



Universidade Federal
de São João del-Rei

Aluno: **Matheus Fernando Rodrigues Santos**
Orientador: **Ben Dêvide de Oliveira Batista**

RELATÓRIO

Revisão e avaliação de desempenho de Procedimentos de Comparações Múltiplas

OURO BRANCO - MG

2021

SUMÁRIO

1	INTRODUÇÃO	2
2	REVISÃO DE LITERATURA	3
3	RESULTADOS DA LITERATURA DE ALGUNS PCMs	11
4	CONCLUSÃO	13
	REFERÊNCIAS	14

1 INTRODUÇÃO

A ideia de uma revisão sobre os procedimentos de comparações múltiplas como um projeto, originou-se de uma sugestão dos editores da revista de Ciência e Agrotecnologia, após a publicação do artigo “*Externally studentized normal midrange distribution*” pelos autores Batista e Ferreira (2014a). Essa distribuição resultou na criação de quatro procedimentos de comparações múltiplas (PCMs), teste TM, teste SNKM, teste MGM e teste MGR. Os três primeiros se basearam nessa distribuição, e o último se baseou na distribuição da amplitude estudentizada externamente, amplamente conhecida na literatura das comparações múltiplas. Os dois primeiros foram publicados por (BATISTA; FERREIRA, 2021), e os dois últimos em fase de submissão. Dois pacotes R foram desenvolvidos, o pacote SMR (BATISTA; FERREIRA, 2014b), para o desenvolvimento numérico da distribuição da *midrange* estudentizada externamente, e o pacote *midrangeMCP* (BATISTA; FERREIRA, 2020).

Alguns trabalhos foram desenvolvidos, no sentido de apresentado o estado atual das comparações múltiplas, como os apresentados O’Neill e Wetherill (1971), e para onde vai as comparações múltiplas, publicado por Tukey (1992). Nesse sentido, esse projeto tenta atualizar as informações sobre as publicações na área dos procedimentos de comparações múltiplas, bem como apresentar algumas informações sobre as avaliações de desempenho dos principais testes. Como complemento do projeto apresentado no Edital 003/2019/PROPE, intitulado *MCP: Um pacote R para procedimentos de comparações múltiplas*, essa revisão será utilizada para implementar os demais testes do pacote MCP, que foi alterado para *MCPtests* (BATISTA; FERREIRA; OLIVEIRA, 2020), por duplicidade de nomes sob o CRAN.

Portanto, o objetivo desse trabalho é apresentar uma revisão de literatura sobre os PCMs, mostrando o estado atual e demais trabalhos relevantes, dando uma evidência sobre o desempenho dos testes quanto a sua utilização. Como complemento, será apresentado também ao longo do texto, que pacotes R podem ser utilizados para os testes apresentados na literatura, bem como outros programas estatísticos.

A estrutura do projeto apresenta a seguinte sequência, seção sobre revisão de literatura, seção sobre avaliação dos testes e por fim, a seção sobre considerações, respectivamente.

2 REVISÃO DE LITERATURA

Antes de mostrar a revisão sobre os testes de comparações múltiplas, é interessante entender como esses métodos são desenvolvidos. Faremos isso, usando as ideias de (BATISTA, 2016). Um dos interesse dos PCMs está em verificar um conjunto de parâmetros originados da combinação linear de médias tomadas duas a duas, expressos por $\theta_l = \mu_i - \mu_{i'}$, $i \neq i' \in \{1, 2, \dots, n\}$, para $l = 1, 2, \dots, N$. A coleção desses parâmetros é chamado de família (HOCHBERG; TAMHANE, 1987, p. 5), sendo o número de parâmetros dado por

$$N = \binom{n}{2} = \frac{n(n-1)}{2}, \quad (1)$$

considerando as diferenças das n médias duas a duas.

Cada parâmetro é avaliado por meio de um teste de hipóteses. E cada teste está associado erros de decisão: erro tipo I e erro tipo II. Assim, o desenvolvimento de um PCM é realizado avaliando o seu desempenho por meio do erro tipo I e o poder, sendo este último o complemento do erro tipo II. Essa avaliação se baseia em um processo de simulação, em que nesse estudo será com base no método Monte Carlo. Para a avaliação do erro tipo I, a simulação realizada supõe hipótese nula global (H_0), isto é, a simulação realizada gera amostras de n tratamentos em um experimento da mesma população, portanto, com mesma média populacional. Define-se a hipótese H_0 da seguinte forma,

Definição 2.1 — Hipótese nula global H_0 . Seja um fator, de efeito fixo e n níveis, então a hipótese nula global é definida por

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_n, \quad (2)$$

sendo $\mu_1, \mu_2, \dots, \mu_n$ as n médias das populações. ■

A avaliação do poder do teste é baseada na hipótese H_1 completa, definida por:

Definição 2.2 — Hipótese H_1 completa. Seja um fator, de efeito fixo e n níveis, então a hipótese H_1 completa é definida por

$$H_1 : \mu_1 \neq \mu_2 \neq \dots \neq \mu_n, \quad (3)$$

sendo $\mu_1, \mu_2, \dots, \mu_n$ as n médias de n populações. ■

A simulação realizada gera amostras de n tratamentos em um experimento de n populações com médias populacionais diferentes. Contudo, na prática essas n médias dos tratamentos não pertencem necessariamente a n populações. O que há na realidade são g grupos de médias pertencentes a g populações, com $g \leq n$. Nesse caso, a hipótese é expressa numa configuração de médias populacionais em que algumas, mas não todas as médias sejam iguais. Essa hipótese é chamada de hipótese nula parcial. Uma definição formal

é apresentada a seguir.

Definição 2.3 — Hipótese nula parcial H_{0_p} . Seja um fator qualquer, de efeito fixo, com n médias e g grupos, em que $g \leq n$, e as p_k médias iguais, para $k = 1, 2, \dots, g$, estão dispostas no j -ésimo grupo, com $j = 1, 2, \dots, p_k$, então a hipótese nula parcial é definida por

$$H_{0_p} : \mu_{11} = \mu_{12} \dots = \mu_{1p_1} \neq \dots \neq \mu_{21} = \mu_{22} \dots = \mu_{2p_2} \neq \dots \neq \mu_{g1} = \mu_{g2} \dots = \mu_{gp_g}, \quad (4)$$

sendo $\sum_{k=1}^g p_k = n$. ■

Observe que essa definição acaba sendo uma hipótese geral em relação às anteriores, pois se for considerado $g = 1$, nada mais é do que a hipótese nula global. Se $g = n$, tem-se a hipótese H_1 completa. Assim, sob H_{0_p} parcial, dentro do j -ésimo grupo, pode ser avaliado o erro tipo I, e entre grupos, o poder do teste.

De um modo formal, define-se o erro tipo I,

Definição 2.4 — Erro tipo I. Seja H_0 uma hipótese verdadeira. A decisão equivocada de rejeitar H_0 , sendo esta verdadeira, é chamada de erro tipo I. O tamanho do erro tipo I é definido pela probabilidade $\alpha = P(\text{Rejeitar } H_0 | H_0 \text{ verdadeira})$. ■

Diz-se que quando, em uma inferência, se rejeita uma hipótese, dado que ela é verdadeira, esta é chamada de inferência falso positiva. O ideal para um teste é manter o tamanho do erro tipo I no nível de significância adotado pelo pesquisador, chamado de nível de significância global. Por outro lado, não rejeitar uma hipótese que realmente deveria ser rejeitada, comete-se o erro tipo II.

Definição 2.5 — Erro tipo II. Seja H_0 uma hipótese falsa. A decisão equivocada de não rejeitar H_0 , sendo esta falsa, é chamada de erro tipo II. O tamanho do erro tipo II é definido pela probabilidade $\beta = P(\text{Não Rejeitar } H_0 | H_0 \text{ falsa})$. ■

Se por um lado, a probabilidade de se incorrer no erro tipo I é conhecida e fixada pelo pesquisador, por outro, a probabilidade do erro tipo II não é conhecida e nem pode ser especificada *a priori*. O que se sabe é que a probabilidade α é inversamente proporcional à probabilidade β . Relacionado ao erro tipo II, uma outra probabilidade que se torna mais fácil de ser entendida é o poder do teste, sendo apresentado a seguir.

Definição 2.6 — Poder do teste. A probabilidade de rejeitar uma hipótese, dado que ela é falsa, denotada por $1 - \beta$, é chamada de poder do teste. ■

A Tabela 1 mostra o resumo dos tipos de erros envolvidos numa tomada de decisão sob a hipótese nula H_0 e nas respectivas probabilidades.

Diante dessas considerações, segundo Rao e Swarupchand (2009, p. 66) o termo “comparações múltiplas” refere-se à aplicação de vários testes de significância estatística das diferenças entre as médias

Tabela 1 Tipos de erros e decisões corretas, com suas respectivas probabilidades em um teste de hipótese.

Decisão	Realidade	
	H ₀ verdadeira	H ₀ falsa
Não Rejeitar H_0	Decisão correta $1 - \alpha$	Erro Tipo II β
Rejeitar H_0	Erro tipo I α	Poder $1 - \beta$

(ou proporções ou variâncias, etc) dentro de um grupo. Esses autores ainda definem os procedimentos de comparações múltiplas, apresentados a seguir.

Definição 2.7 — Procedimentos de comparações múltiplas. Os procedimentos estatísticos designados ao controle adequado de efeitos de multiplicidade são chamados de procedimentos de comparações múltiplas. ■

O problema é que se inúmeras hipóteses, e em especial no tipo $H_{0l} : \mu_i - \mu_{i'}$, para $i \neq i' \in \{1, 2, \dots, n\}$ e $l = 1, 2, \dots, N$, são testadas simultaneamente, nos testes tradicionais ocorre um crescimento no erro tipo I. Este é o efeito de multiplicidade, isto é, o crescimento do erro tipo I com o aumento do número de testes. A forma de medir esse crescimento é realizada pela taxa de erro, definida a seguir.

Definição 2.8 — Taxa de erro em inferências simultâneas. Seja um conjunto de hipóteses H_{0l} verdadeiras e independentes, com $l = 1, 2, \dots, N$. A probabilidade de pelo menos uma hipótese H_{0l} ser rejeitada é chamada de taxa de erro, definida por

$$\text{Taxa de erro} = 1 - \prod_{l=1}^N [1 - P(\text{Rejeitar } H_{0l} | H_{0l} \text{ verdadeira})]. \quad (5)$$

Um exemplo para ilustrar essa situação é:

■ **Exemplo 2.1 — Efeito de Multiplicidade de PCMs.** Considere um conjunto de 10 testes de hipóteses verdadeiras e independentes, com probabilidade de erro tipo I igual a $P(\text{erro tipo I}) = \alpha$, sendo $\alpha = 0,05$. A taxa de erro é dada por:

$$\begin{aligned} \text{Taxa de erro} &= 1 - \prod_{l=1}^N (1 - \alpha) \\ &= 1 - (1 - \alpha)^N \\ &= 1 - (1 - 0,05)^{10} \\ &= 0,4013. \end{aligned} \quad (6)$$

Considerando 30 testes de hipóteses verdadeiras, a taxa de erro é 0,7854. Se forem considerados 60 testes, a taxa de erro será 0,9539. Na Figura 1, observa-se assim, que à medida que o número de testes aumenta, a

taxa de erro aumenta. Essa taxa de erro nada mais é do que o erro tipo I conjunto de todos os testes.

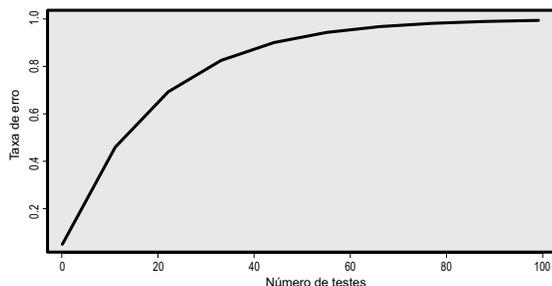


Figura 1 Representação da taxa de erro, em inferências simultâneas, em relação ao número de testes.

Percebe-se que na realidade que as N comparações não são independentes, de modo que essa não é a probabilidade exata, mas uma aproximação conservadora (HOCKING, 2003, p. 655). Diz-se então que esta é a probabilidade máxima ou um limite superior de cometer pelo menos um erro tipo I.

Assim, é com base nas taxas de erro tipo I que os procedimentos de comparações múltiplas são desenvolvidos, com o objetivo de controlar adequadamente a probabilidade de pelo menos uma hipótese verdadeira ser rejeitada, isto é, mesmo que com o aumento dos testes de hipóteses não haja também o aumento na taxa de erro.

Outro ponto importante para a avaliação dos PCMs ser baseada no erro tipo I, é que o erro tipo II não pode ser controlado diretamente pelo pesquisador. Segundo Hochberg e Tamhane (1987, p. 10), a medida de probabilidade do erro tipo I é frequentemente escolhida devido à facilidade de ser analisada e controlada.

Nos PCMs existem duas formas de avaliar o erro tipo I. A primeira avaliação é feita verificando-se a probabilidade de rejeitar uma hipótese verdadeira em todas as possíveis combinações dos níveis dos tratamentos tomados dois a dois, sendo conhecida por taxa de erro tipo I por comparação (TEC), do inglês, “comparisonwise error rate” ou “per-comparison error rate” (RAMALHO; FERREIRA; OLIVEIRA, 2005, p. 86). Segundo O’Neill e Wetherill (1971, p. 221) esse tipo de erro é calculado por:

$$\text{TEC} : \frac{\text{N}^\circ \text{ de comparações rejeitadas indevidamente}}{\text{Número total de comparações}}.$$

A segunda forma é avaliada pela taxa de erro tipo I por experimento (TEE), do inglês “experimentwise error rate” (FERREIRA, 2009, p. 215). Segundo O’Neill e Wetherill (1971, p. 221) esse tipo de erro é calculado por:

$$\text{TEE} : \frac{\text{N}^\circ \text{ exper. com pelo menos uma hipótese rejeitada indevidamente}}{\text{Número total de experimentos}}.$$

A relação entre esses dois tipos de taxas de erro é relatado por Hochberg e Tamhane (1987, p. 8),

dada por

$$TEE = 1 - (1 - TEC)^N, \quad (7)$$

sendo N o número de combinações a serem feitas, conforme expressão (1).

A questão sobre qual das taxas de erros deve ser controlada foi motivo de muitas discussões. Mas, com base em Tukey (1953, cap. 8), Miller (1981, p. 6) e Hochberg e Tamhane (1987, p. 3) a TEE deve ser padrão, pois além de proteger o nível de significância global α , determinado pelo pesquisador, permite calibrar uniformemente procedimentos diferentes para um referencial comum, e assim, comparar as suas características operacionais de maneira mais justa.

Fica evidente pela relação entre essas duas medidas, expressão (7), que o controle da taxa de erro tipo I por experimento controla a TEC de acordo com nível de significância global, porém o contrário não é verdade. Por exemplo, considerando a $TEE = 0,05$, considerando o nível de significância global de $\alpha = 0,05$ para $N = 10$, a taxa de erro por comparação pode ser dada por:

$$TEC = 1 - (1 - TEE)^{1/N} = 1 - (1 - 0,05)^{1/10} = 0,00512,$$

sendo $TEC \leq 0,05$. Já se for considerado $TEC = 0,05$ para $N = 10$, a taxa de erro por experimento é expressa em (6). Assim, é com base na TEE que os testes atualmente são propostos. Pelo fato dessas medidas não terem sido criadas na mesma época, alguns testes foram baseados na TEC, caso clássico como o teste t de *Student* e o teste de Duncan, que controlam a TEC, mas não controlam a TEE.

Com base nessas afirmações, Hsu (1996, p. 27) define a taxa de erro por experimento da seguinte forma:

Definição 2.9 — Taxa de erro por experimento (TEE). A taxa de erro por experimento de um método de comparação múltipla é o supremo, sob H_0 , da probabilidade de se ter ao menos uma afirmativa incorreta, isto é

$$TEE = \sup_{H_0} P_{H_0}(\text{ao menos uma rejeição incorreta}), \quad (8)$$

sendo H_0 expressa na Definição 2.1. ■

Essa taxa de erro por experimento é definida independentemente se uma hipótese nula completa ou parcial está sendo considerada. Portanto, será a TEE o método adotado para a avaliação dos testes propostos nesse trabalho.

Nem todos os testes revisados foram inseridos nesse relatório, devido a limite de páginas necessário. Contudo, faremos uma menção a quatro testes. O primeiro teste apresentado é o teste de Tukey. Este teste requer que todos os níveis de tratamentos tenham o mesmo número de repetições, as médias envolvidas sejam independentes, e haja homogeneidade de variâncias entre os tratamentos envolvidos. Segundo Hinkelmann e Kempthorne (2007, p. 226), o teste de Tukey foi desenvolvido para controlar o erro tipo I por experimento.

A seguir é apresentado o teste.

Proposição 2.1 — Teste de Tukey (1953). Considere n amostras normalmente distribuídas, independentes e homocedásticas, de tamanho r , e \bar{Y}_i , a média da i -ésima amostra (tratamento), $i = 1, 2, \dots, n$. Então a hipótese $H_0 : \mu_i - \mu_{i'} = 0$ para $i \neq i' \in \{1, 2, \dots, n\}$, será testada comparando $|\bar{Y}_i - \bar{Y}_{i'}|$ com $\Delta = q_{(\alpha;n,\nu)} \sqrt{\frac{QME}{r}}$, em que $q_{(\alpha;n,\nu)}$ é o quantil superior a $100\alpha\%$ da distribuição da amplitude estudentizada externamente (BATISTA, 2016). Se

$$|\bar{Y}_i - \bar{Y}_{i'}| \geq q_{(\alpha;n,\nu)} \sqrt{\frac{QME}{r}}, \quad (9)$$

a hipótese H_0 é rejeitada, isto é, μ_i e $\mu_{i'}$ são consideradas significativamente diferentes de zero, no nível de significância α . ■

Equivalentemente, pode-se obter o intervalo de confiança simultâneo $(1 - \alpha)100\%$ para todas as comparações $\mu_i - \mu_{i'}$, da seguinte forma

$$(\bar{Y}_i - \bar{Y}_{i'}) \pm q_{(\alpha;n,\nu)} \sqrt{\frac{QME}{r}}. \quad (10)$$

O segundo teste de SNK foi desenvolvido por Student (1927), Newman (1939) e Keuls (1952). Este teste é similar ao teste de Tukey, com uma pequena diferença que leva em consideração o número de médias abrangidas no contraste entre médias ordenadas.

Proposição 2.2 — Teste de Student-Newman-Keuls (SNK). Considere n amostras normalmente distribuídas, independentes e homocedásticas, de tamanho r , e \bar{Y}_i , a média da i -ésima amostra (tratamento), $i = 1, 2, \dots, n$. Então a hipótese $H_0 : \mu_i - \mu_{i'} = 0$ para $i \neq i' \in \{1, 2, \dots, n\}$, será testada comparando $|\bar{Y}_i - \bar{Y}_{i'}|$ com $\Delta = q_{(\alpha;p,\nu)} \sqrt{\frac{QME}{r}}$, em que $q_{(\alpha;p,\nu)}$ é o quantil superior a $100\alpha\%$ da distribuição da amplitude estudentizada externamente (BATISTA, 2016), sendo $p = 2, 3, \dots, n$. Nesse caso, as médias devem ser ordenadas em primeiro lugar. Se

$$|\bar{Y}_i - \bar{Y}_{i'}| \geq q_{(\alpha;p,\nu)} \sqrt{\frac{QME}{r}}, \quad (11)$$

a hipótese H_0 é rejeitada, isto é, μ_i e $\mu_{i'}$ são consideradas estatisticamente diferentes, no nível nominal de significância α . ■

Apesar de usar o quantil da distribuição da amplitude estudentizada externamente, a diferença mínima significativa varia de acordo com o número p de médias abrangidas, e portanto, um total de $(n - 1)$ valores críticos.

O terceiro teste abordado baseado na amplitude estudentizada externamente, mas que não fará parte dos estudos desse trabalho, apresentado apenas pela sua importância, é o teste de Duncan. Esse teste sofre duras críticas, pois apresenta alta taxa de erro tipo I por experimento (O'NEILL; WETHERILL, 1971).

Proposição 2.3 — Teste de Duncan (1955). Considere n amostras normalmente distribuídas, independentes e homocedásticas, de tamanho r , e \bar{Y}_i , a média da i -ésima amostra (tratamento), $i = 1, 2, \dots, n$.

Então a hipótese $H_0 : \mu_i - \mu_{i'} = 0$ para $i \neq i' \in \{1, 2, \dots, n\}$, será testada comparando $|\bar{Y}_i - \bar{Y}_{i'}|$ com $\Delta = q_{(\alpha_p; p, \nu)} \sqrt{\frac{QME}{r}}$, em que $q_{(\alpha_p; p, \nu)}$ é o quantil superior a $100\alpha_p\%$ da distribuição da amplitude estudentizada externamente (BATISTA, 2016), e $\alpha_p = 1 - (1 - \alpha)^{p-1}$, sendo $p = 2, 3, \dots, n$. Se

$$|\bar{Y}_i - \bar{Y}_{i'}| \geq q_{(\alpha_p; p, \nu)} \sqrt{\frac{QME}{r}}, \quad (12)$$

a hipótese H_0 é rejeitada, isto é, μ_i e $\mu_{i'}$ são consideradas estatisticamente diferentes, no nível nominal de significância α . ■

O teste de Duncan tem uma pequena modificação quanto ao teste de SNK. No cálculo do quantil da distribuição da amplitude estudentizada externamente, o nível de significância é variável de acordo com o número de médias abrangidas, proporcionando um total de $(n - 1)$ valores críticos. Além disso, o valor do quantil da amplitude estudentizada externamente varia conforme o número de p médias abrangidas pelo contraste.

O teste criado por Scott e Knott (1974), tem como base o uso da razão de verossimilhança para testar a existência de mais de um grupo de médias. Esse teste tem vantagem em relação aos PCMs descritos, a de não apresentar ambiguidade em seus resultados. Assim, seguindo o algoritmo apresentado em Ramalho, Ferreira e Oliveira (2005, p. 93-94) e acrescentando algumas atualizações para a aplicação do teste Scott-Knott, tem-se:

1. as médias dos tratamentos devem ser ordenadas: $\bar{Y}_{(1)}, \bar{Y}_{(2)}, \dots, \bar{Y}_{(n)}$. Nessa situação, o número de partições é $n - 1$ (FISHER, 1958, p. 796). Inicialmente o número de tratamentos é $g = n$;
2. determinar a partição entre dois grupos que maximize a soma de quadrados entre grupos, como sugerido por Edwards e Cavalli-Sforza (1965, p. 364). Essa soma de quadrados será definida por β_0 . Sejam T_1 e T_2 os totais dos dois grupos com k_1 e k_2 tratamentos em cada um. Assim, β_0 é estimado da seguinte forma:

$$\beta_0 = \frac{T_1^2}{k_1} + \frac{T_2^2}{k_2} - \frac{(T_1 + T_2)^2}{k_1 + k_2},$$

sendo

$$T_1 = \sum_{i=1}^{k_1} \bar{Y}_{(i)} \quad \text{e} \quad T_2 = \sum_{i=k_1+1}^g \bar{Y}_{(i)},$$

em que $\bar{Y}_{(i)}$ é a média do tratamento da posição ordenada i . Os dois grupos deverão ser identificados por meio da inspeção das somas de quadrados das $g - 1$ partições possíveis, sendo g o número de tratamentos envolvidos no grupo de médias considerado;

3. determinar o valor da estatística λ da seguinte forma:

$$\lambda = \frac{\pi \beta_0}{2\hat{\sigma}_0^2(\pi - 2)},$$

em que $\hat{\sigma}_0^2$ é o estimador de máxima verossimilhança de σ_Y^2 . Seja $S_Y^2 = QME/r$ o estimador não

viesado de $\sigma_{\bar{Y}}^2$ e ν graus de liberdade associados a este estimador. Então,

$$\hat{\sigma}_0^2 = \frac{1}{g + \nu} \left[\sum_{i=1}^g (\bar{Y}_{(i)} - \bar{Y})^2 + \nu S_{\bar{Y}}^2 \right].$$

4. Se $\lambda \geq \chi_{(\alpha; g/(\pi-2))}^2$, rejeita-se a hipótese de que os dois grupos são idênticos em favor da hipótese alternativa de que os dois grupos diferem;
5. No caso de rejeitar essa hipótese, os dois subgrupos formados serão independentemente submetidos aos passos (1) a (3), fazendo respectivamente $g = k_1$ e $g = k_2$. O processo em cada subgrupo se encerra ao não rejeitar H_0 no passo (4) ou se cada subgrupo contiver apenas uma média.

3 RESULTADOS DA LITERATURA DE ALGUNS PCMs

Com base nos resultados obtidos por em simulação e os encontrados na literatura, na Figura 2 (BATISTA, 2016), é apresentado um resumo da avaliação do erro tipo I por experimento, sob H_0 completa, para os testes propostos e os testes citados na literatura para comparação, com base em seu rigor, isto é, qual teste é mais conservativo.

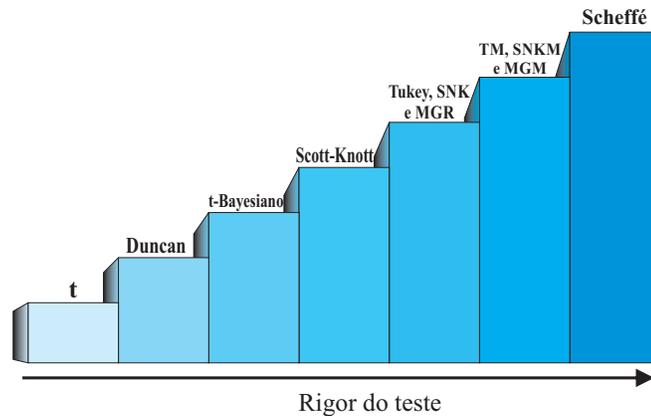
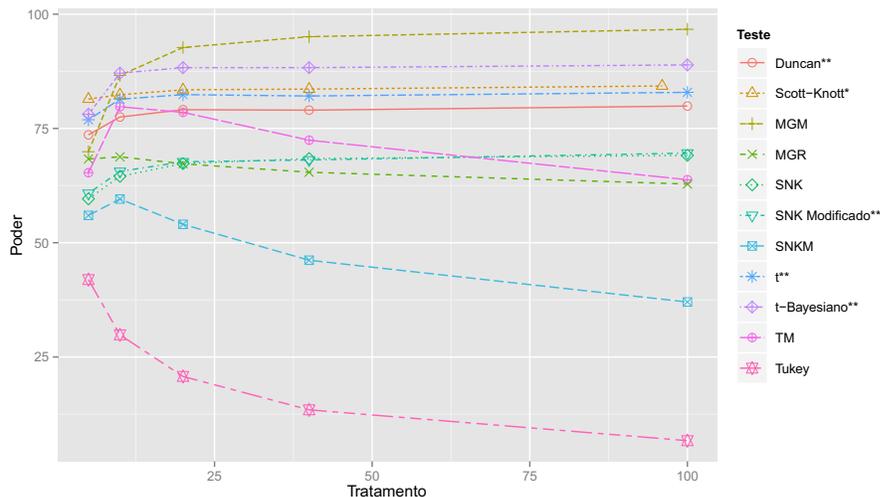


Figura 2 Ordem crescente do rigor do erro tipo I dos procedimentos de comparações múltiplas estudados.

O poder dos testes foi influenciado pelo número de tratamentos. Na Figura 3 é apresentado o poder dos testes no cenário $\delta = 4$, $r = 4$ repetições e $\alpha = 5\%$ de probabilidade. Esse cenário serviu de base para apresentar as demais situações, pois os resultados foram equivalentes.



* Resultado de Silva, Ferreira e Bearzoti (1999) e ** Resultados de Perecin e Malheiros (1989).

Figura 3 Poder dos testes de Duncan, de Scott-Knott, de MGM, de MGR, de SNK, de SNK modificado, de SNKM, t, t-bayesiana, de TM e de Tukey, em porcentagem, sob H_1 completa, para detectar uma diferença entre médias de $4\sigma_{\bar{Y}}$, com $r = 4$ repetições, em função do número de tratamentos, para um $\alpha = 0,05$.

Observou-se que os testes de MGM, t-bayesiano e de Scott-Knott apresentaram o maior poder, sendo

o primeiro teste o que apresentou maior destaque. O teste de Tukey apresentou o pior desempenho. Em relação aos testes propostos, o teste de SNKM foi o que obteve o menor poder. Contudo, o poder desse teste foi superior ao poder do teste de Tukey. Os demais testes apresentaram os valores do poder intermediário entre os do teste de MGM (teste com maior poder) e do teste de Tukey (teste com menor poder). A escolha do melhor teste avaliando apenas o poder, sob H_0 completa, é: para $n \leq 5$ o teste de Scott-Knott é indicado, e para $n > 5$, o teste de MGM é o mais indicado.

4 CONCLUSÃO

Como foi proposto neste trabalho, foram revisados e catalogados um total de 454 documentos sobre os procedimentos de comparações múltiplas, incluindo livros, artigos e outros documentos. O número de procedimentos aqui apresentados foi limitado, não sendo possível apresentar todos devido ao limite de páginas do relatório.

Por fim, esse projeto é parte da submissão de um artigo de revisão que será enviado a revista de Ciência de Agrotecnologia, UFLA/Lavras, MG.

REFERÊNCIAS

- BATISTA, B. D. de O. **Teste de comparações múltiplas baseados na distribuição da *midrange* estudentizada externamente**. 194 p. Tese (Doutorado) — Universidade Federal de Lavras, Lavras, MG, 2016.
- BATISTA, B. D. de O.; FERREIRA, D. F. SMR: An R package for computing the externally studentized normal midrange distribution. **The R Journal**, v. 6, n. 2, p. 123–136, dez. 2014.
- BATISTA, B. D. de O.; FERREIRA, D. F. **SMR**: Externally studentized midrange distribution. Vienna, Austria, 2014. R package version 2.0.1. Disponível em: <<http://cran.r-project.org/web/packages/SMR/index.html>>.
- BATISTA, B. D. de O.; FERREIRA, D. F. **midrangeMCP**: Multiple comparisons procedures based on studentized midrange and range distributions. Vienna, Austria, 2020. R package version 3.1. Disponível em: <<http://CRAN.R-project.org/package=midrangeMCP>>.
- BATISTA, B. D. de O.; FERREIRA, D. F. Alternative to tukey test. **Ciência e Agrotecnologia**, v. 44, n. e008020, p. 1–11, 2021.
- BATISTA, B. D. de O.; FERREIRA, D. F.; OLIVEIRA, D. A. B. J. de. **MCPtests: Multiples Comparisons Procedures**. [S.l.], 2020. R package version 1.0.1. Disponível em: <<https://CRAN.R-project.org/package=MCPtests>>.
- DUNCAN, D. B. Multiple range and multiple F tests. **Biometrics**, v. 11, p. 1–42, 1955.
- EDWARDS, A. W. F.; CAVALLI-SFORZA, L. L. A method for cluster analysis. **Biometrics**, v. 21, n. 2, p. 362–375, 1965.
- FERREIRA, D. F. **Estatística Básica**. 2. ed. Lavras: Editora UFLA, 2009. 664 p.
- FISHER, W. D. On grouping for maximum homogeneity. **Journal of the American Statistical Association**, v. 53, n. 284, p. 789–798, 1958.
- HINKELMANN, K.; KEMPTHORNE, O. **Design and Analysis of Experiments**: Introduction to experimental design. 2. ed. New Jersey: Wiley, 2007. I. 631 p.
- HOCHBERG, Y.; TAMHANE, A. **Multiple comparison procedures**. New York: Wiley, 1987. 450 p.

HOCKING, R. R. **Methods and Applications of Linear Models**: Regression and the analysis of variance. 2. ed. New Jersey: Wiley, 2003. 741 p.

HSU, J. C. **Multiple comparisons**: Theory and methods. London: Chapman & Hall/CRC, 1996. 277 p.

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

MILLER, R. **Simultaneous statistical inference**. 2. ed. New York: Springer-Verlag, 1981. 299 p.

NEWMAN, D. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. **Biometrika**, v. 31, n. 1/2, p. 20–30, 1939.

O’NEILL, R.; WETHERILL, G. B. The present state of multiple comparison methods. **Journal of the Royal Statistical Society**, v. 33, n. 2, p. 218–250, 1971.

PERECIN, D.; MALHEIROS, E. B. Uma avaliação de seis procedimentos para comparações múltiplas. In: ESCOLA SUPERIOR DE LAVRAS. 3º **Simpósio de Estatística aplicada à Experimentação Agônômica**. Lavras, MG, 1989. p. 66.

RAMALHO, M. A. P.; FERREIRA, D. F.; OLIVEIRA, A. C. de. **Experimentação em genética e melhoramento de plantas**. 2. ed. Lavras: Editora UFLA, 2005. 322 p.

RAO, C. R.; SWARUPCHAND, U. Multiple comparison procedure - a note and a bibliography. **Journal of Statistics**, v. 16, p. 66–109, 2009.

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

SILVA, E. C. da; FERREIRA, D. F.; BEARZOTI, E. Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de monte carlo. **Ciência e Agrotecnologia**, v. 23, n. 3, p. 687–696, jul./set. 1999.

STUDENT. Errors in routine analysis. **Biometrika**, v. 19, p. 151–164, 1927.

TUKEY, J. W. The problem of multiple comparisons. Unpublished memorandum in private circulation. 1953.

TUKEY, J. W. Where should multiple comparisons go next? In: HOPPE, F. M. (Ed.). **Multiple Comparisons, Selection, and Applications in Biometry. A Festschrift in Honor of Charles W Dunnett**.

New York: Marcel Dekker, 1992. p. 187–207.