



THIAGO VINCENZI CONRADO

**AJUSTE DO PROCEDIMENTO DE AGRUPAMENTO DE MÉDIAS
SCOTT-KNOTT PARA EXPERIMENTOS NÃO-BALANCEADOS**

LAVRAS - MG

2015

THIAGO VINCENZI CONRADO

**AJUSTE DO PROCEDIMENTO DE AGRUPAMENTO DE MÉDIAS
SCOTT-KNOTT PARA EXPERIMENTOS NÃO-BALANCEADOS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Orientador

Wilson Roberto Maluf, PhD.

LAVRAS - MG

2015

**Ficha catalográfica elaborada pelo Sistema de Geração de Ficha
Catalográfica da Biblioteca Universitária da UFLA, com dados
informados pelo próprio autor.**

Vincenzi Conrado, Thiago .

Ajuste do procedimento de agrupamento de médias Scott-Knott
para experimentos não-balanceados / Thiago Vincenzi Conrado. –
Lavras : UFLA, 2015.

58 p. : il.

Tese(doutorado)–Universidade Federal de Lavras, 2015.

Orientador(a): Wilson Roberto Maluf.

Bibliografia.

1. Scott-Knott. 2. Experimentos desbalanceados. 3. Simulações de
Monte Carlo. 4. Teste de médias. 5. SAS macro. I. Universidade
Federal de Lavras. II. Título.

THIAGO VINCENZI CONRADO

**AJUSTE DO PROCEDIMENTO DE AGRUPAMENTO DE MÉDIAS
SCOTT-KNOTT PARA EXPERIMENTOS NÃO-BALANCEADOS**

Tese apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, área de concentração em Genética e Melhoramento de Plantas, para a obtenção do título de Doutor.

Aprovada em 27 de fevereiro de 2015.

Prof. Dr. Daniel Furtado Ferreira

Prof. Dr. José Airton Nunes

Prof. Dr. Julio Sílvio de Sousa Bueno Filho

Prof. Dra. Nazaré Barata Mateus

Wilson Roberto Maluf, PhD.

Orientador

LAVRAS - MG

2015

*Aos meus pais, Vilson e Leida e,
meus irmão, Pâmella (in memorian), Priscilla, Polliana e Carlos*

DEDICO

AGRADECIMENTOS

Agradeço a Deus por tudo. Dos desafios às conquistas, incluindo as adversidades no meio do caminho que me fizeram uma pessoa melhor.

Aos meus pais, Vilson e Leida que muito me apoiaram e estimularam a sempre fazer algo a mais e também aos meus irmãos, Pâmella (in memorian), Priscilla, Pollyanna e Carlos, pelos suporte e compreensão.

À Universidade Federal de Lavras, ao Programa de Pós-Graduação em Genética e Melhoramento de Plantas, pela oportunidade de cursar o doutorado e, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Fundação de Amparo a Pesquisa de Minas Gerais, pelas bolsas concedidas durante todo o doutorado;

Ao professor Wilson Roberto Maluf; pela orientação, confiança e ensinamentos pessoais preciosos transmitidos ao longo o curso e ao professor Daniel Furtado Ferreira, que atuou decisivamente na orientação deste trabalho.

Aos professores do Programa de Pós-Graduação em Genética e Melhoramento de Plantas, pelo conhecimento transmitido.

À Marcela Carvalho Andrade, que muito me apoio oferecendo suporte integral e me ensinou muito sobre a pesquisa e vida, além de bons momentos.

A todos os amigos do GEN, pelo convívio e amizade.

Gostaria de agradecer também a todos os colegas que participaram comigo desta empreitada e, principalmente aos amigos que tornaram minha vida melhor.

Muito obrigado!

RESUMO

Este trabalho apresenta uma revisão da evolução dos procedimentos de comparação de médias mais difundidos, juntamente com o procedimento de agrupamento de médias proposto por Scott & Knott (1974) e uma solução satisfatória para uso desta análise de agrupamentos em experimentos não balanceados. Uma vez que o teste original foi desenvolvido para atuar em delineamento balanceados, a simples perda de uma parcela pode aumentar em alguns casos significativamente a taxa de Erro Tipo I. Para contornar esse controle possivelmente ineficiente das taxas de erro, na ocorrência de perda de parcelas, propomos um ajuste que deve manter proteção similar ao teste original. A solução proposta foi validada a partir de um conjunto de dados com mais de 40 milhões de experimentos simulados seguindo a metodologia de Monte Carlo. O resultado foi uma perda mínima de Poder com um controle satisfatório dos Erros Tipo I. Também foram mantidas as características originais do teste. Com o objetivo de tornar mais amigável a implementação do ajuste proposto também disponibilizamos uma macro SAS.

Palavras – chave: Erro Tipo I, médias desbalanceadas, Simulações de Monte Carlo, testes de médias, procedimento de agrupamento de médias, SAS macro

ABSTRACT

This paper presents a review about the evolution of the most broadly disseminated means comparison procedures and of the clustering means procedure proposed by Scott & Knott (1974) with a solution to use it even under unbalanced experiment designs. Since the original test was developed to be used specifically under complete designs, the loss of a single plot can increase the Type I Error Rate. To avoid inefficient Error control in the missing plot scenario we propose an adjustment that should maintain a similar protection of the original. The proposed solution was validated over more than 40 million of simulated experiments following the Monte Carlo method. As results a minimal loss of power and a satisfactory Type I Error control was observed with the original procedure features maintained. In order to make easier the implementation of the adjustment, a SAS System macro with minor hardcoding requirement is provided as well.

Keywords: Type I Error Rate, unequal number of observations, Monte Carlo Simulations, means comparisons procedures, means clustering procedures, SAS macro

SUMÁRIO

PRIMEIRA PARTE	10
1 INTRODUÇÃO GERAL.....	10
2 REFERENCIAL TEÓRICO.....	11
2.1 LSD – LEAST SIGNIFICANT DIFFERENCE	11
2.2 BONFERRONI.....	12
2.3 TUKEY	13
2.4 STUDENT-NEWMAN-KEULS - SNK	14
2.5 SCHEFFÉ	15
2.6 DUNCAN.....	15
2.7 DUNNETT	16
2.8 SCOTT-KNOTT	16
2.9 SIMULAÇÕES DE MONTE CARLO.....	22
3 CONSIDERAÇÕES GERAIS	25
4 REFERÊNCIAS.....	26
SEGUNDA PARTE – ARTIGO	33

PRIMEIRA PARTE

1 INTRODUÇÃO GERAL

Considerando que a discriminação de genótipos superiores é um dos maiores desafios de qualquer programa de melhoramento, este trabalho visa apresentar uma proposta de ajuste ao teste proposto por Scott & Knott (1974), que hoje é um dos mais utilizados no melhoramento de plantas no Brasil.

O ajuste em questão visa ampliar a aplicabilidade do teste para qualquer delineamento de experimentos não-balanceado, fornecendo ainda uma rotina de apoio para compensar a dificuldade presente no procedimento original ao ser realizado manualmente e evitar que a complexidade adicional oriunda do ajuste impeça que este atinja seus objetivos.

Inicialmente discutiremos uma visão geral dos procedimentos clássicos de comparação de médias, focando principalmente na evolução dos mesmos, juntamente com a evolução dos questionamentos de como agrupar ou separar médias.

Na sequência apresentaremos a nova proposta e em seguida os resultados de uma avaliação obtida do comparativo entre o teste original e o procedimento ajustado proposto. Esta avaliação será realizada por meio da metodologia de Monte Carlo que vem sendo utilizada para comparar os diferentes testes utilizados bem como suas características como Poder e Taxas de Erro Tipo I, sob nulidade parcial ou completa entre tratamentos.

2 REFERENCIAL TEÓRICO

Um grande passo na direção do desafio de discriminar médias em programas de melhoramento foi dado por Sir Ronald Aylmer Fisher em torno de 1920 com o desenvolvimento do conceito da análise de variância, ainda hoje, amplamente utilizado. A necessidade de desenvolver tão difundida técnica proveio dos interesses de Fisher por estatística e genética quando estudante. Seu trabalho continuou com a proposta da distribuição F que ao ser desenvolvida e tabelada por George W. Snedecor (1934) homenageou a Fisher com a letra F (Fisher, 1935; Savage, 1976).

Uma vez que um conjunto de médias tem a hipótese de igualdade entre médias ou nulidade rejeitada em dado nível de significância pela razão entre a variância dos efeitos controlados (variáveis independentes) e a variância de fatores não controlados (erro ou resíduo), faz-se necessário descobrir quais são os tratamentos com médias diferentes (considerando-se mais de dois tratamentos).

Para tanto diversos métodos estatísticos foram propostos. Estes métodos podem ser classificados como paramétricos ou não paramétricos, (conforme as restrições à natureza dos dados) independentemente de atuarem a priori ou a posteriori da análise global dos dados. Outra característica inerente a cada teste que pode ser identificada é a capacidade do teste em controlar Erros Tipo I (rejeição incorreta da hipótese de nulidade) e Tipo II (não rejeição da hipótese de nulidade quando verdadeira) em cenários específicos e consequentemente o Poder do teste, ou seja, a capacidade de rejeitar corretamente a hipótese de nulidade (Neyman & Pearson, 1933; Fischer, 1955; Mood, 1974 e Lehmann, 1993).

2.1 LSD – Least Significant Difference

Fisher (1935a), propôs o teste *Least Significant Difference* – LSD, ou da tradução livre “diferença mínima significativa”. Conforme Salkind

(2010), este teste é o primeiro teste de comparações emparelhadas entre médias e é amplamente difundido e utilizado principalmente na sequência do teste F de Snedecor (1934). O procedimento consiste basicamente em calcular a mínima diferença significativa entre duas médias como se houvesse apenas estas duas a serem testadas e caso a diferença observada seja maior do que a calculada, indicar as duas médias como diferentes no nível determinado.

Conforme Hayter (1986), o teste LSD é simplesmente uma sequência de testes t-Student's (1908) onde o desvio padrão combinado é computado para todos os grupos de médias simultaneamente ao invés do desvio padrão das duas médias testadas, isto, reduz o valor crítico e aumenta o poder do teste. Hayter (1986) também demonstra a necessidade de controlar melhor o erro cumulativo oriundo de múltiplos testes de hipóteses emparelhados (*Family Wise Error Rate - FWER*) e propõe um ajuste no procedimento para controlar e manter os Erros Tipo I abaixo dos valores nominais ao longo dos múltiplos testes.

2.2 Bonferroni

A correção que utiliza o Princípio das Iniquidades proposta por Bonferroni no teste LSD consiste em considerar p_i o valor-p para o teste H_i que será rejeitada quando $p_i \leq \alpha/m$, onde α é o nível de significância nominal do teste e m é o número de hipóteses a serem testadas, ou seja, difere do LSD apenas ajustando o valor crítico para controlar o FWER.

Simes (1986) argumenta que é possível modificar levemente o teste de Bonferroni, aumentar o Poder do mesmo e simultaneamente manter o controle do FWER, uma vez que aplicando-se apenas a correção, perde-se Poder, conforme o autor. Um exemplo é o procedimento simples proposto por Holm (1979) que inicia ordenando os p-valores a serem testados e respectivas hipóteses ($H_1, \dots, H_m$) testando até a primeira não rejeição utilizando $p_i \leq \alpha/(m + 1 - k)$, onde k é o número de hipótese sendo

testada. Por exemplo, com um $\alpha = 0,05$, $m = 6$ (6 hipóteses a serem testadas no experimento em questão) e $P_4 = 0,04$ (p-valor da quarta hipótese), temos $0,05/(6 + 1 - 4) = 0,0166$, logo $P_4 > 0,016$, não se rejeita a H_4 e nem as subsequentes (H_5 e H_6). Dessa forma utilizando-se Holm-Bonferroni, se observa um controle satisfatório do FWER levando em conta o número de hipóteses que ainda serão testadas, compensando o erro cumulativo FWER e evitando a grande perda de Poder do teste original. Para Marcus (1976) essa abordagem classifica o teste como um procedimento fechado de teste, onde a não rejeição de um hipótese implica na não rejeição das demais hipóteses compreendidas ou dependentes da primeira hipótese.

Devido à flexibilidade do teste, ou seja, características dos conjuntos de dados aos quais se pode aplicar o procedimento, várias outras propostas de ajustes foram realizadas por: Šidák (1967), Simes (1986), Holland & Copenhaver (1987), Hommel (1988) e outros. Fica clara a grande preocupação com controle do Erro Tipo I na forma do FWER e Poder do teste. Olejnik et al. (1997) realizam um estudo mais profundo comparando o controle do Erro Tipo I e Poder entre o teste original e suas modificações, indicando que as mudanças sugeridas foram em grande partes benéficas.

2.3 Tukey

Tukey (1949) propôs um teste para comparar médias e identificar diferenças significativas entre as mesmas, após a rejeição da hipótese de nulidade na ANAVA. Neste teste são realizados comparativos de todas as médias umas com as outras. O mesmo também requer pressupostos como independência das observações, distribuição normal e homogeneidade de variâncias. Embora o teste tenha um “desenho” semelhante ao LSD de Fisher, ao invés de utilizar a distribuição de *t-Student*, o autor utiliza a Distribuição da Amplitude Total Estudentizada, que é semelhante a primeira. Estas diferenças visam basicamente controlar o FWER não controlado pelo LSD de Fisher. A execução do teste é extremamente simples, calcula-se

$q_s = (\bar{Y}_B - \bar{Y}_A)/(SE)$, sendo $\bar{Y}_B$ a menor média e $\bar{Y}_A$ a maior média dentre as duas médias a serem comparadas com N observações; caso $q_s > q_{critical}$, rejeita-se a hipótese de nulidade.

De forma independente Kramer (1956) apresentou um teste semelhante ao de Tukey (1949) contendo também a possibilidade de trabalhar com amostras de tamanhos diferentes (embora as variâncias devam ser comuns).

Por esse motivo, o teste é comumente chamado de Tukey-Kramer, quando se trata de um experimento desbalanceado. Sendo que o ajuste para um número desigual de amostras proposto por Kramer (1956) foi a justificativa de diversos trabalhos para confirmarem a eficiência do mesmo (Dunnett, 1980; Hayter, 1984).

2.4 Student-Newman-Keuls - SNK

A Distribuição da Amplitude Total Estudentizada também é utilizada pelo teste SNK, assim nomeado em homenagem a Student (1908), Newman (1939) e Keuls (1952). Newman (1939), menciona o uso de testes de amplitude em conexão com a análise de variância, ideia prontamente defendida por Keuls (1952) que apresenta o procedimento levemente modificado para reduzir o FWER em experimentos com grande número de médias.

O teste inicia com a ordenação de médias e segue comparando as médias utilizando o princípio stepwise (Zar, 1999). Inicialmente são testadas a menor e a maior média do conjunto; caso a diferença seja significativa serão testadas duas médias com amplitudes nos conjuntos menores. Por exemplo, supondo-se que tenhamos seis médias no conjunto, inicialmente será testado a 1ª vs. a 6ª média; havendo diferença significativa dentro deste intervalo, seriam realizados testes entre intervalos menores, com cinco médias, ou seja, a 1ª vs. a 5ª e a 2ª vs. a 6ª média. O procedimento continua até que a hipótese de nulidade não seja rejeitada para todos os

intervalos de médias ainda não testadas, desta forma também assumindo como não significativamente diferentes todas as médias dentro do intervalo em questão tido como não significativo.

O diferencial do teste está em calcular um valor crítico para cada par de médias e não necessariamente testar todas as médias. Dessa forma o teste possui mais Poder que o teste de Tukey (1949), contudo não possui um controle eficiente do FWER, com exceção para experimentos com 3 tratamentos (Seaman, 1991). Após a discussão da eficiência do FWER por Benjamini & Hochberg (1995) que propõe uma forma diferente de observar o FWER: a teoria do controle da taxa de falsa descoberta (*False Discovery Rate* - FDR), Shaffer (2007) revisou o SNK juntamente com outros testes e afirma que o mesmo se encontra no nível intermediário quando comparado aos demais testes no que tange FDR, contudo mantendo um grande Poder.

2.5 Scheffé

Outro teste de médias muito difundido é o proposto por Scheffé (1953) que conta com a vantagem de permitir contrastes entre grupos não ortogonais e sua aplicação em regressões. Contudo, conforme o próprio autor Scheffé (1959), caso se deseje apenas os contrastes entre os tratamentos, a metodologia de Tukey (1949) deverá apresentar intervalos mais curtos, o que é mais interessante. Posteriormente, Bohrer (1967) discute os amplos intervalos e propõe uma tabela para redução dos mesmos.

2.6 Duncan

Em 1955, o estatístico David Duncan desenvolveu um procedimento para comparações múltiplas como uma modificação do teste SNK, de forma semelhante as publicações de Newman (1939) e Knels (1952), Duncan (1955), testa de forma pareada e metódica todos os contrastes entre a maior média e menor média ranqueada ainda não testada até o momento em que as diferenças entre as médias sejam não significativas.

No decorrer deste procedimento, o teste ajusta o nível de significância conforme o número de médias sendo testadas, de forma a aumentar o Poder do teste, resultando em um procedimento fácil de ser realizado e interpretado.—Carmer & Swanson (1973) ressaltam o alto Poder do teste, enquanto Einot & Gabriel (1975) questionam que este Poder vem acompanhado de um aumento nas taxas de Erro Tipo I. Duncan (1965) aborda novamente o teste de médias pela visão bayesiana, que teve ajuste proposto por Waller & Duncan (1969) e na sequência foi modificado por Shaffer (1999).

2.7 Dunnett

O teste de Dunnett (1955), foi publicado paralelamente ao teste de Duncan (1955) no mesmo ano. O teste apresenta um diferencial ao realizar comparações com um controle e ao permitir que sejam estipulados intervalos de confiança para estas comparações. Esse controle ou intervalos podem atuar tanto na abordagem bi-caudal quando na mono-caudal. Conforme Dunnett (1964), a metodologia proposta é capaz de controlar erros FWER semelhantemente aos testes de Tukey (1949) e Scheffé (1953).

A capacidade de Dunnett (1955) trabalhar focando as comparações com um controle ou intervalo determinado pelo mesmo é algo singular, semelhante a possibilidade de aplicar Scheffé (1953) a regressões. Um exemplo do uso deste artifício é a experimentação de medicamentos na fase final, onde busca-se a melhor dosagem da droga e comparar com um grupo controle para verificar qual é a menor dosagem recomendada.

2.8 Scott-Knott

O teste de agrupamento de médias Scott & Knott (1974) foi publicado pelo professor emérito da Universidade de Auckland, Nova Zelândia, Alastair John Scott e por Martin Knott, professor da London School of Economics and Political Science.

O teste proposto Scott & Knott (1974) apresenta uma filosofia diferente dos testes acima descritos de múltiplas comparações emparelhadas. Em direção oposta a identificar diferenças entre as médias individualmente, o teste visa identificar grupos de médias diferentes. Conforme metodologia de Scott & Knott (1974), doravante referenciado por “teste-SK”, utiliza-se a máxima verossimilhança para checar a existência de dois grupos homogêneos, distintos entre si dentro de um mesmo conjunto de médias e para aferir esta distinção o teste utiliza a estatística λ . Partindo das demonstrações de Hartigan (1972), Scott & Knott (1974) afirmam que por consequência a estatística λ segue assintoticamente a distribuição χ^2 , com $\frac{k}{\pi-2}$ graus de liberdade, sendo k = número de tratamentos.

Tendo em vista as satisfatórias deduções relacionadas a λ , faz-se necessário realizar a partição do conjunto de médias como em um procedimento padrão de agrupamento. Para tanto Scott & Knott (1974) utilizam a metodologia proposta Edwards & Cavalli-Sforza (1965) de agrupamentos consecutivos em árvore e também a proposta de Fisher (1958) para agrupamento de dados complexos: ordenar à priori os dados a serem agrupados por valores e seguir dividindo os novos grupos em 2. Dessa forma o número de agrupamentos possíveis seria reduzido de $2^{k-1} - 1$ para $k - 1$, o que permitiria a realização do teste manualmente. Os autores ainda mencionam, que no caso de apenas três médias a primeira partição poderá ser obtida simplesmente pelo ordenamento das médias e partição entre as duas médias adjacentes que apresentarem maior diferença entre si.

Em suma, a metodologia de agrupamento de Scott & Knott (1974) é direcional e sequencial usando a distribuição χ^2 na estatística λ para decidir se o grupo continua sendo dividido ou não. A tabela 1 ilustra o processo de agrupamento do método proposto por Scott & Knott (1974). Os dados da Tabela 1, são provenientes da altura da primeira inserção de cacho de plantas de tomate tutorado (dados do autor), provenientes de um experimento com quatro repetições e Quadrado médio do resíduo = 0,0582. Os tratamentos são

apresentados ordenados e observou-se que são realizadas sub-partições sucessivamente até o momento onde a nova partição possua apenas um indivíduo (média) no grupo, ou o valor da estatística λ seja não significativo ($\alpha = 0,10$). Repare-se que inicialmente o teste particiona entre a 2ª e 3ª média na primeira iteração (representado pela linha com traço duplo); novamente entre a 1ª e a 2ª e na sequência entre a 17ª e 18ª média, sendo que adiante a partição a 7ª e 8ª média é não significativa (representado pela linha pontilhada).

Tabela 1 - Estágios da particionamento de grupos

Id	Média	1°	2°	3°	4°	5°	6°	7°	Grupo*
01	53.83	T02	T02						A
02	49.35	T01	T01						B
03	46.48	T06		T06	T06				c
04	46.09	T15		T15	T15				c
05	45.55	T05		T05	T05				c
06	45.18	T23		T23	T23				c
07	44.90	T24		T24	T24				c
08	43.20	T25		T25	T25				C
09	42.93	T09		T09	T09				C
10	42.83	T31		T31	T31				C
11	42.73	T08		T08	T08				C
12	42.73	T32		T32	T32				C
13	42.70	T29		T29	T29				C
14	42.45	T17		T17	T17				C
15	42.38	T27		T27	T27				C
16	42.00	T20		T20	T20				C
17	41.85	T22		T22	T22				C
18	41.20	T07		T07		T07	T07		d
19	40.99	T12		T12		T12	T12		d
20	40.93	T04		T04		T04	T04		d
21	40.23	T16		T16		T16	T16		d
22	39.85	T30		T30		T30	T30		d
23	39.75	T28		T28		T28	T28		d
24	39.73	T21		T21		T21	T21		d
25	38.68	T13		T13		T13	T13		D
26	38.58	T03		T03		T03	T03		D
27	38.50	T10		T10		T10	T10		D
28	36.78	T19		T19		T19		T19	e
29	36.55	T18		T18		T18		T18	e
30	36.54	T14		T14		T14		T14	e
31	35.18	T26		T26		T26		T26	E
32	34.39	T11		T11		T11		T11	E

* médias com letras diferentes representam médias significativamente diferentes para o teste de Scott-Knott (1974), com $\alpha = 0,10$, e letras minúsculas enfatizam a provável partição que não significativa com nível de confiança selecionado.

Faz-se notável que a atribuição de letras diferentes indicam a diferenciação da média no nível selecionado. As letras são atribuídas após todas as possíveis partições houver sido realizadas. Nota-se também que este particionamento é semelhante a processo *stepwise* utilizado por Holm (1979).

É neste processo de particionamento que a modificação de Bhering et al. (2008) atua. Na proposta destes autores, após as sucessivas partições significativas atingirem apenas um membro no grupo, ou a nova partição ser qualificada como não significativa, reagrupam-se todas as demais médias que ainda não tem um grupo final definido e se reinicia o processo. Na prática, conforme proposto por Bhering et al. (2008), no exemplo em questão, após separar a primeira média da segunda, as médias entre a segunda e última média comporiam o novo conjunto a ser particionado. Embora esta solução reduza o número de agrupamento, esta também aumenta a homogeneidade entre grupos e permite a confecção de um algoritmo computacional menor.

Conforme Carmer & Walker (1985), procedimentos de agrupamento de médias como o proposto por Scott & Knott (1974) são atrativos a diversos pesquisadores por resultar em grupos distintos mutuamente excludentes sem nenhuma sobreposição, ao contrário dos procedimentos de múltiplas comparações emparelhadas.

De acordo com Willavise (1980), a primeira sugestão do uso do teste-SK para as ciências agrárias foi realizada pelo professor de horticultura Chew (1977). Gates & Bilbro (1978), apresentam uma versão do teste-SK mais detalhada do que o artigo original, comparando os resultados com o teste de Duncan (1955). No artigo os dois autores ressaltam a inexistência de um teste “perfeito” conforme apresentado na publicação de O’Neil & Wetherill (1971). Gates & Bilbro (1978) também reconhecem a dificuldade em calcular manualmente o teste e disponibilizam uma rotina para cálculo do mesmo em Fortran. Por estes motivos é comum encontrar artigos que utilizam o teste-SK e também mencionam a publicação de Gates & Bilbro (1978) (Stahlman & Phillips, et al., 1979; Carmer & Lin, 1983; Conradie, et al., 1992; Manici, et al., 1997; Wood, et al., 1998; Bisognin, et al., 2002.; Jelihovschi, et al., 2014).

Willavise et al. (1980) compararam o teste-SK, o teste LSD protegido pelo teste-F juntamente com outros três procedimentos aglomerativos (*single linkage, complete linkage, unweighted pair of group averages*). Partindo das comparações acima mencionadas, o autor indicou que todos os procedimentos aglomerativos apresentam taxas de Erro Tipo I maiores do que as nominais quando comparado ao LSD, recomendando ainda cautela na substituição dos testes de média clássicos por procedimentos que formam grupos.

Silva (1999), avaliaram o teste-SK, com o Waller-Duncan, SNK-modificado, SNK, LSD, Duncan e Tukey simulando 256.000 experimentos utilizando o método de Monte Carlo. Os autores encontraram grande superioridade do teste-SK sobre os procedimentos de comparações múltiplas emparelhas. Sob o cenário de nulidade parcial (onde em um mesmo experimento co-existem verdadeiramente a rejeição e não-rejeição da hipótese de nulidade) a taxa de Erro Tipo I foi superior à nominal, contudo também houve um controle satisfatório em cenários de nulidade completa (experimentos onde a hipótese de nulidade não deve ser verdadeiramente rejeitada), inclusive nas taxas por experimento. Santos (2001) realizando 960.000 simulações utilizando o método de Monte Carlo comparou Tukey (1949) com os métodos originais e variantes de Bonferroni, Scheffé e Scott-Knott. A partir dos resultados, os autores concluíram que o teste desenvolvido por Scott & Knott (1974) controla adequadamente os erros por comparação e experimento sobre H_0 completa, ou seja, a hipótese de nulidade é verdadeira para todas as médias do experimento. Entretanto sobre H_0 parcial, onde alguns tratamentos verdadeiramente diferem, o teste manteve o poder, contudo sem controlar adequadamente as taxas de Erro Tipo I.

Conforme Santos et al. (2001) ainda não existe teste de comparações adequado para todas as situações e o uso dos mesmos deve depender principalmente do tipo de erro especificamente tolerado pela pesquisa em

questão. Segundo os mesmos autores, o teste de Scheffé modificado apresentou um controle satisfatório do Erro Tipo I em experimentos com até três grupos, incluindo um aumento de Poder à medida que o nível de significância é aumentado, uma vez que o valor real dos mesmos é bem abaixo do nominal. Outra conclusão importante dos autores é que as Taxas de Erro Tipo I mudam conforme o número de tratamentos avaliados.

O teste-SK quando comparado ao teste de Tukey e SNK, sob distribuições normais e não normais, apresenta-se como mais robusto sob nulidade completa, por controlar satisfatoriamente Erros Tipo I por experimento. Em contrapartida, em nulidade parcial, o teste-SK não foi capaz de controlar de Erros Tipo I por comparação e por experimento, ao passo que Tukey e SNK não controlaram apenas por experimento na distribuição log-normal. Todavia, o teste-SK deve ser considerado mais poderoso que Tukey e SNK e é também robusto à violação do pressuposto de normalidade (Borges & Ferreira, 2003).

Tendo em vista as informações expostas sobre o método proposto por Scott & Knott (1974), vale ressaltar a sua grande aplicabilidade na área do melhoramento de plantas. Como o teste possui alto Poder, o mesmo é capaz de separar satisfatoriamente os tratamentos que são realmente diferentes evitando Erros Tipo II, um dos maiores desafios na seleção de novos cultivares para o avanço de gerações, por exemplo. Nesta situação, como considera-se que o material selecionado passará por novos ciclos de avaliação e caso sua seleção trate-se de um Erro Tipo I, esta falha do teste será compensada pelo novo ciclo de avaliação. Esta observação reafirma a conclusão de Santos et al. (2001) que afirma que a aplicação ou não do método depende do tipo de erro tolerado pela pesquisa.

2.9 SIMULAÇÕES DE MONTE CARLO

A primeira menção pública documentada do método de Monte Carlo, com um detalhamento profundo das justificativas para o uso do

método em vista das possíveis alternativas no campo da física, é encontrada no artigo de Metropolis & Ulam (1949). Em síntese, o autor justifica o uso do método como uma opção para solucionar problemas que outras metodologias têm dificuldade em resolver, seja pela falta de informações *a priori*, ou pelo possível número de variações não esperadas de uma variável, ou ainda para evitar um grande volume de trabalho para solucionar um problema pela metodologia tradicional. Conforme Metropolis & Ulam (1949), o método consiste em gerar uma grande sequência de conjuntos de observações que partem de um conjunto inicial já conhecido, ou gerar valores a partir de um conjunto de parâmetros determinados com distribuição igual àquela que governa o parâmetro em questão.

Gass & Assad (2005) apontam que os experimentos de Buffon (1733) e de Galton (1890) são exemplos de uma abordagem idêntica ao Método de Monte Carlo, contudo anterior às bases do método. Os autores afirmam que o método em si tem suas raízes na Segunda Guerra Mundial e utilizou como fundamentos a crescente computação científica e a técnica desenvolvida por Tippet (1925, 1927) para a produção de valores “aleatórios”. Eckhardt (1987) menciona que os primeiros questionamentos que deram origem ao método de Monte Carlo datam de 1946, e proveem de conjecturas de Stan Ulam sobre soluções do jogo de “*Solitaire*”. Estes questionamentos estatísticos de Ulam foram compartilhados posteriormente com John von Neumann e deram origem aos primeiros rascunhos do método, que foi publicado internamente nos Laboratórios de Los Alamos pela primeira vez por Everett & Ulam (1948). Após a primeira publicação, a difusão e uso do método de Monte Carlo cresceram vertiginosamente até meados de 1950 (Gass & Assad, 2005).

Ao se observar os trabalhos publicados, nota-se que a dimensão dos conjuntos de dados simulados vem crescendo juntamente com o desenvolvimento de processadores mais rápidos. Contudo, a qualidade da

implementação dos algoritmos ainda tem papel fundamental e é capaz de capitalizar ganhos semelhantes aos expostos por Fisher (1958).

No início da década de 70 foram publicados os trabalhos-marco por utilizar o Método de Monte Carlo para avaliar as características de métodos de comparação múltiplas emparelhadas (Boardman & Moffitt, 1971; Carmer & Swanson, 1971).

Embora as simulações utilizem intensivamente geradores de valores aleatórios, é possível notar a estabilidade nos resultados obtidos ao longo de diversos trabalhos utilizando a metodologia (Carmer & Swanson, 1971; Carmer & Swanson, 1973; Einot & Gabriel, 1975; Silva et al., 1999; Santos et al., 2001; Borges & Ferreira, 2003; Girardi et al., 2009).

3 CONSIDERAÇÕES GERAIS

Conforme apresentado anteriormente, os procedimentos mais difundidos de comparações entre médias apresentam ao longo do tempo um grande número de revisões, validações e muitas vezes modificações.

É observável que a evolução dos procedimentos revisados advém muitas vezes de novas concepções como o controle do Erro Tipo I por experimento, que era inexistente quando o teste LSD foi proposto. Em outros momentos, a evolução é possibilitada pelo desenvolvimento de uma ferramenta ou difusão da mesma, como é o caso dos computadores. Desta forma, rever procedimentos já exaustivamente utilizados permite a consolidação e melhoria dos mesmos e traz aos pesquisadores mais segurança no uso dos mesmos.

Conforme revisto, apesar das recomendações divergentes nos trabalhos revisados, observa-se características diferenciadas e muitas vezes exclusivas dentre os métodos disponíveis para a diferenciação de médias (Carmer & Swanson, 1971; Girardi et al., 2009). Isto deixa claro que não existe um procedimento perfeito quando se avalia em conjunto o Poder e as Taxas de Erro Tipo I (Borges & Ferreira, 2003). Portanto, o adequado é que o autor, ao planejar os experimentos a serem realizados, faça uma análise cautelosa da sensibilidade da variável ao cometer Erros Tipo I frente ao Poder necessário para atingir os objetivos de pesquisa e a partir desta análise realize a escolha do método a ser utilizado.

4 REFERÊNCIAS

- BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal Statistical Society**, London, v. 57, n. 1, p. 289-300, 1995.
- BHERING, L. L. et al. Alternative methodology for Scott-Knott test. **Crop Breeding And Applied Technology**, New York, v. 8, n. 1, p. 9-16, 2008.
- BISOGNIN, D. A. et al. Half-sib progeny evaluation and selection of potatoes resistant to the US8 genotype of *Phytophthora infestans* from crosses between resistant and susceptible parents. **Euphytica**, Wageningen, v. 125, n. 1, p. 129-138, 2002.
- BOARDMAN, T. J.; MOFFITT, D. R. Graphical Monte Carlo Type I error rates for multiple comparison procedures. **Biometrics**, Washington, v. 27, n. 3, p. 738- 744, 1971.
- BOHRER, R. On sharpening scheffe bounds. **Journal of the Royal Statistical Society**, London, v. 29, n. 1, p. 110-114, 1967.
- BORGES, L. C.; FERREIRA, D. F. Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normal e não normais dos resíduos. **Revista de Matemática e Estatística**, Marília, v. 21, n. 1, p. 67-83, 2003.
- BUFFON, G. L. C. Solution de problèmes qui regardaient le jeu de franc carreau. In: _____. **Histoire de l' académie royale des sciences de Paris**. Paris: Académie Royale des Sciences, 1733. p. 43-45.
- CARMER, S. G.; LIN, W. T. Type I error rates for divisive clustering methods for grouping means in analysis of variance. **Communications in Statistics-Simulation and Computation**, New York, v. 12, n. 4, p. 451-466, 1983.
- CARMER, S. G.; SWANSON, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. **Journal of the American Statistical Association**, New York, v. 68, n. 341, p. 66-74, 1973.
- CARMER, S. G.; SWANSON, M. R. Detection of differences between means: a Monte Carlo study of five pairwise multiple comparison procedures. **Agronomy Journal**, Madison, v. 63, n. 6, p. 940-945, 1971.

CARMER, S. G.; WALKER, W. M. Pairwise multiple comparisons of treatment means in agronomic research. **Journal of Agronomic Education**, Madison, v. 14, n. 1, p. 19-26, 1985.

CHEW, V. **Comparisons among treatment means in an analysis of variance**. Washington: Department of Agriculture, Agricultural Research Service, 1977. 64 p.

CONRADIE, E.; SWART, W. J.; WINGFIELD, M. J. Susceptibility of *Eucalyptus grandis* to *Cryphonectria cubensis*. **European Journal of Forest Pathology**, Hamburg, v. 22, n. 5, p. 312-315, 1992.

DUNCAN, D. B. A Bayesian approach to multiple comparisons. **Technometrics**, Washington, v. 7, p. 171-222, 1965.

DUNCAN, D. B. Multiple range and multiple F tests. **Biometrics**, Washington, v. 11, p. 1-42, 1955.

DUNNETT, C. W. A multiple comparison procedure for comparing several treatments with a control. **Journal of the American Statistical Association**, New York, v. 50, n. 272, p. 1096-1121, 1955.

DUNNETT, C. W. New tables for multiple comparisons with a control. **Biometrics**, Washington, v. 20, n. 3, p. 482-491, 1964.

DUNNETT, C. W. Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. **Journal of the American Statistical Association**, v. 75 n. 372, p. 789-795, 1980.

ECKHARDT, R.; ULAM, S.; von NEUMANN, J. The Monte Carlo method. **Los Alamos Science**, Los Alamos, 1987. Special Issue. Disponível em: <<https://fas.org/sgp/othergov/doe/lanl/pubs/00326867.pdf>>. Acesso em: 22 jan. 2015.

EDWARDS, A. W. F.; CAVALLI-SFORZA, L. L. A method for cluster analysis. **Biometrics**, Washington, v. 21, n. 2, p. 362-375, 1965.

EINOT, I.; GABRIEL, K. R. A study of the powers of several methods of multiple comparisons. **Journal of the American Statistical Association**, New York, v. 70, n. 351, p. 574-583, 1975.

EVERETT, C. J.; ULAM, S. Multiplicative systems. **Proceedings of the National Academy of Sciences of the United States of America**, Washington, v. 34, n. 8, p. 403, 1948.

FISHER, R. A. Statistical methods and scientific induction. **Jornal of the Royal Statistical Society**, Oxford, v. 17, n. 1, p. 69-78, 1955.

FISHER, R. A. **Statistical methods for research workers**. Edinburgh: Oliver and Boyd, 1935a. 239 p.

FISHER, R. A. **The design of experiments**. Edinburgh: Oliver and Boyd, 1935. 252 p.

FISHER, R. A. **The genetical theory of natural selection**. Oxford: The Clarendon, 1930. 466 p.

FISHER, W. D. On grouping for maximum homogeneity. **Journal of the American Statistical Association**, New York, v. 53, n. 284, p. 789-798, 1958.

GALTON, F. Dice for statistical experiments. **Nature**, London, v. 42 p. 13–14, 1890.

GASS, S. I.; ASSAD, A. A. Model world: tales from the time line—the definition of or and the origins of Monte Carlo simulation. **Interfaces**, Providence, v. 35, n. 5, p. 429-435, 2005.

GATES, C. E.; BILBRO, J. D. Illustration of a cluster analysis method for mean separation. **Agronomy Journal**, Madison, v. 70, n. 3, p. 462-465, 1978.

GIRARDI, L. H.; CARGNELUTTI FILHO, A.; STORCK, L. Type I error and power of five multiple comparison procedures for means. **Revista Brasileira de Biometria**, Jaboticabal, v. 27, n. 1, p. 23-36, 2009.

HARTIGAN, J. A. Direct clustering of a data matrix. **Journal of the American Statistical Association**, New York, v. 67, n. 337, p. 123-129, 1972.

HAYTER, A. J. The Maximum Familywise error rate of Fisher's Least significant difference test. **Journal of the American Statistical Association**, New York, v. 81, n. 396, p. 1000–1004, 1986.

HOLLAND, B. S.; COPENHAVER, M. D. an improved sequentially rejective Bonferroni Test Procedure. **Biometrics**, Washington, v. 43, n. 2, p. 417-423, 1987.

HOLM, S. A simple sequentially rejective multiple test procedure. **Scandinavian Journal of Statistics**, Stockholm, v. 6, n. 2, p. 65–70, 1979.

HOMMEL, G. A stagewise rejective multiple test procedure based on a modified Bonferroni test. **Biometrika**, London, v. 75, n. 2, p. 383-386, 1988.

JELIHOVSCHI, E. G.; FARIA, J. C.; ALLAMAN, I. B. ScottKnott: a package for performing the Scott-Knott clustering algorithm in R. **Tema**, Brasília, v. 15, n. 1, p. 3-17, 2014.

KEULS, M. The use of the “studentized range” in connection with an analysis of variance. **Euphytica**, Wageningen, v. 1, n. 2, p. 112-122, 1952.

KRAMER, C. Y. Extension of multiple range teste to group means with unequal number of replications. **Biometrics**, Washington, v. 12, p. 307-311, 1956.

LEHMANN, E. L. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? **Journal of the American Statistical Association**, New York, v. 88, n. 424, p. 1242-1249, 1993.

MANICI, L. M.; LAZZERI, L.; PALMIERI, S. In vitro fungitoxic activity of some glucosinolates and their enzyme-derived products toward plant pathogenic fungi. **Journal of Agricultural and Food Chemistry**, Easton, v. 45, n. 7, p. 2768-2773, 1997.

MARCUS, R.; PERITZ, E.; GABRIEL, K. R. On closed testing procedures with special reference to ordered analysis of variance. **Biometrika**, London, v. 63, n. 3, p. 655–660, 1976.

METROPOLIS, N.; ULAM, S. The Monte Carlo method. **Journal of the American Statistical Association**, New York, v. 44, n. 247, p. 335-341, 1949.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. New York: J. Wiley & Sons, 1974. 564 p.

NEWMAN, D. The distribution of range in samples from a normal population expressed in terms of an independent estimate of standard deviation. **Biometrika**, London, v. 31, p. 20-30, 1939.

NEYMAN, J.; PEARSON, E. S. On the problem of the most efficient tests of statistical hypotheses. **Philosophical Transactions of the Royal Society**, London, v. 231, p. 289–337, 1933.

O'NEILL, R.; WETHERILL, G. B. The present state of multiple comparison methods. **Journal of the Royal Statistical Society**, London, v. 33, p. 218-250, 1971.

OLEJNIK S. et al. Multiple testing and statistical power with modified Bonferroni procedures. **Journal of Educational and Behavioral Statistics**, Washington, v. 22, n. 4, p. 389-406, 1997.

SALKIND, N. J. **Encyclopedia of research design**. Thousand Oaks: Sage, 2010. 326 p.

SANTOS, C.; FERREIRA, D. F.; BUENO FILHO, J. S. S. Novas alternativas de testes de agrupamento avaliadas por meio de simulação Monte Carlo. **Ciência e Agrotecnologia**, Lavras, v. 25, n. 6, p. 1382-1392, 2001.

SAVAGE, L. J. On rereading R A Fisher. **The Annals of Statistics**, Philadelphia, v. 4, n. 3, p. 441-500, 1976.

SCHEFFÉ, H. A method for judging all contrasts in the analysis of variance. **Biometrika**, London, v. 40, n. 1, p. 87-110, 1953.

SCHEFFÉ, H. **The analysis of variance**. New York: J. Wiley & Sons, 1959. 459 p.

SCOTT, A. J.; KNOTT, M. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, Washington, v. 30, n. 3, p. 507-512, 1974.

SEAMAN, M. A.; LEVIN, J. R.; SERLIN, R. C. New Developments in pairwise multiple comparisons: Some powerful and practicable procedures. **Psychological Bulletin**, Washington, v. 110, n. 3, p. 577-586, 1991.

SHAFFER, J. P. Controlling the false discovery rate with constraints: The Newman-Keuls test revisited. **Biometrical Journal**, Berlin, v. 47, p. 136-143, 2007.

SHAFFER, P. J. A semi-Bayesian study of Duncan's Bayesian multiple comparison procedure. **Journal of Statistical Planning and Inference**, Amsterdam, v. 82, n. 1, p. 197-213, 1999.

ŠIDÁK, Z. K. Rectangular confidence regions for the means of multivariate normal distributions. **Journal of the American Statistical Association**, New York, v. 62, n. 318, p. 626-633, 1967.

SILVA, E.C.; FERREIRA, D. F.; BEARZOTTI, E. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. **Ciência e Agrotecnologia**, Lavras, v. 23, n. 3, p. 687-696, 1999.

SIMES, R. J. An improved Bonferroni procedure for multiple tests of significance. **Biometrika**, London, v. 73, n. 3, p. 751-754, 1986.

SNEDECOR, G. W. **Calculation and Interpretation of Analysis of Variance and Covariance**. Ames: Collegiate, 1934. 96 p.

STAHLMAN, P. W.; PHILLIPS, W. M. Inhibition of glyphosate phytotoxicity. **Weed Science**, Champaign, v. 27, n. 5, p. 575-577, 1979.

STUDENT. The probable error of a mean. **Biometrika**, London, v. 6, n. 1, p. 1-25, 1908.

TIPPETT, L. H. C. On the extreme individuals and the range of samples taken from a normal population. **Biometrika**, London, v. 17, p. 364-387, 1925.

TIPPETT, L. H. C. **Random sampling numbers**. Cambridge: Cambridge University, 1927. (Tracts for computers, 15).

TUKEY, J. **Comparing Individual Means in the Analysis of Variance**. **Biometrics**, Washington, v. 5, n. 2, p. 99-114, 1949.

WALLER, R. A.; DUNCAN, D. B. A Bayes rule for the symmetric multiple comparisons problem. **Journal of the American Statistical Association**, New York, v. 64, n. 328, p.1484-1503, 1969.

WILLAVISE, S. A.; CARMER, S. G.; WALKER, W. M. Evaluation of cluster analysis for comparing treatment means. **Agronomy Journal**, Madison, v. 72, n. 2, p. 317-320, 1980.

WOOD, B. W.; GRAUKE, L. J.; PAYNE, J. A. Provenance variation in pecan. **Journal of the American Society for Horticultural Science**, Alexandria, v. 123, n. 6, p. 1023-1028, 1998.

ZAR, J. H. **Biostatistical analysis**. New Jersey: Prentice Hall, 1999. 994 p.

SEGUNDA PARTE – ARTIGO

Artigo formatado conforme as normas da revista
Crop Breeding and applied Biotechnology

Adjusting the Scott-Knott cluster analyses for unbalanced designs

Thiago Vincenzi Conrado^{1*}, Daniel Furtado Ferreira¹, Carlos Alberto Scapim², Wilson Roberto Maluf¹

* tconrado@hotmail.com;

¹ Universidade Federal de Lavras, Av. Doutor Sylvio Menicucci, 1001 – Jardim Kennedy, Lavras - MG, 37200-000 +55 (35) 3829-1342

² Avenida Colombo, 5790 - Jardim Universitário, Maringá - PR, 87020-900

Abstract—The Scott-Knott cluster analysis is an alternative approach to mean comparisons with high Power and no subset overlapping. It is well suited for the statistical challenges in agronomy associated with testing new cultivars, crop treatments or methods. The original Scott-Knott test was developed to be used under balanced designs, therefore, the mere loss of a single plot can significantly increase the rate of Type I Error. In order to avoid Type I Error inflation from missing plots, we propose an adjustment that maintains a similar power from the original test while adding error protection. The proposed adjustment was validated from more than 40 million simulated experiments following the Monte Carlo method. The results indicate a minimal loss of power with a satisfactory

Type I Error control while keeping the original procedure features. A user-friendly SAS macro is provided for this analysis.

Keywords: Type I Error rate, unequal number of observations, Monte Carlo simulations, means clustering procedures, SAS macro

Introduction

A common problem in plant breeding is comparison of new genetic combinations. In order to detect significant difference among treatments, several Multiple Comparisons Procedures (MCP) were developed: LSD (Fisher 1935), Tukey (1949), SNK (Student 1908, Newman 1939, Keuls 1952), Scheffé (1953), and Duncan (1955). Nonetheless, all these procedures can result in groups overlapping, where one treatment ends up belonging to two or more groups simultaneously (Calinski and Corsten 1985). This behavior usually prevents a clear split of the whole set in two or more groups of treatments and also induces to a more complex simultaneous analysis of multiple variables due to the presence of overlapping subsets, thus the selection for advancement of new genetic combinations to the next step in the plant breeding program requires extra effort to overcome this statistical issue.

Cluster analysis is a promising solution to avoid subset overlapping from widely-used MCPs (O'Neill and Wetherill 1971, Plackett 1971). One example of an intuitive and satisfactory approach avoiding the subset

overlapping is the use of the cluster analysis over the generalized distance of Mahalanobis (Rao 1952). Additionally, clustering techniques can be applied for taxonomy purposes since it has high affinity to the Hotelling's Principle Component Analysis and Fisher's Discriminant Analysis (Hotelling 1933, Fisher 1936, Edwards and Cavalli-Sforza 1965).

In 1974, Alastair J. Scott and Martin Knott publicized their idea of using the maximum likelihood ratio test to evaluate the significance of partitions from the cluster analysis of sample treatment means in designs with equal number of observations per treatment (Scott and Knott 1974). The first review of methods for Scott-Knott means separation suggesting its use for agronomics was provided several years following (Chew 1976). The Scott-Knott approach is an alternative to the MCP in a situation in which two or more internally homogenous subsets of sample treatment means are expected. The procedure uses the maximum likelihood ratio as it tests for the existence of two distinct homogeneous groups of treatment means. It uses an univariate form of the divisive clustering procedure (Edwards and Cavalli-Sforza 1965) with a likelihood ratio test for determining when stop the clustering process to create non-overlapping, distinct and exclusive subsets of sampled treatment means. The process order the treatment means to minimize the

number of possible treatment mean partitions to be pondered (Fisher 1958) and then maximizes the sum of squares between clusters to determine the best partitioning. Despite a significant increase on the calculation volume for every additional treatment, it still feasible, even by hand, if the number of treatments stays lower than 12. Indeed, this limitation is a likely reason for the low use of the Scott-Knott test across researchers around the world (Gates and Bilbro 1978).

Many researchers prefer cluster analysis in order to facilitate interpretation and results presentation since it results in non-overlapping, distinct, mutually exclusive groupings of the observed treatment means (Gates and Bilbro 1978, Carmer and Walker 1985). This advantage is very clear when it is necessary to evaluate more than one variable simultaneously, because the test easily allows for a positive selection of primary traits and a negative selection for remaining evaluated traits. Within plant breeding applications, the use of non-overlapping mutually exclusive subsets create a clear cutoff for the genotype advancement procedure, while results with multiple distinct subsets can help with financial management by assigning the right subset to an appropriate testing pipeline. For example, using a cluster analyses over a yield trial with 6 resulting distinct subsets, the breeder could submit only the genotype subset partitioned in the highest category “Group A” to be

tested the in the most accurate and expensive ‘Pipeline I’ (the maximum number of locations in randomized complete blocks design); the ‘Group B’ of genotypes to the intermediate ‘Pipeline II’ (a smaller set of locations) and the “Group C and D” to be tested in the lower cost ‘Pipeline III’ (augmented blocks in the same locations as the ‘Pipeline III’, while discarding the genotypes in the “Groups E and F” (that presents an inferior performance while compared to the commercial checks, clustered in the “Group C”). After harvesting, the breeder can choose to re-test only the superior genotypes from the ‘Pipeline III’ together to the new entries to be tested in the ‘Pipeline II or I’.

Evaluation of the Scott-Knott test with agglomerative procedures exposed a lack of an appropriate experimentwise Type I Error control, thus the method should be used when the experiment has been performed with great precision. The evaluation of the Scott-Knott test through simulations suggests it may be unsuitable for experiments where MCP would be considered inappropriate as those whose the design suggest meaningful, orthogonal, single degree of freedom, linear contrasts among the treatment means. However, the Scott-Knott test has a higher Power than even the LSD test to detect small differences between treatments (Willavise et al. 1980). Moreover, the Scott-Knott test has the highest rate of correct decision and aptitude to improve

performance as treatment number increases when compared to SNK, Duncan, t-student and Tukey tests (Silva et al. 1999, Borges and Ferreira 2003). There are higher Type I Error rates than nominal rates in scenarios where the null hypothesis is false for some treatments (Partial H_0), although in scenarios where the null hypothesis is true for all treatments, the empirical Type I Error rate stays under nominal levels even for the experimentwise Type I Error rate. The Scott-Knott test offers high robustness while compared to MCP tests for mean separation in non-Gaussian distributions (Borges and Ferreira 2003).

The group homogeneity can be improved by changing the clustering pattern from divisive to non-grouped treatments clustering (Bhering et al. 2008). It usually reduces the amount of significantly different clusters—slightly increasing the amount of treatments grouped in each one of the different clusters. In spite of this drawback, this consequence can be useful in plant breeding scenarios where positive selection followed by re-testing is applied, since it can shift a small number of treatments from the second subset to the first, the third to the second and so on.

As it is common that most plant breeding designs are unbalanced, the objective of this research is to adjust and validate the Scott-Knott test to a partially balanced or unbalanced experimental design, since the current

procedure with no adjustments is only applicable for balanced designs. This paper provides a novel solution for the use of the Scott-Knott under unbalanced designs followed by the proposed method validation. Additionally, we provide a user-friendly macro program in SAS, in order to promote better dissemination of adjustment procedure and to compensate the lack of the Scott-Knott test in the SAS Package (Camacho et al. 2015).

Material and Methods

Adjustment Procedure Description

The original Scott-Knott test (1974) initiates by ranking all the k treatment means to be grouped and then by calculating B_0 from the k treatments partitioned in two smaller subsets. The B_0 value is calculated for every $k-1$ possible partition and the partition with the highest value of B_0 is tested using λ as two distinct subsets of treatment means. If the chi-square test with $\left(\frac{k}{\pi-2}\right)$ degrees of freedom, rejects the null hypothesis, the process repeats: each one of these distinct subsets is, in turn, further subdivided until each of the final cluster is show to be homogeneous by a likelihood ratio test on λ .

$$\lambda = \frac{\pi}{2(\pi-2)} B_0 / \hat{\sigma}_0^2 \quad (i)$$

The test statistic λ , represented in equation (i), depends on B_0 which is the maximum value from the sum of squares of all the possible partitions of k treatments into two groups, and on $\hat{\sigma}_0^2$ that is the maximum likelihood estimator of the error variance.

The statistic $\hat{\sigma}_0^2$ is the maximum likelihood estimator of σ^2 for treatments under the null hypothesis. The equation (ii) shows how vs^2 is used where s^2 represents an unbiased estimator of σ^2 associated with v degrees of freedom. The variable n the replication number, or the total of blocks according to the experiment design.

$$\hat{\sigma}_0^2 = \left[\sum_{i=1}^k (y_i - \bar{y})^2 + vs^2 \right] / (k + v); s^2 = \frac{MSE}{n} \quad (ii)$$

Since the Means Square Error (MSE) is a good measure of variance, it can be used as a satisfactory term for the estimation s^2 .

The equation (iii) shows the relation between the unbiased estimator s^2 and Standard Error of the Mean $SE_{\bar{y}}$, where the $RMSE$ is the Root Mean Square Error. It is valid only under an equal number of observations for every treatment ($n_1 = n_2 = \dots = n_k$). Additionally under a balanced experiment design $SE_{\bar{y}}$ presents the very same value for every treatment and leads to the equation (iv), the base of the proposed adjustment, where the mean of the sum of the squares of $SE_{\bar{y}}$, estimates s^2 .

$$s^2 = \frac{\text{MSE}}{n} = \left(\sqrt{\frac{\text{MSE}}{n}} \right)^2, \text{ and } \text{SE}_{\bar{y}} = \frac{\text{RMSE}}{\sqrt{n}} = \sqrt{\frac{\text{MSE}}{n}}, \quad (\text{iii})$$

$$\text{thus } s^2 = (\text{SE}_{\bar{y}})^2$$

$$s^2 = (\text{SE}_{\bar{y}})^2 = \frac{1}{k} \sum_{i=1}^k (\text{SE}_{\bar{y}_i})^2 \quad (\text{iv})$$

Moreover, the equation (iv) used in a balanced experimental design can be modified and expressed as equation (v), where it also permits a different number of observations for every treatment. After the modification, the corrected unbiased estimator of s_c^2 can change according to the $\text{SE}_{\bar{y}_i}$ of treatments in the partitioned set, thus in order to accommodate subsets of treatments with unequal and equal number observations, s_c^2 should be calculated for every null hypothesis before testing the statistic λ against a χ^2 distribution with the associated ν degrees of freedom.

$$s_c^2 = \frac{1}{k} \sum_{i=1}^k (\text{SE}_{\bar{y}_i})^2 \quad (\text{v})$$

The following changes in the original procedure are minimal and are disclosed in the equations (vi). The notation λ_c should be used to identify λ statistics while using the correction, despite that the testing process against the χ^2 distribution remains the same as the original procedure.

$$\lambda_c = \frac{\pi}{2(\pi - 2)} B_0 / \hat{\sigma}_{0c}^2, \quad (\text{vi})$$

$$\text{where } \hat{\sigma}_{0c}^2 = \left[\sum_1^k (y_i - \bar{y})^2 + v s_c^2 \right] / (k + v);$$

The correction increases the $\hat{\sigma}_{0c}^2$ value as the number of observation per treatment reduces—decreasing the final λ_c value, hence leading to a lower probability of rejecting the null hypothesis which protects the test. The unbalanced treatment adjustment maintains the same features and results as the original method in balanced treatment scenarios. Indeed, s_c^2 only changes for clusters in an unbalanced condition (*i.e.*, missing plots). While clustering the same experiment, after partitioning all treatment means with missing plots, the remaining clusters should have the same s_c^2 value. It is important to keep in mind that since the process follows a hierarchical clustering sequence, the very same subset of treatment means with unequal number of observations can be partitioned multiple times before composing the final specific cluster. Consequently, this solution can provide a satisfactory protection for the original Scott-Knott test without a significant power reduction.

Adjustment validation

The s_c^2 deduction can point out how the correction affects the Scott-Knott test, but in spite of that, it is mandatory to quantify and compare Power and Type I Error of the adjustment while using it. In order to

validate the proposed adjustment, the use of Monte Carlo (Metropolis and Ulan 1949) approach is a suitable option to simulate experiments with known parameters and then evaluate the results by comparing the original test against the adjusted solution for unbalanced designs (Carmer and Swanson 1971, Silva et al. 1999, Borges and Ferreira 2003). For this purpose, more than 40 million experiments were simulated for multiple unbalance levels combined with several α values. The simulation scheme is composed of three main branches: complete H_0 ($\mu_1 = \mu_2 = \mu_3 = \dots = \mu_I$), partial H_0 ($\mu_1 = \dots = \mu_{I/2} \neq \mu_{(I/2+1)} = \dots = \mu_I$), complete H_1 ($\mu_1 \neq \mu_2 \neq \mu_3 \neq \dots \neq \mu_I$). The first branch was used only to quantify Type I Error, and the third just to measure Power, while the second branch measures Type I Error and Power.

All three branches contained nine levels of α (0.01, 0.02, 0.05, 0.08, 0.10, 0.12, 0.15, 0.18, and 0.20). Within each α level, there were ten levels of missing data (0.00, 0.01, 0.02, 0.05, 0.08, 0.10, 0.12, 0.15, 0.18, and 0.20). Since the second and third branches were used to evaluate the test power they also presented four (1, 2, 3, and 4) levels of δ (true difference between the treatment means). In order to improve the study robustness, 50,000 experiments were simulated for all 810 Monte Carlo simulation setups across all three branches, culminating in a total of 40.5 million simulated experiments.

Furthermore, every simulated experiment was composed of a random number of blocks (3 to 20) and a random number of treatments (4 to 100). Experiments with a number of observations lower than 50 were replaced to avoid a small number of degrees of freedom after the data removal at random to reach the required missing level. Both block and treatment numbers were from a uniform distribution. The effects of block and observation error were from a normal distribution with a mean of zero and a standard deviation of one. The differences between subsets were defined using the standard error of difference between two means, multiplied by the level of δ . After each experiment was generated, some plot values were removed at random. As the simulation removed plots randomly with no restriction, the minimum number of plots was set to one per treatment to avoid treatments with no plots.

Instead of measuring Type I Error per comparison, the Type I Error rate was measured per experiment, where rejecting a single incorrect null hypothesis in an experiment scores as experimentwise Type I Error. This approach is more severe and general because it does not consider the number of treatments in the experiment (*e.g.*, a higher number of treatments promotes an even higher number of contrasts and it implies a higher probability of Type I Error). However, this approach should be able to perform a better distinction between the original and adjusted

procedures. Converging results were expected for both procedures (original and adjusted) under balanced designs. Thus, contrast can be observed only between balanced and unbalanced designs.

All 40.5 million experiments were simulated in SAS 9.2 using the Interactive Matrix Language Procedure (IML) and evaluated by the Generalized Linear Models Procedure (Proc GLM). The adjusted means output was grouped by a compiled macro. The simulations were performed using a computer with a Intel I7 quad-core processor and a solid state drive. A recursive SAS localhost multithread approach with isolated workplaces was used to speed up the simulation running time. Use of logged error handling capable macros for unexpected error control and processing batches of 5,000 experiments increased the processing stability and allowed the process suspension every twelve minutes if necessary. Six to nine simultaneous SAS executions were performed according to the simulated branch to keep the CPU use over 90%.

Regarding the accuracy of the estimated Type I Error Rates using Monte Carlo simulations, the exact binomial test contrasting the nominal significance level against the obtained empirical rate was applied (Leemis and Trivedi 1996). In scenarios where the exact binomial test rejected the null hypothesis ($p < 0.01$), the performance of the Scott-

Knott test should be considered *conservative* provided the empirical rate is lower than the nominal rate, and was considered *liberal* if higher. In scenarios where the exact binomial test did not reject the null hypothesis, the tests were classified as *accurate* or *precise*. The F-value was obtained using the equation (vii), where y represents the number of experiments with at least one Type I Error and N the number of simulated experiments (50,000). The p -value was found using $v_1 = 2(N - y)$ and $v_2 = 2(y + 1)$ degrees of freedom.

$$F = \left(\frac{y + 1}{N - y} \right) \left(\frac{1 - \alpha}{\alpha} \right) \quad (\text{vii})$$

Adjustment macro

A free compiled SAS macro is available to final users and can be downloaded at <http://www.tconrado.com/sas/sk.zip>. The compressed file also contains a SAS example program to provide better understanding of how to use the program. The macro can work with any *lsmeans* clause in the PROC GLM as any SAS means test and also handles the *class* and *by* statements or combination of both. The macro generates two output SAS datasets, one presenting the means clustering and the second showing the clustering history (p -values of every clustering challenge with the calculated statistics).

The macro uses five parameters (alpha=0.05, homogeneity=0, rep=0, clustering=0, and noprint=0), which can be assigned custom values. The

first parameter indicates the desired α . Setting the homogeneity parameter to 1 specifies the use of non-grouped treatments clustering (Bhering et al. 2008), instead of the default divisive clustering method (Edwards and Cavalli-Sforza 1965). The third option allows the user to specify the n value to force the macro to use the original Scott-Knott clustering procedure to calculate s^2 even under unbalanced number of observations. The last two options allow the user to prevent printing the clustering history and/or the clustered means by setting to 1 the parameters ‘clustering’ and ‘noprint’.

Since the adjusted Scott-Knott macro is not a native SAS procedure there are some use restrictions. It must be loaded into the SAS library and stored in the memory before running. The adjusted Scott-Knott procedure requires SAS version 9.0 or higher and has a positional procedure call disadvantage—meaning it needs to be the last statement of GLM procedure, so any parameter, statement or clause after the macro call will not work and will culminate in an error message.

Results and discussion

Table 1 summarizes the results of 4.5 million simulated experiments. These experiments were simulated under a complete H_0 hypothesis (no real difference among treatments). For experiments with balanced design (no missing plots) as the nominal α level increases, the empirical

experimentwise Type I Error became higher and persisted under experiments with missing plots using the proposed Scott-Knott adjustment, but reduced while the level of unbalance increased.

Moreover, in the half of the simulated combinations the Experimentwise Type I Error was evaluated as significantly different from the nominal value by the exact binomial test. As expected, the adjustment lead to more conservative approach as the level of missing plots increased. This result suggested that the user must ponder the nominal α level when considering the obtained level of missing plots or from the planned unbalance level.

In contrast, the adjusted and non-adjusted (original) Scott-Knott test exhibited a higher empirical experimentwise Type I Error rate than the nominal under Partial H_0 (Table 2). It also showed a small increase of the experimentwise Type I Error rate when the level of missing plots became higher, but the experimentwise Type I Error rate magnitude reduces as the α level increases. This result validated the findings of Silva et al. (1999) and exposed the weakest point of the Scott-Knott test: the lack of control of experimentwise Type I Error under a Partial H_0 Hypothesis.

Additionally, lower values of δ culminated in smaller differences of experimentwise Type I Error rate between the adjusted and non-adjusted

Scott-Knott procedure results (Fig. 1). This trend persisted while increasing the nominal α . Increasing α or δ , lead to a reduction of the difference in the Power among balanced and unbalanced experimental designs (Table 3). The adjusted and non-adjusted tests presented lower Power for $\delta \leq 1$. No significant differences in Power between the adjusted and non-adjusted procedures were noticed for $\delta > 1$. Additionally, the adjusted Scott-Knott test maintained very high Power even with a small α value under a complete H_1 (Fig. 2).

However as the unbalance level gets higher, there was a small loss of Power while using the proposed adjustment. This performance behavior was expected, as missing information causes lower abilities to reject the null hypothesis due to the required additional protection to control the Type I Error.

Therefore, the adjusted Scott-Knott procedure presented a similar performance of the original procedure under unbalanced experimental designs with a minimal loss of Power while kept a satisfactory control of the experimentwise Type I Error with improved performance at $\alpha \geq 0.05$. The adjustment increases the test use spectrum providing to the researcher an alternative to the MCPs even under a significant loss of experimental data (missing plots).

References

Bhering L, Cruz CD, Vasconcelos ES, Ferreira A and Resende MFR (2008) Alternative methodology for Scott-Knott test. **Crop Breeding And Applied Technology** **8**: 9-16

Borges LC and Ferreira DF (2003) Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normal e não normais dos resíduos. **Revista de Matemática e Estatística** **21**: 67-83

Calinski T and Corsten LCA (1985) Clustering Means in ANOVA by Simultaneous Testing **Biometrics** **41**: 39-48

Camacho LRS, Scapim CA, Senhorinho HJC and Conrado, TV. Diallel analysis of popcorn lines and hybrids for baby corn production **Crop Breeding and Applied Biotechnology** **15**: 33-39

Carmer SG and Swanson MR (1971) Detection of differences between means: a Monte Carlo study of five pairwise multiple comparison procedures **Agronomy Journal** **63**: 940-945

Carmer SG and Walker WM (1985) Pairwise multiple comparisons of treatment means in agronomic research **Journal of Agronomic Education** **14**: 19-26

Chew V (1976) Comparing treatment means: a compendium **Hortscience** **11**: 348-357

- Duncan DB (1955) Multiple range and multiple F tests **Biometrics** **11**: 1-42
- Edwards AWF and Cavalli-Sforza LL (1965) A method for cluster analysis **Biometrics** **21**: 362-375
- Fisher RA (1935) **The Design of Experiments**. Oliver and Boyd 252p.
- Fisher RA (1936) The use of multiple measurements in taxonomic problem. **Annals of Eugenics** **7**: 179-188
- Fisher RA (1958) On grouping for maximum homogeneity. **Journal of the American Statistical Association** **55**: 789-98
- Gates CE and Bilbro JD (1978) Illustration of a cluster analysis method for mean separation. **Agronomy Journal** **70**: 462-465
- Hotelling H (1933) Analysis of a complex statistical variables into principal components. **Journal of Educational Psychology** **24**: 417-441
- Keuls M (1952) The use of the “studentized range” in connection with an analysis of variance. **Euphytica** **1**: 112-122
- Leemis L and Trivedi KS (1996) A comparison of approximate interval estimators for the Bernoulli parameter. **The American Statistician Alexandria** **50**: 63-68
- Metropolis N and Ulam S (1949) The Monte Carlo Method. **Journal of the American Statistical Association** **44**: 335-341

Newman D (1939) The distribution of range in samples from a normal population expressed in terms of an independent estimate of standard deviation. **Biometrika** **31**: 20-30

O'Neill R and Wetherill GB (1971) The present state of multiple comparison methods. **Journal of the Royal Statistical Society** **33**: 218-250

Plackett RL (1971) The Discussion on R O'Neill and G B Wetherill Present State of Multiple Comparison Methods. **Journal of the Royal Statistical Society v** **33**: 242-243

Rao CR (1952) **Advanced Statistical Methods in Biometric Research**. John Wiley, 390p.

Scheffé H (1953) A method for judging all contrasts in the analysis of variance. **Biometrika v** **40**: 87-110

Scott AJ and Knott M (1974) A cluster analysis method for grouping means in the analysis of variance **Biometrics** **30**: 507-512

Silva EC, Ferreira DF e Bearzotti E (1999) Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. **Ciência e Agrotecnologia** **23**: 687-696

Student (1908) The probable error of a mean. **Biometrika** **6**: 1-25

Tukey JW (1949) Comparing individual means in the analysis of variance. **Biometrics** **5**: 99-114

Willavise SA, Carmer SG and Walker WM (1980) Evaluation of cluster analysis for comparing treatment means. **Agronomy Journal** **72**: 317-320

Figure 1 - Experimentwise Type I Error of Partial H_0 in three α levels under four contrasting δ

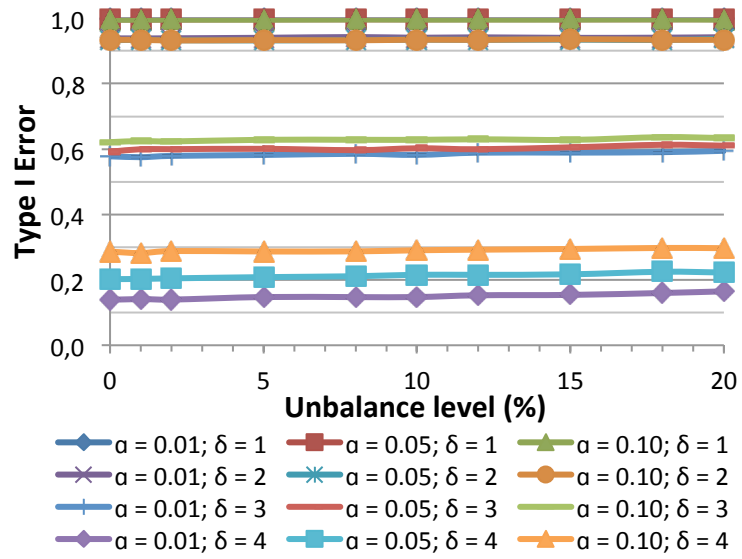
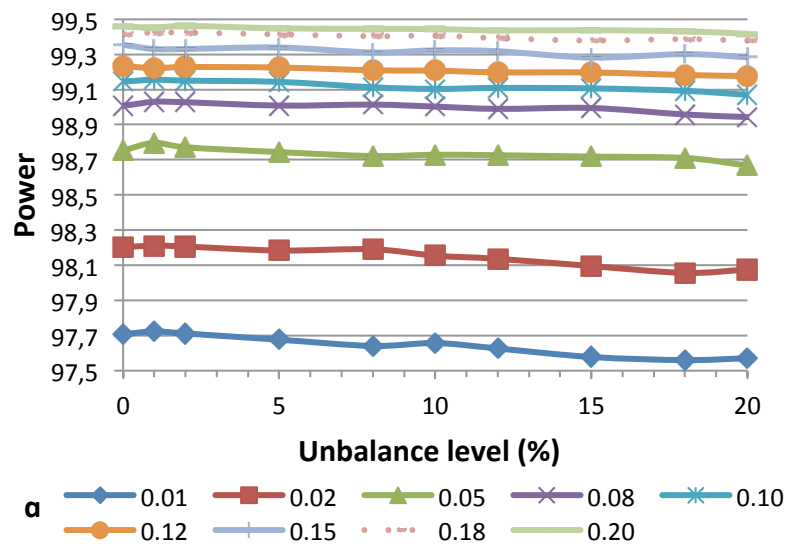


Figure 2 - Power in complete H_1 hypothesis in 90 different combination of α and unbalance level



**Table 1 - Empirical experimentwise Type I Error
under no real difference among treatments**

Nominal Alpha	Unbalance levels				
	0%	1%	2%	5%	8%
1	0.932	0.926	0.834 [†]	0.820 [†]	0.896
2	1.910	1.920	1.768	1.758	1.746 [†]
5	4.854	4.762	4.918	4.914	4.804
8	8.046	8.168	7.832	7.760	7.686 [†]
10	10.184	10.334	10.284	9.830	9.936
12	12.436 [†]	12.374	12.166	12.024	12.018
15	15.366	15.728 [†]	15.430 [†]	15.248	15.052
18	18.686 [†]	18.910 [†]	18.394	18.446 [†]	18.284
20	20.982 [†]	20.900 [†]	20.614 [†]	20.508 [†]	20.370
Nominal Alpha	Unbalance levels				
	10%	12%	15%	18%	20%
1	0.760 [†]	0.776 [†]	0.760 [†]	0.778 [†]	0.672 [†]
2	1.728 [†]	1.724 [†]	1.736 [†]	1.554 [†]	1.692 [†]
5	4.524 [†]	4.358 [†]	4.558 [†]	4.318 [†]	4.316 [†]
8	7.596 [†]	7.556 [†]	7.356 [†]	7.190 [†]	7.106 [†]
10	9.546 [†]	9.634 [†]	9.498 [†]	9.500 [†]	9.514 [†]
12	11.728	11.814	11.576 [†]	11.192 [†]	11.234 [†]
15	15.058	15.062	14.602	14.580 [†]	14.060 [†]
18	18.120	18.200	17.658	17.750	17.382 [†]
20	20.444	19.878	19.706	19.800	19.840

[†] represents scenarios where the exact binomial test rejected the null hypothesis

Table 2 - Empirical experimentwise Type I Error under real difference among treatments of $4\sigma_{\bar{x}_1 - \bar{x}_2}$

Nominal Alpha	Unbalance levels				
	0%	1%	2%	5%	8%
1	13.842	14.136	13.962	14.748	14.722
2	15.124	15.560	15.740	15.870	16.474
5	20.218	20.280	20.456	20.830	21.100
8	25.406	25.408	25.136	25.244	25.944
10	28.676	28.178	28.818	28.674	28.706
12	31.684	31.522	31.628	31.670	31.874
15	36.538	36.356	36.696	36.192	36.470
18	40.600	40.778	40.530	40.698	40.960
20	43.680	43.630	43.438	43.448	43.514

Nominal Alpha	Unbalance levels				
	10%	12%	15%	18%	20%
1	14.740	15.280	15.398	15.986	16.482
2	16.504	16.860	17.132	17.374	17.894
5	21.558	21.532	21.692	22.472	22.246
8	25.798	25.974	26.114	26.246	26.952
10	29.046	29.222	29.440	29.756	29.684
12	31.722	32.100	31.938	32.492	32.448
15	36.186	36.368	36.632	36.688	36.554
18	40.770	40.602	40.984	41.238	41.174
20	43.486	43.530	43.846	43.260	43.600

Table 3 - Power of Adjusted Scott-Knott in Several unbalance levels under partial H_0 hypothesis

δ	Unbalance levels				
	0%	1%	2%	5%	8%
p=0.01					
1	32.525	32.652	32.233	31.884	31.453
2	84.938	84.993	85.082	85.029	85.062
3	96.582	96.574	96.566	96.537	96.516
4	99.519	99.513	99.515	99.484	99.469
p=0.05					
1	48.049	47.38	47.468	47.305	46.570
2	85.206	85.256	85.253	85.295	85.259
3	96.662	96.673	96.605	96.627	96.638
4	99.552	99.546	99.538	99.514	99.500
p=0.10					
1	53.764	53.616	53.668	53.400	53.436
2	85.406	85.338	85.369	85.365	85.362
3	96.792	96.786	96.794	96.757	96.710
4	99.560	99.559	99.542	99.542	99.532
δ	Unbalance levels				
	10%	12%	15%	18%	20%
p=0.01					
1	30.982	31.306	30.748	30.236	29.735
2	85.107	85.015	85.014	85.067	85.027
3	96.513	96.491	96.459	96.452	96.433
4	99.477	99.454	99.438	99.415	99.397
p=0.05					
1	46.849	46.455	46.399	45.756	45.600
2	85.289	85.289	85.292	85.339	85.228
3	96.625	96.596	96.565	96.529	96.534
4	99.475	99.476	99.456	99.430	99.430
p=0.10					
1	52.877	53.281	52.663	52.678	52.265
2	85.396	85.386	85.453	85.439	85.464
3	96.718	96.694	96.690	96.659	96.653
4	99.509	99.502	99.479	99.459	99.459