Ciência e Natura, Santa Maria, v. 36 Ed. Especial, 2014, p. 52–57
Revista do Centro de Ciências Naturais e Exatas - UFSM
ISSN impressa: 0100-8307 ISSN on-line: 2179-460X

CIÊNCIA e NATURA 35 anos

# Meta-Análise caso a caso usando apenas funções de verossimilhança

## Meta-Analysis on sample unit level using just likelihood functions

Carlos Alberto de Bragança Pereira*[1]

[1] Departamento de Estatística, Instituto de Matemática e Estatística,
Universidade de São Paulo, São Paulo, Brasil

**Resumo**

*Este trabalho tem o objetivo de ser uma pequena resenha da Meta-análise caso a caso, a ênfase é dada a mistura de proporções. Comparamos três regiões onde são observados processos de Bernoulli em subáreas amostradas da região. Fazemos uso das verossimilhanças associadas a cada observação e mostramos que a mistura dessas verossimilhanças é mais informativa do que em considerar-se apenas uma observação da estatística suficiente para um único parâmetro. A motivação é fruto de um trabalho que realizamos com animais silvestres ao compararmos três áreas com respeito à "end-points" cito genéticos; Bueno et al. (2000) e Pereira et al. (2002). Embora utilizemos no presente trabalho os mesmos números daqueles artigos, mudamos o problema para comparar três regiões brasileiras quanto à frequência de obesos em alunos de escolas de segundo grau.*

***Palavras-chave**: Processo de Bernoulli, função de verossimilhança, distribuição binomial negativa, distribuição logística normal.*

**Abstract**

*The objective of the present work is reviewing the Meta-analysis for patient level technique. 3 regions for which Bernoulli Processes are observed are compared. The Likelihood Function of each sample is the main tool to build the implementation of the joint information of the sample of each group. This mixture likelihood may be more informative than the sufficient statistic whenever considering a solely parameter for each group. The motivation here is based on the work on wild animals when comparing cito-genetics end-points: Bueno et al. (2000) and Pereira et al. (2002). Although using the same numbers of those articles, we change the problem to the case of comparing regions respect to obesity status of students of high schools in the regions of study.*

***Keywords**: Bernoulli process, likelihood function, negative binomial distribution, logistic-normal distribution.*

*cpereira@ime.usp.br

Recebido: 18/02/2014 Revisado: 14/05/2014

## 1 Introdução

Pereira et al. (2002) apresentam um modelo hierárquico para escolha e comparação do efeito ambiental na saúde de animais silvestres. Mostrou-se neste trabalho que estudar apenas a variabilidade do processo dentro das unidades amostrais impede a consideração relevante da variabilidade entre unidades amostrais: os seres vivos estudados. Usando-se apenas o valor da estatística suficiente dentro das áreas chegava-se a conclusão de uma diferença enorme entre regiões (Bueno et al., 2000). Pereira et al. (2002) mostram que as diferenças não deveriam ser tão significantes. Entendeu-se mais tarde que este artigo apresentava de fato um modelo relevante de meta análise.

Nosso trabalho em Meta-análise foi iniciado com o grupo de pesquisas em Cirurgia Vascular sob a liderança do saudoso Dr. Maximiano Albers da USP. Em uma série de artigos nas revistas mais importantes da área (Maximiano Albers et al., 2001, 2003, 2004, 2005, 2006, 2007, Romiti et al., 2008), utilizamos o que chamamos de Meta-análise caso a caso em análise de sobrevivência. Meta-análise é o estudo estatístico das informações disponíveis na literatura sobre um determinado assunto específico. Ao adicionarmos o termo caso a caso queremos dizer que estamos nos utilizando de todas as unidades amostrais de todos os trabalhos incorporados para a análise. Isto não é comum, pois muitos dos estudos da literatura fazem uso apenas dos índices estatísticos documentados: tamanho da amostra, valor-p ou médias. Na maioria das vezes não se consegue construir as verossimilhanças (fundamental em uma análise Bayesiana) de cada estudo. Muitos confundem o termo Meta-análise com revisão sistemática. De fato a revisão sistemática é a etapa anterior ao estudo estatístico da Meta-análise. Para o estudo sistemático, o pesquisador necessita definir que trabalhos da literatura farão parte de sua análise. Este é um trabalho árduo, pois sabemos que a quantidade de artigos científicos publicados é enorme e todos devem ser analisados para se decidir quais os que devem entrar na Meta-análise final. É claro que se o pesquisador possui o seu próprio conjunto de dados este deve fazer parte do grupo de bancos dados considerados para a composição da informação total. Ponderações podem ser racionalmente construídas para se dar valor maior ou menor a cada um dos trabalhos: Um trabalho mais antigo pode receber um peso menor do que um mais recente. Muitas são as considerações objetivas e subjetivas para a definição das ponderações. Note que embora o estatístico deva fazer parte do grupo de pesquisadores que definem essas premissas, não será ele responsável isoladamente sobre qual trabalho deve ou não entrar no estudo.

Para as comparações usamos aqui as distribuições logísticas normais que estão bem discutidas por Aitchison (2003). As propriedades das distribuições discretas podem ser encontradas em Pereira & Stern (2008). Os artigos liderados pelo Dr. Albers que lidam com a Meta -análise estão listados na bibliografia. As teses de Dutton (2011) e de Martins (2013) são trabalhos que tratam de Meta-análise caso a caso para situações mais gerais e descrevem neste contexto a teoria necessária.

## 2 Um exemplo motivador

Imaginem um estudo sobre obesidade de jovens em três regiões brasileiras. Devido às diferenças culturais das três regiões, estamos interessados na comparação da cultura alimentar nestas três regiões. De fato iremos estudar o efeito dessas alimentações nos jovens estudados. Aqui iremos apenas realisar o trabalho estatístico, a ferramenta que será usada pelos pesquisadores da área alimentar. Lembremos que os dados aqui analisados, embora reais, são de um problema diferente do que se está discutindo. Dados reais para este problema alimentar ainda não estão disponíveis para pesquisadores de outras áreas.

Foram selecionadas 13 escolas de cada região e observações de frequência de obesidade foram feitas em cada escola. Por problemas operacionais uma das diretoras da primeira região não permitiu que fossem realizadas observações em sua escola. Temos assim 12, 13 e 13 escolas, respectivamente, nas regiões 1, 2 e 3. Em cada uma das escolas foram aleatoriamente sendo selecionados jovens até que fossem encontrados 100 deles com IMC acima do recomendável (falha). Contavam-se assim o número de crianças com IMC "normais" (sucesso). Para cada uma das escolas, $K=100$ são as falhas e $X$ os sucessos. O tamanho da amostra de alunos é $n = K+X$ para cada uma das escolas. Os resultados dos sucessos em cada escola estão apresentados na Tabela 1 abaixo.

Tabela 1: Número de jovens com IMC normais até a observação de 100 com IMC alterado

| Escolas | E1 | E2 | E3 | E4 | E5 | E6 | E7 | E8 | E9 | E10 | E11 | E12 | E13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Região 1 | 19 | 52 | 50 | 31 | 68 | 53 | 45 | 28 | 99 | 27 | 31 | 57 | |
| Região 2 | 86 | 31 | 83 | 58 | 31 | 90 | 175 | 124 | 64 | 105 | 120 | 39 | 27 |
| Região 3 | 37 | 131 | 100 | 90 | 56 | 45 | 63 | 46 | 27 | 45 | 37 | 98 | 65 |

Se imaginarmos a permutabilidade completa dentro de cada região poderíamos considerar que a probabilidade de um jovem estar com o IMC normal é $\pi_i$ ($i$ =1,2,3), o mesmo parâmetro para cada escola dentro de cada região $i$. A estatística suficiente seria então a frequência total de cada região. Tabela 2 apresenta estas frequências com o odds amostral de cada região.

A Figura 1 ilustra as três verossimilhanças globais, normalizadas pelas suas respectivas áreas. Como se pode notar, as diferenças entre as regiões passam a ser marcantes. Ao normalizarmos as verossimilhanças na verdade estamos considerando densidades posteriores,

consequência de usarmos distribuições uniforme a priori. Como as amostras são independentes por escolas e por regiões, podemos simplesmente calcular as probabilidades de ordenações específicas dos parâmetros. Isto é: $\Pr\{\pi_1 \leq \pi_2 | \text{dados}\} = 1$, $\Pr\{\pi_1 \leq \pi_3 | \text{dados}\} = 0{,}9996$ e $\Pr\{\pi_3 \leq \pi_2 | \text{dados}\} = 0{,}9924$. A conclusão desta análise seria que a primeira região é a menos saudável e a região dois a mais saudável.

Tabela 2: Números de jovens com IMC normais e alterados por região

| Escolas | Suc. | Falhas | ODDS |
|---|---|---|---|
| Região 1 | 560 | 1200 | 0.4667 |
| Região 2 | 1033 | 1300 | 0.7946 |
| Região 3 | 840 | 1300 | 0.6462 |

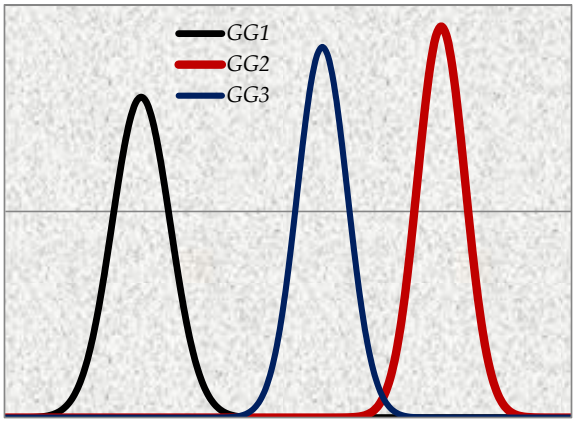


Figura 1: Verossimilhanças normalizadas dos grupos globalizados.

Mostraremos a seguir por meio de uma ilustração que esta não é uma análise adequada, pois desconsideramos toda a variabilidade existente entre as escolas. A Figura 2 é a verossimilhança do primeiro grupo juntamente com as frequências relativas de jovens normais em cada escola. Um exercício para o leitor seria o repetir a Figura 2 para os outros dois grupos e verificar que o problema se repete.

Lembremos que as verossimilhanças normalizadas são densidades beta com parâmetros $(X+1;K)$. Os valores de $K$, considerando-se o resultado global da Tabela 2, são as falhas totais em cada região e evidentemente os sucessos globais definem os primeiros parâmetros do vetor paramétrico. Por problemas de precisão, decidimos considerar o logOdds como parâmetro de trabalho. Para isso, seguimos a recomendação de Aitchison (2003) onde considera que no lugar de trabalhar com $\pi_i$, podemos trabalhar com $\theta_i = \ln[\pi_i/(1-\pi_i)]$. A única consideração feita, baseada em resultados numéricos, é a de que $\theta_i$ tem distribuição aproximadamente normal com média $\psi(X+1) - \psi(K+1)$ e variância $\psi'(X+1) + \psi'(K+1)$.

As funções $\psi$ e $\psi'$ são respectivamente a digamma e a trigamma (Pereira e Stern, 2008).

Considerando-se os resultados globais da Tabela 2, teríamos para os grupos 1, 2 e 3 as seguintes distribuições normais como verossimilhanças (a dispersão aqui é o desvio padrão):

$\theta_1 \sim$ N(-.76;.0512); $\theta_2 \sim$ N(-.23;.0417); e $\theta_3 \sim$ N(-.44;.0443)

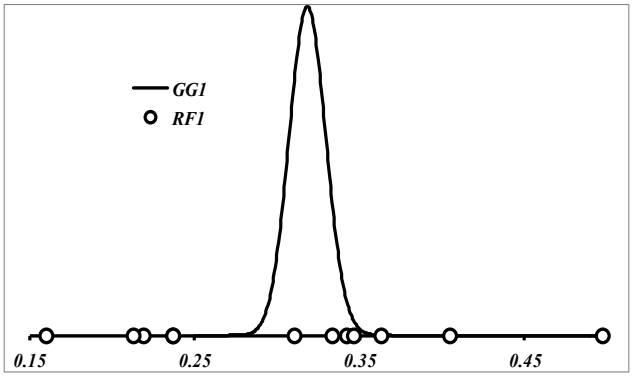


Figura 2: Verossimilhança normalizada da região 1 e frequências relativas de jovens com IMC normais em cada uma das escolas.

Note-se pela Figura 2 que a maioria das escolas apresentou frequências de jovens com IMC normais em regiões cuja verossimilhança é zero. Fica claro então que a contagem global das frequências não deve ser adequada.

## 3 Meta-análise caso a caso

Vamos agora, por meio de outra ilustração, comparar o método que ultiliza as verossimilhanças unificadas de cada região - neste caso considera-se apenas a amostra total de cada região como se fosse apenas uma escola por região - com o presente caso onde se considera uma média ponderada das verossimilhanças. A ponderação correta seria o tamanho das escolas; contudo, com o objetivo apenas de ilustração utilizamos como ponderação o tamanho da amostra observada em cada escola. Novamente tomamos a primeira região como ilustração e logo em seguida apresentamos os gráficos das médias de cada região com as novas probabilidades de ordenação.

A Figura 3 apresenta as verossimilhanças de cada escola da primeira região juntamente com a verossimilhança média multiplicada por seis. A multiplicação tem objetivo apenas artístico. A curva negra é a verossimilhança média multiplicada por seis.

Novamente deixamos para o leitor a construção de figura semelhante para as regiões 2 e 3. Figura 4 apresenta as médias das verossimilhanças de cada região e

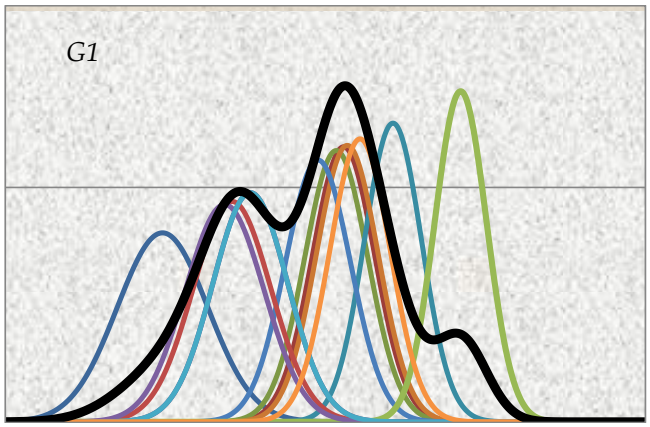


Figura 3: Verossimilhanças normalizadas das escolas e média da primeira região.

ilustra bem a indiferença entre as regiões com repeito aos logOdds das escolas. Nesta figura também consideramos a igualdade das regiões e definimos a verossimilhança global média. As probabilidades comparativas agora se transformam no seguinte:

$$\Pr\{\pi_1 \leq \pi_2|\text{dados}\} = 0{,}73,\ \Pr\{\pi_1 \leq \pi_3|\text{dados}\} = 0{,}67 \text{ e}$$
$$\Pr\{\pi_3 \leq \pi_2|\text{dados}\} = 0{,}52.$$

É claro que, como esperado, a ordem não mudou. Entretanto as diferenças já não são tão drásticas como parecia com a análise anterior.

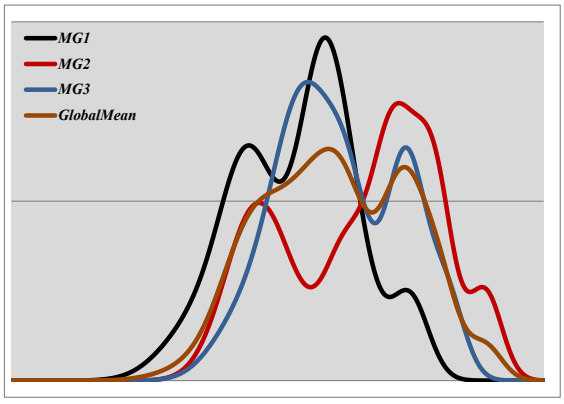


Figura 4: Médias das verossimilhanças por região e global.

Apenas para efeito de ilustração apresentamos a Figura 5 onde usamos uma aproximação para a normal da verossimilhança global média da região 1. Mostramos também para efeito comparativo a verossimilhança quando desconsideramos a variabilidade entre as escolas. Notemos que ao contrário da verossimilhança total, a verossimilhança média possui as observações em regiões de densidade positiva.

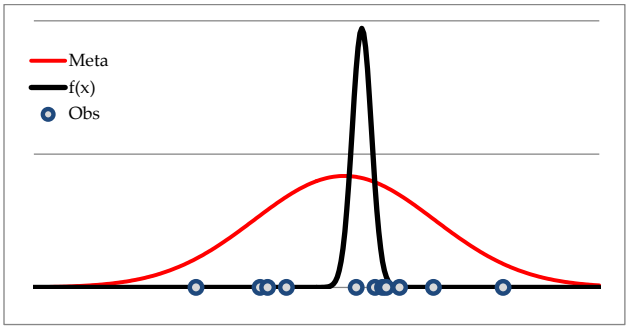


Figura 5: Média das verossimilhanças da região 1 comparada a verossimilhança total (f) desconsiderando a variabilidade entre escolas. As frequências amostrais são mostradas no eixo.

É importante ressaltar que o método aqui descrito pode ser usado em situações diversas, casos em que a variabilidade dentro de unidades amostrais básicas é alta. O exemplo mais popular é o das eleições americanas onde o estatístico pudesse usar os diversos resultados de diferentes pesquisas ao longo dos meses que antecederam as eleições. As regiões aqui seriam os estados americanos e as escolas poderiam ser equivalentes às pesquisas realizadas no estado.

Finalmente, a Figura 6 utiliza a aproximação para a normal das verossimilhanças médias das regiões e mostra uma verossimilhança global considerando todas as escolas de todas as regiões; a verossimilhança média desconsiderando regiões, como se fossem todas as escolas de uma mesma população. Calculamos o intervalo com 95% de credibilidade para a proporção de jovens saudáveis que neste caso é [0,16; 0,60]. Note que as curvas são dos logOdds mas o intervalo é para o parâmetro proporção de crianças saudáveis. Depois de construídas as densidades e calculadas as probabilidades para $\theta$, voltamos ao parâmetro de interesse, $\pi$.

O modelo estatístico utilizado neste trabalho é a binomial negativa com parâmetros K (número na amostra) e $\pi$ (proporção populacional de sucessos). Após a observação $x$ de X (número de sucessos na amostra) obtém-se a função de verossimilhança (uma função apenas de $\pi$) a qual, normalizada pela área sob a curva, se identifica a uma função de densidade Beta com parâmetros $x$+1 e K+1. Com a parametrização utilizada, $\theta = \text{logOdds} = \ln[\pi/(1-\pi)]$ considerou-se a aproximação $\theta$ ~Normal com média igual a $\psi(x+1) - \psi(K+1)$ e variância igual a $\psi'(x+1) + \psi'(x+1)$: $\psi$ e $\psi'$ são as funções digamma (derivada da função gamma) e trigamma (derivada da função digamma), respectivamente.

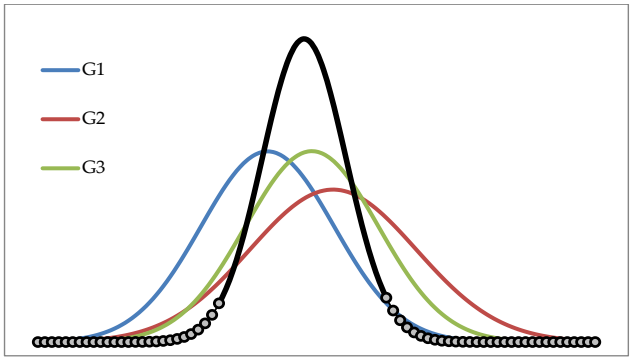


Figura 6: Aproximação para normal das verossimilhanças médias de cada região e da verossimilhança média de todas as escolas amostradas. As caudas possuem uma área de 5% nesta figura.

## 4 Considerações finais

Uma análise estatística de uma coleção de resultados analíticos, estatísticos ou descritivos, com o objetivo de integrar seus respectivos achados, se denomina (estatística) Meta-análise. Meta-análise pode ser considerada um método formal ou racional de sintetizar as informações provindas de uma grande variedade de fontes, principalmente científicas, técnicas e/ou operacionais. Muitas vezes pode ser usada para desenvolver um consenso dentro de parte da comunidade científica, e mesmo profissional. O método é particularmente útil quando há a vontade de não se ignorar muitas das evidências disponíveis.

Estudos independentes sobre a natureza de doenças podem levar à variação de conclusões através de regiões ou grupos. Em tais situações, o método de Meta-análise pode oferecer a oportunidade de reconciliação das diferenças, estimando inclusive o efeito médio do acomodamento dos estudos. São frequentemente usadas em estudos clínicos para estimar o efeito médio dos tratamentos clínicos ao longo dos estudos. Uma afirmação encontrada em vários pontos da literatura resume bem o que se entende por este importante método: "Wide -ranging treatment impact is one of the main goals of Meta-Analysis studies".

Além do uso da Meta-análise em ensaios clínicos, existe um grande potencial para seu uso em estudos observacionais. Estudos observacionais focam-se nas relações entre certos fatores de risco e na doença mais do que no impacto de um tratamento específico. A Meta-análise pode ser usada em tais situações para sintetizar as relações entre estudos, coorte ou outros grupos definidos. Embora haja este potencial, os estudos baseados em Meta-análise com dados observacionais não são atualmente prevalentes na literatura como é o caso dos ensaios clínicos.

Aos leitores deste artigo recomendo uma viagem ao site denominado Crochane Collaboration. Irão encontrar o mais rico material sobre toda a história, o uso e as técnicas desenvolvidas sobre Meta-análise. Creio que tudo que é feito em Meta-análise clássica aparece neste site: http://handbook.cochrane.org/.

## Referências

Aitchison J (2003), The statistical Analysis of Compositional Data, The Blackburn Press, Caldwell, New Jersey.

Albers M; Romiti M; Pereira CAB; Fonseca RLA; Silva-Jr AM (2001), A Meta-analysis of Infrainguinal Arterial Reconstruction in Patients with End-stage Renal Disease, Euro Journal of Vascular & Endovascular Surgery 22(4):294-300.

Albers M; Romiti N; De Luccia N; Battistello VM; Rodrigues; Pereira CAB (2003), Meta-analysis of polytetrafluoroethylene bypass grafts to infrapopliteal arteries, Journal of Vascular Surgery 37(6):1263-1269.

Albers M; Romiti M; Pereira CAB; Antonini M; Wulkan M (2004), Meta-analysis of allograft bypass grafting to infrapopliteal arteries. European Journal of Vascular & Endovascular Surgery 28(5):462-472.

Albers M; Romiti M; Brochado-Neto FC; Pereira CAB (2005), Meta-analysis of alternate autologous vein bypass grafts to infrapopliteal arteries, Journal of Vascular Surgery 42(3):449-455.

Albers M; Romiti M; Brochado-Neto FC; De Luccia N; Pereira CAB (2006), Meta-analysis of popliteal-to-distal vein bypass grafts for critical ischemia, Journal of Vascular Surgery 43(3):498-503.

Albers M; Romiti M; De Luccia N; Brochado-Neto FC; Nishimoto I; Pereira CAB (2007), An updated meta-analysis of infraquinal arterial reconstruction with patients with end-stage renal disease. Journal of Vascular Surgery 45(3):536-542.

Bueno AMS; Pereira CAB; Rabello-Gay MN (2000), Environmental genotoxicity evaluation using cytogenetic end-points in wild rodents, Environmental Health Perspectives 108:1165-1169.

Bueno AMS; Pereira CAB; Rabello-Gay MN; JM Stern (2002), Environmental Genotoxicity Evaluation: Bayesian Approach for a Mixture Statistical Model, "SERRA" Stochastic Environmental Research & Risk Assessment 16(4):267-278.

Dutton M (2011), Individual patient-level data meta-analysis: A comparison of methods for the diverse populations' collaboration data set. PhD Thesis FSU.

Higgins JPT; Green S (2011), (Cochrane Collaboration) Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0, http://handbook.cochrane.org/

Martins CB (2013), Meta-análise caso a caso sob a perspectiva Bayesiana, Tese de doutorado USP.

Pereira CE; Albers M; Romiti M; Brochado-Neto F; Pereira CAB (2006), Meta-analysis of femoropopliteal bypass grafts for lower extremity arterial insufficiency, Journal of Vascular Surgery 44(3):510-517.

Pereira CAB; Stern JM (2008), Special characterizations of standard discrete distributions, REVSTAT Statistics Journal 6:199-230.

Romiti M; Albers M; Brochado-Neto F; Durazzo A; Pereira CAB; DeLuccia N (2008), Meta-analysis of infrapopliteal angioplasty for chronic critical limb ischemia, Journal of Vascular Surgery 47(5); 975-981.