Universidade de São Paulo - USP
Universidade Federal de São Carlos - UFSCar
Universidade Estadual Paulista - UNESP

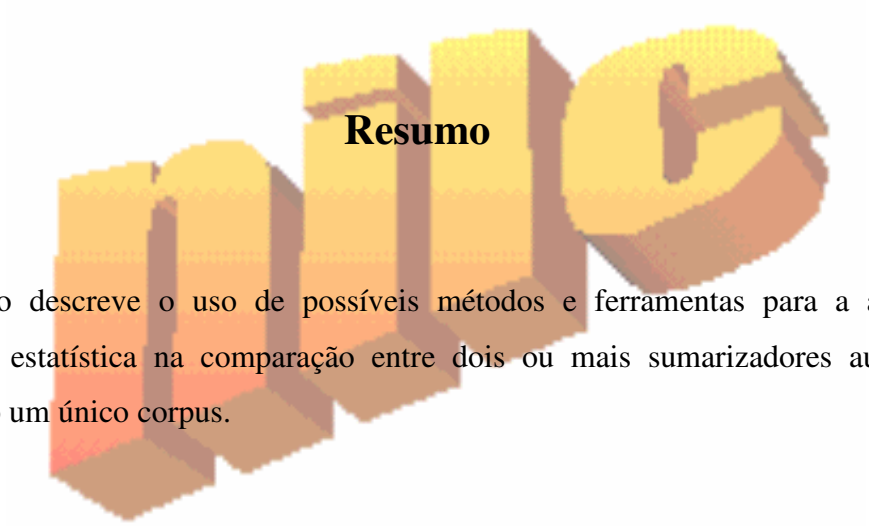
Análise de Significância Estatística na Comparação entre Sistemas de Sumarização Automática

Daniel Saraiva Leite
Lucia Helena Machado Rino

NILC-TR-01-09

Fevereiro, 2009

Série de Relatórios do Núcleo Interinstitucional de Lingüística Computacional
NILC - ICMC-USP, Caixa Postal 668, 13560-970 São Carlos, SP, Brasil



Resumo

Este relatório descreve o uso de possíveis métodos e ferramentas para a análise de significância estatística na comparação entre dois ou mais sumarizadores automáticos considerando um único corpus.

ÍNDICE

FIGURAS

TABELAS

1. Introdução

Para a comparação e a análise dos resultados de sistemas de sumarização automática pode ser útil a verificação de que as diferenças encontradas nos resultados são estatisticamente significantes. Em outras palavras, busca-se provar que as diferenças nos resultados não se devem meramente ao acaso. Em estatística, isso equivale a aplicar um teste de hipóteses para verificar se as médias indicadas pelas amostras (resultados dos sistemas, no caso) são iguais ou não.

Este relatório descreve o uso de possíveis métodos e ferramentas para a análise de significância estatística na comparação entre dois ou mais sumarizadores considerando um único corpus.

Dividiu-se a análise em dois casos:

- para a análise da significância estatística das diferenças entre pares de sistemas, deve-se seguir os passos da Seção 2.
- para a análise da significância estatística considerando, em conjunto, dois ou mais sistemas avaliados no mesmo experimento, um possível processo a ser adotado é descrito na Seção 3.

2. Teste para pares de sumarizadores

O teste para pares de sumarizadores permite analisar a significância estatística no desempenho de dois sumarizadores. Suponha que se tenha utilizado quatro sumarizadores num dado experimento (S1, S2, S3, S4). Os possíveis pares a serem analisados com o teste estatístico são: S1S2, S1S3, S1S4, S2S3, S2S4 e S3S4).

Na análise de um par de sumarizadores, têm-se duas hipóteses, então:

H0: é a hipótese nula, ou seja, a que se quer rejeitar: *As diferenças entre os dois sistemas não são significativas.*

H1: *As diferenças entre os sistemas são significativas.*

O método *t-student* emparelhado (*matched-pair t-student test*) é o mais usado para teste de significância estatística entre pares de modelos, os quais, em nosso caso, são os sumarizadores. O emparelhamento refere-se ao fato de que ambos os sumarizadores trabalham exatamente com os mesmos dados, ou seja, cada texto é sumarizado por ambos. O método deve ser obrigatoriamente usado quando o tamanho da amostra for pequeno ($n < 30$). É recomendado, de maneira geral, para a avaliação de algoritmos de aprendizado no campo do Aprendizado de Máquina, conforme aponta Mitchel (Mitchel 1997).

No contexto da avaliação de sumários (ou extratos) automáticos, as premissas do teste *t-student* emparelhado são:

i) O teste só pode ser aplicado para uma única métrica de avaliação de sumários por vez. Por exemplo, se num experimento utilizou-se a medida *Recall* ROUGE-1[1] e ROUGE-2, serão necessários dois testes. Não é possível realizar o teste para várias métricas simultaneamente. Cada teste foca em apenas uma métrica de avaliação de sumários. Vale lembrar que o resultado do teste pode ser significativo considerando uma métrica de avaliação de sumários e não significativo se considerarmos uma outra métrica.

[1] http://haydn.isi.edu/ROUGE/latest.html

ii) Assume-se uma distribuição normal da população. Isso equivale ao seguinte: Seja *X* a variável aleatória que representa a medida de avaliação (por exemplo, ROUGE-1) de todos os sumários possíveis de serem produzidos por um sistema de sumarização, ou seja, de toda a população de sumários. Então, *X* segue uma distribuição normal. Isso pode ser verificado pelo teste de Shapiro-Wilk (Shapiro and Wilk 1965), que verifica se os dados seguem uma distribuição normal. Entretanto, quando o tamanho da amostra for considerado grande ($n > 30$), o Teorema Central do Limite garante que as médias amostrais serão aproximadamente normalmente distribuídas e tenderão a uma distribuição normal à medida que o tamanho da amostra crescer. Então podemos ter uma variável original com uma distribuição muito diferente da normal, mas se tomarmos várias amostras grandes desta distribuição e então fizermos um histograma das médias amostrais, a forma se parecerá com uma curva normal. Nessa situação, o t-teste ainda produz resultados confiáveis (Kirkman 1996; Mitchel 1997) e podemos ignorar o teste de normalidade de Shapiro-Wilk.

A fórmula do teste *t-student* é a seguinte:

$$t = (\overline{A} - \overline{B}) \sqrt{\frac{n \times (n-1)}{\sum_{i=1}^{n} \left((A_i - \overline{A}) - (B_i - \overline{B}) \right)^2}} \qquad \text{(I)}$$

em que:

n é o tamanho da amostra, ou o número de pares de sumários automáticos gerados pelos sistemas A e B;

(n – 1) é o termo que recebe o nome especial de grau de liberdade do modelo;

$\overline{A}$ e $\overline{B}$ são as médias das medidas de avaliação dos sumários para os sistemas A e B, respectivamente;

A_i e B_i são as medidas de avaliação do sumário *i*, dos sistemas A e B, respectivamente.

Na aplicação do teste, pode-se proceder de duas formas: ou define-se previamente um nível de significância desejado para a tarefa sob teste (Caso I) ou busca-se um valor que indica a probabilidade de significância estatística, o chamado p-valor (Caso II).

Caso I. Define-se um nível de significância estatística, calcula-se o valor *t* usando a fórmula (I) e compara-se com o valor crítico de *t* apresentado na Tabela 1. O uso de uma tabela de valores críticos para *t* é usual em estatística para evitar o cálculo integral envolvido nas fórmulas mais gerais. Geralmente, essas tabelas são trazidas nos livros de Estatística, para alguns tamanhos de amostras; logo, não são completas. Assim, recupera-se um valor crítico (células destacadas) buscando-se a linha que corresponde ao tamanho da amostra em estudo, cujo grau de liberdade é $n - 1$, e cuja coluna é o nível de significância estabelecido. Pode ocorrer de não se encontrar o valor de t crítico mais adequado. Por exemplo, para n = 9 não temos t crítico tabelado. Nessa situação, pode-se optar por fazer uma interpolação entre os t críticos apresentados para os graus de liberdade 5 e 10, presentes na tabela, ou adotar a solução analítica do problema, explorada no Caso II.

Se o valor de t calculado pela fórmula I for maior que o valor crítico encontrado, então as diferenças são estatisticamente significantes e a hipótese nula é

rejeitada. Freqüentemente, o valor α = 0,05 é utilizado para a análise de significância, mas outros valores menos comuns para α podem ser usados, como mostra a Tabela 1.

Tabela 1 – Tabela de valores de t críticos para o teste *t-student*

Grau de liberdade n-1	**Nível de Significância (α)**			
	0,1	**0,05**	**0,025**	**0,005**
1	3,07768	6,31375	1,27062	6,36567
2	1,88562	2,91999	4,30265	9,92484
3	1,63774	2,35336	3,18245	5,84091
4	1,53321	2,13185	2,77645	4,60409
5	1,47588	2,01505	2,57058	4,03214
10	1,37218	1,81246	2,22814	3,16927
30	1,31042	1,69726	2,04227	2,75000
100	1,29007	1,66023	1,98397	2,62589
∞	1,28156	1,64487	1,95999	2,57588

Caso II. Em vez de se utilizar valores tabelados para verificar se há significância estatística, calcula-se diretamente um número que indica a probabilidade de significância, o chamado p-valor, pela seguinte fórmula em que B indica a função Beta (fórmula III) do cálculo integral. Quanto mais próximo de 0 for o p-valor, maior a probabilidade de significância.

$$p-valor = \frac{1}{\sqrt{n-1}\times \mathrm{B}\left(\frac{1}{2},\frac{n-1}{2}\right)} \int_{-t}^{t} \left(1+\frac{x^2}{n-1}\right)^{-\frac{n}{2}} dx \qquad \text{(II)}$$

$$B(x,y) = \int_0^t \left(t^{x-1}(1-t)^{y-1}\right) dt \qquad \text{(III)}$$

A fórmula II é incorporada a vários pacotes estatísticos e, assim, também no Microsoft Excel, sendo de uso bastante frequente para indicar o nível para o qual as diferenças entre os sistemas avaliados são significantes.

Exemplo numérico

Dados os resultados de avaliação de dois sistemas extrativos para um corpus de teste com 3 textos-fonte, verificaremos se as diferenças são significativas num nível de significância de 0,05 (5%), segundo o Caso I. Por fins didáticos, o tamanho da amostra é pequeno (3 extratos) e não segue uma distribuição normal. Dessa forma, não deveríamos confiar num teste t quando apenas 3 extratos forem

produzidos e não for dada nenhuma evidência de que a distribuição segue uma distribuição normal.

Tabela 2 – Dados de exemplo – teste *t-student*

Extrato	Recall ROUGE-1	
	Sistema A	Sistema B
1	0,59	0,39
2	0,58	0,44
3	0,57	0,45

Para:

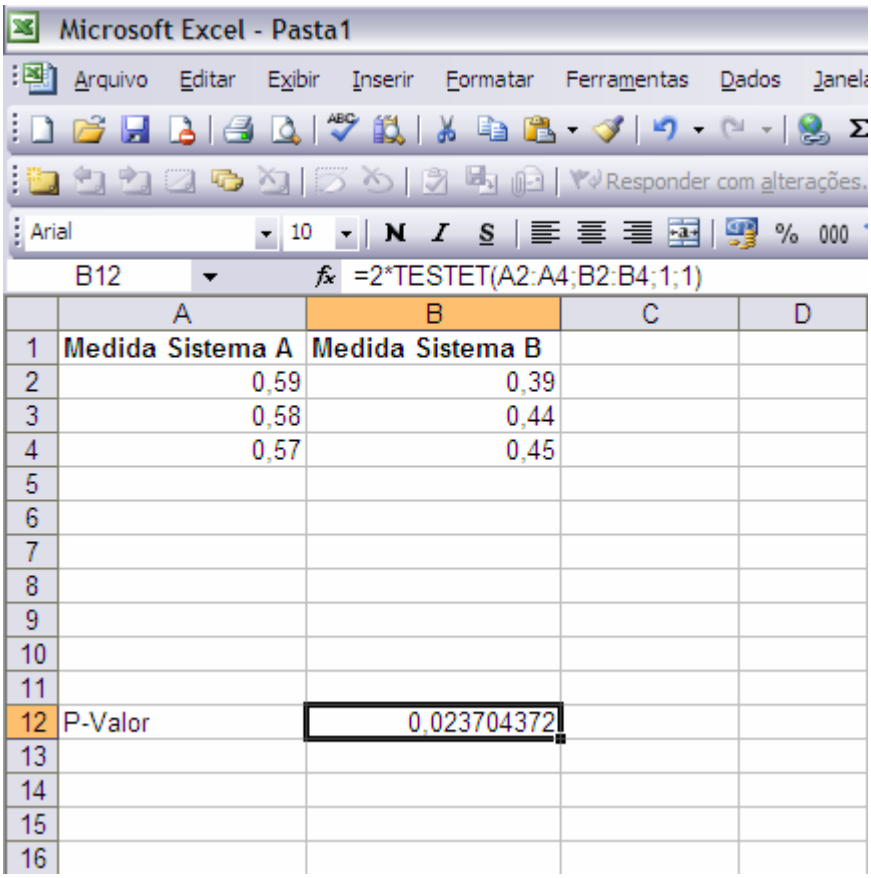
$n = 3$ (número de extratos)
$n - 1 = 2$ (graus de liberdade)
$\overline{A} = 0{,}5800$
$\overline{B} = 0{,}4266$

o valor t calculado pela Fórmula I é $\boldsymbol{t = 6{,}3790}$.

$$t = (0{,}5800 - 0{,}4266) \sqrt{\frac{3 \times (3-1)}{\begin{array}{l}((0{,}59 - 0{,}5800) - (0{,}39 - 0{,}4266))^2 + \\ ((0{,}58 - 0{,}5800) - (0{,}44 - 0{,}4266))^2 + \\ ((0{,}57 - 0{,}5800) - (0{,}45 - 0{,}4266))^2\end{array}}}$$

Buscando o valor crítico na tabela da distribuição t, vemos que ele é menor que o valor encontrado (2,9200 < 6,3790), então se rejeita a hipótese nula e conclui-se que as diferenças são estatisticamente significantes.

Podemos também calcular o p-valor (Caso II). Com o auxílio da função TESTET do Microsoft Excel, obtemos o *p-valor* **0,02370**, conforme Figura 1.



	A	B	C	D
1	**Medida Sistema A**	**Medida Sistema B**		
2	0,59	0,39		
3	0,58	0,44		
4	0,57	0,45		
5				
6				
7				
8				
9				
10				
11				
12	P-Valor	0,023704372		
13				
14				
15				
16				

Figura 1 – Teste *t-student* no Excel

3. Teste para um conjunto de sumarizadores

O teste *t-student* permite comparar pares de sistemas. Entretanto, podemos estar interessados em traduzir em um único número (p-valor) o nível de significância estatística de um experimento que avaliou conjuntamente dois ou mais sumarizadores.

É importante notar que num experimento com 10 sumarizadores, por exemplo, as diferenças entre um dado par podem não ser significantes. Mas se considerarmos o experimento como um todo, isto é, as diferenças encontradas entre os 10 sumarizadores, pode haver significância estatística.

Um possível método a ser utilizado é o ANOVA[2] para medidas repetidas (Elliot and Woodward 2006). Esse método, de maneira geral, é uma extensão do teste *t-student* emparelhado.

No contexto da avaliação de sumários (ou extratos) automáticos, as premissas do teste são as seguintes:

i) Deve-se focar em apenas uma medida por vez. Por exemplo, na medida *Recall* ROUGE-1 também. Não é possível realizar o teste para várias medidas simultaneamente. Cada teste foca em apenas uma medida de avaliação de sumários;

ii) O número de textos sumarizados por todos os sistemas deve ser o mesmo e tdo texto deve ser processado por todos os sistemas, assim como no teste emparelhado descrito na Seção 2;

[2] Sigla de ANalysis Of Variance.

iii) Esfericidade dos Dados. O teste ANOVA exige uma condição chamada de esfericidade. A esfericidade nos diz que a dependência do desempenho de cada sumarizador em relação a um texto fonte é similar para cada um dos sumarizadores. Essa condição deve ser testada pelo teste de (Mauchly 1940). Vale notar, no entanto, que esse teste atesta a não esfericidade. Se o teste de Mauchly por positivo, isto é, a esfericidade for violada, podem ser aplicadas duas correções no resultado do teste ANOVA. Essas correções visam ajustar o teste ANOVA quando a hipótese básica de esfericidade for violada. Temos dois casos, então (Elliot and Woodward 2006):

(a) Correção de Huynh-Feldt, que deve ser utilizada quando a estimativa de esfericidade (ε) for maior que 0,75;
(b) Correção de Greenhouse-Geisser, que deve ser utilizada nos demais casos.

Os cálculos necessários para o teste ANOVA para esse caso são mais complexos que o teste *t-student,* e por isso é necessário utilizar um software estatístico específico. Uma possibilidade é utilizar o software SPSS[3], através de sua opção *Analyze - General Linear Model - Repeated Measures* (Figura 2).

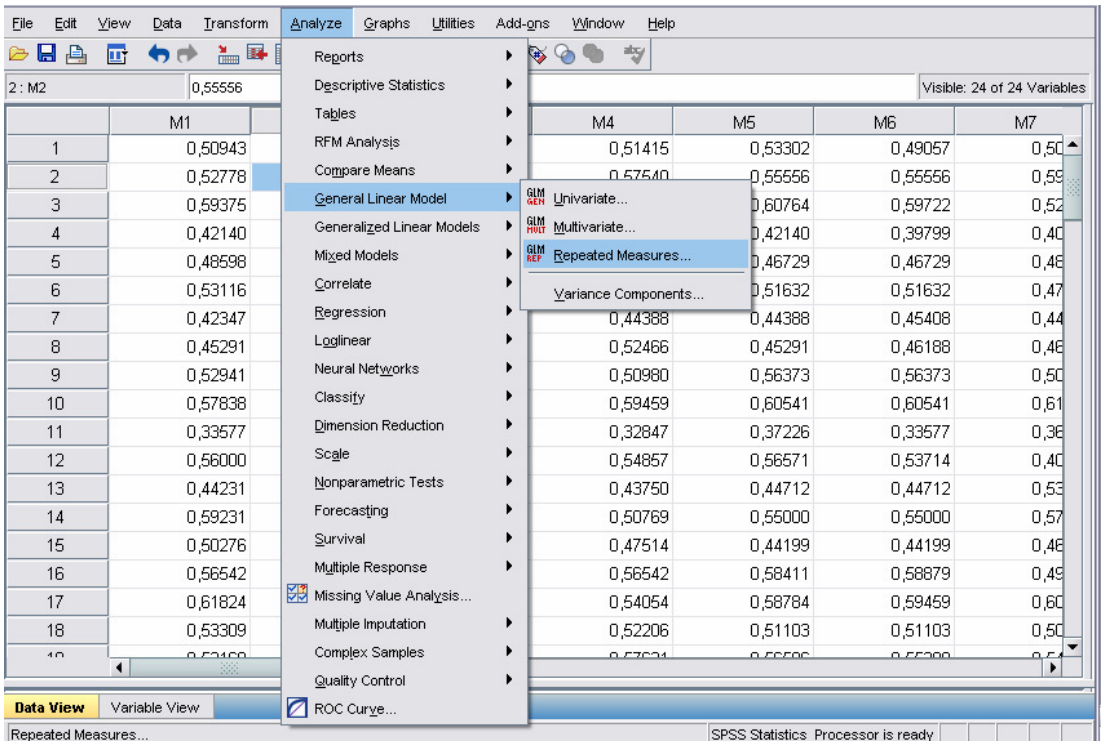


Figura 2 – Teste ANOVA para medidas repetidas no SPSS

A seguir, mostramos as saídas do software SPSS (tabelas 3, 4 e 5) para um teste considerando um experimento envolvendo 24 sumarizadores para um corpus de 100 textos.

[3] http://www.spss.com/

Tabela 3 – Saída SPSS – Teste de Esfericidade

Within Subjects Effect	[a]			
	Mauchly's W	Approx. Chi-Square	df	Sig.
ROUGE	,000	.	275	.

Tabela 4 – Saída SPSS – Cálculo do Épsilon para Determinação do Método de Correção a ser aplicado no ANOVA

Within Subjects Effect	Epsilon[a]		
	Greenhouse-Geisser	Huynh-Feldt	Lower-bound
ROUGE	,216	,228	,043

Tabela 5 – Saída SPSS – Resultado do Teste ANOVA (com métodos de correção)

Source		F	Sig.
	Sphericity Assumed	3,601	,000
	Greenhouse-Geisser	3,601	,003
	Huynh-Feldt	3,601	,003
	Lower-bound	3,601	,061

Os passos para análise dos resultados obtidos no software SPSS são descritos a seguir:

1) O primeiro passo para análise dos resultados é verificar se a esfericidade foi violada. O teste de Mauchly (Tabela 3) indica que sim, pelo fato de o nível de significância (coluna "Sig") ser representado como um ponto ("."). Também a esfericidade será considerada violada quando ele for menor que o nível aceitável (tipicamente 0,05);

2) O próximo passo deve ser considerado apenas se o passo anterior indicou não esfericidade. Na Tabela 4, devemos verificar o parâmetro épsilon para cada um dos métodos de correção (colunas "Greenhouse-Geisser" e "Huynh-Feldt"). Para ambas as colunas temos que o épsilon está abaixo de 0,75 e, portanto, deve-se considerar a correção de Greenhouse-Geisser, conforme estabelecido no item (iii) das premissas do ANOVA;

3) O último passo consiste em verificar o p-valor (coluna "Sig") dado na Tabela 5. Se a esfericidade não tiver sido violada, deve-se adotar o valor dado na linha "Sphericity Assumed". Caso contrário, deve-se adotar a respectiva correção dada no passo 2. Em nosso caso, a correção é a de Greenhouse-Geisser e fornece o p-valor de 0,003.

Como conclusão do exemplo, podemos dizer pelo teste ANOVA que, para um nível de 95% de confiança, as diferenças nas medidas de avaliação dos 24 sistemas são significantes, com p-valor igual a 0,003.

4. Observações finais

Os métodos para análise de significância estatística apresentados neste relatório são de uso geral e, portanto, aplicam-se a sistemas computacionais com diversos fins. Aqui apresentamos seu uso para a avaliação da significância estatística de sumarizadores automáticos somente. Por esse motivo, a interpretação dos resultados estatísticos tem o viés de dizer se os sumarizadores são significantemente comparáveis em relação à medida que se tem em foco.

Demos , na Seção 2, alguns exemplos da medida de informatividade (*recall,* calculada pela ferramenta ROUGE). Neste cenário, dizer que um sumarizador automático produz resultados mais informativos que os de outro sumarizador não basta, se a comparação não for estatisticamente significante. Os métodos de análise estatística, neste caso, servem para garantir a confiabilidade dos resultados da comparação, ou seja, garantir que, de fato, resultados apontados como mais informativos são relevantes para o campo de pesquisa em foco.

Referências Bibliográficas

Elliot, A. and Woodward, W. 2006. *Statistical Analysis Quick Reference Guidebook.* United States: Sage.

Kirkman, T. W. 1996. Statistics to Use. (http://www.physics.csbsju.edu/stats/, Janeiro/2009)

Mauchly, J. W. 1940. Significance Test for Sphericity of a Normal n-Variate Distribution. *The Annals of Mathematical Statistics* 11(2): 204–209.

Mitchel, T. M. 1997. *Machine Learning*. McGraw Hill.

Shapiro, S. S. and Wilk, M. B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52(3 & 4): 591-611.

Student, [William Sealy Gosset]. 1908. The probable error of a mean. *Biometrika* 6(1): 1-25.