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

META-ANÁLISE

LUCIANA NEVES NUNES
ORIENTADORA: JANDYRA MARIA GUIMARÃES FACHEL

Monografia apresentada para obtenção do grau de Bacharel em Estatística

PORTO ALEGRE, JANEIRO DE 1997.

Dedico à Suzi, que foi meu maior

incentivo para chegar até aqui.

AGRADECIMENTO ESPECIAL

Agradeço à minha mãe e minhas irmãs por terem sido os melhores exemplos de vida que uma pessoa poderia ter.

AGRADECIMENTOS

À Professora Jandyra, pela orientação deste trabalho e pela amizade ao longo de todos estes anos.

Aos Professores do Departamento de Estatística que em suas aulas fizeram aumentar cada vez mais a minha paixão pela Estatística.

À Simone e à Stela, por terem me dado o privilégio de conhecê-las melhor e ter descoberto o quão maravilhosas são.

Às minhas sócias, Silvana e Vânia, por segurarem sempre as minhas "pontas" e pelos momentos de descontração que curtimos juntas.

Ao Gio e ao Gu, pelas risadas que compartilhamos e ao Gu, em especial, por ter conseguido os livros consultados para realizar este trabalho.

À todos os colegas que em algum momento cruzaram a minha vida e me fizeram ver o quanto se pode ser feliz neste mundo.

Aos funcionários do Instituto de Matemática que sempre colaboraram e estiveram prontos a me ajudar.

SUMÁRIO

1. INTRODUÇÃO

1.1. Considerações Gerais

Estudos independentes sobre um mesmo tema são realizados em lugares, momentos e por pessoas diferentes. Não raras vezes alguns estudos são realizados com amostras não muito grandes devido à raridade dos eventos e a todo momento sabemos dos resultados destes estudos isoladamente, então por que não generalizarmos os resultados em um único estudo?

Já no início deste século esta necessidade da combinação de resultados de estudos independentes começou a surgir, aos poucos foram aparecendo técnicas de agregação destes resultados, e na década de 70 surge, então, a Meta-análise, que é um método quantitativo de combinar os resultados de estudos independentes.

A técnica Meta-análise ainda está em desenvolvimento e por ser um assunto relativamente novo e de tanta importância, pois está sendo cada vez mais utilizado nas mais diversas áreas, com destaque

para a área da Medicina, é relevante que se explore ao máximo este assunto.

A presente monografia tem por principal objetivo fazer uma revisão de literatura que servirá como referencial teórico sobre Meta-análise, servindo como uma espécie de guia para quem se interessar em aprender um pouco sobre a técnica.

Neste trabalho estão presentes um pequeno histórico da Meta-análise, os testes combinados mais utilizados, métodos de combinação de comparação de grupos de vários estudos, e combinação das correlações de Pearson de vários estudos. Também apresentamos as fontes de viés e como examiná-las e reduzi-las da análise, testes de homogeneidade dos *effect sizes* e das estatísticas dos testes utilizados nos estudos combinados, validade e fidedignidade da análise, métodos não-paramétricos e ainda, críticas à utilização da Meta-análise.

Esta monografia teve como referência básica o livro "Meta-analysis - Quantitative Methods for Research Synthesis" de Frederic M. Wolf (1986), e muitos dos exemplos nele apresentados serão ilustrados aqui.

1.2. Histórico

A Meta-análise surgiu da necessidade de se sintetizar resultados de estudos que eram realizados sobre o mesmo assunto, mas em lugares e épocas diferentes em cada pesquisa realizada. A partir desta necessidade surgiu a Meta-análise que tem como principal objetivo analisar e combinar os resultados de estudos já realizados.

O nome "Meta-análise" surgiu em 1976 e foi atribuído por Gene Glass, que foi a primeira pessoa a se referir à técnica com este nome, mas há estudos anteriores a esta data que podem ser considerados meta-análises. Desde então esta técnica vem desenvolvendo-se a passos largos e sua aplicação vem se tornando cada vez mais freqüente, principalmente na Medicina.

Segundo Glass (1976) há distinção entre análises primária, secundária e Meta-análise:

> "*Análise primária* é a análise original dos dados realizada numa pesquisa... *Análise secundária* é a re-análise dos dados com o propósito de responder a questão original da pesquisa com técnicas estatísticas melhores, ou responder novas questões com velhos dados... *Meta-análise* se refere à análise das análises... a análise estatística de uma grande coleção de análises resulta num estudo individual com o propósito de integrar os achados. Isto implica uma rigorosa alternativa ao casual, discussões narrativas de estudos de pesquisa, os quais representam nossas tentativas de fazer senso de rapidamente expandir a pesquisa literária."

Gene Glass foi o primeiro a publicar um artigo sobre Meta-análise, e por isso, ele é conhecido como "papa" da Meta-análise. No entanto outras pessoas estavam trabalhando no assunto na mesma época (Hunter e Schmidt), mas publicaram seus trabalhos posteriormente. Para lançar um livro sobre Meta-análise, novamente Glass foi mais rápido e lançou o primeiro livro no assunto, em 1981, e somente em 1982 Hunter, Schmidt e Jackson conseguem lançar seu livro. Desde então estes livros têm sido extensivamente utilizados em várias áreas com algum destaque para a área de pesquisa organizacional industrial.

O aumento da utilização da Meta-análise foi tão grande que foi se tornou tema para estudos. Um destes estudos foi verificar o número de artigos e dissertações que trabalharam com Meta-análises, publicados entre os anos de 1977 e 1984 (Guzzo, Jackson e Katzell, 1986). Os dados por eles encontrados foram os seguintes:

1977	1978	1979	1980	1981	1982	1983	1984
2	4	6	9	18	32	55	63

Segundo a *National Library of Medicine*, Meta-análise foi definida como sendo um método quantitativo de combinar os resultados de estudos independentes e sintetizar as conclusões em um único estudo.

Então, Meta-análise é uma técnica que criticamente revê e estatisticamente combina os resultados de estudos já realizados. Os

objetivos da Meta-análise podem ser citados da seguinte forma (Wolf, 1986):

1. Aumentar o poder estatístico para grupos e subgrupos de estudos;
2. Resolver incertezas quando são relatadas discordâncias;
3. Melhorar as estimativas dos efeitos de medidas;
4. Responder questões não afirmadas nos estudos individuais.

Para que não ocorram problemas de vieses na escolha e seleção de estudos para Meta-análise, pois muitas vezes os próprios revisores geram viés ao selecionarem apenas os trabalhos que resultaram em significância e também por não terem critérios pré-definidos para esta escolha nas revisões de literatura de estudos independentes, é preciso que siga-se um guia para validar-se revisões, integrações e sínteses de estudos que tenham similares questões de pesquisa. Procedimentos empregados em Meta-análise permitem revisões quantitativas e síntese da pesquisa literária que trata dos resultados.

Baseando-nos em várias abordagens sobre meta-análise, podemos generalizar os seguintes passos para que tenhamos uma "boa" Meta-análise:

1. Definir o problema e critérios de seleção dos trabalhos;
2. Localizar os trabalhos na literatura ou mesmo trabalhos não publicados;

3. Classificar e codificar as características dos trabalhos;

4. Quantificar as características dos trabalhos em uma escala comum;

5. Agregar os achados e relatá-los (análise e interpretação);

6. Relatar os resultados.

1.3. Contextos Metodológicos

Segundo Normand (1995) em geral, há dois contextos metológicos para se sintetizar informações de um grupo de estudos: a abordagem de efeitos fixos e a abordagem de efeitos aleatórios. A escolha da abordagem se dá baseada na interpretação dos resultados e também na questão de interesse da pesquisa.

No contexto de efeitos fixos a Meta-análise é realizada quando os estudos selecionados têm por objetivo fazer uma inferência sobre um único parâmetro, supondo que há homogeneidade entre os parâmetros populacionais estimandos em cada estudo. Laird e Mosteller (1990) indicam um método para trabalhar neste contexto. Neste caso a única fonte de incerteza é das amostras de cada estudo.

O contexto de efeitos aleatórios pode ser conceituado como um procedimento em duas etapas. O conjunto de estudos em consideração representam uma amostra aleatória de um população de estudos e cada efeito específico de um estudo é uma amostra de uma população de efeitos ou parâmetros. Assim, no contexto de efeitos

aleatórios, cada estudo tem seu próprio efeito amostral e este é uma estimativa do efeito populacional. Em contraste ao contexto de efeitos fixos, nesta abordagem há duas fontes de variabilidade: uma devido a variabilidade nas estimativas dos parâmetros e outra devido a variabilidade das amostras em cada estudo. Uma alternativa para se trabalhar neste contexto é adotar a filosofia Bayesiana, especialmente dentro do paradigma Bayesiano que diz: a variabilidade na estimativa do parâmetro representa a variação dos efeitos populacionais.

A partir daí nasce a dúvida sobre qual a abordagem que deve ser utilizada e a respeito disto alguns autores argumentam que é responsabilidade do meta-analista identificar as fontes de variabilidade dos estudos. Hedges (1994) argumenta que se o modelo de efeitos fixos explica toda a variação nas estimativas dos parâmetros, então este é o modelo correto.

Raudenbush (1994) defende a abordagem de efeitos aleatórios porque como existe um vasto número de variáveis moderadoras (diferentes características de cada estudo) um estudo de efeito populacional deve ser aleatório. Sob o ponto de vista Bayesiano, como há incerteza sobre os processos que geram os efeitos, a incerteza do pesquisador pode ser incorporada, vendo todos os parâmetros desconhecidos como variáveis aleatórias.

2. TESTES COMBINADOS

2.1. Introdução

Desde que Fisher (1932) e Pearson (1933) começaram a tratar estatisticamente de sínteses de resultados independentes de testes da mesma hipótese, o interesse nestes tipos de procedimentos tem continuado. Os métodos descritos nesta monografia são análises de resultados de testes de hipóteses de estudos independentes utilizados para gerar um teste de hipóteses geral.

Os testes combinados são formas de se agregar quantitativamente os resultados de testes de hipóteses de estudos independentes, gerando um único resultado, isto é, podemos concluir sobre a hipótese de forma geral, como se agora tivéssemos um único estudo.

2.2. Teste combinado de Fisher

Com a questão de combinar os resultados de um número de testes independentes que foram planejados para testarem a mesma hipótese (comparação de médias, por exemplo), Fisher descreveu um método baseado na multiplicação de probabilidades (p-value´s) de diferentes experimentos. Se os logaritmos destas probabilidades são calculados, multiplicados por menos dois (-2), e então somados, uma estatística χ^2 com os graus de liberdade igual a duas vezes o número de testes combinados (2n) é obtida. A transformação logarítmica permite uma função de soma melhor que uma função multiplicativa e, com isso, facilita os cálculos. Logo temos a seguinte expressão:

$$\chi^2 = -2\sum_{i=1}^{n} \log_e p \qquad (1)$$

onde,
n = número de testes combinados;
p = probabilidade unilateral associada a cada teste.

Este procedimento tem se mostrado melhor assintoticamente que outros métodos de combinação de estudos (Koziol e Pearlman, 1978; Littel e Folks, 1973), e segundo Mosteller e Bush (1954) ao compararmos o método de Fisher com o teste do sinal, é possível verificar-se resultados inconsistentes, pois para determinados valores de p (valores

próximos de 0,5, por exemplo) o teste do sinal pode facilmente rejeitar a hipótese de nulidade e o teste de Fisher não. Podemos dizer que o teste de Fisher é mais conservador em seus resultados do que o teste do sinal.

Uma desvantagem do procedimento de Fisher é o "apoio" à significância dos resultados dos estudos, mesmo quando as significâncias de dois estudos estão igualmente e fortemente em direções opostas. Tirando estas limitações, este é o melhor procedimento conhecido e aplicado.

2.3. Teste combinado de Winer (combinação de testes t)

Foi apresentado por Winer (1971) um procedimento para combinação de testes t independentes provenientes de amostras independentes em que as estatísticas t associadas a cada teste são somadas e divididas pela raiz quadrada da soma dos graus de liberdade (gl) associado com cada t depois de cada gl ter sido dividido por gl-2. Podemos expressar da seguinte forma:

$$Z_c = \frac{\sum t}{\sqrt{\sum [gl/(gl-2)]}} \qquad (2)$$

Como gl/(gl-2) é a variância de uma distribuição t, temos que Z_c tem uma distribuição aproximadamente normal padrão [N(0,1)] quando

gl>10. A partir disto verificamos que o procedimento não é adequado quando as amostras são muito pequenas, ou seja, menores que 10. Entretanto como geralmente para testes de significâncias não são usadas amostras pequenas, esta desvantagem é minimizada.

2.4. Teste combinado de Stouffer

Este procedimento é similar ao teste de Winer, e foi atribuído por Stouffer et al. (1949) e mais completamente descrito por Mosteller e Bush (1954) e Rosenthal (1978), com a exceção de que os valores de p são convertidos em z's ao invés de t's, e então somados. O denominador é simplesmente a raiz quadrada do número de testes combinados, expressado da seguinte forma:

$$Z_c = \frac{\sum z}{\sqrt{N}} \qquad (3)$$

Este teste é baseado na soma dos desvios de Normais, sendo ele mesmo um desvio de uma Normal, com a variância igual ao número de observações somadas.

Há várias vantagens em se utilizar este teste. Seu cálculo é mais direto que o cálculo do teste combinado de Fisher, que necessita de transformações logarítmicas, e que o teste combinado de Winer que exige ajustamento dos graus de liberdade. Mais ainda, os resultados dos

procedimentos "z", enquanto são elegantemente mais poderosos, são virtualmente idênticos aos resultados dos procedimentos t (Wolf e Spies, 1981). Isto é verdadeiro quando as estatísticas somadas são derivadas de amostras grandes, isto é, na vizinhança de gl/(gl-2).

Exemplo numérico

Suponha que queremos fazer um revisão dos estudos que testaram a hipótese de que exercícios podem aumentar a auto-estima individual e que encontramos quatro estudos à respeito disto. A Tabela 1 apresenta os resultados destes quatro estudos fictícios. Os estudos A e C usaram o "Coopersmith Self-Steem Inventory" para medir a auto-estima, enquanto o estudo B usou a "Tenessee Self-Concept Scale" e o estudo D usou a "Rosenberg Self-Esteem Scale".

Tabela 1 - Resultados hipotéticos de quatro estudos que examinam os efeitos de exercícios na auto-estima.

	Grupo Controle		Grupo Experimental			
Estudo	n	$\bar{X}$	n	$\bar{X}$	DP combinado	t
A	41	11	41	17	16	2,72**
B	29	225	33	175	100	1,95
C	104	9	98	12	7	2,03*
D	11	23	11	31	12	1,56

*p<0,05, teste bilateral; **p<0,01, teste bilateral

Os resultados destes estudos mostram que houve diferença significativa entre as médias dos grupos controle e experimental nos estudos A e C, mas não houve diferença nos estudos B e D. Resultados diferentes de estudos são comuns nas ciências do comportamento e a questão permanece: Exercícios podem aumentar a auto-estima? Numa revisão de literatura tradicional, julgamentos superficiais podem ser feitos sobre cada um dos estudos. Alguns estudos podem ser considerados mais valiosos que outros, e, assim, pesar mais nas conclusões que serão descritas. Usando um teste combinado será possível conseguir alguma evidência comum aos quatro estudos.

Os resultados da Tabela 1 estão resumidos na Tabela 2 de tal forma que facilite os cálculos dos testes combinados. Pode-se notar que os sinais que precedem os t´s e Z´s indicam a direção dos resultados, sendo que o sinal negativo indica que o resultado foi inconsistente com a maioria dos resultados.

Tabela 2 - Resultados dos quatro estudos independentes utilizados para calcular os testes combinados.

Estudo	t	gl	p, unilateral	Z	$-2\log_e p$
A	2,72	80	0,004	2,65	11,04
B	-1,95	60	0,970	-1,88	0,06
C	2,03	200	0,024	1,98	7,46
D	1,56	20	0,060	1,52	5,63

Aplicando a fórmula 1 correspondente ao procedimento de Fisher, nos nossos resultados, temos

$$\chi^2 = 11{,}04 + 0{,}06 + 7{,}46 + 5{,}63 = 24{,}19 \qquad (4)$$

Como temos quatro testes independentes para esta hipótese, um para cada estudo, então são 2n=(2)(4)=8 graus de liberdade, e um valor do χ^2=24,19 é associado com um p<0,01.

Similarmente, quando aplicamos a fórmula 2 correspondente ao procedimento de Winer aos mesmos dados, o seguinte resultado é obtido:

$$Z_c = \frac{2{,}72 - 1{,}95 + 2{,}03 + 1{,}56}{\sqrt{(80/78) + (60/58) + (200/198) + (20/18)}} = \frac{4{,}29}{\sqrt{4{,}18}} = 2{,}10 \qquad (5)$$

A probabilidade de obter este valor de Z_c ou maior é P(≥2,10)<0,018, unilateral. O teste de hipótese unilateral está sempre sendo usado porque a direção natural da hipótese resulta das direções já conhecidas da maioria dos resultados dos estudos individuais combinados nesta análise. Rosenthal (1980) discute isto mais detalhadamente.

Quando aplicamos a fórmula 3 correspondente ao procedimento de Stouffer aos dados, é preciso converter os valores de p em Z´s respectivos a estes valores e então somá-los e dividir pela raiz quadrada dos número de testes somados. Para que se consiga valores de p razoavelmente exatos, é preciso que se use Tabelas extensas da distribuição de t. Estas Tabelas podem ser consultadas em Federighi (1959) e em Rosenthal e Rosnow (1984).

$$Z_c = \frac{2{,}65 - 1{,}88 + 1{,}98 + 1{,}52}{\sqrt{4}} = \frac{4{,}27}{2} = 2{,}135 \qquad (6)$$

A probabilidade de obter este valor de Z_c ou maior é $P(Z \geq 2{,}135) < 0{,}017$, unilateral.

Independentemente de qual destes testes é usado, a evidência combinada destes quatro estudos indica que a hipótese nula de que exercícios não aumetam a auto-estima deve ser rejeitada se a extensão da inferência é com respeito as populações combinadas. Curiosamente, exercícios parecem afetar positivamente a auto-estima, mesmo que resultados de dois dos quatro estudos investigados tenham dado resultados diferentes quando examinados independentemente. Razões para isto são discutidas nos capítulos 3 e 4.

2.5. A escolha de um teste combinado

Na prática, os resultados dos testes aplicados a estudos combinados são consistentes entre si. Talvez um fato a se considerar seja a facilidade de cálculo, em que o teste de Winer desponta.

Se os estudos independentes apresentarem estatísticas diferentes, para se combinar os resultados é necessário que se tenha muito cuidado nas transformações dos valores de p. O procedimento de Fisher tem a vantagem de ser o mais eficiente assintoticamente dos testes combinados, mas isto pode ser relativamente ponderado pela facilidade do cálculo do teste de Stouffer. De maneira prática, a diferença entre os resultados dos testes não é importante.

3. MEDIDAS DO EFEITO POPULACIONAL (*EFFECT SIZE* - ES)

3.1. Definição de Efeito Populacional

O efeito populacional, θ, é escolhido para representar qualquer quantidade de relevância à particular questão em estudo. O efeito pode representar a verdadeira diferença entre tratamento e controle em um estudo ou o verdadeiro *odds ratio* num estudo de caso controle, ou ainda a correlação num estudo correlacional. Para cada parâmetro de interesse, algumas estatísticas resumo correspondentes podem ser utilizadas para estimá-lo.

Mas, frequentemente, escalas diferentes de mensuração são utilizadas em estudos diferentes para estimar o mesmo parâmetro. Nesta situação cada estatística resumo pode ser transformada numa estatística padronizada (*scale-free*) denotada por Efeito Populacional (*Effect Size)*.

Poderíamos expressar ES como sendo "o grau em que o fenômeno está presente na população", ou, "o grau de falsidade da hipótese nula". Uma possível tradução seria efeito populacional, mas

para tornar mais claro e geral usaremos igualmente o termo original em inglês: *effect size*.

Exemplos

1) Se a porcentagem observada de pacientes portando esquizofrenia paranóica na população de pacientes psiquiátricos é 52%, e se a hipótese nula é p=50%, então o efeito é medido como uma distância da suposição (hipótese nula) de 50%, o seu valor é de 2%. Já se a porcentagem, ao invés de 52%, fosse 60%, o ES seria 10%, um valor bastante superior.

2) Em um mercado consumidor, para determinar se a preferência por uma marca A sobre sua principal concorrente B está relacionada com o nível de renda do consumidor, a hipótese nula poderia ser: a diferença na renda média familiar dos usuários da marca A e da marca B é zero, ou, equivalentemente, que o tamanho do efeito ou "*effect size*" da renda sobre a preferência da marca é zero.

Na exposição de Glass a aplicação de Meta-análise conta fortemente com o uso de medidas de ES e isto foi muito bem resumido por Cohen (1977):

> "Sem a suposição necessária de casualidade, é conveniente o uso da frase *effect size* para significar "o grau no qual o fenômeno está presente na população", ou, "o grau no qual a hipótese nula é falsa". Qualquer que seja a maneira de representação

> do fenômeno numa pesquisa em particular no presente tratamento, a hipótese nula sempre significa que o *effect size* é zero."

Cohen (1965, 1977) mostra medidas de ES para muitos testes estatísticos comuns. Faremos referências às medidas de ES apropriados para os seguintes testes estatísticos: (1) diferença entre dois grupos avaliada pelo teste t de Student; (2) o grau de associação entre duas variáveis medido pela Correlação Cruzada de Pearson.

3.2. Diferença entre grupos

O objetivo é obter um número padronizado, livre de sua unidade de medida original alternativamente chamado de "divergência das hipóteses nula e alternativa", ou, o ES que queremos detectar.

Pode-se mostrar seu cálculo da seguinte forma:

$$d = \frac{\mu_1 - \mu_2}{\sigma}, \text{ para testes unilaterais}$$

$$d = \frac{|\mu_1 - \mu_2|}{\sigma}, \text{ para testes bilaterais} \qquad (7)$$

onde,

d = o ES para testes t de médias, expresso em unidade padrão

μ_1 e μ_2 = médias populacionais em unidades de medida originais para os grupos comparados

σ = desvio-padrão comum às populações (homogeneidade de variâncias é assumido)

Médias e desvios-padrão amostrais são utilizados como aproximações de suas medidas populacionais,

$$d = \frac{\overline{X}_1 - \overline{X}_2}{DP}, \text{ para testes unilaterais}$$

$$d = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{DP}, \text{ para testes bilaterais} \qquad (8)$$

onde,

d = o ES para testes t de médias, expresso em unidade padrão

$\overline{X}_1$ e $\overline{X}_2$ = médias amostrais em unidades de medida originais para os grupos comparados 1 e 2

DP = desvio-padrão comum às amostras (homogeneidade de variâncias é assumido)

Assim consegue-se obter uma escala invariante estimada para θ. Tipicamente utiliza-se o desvio-padrão do grupo controle ou do pré-teste. Alternativamente, o desvio-padrão combinado (*pooled*) da população pode ser utilizado e até é preferido por alguns pesquisadores.

Exemplo numérico

Retornando ao nosso exemplo que testam as hipóteses de pesquisa de exercícios que podem aumentar a auto-estima, os resultados dos quatro estudos independentes da Tabela 1 são sumarizados

novamente na Tabela 3. O *Effect Size* (d) é calculado separadamente para cada um dos quatro estudos. Aplicando a equação 8 nos resultados do estudo A, por exemplo, nós obtemos

$$d = \frac{|11 - 17|}{10} = \frac{6}{10} = 0{,}60 \qquad (9)$$

O valor de d é calculado de forma similar para os outros estudos. O valor absoluto de d (diferença padronizada entre duas médias) é obtido e então especificado como positivo (+) ou negativo (-).

Tabela 3 - Resultados e d calculado para quatro estudos independentes

	Média do grupo				
Estudo	Controle	Experimental	DP combinado	d	U_3 (%)
A	11	17	10	0,60	72,6
B	225	175	100	-0,50	30,9
C	9	12	7	0,43	66,6
D	23	31	12	0,75	77,3
Média				0,32	62,5

Valores positivos especificam valores de d associados a resultados que favorecem o grupo experimental, enquanto valores negativos de d estão associados a resultados que favorecem o grupo controle. em nosso exemplo, os d´s para os estudos A, C e D são valores positivos, enquanto o d do estudo B é um valor negativo. Calcularemos o d médio de acordo com a fórmula 10 e este representará a estimativa do ES cruzado dos quatro estudos.

$$d_{medio} = \frac{\sum d}{n} \qquad (10)$$

onde,

d = o *effect size* para cada estudo independente

n = número de estudos

Assim, para nosso exemplo na Tabela 3, encontramos

$$d_{médio} = \frac{0,60 - 0,50 + 0,43 + 0,75}{4} = 0,32 \qquad (11)$$

Baseado em nossos achados para $d_{médio}$, nós podemos dizer que exercícios melhorariam a auto-estima em aproximadamente 0,32 unidades de desvio-padrão. Isto seria a melhor estimativa independentemente de como a auto-estima é medida, e vale lembrar que utilizamos três diferentes medidas como variáveis dependentes nos quatro estudos: a medida de Coopersmith nos estudos A e C, a de Tenessee no estudo B e a de Rosenberg no estudo C. As três medidas diferem em número e forma dos itens que contêm.

3.3. Interpretando Effect Sizes para estudos de diferença de grupos

Os métodos utilizados na Meta-análise têm sido desenvolvidos por Cohen (1977), Haase et al. (1982) e outros estudiosos.

Tendo calculado o ES, é importante agora que saibamos interpretá-lo. Um método para proceder esta interpretação é construirmos intervalos de confiança de 95 e 99% de confiança e verificarmos se algum destes inclui o zero em seus intervalos. Para calcularmos o intervalo de confiança para o d médio dos nossos estudos combinados devemos calcular também o desvio-padrão, que também nos fornece uma idéia de sua variabilidade. Em nosso exemplo o desvio-padrão

associado ao d de 0,32 é 0,56, então o intervalo de confiança de 95% para o d médio ficou (-0,23;0,87), ou seja, inclui o zero em seu intervalo. Seria desejável que o intervalo não incluísse o zero, pois nos daria mais certeza de que há efeito significativo entre os grupos para os estudos combinados.

Frequentemente não conseguimos definir uma distribuição padrão para d, ou seja, termos valores de comparação de d, e é precisamente neste momento que o conceito de *Effect Size* surge, pois ao invés de definirmos um distribuição exata para d, podemos posicioná-lo diretamente de tal forma que teremos uma classificação convencional de seus valores.

Os termos "pequeno", "médio" e "grande" são relativos, não somente para cada indivíduo, mas também para cada método utilizado em uma dada investigação. Em vista desta relatividade, há um certo risco inerente em oferecer definições operacionais convencionais para estes termos. Muitas vezes, entretanto, este risco é aceito, na crença de que mais é ganho do que seria perdido caso essas definições não fossem feitas.

Cohen (1977) sugere uma classificação de d médio para diferença de grupos para servir como um guia:

- *Effect Size* pequeno → d=0,2
- *Effect Size* médio → d=0,5
- *Effect Size* grande → d=0,8

3.4. Effect Size (d) como um percentual de não-intersecção das curvas de probabilidade (*Nonoverlap*)

Cohen (1977) sugere uma Tabela de conversão de ES (d) em medidas de *nonoverlap*, a qual Glass (1976, 1977) introduziu na abordagem meta-analítica das pesquisas.

Desta forma, o ES médio é transformado em uma representação gráfica do efeito do grau de sobreposição (*overlap*) entre os grupos controle e experimental. Para fazermos esta sobreposição precisamos aceitar as suposições de que os grupos a serem comparados são normais e com mesma variabilidade. É possível definir-se medidas de não-sobreposição (*nonoverlap*) como sendo a área de não intersecção entre curvas, U, associadas com d.

Cohen (1977) fornece os percentis de *nonoverlap* U_3 correspondente a vários valores de d em seu texto de análise do poder (Ver Anexo). Ao consultarmos uma Tabela de valores da curva Normal e olharmos a área sob a curva associada ao nosso valor de d, obteremos o percentual do grupo experimental que excede a metade superior de casos de nossa população do grupo controle.

Por exemplo, a área sob a curva normal associada ao valor de $d_{médio}$=0,32 é de 0,625. Isto significa que em média as pessoas que fazem exercícios teriam um escore de auto-estima maior que 62,5% das pessoas que não tiveram os exercícios. Assim seria esperado que os exercícios fizessem com que uma pessoa do grupo controle que está no

50º percentil passasse para o 62,5º percentil de auto-estima. Esta sobreposição gráfica dos grupos pode ser verificada através da Figura 1.

Outros métodos gráficos têm sido utilizados para resumir *effect sizes* e têm sido típico estes desenvolvimentos de distribuições do ES para coleções de estudos que são sintetizados.

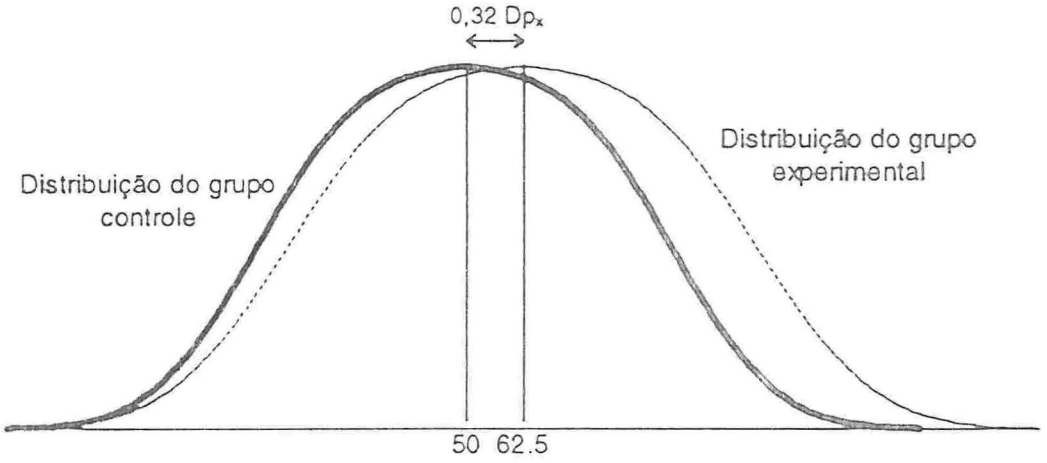

Figura 1 - *Effect size* médio em unidades de desvio-padrão (Dp_x)

3.5. Correlação

Os métodos para sintetizar os resultados de estudos de correlação entre duas variáveis que respondem a mesma questão de pesquisa, são bastante confiáveis. Basicamente, é feita uma média entre os coeficientes de correlação (Coeficiente r de Pearson), que são calculados separadamente em cada estudo. Isto é feito através da aplicação da fórmula 12, que segue:

$$\bar{r} = \frac{\sum r}{n} \qquad (12)$$

onde,

r = correlação de Pearson para cada estudo

n = número de coeficientes de correlação combinados

Exemplo numérico

Suponha que queremos rever estudos prévios que testaram a hipótese de que a renda estava significantemente relacionada com a auto-estima, e nossa revisão de literatura descobriu que somente quatro estudos respondiam esta questão. Os resultados destes estudos estão resumidos na Tabela 4.

Aplicando a fórmula 12 a estes estudos, encontramos

$$\bar{r} = \frac{0{,}13 + 0{,}56 - 0{,}24 + 0{,}67}{4} = 0{,}28 \qquad (13)$$

Tabela 4 - Correlações entre renda pessoal e auto-estima em quatro estudos fictícios

Estudo	n	r_{xy}
A	16	0,13
B	82	0,56**
C	102	-0,24*
D	47	0,67**

*p<0,05, teste bilateral; **p<0,01, teste bilateral.

Frequentemente distribuições de frequência ou diagramas de ramos e folhas são usados para resumir um grande número de correlações encontradas numa revisão de literatura. Suponha que encontramos 15 estudos que mostram as seguintes correlações (r) entre renda e auto-estima: 0,20; 0,17; 0,41; -0,24; 0,27; 0,34; 0,37; -0,06; 0,26; 0,67; 0,37; 0,23; 0,38; 0,35 e 0,40, respectivamente. Um diagrama de ramos e folhas (Tukey, 1977) é similar ao que é mostrado na Tabela 5 que pode ser construído para resumir os resultados de nossa Meta-análise dos 15 estudos. Por exemplo, a maior das quinze correlações, 0,67, será descrita em nosso diagrama colocando o primeiro decimal (,6) como um ramo e o segundo decimal (,07) como uma folha. A Tabela é construída incluindo ramos para todos os valores entre a maior e a menor correlação (-0,24 em nosso exemplo). Cada correlação é então incluída na Tabela da mesma maneira descrita para 0,67.

Medidas de tendência central e de variabilidade normalmente são incluídas num diagrama de ramos e folhas para facilitar a interpretação do conjunto de correlações.

Tabela 5 - Diagrama de ramos e folhas das correlações fictícias entre renda pessoal e auto-estima e Estatísticas de Resumo

r		Estatísticas de Resumo	
,6	7		
,5		r máximo	0,67
,4	0 1	Terceiro quartil (Q_3)	0,38
,3	4 5 7 7 8	Mediana (Q_2)	0,34
,2	0 3 6 7	Primeiro quartil (Q_1)	0,20
,1	7	r mínimo	-0,24
0		Média ($\bar{r}$)	0,27
-,0	6	Desvio-Padrão de $\bar{r}$	0,21
-,1		Média ponderada	0,28
-,2	4		

O assunto sobre a interpretação do que constitui um grande efeito é similar ao já discutido previamente no item comparação de grupos. Pela impossibilidade de padronizarmos uma interpretação para o ES, o que pode ser sugerido é a idéia de Cohen (1977) em estabelecer o seguinte guia para os valores de ES para Correlação:

- *Effect Size* pequeno → r=0,1
- *Effect Size* médio → r=0,3
- *Effect Size* grande → r=0,5

Assim, uma correlação média de 0,28 sugere um *Effect Size* médio para a relação entre renda pessoal e auto-estima. É preciso que se tenha em mente que este guia é arbitrário, mas que sendo utilizado de forma sensata pode ser de grande ajuda na interpretação dos dados.

3.6. Selecionando uma métrica comum

Frequentemente ao fazermos a revisão de literatura para a seleção dos estudos que farão parte da Meta-análise, defrontamo-nos com o problema de que os resultados de cada estudo estão apresentados de forma diferente, isto é, análises diferentes são realizadas em cada estudo. Podemos ter uma coletânea de estudos que estão avaliando a mesma questão, mas que os resultados das diferenças dos grupos são apresentados usando t, F ou outras estatísticas, enquanto outros estudos apresentam resultados de associação entre grupos usando r, χ^2 ou outras estatísticas. Para procedermos com a Meta-análise é necessário que se converta estas várias estatísticas em uma escala comum que pode ser de d ou r.

Nas Tabelas 6 e 7 a seguir há sugestões de conversões das estatísticas mais comuns em valores de r e d, respectivamente. Cohen (1965, 1977), Friedman (1968), Glass et al. (1981) e Rosenthal (1984) discutem estes procedimentos e sugerem conversões de estatísticas menos comuns. Uma vez estando selecionada a métrica comum a ser

utilizada: r ou d, cada estatística é convertida para esta métrica e os resultados de cada estudo são agregados usando os métodos previamente descritos.

Tabela 6 - Guia para conversão de estatísticas de testes em r

Estatísticas para converter	Fórmula para transformação em r	Comentário
t	$r = \sqrt{\frac{t^2}{t^2 + gl}}$	
F	$r = \sqrt{\frac{F}{F + gl(erro)}}$	Usar somente para a comparação de duas médias (i.é., gl=1)
χ^2	$r = \sqrt{\frac{\chi^2}{n}}$	n=tamanho da amostra Usar somente para Tabela de contingência 2x2 (gl=1)
d	$r = \frac{d}{\sqrt{d^2 + 4}}$	

Tabela 7- Guia para conversão de estatísticas de testes em d

Estatísticas para converter	Fórmula para transformação em d	Comentário
t	$d = \frac{2t}{\sqrt{gl}}$	
F	$d = \frac{2\sqrt{F}}{\sqrt{gl(erro)}}$	Usar somente para a comparação de duas médias (i.é., gl=1)
r	$d = \frac{2r}{\sqrt{1-r^2}}$	

3.7. O "efeito" combinado de procedimentos errôneos de avaliar a consistência dos efeitos

O efeito combinado de duas "fraudes" no uso de testes de significância estatística tem levado os pesquisadores a conclusões muito pessimistas dos resultados. Amostras e *effect sizes* pequenos, que são normalmente encontrados em pesquisa educacional, por exemplo, levam a situações na qual muitos, senão todos estudos falham na rejeição da hipótese nula ao nível de significância α=0,05.

Na Figura 2 a proporção esperada de resultados significativos (comparação de duas amostras) é mostrada como função do tamanho da amostra (assumindo que os estudos têm mesmo tamanho de amostra) para três valores de d. Os valores de d são: d pequeno=0,2, d médio=0,5 e d grande=0,8. As curvas mostram que amostras menores que 100 e pequeno *effect size*, a proporção esperada de resultados significativos nunca execedem 0,20, e também, quando d=0,5, a

proporção esperada de resultados significativos excede 0,50 somente se a amostra for por volta de 70.

Uma conclusão é que se muitos efeitos forem de pequenos à médios e as amostras forem moderadas, então haverá predominância de resultados não significativos. O uso do método de contagem de votos (Hedges e Olkin, 1985) decide qual efeito do tratamento que é "não zero" que conduz o efeito médio para "não zero". Alternativamente, o pesquisador deve considerar os pequenos resultados significativos como evidência de que o efeito do tratamento, embora em grande parte desprezível, tem um efeito significativo. Assim, o efeito combinado de dois procedimentos errôneos sugerem um pessimismo não garantido sobre a magnitude e possivelmente a significância do efeito considerado.

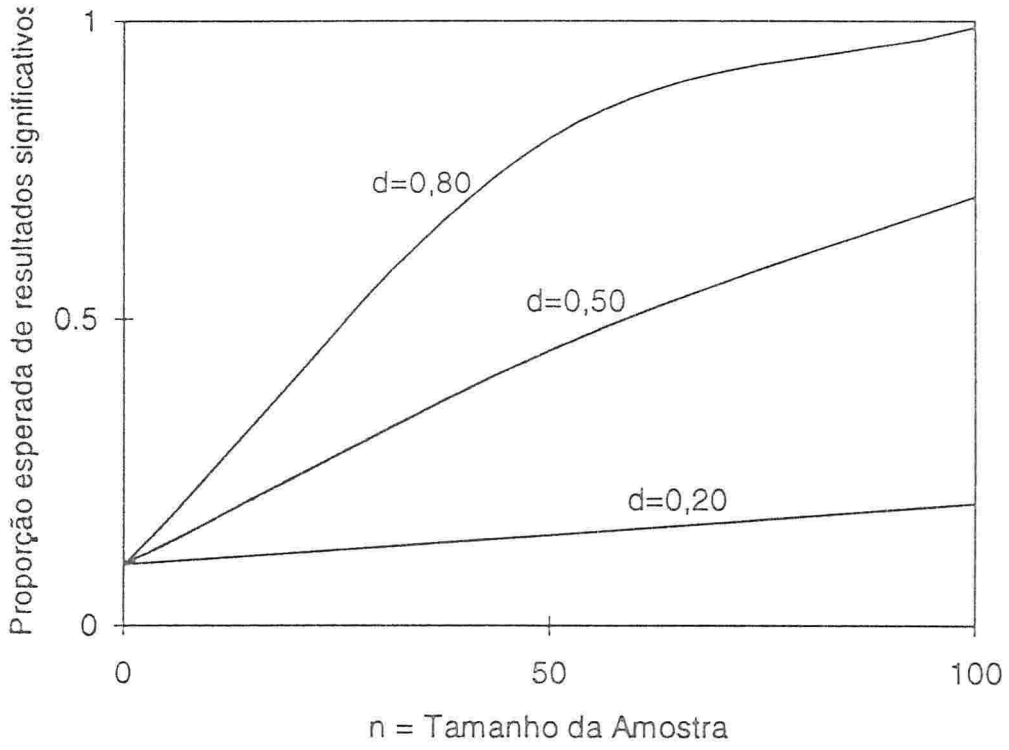

Figura 2 - A proporção esperada de resultados significativos como uma função do tamanho da amostra e do *effect size*.

4. EXAMINANDO E REDUZINDO O VIÉS

4.1. Introdução

Em estudos realizados, tanto qualitativos quanto quantitativos, há procedimentos que podem produzir viés nos resultados, como por exemplo a ponderação atribuída aos resultados de cada estudo examinado, que deve ser feita com cuidado para todos os estudos caso contrário pode-se não ter assegurado um alto nível de concordância ou fidedignidade entre os "juizes" que atribuem pesos aos estudos conforme suas características.

Cada uma destas fontes de viés são um problema e dificultam o trabalho dos meta-analistas, mas estratégias para se contornar estes problemas estão sendo desenvolvidas e propostas, pois desde que a Meta-análise surgiu muitos avanços estão sendo feitos neste sentido.

4.2. N salvo de falhas (Fail-Safe N)

Não podemos incluir TODOS os trabalhos realizados sobre um mesmo tema em uma Meta-análise, e então surge o problema de que geralmente os trabalhos publicados são aqueles cujos resultados foram significativos, pois os que não foram, muitas vezes, têm dificuldade de serem publicados e isto pode gerar um viés em nossa Meta-análise, pois aí estaremos com um viés já na seleção dos estudos.

Muitos Meta-analistas fazem análises separando os estudos em dois grupos: os publicados e os não publicados e testam as diferenças dos *effect size* de acordo com a fonte do estudo. Rosenthal tratou do problema analiticamente e sugeriu um cálculo para o número de estudos que seriam necessários ser incluídos na Meta-análise para que mudasse a conclusão a respeito da significância. Cooper (1979) chamou isto de *Fail Safe N* (N_{fs}) e baseado nos valores de p usuais de 0,05 e 0,01 podemos expressar este cálculo da seguinte forma:

$$N_{fs,05} = \left(\frac{\sum Z}{1{,}645} \right)^2 - N \qquad (14)$$

$$N_{fs,01} = \left(\frac{\sum Z}{2{,}33} \right)^2 - N \qquad (15)$$

onde,

ΣZ = soma dos escores Z individuais;

N = número de estudos combinados

Se aplicarmos a equação 14 em nosso exemplo de quatro estudos fictícios da Tabela 2 que testam a hipótese de que exercícios afetam positivamente na auto-estima, obteremos o seguinte resultado:

$$N_{fs,05} = \left(\frac{2{,}65 - 1{,}88 + 1{,}98 + 1{,}52}{1{,}645}\right)^2 - 4 = 6{,}75 - 4 = 2{,}75 \qquad (16)$$

Assim, nós precisamos aproximadamente de mais 3 estudos, cada um mostrando efeito nulo (i.é., Z=0,0) ou somando efeito nulo (i.é., ΣZ=0) para reverter nossa conclusão de que exercícios afetam positivamente na auto-estima a 5%.

Orwin (1983) desenvolveu um cálculo similar baseado na média dos *effect size* que pode ser obtido a partir da fórmula:

$$N_{fs} = \frac{N\left(\bar{d} - \bar{d}_c\right)}{\bar{d}_c} \qquad (17)$$

onde,

N = número de estudos incluídos na Meta-análise;

$\bar{d}$ = o *effect size* médio dos estudos sintetizados;

$\bar{d}_c$ = valor crítico selecionado que seria igual a $\bar{d}$ se um número hipotético de estudos fossem incluídos na Meta-análise.

Quando calculamos o *fail-safe N* baseado na significância estatística do teste combinado de Stouffer, tradicionalmente o valor padrão de p é 0,05 e 0,01. Por causa da falta de concordância sobre os valores padrões de d, pode se usar a sugestão de Orwin (1983) que baseia-se em Cohen (1977) que é: d=0,2 (pequeno efeito), d=0,5 (médio efeito) e d=0,8 (grande efeito). Quando selecionamos d=0,2 como nosso valor crítico ($\bar{d}_c$), encontramos que seriam necessários 2 ou 3 estudos adicionais para que conseguíssemos baixar nosso efeito médio obtido na Meta-análise que era d=0,32 para d=0,2, ou seja, um efeito pequeno. Isto pode ser verificado no cálculo abaixo,

$$N_{fs0,2} = \frac{(4)(0,32 - 0,20)}{0,20} = 2,4 \qquad (18)$$

4.3. Ponderando estudos pelo tamanho da amostra

Pode-se discutir que nem todos os trabalhos incluídos em uma Meta-análise deveriam ter a mesma ponderação, no sentido de que alguns estudos são baseados em amostras muito pequenas ou não representativas, enquanto que outros estudos são muito bem planejados e possuem uma "boa" amostra. Dar o mesmo peso a todos estudos faz com que os menos representativos contribuam da mesma forma para os resultados de nossa Meta-análise que os muito bem planejados ou de maior poder.

Mosteller e Bush (1954) sugerem utilizarmos a ponderação dos desvios de normais padronizadas (Z's), também usados no teste combinado de Stouffer, pelo tamanho da amostra, para darmos pesos a cada um dos estudos. Isto pode ser feito pelos gl associados a cada teste estatístico usado na Meta-análise, de acordo com a fórmula 19:

$$Z_c\text{ponderado} = \frac{\sum glZ}{\sqrt{\sum gl^2}} \qquad (19)$$

onde,

Z = desvio da normal padrão associado com o valor de p unilateral de cada estatística sintetizada

gl = graus de liberdade associados com a estatística

Em nosso exemplo dos quatro estudos fictícios da Tabela 2, obtemos o seguinte $Z_{c\text{ponderado}}$:

$$Z_c\text{ponderado} = \frac{(80)(2{,}65)+(60)(-1{,}88)+(200)(1{,}98)+(20)(1{,}52)}{\sqrt{(80)^2+(60)^2+(200)^2+(20)^2}}$$

$$= \frac{525{,}6}{224{,}5} = 2{,}34 \qquad (20)$$

A probabilidade de obtermos este valor de $Z_{c\text{ponderado}}$ ou maior é $P(Z \geq 2{,}34) < 0{,}010$, unilateral. Este resultado pode ser comparado ao valor de Z_c não ponderado obtido anteriormente para estes estudos na equação 6, onde $p(Z \geq 2{,}13) < 0{,}017$, unilateral. Por causa do tamanho da

amostra do estudo C (n=200), o valor de Z_c não ponderado mudou para o valor de Z_c ponderado na direção dos resultados para este estudo. Neste caso a diferença entre os valores de Z_c ponderado e não ponderado não foi grande ($Z_{c\text{ponderado}}$=2,13; $Z_{c\text{ não ponderado}}$=2,34), mas não é sempre o caso. Em geral, é recomendado que se calcule ambos os Z_c, ponderado e não ponderado, e muitas Meta-análises mostram os dois valores.

4.4. Estimativa de ES não-viciada

Hedges (1982) e Rosenthal e Rubin (1982) apresentaram métodos para obter uma estimativa de θ não-viciado. Hedges (1981) desenvolveu uma distribuição amostral de d e mostrou que d tem um leve viés ao estimar θ. Hedges (1982) mostrou que um estimador ponderado de θ é assintoticamente eficiente e inteiramente preciso quando os grupos experimental e controle têm os tamanhos de amostra maiores que 10. A fórmula 21 pode ser usada para obter $\bar{d}$ ponderado:

$$\bar{d} = \frac{\sum wd}{\sum w} \qquad (21)$$

onde,

d= ES não ponderado

w = variância estimada do d recíproco de cada estudo a ser agregado na Meta-análise

Hedges (1982) e Rosenthal e Rubin (1982b) sugerem fórmulas para o cálculo da variância estimada. Quando os tamanhos de amostra dos grupos controle e experimental forem aproximadamente iguais e maiores que 10, Rosenthal e Rubin (1982b) sugerem a fórmula abaixo para estimar w de cada estudo independente:

$$w = \frac{2N}{8 + d^2} \tag{22}$$

onde,

d = ES não ponderado

N = tamanho da amostra total no estudo para ambos os grupos, controle e experimental

Por exemplo, para o estudo A, previamente sumarizado nas Tabelas 2 e 3, nós temos a seguinte estimativa para w a partir da equação 22,

$$w = \frac{2(82)}{8 + (0{,}60)^2} = 19{,}6 \tag{23}$$

As estimativas de w para os estudos B, C e D foram obtidos da mesma forma e aparecem na Tabela 8. Então podemos calcular nosso ES médio ponderado não viciado de acordo com a equação 21,

$$\bar{d} = \frac{(19{,}6)(0{,}60) + (15{,}0)(-0{,}50) + (49{,}4)(0{,}43) + (5{,}1)(0{,}75)}{19{,}6 + 15{,}0 + 49{,}4 + 5{,}1} = \frac{29{,}33}{89{,}1} = 0{,}33 \tag{24}$$

Tabela 8 - Resultados de quatro estudos fictícios usados para obter um $\bar{d}$ ponderado e não viciado

Estudo	N	d	w
A	82	0,60	19,6
B	62	-0,50	15,0
C	202	0,43	49,4
D	22	0,75	5,1

Nota: w é calculado pela equação 22.

Assim, nosso $\bar{d}$ ponderado e não viciado de 0,33 é um pouco maior que o d não ponderado de 0,32 obtido com a equação 11. Desta forma, Green e Hall (1984) notaram que este fator de correção para o leve viés de d está muito perto de um, não importando muito a diferença quando o tamanho da amostra (n) é maior que 10 sujeitos.

4.5. Testes de homogeneidade

Em nosso exemplo da Tabela 1, nós queremos sintetizar os resultados dos estudos A, B, C e D que testam a hipótese de que exercícios podem conduzir a um aumento na auto-estima. Para que se possa resumir estes resultados quantitativamente em uma Meta-análise, é suposto que todos os estudos forneçam uma estimativa do efeito populacional que é estimativo de θ.

Se uma série de estudos independentes fornecerem uma estimativa comum (homogênea) do efeito populacional, então é provável

que os estudos estejam testando a mesma hipótese. Já se as estimativas são heterogêneas, então levanta-se a questão de que os estudos não estão testando o mesmo efeito. Heterogeneidade indica que não é apropriado combinar e sintetizar os resultados de todos os estudos em uma Meta-análise. Neste caso um contexto de efeitos aleatórios deveria ser priorizado. O pesquisador deverá decidir quais estudos poderão formar subgrupos com resultados homogêneos para que se conduzam Meta-análises separadas.

Um problema prático em qualquer Meta-análise é decidir sobre a inclusão ou não de um estudo na coleção de estudos que serão agregados e sintetizados. Este ponto será discutido no item Validade e Fidedignidade, mas os métodos fornecidos neste item podem ser usados para testar a homogeneidade dos estudos e ajudar a detectar os *outliers* ou estudos fora do contexto. Os estudos devem ser examinados para poder se verificar o que pode estar deixando-os diferentes. Uma sugestão é que se faça um gráfico dos valores das estimativas de θ dos estudos, sendo possível que se tenha que fazer Meta-análises separadas para os *outliers*. Light e Pillemer (1984) sugerem um interpretação para este gráfico.

Há uma discussão sobre o que fazer com os estudos que tiveram seus resultados heterogêneos. Hedges (1982a) e Hunter et al. (1982) sugerem que é inapropriado incluí-los na Meta-análise. Harris e Rosenthal (1985) argumentam que a heterogeneidade é análoga as

diferenças encontradas entre os indivíduos em cada estudo e é comum quando muitos estudos são examinados por diferentes investigadores usando métodos diferentes. Light e Pillemer (1984) acreditam que estes estudos diferentes e os *outliers* possam trazer acréscimo ao entendimento e até gerar novas hipóteses. Até Hedges (Becker e Hedges, 1984) admitiu: "Não é necessariamente desaconselhável que se faça inferências dos efeitos heterogêneos".

4.5.1. Homogeneidade dos testes estatísticos

Rosenthal (1983, 1984) sugere testes estatísticos para avaliar a homogeneidade/heterogeneidade dos desvios da normal padrão Z correspondentes ao valor de p unilateral dos estudos que queremos combinar. Tabelas extensas da distribuição normal são frequentemente necessárias e podem ser encontradas em Rosenthal e Rosnow (1984) ou Federighi (1959). Para o caso de comparação de somente dois grupos, o procedimento é bastante fidedigno e é dado da seguinte forma:

$$Z = \frac{Z_1 - Z_2}{\sqrt{2}} \qquad (25)$$

A diferença entre Z_1 e Z_2 é distribuída como Z, quando dividida por $\sqrt{2}$.

Por exemplo, suponha que queremos testar se o resultado do estudo A (Z=2,65) é significativamente diferente do resultado do estudo B (Z=-1,88), como está na Tabela 2, no capítulo 2. Encontramos o seguinte resultado:

$$Z = \frac{(2,65) - (-1,88)}{\sqrt{2}} = 3,20 \qquad (26)$$

Nosso Z resultante de 3,20 tem um valor de $p<0,001$, unilateral, ou $p<0,002$, bilateral. Isto significa que temos que explorar porque os dois estudos têm resultados diferentes, e mais, não podemos incluir ambos na Meta-análise.

Há várias explicações plausíveis para a diferença significativa encontrada entre os estudos. Primeiro, os dois estudos utilizaram medidas de auto-estima diferentes. O estudo A usou a medida de Coopersmith, enquanto o estudo B usou a medida de Tenessee desenvolvida para crianças. Os resultados podem ser diferentes porque as amostras diferem muito de estudo para estudo, por exemplo, se o estudo A foi conduzido com estudantes universitários ou adultos enquanto o estudo B incluiu somente crianças pequenas. E ainda pode haver diferenças grandes entre os planejamentos experimentais de cada estudo. O ideal é que se faça um estudo atento sobre todas estas características metodológicas antes de chegarmos a esta etapa, ou alternativamente, realizar Meta-análises em separado para os estudos com adultos e os estudos com crianças, por exemplo, ou qualquer outra característica de diferença entre os estudos.

Quando desejamos testar a homogeneidade de mais de dois estudos independentes, um teste geral, ou o que Rosenthal (1983) chamou de teste difuso, deve ser usado para verificar quais estudos são homogêneos ou não. Se estes estudos apresentarem heterogeneidade significativa, é imperativo que se examine os *outliers* da distribuição dos estudos e se verifique o que está contribuindo para esta heterogeneidade. A equação 27 pode ser utilizada para realizar o teste geral difuso:

$$\chi^2 = \sum (Z - \bar{Z})^2 \qquad (27)$$

com graus de liberdade (gl)=k-1 onde k = o número de Z´s dos estudos independentes. Voltando ao nosso exemplo, podemos testar a homogeneidade dos quatro estudos:

$$\chi^2 = (2{,}65 - 1{,}07)^2 + (-1{,}88 - 1{,}07)^2 + (1{,}98 - 1{,}07)^2 + (1{,}52 - 1{,}07)^2 = 12{,}23 \qquad (28)$$

Um χ^2 de 12,23 com k-1=4-1=3 graus de liberdade tem significância de $p<0{,}01$, indicando que os quatro estudos são significativamente heterogêneos. Para explorar as causas desta heterogeneidade é razoável que se comece pelo estudo B no qual os resultados foram inconsistentes com a direção dos resultados dos outros estudos.

4.5.2. Homogeneidade do Effect Size

Podemos testar a homogeneidade/heterogeneidade dos *effect sizes* de maneira análoga ao teste de homogeneidade para testes estatísticos descrito na seção anterior. A equação 29, Rosenthal e Rubin (1982b, 1982c) pode ser usado para este propósito, onde a hipótese nula é de igualdade dos ES de todos estudos:

$$\chi^2 = \sum\left(w(d-\bar{d})^2\right) \tag{29}$$

onde,

$\bar{d}$ = d médio ponderado dos estudos a serem agregados

d = ES de cada estudo

w = variância estimada recíproca de cada d

O resultado é um qui-quadrado com k-1 graus de liberdade onde k é o número de estudos a serem agregados. O $\bar{d}$ e w podem ser obtidos a partir das fórmulas 21 e 22 descritas anteriormente, onde N é o tamanho da amostra no estudo.

Para o nosso exemplo, usando os estudos A, B, C e D na Tabela 8, obtemos o seguinte resultado pela equação 29 e usando $\bar{d}$=0,33 obtido na equação 24,

$$\chi^2 = (19,6)(0,60-0,33)^2 + (15,0)(-0,50-0,33)^2 + (49,4)(0,43-0,33)^2 + (5,1)(0,75-0,33)^2 \tag{30}$$

$$= 14,44$$

Um χ^2 de 14,44 com k-1=4-1=3 graus de liberdade tem significância de p<0,001, o que indica que há heterogeneidade entre os ES obtidos nos quatro estudos. Novamente, devemos investigar o estudo B para saber porque ele é diferente dos outros estudos.

4.6. Estudos com mais de uma estatística

Muitos estudos podem fornecer mais do que um teste de significância para uma hipótese que está sendo estudada em uma Meta-análise, e aí surge uma nova discussão sobre incluir ou não estes múltiplos testes de um mesmo estudo numa Meta-análise. Glass e seus colegas (Smith and Glass, 1980; Glass et al., 1981) defendem que devem ser incluídos enquanto outros argumentam que não devem ser incluídos (Kulik et al., 1980; Mazzuca, 1982; Findley and Cooper, 1983; Steinkamp and Maehr, 1983; Harris e Rosenthal, 1985).

Kulik (1983) argumenta que a inclusão de resultados múltiplos de um estudo inflaciona os efeitos dos estudos independentes. Enquanto pode aumentar o poder de nossa Meta-análise, torna difícil determinar o valor do erro nas estatísticas da coleção dos estudos sintetizados. Kulik (1983), Rosenthal (1984) e outros recomendam que é melhor realizar Meta-análises separadas para cada variável dependente do que aglutinar os diferentes resultados em uma só análise. Kulik (1983)

mantém que a aglutinação dos resultados podem gerar uma confusão conceitual. Rosenthal (1984) sugere que no mínimo deveria se examinar o efeito da variável independente sobre cada categoria das variáveis dependentes para detectar qual delas é a mais ou menos afetada.

Conduzir Meta-análises separadamente para diferentes classes de resultados de variáveis dependentes que tratam das hipóteses de interesse é uma solução prática para a crítica à Meta-análise conhecida como "problema das maçãs e laranjas", que veremos a seguir no capítulo 7. Esta abordagem aumenta a clareza conceitual e de interpretação dos resultados da Meta-análise em cada categoria de resultados ou variáveis dependentes que são analisadas separadamente. Assim, maçãs são tratadas como maçãs e laranjas como laranjas.

Strube (1985) previne que meta-analistas deveriam evitar estudos que impeçam um exame analítico das diferenças e similaridades dos resultados para diferentes categorias de resultados. Strube (1985) desenvolveu um procedimento computacional para ajustar o teste combinado de Stouffer para testes de hipóteses não independentes. Seu procedimento resulta num teste combinado mais conservador que meramente inclui todos os resultados não independentes, devendo ser evitado, pois há uma tendência de que o Erro Tipo I da Meta-análise seja inflacionado.

Tracz et al. (1985) demonstram uma regressão linear múltipla numa abordagem de Meta-análise que leva em conta múltiplos

tratamentos, não independência, interação e covariância construindo modelos completos e restritos. Este tema de não independência ou múltiplas variáveis dependentes é bastante complexo e importante, e apenas está começando a ser tratada mais sofisticadamente.

4.7. Validade e Fidedignidade

Vários artigos têm discutido sobre a fidedignidade na Meta-análise (Glass et al., 1981; Green and Hall, 1984; Hunter et al., 1982; McGuire et al., 1985; Orwin and Cordray, 1985; Rosenthal, 1984; Stock et al., 1982). Quando realizamos uma Meta-análise temos que avaliar a consistência dos resultados, pois os resultados meta-analíticos podem ser afetados se os estudos selecionados não passaram por um processo rigoroso e consistente. Outro tema crítico levantado é o grau de fidedignidade na codificação das características dos estudos a serem incluídos na análise.

Algumas questões foram formuladas e tentar-se-á responde-las, indicando alguns caminhos encontrados pela bibliografia consultada:

- Como é que normalmente meta-analistas independentes conseguem localizar e incluir os mesmos estudos para realizar uma Meta-análise?
- Quão ampla (completa) é a coleção de estudos incluídos em uma Meta-análise?

Stock et al. (1982) dá sete sugestões de aumentar a fidedignidade entre os codificadores:

1. Fazer um teste piloto para desenvolver os códigos antes de codificar as características para a Meta-análise.
2. Criar um detalhado e explícito manual com as formas de codificação.
3. Oferecer um treinamento aos codificadores baseado no manual e nas formas de codificação.
4. Medir e relatar a fidedignidade entre os codificadores como parte da Meta-análise.
5. Revisar o manual e formas de codificação e dar novo treinamento aos codificadores se necessário.
6. Desenvolver procedimento para admitir novos codificadores.
7. Encorajar os codificadores a se envolver em discussões e decisões sobre as formas de codificação.

Wortman (1983) comenta sobre a validade externa, interna, de constructo e das conclusões estatísticas das Meta-análises. Validades externa e de constructo recaem no "problema das maçãs e laranjas" e tentam determinar quais os trabalhos que devem ser agregados na Meta-análise. A codificação das características dos estudos e os testes de homogeneidade ajudam a examinar e aumentar a validade externa.

A validade interna numa Meta-análise está relacionada com quais variações na qualidade do planejamento influencia nos resultados

da Meta-análise. Este fator também pode ser codificado na Meta-análise e empiricamente estudado. Algumas análises encontram estudos que tem alta qualidade (i.é., tamanho grande de amostra, amostras bem controladas, etc.) e que resultam em ES mais baixos que estudos de baixa qualidade, enquanto outras Meta-análises tem mostrado que a qualidade de planejamento está relacionada com o ES.

Glass (1983) diz que "é desejável que Meta-análise inclua um exame empírico de validade interna". Este exame empírico é desejável no sentido de determinar o grau de validade interna de um conjunto particular de estudos. Green e Hall (1984) sugerem que o grau de "cegueira" do pesquisador, aleatorização, tamanho da amostra, controle dos erros, tipo de variável dependente (tipo: subjetiva versus objetiva), e viés de publicação são áreas metodológicas e de qualidade de planejamento que uma boa Meta-análise deve examinar.

5. MÉTODOS NÃO-PARAMÉTRICOS

5.1. Introdução

O interesse em medidas não-paramétricas do θ tem crescido desde que Glass et al. (1981) abordou o assunto de estimar θ a partir de estatísticas não-paramétricas e dados ordinais ou dicotômicos. É desejável que se utilize estimador não-paramétrico de θ quando temos dados não-normais ou quando algumas observações são *outliers*. Kraemer e Andrews (1982) salientam as limitações do ES estimado (d) paramétrico, o qual inclui que a interpretação de d depende da suposição de que os escores do grupo controle são normalmente distribuídos e que todos os sujeitos do grupo experimental têm os mesmos benefícios do tratamento (suposição de aditividade), e que d é não invariante sob todas as transformações de escalas monotônicas (por exemplo, transformação logarítmica).

Quando alguma destas suposições não é satisfeita, temos algumas alternativas para tentar resolver. Podemos tentar uma transformação nos dados que faça com que eles fiquem normalmente

distribuídos ou eliminar os *outliers* antes de realizar a Meta-análise com os procedimentos paramétricos já descritos.

Uma abordagem indica que o procedimento seria obter os desvios da normal padrão (Z) associados aos valores exatos de p das estatísticas não-paramétricas relatadas nos estudos. Outra abordagem, (Kraemer e Andrews, 1982) ainda, indica que deveria se usar o ES mediano ao invés de usar o ES médio como medida de tendência central para descrever o efeito populacional, e muitos Meta-analistas têm feito desta forma (Hyde, 1981; Mazzuca, 1982).

Rosenthal, entretanto, alerta "o uso de medianas em trabalhos meta-analíticos tendem a dar resultados que beneficiam o Erro tipo II, isto é, resultados que conduzem estimativas em favor da hipótese nula".

Todas estas discussões estão presentes nos trabalhos de vários autores e recomenda-se para quem tiver interesse em examinar a questão dos métodos não-paramétricos mais profundamente que os consulte: Hedges e Olkin, 1984; Krauth, 1983; Kraemer, 1984; Glass et al., 1981; Katz et al., 1985; Hodges e Lehman, 1962.

5.2. Effect size não-paramétrico

O ES não-paramétrico (D) pode ser obtido a partir da equação 31:

$$D = \phi^{-1}(p) \qquad (31)$$

onde D é o desvio da normal padrão, ϕ^{-1} é o inverso da função distribuição acumulada da normal padrão e p é a proporção de sujeitos do grupo controle que têm seu valor da variável dependente menor que a mediana do grupo experimental. Isto possibilita que p seja 0 ou 1, então, nestes casos, Kraemer e Andrews (1982) recomendam definir p como $\frac{1}{n+1}$ quando p=0 e como $\frac{n}{n+1}$ quando p=1 para evitar ES extremos.

Exemplo numérico

Este exemplo servirá como auxílio para entender como D é calculado a partir da equação 31. Suponha que queremos calcular o ES não-paramétrico D para o impacto de um programa de exercícios em crianças com baixa auto-estima. Nossa amostra de 40 crianças da Tabela 9 foi aleatoriamente separada em grupos controle e experimental e questionadas através do instrumento completo da Escala de Tenessee antes do grupo experimental receber a intervenção dos exercícios.

Tabela 9 - Resultados hipotéticos de escores de auto-estima para dois grupos de crianças.

	Grupo controle (n=20)	Grupo experimental (n=20)
	130	130
	131	131
	135	144
	136	146
	136	128
	138	156
	124	161
	126	162
	104	160
	142	131
	114	158
	166	166
	153	150
	169	186
	127	188
	130	153
	120	144
	121	147
	149	169
	150	170
Média	135	154
DP	16,3	17,0
Mediana	133	154,5

Se examinarmos a os escores do grupo controle na Tabela 9 podemos verificar que 18 das 20 crianças têm escores abaixo de 154,5 que é a mediana do grupo experimental, portanto p=18/20=0,90. Aplicando a equação 31 em nosso exemplo, temos:

$$D = \phi^{-1}(0,90) = 1,28 \qquad (32)$$

Podemos comparar o valor de D com o valor de d obtido a partir da equação 8 que era d=1,17 e, assim, verificamos que o ES não-

paramétrico é levemente maior que o ES paramétrico, refletindo a assimetria dos dados.

Kraemer e Andrews (1982) sugerem que se calcule ambos ES, paramétrico e não-paramétrico, pois quanto mais próxima a simetria dos dados em zero, isto é, normalmente distribuídos, mais próximos estão os ES. Na prática é difícil para os meta-analistas calcularem o ES não-paramétrico, pois normalmente não são publicadas as informações necessárias para isto.

Kraemer e Andrews (1982) também sugerem que ao calcular-se o ES não-paramétrico agrupado dos estudos independentes é melhor que os ES de cada estudo seja ponderado pelo tamanho da amostra, tal como feito para o ES paramétrico da equação 21.

6. COMENTÁRIOS SOBRE OUTROS MÉTODOS E PROGRAMAS COMPUTACIONAIS

6.1. Outras estatísticas combinadas: o caso do odds ratio e riscos relativos

Existem outros métodos de combinação de estudos, principalmente os de pesquisas clínicas, utilizados em Meta-análise que citaremos neste capítulo.

Segundo L'Abbé, Detsky e O'Rourke (1987) quando temos dados categóricos, ou especificamente variáveis dicotômicas relativas a um evento (sucesso ou fracasso) em um certo número de estudos considerados homogêneos, os resultados podem ser agregados usando a técnica de Mantel-Haenszel do risco relativo combinado, ou *odds ratio* combinado.

Esta técnica é computacionalmente mais simples e tem ampla aplicabilidade em pesquisa clínica. Proporções de sucessos são adicionadas para todos os estudos assim como estas proporções são calculadas em cada estudo quando considerados vários locais. Uma estatística qui-quadrado com um grau de liberdade pode ser usada para

julgar a significância da estatística resumo que estima o efeito do tratamento, e um erro padrão da estimativa pode ser usado para calcular um intervalo de confiança para o risco relativo. Entretanto, com o avanço computacional da estatística podemos fazer regressões logísticas, calculando o estimador agregado de máxima verossimilhança do *odds ratio*, que é uma estimativa do risco relativo.

6.2. Programas computacionais para Meta-análise

No artigo de Normand (1995) são mencionados três pacotes estatísticos comerciais: DSTAT, TRUE EPISTAT e FAST*PRO. Os programas DSTAT e TRUE EPISTAT são similares em seus contextos de Meta-análise , ambos fazendo análises somente para os contextos de efeitos fixos. Já o FAST*PRO diferencia-se, pois podemos fazer análises tanto em contextos de efeitos fixos como aleatórios.

O pacote DSTAT foi inteiramente desenvolvido para Meta-análise e através de seu manual pode-se saber como são feitos os cálculos em suas rotinas. Este pacote combina estatísticas t, estatísticas z, valores de p, estatísticas F, estatísticas qui-quadrado, coeficientes de correlação, médias amostrais (e seus desvios-padrão correspondentes) e misturas de estatísticas (por exemplo, combina estatísticas t e valores de p).

O pacote TRUE EPISTAT contém uma variedade de procedimentos para Meta-análise. Realiza todas as análises já mencionadas sobre o DSTAT e ainda fornece ao usuário a possibilidade de trabalhar com experimentos de caso controle calculando o inverso da variância ponderada do risco relativo como uma estimativa geral para estudos de caso controle.

O pacote FAST*PRO também foi desenvolvido para Meta-análise. Com o FAST*PRO é possível combinar estatísticas de testes observados ou combinar estimativas de parâmetros (por exemplo, médias, riscos relativos ou *odds ratios*). Um manual vem incluído contendo informações de instalação e instruções de procedimentos, bem como vários outros itens que são relevantes para a combinação de dados, mas as fórmulas dos cálculos que ele realiza não são mencionadas, podendo se obter estas informações através dos autores (Eddy, Hasselblad e Schachter, 1992).

Completos detalhes sobre os programas são encontrados no do artigo de Normand (1995).

7. CRÍTICAS À META-ANÁLISE

É importante que se tenha conhecimento sobre as críticas feitas à técnica de Meta-análise, pois é interessante que o leitor seja alertado sobre os possíveis problemas que podem ocorrer, identificando-os em uma leitura ou até mesmo ao se realizar uma Meta-análise. Segundo Glass et al. (1981) é possível dividir as críticas em quatro grupos:

1. Conclusões lógicas não podem ser feitas pela comparação e agregação de estudos que incluam técnicas diferentes de mensuração, formas diferenciadas de definição de variáveis ou sujeitos, porque são muito dissimilares ("Problema das maçãs e laranjas");
2. Os resultados da Meta-análise não são interpretáveis se provêm da combinação de estudos que tiveram um planejamento "ruim" ou são "não poderosos" com estudos que tiveram um planejamento "ótimo" ou são "poderosos";
3. Combinação de pesquisas publicadas com viés, no sentido de serem selecionados apenas os trabalhos com resultados significativos,

porque os de não-significância raramente são publicados, geram uma Meta-análise viesada;

4. Resultados múltiplos de um mesmo estudo são freqüentemente usados e invalidam a Meta-análise, pois os resultados não são independentes, fazendo os resultados parecerem mais fidedignos do que realmente são.

Outra crítica feita à Meta-análise é que os efeitos das interações são muitas vezes ignorados, fazendo com que as conclusões, na verdade, não sejam verdadeiras (Cook e Levinton, 1980; Slavin, 1983).

8. CONCLUSÕES

A intenção deste trabalho foi dar ao leitor uma idéia do que é a técnica de Meta-análise, pois a importância desta técnica é muito grande, e tende a aumentar cada vez mais a sua aplicabilidade, ou seja, toma-se essencial que se tenha pelo menos um conhecimento básico da técnica.

É claro que, por ser uma monografia, é um trabalho básico, que mostra parte da literatura existente sobre o assunto. Há muito mais que pode ser consultado e explorado.

O leitor interessado poderá encontrar uma vasta gama de aplicações e técnicas mais aprofundadas nas referências bibliográficas citadas.

9. REFERÊNCIAS BIBLIOGRÁFICAS

BECKER, B.J. and HEDGES, L.V. (1984). Meta-analysis of cognitive gender differences: A comment on an analysis by Rosenthal and Rubin. **Journal of Educational Psychology**. 76:583-587.

COHEN, J. (1965). Some statistical issues in psychology research. In B. Wolman (ed.) **Handbook of Clinical Psychology**. New York: McGraw-Hill.

COHEN, J. (1977). **Statistical Power Analysis for the Behavioral Sciences** (revised ed.). New York: Academic Press.

COOPER, H.M. (1979). Statistically combining independent studies: A meta-analysis of sex differences in conformity research. **Journal of Personality and Social Psychology**. 37:131-146.

DEAR, K.B.G.; BEGG, C.B. (1992). An approach for assessing publication bias prior to performing a Meta-analysis. **Statistical Science**. Vol. 7, N.2, 237-245.

EDDY, D.M.; HASSELBLAD, V.; SHACHTER, R. (1992). **Meta-analysis by the Confidence Profile Method: The Statistical Synthesis of Evidence**. New York: Academic Press.

FEDERIGHI, E.T. (1959). Extended tables of the percentage points of Student´s t-distribution. **Journal of the American Statistical Association**. 54:683-688.

FINDLEY, M. and COOPER, H. (1983). Locus of control and academic achievement: A literature review. **Journal of Personality and Social Psychology**. 44:419-427.

FISHER, R.A. (1932). **Statistical Methods for Research Workers**. London: Oliver and Boyd.

FRIEDMAN, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. **Psychological Bulletin**. 70:245-251.

GLASS, G. (1976). Primary, secondary, and meta-analysis of research. **Educational Researcher**. 5:3-8.

GLASS, G. (1977). Integrating findings: The meta-analysis of research. **Review of Research in Education**. 5:351-379.

GLASS, G. (1983). Synthetizing empirical research: Meta-analysis. In: S. A. Ward and L. J. Reed (eds.). **Knowledge Structure and Use: Implications for Synthesis and Interpretation**. Philadelphia: Temple University Press.

GLASS, G.; McGAW, B.; SMITH, M.L. (1981). **Meta-Analysis in Social Research**. Beverly Hills, CA: Sage.

GREEN, B. and HALL, J. (1984). Quantitative methods for literature review. **Annual Review of Psychology**. 35:37-53.

GUZZO, R.A.; JACKSON, S.E.; KATZELL, R.A. (1986). Meta-analysis analysis. In: L. L. Cummings & B. M. (eds). **Research in Organizational Behavior**. (Vol 9). Greenwich, CT: JAI Press.

HAASE,R.; WAECHTER, D.; SOLOMON, G. (1982). How siginificant is a significant difference? Average effect size of research in counseling. **Journal of Counseling Psychology**. 29:58-65.

HARRIS, M.J. and ROSENTHAL, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-análises. **Psychological Bulletin**. 97:363-386.

HEDGES, L. (1981). Distribution theory for Glass´s estimator of effect size and related estimators. **Journal of Educational Statistics**. 6:107-128.

HEDGES, L. (1982a). Estimation of effect size from a series of independent experiments. **Psychocological Bulletin**. 92:490-499.

HEDGES, L. (1992). Modeling publication selection effects in Meta-analysis. **Statistical Science**. Vol.7, N.2, 246-255.

HEDGES, L. and OLKIN, I. (1984). Nonparametric estimators of effect size in meta-analysis. **Psychocological Bulletin**. 96:573-580.

HEDGES, L. and OLKIN, I. (1985). **Statistical Methods for Meta-analysis**. San Diego, CA: Academic Press.

HODGES, J.L. and LEHMAN, E.L. (1962). Rank methods for combination of independent experiments in analysis of variance. **Annals of Mathematical Statistics**. 33:482-497.

HUNTER, J.E.; SCHMIDT, F.L.; JACKSON, G.B. (1982). **Meta-analysis: Cumulating Research Findings Across Studies**. Beverly Hills, CA: Sage.

HUNTER,J.E.; SCHMIDT, F.L. (1990). **Methods of Meta-analysis: Correcting Error and Bias in Research Findings**. Newbury Park, CA: Sage.

HYDE, J. (1981). How large are cognitive gender differences? **American Psychologist**. 36:892-901.

KATZ, B.M.; MARASCUILO, L.A.; McSWEENEY, M. (1985). Nonparametric alternatives for testing main effects hypotesis: A model for combining data across independent studies. **Psychocological Bulletin**. 98:200-208.

KOZIOL, J.A. and PEARLMAN, M.D. (1978). Combining independent chi-squared tests. **Journal of the American Statistical Association**. 73:753-763.

KRAEMER, H.C. (1984). Nonparametric effect size estimation: A reply. **Psychocological Bulletin**. 96:569-572.

KRAEMER, H.C. and ANDREWS, G. (1982). A nonparametric technique for meta-analysis effect size calculation. **Psychocological Bulletin**. 91:404-412.

KRAUTH, J. (1983). Nonparamatric effect size estimation: A comment on Kraemer and Andrews. **Psychocological Bulletin**. 94:190-192.

KULIK, J. (1983). Book review. **Review of G. V. Glass et al., Meta-analysis in Social Research (Sage, 1981)**. Evaluation News. 4:101-105.

KULIK, J.; KULIK, L.C.; COHEN, P.A. (1980). Effectiveness of computer-based college teaching: A meta-analysis of findings. **Review of Educational Research**. 50:525-544.

L´ABBÉ, K.A.; DETSKY, A.S.; O´ROURKE, K. (1987). Meta-analysis in clinical research. **Annals of Internal Medicine**. 107:224-233.

LIGHT, R.J. and PILLEMER, D.B. (1984). **Summing Up: The Science of Reviewing Research**. Cambridge, MA: Harvard University Press.

LITTEL, R.C. and FOLKS, J.I. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. **Journal of the American Statistical Association**. 68:193-194.

MAZZUCA, S. (1982). Does patient education in chronic disease have therapeutic value? **Journal of Chronic Diseases**. 35:521-529.

McGAW, B. and GLASS, G. (1980). Choice of the metric for effect size in meta-analysis. **American Educational Research Journal**. 17:325-337.

McGUIRE, J.; BATES, G.W.; DRETZKE, B.J.; McGIVERN, E.; REMBOLD, K.L.; SEABOLD, D.R.; TURPIN, B.M.; LEVIN, J.R. (1985). Methodological quality as a component of meta-analysis. **Educational Psychologist**. 20:1-5.

MOSTELLER, F.; CHALMERS, T.C. (1992). Some progress and problems in Meta-analysis of clinical trials. **Statistical Science**. Vol.7, N.2, 227-236.

MOSTELLER, F.M. and BUSH, R.R. (1954). Selected quantitative techniques. In: G. Lindzey (ed.) **Handbook of Social Psychology**. Vol.1. Cambridge, MA: Addison-Wesley.

NORMAND, S.T. (1995). Meta-analysis software: A comparative review. **The American Statistician**. August, Vol. 49, Nº 3. p. 298-309.

OLKIN, I. (1992). Meta-analysis: Methods for combining independent studies. **Statistical Science**. Vol. 7, N.2, 226.

ORWIN, R.G. (1983). A fail-safe N for effect size. **Journal of Educational Statistics**. 8:157-159.

ORWIN, R.G. and CORDRAY, D.S. (1985). Effects of deficient reporting on meta-analysis: A conceptual framework and reanalysis. **Psychological Bulletin**. 97:134-147.

PEARSON, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a know probability integral has probabily been drawn at random. **Biometrika**. 25:379-410.

ROSENTHAL, R (1978a). Combining results of independent studies. **Psychological Bulletin**. 65:185-193.

ROSENTHAL, R (1980a). On telling tails when combining results of independent studies. **Psychological Bulletin**. 88:496-497.

ROSENTHAL, R. (1983). Assessing the statistical and social importance of the effects of psychotherapy. **Journal of Consulting and Clinical Psychology** 51:4-13.

ROSENTHAL, R. (1984). **Meta-analytic Procedures for Social Research**. Beverly Hills, CA:Sage.

ROSENTHAL, R. and D. RUBIN (1982a). A simple, general purpose display of magnitude of experimental effect. **Journal of Educational Psychology** 74:166-169.

ROSENTHAL, R. and D. RUBIN (1982b). Comparing effect sizes of independent studies. **Psychological Bulletin**. 92:500-504.

ROSENTHAL, R. and D. RUBIN (1982c). Further meta-analitic procedures for assessing cognitive gender differences. **Journal of Educational Psychology**. 74:708-712.

ROSENTHAL, R. and ROSNOW, R.L. (1984). **Essentials of Behavioral Research**. New York: McGraw-Hill.

SMITH, M. and GLASS, G. (1980). Meta-analysis of research on class size and its relationship to attitudes and instruction. **American Educational Research Journal**. 17:419-433.

STEINKAMP, M. and MAEHR, M. (1983). Affect, ability and science achievement: A quantitative synthesis of correlational research. **Review of Educational Research**. 53:369-396.

STOCK, W.; OKUM, M.; HARING, M.; MILLER, W.; KINNEY, C.; CEURVORST, R. (1982). Rigor in data synthesis: A case study of reability in meta-analysis. **Educational Researcher**. 11(6):10-20.

STOUFFER, S.A.; SUCHMAN, E.A.; De VINNEY, L.C.; STAR, S.A.; WILLIAMS, R.M. Jr. (1949). **The American Soldier: Adjustment During Army Life**. Vol.1, Princeton, NJ: Princeton University Press.

STRUBE, M.J. (1985). Combining and comparing significance levels from nonindependent hypotesis tests. **Psychological Bulletin**. 97:334-341.

THACKER, S.B. (1988). Meta-analysis: A quantitative approach to research integration. **JAMA**. Vol.259, N.11, 1685-1689.

TRACKZ, S.M.; NEWMAN, I.; McNEIL, K. (1985). **Regression Techniques and Dependence of Data in Meta-analysis.** Apresentado no Annual Meetings of the Mid-Western Educational Research Association, Chicago.

TUKEY, J.W. (1977). **Exploratory Data Analysis.** Reading, MA: Addison-Wesley.

WINER, B.J. (1971). **Statistical Principals in Experimental Design.** New York: McGraw-Hill.

WOLF, F.M. (1986). **Meta-analysis: Quantitative Methods for Research Synthesis.** Beverly Hills: Sage Publications.

WOLF, F.M. and SPIES, C.J. (1981). Assessing the consistency of cross-lagged panel effects with the Fisher Combined Test. **Proceedings of the American Statistical Association Social Statistics Section.** 24:506-511.

WORTMAN, P. (1983). Evaluation research: A methodological perspective. **Annual Review of Psychology.** 34:223-260.

ZIEGELMAN, F.A. (1993). **Análise do Poder Estatístico.** Monografia apresentada para obtenção do grau de Bacharel em Estatística. Porto Alegre.

ANEXO 1

Percentis de não intersecção de áreas (*nonoverlap*) correspondentes a **d**

d	U_1	U_2	U_3	r
0	0,0%	50,0%	50,0%	,000
0,1	7,7	52,0	54,0	,050
0,2	14,7	54,0	57,9	,100
0,3	21,3	56,0	61,8	,148
0,4	27,4	57,9	65,5	,196
0,5	33,0	59,9	69,1	,243
0,6	38,2	61,8	72,6	,287
0,7	43,0	63,7	75,8	,330
0,8	47,4	65,5	78,8	,371
0,9	51,6	67,4	81,6	,410
1,0	55,4	69,1	84,1	,447
1,1	58,9	70,9	86,4	,482
1,2	62,2	72,6	88,5	,514
1,3	65,3	74,2	90,3	,545
1,4	68,1	75,8	91,9	,573
1,5	70,7	77,3	93,3	,600
1,6	73,1	78,8	94,5	,625
1,7	75,4	80,2	95,5	,648
1,8	77,4	81,6	96,4	,669
1,9	79,4	82,9	97,1	,689
2,0	81,1	84,1	97,7	,707
2,2	84,3	86,4	98,6	,740
2,4	87,0	88,5	99,2	,768
2,6	89,3	90,3	99,5	,793
2,8	91,2	91,9	99,7	,814
3,0	92,8	93,3	99,9	,832
3,2	94,2	94,5	99,9	,848
3,4	95,3	95,5	*	,862
3,6	96,3	96,4	*	,874
3,8	97,0	97,1	*	,885
4,0	97,7	97,7	*	,894

* Maior que 99,95

Fonte: COHEN, J. (1977). Statistical Power Analysis for the behavioral Sciences