UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

UTILIZAÇÃO DA DIFERENÇA DE MÉDIAS PADRONIZADAS COMO MEDIDA DE EFEITO

DANIELA BENZANO BUMAGUIN

Porto Alegre
2016

DANIELA BENZANO BUMAGUIN

UTILIZAÇÃO DA DIFERENÇA DE MÉDIAS PADRONIZADAS COMO MEDIDA DE EFEITO

Trabalho de Conclusão de Curso submetido como requisito parcial para a obtenção do grau de Bacharel em Estatística

Orientadora: Patricia Klarmann Ziegelmann
Daniela Benzano Bumaguin

Porto Alegre
2016

Instituto de Matemática e Estatística
Departamento de Estatística

Utilização da diferença de médias padronizadas como medida de efeito

Daniela Benzano Bumaguin

Banca examinadora:

Professor Álvaro Vigo
UFRGS

Professora Patrícia Klarmann Ziegelmann
UFRGS

AGRADECIMENTOS

A minha mãe, cuja profunda e discreta presença orientaram meu caminho e o da minha filha Clara desde o dia em que nasceu;

A Clara, por me dar “boa aula mamãe!”assim que começou a falar;

Ao meu entranhável amigo Ivan, que me deu a sua mão e a sua luz;

Ao meu amigo Filipe, por tudo o que soube me ensinar com cuidado e criatividade;

A meu primeiro Professor e amigo Mário, que me incentivou sempre com carinho a seguir este caminho;

A minha Professora Patrícia, que me orientou com suavidade e competência;

A minha querida Professora Jandyra pelo seu exemplo e o seu carinho;

A Rodrigo, meu terapeuta, porque sua presença foi muito importante;

A todos os meus professores e as minhas amigas estatísticas (a Vânia, a Suzi, a Lu), pessoas felizes com esta profissão. Cada um deles foi muito importante no meu caminho.

A UFRGS, por ter me acolhido.

DEDICATÓRIA

Ao meu pai, e à falta que ele me faz.

EPÍGRAFE

"Sólo el que sabe es libre y más libre el que más sabe".

Miguel de Unamuno

RESUMO

Há mais de meio século que pesquisadores da área da saúde são estimulados a apresentar medidas de tamanho de efeito nas suas pesquisas. O tamanho de efeito é entendido como a medida ou força em que os fenômenos acontecem nas populações. Uma das comparações feitas frequentemente nos estudos é a comparação de duas médias, e esta pode ser realizada através da diferença bruta de médias ou da diferença padronizada de médias. Em 1962, com o objetivo de calcular o poder dos estudos publicados em 1960 no *Journal of Abnormal and Social Psychology*, Cohen constatou que os pesquisadores tinham grandes dificuldades para definir uma diferença clinicamente relevante para as suas variáveis, ou seja, muitos dificilmente conheciam o significado de uma diferença bruta de médias. E neste contexto o autor definiu o *d* de Cohen, a diferença bruta de médias dividida pela média ponderada dos desvios padrão nos dois grupos. Ele apresentou esta medida como uma medida interessante: já que não teria unidades da medida original, poderia ser utilizada para definir, através de pontos de corte arbitrários, o que seria um tamanho de efeito pequeno, moderado ou grande. Surgiu, então, no intuito de facilitar a discussão sobre o poder (que resultou na sua revisão, extremadamente baixo). O objetivo deste trabalho é discutir o uso do *d* de Cohen como medida de efeito e fazer o contraponto entre esta medida e a diferença bruta de médias. Na criação de diferentes cenários, é possível observar que diferenças brutas crescentes podem gerar valores de *d* de Cohen iguais. Assim,um *d* de Cohen classificado como efeito grande pode ser oriundo de diferenças brutas entre as médias pequenas. São vários os autores que convergem na idéia de que a diferença bruta de médias é uma medida que deve aparecer nas pesquisas, já que é calculada na unidade original da variável estudada, não depende da variabilidade dos dados e é simples de ser calculada. O uso do *d* de Cohen deve ser no contexto de comparação de desfechos medidos em diferentes escalas, e sempre é necessário contextualizar colocando a diferença bruta de médias na sua medida original. Valores de *d* de Cohen pequenos podem significar resultados clinicamente relevantes e por isso é importante a apresentação da medida bruta. Velar o desconhecimento das variáveis estudadas sob uma unidade de desvio padrão parece não ser uma boa estratégia quando o assunto é medida de efeito. Nem parece ter sido a intenção de Cohen que a diferença bruta de médias fosse substituída pelo *d*. Este artigo sugere a existência de um mal-entendido entre o pontuado por Cohen e o utilizado na literatura atual. A conclusão é que a diferença bruta de médias, acompanhada do intervalo de confiança para medir a incerteza devida à amostragem é a melhor abordagem quando o objetivo é apresentar uma medida de efeito. O *d* de Cohen vem complementar permitindo que o efeito observado no estudo possa ser comparado entre diferentes desfechos ou entre o mesmo desfecho medido em diferentes escalas.

Palavras chave: Tamanho de efeito. Diferença bruta de médias. *d* de Cohen.

ABSTRACT

For over half a century, health researchers have been encouraged to present effect size measures in their research. Effect size is understood as the measure or force in which phenomena occur in populations. One of the comparisons made frequently in the studies is the comparison of two means, and this can be done through the raw difference of means or the standardized difference of means. In 1962, with the purpose of calculating the power of studies published in 1960 in the Journal of Abnormal and Social Psychology, Cohen found that the researchers had great difficulties in defining a clinically relevant difference for their variables, that is, many hardly understood the meaning of a raw mean differences. And in this context the author defined or d of Cohen, the raw raw mean difference divided by the pooled standard deviations in the two groups. He presented this measure as an interesting measure: since it would not have units of the original measure, it could be used to define, through arbitrary cut-offs, what would be a small, moderate, or large effect size. It arose, then, in order to facilitate the discussion about power (which resulted in its extremely low revision). The aim of this work is to discuss the use of Cohen's d as an effect measure and to make the comparison between this measure and the raw mean differences. In creating different scenarios, it is possible to observe that increasing raw differences can generate equal values of Cohen d. Thus, a Cohen d classified as a large effect may be derived from small raw mean differences. There are several authors who converge on the idea that the raw difference between means is a measure that must appear in the researches, since it is calculated in the original unit of the studied variable, does not depend on the variability of the data and is simple to be calculated. The use of Cohen's d must be in the context of comparing outcomes measured at different scales, and it is always necessary to contextualize by placing the gross difference of means in their original measure. Small Cohen's d values may mean clinically relevant results and so it is important to present the raw measure. Ensuring ignorance of the studied variables under a unit of standard deviation does not seem to be a good strategy when the subject is measure of effect. Nor does it seem to have been Cohen's intention that the raw mean difference be replaced by d. This article suggests the existence of a misunderstanding between the one punctuated by Cohen and that used in the current literature. The conclusion is that the raw mean difference with the confidence interval to measure the uncertainty due to sampling is the best approach when the objective is to present an effect measure. Cohen's d is complementary, allowing the effect observed in the study to be compared between different outcomes or between the same outcome measured at different scales.

Keywords: Effect size, Raw mean differences, Cohen´s d.

SUMÁRIO

1 INTRODUÇÃO

As ciências da saúde dedicam seus esforços ao estudo das características dos indivíduos que, por natureza, são variáveis. Inferências para populações são realizadas através da observação de amostras e o uso da significância estatística para a sua avaliação recebe críticas na literatura há pelo menos meio século (COHEN, 1962). Em estatística a palavra significância é utilizada para definir a probabilidade de que um efeito observado não seja devido a um possível acidente de amostragem e pode ser expressa por uma medida chamada de p-valor. Uma diferença estatisticamente significativa está relacionado com o tamanho da amostra, mas o tamanho da diferença entre os grupos mede o efeito. Tamanho de efeito é simplesmente uma maneira de quantificar o tamanho da diferença entre dois grupos (COE, 2002). O conceito de tamanho do efeito foi desenvolvido como resposta as críticas sobre o uso da significância estatística como medida da relevância clinica ou social e aparece com vários significados (CONBOY, 2003). Nakagawa (2007) cita algumas acepções. Uma delas seria qualquer estatística (risco relativo, diferença bruta de médias, diferença padronizada de médias, etc.) que expresse a magnitude da diferença entre, por exemplo, o efeito de duas intervenções ou o valor de uma característica (nível de depressão, por exemplo) entre duas populações. Outra refere-se ao valor calculado para esta estatística. E uma terceira, refere-se ao tamanho de efeito biológico, clínico ou social. Uma estatística que expresse uma magnitude de diferença, o seu valor calculado ou o efeito biológico/social parecem ser definições que fazem parte do mesmo conceito e devem ser entendidas como algo único. O importante a destacar é que as medidas de tamanho de efeito a serem utilizadas devem depender do tipo de estudo, dos desfechos em questão e o tipo de comparação desejadas.

Com frequência o interesse é a comparação de desfechos do tipo quantitativo entre diferentes grupos, o que pode ser resumido em uma comparação de médias. Para expressar esta comparação são utilizadas na literatura duas medidas de tamanho de efeito: a diferença bruta de médias e a diferença padronizada de médias.

A diferença bruta de médias é uma estatística que mede a diferença absoluta entre o valor da média em um grupo e o valor da média em outro. A diferença padronizada de médias é a diferença bruta dividida pela média ponderada dos desvios padrão dos dois grupos. Foi proposta por Jacob Cohen em 1962 e ficou então conhecida por *d* de Cohen. Surgiu como resposta à necessidade de quantificar a diferença clinicamente relevante entre médias de desfechos quantitativos que não são completamente conhecidos. Esta quantificação é

importante para o cálculo do poder dos estudos, que na visão de Cohen encontrava-se negligenciado. Frequentemente o valor de p era citado como sinónimo de magnitude de efeito (COHEN, 1990). E neste contexto, Cohen apresentou o *d* de Cohen como uma medida que definiria de forma mais clara e padronizada a magnitude clinicamente relevante de uma diferença (COHEN, 1992). A diferença padronizada de médias também iria possibilitar a comparação entre desfechos medidos através de escalas diferentes, e acumular conhecimento no contexto de metanálises.

As duas medidas de efeito atualmente utilizadas na literatura, a diferença bruta de médias e a diferença padronizada *d* de Cohen parecem ter propósitos diferentes quando apresentadas. Na literatura é frequente encontrar autores se posicionando a favor de uma ou outra (COE, 2002; BAGULEY,2009). Os argumentos a favor de cada uma delas são variados. A facilidade de sua compreensão, o fato de ser expressa na unidade original do estudo e de ser independente da variância, aparecem em destaque como argumentos em favor da diferença bruta de médias. Aqueles que defendem a padronizada sugerem que esta seria melhor para a compreensão do tamanho das diferenças de médias de desfechos medidos com escalas desconhecidas. Frequentemente na literatura quando as variáveis são medidas através de escalas, os autores descrevem o valor do *d* de Cohen como medida de tamanho de efeito e ignoram a a diferença bruta de médias. E neste contexto polêmico, aparece uma mistura nas informações escolhidas para descrever as comparações, optando muito frequentemente pela apresentação de médias e desvios padrão acompanhados pelo *d* de Cohen como medida de tamanho de efeito.

O objetivo deste estudo é discutir a utilização do *d* de Cohen como medida de efeito e compará-la com a diferença bruta de médias como medida para a descrição do tamanho do efeito clínico, biológico ou social.

2 EXEMPLO

Para ilustrar as discussões sobre a comparação entre a diferença bruta de médias e diferença padronizada de médias *d* de Cohen, será utilizado um exemplo envolvendo dados hipotéticos do escore "*Beck Depression Inventory*" (BDI-II). O BDI-II é um instrumento clínico de 21 itens, amplamente utilizado e validado que mede aspectos cognitivos, afetivos e fatores fisiológicos para avaliar a gravidade da depressão. O escore tem um mínimo de 0 e um máximo de 63 pontos, sendo até 13 pontos depressão mínima, de 14 a 19 depressão leve, de 20 a 28 moderada e acima de 29 depressão grave (BECK, 1996).

O exemplo hipotético avalia o efeito da institucionalização em asilos para idosos no nível de depressão. O referencial teórico para este assunto aponta que o estresse relacionado à institucionalização parece ser um dos principais fatores ambientais associado ao aumento do nível de depressão em idosos (SOARES, 2012). A hipótese de pesquisa é que a institucionalização está associada com o aumento dos níveis de depressão ou seja, que o escore BDI-II dos idosos institucionalizados seja maior. A amostra é composta de sessenta idosos sendo metade deles institucionalizados. Os valores do escore BDI-II foram calculadas para cada idoso e são apresentadas na Tabela 1.

Tabela 1.-Tabela comparativa do BDI-II de pacientes institucionalizados e não institucionalizados.

Escore	Institucionalizados n=30	Não institucionalizados n=30
BDI-II	15,43±2,90	8,32±3,10

Dados apresentados pela média±desvio padrão.

3 MEDIDAS DE TAMANHO DE EFEITO PARA DESFECHOS QUANTITATIVOS

A medida de tendência central mais indicada para dados quantitativos que apresentam comportamento normal é a média aritmética simples. Assim, uma medida de efeito direta e intuitiva para comparação de dois grupos é a diferença bruta de médias. Outra medida citada na literatura é a diferença padronizada de médias. Estas duas medidas serão definidas nesta sessão com exemplificação e interpretações no contexto do exemplo sobre os idosos e a sua condição de institucionalização.

3.1 Diferença bruta de médias

A definição da diferença bruta de médias é direta: considere que $\bar{x}_1$ é a média obtida em uma das amostras e $\bar{x}_2$ a média obtida na outra amostra. Assim, a diferença bruta de médias é dada por:

$$\text{dif} = \bar{x}_1 - \bar{x}_2$$

A dif é uma estimativa pontual para a diferença bruta de médias das duas populações utilizando as amostras de tamanho $n_1 \, e \, n_2$. Para construir um intervalo de confiança é necessário que seja cumprida a suposição de normalidade dos dados. Ainda, a construção do intervalo será diferente na presença de amostras independentes ou amostras pareadas. Ainda, no caso de amostras independentes, a construção do intervalo irá depender de se é possível supor variâncias populacionais iguais.

Amostras independentes com suposição de variâncias iguais: os limites inferior e superior do intervalo de confiança são apresentados nas expressões abaixo.

$$LI = (\bar{x}_1 - \bar{x}_2) - t_{(gl;\alpha/2)} \sqrt{s_0^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$LS = (\bar{x}_1 - \bar{x}_2) + t_{(gl;\alpha/2)}\sqrt{s_0^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)};$$

sendo s_1 e s_2 os desvios padrão , $s_0^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}$ e $gl = (n_1 + n2 - 2)$.

Amostras independentes sem suposição de variâncias iguais: os limites inferior e superior do intervalo de confiança são apresentados nas expressões abaixo.

$$LI = (\bar{x}_1 - \bar{x}_2) - t_{(gl;\alpha/2)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$LS = (\bar{x}_1 - \bar{x}_2) + t_{(gl;\alpha/2)}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}};$$

onde

$$gl = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{(n_1-1)} + \frac{(s_2^2/n_2)^2}{(n_2-1)}}$$

Amostras pareadas: em caso de amostras pareadas, a diferença intrapar para as variáveis x_1 e x_2 é dada por:

$d_i = x_{1i} - x_{2i}$ para cada par i de n observações, cuja média será $\overline{\boldsymbol{x}} = \frac{\sum_{i=1}^{n} di}{n} = \frac{\sum_{i=1}^{n} \boldsymbol{xi_1}}{n} - \frac{\sum_{i=1}^{n} \boldsymbol{xi_2}}{n} = \overline{\boldsymbol{x_1}} - \overline{\boldsymbol{x_2}}$. Os limites inferior e superior do intervalo de confiança para a média da diferença intrapar $\bar{x}$ são dados nas expressões abaixo.

$$LI = \bar{x} - t_{\left(gl;\frac{\alpha}{2}\right)}\frac{s}{\sqrt{n}}$$

$$LS = \bar{x} + t_{\left(gl;\frac{\alpha}{2}\right)}\frac{s}{\sqrt{n}};$$

sendo s o desvio padrão da diferença intrapar, n o número de pares de observações e $gl = (n-1)$.

Exemplo: o grupo institucionalizado obteve uma média de 15,43 pontos no BDI-II e o grupo não institucionalizado uma média de 8,32 pontos perfazendo uma diferença bruta de médias de 7,11 pontos. Após realizar o teste F concluímos não há diferença estatisticamente significativa entre as variâncias das duas populações os dois grupos (P=0,828). Assim, supondo normalidade dos dados, amostras independentes, variâncias populacionais iguais, o intervalo de 95% de confiança para a diferença bruta de médias será:

$$LI = 7{,}11 - 2{,}00\sqrt{\left(\frac{(30-1)2{,}90^2 + (30-1)3{,}10^2}{(30+30-2\,)}\right)^2\left(\frac{1}{30}+\frac{1}{30}\right)} = 5{,}56;$$

$$LS = 7{,}11 + 2{,}00\sqrt{\left(\frac{(30-1)2{,}90^2 + (30-1)3{,}10^2}{(30+30-2\,)}\right)^2\left(\frac{1}{30}+\frac{1}{30}\right)} = 8{,}66$$

Estima-se que a diferença entre a média do BDI-II de indivíduos institucionalizados e não institucionalizados seja de 7,11 pontos (IC95%:5,56-8,66). Esta diferença é estatisticamente significativa (P<0,001) indicando que indivíduos institucionalizados tem maior depressão. Para avaliar se esta estimativa é clinicamente relevante é necessário definir se uma diferença de 5,56 ou 8,86 pontos na escala BDI-II é relevante.

3.2 Diferença padronizada de médias

Cohen (1977) propôs uma padronização da diferença bruta de médias. Esta padronização ficou conhecida por diferença padronizada de médias *d* de Cohen ou, simplesmente, diferença padronizada. Existem autores como por exemplo Coe que utilizam a expressão "*effect size*" para se referir à diferença padronizada de médias (COE, 2002). Para o cálculo do d de Cohen é necessário identificar se médias são oriundas de amostras independentes ou pareadas.

Amostras Independentes: o *d* de Cohen é obtido pela expressão:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}},$$

sendo n_1 e n_2 os tamanhos das amostras e s_1 e s_2 os desvios padrão.

Os limites inferior e superior do intervalo de 95% de confiança para o *d* de Cohen são dados pelas expressões seguintes:

$$LI = d - t_{(gl;\alpha/2)}.EP$$

$$LS = d + t_{(gl;\alpha/2)}.EP;$$

sendo o erro padrão: $\text{EP} = \sqrt{\frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}}$, e $gl = (n_1 + n2 - 2)$.

Amostras pareadas: Nas situações nas quais o objetivo é comparar grupos pareados ou observações antes e depois dentro do grupo as expressões para obter o d de Cohen é a seguinte:

$$\text{d} = \frac{\bar{y}_{\text{diff}}}{\frac{s_{\text{diff}}}{\sqrt{2\,(1-r)}}} = \frac{\bar{y}_1 - \bar{y}_2}{\frac{s_{\text{diff}}}{\sqrt{2\,(1-r)}}},$$

considerando s_{diff} o desvio padrão da diferença e r o coeficiente de correlação entre os pares de observações.

Os limites inferior e superior do intervalo de 95% de confiança para o *d* de Cohen são dados pelas expressões seguintes:

$$LI = d - t_{\left(gl;\frac{\alpha}{2}\right)}.EP$$

$$LS = d + t_{\left(gl;\frac{\alpha}{2}\right)}.EP;$$

considerando n o número de pares de observações, $gl = (n - 1)$, e o erro padrão:

$$EP = \sqrt{\left(\frac{1}{n} + \frac{d^2}{2n}\right) 2\,(1 - r)}$$

Correção de Hedges: Há casos em que as amostras são pequenas e geralmente acontece um viés que pode ser corrigido, resultando no g proposto por Hedges em 1981, e que é dado pela expressão seguinte:

$$g = d.J,$$

sendo $J = \left(1 - \frac{3}{4\text{gl}-1}\right)$ um fator de correção e $gl = (n_1 + n2 - 2)$.

Os limites inferior e superior do intervalo de 95% de confiança para o g de Hedges são dados pelas expressões seguintes:

$$LI = g - t_{\left(gl;\frac{\alpha}{2}\right)}.EP$$

$$LS = d + t_{\left(gl;\frac{\alpha}{2}\right)}.EP$$

sendo $\text{EP} = \sqrt{J^2.\left(\frac{\text{n}_1+\text{n}_2}{\text{n}_1\text{n}_2} + \frac{\text{d}^2}{2(\text{n}_1+\text{n}_2)}\right)}$ e $gl = (n_1 + n2 - 2)$.

Na escala proposta por Cohen um d de Cohen de 0,2 a 0,5 é considerado pequeno, um de 0,5 a 0,8 será moderado e um acima de 0,8 será grande.

Exemplo: Considerando amostras independentes, a diferença padronizada de médias é dada por:

$$d = \frac{15{,}43 - 8{,}32}{\sqrt{\frac{(30-1)2{,}90^2 + (30-1)3{,}10^2}{(30+30-2)}}} = 2{,}37$$

Os limites do intervalo de 95% de confiança para o *d* de Cohen é o seguinte:

$$LI = 2{,}37 - 2.\sqrt{\frac{30+30}{30.30} + \frac{2{,}37^2}{2(30+30)}} = 1{,}69$$

$$LS = 2{,}37 + 2.\sqrt{\frac{30+30}{30.30} + \frac{2{,}37^2}{2(30+30)}} = 3{,}05$$

Estima-se que a diferença padronizada de médias comparando idosos institucionalizados e não institucionalizados seja de 2,37 (IC95%:1,69-3,05). Ou seja, uma diferença classificada como grande segundo os pontos de corte na escala proposta por Cohen.

3.3 O surgimento do *d* de Cohen como medida de efeito

A primeira referência da diferença padronizada de médias *d* de Cohen data de 1962 no artigo "*The statistical power of abnormal-social psychological research: A review*" (COHEN, 1962). Neste trabalho, Jacob Cohen, estatístico e psicólogo, publicou a primeira revisão sistemática com metanálise de poder dos estudos (SEDLMEIER e GINGERNEZER, 1989). Nesta revisão, ele calculou o poder dos artigos publicados no volume de 1960 da revista "Journal of Abnormal and Social Psychology". Para o cálculo do poder dos estudos precisou definir qual seria uma diferença que fosse clinicamente relevante para os desfechos estudados em cada estudo. E assim surge o *d* de Cohen, como medida para quantificar uma diferença que os pesquisadores não conseguiam definir. Surge para suprir a necessidade de se saber qual diferença era relevante quando os pesquisadores não sabiam definir nas suas métricas (não sabiam, por exemplo, quantos pontos na escala BDI-II são necessários para dizer que o nível médio de depressão em um grupo é clinicamente maior que em outro grupo). Foi constatado que os pesquisadores utilizavam o p-valor do teste estatístico para avaliar a importância dos achados do estudo, ou seja, quanto mais significativo era o resultado do teste estatístico, mais importância adquiria o resultado do estudo na literatura. Desconsideravam, completamente, a importância clínica do achado. Preocupado com esta situação e com o objetivo de calcular o poder dos estudos Cohen propôs uma padronização na diferença bruta de médias. Sua ideia era que a diferença padronizada de médias ajudaria os pesquisadores a definir o quanto seria uma diferença clinicamente relevante e também serviria para comparar resultados de estudos que utilizassem diferentes métricas para medir o desfecho (por exemplo, diferentes escalas para medir depressão). Foi neste contexto que Cohen desenvolveu o seu livro "*Statistical Power Analysis for the Behavioral Sciences*" publicado em 1977 e reeditado em 1988. Neste, Cohen teve o propósito de apresentar um tratamento abrangente da análise do poder dos estudos de um ponto de vista aplicado. Logo no primeiro capítulo, ele define o tamanho do efeito como o grau em que o fenômeno é presente na população e o descreve como um parâmetro fundamental dentre três que influenciam o poder do estudo. Os outros dois parâmetros são o nível de significância e a precisão do resultado da amostra. Segundo Cohen o poder via-se negligenciado e o valor de p supervalorizado e nos estudos um p-valor pequeno era usado como sinônimo de

resultado importante do ponto de vista da sua magnitude. Ele detalha que quando é realizado um teste de hipótese haverá a possibilidade de a hipótese nula ser verdadeira na população (ausência de efeito) ou ser falsa (presença de efeito). Quando são comparadas duas populações a hipótese nula expressa que a diferença em relação ao parâmetro entre estas populações é igual a zero e a hipótese alternativa expressa que esta diferença é diferente de zero. O quanto que essa diferença se afasta do valor zero expressa o tamanho de efeito. A relação entre tamanho de efeito e poder do estudo torna-se evidente. Para um mesmo nível de significância, e mesmo tamanho de amostra, quanto maior é o tamanho de efeito maior será o poder do teste. Uma situação similar acontece para o tamanho de amostra necessário. Considerando uma mesma significância e poder desejados se existe um tamanho de efeito grande será necessário um tamanho de amostra menor para detectar a diferença. Várias situações de comparação existem como por exemplo a comparação entre proporções, comparação de médias e cada teste terá a sua medida apropriada de tamanho de efeito.

O Cohen observou que o pesquisador da área do comportamento apresentava dificuldades para determinar o que seria uma diferença pequena, média ou grande após rejeitar a hipótese nula (COHEN,1988). Na sua visão, a teoria deveria responder qual valor seria considerado um efeito nulo e qual um efeito não nulo. Mas nesse contexto de desconhecimento e dificuldade ele irá propor a sua medida de efeito na comparação de médias, o *d* de Cohen, assim como uma convenção para a interpretação da sua magnitude, alertando para a natureza arbitraria dos pontos de corte (COHEN, 1988). A falta de rigor que revisores e editores de jornais tinham em relação à avaliação do poder dos estudos era para Cohen uma preocupação. Segundo ele, esta carência era em parte devida ao baixo nível de consciência que os autores tinham sobre a magnitude do seu efeito (COHEN, 1992). E por isto que uma medida como o *d* ajudaria a definir esta magnitude.

Pontos de corte arbitrários em estatística não seriam novidades, uma vez que geralmente o nível de significância é fixado em 1 ou 5% e o poder do estudo em 80 ou 90%. Então no contexto do cálculo do poder, e do cálculo do tamanho da amostra parece que convencionar o quanto seria uma diferença pequena, média e grande seria uma estratégia conhecida e aceita. O tamanho de efeito seria uma forma de caracterizar a magnitude absoluta da diferença, e o significado do mesmo dependerá do contexto no qual se assenta. Isto é necessário porque na área do comportamento por exemplo, dificilmente se trabalha com unidades de medida mais conhecidas como reais, quilogramas ou meses.

A convenção proposta por Cohen sobre valores de diferença padronizada de médias *d* considerados “pequenos”, “médios” e “grandes” surge como forma de quantificar ao

pesquisador a magnitude da sua diferença, havida conta da dificuldade existente. Médio seria um efeito detectável a olho nu por um observador cuidadoso, pequeno seria visivelmente menor mas não trivial, e grande seria a mesma distância entre um efeito pequeno e médio, mas no sentido inverso. Os pontos de corte para as três categorias de *d* de Cohen seriam *d* igual a 0,20, 0,50 e 0,80. Para Cohen *"for the test that two population means are equal, the effects sizes (ES) in the same order are d=0,20, 0,50 e 0,80. The 0,20 ES is exemplified by the mean IQ difference between twins and non twins (the later being larger), the0,50 ES by the IQ difference between clerical and semi skilled workers, and the 0,80 effect size by the mean IQ diference between PhDS and college freshmen"*(Cohen, 1992). A diferença encontrada no exemplo, cujo *d* de Cohen calculado foi 2,37 seria classificada segundo esta escala como uma diferença grande. Porém, reiterando, o mesmo Cohen alerta ao caráter arbitrário da escala e autores como Lenth (2001) sugerem evitar as "medidas enlatadas".

4 COMPARAÇÃO DAS MEDIDAS DE EFEITO

O objetivo desta sessão é comparar a diferença bruta de médias com a diferença padronizada de médias. Esta comparação será primeiramente realizada através de um exemplo numérico. Acredita-se que o entendimento deste exemplo facilita ao leitor o acompanhamento das discussões geradas por diferentes autores e apresentadas na Seção 4.2.

4.1 Exemplo numérico

Serão descritos diferentes cenários e calculadas as diferenças brutas de médias de BDI-II entre indivíduos institucionalizados (em diferentes instituições) e indivíduos não institucionalizados. Será calculado também o *d* de Cohen para cada uma das situações. Estes cenários levam a resultados de *d* de Cohen iguais, porém com diferenças brutas distintas.

A Figura 1 apresenta um gráfico com as estimativas das diferenças brutas e padronizadas derivadas de três diferentes cenários de pontuação no BDI-II. Todos os resultados são acompanhados dos seus respectivos intervalos de 95% de confiança. Os dados utilizados neste exemplo são fictícios. Para embasar a discussão pense, por exemplo, que o cenário A envolve uma instituição referenciada como de alta qualidade (com uma boa infraestrutura e uma equipe de profissionais de saúde mental e física disponíveis 24 horas para a atenção dos seus internados). O cenário B envolveria uma instituição de qualidade média (com uma boa infraestrutura mas sem equipe disponível 24 horas) e o cenário C uma instituição de baixa qualidade (com escassez de infraestrutura e de equipe disponível). Observe que a diferença padronizada de médias estimada é a mesma para os três cenários e o valor resultante igual a 2,37 é classificado como de magnitude forte, segundo Cohen. No cenário A, a diferença bruta foi de 1,19 pontos, no cenário B, foi de 4,74 e no cenário C, foi de 7,11 pontos. Os desvios padrão foram aumentando em seu valor indicando que a variabilidade observada é maior para os resultados observados na instituição de baixa qualidade (cenário C). Uma diferença entre as médias de 1,19 pontos representa uma diferença de 1,89% da amplitude total da escala BDI-II enquanto que uma diferença de 7,11 pontos representa uma diferença de 11,29%. Isto faz pensar que, do ponto de vista clínico, os três tipos de instituição geram um efeito diferenciado na depressão dos idosos. Tendo em conta que a diferença bruta entre idosos em instituição de baixa qualidade comparados com os não institucionalizados é de 7,11 pontos será possível concluir que este é o pior cenário, e os pacientes que estão neste tipo de Instituição sofrem uma

depressão mais grave que a sofrida pelos idosos que não estão institucionalizados. Esta instituição (baixa qualidade) é aquela que terá maior impacto nos níveis de depressão dos seus internados se a diferença bruta de médias for descrita. Assim, a instituição de qualidade média (cenário B) seria a segunda pior e a instituição de alta qualidade (cenário A) seria a melhor de todas, em relação as pontuações de BDI-II dos seus institucionalizados. Porém, a diferença padronizada de médias *d* de Cohen mantém-se a mesma, e em todos os cenários será descrita como grande.

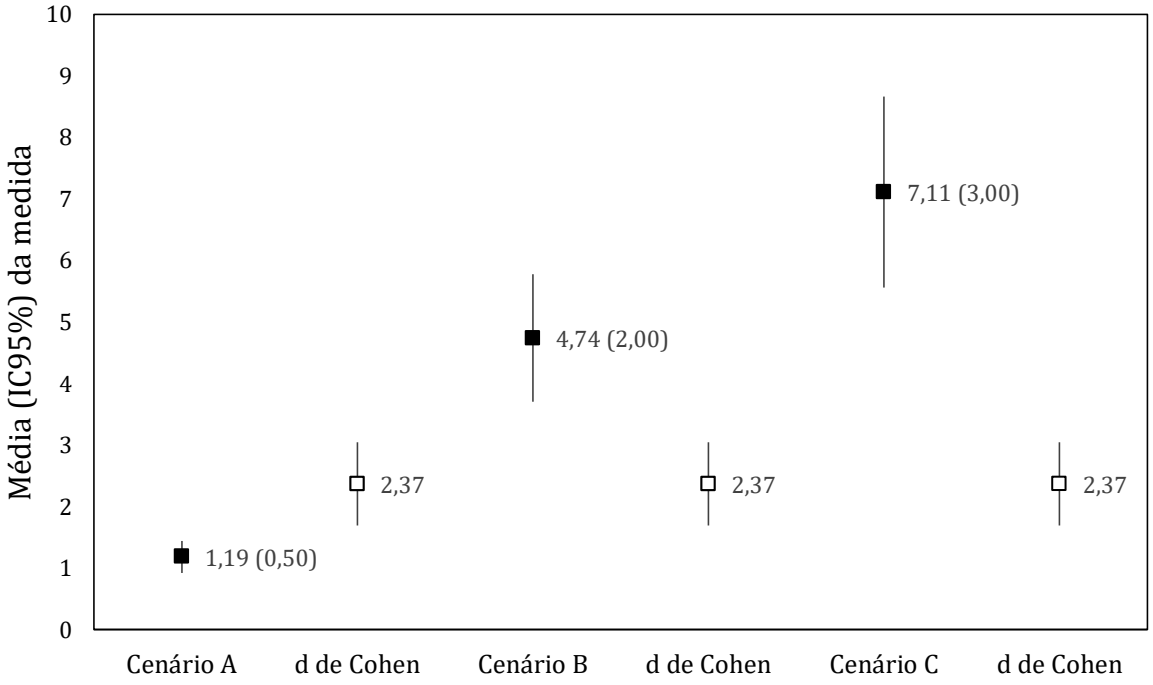

Figura 1. Médias brutas (desvio padrão) e padronizada do BDI-II comparando idosos institucionalizados em diferentes tipos de Instituição e idosos não institucionalizados.

A análise dos resultados apresentados no Figura 1 sugerem que a compreensão sobre o efeito da internação em uma ou outra instituição será diferente, quando se utiliza o *d* de Cohen como medida de efeito e quando se utiliza a diferença bruta de médias. Apresentando o *d* de Cohen será compreendido que qualquer que seja a instituição, a institucionalização tem um efeito grande na depressão dos idosos. Ao ser apresentado unicamente o *d* de Cohen pareceria que é igual, do ponto de vista do impacto na depressão, a institucionalização em uma clínica de alta, média ou baixa qualidade. Quando a diferença bruta de médias é apresentada, parece que estar internado em uma Instituição de alta qualidade não é tão diferente de estar em casa. Já estar internado em uma Instituição de qualidade média teria um impacto um pouco maior, e

ainda maior parece ser o impacto no BDI-II o fato de estar internado em uma Instituição de baixa qualidade.

4.2 Discussão sobre a medida de efeito na visão de outros autores

Vários autores discutem o uso da diferença padronizada de médias como medida de efeito clínico e concordam em que a diferença bruta de médias não pode estar ausente. Alguns autores se posicionam em relação ao uso de uma ou outra medida de efeito.

A *American Psychological association* (APA) na quarta edição do seu *publication manual* (1994) cita pela primeira vez a necessidade de acrescentar nos achados o poder do estudo e a magnitude dos efeitos. Nesta edição ela encorajou os autores a reportarem medidas de tamanho de efeito mas o atendimento da sugestão ocorreu lentamente. Wilkinson destaca a importância desta apresentação, e destaca principalmente a sua importância para o cálculo do poder dos estudos e possível participação dos mesmos em futuras metanálises (WILKINSON, 1999). Mais tarde, na sua quinta edição (2001) o *APA publication manual* ainda recomendou o uso de medidas de magnitude de diferença na escala original sempre que possível, já que seria uma maneira muito fácil de julgar o efeito e de comparar os estudos, se todos eles usarem a mesma medida de desfecho. Não entanto, foi destacada a padronização como necessária em caso de haver diferentes medidas, a fim de compará-las. Porém colocou, e é importante destacar, que não existe uma relação direta entre magnitude de um efeito e o seu valor clínico. Dependendo das circunstâncias um efeito de menor magnitude em um resultado pode ser mais importante que um efeito de maior magnitude em outro contexto. Uma magnitude de efeito pode significar situações diferentes conforme a pesquisa que o produziu e os resultados de estudos anteriores semelhantes. Incentivou nesta edição ao uso de intervalos de confiança.

Coe (2002) faz um paralelismo entre o *d* de Cohen e o escore z onde por exemplo a interpretação de um valor de 0,8 significaria que uma pessoa média do grupo experimental está 0,8 desvios padrões acima de uma pessoa média do grupo controle mas isto dificulta a comparação entre os estudos. Esta interpretação parece ser fácil para estatísticos porém pode gerar algumas dificuldades para os pesquisadores da área da saúde. No intuito de fazer não-estatísticos entenderem o conceito de tamanho de efeito, os autores McGraw e Wong (1992) definiram o conceito de "*Common Language Effect Size*" (CLES) que seria a probabilidade de uma pessoa do grupo experimental ter um valor mais alto na variável sendo estudada quando comparada com uma pessoa do grupo controle, se ambos fossem escolhidos aleatoriamente (COE, 2002). Os autores exemplificam que em caso de uma diferença entre as médias de altura

cujo *d* de Cohen fosse 2, então o CLES valeria 0,92 (cálculo feito a partir de sobreposições das curvas normais dos dois grupos) o que significaria que em 100 encontros entre homens e mulheres, em 92% dos casos o homem seria mais alto que a mulher. Parece que esta medida pode aproximar uma visão clínica ao uso do *d* de Cohen, que requer um raciocínio mais estatístico para a sua compreensão. Neste contexto de medidas mais ou menos compreensíveis do ponto de vista clínico, Coe sugere o uso da diferença bruta de médias com o intervalo de confiança quando o desfecho é medido com uma escala familiar, a amostra tem pouca variabilidade, e os grupos tem diferentes desvios padrões (COE, 2002).

Durlak cita três pontos importantes a destacar no contexto da medida de efeito. Primeiro é de suma importância considerar a qualidade da pesquisa (nova e prévia) que gerou este efeito. Segundo, é necessário fazer comparações em condições de pesquisa semelhantes em relação à medida de resultado. Terceiro, é fundamental considerar o significado clínico ou prático dos achados (DURLAK, 2009).

Baguley (2009) cita "*The robust beauty of simple effect size*" e sintetiza a existência de três vantagens ao serem comparadas a diferença bruta de médias com a diferença padronizada de médias como o *d* de Cohen. Em primeiro lugar, a diferença bruta de médias independe da variância dos dados, o que evitaria problemas decorrentes da relação entre o *d* de Cohen e a variância. Estes problemas são a potencial falta de confiabilidade entre os instrumentos de medida para medir variáveis iguais nos diferentes grupos, a seleção da amostra que irá se traduzir na amplitude nos dados observados e a influência do desenho do estudo, que poderá ter diferentes variabilidades conforme sejam utilizadas amostras independentes ou pareadas. Por estes motivos a medida bruta será mais robusta que a medida padronizada. Outra vantagem da diferença bruta de médias é que esta diferença é expressa em unidades da variável em estudo o que quase sempre é mais compreensível que a diferença padronizada de médias. Parece ainda que se a medida não for muito conhecida a padronização do efeito poderia inclusive velar o desconhecimento sobre a mesma, o que não seria desejável. E em terceiro lugar Baguley cita uma vantagem de ordem prática, que é a simplicidade do cálculo das medidas de efeito brutas, que além de facilitar as contas, evita possíveis erros.

Em 2010 a APA *publication manual* na sua sexta edição volta a orientar aos autores a incluir alguma medida de tamanho de efeito para que o leitor possa apreciar a magnitude ou a importância das descobertas de um estudo. Cita que os tamanhos de efeito podem ser expressos nas unidades originais, que são muitas vezes mais facilmente compreendidos. E indica que pode ser útil relatar um tamanho de efeito também em algumas unidades padronizadas ou sem

unidades (por exemplo, o *d* de Cohen). Sempre que possível, deveria o autor fornecer um intervalo de confiança para cada tamanho de efeito relatado, para indicar a precisão da estimativa (FIDLER, 2010).

Lipsey cita que no intuito de representar a diferença entre duas médias os pesquisadores utilizam frequentemente o *d* de Cohen. Ele discute que esta forma de medir o do tamanho do efeito tem muito menor significado inerente do que se fosse citada a diferença bruta de médias. Para Lipsey esta forma padronizada é enganosa para avaliar tamanhos de efeito em relação a sua importância prática. O maior problema seria a existência de pontos de corte na escala do *d* de Cohen, que divide o efeito em pequeno, médio e grande. Seria como caracterizar a altura de uma criança como pequena, média ou grande, em relação a todos os mamíferos vertebrados em lugar de em referência à distribuição de valores para crianças de idade e sexo semelhantes. Utilizar estes pontos de corte para intervenções por exemplo de educação, onde pequenos efeitos são importantes, seria inadequado. Além do mais o autor considera que a padronização não respeitaria o construto do desfecho, a forma de ser medido e a forma de ser analisado (LIPSEY, 2012).

Na visão de Lakens o principal problema do *d* de Cohen aparece vinculado aos pontos de corte arbitrários propostos por ele. Este autor destaca que as vezes pequenos tamanhos de efeito estatístico podem ter grandes consequências. A única razão para o uso isolado desta medida seria no contexto de inexistência de achados parecidos para realizar comparações. A melhor forma de interpretar o *d* de Cohen seria, para Lakens relacionar o mesmo com outros efeitos existentes na literatura e, se possível, acrescentar uma explicação das consequências práticas do efeito. Infelizmente, não há uma recomendação clara de como fazê-lo (LAKENS, 2013).

Em 2013 Higgins e Green (editores da Cochrane) apontaram que a diferença de médias padronizada devia ser usada como uma estatística de resumo na metanálise quando os estudos incluídos medissem o mesmo desfecho, porem com diferentes instrumentos. Neste caso, uma medida padronizada ajudaria a poder comparar e combinar os resultados oriundos de estudos diferentes, portanto seria necessário padronizar os resultados dos estudos a uma escala uniforme para este fim.

5 CONSIDERAÇÕES FINAIS

Há mais de meio século Cohen propunha uma diferença de medias padronizada, que chamou *d* de Cohen, no contexto de um estudo sistemático sobre o poder das pesquisas na área da psicologia. No intuito de quantificar a diferença entre médias para poder calcular o poder dos estudos Cohen propôs a média padronizada *d* de Cohen. Era o objetivo desta medida definir a diferença entre a hipótese nula e alternativa para a variável em estudo, já que existia falta de conhecimentos dos autores sobre o que era uma diferença pequena, moderada e grande. Ele estabeleceu pontos de corte para esta diferença media padronizada, alertando para a sua natureza arbitrária. Sabe-se que além de ajudar pesquisadores que tinham dúvidas em relação as suas medidas, a padronização proposta ajudaria no caso de desfechos medidos com escalas diferentes, para poder comparar. Esta medida padronizada, que seria livre de unidades ao seu entender, teria realmente como unidade a média ponderada dos desvios padrão dos dois grupos, desvio que depende de, entre outras considerações, a forma de coleta dos resultados e a amostra escolhida. Portanto, parece que a ausência de unidades citada é uma ilusão, e esta dependência do *d* de Cohen da variabilidade das pesquisas só velaria o desconhecimento sobre a variável que está sendo trabalhada. Já em 1969 Tukey colocava ao se referir a outra medida de efeito, que não seria desejável que os pesquisadores estivessem tão desinteressados sobre as suas variáveis como para não se preocupar com as suas unidades, e que se basear nas medidas de efeito para a interpretação dos resultados seria "varrer a poeira para baixo do tapete".

Em vários cenários de comparação de médias, diferenças brutas muito variadas podem resultar em um mesmo *d* de Cohen que, seguindo os pontos de corte sugeridos pelo autor definirão a diferença como pequena média ou grande. Porém, ao olhar para a diferença bruta de médias, na escala original, estas diferenças podem ser mais ou menos distantes e a situação clínica ou social envolvida mais ou menos grave.

Vários autores concordam em que o *d* de Cohen é importante para comparar escalas diferentes, já que utilizaria a mesma unidade, e é importante no contexto de uma metanálise pelo mesmo motivo. Porém vemos frequentemente nas pesquisas o uso do *d* de Cohen como substituto da diferença bruta de médias, e sendo interpretado como medida da relevância clinica ou social.

Nas pesquisas da área da saúde é muito frequente a medida dos fenômenos em escalas que não são muito conhecidas, porem seria importante não medir esforços para ter um real conhecimento do que significa clinicamente uma quantidade de pontos de uma escala na qual

se trabalha e se está interessado. Dizer que não é de conhecimento do pesquisador o quanto são 7,11 pontos na escala BDI-II, se este está estudando a depressão e escolheu esta escala já validada, se a mesma tem pontos de corte que definem inclusive eventos clínicos, parece um contrasenso. Se assim fosse seria questionável o fato dele realmente conhecer o que significa uma média de 15,43 pontos. Ainda mais difícil de acreditar seria que o pesquisador saiba que não é igual ter uma pontuação compatível com depressão leve (até 13 pontos no escore) e uma pontuação compatível com uma depressão grave (mais de 29 pontos no escore).

Utilizar o *d* de Cohen como uma medida da relevância clinica quando na sua origem foi uma medida proposta para fazer o cálculo do poder dos estudos, compará-los e incluí-los em metanálises, parece ter sido um mal-entendido. As discussões sobre o uso de uma ou outra medida por vários autores, inclusive neste artigo, denotam o quanto foi entendido que uma medida surgiu como substituta da outra. Mas tratava-se de outra proposta na origem, e parece que foi outra a intenção de Cohen ao apresentá-la. É possível se deduzir que não teria nascido esta como um potencial concorrente da diferença bruta de médias e muito menos seria a sua substituta.

Por ser uma medida que não depende da variabilidade dos resultados da pesquisa, ser simples de calcular e ser expressa na unidade da variável estudada, a diferença bruta de médias é uma medida de efeito que deve estar presente nos resultados. O fato do pesquisador não conhecer a sua variável não parece desculpa válida para não apresentá-la, já que é necessária a compreensão do seu desfecho na interpretação, inclusive, do valor da média. Por estes motivos, é sugerida a utilização da diferença bruta de médias para apresentação dos resultados acompanhada do intervalo de confiança da mesma, para fugir da automaticidade da significância estatística como medida da força do fenômeno (GOODMAN, 1999).

Em relação à inquietação pela possível inclusão da pesquisa em uma metanálise, de posse das médias, dos desvios padrão e do tamanho das amostras dos grupos, pesquisadores poderão calcular o *d* de Cohen se necessário (ou melhor ainda, o *g* de Hedges porque seu valor corrigido que o torna mais exato) e poderão também contextualizar o valor do mesmo encontrado com a diferença existente na unidade de origem.

Após uma revisão dos pontos de vista de diferentes autores parece que na visão da maioria, o uso do *d* de Cohen como medida de efeito seria limitado na interpretação da significância prática ou clinica dos resultados e que seria uma medida a ser usada como complementar à diferença bruta de médias e o seu intervalo de confiança.

REFERÊNCIAS

APA. **Publication Manual of the American Psychological Association**. 5th Edition. Washington DC: American Psychological Association, 2001.

APA. **Publication Manual of the American Psychological Association**. 6th Edition. Washington DC: American Psychological Association, 2010.

Baguley, T. Standardized or simple effect size: what should be reported? **British Journal of Psychology**. *v.* 100, n. Pt 3, p. 603-617, 2009.

Beck, A.T.; Steer, R.A.; Brown, G.K. **Manual for Beck Depression Inventory-II**. San Antonio: The Psychological Corporation, 1996.

Borenstein, M. et al. **Introduction to Meta-analysis**. John Wiley &Sons, 2009

Coe, R. (2002), "*It's the effect size, stupid: What effect size is and why it is important*," In: The Annual Conference of the British Educational Research Association, University of Exeter. 2002, England.

Cohen, J. The statistical power of abnormal-social psychological research: A review. **Journal of Abnormal and Social Psychology**, v. 65, n. 3, p. 145-153, 1962.

Cohen, J. **Statistical Power Analysis for the Behavioral Sciences**, 2nd Edition. Hillsdale, N.J.: Lawrence Erlbaum, 1988.

Cohen, J. Things I have learned (so far). **American Psychologist**. v. 45, n. 12, p. 1304-1312, 1990.

Cohen, J. A Power Primer. **Psychological Bulletin**. v. 112, n. 1, p. 155-159, 1992.

Cohen, J. Statistical power analysis. **Current Directions in Psychological Science**. *v.* 1, n. 3, p. 99-101, 1992.

Conboy, J. Algumas medidas típicas univariadas da magnitude do efeito. **Análise Psicológica**. v. 2, n. XXI, p. 145-158, 2003.

Durlak, J. How to Select, Calculate, and Interpret Effect Sizes. **Journal of Pediatric Psychology**. v. 34, n. 9, p. 917–928, 2009.

Fidler, F. **The American Psychological Association Publication Manual**. Sixth edition: implications for statistics education. International Association of Statistical Education (IASE), 2010.

Goodman, S. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy, **Annals of. Internal. Medicine**. v. 130, n. 12, p. 995-1004, 1999.

Higgins, J.P.T., Green, S. (editors). **Cochrane Handbook for Systematic Reviews of Interventions**. Version 5.1.0 [updated March 2011]. The Cochrane Collaboration, 2011.

Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs, **Frontiers in Psychology**. v. 4, n. 863, p. 1-12, 2013.

Lenth, R. Some Practical Guidelines for Effective Sample Size Determination. **The American Statistician**. v. 55, n. 3, p. 187-193, 2001.

Lipsey, M. et al. **Translating the Statistical Representation of the Effects of Education Interventions Into More Readily Interpretable Forms**. Report prepared for the National Center for Special Education Research, Institute of Education Sciences under Contract ED-IES-09-C-0021, 2012.

Nakagawa, S. Effect size, confidence interval and statistical significance: a practical guide for biologists. **Biological Reviews**. v. 82, p. 591-605, 2007.

Sedlmeier, P.; Gigerenzer, G. Do studies of statistical power have an effect on the power of studies? **Psychological Bulletin**. v. 105, n. 2, p. 309-316, 1989.

Soares, E. Capacidade funcional, declínio cognitivo e depressão em idosos institucionalizados: possibilidade de relações e correlações. **Revista Kairós Gerontologia**. v. 15, n. 5, p. 117-139, 2012.

Wilkinson, L.; The Taskforce on Statistical Inference. Statistical methods in psychology journals: Guidelines and expectations. **American Psychologist**. v. 54, n. 8, p. 594-604, 1999.