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

Gasto Ótimo da Probabilidade do Erro Tipo I em Processos Estocásticos Binomiais

Natália da Conceição Figueiredo dos Anjos

Ouro Preto - MG
Julho 2019

Natália da Conceição Figueiredo dos Anjos

Gasto Ótimo da Probabilidade do Erro Tipo I em Processos Estocásticos Binomiais

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

Prof. Dr. Ivair Ramos Silva

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

Ouro Preto-MG

Julho 2019

A58g Anjos, Natália da Conceição Figueiredo dos.
Gasto ótimo da probabilidade do erro tipo I em processos estocásticos binomiais [manuscrito] / Natália da Conceição Figueiredo dos Anjos. - 2019.

24f.: il.: color; grafs; tabs.

Orientador: Prof. Dr. Ivair Ramos Silva.

Monografia (Graduação). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Departamento de Estatística.

1. Gasto de alfa . 2. Análise sequencial. I. Silva, Ivair Ramos. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 519.23

Catalogação: ficha.sisbin@ufop.edu.br

Monografia de Graduação sob o título *Gasto Ótimo da Probabilidade do Erro Tipo I em Processos Estocásticos Binomiais* apresentada por Natália da Conceição Figueiredo dos Anjos e aceita pelo Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto, sendo aprovada por todos os membros da banca examinadora abaixo especificada:

Prof. Dr. Ivair Ramos Silva
Orientador
Departamento de Estatística
Universidade Federal de Ouro Preto

Prof. Dr. Tiago Martins Pereira
Departamento de Estatística
Universidade Federal de Ouro Preto

Prof. Dr. Eduardo Bearzoti
Departamento de Estatística
Universidade Federal de Ouro Preto

Ouro Preto-MG, 3 de Julho de 2019.

Agradecimentos

Gostaria de agradecer minha família, especialmente minha mãe, que fez de tudo para tornar os momentos difíceis mais brandos. Obrigado, meu irmão Lucas, por ser tão companheiro e sempre me incentivar.

Agradeço as minhas amigas Carolina, Débora e Juliana, que do início ao fim tornaram este processo mais prazeroso e cheio de companheirismo. Vocês foram fundamentais para minha formação, por isso merecem o meu eterno agradecimento.

Agradeço a todos os professores do departamento de estatística por me proporcionar o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional. Agradeço em especial ao meu orientador pela confiança que depositou em mim e por me proporcionar um conhecimento maior.

Agradeço, enfim, a todos os colegas, amigos e professores, que contribuíram direta ou indiretamente com minha formação

Gasto Ótimo da Probabilidade do Erro Tipo I em Processos Estocásticos Binomiais

Autor: Natália da Conceição Figueiredo dos Anjos
Orientador: Prof. Dr. Ivair Ramos Silva

Resumo

A função gasto de alfa desempenha um papel importante na análise estatística sequencial, pois é utilizada para estabelecer, de antemão, a intensidade de como devemos utilizar a probabilidade do erro tipo I, ao longo dos múltiplos testes. A solução ótima para a função gasto de alfa pode ser obtida através de programação linear, ou seja, obtém-se uma solução factível que apresenta o melhor valor da função gasto de alfa por meio da programação linear. No entanto, apesar de ser exata, este tipo de manipulação matemática não é trivial e tem um custo computacional elevado. Por isso, uma questão em aberto é a identificação de possíveis funções analíticas que possam aproximar a função gasto ótimo de alfa de maneira satisfatória. Diante disso, este trabalho teve como objetivo identificar possíveis funções analíticas que possam obter tal aproximação da função gasto ótimo de alfa de maneira satisfatória. Assim sendo, os principais objetivos deste projeto foram: (i) fazer um levantamento bibliográfico sobre as tentativas recentes de se obter o procedimento para teste sequencial ótimo em processos binomiais; (ii) estudar as aplicações do teste sequencial em processos binomiais; (iii) implementar computacionalmente os resultados analíticos para avaliação numérica utilizando linguagem R; (iv) identificar as possíveis funções analíticas que possam aproximar a função gasto ótimo de alfa de maneira satisfatória. Todos os objetivos deste projeto foram alcançados, podendo-se evidenciar o principal destes que era obter uma expressão analítica que possibilite uma aproximação da função gasto ótimo de alfa de maneira simplificada e eficiente. Tal aproximação foi obtida de forma satisfatória utilizando a distribuição Gumbel.

Palavras-chave: Gasto de alfa, Gumbel, Análise sequencial.

Optimum spending of Probability of Type I Error in Binomial Stochastic Processes

Author: Natália da Conceição Figueiredo dos Anjos
Advisor: PhD. Ivair Ramos Silva

Abstract

The alpha spending function plays an important role in sequential statistical analysis, because it is used to establish in advance the intensity of how we should use the probability of type I error over of multiples tests. The optimal solution for the alpha spending function can be obtained through linear programming, that is, a feasible solution is obtained that presents the best value of the alpha spending function through linear programming. However, while accurate, this type of mathematical manipulation is not trivial and has a high computational cost. Therefore, an open question is the identification of possible analytical functions that can approximate the optimal alpha spending function satisfactorily. Therefore, the objective of this work was to identify possible analytical functions that can obtain such an approximation of the optimal alpha spending function in a satisfactory way. Thus, the main objectives of this project were: (i) to make a bibliographical survey about the recent attempts to obtain the procedure for optimal sequential testing in binomial processes; (ii) studying the applications of the sequential test in binomial processes, (iii) computationally implementing the analytical results for numerical evaluation using language R; (iv) identify possible analytical functions that can approximate the optimal alpha spending function satisfactorily. However, all the objectives of this project were achieved, and it was possible to highlight the main one of these that was to obtain an analytical expression that allows an approximation of the alpha spending function in a simplified and efficient way, and such an approximation was obtained in a satisfactory way using the Gumbel distribution.

Keywords: Alpha spending, Gumbel Distribution, Sequential Analysis.

Lista de figuras

Lista de tabelas

Sumário

1 Introdução

A análise sequencial consiste em monitorar informações amostrais que estão sendo acumuladas ao longo do tempo. Esse monitoramento é realizado através de técnicas de inferência estatística, tais como testes de hipóteses, estimação pontual ou intervalar. No caso dos testes, estes são aplicados repetidas vezes de acordo com que os dados vão se acumulando. Este tipo de teste não baseado em amostra promove inferências mais precisas, devido à possibilidade de obter decisões mais rápidas ao compararmos com testes que são fundamentados em amostras. Por esta razão a análise sequencial é muito utilizada em diversas áreas da ciência como, por exemplo, ensaios clínicos (JENNISON; TURNBULL, 2000), vigilância de drogas e vacinas pós-comercializadas (DAVIS et al., 2005; LIEU et al., 2007; YIH et al., 2011; LEITE; ANDREWS; THOMAS, 2016).

A distribuição binomial apresenta um desempenho relevante em muitas aplicações de análise sequencial, pois tal distribuição é útil para estudos de coortes equivalentes, incluindo projetos adequados de propensão e também é muito importante para a análise autocontrolada, em que o objetivo é comparar o período em que um indivíduo é exposto a um medicamento administrado contra um período não exposto, antes de administrar o medicamento, no mesmo indivíduo. Outra situação em que um modelo binomial é aplicável é o caso em que os pacientes expostos a uma droga são comparados com pacientes não expostos pareados.

Convencionalmente a forma abordada em testes sequenciais com dados binomiais utiliza monitoramento de uma estatística de teste em comparação com o valor crítico, chamado de "limiar de sinalização". Este tipo de abordagem é utilizada em métodos clássicos, como o teste SPRT de Wald (WALD, 1945), o teste de Pocock (POCOCK, 1977) e o teste de O'Brien-Fleming (O'BRIEN; FLEMING, 1979). Alternativamente, existe uma abordagem em que a regra de decisão é construída através do gasto do erro tipo I, esta abordagem é extensamente explorada no livro do Jennison e Turnbull (2000). O método desenvolvido por Silva e Kulldorff (No prelo) tem como ideia utilizar esta abordagem, com o intuito de obter o limiar de sinalização em alguma escala desejável.

O gasto do erro tipo I é uma função não decrescente, digamos S (n), tomando valores no intervalo de [0; α], no qual α é o nível de significância global até o final da análise, e n é o índice de tempo definido como uma fração do total do monitoramento, N. Dita S (n), antecipadamente, como a taxa na qual a probabilidade de erro tipo I deve ser usada no decorrer dos vários testes sequenciais. Uma função bastante utilizada para isto é a forma do tipo potência:

$$S(n) = \alpha \times \left(\frac{n}{N}\right)^{\rho}, \rho > 0,\ n \in (1, N]. \tag{1.1}$$

A função gasto de alfa é utilizada para estabelecer a intensidade de como devemos utilizar a probabilidade de erro tipo I em cada teste realizado durante o monitoramento sequencial, ou seja, nela define-se um limite superior para a probabilidade associada à ocorrência de erro tipo I para cada teste efetuado. Esta função é frequentemente utilizada na análise sequencial em ensaios clínicos, mas recentemente cada vez mais utiliza-se também no monitoramento de drogas e vacinas pós-comercializados.

No método abordado no artigo *"Optimal Alpha Spending for Sequential Analysis With Binary Data"* (SILVA; KULLDORFF, No prelo) tem-se que a função gasto ótimo de alfa é calculada por meio de um algoritmo de programação linear, a partir do pacote *"Sequential"* (SILVA; KULLDORFF, 2019) do software R (R Core Team, 2017), mas este método pode tomar muito tempo para tempos grandes de monitoramento e, além disso, é teoricamente complexo. Assim, é de interesse estudar formas analíticas simples que possam oferecer boas aproximações para a função gasto ótimo de alfa. Por isso, esse trabalho tem como objetivo encontrar uma expressão analítica que aproxime a função gasto ótimo de alfa de forma eficiente e simplificada. Para realizar tal fato foi necessário inicialmente fazer um levantamento bibliográfico sobre as tentativas recentes de se obter o procedimento para teste sequencial ótimo em processos binomiais, para que, através das implementações computacionais dos resultados analíticos, seja possível identificar prováveis funções analíticas que possam aproximar a função gasto ótimo de alfa de maneira satisfatória.

Esse trabalho está organizado da seguinte maneira: o capítulo a seguir apresenta os métodos utilizados de base para esta pesquisa. O capítulo 3 mostra os cenários utilizados e os resultados obtidos a partir desses cenários. Por último são apresentadas as considerações finais.

2 Referencial Teórico

Apresenta-se neste capítulo os principais métodos mencionados anteriormente, que serviram como base para este trabalho.

2.1 Teste Sequencial da Razão de Probabilidade de Wald (SPRT)

O teste sequencial da razão de probabilidade é um método sequencial contínuo. Este teste foi introduzido em 1945 por Wald (1945), sendo baseado na razão de verossimilhança. Um ponto positivo do SPRT é que pode ser usado para distintas distribuições de probabilidade. Uma grande contribuição de Wald foi instituir uma regra exata para escolha dos limites de sinalização superiores e inferiores, no qual estes limites estão em função das probabilidades dos erros tipo I e tipo II.

Wald e Wolfowitz (1948) provaram que o teste SPRT é minimax, ou seja, é um método para minimizar a possível perda máxima no sentido do tamanho da amostra esperada. Em aplicações reais, nas quais normalmente o truncamento é estabelecido por conveniência, esta propriedade minimax não é válida e, além disso, a análise pode ser interrompida em um tempo inviável para algumas aplicações, e isto acaba sendo uma limitação desse método.

2.2 Gasto ótimo de alfa para dados binários

Para estabelecer uma terminologia amigável, a motivação subjacente aos estudos de caso-controle de experimentos clínicos será adotada neste trabalho, o que ocorre principalmente porque os termos "expostos", "não expostos", "casos"e "controles"são familiares para os profissionais de análises sequenciais em geral. Mas, todos os resultados são igualmente válidos para qualquer problema de teste sequencial utilizando processos estocásticos

binomiais.

A solução sugerida por Silva e Kulldorff (No prelo) vem do fato que os métodos convencionais são apenas casos particulares de uma estrutura mais geral, a abordagem do limiar de sinalização aleatória. A partir desse fato é possível identificar uma reparametrização que transforma o problema original, que é de otimização inteira não-linear em alta dimensão e múltiplas restrições, em um problema de programação linear regular.

Com base no ajuste do limiar aleatório, o critério de teste pode ser projetado de forma que cada resultado possível do processo binomial possa indicar a rejeição de H_0 com uma certa probabilidade. Por meio da "estratégia da moeda"isso foi construído, funcionando da seguinte forma: no primeiro teste, realizado no tempo n_1, é "arremessada uma moeda"com probabilidade $\theta_{n_1,c}$ para cara, na qual c corresponde ao número de casos observados em n_1. Rejeite H_0 se cara for o resultado; caso contrário, continue com a vigilância. Este procedimento é repetido nos tempos de teste posteriores, ou seja, considerando que H_0 não foi rejeitada até o teste $(n-1)$ e, denotando c como o número de casos observados até o n-ésimo teste, arremessa-se a moeda com probabilidade $\theta_{n,c}$ para cara. Rejeite H_0 se cara for o resultado; caso contrário, continue com a vigilância se $n < N$, ou pare e não rejeite H_0 se $n = N$.

A estratégia da moeda é utilizada como estrutura matemática para converter a função poder e as medidas estatísticas relacionadas em expressões lineares mais simples. A ideia é utilizar os gastos ótimos de alfa obtidos a partir dessa técnica para obter o limiar não aleatório convencional em uma escala desejável, como a estatística do teste da razão de verossimilhança ou a escala das contagens binomiais. O critério dos métodos convencionais é um caso particular da estratégia da moeda, em que b_n é um limiar de sinalização, e a configuração implícita é dada por:

$$\theta_{n,c} = I(c \geq b_n).$$

.

Seja X_n o número de eventos da população exposta (casos) quando um total de n eventos é observado. Suponha que $Y_n = X_n - X_{n-1} \sim Bernoulli\,(p(RR))$, no qual $Y_1, \cdots, Y_n$ são independentes para cada n, e $Y_0 := 0$, no qual:

$$p(RR) = 1/(1 + z/RR). \tag{2.1}$$

A dependência de $p(RR)$ após RR é explicita. Assim, buscando uma notação sem perda

de generalidade, a probabilidade de Bernoulli $p(RR)$ será denotada simplesmente por p ao longo deste trabalho.

Observe que z representa a relação população não-exposta/população exposta para cada evento. Quando se quer comparar a hipótese de um risco relativo pequeno a aceitável (RR ≤ 1), contra um risco possivelmente elevado (RR> 1), a ideia é rejeitar H_0 apenas para grandes valores de X_n. Neste sentindo, define-se:

$$\begin{aligned} k_n &= \min\{c \in \mathbb{N} : Pr[X_n \geq c|RR = 1] \leq \alpha\}, \\ n_1 &= \min\{n \in \mathbb{N} : Pr[X_n \geq n|RR = 1] \leq \alpha\}. \\ \text{Para } n &\geq n_1 \text{ e } c \geq k_n, \text{ define } 0 \leq \theta_{n,c} \leq 1, \text{ mas faça } \theta_{n,c} = 0, \text{ caso contrario.} \end{aligned} \tag{2.2}$$

O termo k_n pode ser interpretado como o limite de sinalização de um teste não sequencial com amostra fixa de tamanho n, e n_1 é o número mínimo de casos antes de ocorrer a rejeição de H_0. Dessa forma, a probabilidade geral de rejeitar a hipótese nula pode ser expressa como uma soma de probabilidades parciais da seguinte maneira:

$$\beta(RR) = \beta_{n_1}(RR) + \beta_{n_1+1}(RR) \cdots + \beta_N(RR). \tag{2.3}$$

O poder pontual no n-ésimo teste é denotado por $\beta_n(RR)$, mas este termo também é o gasto alfa pontual, ao avaliar RR $= 1$. A chave para a solução ótima obtida por Silva e Kulldorff vem do fato de que se pode cortar $\beta_n(RR)$ para cada c como uma função de termos sub-parciais $\beta_{n,c}$, isto é:

$$\beta_n(RR) = \sum_{c=k_n}^{n} \beta_{n,c}(RR). \tag{2.4}$$

Para isto, e definindo para $n = n_1$, a probabilidade parcial $\beta_{n_1,c}(RR)$ como:

$$\begin{aligned} \beta_{n_1,c}(RR) &= Pr[\{X_{n_1} = c\} \cap \{H_0 \text{ rejeitada até o tempo } n_1\}|RR] \\ &= \theta_{n_1,c} \times P[X_{n_1} = c|RR], \end{aligned} \tag{2.5}$$

e para $n > n_1$:

$$\begin{aligned} \beta_{n,c}(RR) &= Pr[\{H_0 \text{ aceita até o tempo } (n-1)\} \cap \{X_n = c\} \cap \{H_0 \text{ rejeitada no tempo n}\}|RR] \\ &= \theta_{n,c} \times \bar{\beta}_{n-1,c}(RR) \times (1-p) + \theta_{n,c} \times \bar{\beta}_{n-1,c-1}(R) \times p, \end{aligned} \tag{2.6}$$

no qual:

$$\bar{\beta}_{n-1,c}(RR) = Pr[\{H_0 \text{ não rejeitada até o tempo } (n-1)\} \cap \{X_{n-1} = c\}|RR].$$

Então, a probabilidade geral de rejeitar H_0 pode ser reescrita como:

$$\beta(RR) = \sum_{n=n_1}^{N} \sum_{c=k_n}^{n} \beta_{n,c}(RR). \tag{2.7}$$

Ao denotar S como o tempo até a rejeição de H_0, então o tempo esperado para sinalizar é dado por:

$$E[S|RR] = \frac{\sum_{n=n_1}^{N} \sum_{c=k_n}^{n} n \times \beta_{n,c}(RR)}{\beta(RR)}. \tag{2.8}$$

Define-se $\vec{\theta} = (\theta_{n_1,k_{n_1}}, \cdots, \theta_{N,N})$ como o vetor de linha. Dado o risco relativo alvo $RR = r$, nível de significância α, e poder alvo γ, tem-se como objetivo resolver o seguinte sistema não-linear em $\vec{\theta}$:

$$\begin{cases} \min_{\vec{\theta}} E[S|RR = r], \\ \beta(1) = \alpha, \\ \beta(r) = \gamma, \\ 0 \leq \theta_{n,c} \leq 1, \text{ para cada } n = n_1, \cdots, N. \end{cases} \tag{2.9}$$

A abordagem feita por Silva e Kulldorff consiste na reparametrização do sistema não-linear (2.9) com o intuito de obter um novo sistema com funções lineares objetivas e de restrição. Assim, o sistema (2.9) se torna um problema de programação linear. Com isso tem-se o seguinte teorema:

Teorema 1. *Existem sempre números positivos $\pi_{n,c}$, $n = n_1, \cdots, N$, não dependentes do risco relativo RR, de modo que a probabilidade parcial $\beta_{n,c}(RR)$ em (2.5) e (2.6) sempre possam ser rescritas da seguinte maneira:*

$$\beta_{n,c}(RR) = \pi_{n,c} \times Pr[X_n = c|RR], \tag{2.10}$$

para cada $n = n_1, \cdots, N$.

O teorema 1 não será demonstrado neste trabalho, mas pode ser encontrado no seguinte artigo *"Optimal Alpha Spending for Sequential Analysis With Binary Data"* (SILVA; KULLDORFF, No prelo).

Devido à restrição $\beta(r) = \gamma$ no sistema (2.9), a função objetivo a ser minimizada é,

na verdade, o numerador de (2.8), que também é linear em $\pi_{n,c}$. A partir do Teorema 1 o sistema não linear em (2.9) pode sempre ser reescrito como um problema de otimização linear. O termo $\pi_{n,c}$ pode ser interpretado como o peso atribuído à probabilidade de rejeitar H_0 com exatamente c casos no n-ésimo teste. Para tanto, define-se o vetor $\vec{\pi} = (\pi_{n_1,k_{n_1}}, \cdots, \pi_{N,N})$. A reparametrização do sistema (2.9) em termos de um sistema linear pode ser expressa da seguinte forma:

$$\begin{cases} \min_{\vec{\pi}} E[S|RR = r] = \min_{\vec{\pi}} \sum_{n=n_1}^{N} \sum_{c=k_n}^{n} n \times \pi_{n,c} \times Pr[X_n = c|RR = r], \\ \beta(1) = \sum_{n=n_1}^{N} \sum_{c=k_n}^{n} \pi_{n,c} \times Pr[X_n = c|RR = 1] = \alpha, \\ \beta(r) = \sum_{n=n_1}^{N} \sum_{c=k_n}^{n} \pi_{n,c} \times Pr[X_n = c|RR = r] = \gamma, \\ 0 \leq \pi_{n,c} \leq I(c \leq n) \times I(c \geq k_n), \text{ para } n = n_1, \\ 0 \leq \pi_{n,c} \leq \left[\binom{n-1}{c}\bar{\pi}_{n-1,c} + \binom{n-1}{c-1}\bar{\pi}_{n-1,c-1}\right]\binom{n}{c}^{-1} \times I(c \leq n) \times I(c \geq k_n), \\ \text{para } n = n_1 + 1, \cdots, N. \end{cases} \tag{2.11}$$

A segunda e a terceira linha do sistema acima são restrições arbitrárias, enquanto a quarta e quinta linhas são condições necessárias para a admissibilidade de cada $\pi_{n,c}$.Um método bem conhecido para resolver problemas de programação linear é o algoritmo simplex(DANTZIG, 1951). Com as soluções ótimas $pi_{n,c}$, digamos $\pi^*_{n,c}$, pode-se calcular facilmente o gasto ótimo de alfa para o tempo n, que é dado por:

$$S(n) = \sum_{c=k_n}^{n} \pi^*_{n,c} \times Pr[X_n = c|RR = 1], \tag{2.12}$$

para $n = n_1, \cdots, N$, ou $S(n) = \alpha$ para $n \geq N$, e $S(n) = 0$, caso contrario.

Este procedimento ótimo está implementado no pacote *"Sequential"* (SILVA; KULLDORFF, 2019), sendo de livre utilização por meio do software R (R Core Team, 2017).

2.3 Distribuição Gumbel

Como já mencionado previamente, uma das funções que pode ser utilizada para obter o limiar de sinalização por meio da abordagem gasto do erro tipo I é a função do tipo potência. Nesse trabalho ao tentarmos utilizá-la com intuito de relacioná-la à solução exata obtida por Silva e Kulldorff (No prelo) não se obteve uma boa aproximação. Além disso, foram testadas varias outras funções, incluindo algumas que não são funções de distribuição, mas também não se obteve um bom resultado. Com isso chegou-se a distribuição

Gumbel, devido ao fato desta distribuição ter uma forma fechada simples, ou seja, de fácil manipulação, além de ser flexível. Dada a importância desta distribuição para esse trabalho, esta será apresentada a seguir.

A teoria de Valores Extremos foi proposto por Leonard Henry Caleb Tippet, enquanto trabalhava em uma empresa têxtil. Ele percebeu que a fragilidade de uma linha era influenciada pela fragilidade de uma fibra considerada mais fraca. Com a ajuda de Ronald Aylmer Fisher, foram obtidos por eles os três limites assintóticos que compõem as distribuições dos valores extremos, o mínimo e o máximo. Esta teoria foi formalizada por Gumbel em seu livro *"Statistics of Extremes"* (GUMBEL, 2004), no qual é apresentada a distribuição Gumbel, que recebeu esse nome em sua homenagem. Gumbel mostrou que conforme se aumenta o tamanho da amostra de uma variável aleatória com distribuição exponencial o máximo desta amostra vai se aproximando de uma distribuição Gumbel.

A distribuição Gumbel, também conhecida por distribuição de valores extremos tipo I, tem sua função densidade de probabilidade definida por:

$$f(x|\Theta) = \frac{1}{\beta}\, e^{-\frac{x-\mu}{\beta}\, -\, e^{\left(-\frac{x-\mu}{\beta}\right)}} \tag{2.13}$$

em que $\Theta = (\mu, \beta)$ tal que $\mu \in \mathbb{R}$ é o parâmetro de locação e $\beta > 0$ o parâmetro de escala. A função de distribuição acumulada é estabelecida da seguinte forma:

$$F(x|\Theta) = e^{-e^{\left(-\frac{x-\mu}{\beta}\right)}} \tag{2.14}$$

3 Resultados e Discussão

Uma questão importante na análise sequencial é como controlar a probabilidade de erro tipo I. Para isso utiliza-se a função gasto ótimo de alfa, podendo obtê-la através do pacote *"Sequential"* (SILVA; KULLDORFF, 2019) do software R. Mas, dependendo do tempo total de monitoramento de interesse o tempo de processamento pode ser elevado ou até ser impossível de ser obtido a função gasto ótimo de alfa, devido ao alto custo computacional. Portanto, para tentar encontrar uma função simples que possa aproximar a função gasto ótimo de alfa de modo satisfatória, inicialmente foi necessário obter a função gasto ótimo de alfa dos seguintes cenários descritos na Tabela 1, nos quais N representa o tempo total de vigilância, Z o número de controles correspondentes a cada caso, α o alfa total, γ o poder alvo e RR o risco relativo alvo.

Após ter sido calculado o gasto ótimo de alfa dos cenários citados anteriormente, foram testadas algumas funções pretendendo encontrar a que obtivesse a melhor aproximação. Dentre essas funções testadas a função potência não apresentou uma boa aproximação. Por isso começou-se a procurar uma distribuição que fosse mais flexível, e assim chegou-se à distribuição Gumbel, sendo que esta distribuição foi utilizada por ter uma forma fechada simples, ou seja, de fácil manipulação. Com essa distribuição obteve-se uma boa aproximação e com isso foi necessário desenvolver linhas de código que futuramente serão implementadas no pacote *"Sequential"* (SILVA; KULLDORFF, 2019) para então encontrar os parâmetros ótimos desta distribuição. A obtenção dos parâmetros ótimos da Gumbel foi realizada da seguinte forma: inicialmente estabelece uma grade de valores para os dois parâmetros da Gumbel, esta grade de valores vai de 0,01 a 10, de 0,01 em 0,01. Cada combinação de valores desta grade foi utilizada para o cálculo da performance do teste sequencial associado e a combinação que levou ao valor de performance ótimo foi considerada a escolha ótima para os dois parâmetros em cada cenário.

Ao utilizar a distribuição Gumbel conseguiu-se obter uma boa aproximação da função gasto ótimo de alfa. A Tabela 1 contém os parâmetros ótimos da distribuição Gumbel obtidos para cada um dos cenários utilizados. Pode-se notar um certo padrão nos valores

dos parâmetros ótimos da distribuição Gumbel, sendo que estes valores não apresentaram uma grande variação.

Tabela 1: Parâmetros ótimos dos cenários utilizados

N	**Z**	$\boldsymbol{\alpha}$	$\boldsymbol{\gamma}$	**RR**	$\boldsymbol{\mu}$	$\boldsymbol{\beta}$
70	1	0,05	0,2	1,3	0,29	0,14
70	2	0,05	0,2	1,3	0,27	0,15
70	3	0,05	0,2	1,3	0,31	0,17
70	1	0,05	0,95	3	0,18	0,14
70	2	0,05	0,95	3	0,17	0,11
70	3	0,05	0,95	3	0,14	0,14
80	1	0,01	0,2	1,5	0,43	0,15
80	2	0,01	0,2	1,5	0,42	0,15
80	3	0,01	0,2	1,5	0,44	0,18
80	1	0,01	0,95	3	0,29	0,15
80	2	0,01	0,95	3	0,23	0,15
80	3	0,01	0,95	3	0,25	0,15
110	1	0,05	0,5	1,5	0,31	0,15
110	2	0,05	0,5	1,5	0,3	0,16
110	3	0,05	0,5	1,5	0,33	0,18
120	1	0,05	0,9	2	0,23	0,15
120	2	0,05	0,9	2	0,2	0,15
120	3	0,05	0,9	2	0,21	0,18
140	1	0,01	0,9	2	0,37	0,2
140	2	0,01	0,9	2	0,33	0,19
140	3	0,01	0,9	2	0,42	0,27
150	1	0,01	0,5	1,5	0,58	0,24
160	2	0,01	0,5	1,5	0,51	0,22
190	3	0,01	0,5	1,5	0,46	0,2

Após obtenção dos parâmetros percebeu-se a necessidade de visualizar essa aproximação. Por isso foram feitos gráficos para comparar a aproximação com gasto ótimo de alfa com a obtido por meio do algoritmo de programação. Aqui serão somente apresentados alguns deles. Tais gráficos são apresentados a seguir, nos quais a linha vermelha indica a aproximação da função gasto de alfa utilizando a distribuição Gumbel e a linha preta

aponta a função gasto de alfa obtida através do pacote *"Sequential"*.

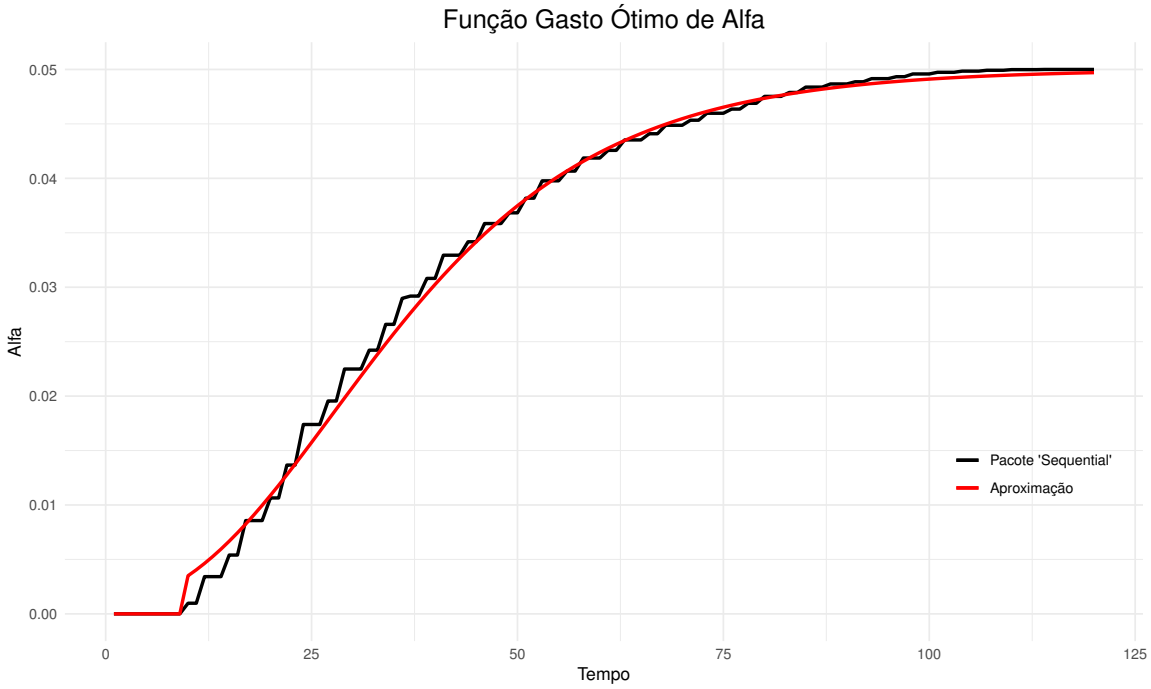

Figura 1: Comparação do gasto ótimo de alfa utilizando o pacote *"Sequential"* com o obtido através da aproximação utilizando a distribuição Gumbel, em função do tempo para o cenário N=120, Z=1, α=0,05, γ=0,9 e RR = 2

Na Figura 1 tem-se uma comparação do gasto ótimo de alfa utilizando o pacote *"Sequential"* com o obtido através da aproximação utilizando a distribuição Gumbel,nos quais os parâmetros ótimos são $\mu = 0{,}23$ e $\beta = 0{,}15$, em função do tempo para o cenário N=120, Z=1, α=0,05, γ=0,9 e RR = 2.Por meio deste gráfico pode-se perceber uma excelente aproximação da função gasto ótimo de alfa. Além disso, observa-se também que quanto mais o tempo cresce a aproximação se torna mais parecida com a curva original.

Já na Figura 2 é apresentada uma comparação do gasto ótimo de alfa utilizando o pacote *"Sequential"* com o obtido através da aproximação utilizando a distribuição Gumbel,nos quais os parâmetros ótimos são $\mu = 0{,}37$ e $\beta = 0{,}20$, em função do tempo para o cenário N=140, Z=1, α=0,01, γ=0,9 e RR = 2. Neste gráfico pode-se notar também uma boa aproximação.

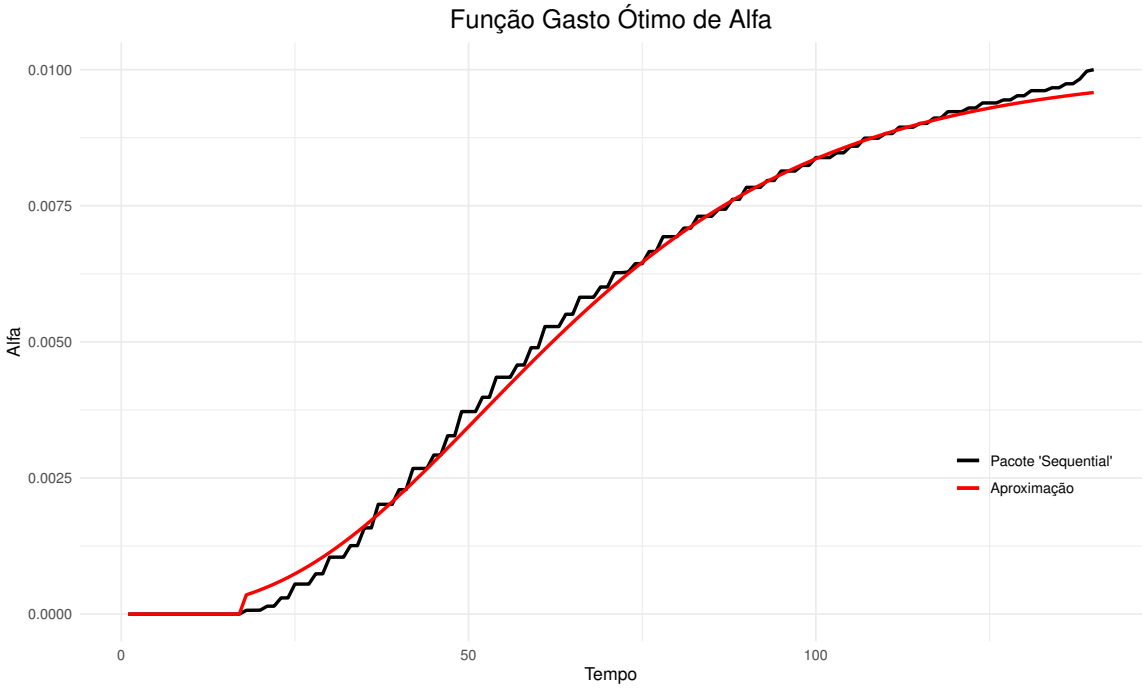

Figura 2: Comparação do gasto ótimo de alfa utilizando o pacote *"Sequential"* com o obtido através da aproximação utilizando a distribuição Gumbel, em função do tempo para o cenário N=140, Z=1, α=0,01, γ=0,9 e RR = 2

A partir da avaliação gráfica dessa aproximação percebeu-se que a distribuição Gumbel se mostra adequada para tal função. Portanto, pode-se dizer que uma boa opção para poder aproximar a função gasto ótimo de alfa de forma eficiente e simplificada é a distribuição Gumbel.

4 Considerações finais

Com o intuito de reduzir o custo computacional para encontrar a função gasto ótimo de alfa, utilizada para controlar a probabilidade de erro tipo I na análise sequencial, foram utilizados alguns cenários para tentar encontrar uma função simples capaz de produzir uma aproximação satisfatória da função gasto ótimo de alfa. A função gasto ótimo de alfa pode ser obtida através do método Silva e Kulldorff (No prelo), já implementado no pacote *"Sequential"* (SILVA; KULLDORFF, 2019) do software R, mas este método pode tomar muito tempo para valores grandes de monitoramento e, além disso, é teoricamente complexo. A partir dessa pesquisa foi possível perceber a capacidade da distribuição Gumbel de fazer uma aproximação satisfatória de tal função, tornando mais simples sua obtenção. Tal distribuição foi utilizada devido ao fato dela ter uma forma fechada simples, ou seja, de fácil manipulação, além de ser flexível.

Os resultados da pesquisa aqui apresentados contribuem para o conhecimento estatístico, através da área de análise sequencial, pois possibilita a obtenção da função gasto ótimo de alfa de maneira simplificada, que também pode ser calculada através do pacote *"Sequential"* (SILVA; KULLDORFF, 2019) do software R, mas existe um alto custo computacional para este cálculo. Este resultado é de grande importância para diversas áreas como, por exemplo, a vigilância de vacinas e drogas pós comercializadas, podendo assim ser utilizado para facilitar a identificação de eventos adversos.

Referências

DANTZIG, G. B. Maximization of a linear function of variables subject to linear inequalities. *Activity analysis of production and allocation*, v. 13, p. 339–347, 1951.

DAVIS, R. L. et al. Active surveillance of vaccine safety: a system to detect early signs of adverse events. *Epidemiology*, JSTOR, p. 336–341, 2005.

GUMBEL, E. *Statistics of Extremes*. Dover Publications, 2004. (Dover books on mathematics). ISBN 9780486436043. Disponível em: <https://books.google.com.br/books?id=kXCg8B5xSUwC>.

JENNISON, C.; TURNBULL, B. W. *Group sequential methods with applications to clinical trials*. [S.l.]: Chapman and Hall/CRC, 2000.

LEITE, A.; ANDREWS, N. J.; THOMAS, S. L. Near real-time vaccine safety surveillance using electronic health records?a systematic review of the application of statistical methods. *pharmacoepidemiology and drug safety*, Wiley Online Library, v. 25, n. 3, p. 225–237, 2016.

LIEU, T. A. et al. Real-time vaccine safety surveillance for the early detection of adverse events. *Medical care*, LWW, v. 45, n. 10, p. S89–S95, 2007.

O'BRIEN, P. C.; FLEMING, T. R. A multiple testing procedure for clinical trials. *Biometrics*, JSTOR, p. 549–556, 1979.

POCOCK, S. J. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, Oxford University Press, v. 64, n. 2, p. 191–199, 1977.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>.

SILVA, I. R. Análise sequencial usando r para monitoramento de vacinas e drogas pós-comercializadas. 2014.

SILVA, I. R.; KULLDORFF, M. *Sequential: Exact Sequential Analysis for Poisson and Binomial Data*. [S.l.], 2019. R package version 3.0.1. Disponível em: <https://CRAN.R-project.org/package=Sequential>.

SILVA, I. R.; KULLDORFF, M. Optimal alpha spending for sequential analysis with binary data. No prelo.

WALD, A. Sequential tests of statistical hypotheses. *The annals of mathematical statistics*, JSTOR, v. 16, n. 2, p. 117–186, 1945.

WALD, A.; WOLFOWITZ, J. Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, The Institute of Mathematical Statistics, v. 19, n. 3, p. 326–339, 09 1948. Disponível em: <https://doi.org/10.1214/aoms/1177730197>.

YIH, W. K. et al. Active surveillance for adverse events: the experience of the vaccine safety datalink project. *Pediatrics*, Am Acad Pediatrics, v. 127, n. Supplement 1, p. S54–S64, 2011.