I ou erro de experimento.

Corrigindo o valor de p pelo post hoc de Bonferroni?

Normalmente, em pesquisas científicas o valor de $p \leq 0,05$ é adotado[5]. Então, ao rodar um teste estatístico como o teste de comparações, pode-se obter para os resultados um valor de $p \leq$ ou $>$ que 0,05. O que ficar abaixo ou igual a 0,05 é considerado significativo e o que ficar acima disso é considerado não significativo.

No exemplo do aluno de IC (tabela 1) que rodou o teste t dependente comparando os momentos 1 versus 2 encontrando um $p = 0,002$ concluímos que o TF diminuiu significativamente a PAS das idosas ao final do treinamento. No entanto, como ele decidiu também avaliar depois de 28 semanas, ele comparou o momento 2 versus momento 3 e comparou o momento 1 versus momento 3 e obteve valores de $p = 0,03$ e 0,001, respectivamente.

Na descrição dos resultados do trabalho de conclusão de curso o aluno presumidamente reportou que o protocolo de TF adotado por ele foi significativamente efetivo em reduzir os valores da PAS de mulheres idosas hipertensas. Sendo que a PAS no momento 2 foi significativamente inferior em relação ao momento 1 e ao momento 3 e que o momento 3 foi significativamente inferior em relação ao momento 1.

Agora, avaliando criticamente, fica claro que testes múltiplos foram realizados e que as diferenças encontradas na verdade podem não existir (erro tipo I).

Como então aplicar a correção pelo post hoc de Bonferroni nesses dados?

Para evitar a inflação do erro tipo I em pesquisas da área ciências da saúde sugere-se que uma correção do valor de p seja realizada[6]. A sua correção é muito simples. Apenas divide-se o valor de p pelo número de comparações realizadas ($p = 0,05/3$). Portanto, o valor de $p \leq 0,01$ será o novo nível de significância adotado no exemplo do aluno de IC. Agora, analisando novamente os resultados verifica-se que algumas das diferenças encontradas na verdade não são reais e que ficaram acima da probabilidade de corte.

Em estudos com delineamentos correlacionais onde correlações múltiplas são realizadas, a correção de p pelo post hoc de Bonferroni pode também ser aplicada[7].

O que há de errado com as correções feitas pelo post hoc de Bonferroni?

Esse tópico sobre ajustes do valor de p parece ser crucial para alguns praticantes de estatística e irrelevante para outros[8]. Embora eu pertença ao grupo que considera crucial, acho importante e concordo com a opinião de autores que pensam ao contrário.

Umas das limitações dos ajustes de p é que muitas hipóteses nulas (de que não há diferenças entre os momentos) são aceitas quando na verdade estão erradas (erro tipo II)[9,10]. Portanto, o problema de se utilizar a correção pelo post hoc de Bonferroni é que o controle da taxa de erro pode acarretar a perda do poder estatístico, ou seja, deixamos de encontrar uma diferença genuína nos dados[4]. No entanto, muitos artigos científicos adotam a correção pelo post hoc de Bonferroni. Sugiro aqui a leitura criteriosa da análise estatística de dois

artigos publicados na área de TF com o uso do post hoc de Bonferroni[11,12.]

Cuidados na interpretação do valor de p após as correções pelo post hoc

Uma distorção comumente verificada nos artigos que são publicados é que o termo "diferença estatística" é utilizado ambiguamente, ou seja, é difícil entender pelos resultados de um estudo que o autor quis se referir à diferença estatística ou a importância clínica/prática dos resultados. Verificamos que o valor de p no exemplo 1 ficou abaixo de 0,05 e que a diferença foi também significativa, mas é importante entender que o valor de p muitas vezes não estabelece medida de importância e não fornece mais informação do que isso[13].

No entanto, cito um exemplo abaixo de um estudo no qual o pesquisador distinguiu muito bem a diferença estatística de diferença clínica em seu estudo.

No estudo publicado pelo pesquisador Jannig e colaboradores[14], foram comparados três protocolos e foi avaliado o efeito sobre a pressão arterial de idosos hipertensos. O primeiro protocolo seguiu a ordem de exercícios de membros inferiores para superiores, o segundo fazia o inverso e o terceiro alternava membros superiores e inferiores. Após cada protocolo, a pressão arterial foi verificada em intervalos de 10 minutos, até 60 minutos pós-exercício.

Nos resultados do estudo do pesquisador Jannig e colaboradores[14], somente o protocolo que alternava exercícios apresentou diferenças significativas em relação ao repouso. No entanto, qual a aplicabilidade prática dos resultados?

Os pesquisadores citaram a importância clínica dos resultados e acrescentaram que a redução média da pressão arterial sistólica (PAS) e pressão arterial diastólica (PAD) de 11,3 mmHg e 4,4 mmHG encontradas em seu estudo corrobora com a citação na literatura que decréscimos de aproximadamente 2 mmHg na PAS e PAD reduzem os riscos de acidente vascular encefálico em 14% e 17%, e risco de doença arterial coronariana em 9% e 6%, respectivamente.

Através desse exemplo, entendemos que o valor de p dá uma estimativa de que a probabilidade das diferenças observadas podem ser explicadas pela chance, mas não falam sobre a aplicabilidade prática do resultado[15], ou seja, o quanto que a redução da PAS observada no estudo do pesquisador Jannig e colaboradores14 será importante para o paciente[15].

A interpretação da importância clínica ou prática deve ser realizada pelo pesquisador e leitor. Foi demonstrado, por exemplo, que declínios na PAS de 10 mmHg identificados no exemplo 1 utilizado nesse artigo diminuem em 34% os riscos de acidente vascular encefálico em uma população na faixa etária de 60 – 69 anos de idade[16]. Então, verifica-se que apesar das diferenças encontradas não serem estatisticamente significativas, do ponto de vista clínico/prático, o declínio na PAS após 14 semanas de treinamento foi muito importante para a amostra de idosas hipertensas. A grande mensagem aqui para o leitor é que o valor de p é um pedacinho de uma informação limitada presente em um artigo científico.

Outro fator de grande relevância que se deve controlar em um estudo é o erro de medida. É de fundamental importância assegurar que mudanças provocadas sobre a PAS após 14 semanas de TF sejam reais e não resultantes do erro de medida. Fatores de confundimento como motivação do participante, horário da medida e estado emocional terão diferentes efeitos sobre os resultados de cada score[17].

Nesse sentido, para evitar uma interpretação inadequada dos resultados da pesquisa, o erro padrão de medida ou standard error of measurement (SEM) deve ser aplicado[18]. Esse procedimento estatístico representa o agregado de variáveis (avaliador, participante, protocolo de teste e condições ambientais) que coletivamente podem afetar o verdadeiro valor medido. Para isso, a menor diferença real ou smallest real diference (SRD) deve ser calculada. A SRD trata de um modelo de distribuição que compara as mudanças do score de um participante com uma medida de variabilidade (desvio padrão e SEM) e determinará o que será uma diferença clínica importante de um erro de medida.

Análises de diferenças entre três ou mais condições e procedimento post hoc

Outra diferente forma ou tratamento estatístico que melhor adere aos dados do exemplo 1 é a ANOVA de medidas repetidas[19]. A ANOVA procura verificar a existência de diferenças entre as médias dos grupos ou momentos. A estatística teste F na ANOVA é a forma de sabermos que o resultado encontrado em nosso estudo se deve ao efeito do tipo de treinamento utilizado e não uma variação não explicada pelo modelo ou erro (ao acaso).

Como são as mesmas participantes analisadas em todos os momentos. A ANOVA de medidas repetidas é o teste que apresenta aderência a esse tipo de delineamento[20]. No entanto, ela apenas nos informa que existe uma diferença entre os momentos e não onde está a diferença. Para encontrar exatamente onde essas diferenças estão, executamos testes adicionais para confirmar a nossa hipótese entre os momentos 2 e 1, 3 e 1 ou entre 2 e 3.

Procedimentos post hoc

O procedimento de comparações planejadas ou múltiplas é utilizado para verificar qual momento difere de outros momentos. Muito parecido com o que fizemos anteriormente. Só que, ao invés de aplicar múltiplos testes t para encontrar onde está a diferença realizamos primeiramente a ANOVA e posteriormente os testes post hoc com correção de Bonferroni.

Considerando os três momentos no projeto de IC, ao realizar a ANOVA de medidas repetidas saberemos que diferenças entre os momentos existem, uma das formas de se verificar exatamente onde está essa diferença é por meio da aplicação de múltiplos testes t com a correção dos valores de p pelo post hoc de Bonferroni para não inflacionar o erro do tipo I. Procedimento similar é aplicado quando uma ANOVA de Friedman (mesma coisa que a ANOVA de medidas repetidas, só que aplicada a dados não normalmente distribuídos) é a opção de teste[20].

Em relação a sugestões estatísticas para o uso da correção pelo post hoc de Bonferroni é recomendado com o intuito de corrigir a taxa do experimental-wise error (quando se realiza múltiplos testes envolvendo o mesmo procedimento como t ou r) e corrigir o family-wise error após o uso da ANOVA[8].

Como está além do objetivo desta revisão descrever todas as formas de comparações planejadas, considera-se importante citar e dar uma breve explicação que em outras ANOVAs com mais de uma variável independente, as comparações planejadas recebem outro nome conhecido como análises de efeitos simples. Quando uma ANOVA com mais de uma variável independente é utilizada, interações entre as variáveis podem ocorrer e depois de verificadas as interações, análises de efeitos simples são utilizadas e o mesmo procedimento de correção do valor de p é aplicado.

O exemplo do aluno de IC pode ser usado para que o leitor entenda melhor o conceito de interação e variável dependente. Vamos supor que além do TF outro grupo com treinamento aeróbio (TA) seja também utilizado para saber o seu efeito sobre a PAS em idosas hipertensas. Os grupos, ou melhor, os tipos de treinamento são as variáveis que o pesquisador manipula para saber seus efeitos sobre a PAS (variável dependente) e o tempo (momentos) é também outra variável independente que o pesquisador manipula.

Uma interação ocorre quando o efeito de uma variável independente (tipos de treinamento) sobre a variável dependente muda dependendo do nível de outra variável independente (momentos). No exemplo do aluno de IC isso é equivalente a perguntar quando os efeitos dos tipos de treinamento mudam dependendo do tempo exposto ao treinamento?

Se o efeito do tipo de treinamento no momento 1 (pré-treino) é diferente do efeito do tipo de treinamento no momento 2 (pós-treino), então tivemos uma interação.

Testes post hoc

Considerando que vários procedimentos para comparações estão disponíveis, muitos estatísticos ainda têm dúvidas sobre qual procedimento recomendar. Além disso, pesquisadores na área das ciências da saúde e especialmente da Educação Física ficam confusos quando indagados a escolher qual tipo de teste post hoc aplicar em seus dados. Dessa forma, apresentamos nessa revisão uma explicação didática sobre alguns dos

procedimentos post hoc mais utilizados.Em relação aos testes post hoc, alguns são conservadores e outros são liberais. Para os testes conservadores o seu poder estatístico é maior. O que implica que ele somente identificará uma diferença genuína entre as médias ou momentos quando ela realmente existir. Em outras palavras, que o TA é melhor que o TF como estratégia de intervenção não farmacológica no controle da PA em idosos. Os testes post hoc conservadores apresentam uma menor probabilidade de erro tipo I (falar que o TA é melhor que o TF quando na verdade não é!). Em contrapartida, tanto poder nem sempre é adequado. O aumento do poder também está associado ao erro tipo II (afirmar que o TF não é melhor que o TA quando na verdade é!). Para um melhor entendimento do poder estatístico no contexto da pesquisa aplicada, sugere-se a leitura do artigo de Baguley[21].

Tipos de procedimentos post hoc

1. Post Hoc de Bonferroni: esse teste é muito conservador e controla muito bem o erro tipo I, especialmente quando número de comparações é maior do que cinco[22]. Por isso, apresenta bastante poder.
2. Post Hoc de Fisher's least significant difference (LSD): esse procedimento é um dos mais antigos testes de comparações sendo o primeiro a ganhar aceitação. Recomenda-se que sua aplicação seja realizada quando você tem até três médias para comparação[23]. Acima disso, sugere-se não aplicar. É um teste que não controla o erro tipo I e é equivalente à execução de múltiplos testes t[22]. A diferença entre o LSD e o teste t é que o primeiro requer que a estatística F seja significativa[3].
3. Studentized Range Statistic (q): o teste de q é bastante parecido com o teste t. Quando queremos comparar duas médias (TA versus TF), o teste t é o mais apropriado. Vamos supor agora que temos mais de duas médias e de forma deliberada escolhemos a maior (TA) e menor média (GC) para testar. É aparente que modificamos totalmente a probabilidade do erro tipo I. Considerando que a hipótese nula seja verdadeira (de que não há diferenças entre os tipos de treinamento). A probabilidade de encontrar uma diferença genuína entre a maior e menor média aumenta. Esse é o ponto que o teste q é útil, essa é a sua finalidade[23].
4. Post Hoc de Dunn-Sidák: esse teste é uma variação do procedimento de Bonferroni e também controla muito bem o erro tipo I quando o número de comparações excede cinco[22,24]. Uma comparação de poder entre Sidák e Bonferroni demonstra uma pequena diferença a favor do procedimento de Sidák. No entanto, recomenda-se o uso de Bonferroni pelo seu maior uso[23].
5. Post Hoc de Scheffé: em caso de comparações com outros procedimentos de comparações múltiplas, o mesmo apresenta uma maior proteção contra o erro tipo I e essa superproteção torna-se drástica quando o número de grupos excede a cinco. Não parece ter grandes vantagens quando comparado com Bonferroni ou Dunn-Sidák[23].
6. Ryan, Einot, Gabriel e Welsh Q (REGWQ): sugere-se o uso quando os grupos apresentam dimensões semelhantes, variâncias populacionais iguais e quando se quer comparar todos os pares de médias. É um teste que apresenta muito poder e controle sobre o erro tipo I[20].
7. Tukey: apresenta a mesma aplicabilidade que o teste anteriormente citado (REGWQ). É um teste conservador e quando se quer comparar muitos pares de médias. O teste de Tukey é sensível para detectar diferenças entre os grupos[23].
8. Studentized Newman-Keuls (SNK): é um teste aplicado quando o número de participantes em cada grupo é diferente. Compara pares de médias de forma Stepwise. Separa as médias dos grupos da maior para a menor e as maiores diferenças são primeiramente testadas. Não é um teste conservador e não controla muito bem o erro tipo I. Por essa razão, o seu uso não é recomendado[22].
9. Hochberg GT2: sugere-se o seu uso quando os números de participantes entre os grupos forem diferentes, mas com variâncias populacionais iguais. O teste de Gabriel é mais conservador, mas pode não controlar o erro tipo I quando os números de participantes entre os grupos forem muito diferentes.
10. Dunnett's: é usado para comparar um grupo (normalmente o grupo controle) com os outros grupos. Tem muito poder quando realizado para esse propósito[23].
11. Games-Howell: sugere-se o seu uso quando os números de participantes entre os grupos são diferentes e a variância também. Para amostras pequenas o teste pode não ser conservador.
12. Tamhane's T2, Dunnett's T3 e Dunnett's C: sugere-se o seu uso quando as variâncias dos grupos não são iguais. Os testes são conservadores e apresentam bom controle para o erro tipo I.

Considerações Finais

Dos procedimentos anteriormente citados. Cinco promovem excelente proteção contra o erro tipo I. Dentre eles o Bonferroni, Dunn-Sidak, Sheffé, Tukey e procedimento de Dunnet. Em relação às escolhas de procedimentos. O Post Hoc ideal será aquele que protege contra a inflação excessiva do erro tipo I e apresenta também conservadorismo (poder). Mas ao mesmo tempo o procedimento não deve diminuir a probabilidade de erro tipo I abaixo do nível pré-determinado pelo investigador ($p \leq 0,05$). Caso isso aconteça o procedimento é muito conservador e diferenças genuínas podem ser negligenciadas (erro tipo II).

Por exemplo, em um estudo, uma ANOVA de medidas repetidas pode ser realizada para avaliar se o TF pode diminuir a PA em um grupo de idosas. Se você achar que as diferenças proporcionadas pelo TF são importantes, um teste post-hoc mais liberal pode ser utilizado como o post hoc de Fisher's least significant difference ao invés do post hoc de Bonferroni[8].

Um exemplo real que podemos citar, é o estudo do pesquisador Carvalho e colaboradores25. Nesse estudo, após ser aplicada a ANOVA, o post hoc de Tukey foi utilizado. Considerando que o post hoc de Tukey é conservador, o erro do tipo II foi inflacionado e consequentemente diferenças genuínas dos efeitos das diferentes intensidades do TF sobre a pressão arterial podem ter sido negligenciadas.

Com isso, será que o resultado do estudo do pesquisador Carvalho e colaboradores[25] seria diferente se fosse utilizado um post hoc mais liberal como o LSD?

Nesse caso, seria melhor não perder um possível efeito do treino, ter evitado o erro do tipo II e não utilizar o post hoc de Tukey.

Certamente existe uma grande dificuldade de encontrar o equilíbrio entre os dois tipos de erros. No entanto, os procedimentos mais recomendados são os conservadores.

Uma mensagem importante é que a presente revisão apresenta sugestões e reflexões sobre os procedimentos, ao passo que, sugestões não devem ser confundidas com regras.

Referências

1. Weissgerber TL, Garovic VD, Milin-Lazovic JS, Winham SJ, Obradovic Z, Trzeciakowski JP, et al. Rein Bio Educ Bas Sci. PLoS Biol 2016;14(4):e1002430.
2. Ryan TA. Multiple comparisons in psychological research. Psy Bull 1959;56(1):26-47.
3. Carmer SG, Swanson MR. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. J Ame Stat Ass 1973;68(341):66-74.
4. da Cunha Nascimento D, Tibana RA, de Melo GF, Prestes J. Testes de normalidade em análises estatísticas: uma orientação para praticantes em ciências da saúde e atividade física. Rev Mack Educ Fís Esp 2017;14(2).
5. Fethney J. Statistical and clinical significance, and how to use confidence intervals to help interpret both. Australian critical care: official journal of the Confederation of Australian Cri Car Nur 2010;23(2):93-7.
6. Ottenbacher KJ. Statistical conclusion validity. Multiple inferences in rehabilitation research. Am J Phy Med Reha 1991;70(6):317-22.
7. Curtin F, Schulz P. Multiple correlations and Bonferroni's correction. Biol Psy 1998;44(8):775-7.
8. Armstrong RA. When to use the Bonferroni correction. Opht Phy Opt 2014;34(5):502-8.
9. Perneger TV. What's wrong with Bonferroni adjustments. Bmj 1998;316(7139):1236-8.
10. Rothman KJ. No adjustments are needed for multiple comparisons. Epidemiology. 1990:43-6.
11. Sampson JA, McAndrew D, Donohoe A, Jenkins A, Groeller H. The effect of a familiarisation period on subsequent strength gain. J Spo Scien 2013;31(2):204-11.
12. Andersen LL, Andersen JL, Magnusson SP, Aagaard P. Neuromuscular adaptations to detraining following resistance training in previously untrained subjects. Eur J App Phy 2005;93(5-6):511-8.
13. Kalinowski P, Fidler F. Interpreting significance: the differences between statistical significance, effect size, and practical importance. New Inf Nurs Rev 2010;10(1):50-4.

14. Jannig PR, Cardoso AC, Fleischmann E, Coelho CW, de Carvalho T. Influência da ordem de execução de exercícios resistidos na hipotensão pós-exercício em idosos hipertensos Influence of resistance exercises order performance on post-exercise hypotension in hypertensive elderly. Rev Bras Med Esp 2009;15(5):338-41.
15. Houle TT, Stump DA, editors. Statistical significance versus clinical significance. Seminars in Cardiothoracic and Vascular Anesthesia; 2008: SAGE Publications Sage CA: Los Angeles, CA.
16. Lawes CM, Bennett DA, Feigin VL, Rodgers A. Blood pressure and stroke: an overview of published reviews. Stroke 2004;35(3):776-85.
17. Mann TN, Lamberts RP, Lambert MI. High responders and low responders: factors associated with individual variation in response to standardized training. Spor Med 2014;44(8):1113-24.
18. Weir JP. Quantifying test-retest reliability using the intraclass correlation coefficient and the SEM. J Stren Cond Res 2005;19(1):231.
19. Ludbrook J. Statistics in biomedical laboratory and clinical science: applications, issues and pitfalls. Med Prin Prac. 2008;17(1):1-13.
20. Field A. Discovering statistics using IBM SPSS statistics: sage; 2013.
21. Baguley T. Understanding statistical power in the context of applied research. App Ergo 2004;35(2):73-80.
22. Ludbrook J. On making multiple comparisons in clinical and experimental pharmacology and physiology. Cli Exp Phar Phys 1991;18(6):379-92.
23. Howell DC. Statistical methods for psychology: Cengage Learning; 2012.
24. Šidák Z. Rectangular confidence regions for the means of multivariate normal distributions. J Amer Stat Asso 1967;62(318):626-33.
25. Canuto PMdBC, Nogueira IDB, da Cunha ES, Ferreira GMH, de Mendonça KMPP, da Costa FA, et al. Influência do treinamento resistido realizado em intensidades diferentes e mesmo volume de trabalho sobre a pressão arterial de idosas hipertensas. Rev Bras Med Esp 2011;17(4):246-9.