Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas
Departamento de Estatística
Bacharelado em Estatística

Adaptação do Método Aglomerativo de SCOTT-KNOTT a Dados de Contagem

Leticia Gauna dos Santos

Ouro Preto-MG
Janeiro 2022

Leticia Gauna dos Santos

Adaptação do Método Aglomerativo de SCOTT-KNOTT a Dados de Contagem

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador(a)

Eduardo Bearzoti

Universidade Federal de Ouro Preto – UFOP
Departamento de Estatística – DEEST

Ouro Preto-MG

Janeiro 2022

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Leticia Gauna dos Santos

Adaptação do método aglomerativo de SCOTT-KNOTT a dados de Contagem

Monografia apresentada ao Curso de Estística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 14 de janeiro de 2022

Membros da banca

Dr. Eduardo Bearzoti - Orientador - Universidade Federal de Ouro Preto
Dr. Marcelo Carlos Ribeiro - Universidade Federal de Ouro Preto
Dr. Tiago Martins Pereira - Universidade Federal de Ouro Preto

Professora Dr. Eduardo Bearzoti, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 14/01/2022

Documento assinado eletronicamente por **Eduardo Bearzoti**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 19/01/2022, às 16:41, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Tiago Martins Pereira**, **COORDENADOR(A) DE CURSO DE ESTATÍSTICA**, em 27/01/2022, às 10:31, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Marcelo Carlos Ribeiro**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 28/01/2022, às 11:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0267478** e o código CRC **ED0F729C**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35400-000
Telefone: - www.ufop.br

Ao meu orientador Eduardo, dedico esse trabalho pela sua atenção dedicada ao longo de todo o projeto da minha monografia.

Agradecimentos

Em primeiro lugar agradeço ao meu pai, por me proporcionar a oportunidade de estudar em uma universidade de qualidade e me incentivar ao estudo e seguir em frente por todos esses anos. À minha tia Paula, por me guiar e me orientar em todos os momentos. À minha querida república *Diferença*, por me mostrarem o significado de família, estarem comigo em todos os momentos de dificuldade, e me proporcionarem momentos de muita felicidade durante esse ciclo.

Aos meus amigos de curso, Felipe, Iara, Milele, Yuri e Gustavo, por me acompanharem e me incentivarem durante todo o curso, e por estar comigo em todos os momentos do dia a dia especiais durante a vida acadêmica, trazendo sempre muita felicidade durante essa jornada.

Por fim a todos os professores do DEEST, por me proporcionarem um ensino de qualidade e gratuito, em especial ao meu orientador Eduardo Bearzoti, pela ótima orientação e acompanhamento durante o processo da construção da monografia, tornando esse momento possível.

"A persistência é o caminho do êxito."

Charles Chaplin

Adaptação do método aglomerativo de SCOTT-KNOTT a dados de contagem

Autor: Leticia Gauna dos Santos
Orientador(a): Prof. Dr. Eduardo Bearzoti

Resumo

O método aglomerativo de Scott-Knott é uma alternativa interessante aos procedimentos de comparações múltiplas utilizados no Planejamento de Experimentos, pois aqui não há sobreposição de médias de tratamentos, o que poderia dificultar a interpretação dos resultados. O método foi originalmente proposto para dados com distribuição normal, mas seu uso já foi sugerido para outras distribuições. A presente pesquisa teve como objetivos avaliar e ilustrar o uso do método para dados de contagem, considerando a distribuição Poisson. A técnica foi avaliada quanto ao controle do erro tipo I, utilizando simulação computacional, e ilustrada com conjuntos de dados reais, referente ao número de óbitos por Covid-19 em 20 municípios de Minas Gerais. Diferentes experimentos eram simulados sob H_0 (sem diferenças entre tratamentos), considerando diferentes configurações quanto ao número t de tratamentos, número r de repetições, e valor μ do parâmetro Poisson. Duas variações do método foram consideradas, pelo uso de uma correção do nível de significância, bem como o uso de uma proteção, utilizando o algoritmo aglomerativo somente se a rejeição de H_0 já havia acontecido com o modelo linear generalizado levando em conta os tratamentos (ou seja, sem grupos de tratamentos). Cada configuração foi simulada 5000 vezes, calculando-se taxas de erro por experimento (EER) e por comparação (CER). Os resultados mostraram que a taxa EER para o método de Scott-Knott apresentou uma tendência geral de ser superior ao nível nominal de 5%; já sua taxa CER apresentou uma tendência geral de ser inferior ao nível nominal de 5%, à exceção para situações com $r = 1$, elevados números de tratamentos, e baixo valor de μ. Os resultados também sugerem que o uso da correção do α pode corresponder a um procedimento interessante para a melhoria do controle do erro tipo I por experimento, em situações com baixos valores de μ e números de tratamentos até 10. Para baixos valores de μ recomenda-se utilizar a correção

do α, sendo também interessante proteger o método de Scott-Knott, quando há elevados números de tratamentos (acima de 10). Para valores intermediários ou maiores de μ, recomenda-se utilizar o método de Scott-Knott protegido. A aplicação do método a dados de óbitos por Covid-19 em 20 municípios de Minas Gerais ilustrou as potencialidades da técnica, formando grupos com taxas de óbitos semelhantes.

Palavras-chave: Scott-Knott, dados de contagem, Poisson.

Adaptation of the SCOTT-KNOTT agglomerative method to count data

Author: Leticia Gauna dos Santos
Advisor: Prof. Dr. Eduardo Bearzoti

Abstract

The Scott-Knott agglomerative method is an interesting alternative to the multiple comparison procedures used in Design of Experiments, as there is no overlapping of treatment means here, which could make interpretation of the results difficult. The method was originally proposed for normally distributed data, but its use has already been suggested for other distributions. The present research aimed to evaluate and illustrate the use of the method for count data, considering the Poisson distribution. The technique was evaluated for type I error control, using computer simulation, and illustrated with real data sets, referring to the number of deaths by Covid-19 in 20 municipalities in Minas Gerais. Different experiments were simulated under H_0 (no differences between treatments), considering different configurations regarding the number t of treatments, number r of replications, and the value μ of the Poisson parameter. Two variations of the method were considered, using a correction of the significance level, as well as the use of a protection, using the agglomerative algorithm only if the rejection of H_0 had already happened with the generalized linear model taking into account the treatments (*i.e.* without treatment groups). Each configuration was simulated 5000 times, calculating experimentwise (EER) and comparisonwise (CER) error rates. The results showed that the EER rate for the Scott-Knott method showed a general tendency to be higher than the nominal level of 5%; its CER rate presented a general tendency to be below the 5% nominal level, except for situations with $r = 1$, high number of treatments, and low values of μ. The results also suggest that the use of α correction may be an interesting procedure to improve the control of type I error by experiment, in situations with low values of μ and number of treatments up to 10. For low values of μ it is recommended to use α correction, and it is also interesting to protect the Scott-Knott method when there are high numbers of

treatments (above 10). For intermediate or greater values of μ, it is recommended to use the protected Scott-Knott method. The application of the method to data on deaths by Covid-19 in 20 municipalities in Minas Gerais illustrated the potential of the technique, forming groups with similar death rates.

Keywords: Scott-Knott, count data, Poisson.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

No Planejamento de Experimentos, a ferramenta de análise talvez mais utilizada consiste na chamada Análise de Variância (abreviação em inglês: ANOVA), que basicamente consiste, através de um teste F, verificar a hipótese de igualdade de efeitos dos tratamentos sendo avaliados. Quando são apenas dois os tratamentos, o próprio teste F, caso seja significativo, já os diferencia. Na maioria das situações, contudo, o pesquisador tem interesse em comparar mais do que dois tratamentos. Em tais casos, o teste F possibilita indicar se ocorre alguma diferença entre os efeitos dos tratamentos, porém sem especificar onde tais diferenças residem. Assim, com mais de dois tratamentos, após o teste F ter sido significativo, faz-se necessário um teste de comparações múltiplas entre médias.

Comparações múltiplas são essenciais no planejamento para promover um bom controle do erro tipo I, e evitar atribuir diferenças significativas quando elas realmente não existem. Pois, se apenas controlamos o erro tipo I em cada comparação entre dois tratamentos, isoladamente, o erro tipo I no conjunto de todas as comparações será maior do que o especificado em cada comparação. Dentro do Planejamento, temos diversos métodos de comparações múltiplas disponíveis, sendo os mais populares os testes de Tukey, SNK (Student-Newman-Keuls) e Duncan.

Tais métodos, contudo, apresentam uma desvantagem potencial, que é a possibilidade de uma sobreposição de médias, o que muitas vezes pode não ser eficiente e dificulta o entendimento dos resultados. Por exemplo, o primeiro tratamento (quanto ao desempenho) pode ter sua média estatisticamente igual à do segundo tratamento, e a do segundo estatisticamente igual à do terceiro, mas com a do primeiro sendo estatisticamente diferente da do terceiro. Esta sobreposição dificulta a interpretação, tanto mais quanto maior for o número de tratamentos.

Como alternativa a estes testes de comparações múltiplas, temos o método de agrupamento de Scott-Knott, sendo útil e com uma interpretação mais fácil e eficiente, uma vez que resulta na formação de grupos de médias, dentro dos quais são consideradas

estatisticamente iguais, e os grupos estatisticamente diferentes.

Da forma como foi proposto por SCOTT & e KNOTT (1974), o método baseia-se em razões de verossimilhanças, sendo válido apenas para dados com distribuição de probabilidade normal.

Quando os dados não têm distribuição normal, por exemplo quando a variável resposta é discreta, então o que geralmente se utiliza é a chamada teoria de modelos lineares generalizados (abreviação em inglês: GLM). No contexto dos GLM, já existem procedimentos disponíveis de comparações múltiplas semelhantes ao teste de Tukey, por exemplo, em ferramentas computacionais como a linguagem **R**. No entanto, a adaptação da metodologia de Scott-Knott para dados de contagem também já foi proposta (por exemplo, BARRETO, 1993), e assim seria interessante avaliar a qualidade desta metodologia adaptada, quanto ao controle do erro tipo I, uma vez que se baseia em uma estatística de teste com distribuição não exata.

Sendo assim, a presente pesquisa teve como objetivos avaliar e ilustrar o uso do método aglomerativo de Scott-Knott para dados de contagem, considerando a distribuição Poisson. A técnica foi avaliada quanto ao controle do erro tipo I, utilizando simulação computacional, e ilustrada com conjuntos de dados reais, referente ao número de óbitos por COVID-19 em 20 municípios de Minas Gerais.

2 Referencial Teórico

2.1 Comparações Múltiplas

A chamada Análise de Variância (comumente referida pela sua abreviação inglesa, ANOVA) consiste na ferramenta básica de dados experimentais, e que julga a hipótese nula de que não há diferenças entre os tratamentos avaliados. Quando rejeitamos esta hipótese, somos levados a concluir que há diferença significativa entre pelo menos um par de médias dos grupos ou tratamentos analisados. Se o experimento apresenta três ou mais tratamentos, em geral é conveniente realizar um estudo subsequente, para identificar onde residem tais diferenças, ou seja, identificar quais pares de tratamentos poderiam ser considerados como tendo médias diferentes. Quando a natureza dos tratamentos sugere a formação de comparações de interesse, é conveniente utilizar a técnica de contrastes ortogonais. Porém, quando isto não ocorre, uma alternativa consiste na técnica de comparações múltiplas (ver, por exemplo, PIMENTEL GOMES, 2009).

A técnica de comparações múltiplas é um procedimento pelo qual são avaliadas as diferenças entre todos os pares de médias dos tratamentos analisados, dado que a realização da ANOVA tenha mostrado evidências de que pelo menos um par de tratamentos apresentam médias diferentes. Trata-se de uma técnica classificada pela abreviação latina *post hoc*, ou seja, após os dados terem sido observados. Isto difere da técnica dos contrastes ortogonais, onde as comparações de interesse são formuladas antes mesmo de o experimento ser instalado. Em virtude disso, todos os testes de comparações múltiplas procuram de alguma maneira fazer um controle do erro tipo I para o conjunto total de comparações, levando-se em conta que se tratam de comparações *post hoc* (MONTGOMERY, 2012).

Dentre as técnicas de comparações múltiplas, existem diversos testes que nos permitem realizar a comparação entre todos os grupos, como os testes de Duncan, Student-Newman-Keuls (Teste "SNK"), Sheffé, Tukey, dentre outros. Dentre estes citados, é possível que o mais utilizado seja o teste de Tukey, também conhecido como Teste de Tukey da Diferença Honestamente Significativa, por propiciar um bom controle do Erro Tipo I, embora

eventualmente seja considerado um teste muito rigoroso. Dada sua grande popularidade, discorreremos brevemente sobre esta técnica, bem como um pouco sobre suas limitações. Maiores detalhes podem ser encontrados na literatura especializada em Planejamento de Experimentos, como MONTGOMERY (2012), e PIMENTEL GOMES (2009).

O teste de Tukey realiza a comparação entre pares de médias, sendo um dos testes mais populares, por ser de fácil aplicação. É um teste rigoroso, podendo ser um teste exato, quando se trata de grupos de mesmo tamanho. Tem como taxa de erro exatamente α e o intervalo de confiança sendo exatamente 1-α para grupos de mesmo tamanho. O teste de Tukey se baseia na diferença mínima significativa (D.M.S.), representada aqui por Δ, considerando os percentis do grupo. O cálculo da D.M.S. é realizada da seguinte forma:

$$\Delta = q_{\alpha}(\tau, \nu)\sqrt{\frac{QMErro}{r}}$$

Em que:

- q é o valor da amplitude total estudentizada, em função do número de médias a serem comparadas (τ) e do número de graus de liberdade do resíduo (ν).
- QMErro é o quadrado médio do erro.
- r é o número de repetições do tratamento.

Dada a diferença mínima significativa, iremos rejeitar a igualdade das médias entre dois tratamentos i e j, para um certo α, quando

$$|\bar{y}_i - \bar{y}_j| > \Delta$$

Como ilustração, iremos considerar um exemplo apresentado em BEARZOTI (2021), onde são avaliados cinco tipos de recipientes no desenvolvimento de mudas de eucalipto, sendo eles: laminado de madeira, torrão paulista, pote fértil, tubo de papel e saco plástico. Tratou-se de um Delineamento Inteiramente ao Acaso, com 6 repetições (mudas). Estes dados estão apresentados na Tabela 1.

Para esse exemplo, a análise de variância está apresentada na Tabela 2.

Tabela 1: Dados referentes à altura de mudas de eucalipto, cultivadas em cinco tipos de recipientes: laminado de madeira (LM), torrão paulista (TP), pote fértil (FP), tubo de papel (TP) e o saco plástico (SC).

	Tipo de Recipiente				
Rep.	LM	TP	SC	TU	FP
1	1,5	1,4	1,0	1,1	1,4
2	1,4	1,4	1,1	1,3	1,3
3	1,6	1,3	0,9	1,1	1,3
4	1,7	1,2	1,0	1,2	1,2
5	1,8	1,3	1,0	1,1	1,0
6	1,9	1,2	1,0	1,1	1,0

Tabela 2: Análise de variância referente à altura de mudas de eucalipto cultivadas em cinco tipos de recipientes.

Causas de Variação	GL	SQ	QM	F
Recipientes	4	1,422	0,3555	21,68**
Erro	25	0,410	0,0164	
Total	29	1,832		

Com essas informações conseguimos obter a diferença mínima significativa e realizar a comparações entre as médias dos tratamentos:

$$\text{DMS} = \Delta = 4,15\sqrt{\frac{0,0164}{6}} = 0,22$$

Em posse da DMS, pode-se comparar as médias duas a duas. Qualquer diferença entre duas médias igual ou superior à DMS, então os tratamentos em questão são considerados diferentes. Ainda é relativamente comum a apresentação do Teste de Tukey utilizando identificadores (em geral letras) que apontam se duas quaisquer médias de tratamentos poderiam ser consideradas iguais ou não. O resultado do Teste de Tukey para este exemplo está apresentado na Tabela 3.

A Tabela 3 aponta que o laminado de madeira se diferenciou dos demais tratamentos; a média do torrão paulista poderia ser considerada estatisticamente igual às do pote fértil e tubo de papel, enquanto que as médias destes dois últimos poderiam ser consideradas iguais à média do saco plástico.

Com esse exemplo, verifica-se a sobreposição de médias, que pode acontecer com o teste de Tukey. Por exemplo, poderíamos com o teste admitir que os pares de tratamentos:

Tabela 3: Alturas médias de mudas de eucalipto cultivadas em 5 recipientes.

Recipientes	Médias[1]	
Laminado de madeira	1,65	a
Torrão paulista	1,30	b
Pote fértil	1,20	bc
Tubo de papel	1,15	bc
Saco plástico	1,00	c

[1] Médias seguidas de mesma letra são estatisticamente iguais pelo teste de Tukey ao nível de 5% de probabilidade.

torrão paulista e pote fértil, bem como: pote fértil e saco plástico, seriam considerados iguais. Porém, a comparação entre o torrão paulista e o saco plástico indica uma diferença significativa, acarretando uma aparente contradição, que dificulta a interpretação.

Deve-se ressaltar, contudo, que este tipo de apresentação tem sido muito criticado nos últimos anos. Basicamente, o motivo é o de que pares de médias seguidas de mesma letra podem ter graus de evidência muito diferentes se são iguais ou não (podem apresentar valores-p muito diferentes). Por exemplo, o pacote *emmeans* da linguagem **R** (LENTH, 2020) possuía anteriormente uma função, *cld*, que gerava a apresentação utilizando letras. Esta função não está mais disponível. Este pacote apresenta as comparações entre as médias, duas a duas, com a apresentação de seus valores-p

2.2 Método de Scott-Knott

Outro método proposto para identificar onde residem as diferenças entre tratamentos de um experimento é procedimento descrito por SCOTT & KNOTT (1974). Ao contrário dos procedimentos de comparações múltiplas, o método de Scott-Knott é uma técnica de agrupamento. Baseia-se no teste da razão de verossimilhanças, visando minimizar a soma de quadrados dentro de cada grupo de médias, e maximizar a soma de quadrados entre grupos.

O método de Scott-Knott é um caso particular da técnica de agrupamento conhecida como AID (*Automatic Interaction Detection*), proposta originalmente por MORGAN & SONQUIST (1963). Basicamente, consiste em uma análise de agrupamento divisivo hierárquico dicotômico para distribuir as médias dos tratamentos em grupos relativamente homogêneos, para dados balanceados. Foi originalmente concebido para dados com distri-

buição normal.

Ou seja, admitem-se variáveis aleatórias Y_{ij} com $i = 1, 2, \ldots, t$ e $j = 1, 2, \ldots, r$, com distribuição normal independente e com variância comum σ^2, onde:

- t representa o número de tratamentos avaliados.
- r representa o número de repetições em cada tratamento.

Em posse de um conjunto de médias $\bar{y}_i$, $i = 1, 2, \ldots t$ de tratamentos, é possível formar, num primeiro momento, dois grupos de médias, podendo-se calcular uma soma de quadrados entre os grupos (SQGrupos), que necessariamente será menor que a soma de quadrados entre tratamentos (SQTrat). Defina-se B_0 como sendo o valor máximo da SQGrupos, considerando todas as possíveis partições dos t tratamentos em dois grupos.

O método de Scott e Knott é baseado em um teste de razão de verossimilhanças (ver, por exemplo, CASELLA & BERGER, 2010), tendo no numerador a verossimilhança sob a H_0: $\mu_i = \mu, \quad \forall i$, e no denominador a verossimilhança considerando H_1: $\mu_i = m_1$ ou $\mu_i = m_2$, sendo m_1 e m_2 as médias dos grupos 1 e 2, e com ao menos uma μ_i em cada grupo.

Defina-se $\hat{\sigma}_0^2$ como sendo o estimador de máxima verossimilhança de σ^2 sob H_0, o qual é dado por:

$$\hat{\sigma}_0^2 = \frac{\text{SQTrat} + \text{SQErro}}{t + \nu}$$

sendo ν o número de graus de liberdade associado à soma de quadrados de erros ou resíduos (SQErro).

Os autores apontam que a razão de verossimilhanças assim definida é uma função monotônica de $B_0/\hat{\sigma}_0^2$, e que assim, para realizar o teste da razão de verossimilhanças, poderíamos nos basear numa distribuição associada à estatística $B_0/\hat{\sigma}_0^2$. Finalmente, eles mostram que a estatística:

$$\lambda = \frac{\pi}{2(\pi - 2)} \frac{B_0}{\hat{\sigma}_0^2} \tag{2.1}$$

tem distribuição assintótica (à medida que t cresce) de qui-quadrado, com número de graus de liberdade ν_0 dado por:

$$\nu_0 = \frac{t}{\pi - 2} \tag{2.2}$$

Os autores ainda mostram que esta aproximação é muito boa, mesmo para um número reduzido de tratamentos t.

Quando temos um número muito grande de tratamentos, há um crescimento exponencial no número de grupos possíveis de serem formados, dificultando a aplicação do teste, já que existem 2^{t-1} partições possíveis das t médias entre dois grupos distintos. No entanto, os autores apontam que não há necessidade de se calcular a SQGrupos para toda partição possível entre os t tratamentos em 2 grupos, bastando apenas considerar as $t-1$ partições possíveis ao se trabalhar com as médias ordenadas.

No Brasil, o método de agrupamento de Scott-Knott se popularizou após a sua implementação no *software* SISVAR pelo Prof. Daniel Furtado Ferreira, da Universidade Federal de Lavras (FERREIRA, 2011).

Como exemplo numérico utilizaremos um exemplo apresentado por RAMALHO *et al.* (2000), no qual se estudou a absorção de água em diferentes variedades de feijão. Quanto maior a capacidade de absorção do grão de feijão em um menor espaço de tempo, mostra que o feijão tem uma melhor propriedade culinária. O delineamento utilizado foi o inteiramente casualizado (DIC). Os resultados estão apresentados na Tabela 4, e a análise de variância apresentada na Tabela 5.

Tabela 4: Porcentagem de absorção de água em 10 variedades de feijão.

	Variedade de Feijão									
Rep.	A	B	C	D	E	F	G	H	I	J
1	92,3	86,8	72,4	26,4	108,9	92,2	101,3	50,3	89,0	101,7
2	96,7	88,8	70,1	24,0	107,5	90,7	98,7	47,4	90,4	100,4
3	97,4	87,8	68,6	28,4	108,3	87,4	104,0	51,9	89,9	99,5
$y_{i.}$	286,4	263,4	211,1	78,8	324,7	270,3	304,0	149,6	269,3	301,6
$\bar{y}_{i.}$	95,5	87,8	70,4	26,3	108,2	90,1	101,3	49,9	89,8	100,5

Tabela 5: Análise de variância de um experimento comparando 10 variedades de feijão (tratamentos) quanto à absorção de água.

Causas de Variação	GL	SQ	QM	F
Tratamentos	9	18061,37	2006,82	533,21**
Erro	20	75,27	3,76	—
Total	29	18136,64	—	—

** significativo a 1% de probabilidade.

Verifica-se, pela Tabela 5, que houve diferenças significativas entre as 10 variedades de feijão. Se fôssemos utilizar um teste de Tukey, teríamos o resultado apresentado na Tabela 6.

Tabela 6: Absorção média de água em 10 variedades de feijão.

Variedade	Médias[1]	
E	108,23	a
G	101,33	b
J	100,53	bc
A	95,47	cd
F	90,10	de
I	89,77	e
B	87,80	e
C	70,37	f
H	49,87	g
D	26,27	h

[1] Médias seguidas de mesma letra são estatisticamente iguais pelo teste de Tukey ao nível de 5% de probabilidade.

Conforme a Tabela 6, percebe-se aqui o mesmo fenômeno de sobreposição de médias, que dificulta a interpretação dos resultados. Isto tende a se tornar mais relevante, para números elevados de tratamentos. A interpretação é facilitada com o uso do método de Scott-Knott, apresentado na Tabela 7.

Tabela 7: Absorção média de água em 10 variedades de feijão.

Variedade	Médias[1]	
E	108,23	a
G	101,33	b
J	100,53	b
A	95,47	c
F	90,10	d
I	89,77	d
B	87,80	d
C	70,37	e
H	49,87	f
D	26,27	g

[1] Médias seguidas de mesma letra são estatisticamente iguais pelo método de agrupamento de Scott-Knott, ao nível de 5% de probabilidade.

Como podemos analisar, temos resultados de mais fácil interpretação, utilizando o Scott-Knott. Temos que as variedades E, A, C, H e D não são equivalentes a nenhuma outra variedade de feijão. Já as variedades G e J e as variedades F, I e B são equivalentes entre si com um nível de significância de 0,05.

2.3 GLM para Dados de Contagem

Em muitas situações, experimentais ou não, tem-se que a variável resposta não pode ser admitida como tendo distribuição normal, por exemplo, quando se tem dados de contagens. Nestes casos, uma abordagem paramétrica consiste na utilização de um Modelo Linear Generalizado (NELDER & WEDDERBURN, 1972), geralmente abreviado por sua sigla inglesa, GLM.

Na formulação de um modelo linear generalizado consideramos três componentes. Primeiramente um componente aleatório, onde especificamos a distribuição de probabilidade da variável de resposta. O segundo componente corresponde a um preditor linear de parâmetros a serem estimados (componente sistemático) e, por fim, uma função de ligação, que nós permite fazer a conexão entre a esperança das observações e a parte sistemática.

Com dados de contagem, geralmente é considerada a distribuição Poisson, que possui um parâmetro μ, com $\mu > 0$, correspondente à taxa média de ocorrências em um determinado intervalo de tempo ou unidade de espaço, como por exemplo número de acidentes de carro por mês, número de mortes por determinada doença por cidade, número de casos de dengue em uma determinada região, entre outros.

Em um experimento com dados de contagem, o efeito da constante do modelo, bem como os efeitos dos tratamentos, estão contidos dentro do preditor linear. Com a distribuição de Poisson, é comum a utilização da função de ligação logarítmica, e assim teríamos:

$$\log(\mu_i) = \beta_0 + \tau_i \tag{2.3}$$

sendo β_0 e τ_i os efeitos da constante do modelo e dos efeitos de tratamento, respectivamente, e μ_i é o valor esperado para os dados do tratamento i.

Outra função de ligação proposta para a distribuição de Poisson é a função raiz quadrada. Esta função, embora de mais difícil interpretação em um ajustamento, por vezes pode ser útil, quando o ajuste utilizando a função logarítmica não apresenta convergência.

2.3.1 Dados em Taxas

Quando lidamos com dados de contagem, é relativamente comum que a unidade de tempo ou de espaço varie ao longo das unidades amostrais. Por exemplo, podemos estar estudando o número de casos de COVID-19 por município. Como diferentes municípios apresentam diferentes quantidades de habitantes, é mais conveniente se trabalhar com

taxas. Por exemplo, poderíamos considerar o número de casos por 1000 habitantes.

Nestas situações, ao invés de se trabalhar diretamente com a variável resposta Y, trabalha-se com taxas Y/t, tornando comparáveis as unidades de amostragem. O valor esperado de uma taxa é dado por μ/t, onde $\mu = E(Y)$. Para modelar a taxa, poderíamos considerar o modelo loglinear:

$$\log(\mu/t) = \beta_0 + \tau_i$$

ou seja:

$$\log \mu - \log t = \beta_0 + \tau_i$$

O termo $\log t$ é conhecido pela denominação inglesa *offset*, e os pacotes computacionais que ajustam GLMs em geral preveem a possibilidade de que haja um termo *offset*.

2.3.2 Conceito de Deviance

Seja um conjunto de dados $\mathbf{y} = (y_1, y_2, \ldots y_n)$, aos quais será ajustado um GLM. O método de ajustamento geralmente utilizado é o da máxima verossimilhança (ver, por exemplo, AGRESTI, 2002), sendo que a verossimilhança é dada por $L(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta})$, sendo $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_n)$ os parâmetros naturais da família exponencial.

Ao invés de uma função dos parâmetros naturais θ_i, podemos expressar a verossimilhança de maneira alternativa, em termos das esperanças de Y_i, que estamos representando por μ_i. Vamos representar esta parametrização alternativa como $L(\boldsymbol{\mu}|\mathbf{y})$, sendo $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots \mu_n)$.

Vamos representar ainda $L(\hat{\boldsymbol{\mu}}|\mathbf{y})$ como sendo o máximo da verossimilhança para o GLM. Dentre todos os possíveis modelos para ajustar os dados $\mathbf{y}$, o maior valor possível para a verossimilhança seria $L(\mathbf{y}|\mathbf{y})$, ou seja, um modelo no qual $\hat{\boldsymbol{\mu}} = \mathbf{y}$ Este modelo é chamado de *modelo saturado*. Não é um modelo útil por si só, uma vez que não promove redução de dados, mas serve como uma referência para avaliar o ajustamento de outros modelos.

Assim, para a Poisson, a *deviance* do GLM consiste na seguinte estatística de razão de verossimilhanças:

$$-2\left[\ln L(\hat{\boldsymbol{\mu}}|\mathbf{y}) - \ln L(\mathbf{y}|\mathbf{y})\right] \tag{2.4}$$

Esta estatística tem distribuição aproximada de qui-quadrado com $n - p$ graus de liberdade (sendo p o número de parâmetros do GLM). No caso da distribuição de Poisson, esta aproximação é tida como razoável, ao menos quando as estimativas $\hat{\mu}_i$ não sejam

muito próximas de 0 (BILDER, 2015). Assim, caso o ajustamento ao modelo tenha sido satisfatório, espera-se que a deviance seja não significativa.

A deviance é utilizada não apenas para verificar a adequação ao modelo, mas também para comparar diferentes modelos hierarquizados entre si. Assim, vamos considerar dois modelos M_0 e M_1, sendo M_0 um caso particular de M_1, e com verossimilhanças L_0 e L_1, respectivamente.

A diferença entre as deviances dos dois modelos é por sua vez uma estatística de razão de verossimilhanças:

$$-2\left[\ln L_0 - \ln L(\mathbf{y}|\mathbf{y})\right] - \left\{-2\left[\ln L_1 - \ln L(\mathbf{y}|\mathbf{y})\right]\right\} = -2\left[\ln L_0 - \ln L_1\right]$$

a qual pode ser utilizada para verificar se os parâmetros a mais em M_1, em relação ao modelo mais simples M_0, são significativos.

O procedimento de se ir calculando estas diferenças para modelos sucessivos, sequencialmente, é chamado de *Análise de Deviance*, sendo um análogo da ANOVA, quando trabalhamos com dados de distribuição normal. Estas diferenças de deviances (que também podem ser chamadas, por sua vez, de deviances) são os correspondentes às estatísticas F da ANOVA.

2.3.3 Técnica AID para Dados Não Normais

A técnica de agrupamento AID (e consequentemente o método de Scott-Knott) tem sido estendida para dados com outras distribuições que não a normal. Um detalhamento sobre essa generalização pode ser encontrada, por exemplo, em BARRETO (1993). Essencialmente, a generalização consiste em se utilizar razões de verossimilhanças considerando outras distribuições de probabilidade. Reconsidere-se a estatística de teste do método de Scott-Knott, apresentada em 2.1:

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

O primeiro termo de λ, ou seja, $\frac{\pi}{2(\pi-2)}$, é um termo constante, enquanto que o segundo termo, $\frac{B_0}{\hat{\sigma}_0^2}$, corresponderia à razão de verossimilhanças propriamente dita.

Assim, a extensão para outras distribuições (por exemplo, a Poisson), consiste em substituir este segundo termo por uma deviance (que, no fundo, é uma razão de verossimilhanças), dentro do algoritmo de procura pela melhor formação de dois grupos de tratamentos, em cada etapa do método. O número de graus de liberdade da estatística

continua sendo dada como em 2.2, e a aproximação de qui-quadrado aparentemente é satisfatória, como apontam os resultados de BARRETO & DEMÉTRIO (1998).

3 Metodologia

Atendendo aos objetivos deste trabalho, foram gerados dois conjuntos de resultados. Em um primeiro momento foi realizado um estudo de simulação computacional para avaliar a qualidade do controle do erro tipo I ao se utilizar o método de Scott-Knott adaptado a dados com distribuição Poisson. Em seguida, o método foi ilustrado utilizando dados de óbitos provocados pela Covid-19 em 20 municípios de Minas Gerais. Estas duas abordagens são descritas com maior detalhamento a seguir.

3.1 Avaliação do Controle do Erro Tipo I

Rotinas computacionais foram desenvolvidas para simular amostras correspondentes a experimentos, contendo t tratamentos, cada qual repetido r vezes, sempre utilizando um delineamento inteiramente casualizado. Tais amostras foram simuladas sob H_0, ou seja, sem haver diferenças entre tratamentos.

Dessa forma, cada um dos $t \times r$ elementos de cada amostra foram gerados conforme uma mesma distribuição Poisson, de mesmo parâmetro μ. Isto foi feito pela função *rpois()* da linguagem **R** (R CORE TEAM, 2020).

Cada amostra simulada era então submetida a dois ajustamentos:

1. Modelo doravante denominado "GLM", correspondente a um modelo generalizado tendo em seu preditor linear o efeito dos t tratamentos (ou seja, com o preditor linear apresentado em 2.3);

2. Agrupamento de tratamentos pelo método de Scott-Knott (doravante denominado "SK") adaptado a dados com distribuição Poisson. O algoritmo para esse agrupamento será descrito mais adiante;

Contudo, simulações exploratórias iniciais revelaram algumas situações em que o método SK apresentou altas taxas de erro tipo I, consideravelmente acima do nível nominal.

Assim, na expectativa de se encontrar critérios de decisão com melhor controle do erro tipo I, foram definidos ainda outros dois ajustamentos:

3. Scott-Knott com correção do nível de significância α;
4. Scott-Knott protegido (doravante denominado "SK_Prot").

A motivação da abordagem (3) se baseou em um apontamento feito por SCOTT & KNOTT (1974), em seu trabalho seminal. Os autores afirmam que, ao se tentar formar j grupos homogêneos dentre os t tratamentos, em alguma etapa do método de agrupamento, a probabilidade (sob H_0) de se obter pelo menos uma subdivisão em grupos não é maior que $\alpha^* = 1 - (1 - \alpha)^t$. Como α^* é um valor maior que α, optou-se por trabalhar com α^*, na expectativa de redução das taxas de erro tipo I. Tratou-se de um procedimento empírico, uma vez que aparentemente não haveria uma justificativa teórica que indicasse que o uso de α^*, ao longo de todo o processo aglomerativo, pudesse promover um nível de significância global (ou seja, a formação de pelo menos dois grupos ao longo de todas as etapas) mais próximo do nível nominal α.

Já a abordagem (4) correspondeu a implementar o método de Scott-Knott somente se o ajuste ao modelo GLM foi significativo, previamente. Isto serviria como uma "proteção", semelhante a alguns procedimentos de comparações múltiplas, em que são utilizados somente se o Teste F da ANOVA foi significativo.

As amostras foram simuladas considerando diferentes configurações quanto ao valor de μ, o número de tratamentos t, o número de repetições r, e se o nível de significância era corrigido ou não. As especificações destes parâmetros de simulação são dadas abaixo:

- μ: 1, 3, 5 e 10
- t: 3, 5, 10 e 20
- r: 1, 2 e 4
- Correção do α: não ou sim

Com isto, definiu-se um total de 96 situações, correspondentes a todas as combinações entre os parâmetros acima. Para cada situação, eram simuladas 5000 amostras, sendo sempre considerado um nível de significância nominal de 5%.

Valores mais baixos de μ correspondem a distribuições de Poisson mais assimétricas, o que poderia comprometer a aproximação de qui-quadrado da estatística dos ajustamentos. Quanto ao número de tratamentos, partiu-se de um valor mínimo de 3 tratamentos, pois não haveria sentido em se fazer um agrupamento em uma situação contendo apenas 2 tratamentos.

Em relação ao número de repetições, pode parecer estranho ter-se considerado o valor mínimo $r = 1$. Com apenas 1 repetição não seria possível, por exemplo, realizar uma ANOVA. Porém, ao se considerar a distribuição de Poisson, por esta não apresentar um parâmetro de dispersão, é possível realizar o ajustamento, bem como uma análise de deviance. De certa maneira, isto poderia ser considerado uma "vantagem" para aquelas situações em que se considera uma distribuição de Poisson.

A avaliação das taxas de Erro Tipo I foram avaliadas considerando duas grandezas:

- Taxa de Erro por Experimento, abreviada doravante como EER (da sigla inglesa, *experimentwise error rate*), correspondendo à proporção de experimentos (amostras) em que H_0 foi rejeitada;
- Taxa de Erro por Comparação, abreviada doravante como CER (da sigla inglesa, *comparisonwise error rate*), correspondendo à proporção de comparações entre tratamentos em que o par de tratamentos era considerado diferente.

A EER foi calculada para os ajustamentos: GLM, SK e SK_Prot (com e sem correção do α). No caso dos dois últimos, a rejeição de H_0 correspondia à formação de pelo menos 2 grupos de tratamentos. Já no caso do GLM, a rejeição de H_0 correspondia à rejeição da hipótese de igualdade entre os t tratamentos.

A taxa CER foi calculada apenas para o ajustamento SK. Para um dado número t de tratamentos, existem $\frac{t(t-1)}{2}$ comparações possíveis, dois a dois. Assim, em cada experimento simulado, era computada a proporção dessas comparações em que o método SK identificava dois tratamentos como pertencentes a dois grupos diferentes. A taxa final CER foi então calculada considerando o valor médio destas proporções, ao longo de todas as 5000 simulações.

Em todas as 96 configurações, foi avaliada a qualidade do controle do erro tipo I, considerando sempre um valor de referência para o nível de significância como sendo $\alpha = 0,05$, um valor frequentemente utilizado nas mais diversas áreas do conhecimento. Para verificar se uma dada taxa EER ou CER era ou não significativamente maior que este

valor, construiu-se um intervalo de confiança unilateral em torno de 0,05, considerando um índice de confiança de 95%. Ou seja, um intervalo de confiança com limite inferior igual a zero, e limite superior dado por:

$$0,05 + 1,645 \times \sqrt{\frac{0,05(1-0,05)}{5000}} \tag{3.1}$$

Uma dada taxa EER ou CER, em qualquer das 96 configurações, era considerada significativamente superior a 0,05 caso fosse superior ao limite 3.1.

Algoritmo Scott-Knott Para Dados Poisson

Para o ajustamento SK (e, claro, também o método SK_Prot), foi desenvolvida uma função na linguagem **R** para a adaptação do método de Scott-Knott para dados de contagem com distribuição Poisson. Esta função é apresentada no Apêndice deste trabalho, e os passos de seu algoritmo são brevemente descritos a seguir. Para um maior entendimento do algoritmo, contudo, é importante ressaltar que o método de agrupamento, ao longo de suas etapas, está sempre considerando dois modelos, com os seguintes preditores lineares:

$$\log(\mu_i) = \beta_0 + \tau_i \tag{3.2}$$

$$\log(\mu_{ij}) = \beta_0 + \gamma_j + \tau_{(j)i} \tag{3.3}$$

O primeiro modelo só leva em conta os tratamentos, sem a formação de grupos, enquanto que o segundo modelo leva em conta o efeito γ_j de um certo número de grupos formados. Neste modelo, os tratamentos estão *hierarquizados* dentro de grupos, e daí a notação $(j)i$ na identificação de cada tratamento. Ressalta-se que tanto tratamentos como grupos são *fatores*, e não covariáveis.

O algoritmo basicamente procura a melhor (no sentido de maior deviance) formação de grupos, verificando se há diferença significativa entre estes. No modelo 3.3, considerando um ajuste sequencial, tem-se a seguinte interpretação ao se testar os tratamentos (por exemplo na análise de deviance): trata-se do efeito dos tratamentos, *tendo-se removido* o efeito dos grupos (ajustado para grupos). Em outras palavras, trata-se da variação remanescente entre tratamentos, após a formação dos grupos. Ou seja, trata-se da variação entre tratamentos dentro dos grupos. Se esta variação é significativa, o método prossegue na busca por novos grupos.

Por exemplo, em uma dada etapa do método aglomerativo, tendo sido já identificado um certo número de grupos, a busca prossegue, considerando uma nova partição de gru-

pos, subdividindo um dos grupos anteriormente formados em dois. Ou seja, dentro do algoritmo, é considerado o ajuste a um modelo com o seguinte preditor:

$$\log(\mu_{ijk}) = \beta_0 + \gamma_j + \delta_{(j)i} + \tau_{(ij)k} \tag{3.4}$$

sendo $\delta_{(j)i}$ o fator correspondente à nova configuração de grupos, contendo um grupo a mais. No ajuste sequencial deste modelo, é possível verificar se a nova configuração de grupos é significativa ou não.

Os passos do algoritmo são:

1. Defina o quantil da distribuição de qui-quadrado a ser considerado no julgamento de H_0 (α ou α^*);

2. Ajuste o modelo GLM, e obtenha as médias estimadas dos t tratamentos. Ordene estas médias estimadas, e as identifique com $i = 1, 2, \ldots t$;

3. Forme um primeiro par de grupos, sendo o primeiro constituído pelo tratamento 1, e o segundo grupo constituído pelos demais tratamentos. Calcule a deviance devida a grupos (do modelo 3.3), e armazene este valor;

4. Forme um segundo possível par de grupos, tendo o primeiro grupo os dois primeiros tratamentos, e o segundo grupo os demais tratamentos. Calcule e armazene a deviance devida aos grupos. Prossiga dessa maneira, formando grupos de maneira sequencial, até à última formação, em que o primeiro grupo contém os $t-1$ primeiros tratamentos, e o segundo grupo contém o último tratamento;

5. Tome aquela partição de dois grupos que produziu o maior valor para a deviance entre grupos. Verifique se este valor está acima do valor crítico de qui-quadrado. Se não estiver, encerre o processo. Se estiver, vá ao passo seguinte;

6. Teste a significância da variação remanescente dos tratamentos dentro de grupos. Se não for significativa, encerre o processo, ficando com os grupos formados até o momento. Se for significativa, volte ao passo 3, mas aplicando este e os passos seguintes *dentro* de cada um dos grupos já formados.

No Apêndice deste trabalho, além da função para o método aglomerativo de Scott-Knott, também é apresentado um exemplo de rotina de simulação, para uma dada configuração de parâmetros de simulação.

3.2 Dados Covid-19

Na segunda parte deste trabalho, procurou-se ilustrar o método de Scott-Knott aplicado a dados de contagem com conjuntos de dados reais. Estes conjuntos se referiram ao número de óbitos relacionados à Covid-19, para os meses de janeiro e fevereiro do ano de 2021, bem como o total de óbitos até aquele momento, em vinte cidades do estado de Minas Gerais correspondentes àquelas com o maior número de estudantes da Universidade Federal de Ouro Preto. Estes dados foram coletados pelo site de notícias G1[1], sendo que os números de habitantes por cidade foram coletados no site oficial do IBGE[2].

Os dados originais estão apresentados na Tabela 8.

Tabela 8: Número de óbitos por covid-19.

Cidade	Óbitos janeiro	Óbitos fevereiro	Total óbitos	População
Barão de Cocais	5	2	66	32866
Belo Horizonte	408	480	6234	2521564
Betim	408	69	6234	444784
Congonhas	10	4	120	55309
Conselheiro Lafaiete	12	19	269	129606
Contagem	119	94	1789	668949
Coronel Fabriciano	31	19	329	110290
Divinópolis	41	24	569	240408
Governador Valadares	69	84	1183	281046
Ipatinga	76	41	886	265409
Itabira	13	22	361	120904
Itabirito	12	13	146	52446
João Monlevade	35	19	209	80416
Mariana	8	2	98	61288
Ouro Branco	7	6	51	39867
Ouro Preto	6	8	119	74558
Ponte Nova	11	12	198	59875
Santa Bárbara	2	0	36	31604
Sete Lagoas	6	17	575	241835
Timóteo	30	27	327	90568

É interessante notar que esta é uma situação com $t = 20$ "tratamentos", sendo $r = 1$ repetição por tratamento. Outra particularidade é o fato de as cidades terem números diferentes de habitantes, tornando necessário o uso do *offset* no ajustamento do modelo e na formação dos grupos.

[1]URL `https://especiais.g1.globo.com/bemestar/coronavirus/2021/mapa-cidades-brasil-mortes-covid/`

[2]URL `https://www.ibge.gov.br/cidades-e-estados/mg/`

4 Resultados e Discussão

Conforme relatado no Capítulo anterior, este trabalho apresenta dois conjuntos de resultados. No primeiro, a qualidade do controle do erro tipo I é avaliada para o método de Scott-Knott adaptado a dados de contagem, através de simulação computacional. O segundo conjunto de resultados foram obtidos a partir da aplicação deste método aglomerativo ao número de óbitos relacionados à Covid-19 em 20 municípios de Minas Gerais.

4.1 Avaliação do Controle do Erro Tipo I

4.1.1 Problemas de Convergência

Antes da apresentação dos resultados das simulações computacionais, cabe aqui uma observação sobre problemas de convergência. Observou-se, em algumas configurações, uma certa proporção de amostras simuladas para as quais o ajuste não apresentava convergência, em ao menos alguma etapa do método aglomerativo de Scott-Knott em que era utilizado o preditor linear 3.4. Isto aconteceu em cerca de um terço das configurações, sendo que a falta de convergência foi mais frequente em situações com números elevados de tratamentos e pequeno número de repetições (1 ou 2). Para a semente aleatória utilizada, a proporção de amostras simuladas sem convergência apresentou um valor máximo de 11%, para a configuração com $t = 20$, $r = 1$ e $\mu = 1$.

A título de ilustração (utilizando a linguagem **R**), vamos considerar uma amostra simulada em que isso aconteceu. Trata-se de uma situação com $t = 5$ e $r = 1$:

```
   trat  y
1:    1  0
2:    2  3
3:    3  5
4:    4  2
5:    5 12
```

sendo *trat* o fator que identifica os tratamentos, e y a variável resposta.

O ajuste ao modelo GLM produz a seguinte análise de deviance:

```
> mod <- glm(y ~ trat,family = poisson(link = "log"),
+             data=dados)
> anova(mod, test = "Chisq")

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     4     19.906
trat  4   19.906         0      0.000 0.0005212 ***
```

observando-se que o modelo convergiu, e tendo sido observadas diferenças significativas entre os tratamentos.

Suponha que, em uma dada etapa do método aglomerativo de Scott-Knott, tenham sido já formados dois grupos: um primeiro grupo, contendo os tratamentos 1, 2, 3 e 4, e um segundo grupo, contendo apenas o tratamento 5. Suponha ainda que, nesta etapa do algoritmo, esteja-se investigando uma nova configuração, contendo agora 3 grupos: um primeiro grupo contendo apenas o tratamento 1, um segundo grupo contendo os tratamentos 2, 3 e 4, e o último grupo contendo apenas o tratamento 5. Ou seja, trata-se de uma tentativa de subdivisão do primeiro grupo anteriormente formado em dois novos grupos.

Vamos dispor a configuração da etapa anterior em um fator denominado *Gant*, e a nova configuração em um fator denominado *Gr*, ou seja, por exemplo em um *data.table* como o seguinte:

```
   trat Gant Gr  y
1:    1    2  3  0
2:    2    2  2  3
3:    3    2  2  5
4:    4    2  2  2
5:    5    1  1 12
```

Ao ajustarmos o modelo contendo apenas os fatores *Gant* e *Gr* (sem o fator *trat*), tem-se o seguinte resultado:

```
> mod2 <- glm(y ~ Gant + Gr,family = poisson(link = "log"),
+             data=dados)
> anova(mod2, test = "Chisq")

     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     4    19.9058
Gant  1  12.7730         3     7.1328 0.0003517 ***
Gr    1   5.7536         2     1.3792 0.0164545 *
```

indicando que a nova configuração é significativa. A deviance relativa ao fator *trat*, ajustado tanto para o fator *Gant* como o fator *Gr*, poderia ser obtida por diferença de deviances dos dois ajustes:

$$19,906 - (12,7730 + 5,7536) = 1,379$$

No entanto, se tentarmos ajustar um modelo contendo todos os 3 fatores, ocorrem problemas de convergência:

```
> mod <- glm(y ~ Gant + Gr + trat,family = poisson(link = "log"),
+             data=dados)

Erro: loop interno 1; não é possível corrigir o tamanho do passo
Além disso: Warning message:
tamanho do passo truncado devido a divergência
```

Ou seja, com essa amostra simulada em particular, o modelo com o preditor 3.4 não converge. Inicialmente, as rotinas computacionais foram configuradas para, em uma situação sem convergência, admitir que a hipótese H_0 não foi rejeitada, para o método SK ou SK_Prot. No entanto, isto poderia ter resultado em taxas subestimadas de rejeição de H_0.

Posteriormente, investigando amostras que apresentavam tais problemas, verificou-se que a mudança da função de ligação, utilizando a raiz quadrada ao invés da função logarítmica, resultava em ajustes sem problemas de convergência. No exemplo anterior:

```
> mod <- glm(y ~ Gant + Gr + trat,family = poisson(link = "sqrt"),
+             data=dados)
```

```
     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     4    19.9058
Gant  1  12.7730         3     7.1328 0.0003517 ***
Gr    1   5.7536         2     1.3792 0.0164545 *
trat  2   1.3792         0     0.0000 0.5017804
```

resultando naqueles valores de deviance obtidos anteriormente.

Dessa forma, para evitar uma possível subestimação das taxas de rejeição de H_0, aquelas configurações com problemas de convergência foram refeitas, considerando a função de ligação raiz quadrada.

4.1.2 Taxas de Erro Tipo I

As taxas de rejeição de H_0 por experimento (EER), para os métodos designados anteriormente como GLM, SK e SK_Prot, e as taxas CER para o método SK, considerando todas as 96 configurações estudadas, variando o número de tratamentos, número de repetições, valor de μ, e correção ou não do nível de significância, estão apresentadas nas Figuras 1, 2, 3 e 4.

Tais Figuras representam situações, respectivamente, com μ igual a 1, 3, 5 e 10. Cada uma dessas Figuras apresenta duas colunas de gráficos, sendo que a coluna da esquerda se refere às situações sem correção do nível de significância para os métodos SK e SK_Prot, enquanto que nos gráficos da coluna da direita tem-se as situações com o nível de significância corrigido. Para o ajuste GLM, em nenhum momento foi realizada esta correção. Assim, como eram utilizadas as mesmas sementes aleatórias nas diferentes simulações, os segmentos de reta referentes ao ajuste GLM são iguais, considerando um gráfico da coluna da esquerda com o seu correspondente na coluna da direita. Por exemplo, o gráfico (a) referente à taxa EER para o método GLM é o mesmo gráfico que está apresentado em (b). Apesar desta redundância de informação, optou-se por manter esta disposição, para uma melhor comparação entre os métodos.

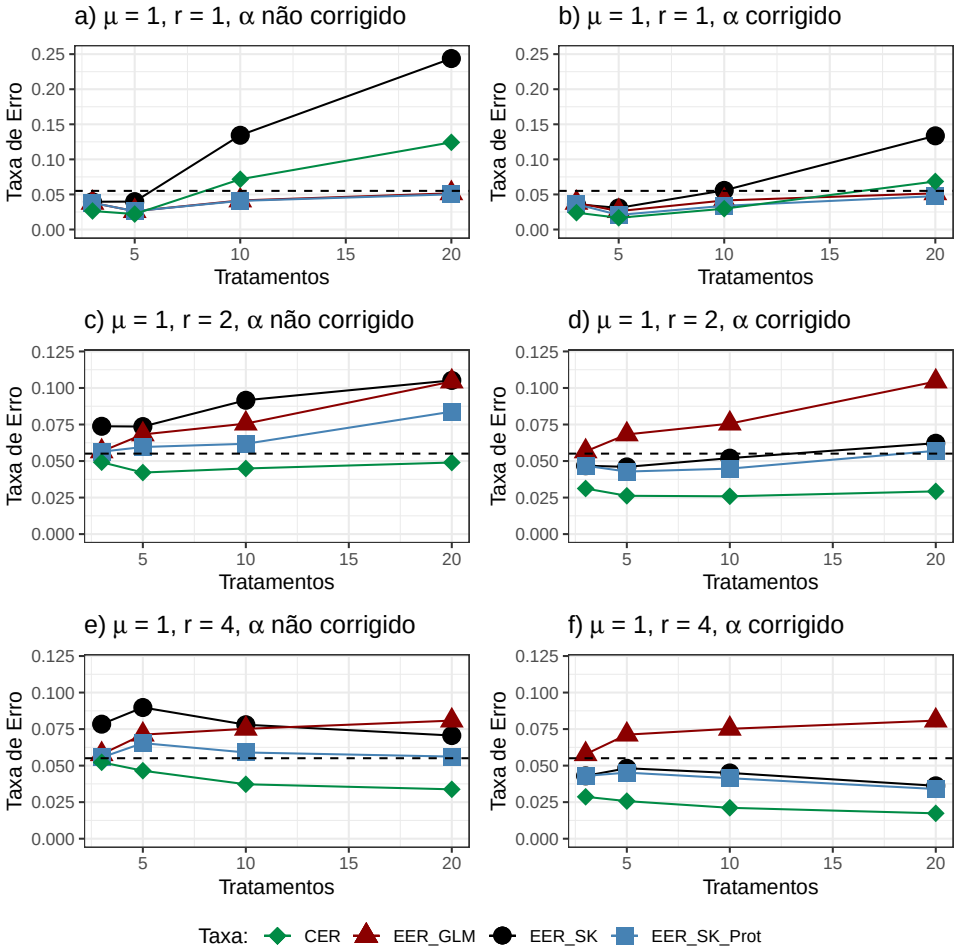

Figura 1: Taxas de erro tipo I avaliadas em 5000 simulações sob H_0, não utilizando e utilizando correção para o nível de significância, e considerando diferentes números de tratamentos e repetições, e o parâmetro Poisson $\mu = 1$ (EER_GLM: taxa de rejeição de H_0 para o modelo GLM; EER_SK: taxa de rejeição de H_0 para o método de Scott-Knott; EER_SK_Prot: taxa de rejeição de H_0 para o método de Scott-Knott protegido; CER: taxa de erro por comparação, para o método de Scott-Knott.

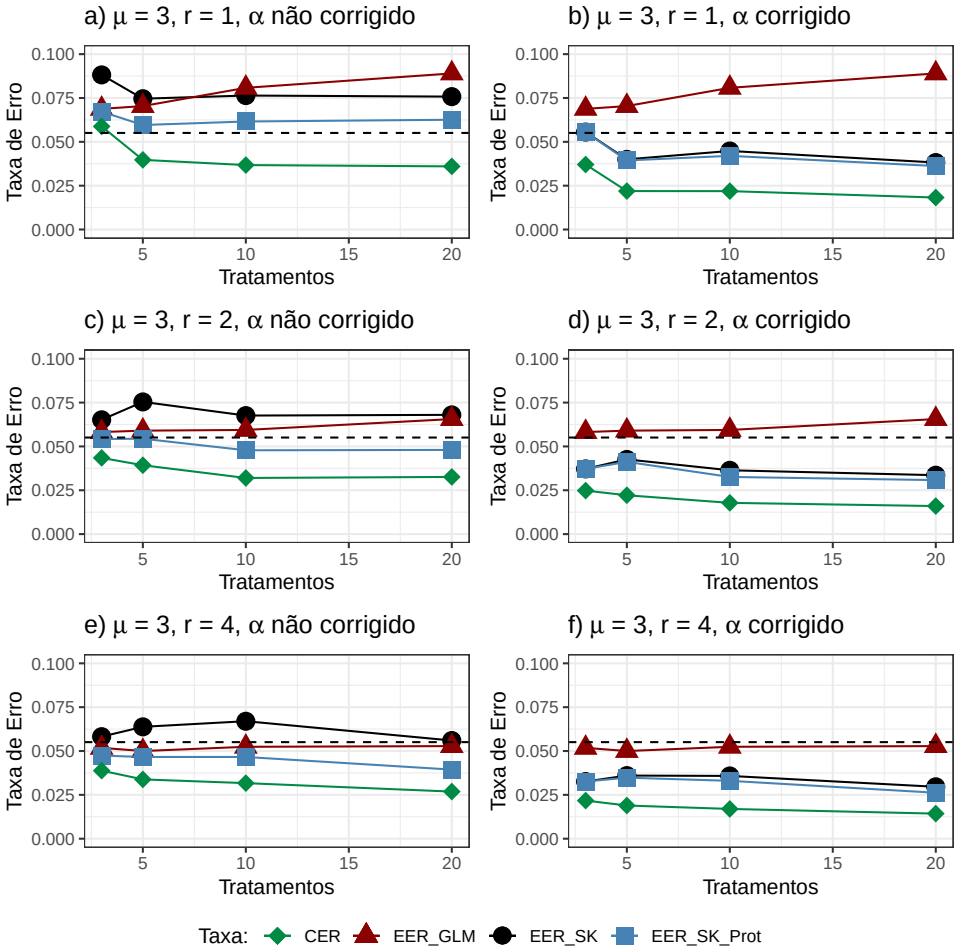

Figura 2: Taxas de erro tipo I avaliadas em 5000 simulações sob H_0, não utilizando e utilizando correção para o nível de significância, e considerando diferentes números de tratamentos e repetições, e o parâmetro Poisson $\mu = 3$ (EER_GLM: taxa de rejeição de H_0 para o modelo GLM; EER_SK: taxa de rejeição de H_0 para o método de Scott-Knott; EER_SK_Prot: taxa de rejeição de H_0 para o método de Scott-Knott protegido; CER: taxa de erro por comparação, para o método de Scott-Knott.

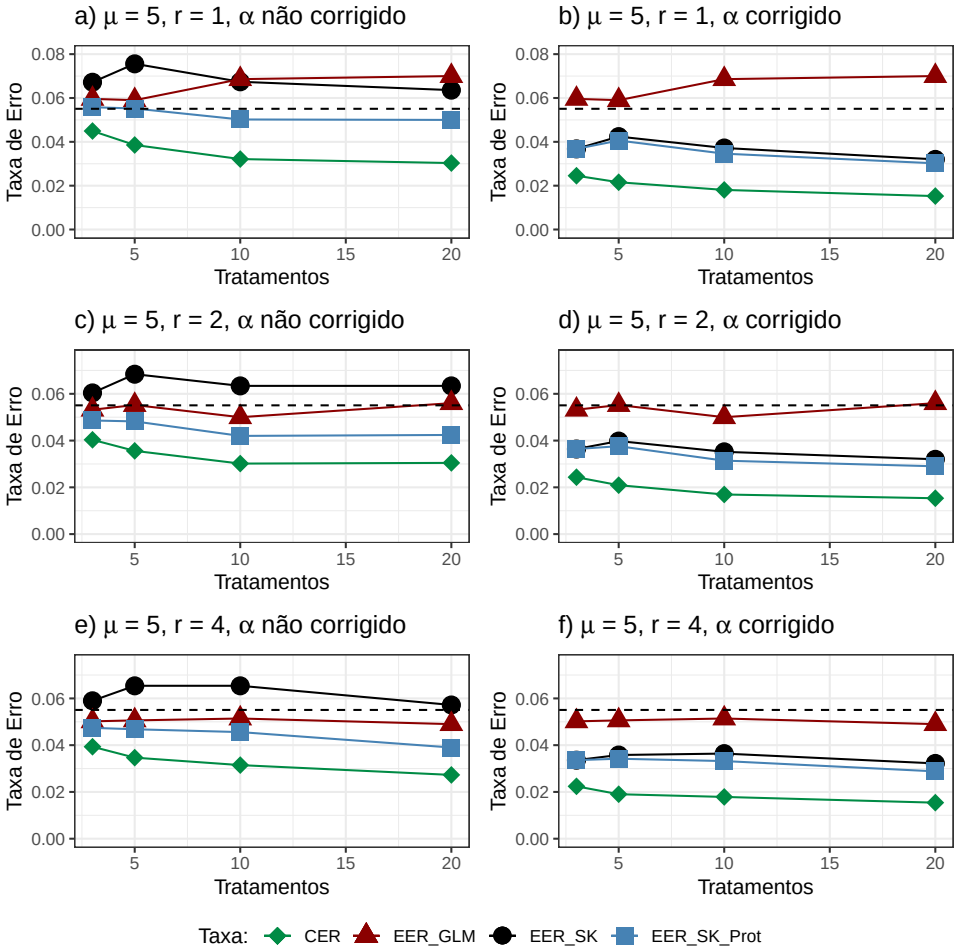

Figura 3: Taxas de erro tipo I avaliadas em 5000 simulações sob H_0, não utilizando e utilizando correção para o nível de significância, e considerando diferentes números de tratamentos e repetições, e o parâmetro Poisson $\mu = 5$ (EER_GLM: taxa de rejeição de H_0 para o modelo GLM; EER_SK: taxa de rejeição de H_0 para o método de Scott-Knott; EER_SK_Prot: taxa de rejeição de H_0 para o método de Scott-Knott protegido; CER: taxa de erro por comparação, para o método de Scott-Knott.

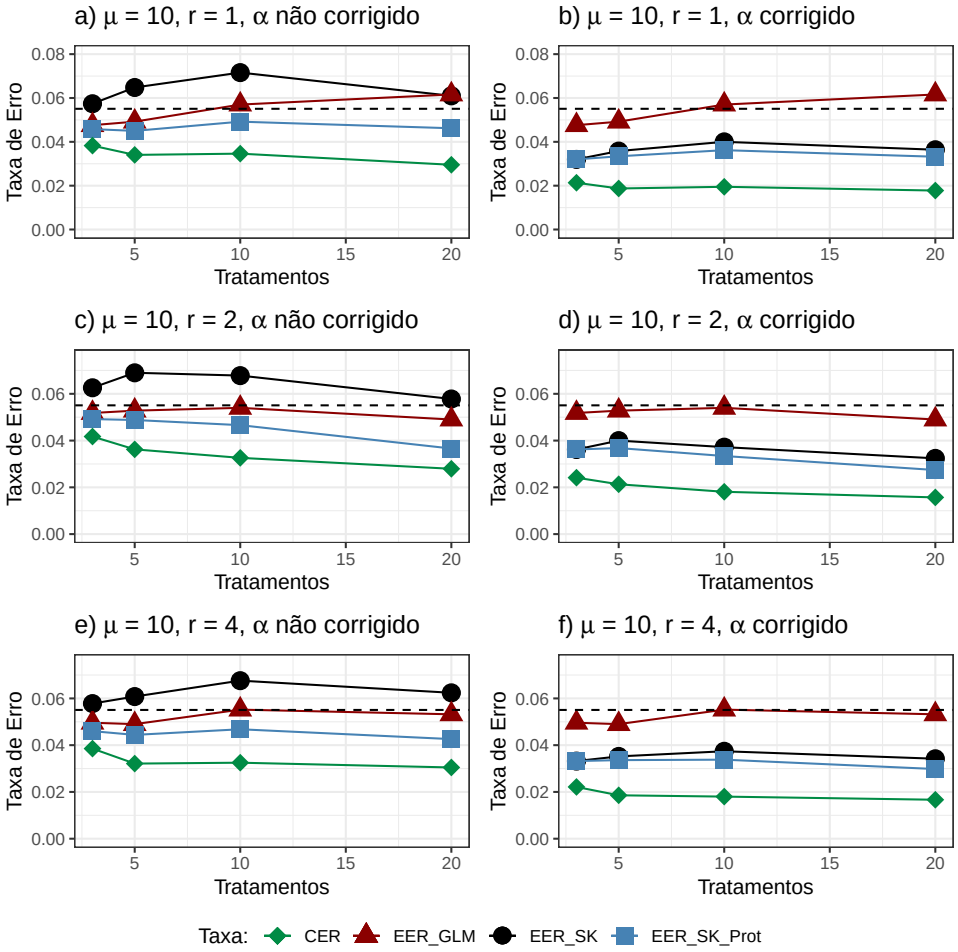

Figura 4: Taxas de erro tipo I avaliadas em 5000 simulações sob H_0, não utilizando e utilizando correção para o nível de significância, e considerando diferentes números de tratamentos e repetições, e o parâmetro Poisson $\mu = 10$ (EER_GLM: taxa de rejeição de H_0 para o modelo GLM; EER_SK: taxa de rejeição de H_0 para o método de Scott-Knott; EER_SK_Prot: taxa de rejeição de H_0 para o método de Scott-Knott protegido; CER: taxa de erro por comparação, para o método de Scott-Knott.

Considere-se, em um primeiro momento, o ajuste ao modelo GLM. Com $\mu = 1$ (Figura 1), este ajuste apresentou taxas de erro próximas do nível nominal de 5%, quando $r = 1$, mas apresentou valores elevados, consideravelmente acima do valor nominal, com $r = 2$ ou 4, notadamente com números maiores de tratamentos. Esta tendência se inverte, com valores maiores de μ (Figuras 2, 3 e 4): taxas significativamente maiores que 5% para $r = 1$, e um melhor controle do erro tipo I para maiores números de repetições. Isto ilustra como as taxas de rejeição de H_0 apresentam uma notável interação com os parâmetros da simulação, com um comportamento variável, conforme a combinação entre eles.

De qualquer maneira, é notável que haja situações em que a estatística de teste, para o modelo GLM, se desvie assim consideravelmente da distribuição teórica de qui-quadrado, ao menos no que se refere ao seu quantil 0,95 (pois se está considerando um nível de significância de 5%). A julgar pela recomendação prática de BILDER (2015), a aproximação de qui-quadrado pode não ser satisfatória para baixos valores de μ, mas, a julgar pelos resultados destas simulações, isto pode acontecer mesmo para maiores valores de μ, desde que $r = 1$ (Figuras 2, 3 e 4).

Quanto ao método Scott-Knott não protegido (método SK), e sem nenhuma correção do nível de significância, temos os valores das taxas EER e CER sempre nas colunas da esquerda das Figuras 1, 2, 3 e 4. Considere-se inicialmente a taxa EER. Chama a atenção de que esta taxa apresentou uma tendência consistente de ser significativamente superior ao nível nominal de 5%. Em alguns casos esta taxa foi notavelmente elevada, como na situação com $\mu = 1$, $r = 1$, e números elevados de tratamentos (Figura 1, gráfico (a)).

Este resultado causa alguma surpresa, dado que trabalhos anteriores (BARRETO & DEMÉTRIO, 1998) evidenciaram uma proximidade entre a distribuição da estatística de teste e a distribuição teórica de qui-quadrado. Porém, deve-se salientar que o estudo de BARRETO & DEMÉTRIO (1998) considerou a distribuição da estatística como um todo, enquanto que o presente trabalho abordou apenas um dos aspectos da distribuição (o quantil da cauda superior), e para um único valor de nível de significância (5%).

Contudo, em relação à taxa CER, também para o método SK com α não corrigido, tem-se que, de uma maneira bastante geral, tais taxas estiveram quase sempre próximas ou bem abaixo do valor nominal de 5% (colunas da esquerda das Figuras 1, 2, 3 e 4). Isto, de certa forma, promove uma certa tranquilidade no controle do erro tipo I, ao menos quando os objetivos da pesquisa residem mais nas comparações entre tratamentos, dois a dois, e não simplesmente na informação sobre se houve ou não a formação de grupos (à qual se refere a taxa EER).

De qualquer maneira, os resultados referentes às taxas EER sugerem a necessidade de algum procedimento para melhoria do controle do erro tipo I para o método SK, e a partir deste fato é que foram consideradas variações, a saber, o método SK com α corrigido, bem como o método de Scott-Knott protegido (SK_Prot, com ou sem a proteção do α).

Vamos considerar inicialmente o método SK com α corrigido. Com $\mu = 1$ (coluna da direita da Figura 1), este procedimento empírico de correção do α aparentemente melhorou consideravelmente o controle do erro tipo I. A única exceção correspondeu à situação com $t = 20$ e $r = 1$ (gráfico (b) da Figura 1). Estes resultados sugerem que o uso da correção do α possa corresponder a um procedimento interessante para a melhoria do controle do erro tipo I por experimento, em situações com baixos valores de μ e números de tratamentos até 10.

Contudo, considerando maiores valores de μ, o uso da correção do α promoveu consideráveis reduções das taxas EER, para o método SK, de uma forma geral (colunas da direita das Figuras 2, 3 e 4). Isto também não é desejado, pois taxas de rejeição de H_0 consideravelmente menores que o valor nominal geralmente estão associadas a um poder do teste reduzido. Ou seja, em situações em que *há* diferenças entre os tratamentos, a correção do α poderá dificultar sobremaneira a sua detecção.

O outro procedimento investigado, objetivando uma melhoria do controle do erro tipo I, foi o método de Scott-Knott protegido (SK_Prot). Neste procedimento, o algoritmo de agrupamento de médias só era considerado caso já tivesse ocorrido a rejeição de H_0 com o modelo GLM. Este procedimento foi avaliado tanto considerando como não considerando a correção do α.

As taxas de erro (EER) para o método SK_Prot, sem considerar, e considerando correção do α, estão apresentadas nas Figuras 1, 2, 3 e 4.

Com $\mu = 1$, e sem proteção do α, de uma maneira geral este procedimento também promoveu uma considerável melhoria do controle do erro tipo I (coluna da esquerda da Figura 1). A exceção que mais chama a atenção foi a situação com $t = 20$ e $r = 2$ (gráfico (c) da Figura 1). As taxas de erro para este método, com correção do α, estão apresentadas nos gráficos (b), (d) e (f) da Figura 1. Houve uma tendência geral de tais taxas acompanharem aquelas do método SK. A exceção marcante acontece com $t = 20$ e $r = 1$. Somando-se à recomendação anterior de se corrigir o α com baixos valores de μ, este resultado sugere que, com número elevados de tratamentos (acima de 10), além da correção do α, talvez seja também interessante se trabalhar com o método de Scott-Knott protegido.

Com valores maiores de μ, contudo, e de maneira semelhante ao acontecido com o método SK, a correção do α parece ter diminuído excessivamente as taxas de erro para o método SK_Prot (gráficos (b), (d) e (f) das Figuras 2, 3 e 4). Isto, conforme comentado anteriormente, não é desejado. Por outro lado, as taxas deste método *sem* a correção do α (gráficos (a), (c) e (e) das Figuras 2, 3 e 4), estiveram bem mais próximas do nível nominal, e geralmente abaixo deste. A exceção ocorre com $r = 1$ e $\mu = 3$ (gráfico (a) da Figura 2), mas mesmo aqui os valores, embora superiores ao nível nominal, estiveram relativamente próximos. Assim, estes resultados sugerem que, para valores não baixos de μ, o uso do método SK_Prot (sem correção do α), pode corresponder a uma alternativa interessante para um melhor controle do erro tipo I.

4.2 Ilustração: Dados COVID-19

Finalmente, o método aglomerativo de Scott-Knott adaptado a dados com distribuição Poisson foi ilustrado utilizando dados referentes a óbitos devido à Covid-19, nos 20 municípios de Minas Gerais com mais alunos da Universidade Federal de Ouro Preto. Por se tratarem de dados com número de "tratamentos" (municípios) elevado, utilizou-se aqui o método de Scott-Knott protegido.

A Tabela 9 se refere a este ajustamento para o número total de óbitos, para estes vinte municípios, até a data de 02/08/2021. Na segunda coluna, são apresentadas as estimativas por habitante, enquanto que a última coluna as apresenta no formato de taxas por 10.000 habitantes.

Como podemos observar nesta primeira análise, o método dividiu os 20 municípios em 7 grupos distintos, sendo que os municípios contidos em um mesmo grupo são considerados estatisticamente iguais, ao nível de 5% de probabilidade.

A mesma analise foi realizada para o número de óbitos apenas do mês de janeiro de 2021, cujos resultados estão apresentados na Tabela 10. Conforme se observa, aqui houve a subdivisão dos 20 municípios em 4 grupos distintos.

Finalmente, o método de Scott-Knott protegido foi aplicado ao número de óbitos apenas do mês de fevereiro de 2021, cujos resultados estão apresentados na Tabela 11. Aqui houve a formação de 3 grupos distintos. Esta redução no número de grupos, de janeiro para fevereiro, aparentemente se deu devido ao aumento do número de óbitos por Covid-19 no município de Sete Lagoas, aproximando sua estimativa de outros municípios, deixando de compor um grupo em separado.

Tabela 9: Número total de óbitos por Covid-19 até a data de 02/08/2021 nas 20 cidades com o maior número de alunos da Universidade Federal de Ouro Preto.

Tratamento	Estimativas/hab.	Scott-Knott	População	Taxa
Santa Bárbara	0,0011391	1	31604	11,39
Ouro Branco	0,0012793	1	39867	12,79
Ouro Preto	0,0015961	1	74558	15,96
Mariana	0,0015990	1	61288	15,99
Barão de Cocais	0,0020082	2	32866	20,08
Conselheiro Lafaiete	0,0020755	2	129606	20,76
Congonhas	0,0021696	2	55309	21,7
Divinópolis	0,0023668	3	240408	23,67
Sete Lagoas	0,0023777	3	241835	23,78
Belo Horizonte	0,0024723	3	2521564	24,72
João Monlevade	0,0025990	3	80416	25,99
Contagem	0,0026743	4	668949	26,74
Itabirito	0,0027838	4	52446	27,84
Betim	0,0029228	5	444784	29,23
Coronel Fabriciano	0,0029830	5	110290	29,83
Itabira	0,0029858	5	120904	29,86
Ponte Nova	0,0033069	6	59875	33,07
Ipatinga	0,0033382	6	265409	33,38
Timóteo	0,0036105	6	90568	36,11
Governador Valadares	0,0042093	7	281046	42,09

Tabela 10: Número de óbitos por Covid-19 no mês de janeiro do ano de 2021 nas 20 cidades com o maior número de alunos da Universidade Federal de Ouro Preto.

Tratamento	Estimativas/hab.	Scott-Knott	População	Taxa
Sete Lagoas	0,0000248	1	241835	0,25
Santa Bárbara	0,0000633	2	31604	0,63
Ouro Preto	0,0000805	2	74558	0,8
Conselheiro Lafaiete	0,0000926	2	129606	0,93
Itabira	0,0001075	3	120904	1,08
Mariana	0,0001305	3	61288	1,31
Barão de Cocais	0,0001521	3	32866	1,52
Belo Horizonte	0,0001618	3	2521564	1,62
Divinópolis	0,0001705	3	240408	1,71
Betim	0,0001731	3	444784	1.73
Ouro Branco	0,0001756	3	39867	1,76
Contagem	0,0001779	3	668949	1,78
Congonhas	0,0001808	3	55309	1,81
Ponte Nova	0,0001837	3	59875	1,84
Itabirito	0,0002288	4	52446	2,29
Governador Valadares	0,0002455	4	281046	2,46
Coronel Fabriciano	0,0002811	4	110290	2,81
Ipatinga	0,0002864	4	265409	2,86
Timóteo	0,0003312	4	90568	3,31
João Monlevade	0,0004352	4	80416	4,35

Tabela 11: Número total de óbitos por Covid-19 no mês de fevereiro do ano de 2021 nas 20 cidades com o maior número de alunos da Universidade Federal de Ouro Preto

Tratamento	Estimativas/hab.	Scott-Knott	População	Taxa
Santa Barbara	6.52e-15	1	31604	6.52e-11
Mariana	0,0000326	1	61288	0,33
Barão de Cocais	0,0000609	1	32866	0,61
Sete Lagoas	0,0000703	1	241835	0,7
Congonhas	0,0000723	1	55309	0,72
Divinópolis	0,0000998	1	240408	1
Ouro Preto	0,0001073	1	74558	1,07
Contagem	0,0001405	2	668949	1,41
Conselheiro Lafaiete	0,0001466	2	129606	1,47
Ouro Branco	0,0001505	2	39867	1,51
Ipatinga	0,0001545	2	265409	1,54
Betim	0,0001551	2	444784	1,55
Coronel Fabriciano	0,0001723	2	110290	1,72
Itabira	0,0001820	2	120904	1,82
Belo Horizonte	0,0001904	2	2521564	1,9
Ponte Nova	0,0002004	2	59875	2
João Monlevade	0,0002363	3	80416	2,36
Itabirito	0,0002479	3	52446	2,48
Timóteo	0,0002981	3	90568	2,98
Governador Valadares	0,0002989	3	281046	2,99

Ao longo destas 3 análises, é interessante notar que os pares de municípios: a) Santa Bárbara e Ouro Preto; b) Barão de Cocais e Congonhas; e c) Betim e Itabira, sempre estiveram presentes em um mesmo grupo, corroborando que se trate de pares de municípios muito semelhantes quanto à ocorrência de óbitos por Covid-19.

Deve-se ressaltar que a presente abordagem de agrupamento de municípios não tem, em princípio nenhum caráter espacial, ou seja, não se trata de uma técnica de Estatística Espacial. Os municípios pertencentes a um mesmo grupo provavelmente guardam semelhanças em relação a outras variáveis (políticas públicas semelhantes, por exemplo), que não a maior ou menor proximidade espacial.

5 Considerações Finais

Este trabalho procurou avaliar e ilustrar o método aglomerativo de Scott-Knott para dados de contagem, com distribuição Poisson.

Os resultados de simulação sob H_0, ao menos considerando um nível de significância de 5%, sugerem algumas conclusões, tais como:

1. Em várias situações, a taxa de erro por experimento (EER) do ajuste ao modelo GLM se mostrou significativamente superior ao nível nominal de 5%, notadamente para $r = 1$, maiores números de tratamentos, e maiores valores de μ.

2. A taxa EER para o método de Scott-Knott apresentou uma tendência geral de ser superior ao nível de significância nominal de 5%.

3. A taxa CER para o método de Scott-Knott apresentou uma tendência geral de ser inferior ao nível de significância nominal de 5%, à exceção para situações com $r = 1$, elevados números de tratamentos, e baixo valor de μ.

4. O uso da correção do α pode corresponder a um procedimento interessante para a melhoria do controle do erro tipo I por experimento, em situações com baixos valores de μ e números de tratamentos até 10.

5. Para baixos valores de μ recomenda-se utilizar a correção do α, sendo também interessante proteger o método de Scott-Knott, com elevados números de tratamentos (acima de 10).

6. Para valores intermediários ou maiores de μ, recomenda-se utilizar o método de Scott-Knott protegido.

Como limitações do estudo de simulação realizado, deve-se apontar que seria interessante um aprofundamento da investigação, considerando outros valores de α, além de

5%. Também seria interessante simular situações sob H_0 falsa, para um estudo acerca do poder do teste, também considerando diferentes valores de α.

A aplicação do método a dados de óbitos por Covid-19 em 20 municípios de Minas Gerais ilustrou as potencialidades da técnica, mesmo para situações não experimentais.

6 Referências Bibliográficas

AGRESTI, A. **Categorical Data Analysis**, 2.ed. New Jersey, Editora John Wiley and Sons, 2002. 721p.

BARRETO, M.C.M. **Uma extensão da técnica AID em modelos lineares generalizados**. Piracicaba, ESALQ, 1993. 200p. (Tese de Doutorado)

BARRETO, M.C.M. & DEMÉTRIO, C.G.B. Um estudo da distribuição assintótica da medida de homogeneidade entre grupos da extensão da técnica de agrupamento AID considerando a distribuição Poisson. **Revista de Matemática e Estatística**, São Paulo, v.16, p.191-207, 1998.

BEARZOTI, E. **Apostila: Planejamento de Experimento I**. Ouro Preto: 2021. 220p.

BILDER, C.R. **Analysis of Categorical Data With R**. Boca Raton: CRC Press, 2015.

CASELLA, G. & BERGER, R.L. **Inferência Estatística**. São Paulo, Cengage Learning, 2010.

FERREIRA, D. F. Sisvar: a computer statistical analysis system. **Ciência e Agrotecnologia**, n.35, p.1039-1042, 2011.

LENTH, R.V. (2020). **emmeans: Estimated Marginal Means, aka Least-Squares Means**. R package version 1.5.3. `https://CRAN.R-project.org/package=emmeans`.

MONTGOMERY, D.C. **Design and Analysis of Experiments**. 8.ed. Wiley, 2012.

MORGAN, J.N. & SONQUIST, J.A. Problems in the analysis of survey data. **Journal of the American Statistical Association**, Washington, 58: 415-434, 1963.

NELDER, J.A, & WEDDERBURN, R.W.M. Generalized linear models. **Journal of the Royal Statistical Association A**, Londres, n.135, p.370-384, 1972.

PIMENTEL GOMES,F. **Curso de Estatística Experimental**. 15 ed. Piracicaba: FEALQ, 2009.

RAMALHO, M.A,P.; FERREIRA, D.F. & OLIVEIRA, A.C. **Experimentação em Genética e Melhoramento de Plantas**. Lavras: Editora UFLA, 2000. 326p.

R CORE TEAM (2020). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

SCOTT, A. J., & KNOTT, M. A. A cluster analysis method for grouping means in the analysis of variance. **Biometrics**, Alexandria, 30: 507-512, 1974.

APÊNDICE A – Códigos R

Função Scott-Knott

O rotina (em linguagem **R**) para o método aglomerativo de Scott-Knott, adaptado a dados Poisson, é apresentado abaixo. Esta função foi construída considerando conjuntos de dados balanceados (mesmo número r de repetições para cada tratamento) e o delineamento inteiramente casualizado.

A função não foi elaborada considerando aspectos para uma melhor interface com o usuário. A entrada de dados deve sempre ser feita através de um *data.table*, em que os tratamentos correspondem a um fator de nome *trat*, cujos níveis devem ser identificados como $1, 2, \ldots t$, e a variável resposta identificada como y.

O usuário deve fornecer o nível de significância (corrigido ou não) em um dos argumentos da função.

Nesta rotina apresentada, está sendo considerada uma situação em que não há necessidade de termo *offset*, e a função de ligação utilizada foi a logarítmica.

Um típico conjunto de dados, no formato de entrada para a função, corresponderia ao exemplo fictício abaixo, considerando uma situação com $t = 4$ tratamentos e $r = 3$ repetições.

```
library(data.table)
Dados <- data.table(trat = as.factor(c(1,1,1,
                                       2,2,2,
                                       3,3,3,
                                       4,4,4)),
                    y = c(4,5,5,
                          3,2,5,
                          8,10,7,
                          12,9,10))
```

Para um certo nível de significância *alfa* (corrigido ou não), a função é dada por:

```
scottknott <- function(dados,alfa) {
  # numero de repeticoes e de tratamentos
  reps <- dados[trat == "1",.N]
  ntrat <- nlevels(dados$trat)

  # analise principal:
  mod <- glm(y ~ trat,family = poisson(link = "log"), data=dados)
  # a matriz abaixo é para "captar" as médias dos tratamentos:
  A <- rbind(t(c(1,rep(0,ntrat-1))),cbind(rep(1,ntrat-1),diag(ntrat-1)))
  # vetor de médias de tratamento já ordenadas:
  Medias <- data.table(trat = as.factor(seq(1:ntrat)),
                       meds = exp(A%*%mod$coefficients),Gr = rep(1,ntrat))
  Medias <- Medias[order(meds.V1),]

  # Para iniciar a partição em grupos:
  ngrupos <- 1
  Sig <- TRUE
  MediasF <- Medias

  while (Sig) {
    Sig <- F
    lk0 <- 0
    dev2 <- 0
    for (i in 1:ngrupos) {
      medias <- Medias[Gr==i,]

      if (nrow(medias) > 1) {

        GLp <- nrow(medias)/(pi-2)
        K <- pi/2/(pi-2)

        for (ii in c(1:(nrow(medias)-1))) {
          MediasTemp <- Medias
          mediasTemp <- medias
          MediasTemp$Gant <- Medias$Gr

          mediasTemp[1:ii,Gr:=ngrupos+1]

          for (j in 1:nrow(mediasTemp) ) {

            for (k in 1:nrow(MediasTemp)) {
```

```
            if (as.numeric(mediasTemp[j,trat])==as.numeric(MediasTemp[k,trat])) {
              MediasTemp$Gr[k] <- mediasTemp$Gr[j]
            }
          }
        }
        dadosTemp <- dados
        if (nlevels(as.factor(MediasTemp$Gant)) == 1) {
          dadosTemp$Gr <- expand.grid(rep(1,reps),MediasTemp[order(trat),]$Gr)$Var2
          modt <- glm(y ~ as.factor(Gr)+trat,family = poisson(link = "log"),
                      data=dadosTemp)
          if ((anova(modt, test = "Chisq")$Deviance[2] > lk0) &
              (pchisq(anova(modt, test = "Chisq")$Deviance[2]*K,
                      GLp,lower.tail = F) < alfa)) {
            lk0 <- anova(modt, test = "Chisq")$Deviance[2]
            dev2 <- anova(modt, test = "Chisq")$Deviance[3]
            Sig <- TRUE
            MediasF <- MediasTemp
          }
        }
        else {
          dadosTemp$Gr <- expand.grid(rep(1,reps),MediasTemp[order(trat),]$Gr)$Var2
          dadosTemp$Gant <- expand.grid(rep(1,reps),MediasTemp[order(trat),]$Gant)$Var2
          modt <- glm(y ~ as.factor(Gant)+as.factor(Gr)+trat,family = poisson(link = "log"),
                      data=dadosTemp)
          if ((anova(modt, test = "Chisq")$Deviance[3] > lk0) &
              (pchisq(anova(modt, test = "Chisq")$Deviance[3]*K,
                      GLp,lower.tail = F) < alfa)) {
            lk0 <- anova(modt, test = "Chisq")$Deviance[3]
            dev2 <- anova(modt, test = "Chisq")$Deviance[4]
            Sig <- TRUE
            MediasF <- MediasTemp
          }
        }
      }
    }
  }
  if (Sig) {
    ngrupos <- ngrupos+1
    Medias <- MediasF
  }
}
Medias$G3 <- rep(0,nrow(Medias))
Medias$G3[1] <- 1
```

```
  grupo <- 1
  anterior <- Medias$Gr[1]
  for (i in 2:nrow(Medias)) {
    if (Medias$Gr[i] != anterior) {
      grupo <- grupo + 1
      anterior <- Medias$Gr[i]
    }
    Medias$G3[i] <- grupo
  }
  result <- data.table(Tratamento = Medias$trat, Estimativas = Medias$meds.V1,
                       ScottKnott = Medias$G3)
}

# no exemplo, 2 grupos sao formados:
print(scottknott(dados = Dados,alfa = 0.05))
```

```
   Tratamento Estimativas ScottKnott
1:          2    3.333333          1
2:          1    4.666667          1
3:          3    8.333333          2
4:          4   10.333333          2
```

Rotina de Simulação

Como exemplo de rotina de simulação utilizada, é apresentada abaixo a rotina referente à situação com 20 tratamentos, 4 repetições por tratamento, $\mu = 10$, e sem correção do α.

```
library(data.table)
require(combinat)
library(berryFunctions)

# Parametros
situacao <- 99
stalfa <- "Nao"
simulas <- 5000
arqq <- "/Meu_path/Simula99_Ruins.csv"
```

```
ruins <- data.frame(amostra=0.01,ConvSK=0.01,ConvGLM=0.01, Tratamento=0.01,Y=0.01)
write.table(ruins, file = arqq, sep=";", row.names = F)

ntrats <- 20
nreps <- 4
theta <- 10
nivel <- 0.05

# Inicializadores

# controle de convergencia:
EventA <- 0
EventB <- 0
EventC <- 0
EventD <- 0
EventE <- 0
EventF <- 0
EventG <- 0

# taxa de erro por comparacao (Scott-Knott):
cer <- 0

# para sempre dar os mesmos valores simulados:
set.seed(64573189)
fechou <- TRUE
contador<-0
while (fechou) {
  # "Esqueleto" do data.frame
  contador <- contador+1
  dados <- data.frame(trat = as.factor(expand.grid(rep(1,nreps),seq(1:ntrats))$Var2),
                      y = rep(0,ntrats*nreps))
  dados <- as.data.table(dados)

  for (i in 1:(ntrats*nreps)){
    dados$y[i] <- rpois(1,theta)
  }
  nivel2 <- 1-sqrt(1-nivel)
  if (is.error(scottknott(dados,nivel))==FALSE) {
    modt <- glm(y ~ trat,family = poisson(link = "log"),data=dados)
    medias <- scottknott(dados,nivel)
    medias$grupo <- as.factor(medias$ScottKnott)
    if (nlevels(medias$grupo) > 1) {
      if (pchisq(anova(modt, test = "Chisq")$Deviance[2],
```

```
        anova(modt, test = "Chisq")$Df[2],lower.tail = F) < nivel) EventE <- EventE+1
      else EventD <- EventD+1
    }
    else {
      if (pchisq(anova(modt, test = "Chisq")$Deviance[2],
        anova(modt, test = "Chisq")$Df[2],lower.tail = F) < nivel) EventC <- EventC+1
      else EventB <- EventB+1
    }

    for (i in 1:(nrow(medias)-1)) {
      for (j in (i+1):nrow(medias)) {
        if (medias$grupo[i]!=medias$grupo[j]) cer <- cer +1
      }
    }
  }
  else {
    if (is.error(glm(y ~ trat,family = poisson(link = "log"),data=dados))==FALSE) {
      modt <- glm(y ~ trat,family = poisson(link = "log"),data=dados)
        ruins2 <- data.frame(amostra=rep(contador,nrow(dados)),
                             ConvSK=rep(0,nrow(dados)),
                             ConvGLM=rep(1,nrow(dados)),
                             Tratamento = dados$trat,
                             Y = dados$y)
        write.table(ruins2, file = arqq, sep=";", row.names = F,append = TRUE)
      if (pchisq(anova(modt, test = "Chisq")$Deviance[2],
        anova(modt, test = "Chisq")$Df[2],lower.tail = F) < nivel) EventG <- EventG+1
      else EventF <- EventF+1
    }
    else {
      EventA <- EventA+1
        ruins2 <- data.frame(amostra=rep(contador,nrow(dados)),
                             ConvSK=rep(0,nrow(dados)),
                             ConvGLM=rep(0,nrow(dados)),
                             Tratamento = dados$trat,
                             Y = dados$y)
        write.table(ruins2, file = arqq, sep=";", row.names = F,append = TRUE)

    }
  }

  if (contador==simulas) fechou <- FALSE

}
```

```
EER_SK <- (EventD+EventE)/simulas
EER_GLM <- (EventC+EventE+EventG)/simulas
EER_GLM_Prot <- EventE/simulas
Total <- EventA+EventB+EventC+EventD+EventE+EventF+EventG
CER <- cer/(dim(combn(ntrats,2))[2]*simulas)

arq <- "/Meu_Path/Resultados2.csv"
saidas  <- read.table(arq,  sep = ";", dec =".",
                          header=T, colClasses = c("numeric","numeric","numeric","numeric",
                                                   "numeric","numeric","numeric","numeric",
                                                   "character"))
saidas$Alfa <- as.character(saidas$Alfa)
saidas <- rbind(saidas, list(situacao,ntrats,nreps,theta,EER_SK,EER_GLM,
                             EER_GLM_Prot,CER,stalfa,EventA,EventB,
                             EventC,EventD,EventE,EventF,EventG))
write.table(saidas, file = arq, sep=";", row.names = F)
```