Monitoramento de séries de contagem por meio de gráficos de controle

Orlando Yesid Esparza Albarracín

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientador: Profa. Dra. Airlane Pereira Alencar

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES

São Paulo, fevereiro de 2014

Monitoramento de séries de contagem por meio de gráficos de controle

Esta é a versão original da dissertação/tese elaborada pelo candidato Orlando Yesid Esparza Albarracín, tal como submetida à Comissão Julgadora.

Monitoramento de séries de contagem por meio de gráficos de controle

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 10/03/2014. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Airlane Pereira Alencar (orientadora) - IME-USP
- Profª Drª Denise Aparecida Botter / IME - USP
- Profª Drª Linda Lee Ho / EP - USP

Agradecimentos

Cada oportunidad en la vida me ha permitido descubrir matices de mi personalidad antes ocultas a mi razón. En el pasado muy seguramente el entorno provocaba en mí cierta sensación de amaestramiento ligado a un sin número de paradigmas. Hoy, en las tierras que me adoptan y que en un comienzo las percibía grises, pero que al pasar del tiempo se han tornado en un campo de batalla pintoresco, descubro que los límites solo son mentales y que es posible que el sueño y la realidad sean uno. Estoy seguro de que cada situación manifestada, cada sensación sentida y apasionadamente sufrida, en la que me he visto envuelto, ha contribuido para que esta formación, más que académica, haya sido del ser.

Gracias y siempre gracias.

Un Agradecimiento especial a mi Esperanza, a mi Papá y mis hermanos.
A mi maletera preferida, a mi nona Chela y a la Luz que siguió constantemente esta novela.
E a minha orientadora Airlane Pereira.

Resumo

ALBARRACIN. O.Y.E **Monitoramento de séries de contagem por meio de gráficos de controle**. 2014. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

Na área da saúde, várias abordagens nos últimos anos têm sido propostas baseadas nos gráficos de controle CUSUM para a detecção de epidemias infecciosas em que a caraterística a ser monitorada é uma série temporal de dados de contagem, como o número de internações. Neste trabalho foram implementados os modelos lineares generalizados (MLG) no monitoramento, por meio dos gráficos CUSUM e Shewhart, da série do número diário de internações por causas respiratórias para pessoas com 65 anos ou mais residentes no município de São Paulo.

Por meio de simulações, avaliaram-se a eficiência de cinco estatísticas diferentes para detectar mudanças na média em séries de contagem. Uma das abordagens consistiu na implementação de três transformações normalizadoras simples que dependem unicamente dos parâmetros das distribuições Poisson e binomial negativa: a transformação Rossi para dados com distribuição Poisson, a transformação Jorgensen para dados com distribuição binomial negativa e os sesíduos de Anscombe para modelos lineares generalizados. As duas últimas estatísticas já foram propostas como gráficos CUSUM: o Método Rogerson e Yamada (2004) é apresentado para dados com distribuição Poisson e neste trabalho foi proposto um novo parâmetro k_t para dados binomial negativa; já o método proposto por Hohle (2007) é baseado na função de verossimilhança da distribuição binomial negativa. Utilizando limites de controle para obter um valor $ARL_0 = 500$ sob normalidade, monitorou-se via simulação a série de interesse, implementando as transformações normalizadoras. Entretanto, utilizando-se esses limiares observa-se um maior número de alarmes falsos para as três estatísticas. Modificando o parâmetro k do gráfico CUSUM permitindo que variasse ao longo do tempo a série foi monitorada e foram obtidos valores ARL_0 próximos a 500. Os gráficos CUSUM baseados no método Rogerson e Yamada e na estatística da razão de verossimilhanças para dados com distribuição binomial negativa mostraram, via simulação, bons resultados para detectar mudanças na média. As suposições de normalidade e independência das estatísticas normalizadoras, em geral omitidas em trabalhos publicados na literatura, foram avaliadas e comprova-se que as transformações não normalizam os dados, porém são independentes e estacionárias. Analisando os dados reais, as estatísticas apresentaram autocorrelação significativa no lag 7. Devido à persistência desta autocorrelação, foi proposta uma abordagem baseada no ajuste do modelo GARMA.

Palavras-chave: CUSUM, Shewhart, séries de contagem.

Abstract

ALBARRACIN. O.Y.E **Monitoring time series of counts using control charts.**. 2014. 120 f. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2014.

In public health, several approaches have been proposed in order to detect outbreaks of infectious diseases. The monitored characteristics is the time series of count data as the number of hospitalizations, where the population and the expected rate of admissions change over time. In this work, we fitted generalized linear models (GLM) and implemented Shewhart and CUSUM control charts for monitoring the daily number of hospital admissions due to respiratory diseases for people aged 65 and older in the city of São Paulo.

Through simulations, the efficiency of implementing five different statistics for detecting changes in time series of count was evaluated. One approach consists in the implementation of three transformations that only depend on the parameters of negative binomial and Poisson distributions: the transformations of Rossi for data with Poisson distribution, the transformation proposed by Jorgensen for data with negative binomial distribution and residuals proposed by Anscombe for generalized linear models. The other two statistics have been proposed as CUSUM charts: the method of Rogerson e Yamada (2004) was presented for data with Poisson distribution and in this work we proposed a new parameter k_t for negative binomial distribution and the method proposed for Hohle (2007) that uses the likelihood ratio statistic. The time series of interest was monitored using control limits that, assuming normality,are obtained value $ARL_0 = 500$. However, it was observed a greater number of alarms for the three transformations using these limits. Values of ARL_0 close to 500 were obtained by modifying the parameter k of the CUSUM control charts allowing change this parameter over time. The CUSUM control charts for the methods of Rogerson and Yamada and Holhe for data with negative binomial distribution showed, by simulation, good results for detecting variations in the average. The assumptions' evaluation of normality for the statistics proposed by Rossi, Jorgensen and Anscombe generally is omitted in published studies. In this work, these assumptions were evaluated indicating that the statistics are not normal using the real dataset but are independent and stationary. By analyzing real data, due to the persistence correlation for the normalized statistics, an approach based on setting GARMA model was proposed. This method brought good results once the residuals of fitted model were normal and independent. Due to the persistence of correlation for the normalized statistics, an approach based on setting GARMA model was proposed. This method showed good results once the residuals of the fitted model were normal and independent.

Keywords: CUSUM, Shewhart, time series of counts.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

O conceito de qualidade permeia tanto as áreas de produção de produtos, quanto de serviços. O monitoramento de processos pode evitar grandes prejuízos se problemas forem detectados tão logo ocorram. Na área da saúde pública, esse conceito também pode ser utilizado na vigilância de doenças, acompanhando, por exemplo, a morbidade ou mortalidade de doenças para detectar epidemias, tanto para a tomada de medidas urgentes, quanto para o planejamento de serviços de saúde. A principal motivação desse trabalho é o monitoramento de séries temporais de internações visando a detecção de aumentos no número médio de internações para que sejam tomadas medidas em políticas de saúde.

Primeiro, apresentaremos um histórico dos gráficos de controle mais utilizados em controle estatístico de qualidade, antes de apresentar os métodos que utilizaremos nesse trabalho. A principal referência utilizada que apresenta os gráficos de controle tipo Shewhart e CUSUM é Montgomery (2009).

O controle estatístico da qualidade de processos iniciou-se com Walter Shewhart, autor de "Economic Control of Quality Manufacture Product", na década de 20, quando trabalhava na Bell Telephone, com o desenvolvimento das cartas de controle (também denominados gráficos de controle). Posteriormente, durante a Segunda Guerra Mundial, surgiram diferentes pesquisas para o desenvolvimento destas cartas, em parte, devida à grande quantidade de material bélico produzido e à deficiência e deterioramento da mão de obra. Outras pesquisas surgiram em empresas de produção nos Estados Unidos, que precisavam fabricar produtos e equipamentos eletrônicos com um elevado nível de qualidade.

Segundo Shewhart, todo e qualquer processo, por mais bem projetado e controlado que seja, possui um componente de variabilidade natural impossível de ser eliminado, proveniente de causas aleatórias. Quando o processo sofre somente a influência de causas aleatórias, diz-se que ele está sob controle. Além dessa influência, os processos podem estar sujeitos à ocorrência ocasional de uma variabilidade extra, que aumenta a dispersão e/ou tira a característica de qualidade de interesse de seu valor especificado (valor-alvo ou valor em controle). Quando, além das causas aleatórias, causas especiais de variabilidade estiverem presentes, diz-se que o processo está fora de controle. Desse modo, o processo deve ser monitorado continuamente para detectar a ocorrência de causas especiais e implementar ações corretivas, para não comprometer a qualidade final dos produtos. A principal ferramenta utilizada no monitoramento dos processos e na sinalização da presença de causas especiais são os gráficos de controle.

Um gráfico de controle consiste basicamente de uma linha que apresenta alguma estatística calculada a partir de amostra de observações ao longo do tempo e três linhas de referência: um Limite de Controle Superior (LCS), um Limite de Controle Inferior (LCI) e a Linha Central (LC) que representa um valor-alvo da característica a monitorar. O processo está sob controle ou fora

de controle de acordo com a posição da estatística de interesse, calculada para cada amostra e registrada no gráfico de controle. Se o valor da estatística estiver acima do LCS ou abaixo do LCI, os limites preestabelecidos, diz-se que o processo está fora de controle. Por outro lado, se a estatística registrada no gráfico estiver dentro dos limites de controle o processo é considerado sob controle.

No monitoramento das características de qualidade de interesse, os gráficos de controle utilizados para variáveis contínuas são denominados gráficos de controle por variáveis. Um exemplo é o gráfico da média $\overline{X}$. Nesse gráfico se verifica se a média amostral está entre os limites de controle para a média populacional. Em muitos processos, as características de qualidade são medidas pela proporção de produtos não-conformes (ou defeituosos) produzidos ou pela frequência de ocorrência de não-conformidades (ou defeitos). Os tipos de gráficos que monitoram estas medidas de qualidade são denominados gráficos de controle por atributos.

Os gráficos de controle de Shewhart levam em consideração somente a informação da última amostra. Este tipo de gráfico é bastante simples de ser utilizado e é satisfatoriamente rápido para detectar alterações de grande magnitude na característica de qualidade de interesse (por exemplo, uma alteração de 1,5 desvio padrão no caso do gráfico de $\overline{X}$) mas é mais lento para detectar alterações de menor magnitude (Montgomery , 2009).

Os métodos de controle estatístico de processo nos últimos anos têm uma longa lista de aplicações a problemas de vigilância em saúde pública (Woodall , 2006). Nos últimos anos várias abordagens, baseadas nestes métodos, têm sido propostas para a detecção de epidemias de doenças infecciosas. Esse tipo de monitoramento é semelhante à detecção de alterações em processos de produção industrial como apresentado em Unkel *et al.* (2012). Entretanto, na vigilância em saúde, a característica a ser monitorada é uma série temporal de dados de contagem, como o número de internações, onde o tamanho da população de interesse e a contagem média podem mudar ao longo do tempo e, portanto, devem ser tomadas medidas para ajustar essas mudanças ao monitorar a série. Um exemplo de alteração ao longo do tempo é a possível sazonalidade, por exemplo, com mais internações durante o inverno. Vários métodos para a vigilância em saúde têm sido sugeridos na literatura baseados na detecção de alterações de pequenas magnitudes. Pesquisas amplas e bibliografias sobre vigilância nesta área são apresentadas por Lai (1995), com foco em propriedades minimax de regras de parada, e por Woodall e Montgomery (1999), que se concentram em gráficos de controle por meio das técnicas de controle estatístico de processo (CEP). (Unkel *et al.* , 2012) apresenta uma extensa revisão de vários métodos de controle nessa área.

No controle estatístico para vigilância em saúde com frequência são implementados os gráficos de soma acumuladas, chamados doravante de CUSUM, que são alternativas viáveis aos gráficos de controle de Shewhart, devido à sua eficiência e rapidez para detectar pequenas mudanças na série monitorada (Shmueli e Burkom , 2010). Uma vantagem destes gráficos comparados aos gráficos Shewhart é que guardam as informações acumuladas de toda a sequência de observações. Autores como Lucas (1985) e Hawkins (1981) afirmam que estes gráficos são mais eficientes que os gráficos de Shewhart na detecção de pequenas mudanças. Na literatura Rogerson e Yamada (2004), Woodall (2006) e Hohle e Mazick (2010) são alguns exemplos de trabalho que propõem gráficos CUSUM para dados de contagem.

Recentemente, Hohle e Paul (2008) comparam diversos métodos para a vigilância de doenças infecciosas e comentam que é importante no monitoramento das séries de contagens levar em conta as variações sazonais na média, o ajuste para a população em situação de risco e/ou outras variáveis explicativas. Basicamente, ele recomenda utilizar modelos de regressão baseados em modelos lineares generalizados (MLG) no monitoramento das séries. Gráficos de controle com modelos de regressão para resposta normal são encontrados na literatura de engenharia. Algumas pesquisas que implementam modelos MLG no controle estatístico são encontrados em Skinner e Runger (2003)

e na literatura de vigilância em saúde em Rossi *et al.* (1999) e Rogerson e Yamada (2004). Na literatura, é mais comum adotar a distribuição Poisson para modelar séries de contagem como em Hohle e Paul (2008) e Rossi e Marchi (2010)) e mais recentemente alguns trabalhos utilizam a distribuição binomial negativa Hohle e Mazick (2009). A maioria dos trabalhos como Rossi e Marchi (2010), (Hohle e Paul , 2008) e Rossi *et al.* (1999) baseiam-se na construção de gráficos de controle utilizando transformações normalizadoras e constroem os limites dos gráficos utilizando a distribuição normal. Apesar da importante suposição de normalidade para a validade dessa técnica, a maioria dos trabalhos não apresentam uma análise de resíduos para verificar essa suposição.

Esta dissertação tem como tema principal o estudo dos gráficos CUSUM como uma ferramenta eficaz para o monitoramento de séries temporais de dados de contagem com efeitos sazonais. A média da série de tempo será ajustada com variáveis explicativas com distribuição binomial negativa com função de ligação logaritmo e a população será incluída (como offset). Em particular, será analisada a série de tempo do número diário de internações por causas respiratórias de pessoas com 65 anos ou mais residentes no município de São Paulo no ano de 2011. Será apresentada uma comparação dos gráficos de Shewhart e CUSUM para avaliar a rapidez e efetividade destes gráficos para detectar pequenas mudanças.

1.1 Objetivos

O objetivo principal desta dissertação consiste em monitorar séries de contagem com efeitos sazonais por meio dos gráficos CUSUM avaliando por meio de simulações a eficiência e rapidez de cinco estatísticas diferentes utilizadas nestes gráficos.

Os objetivos específicos deste trabalho são os seguintes:

1. Avaliar as suposições de normalidade e independência das estatísticas normalizadoras que serão implementadas nos gráficos tipo Shewhart e CUSUM.

2. Comparar a rapidez e efetividade dos gráficos CUSUM e gráficos de Shewhart para detectar pequenas mudanças na média em séries de contagem com efeitos sazonais modelando a média com variáveis explicativas.

3. Monitorar a série de tempo do número diário de internações por causas respiratórias de pessoas com 65 anos ou mais residentes no município de São Paulo no ano 2011 mediante os gráficos CUSUM e de Shewhart.

4. Propor e avaliar os resíduos de Anscombe como estatística alternativa para os gráficos CUSUM e de Shewhart.

1.2 Contribuições

Nesta dissertação o monitoramento de séries temporais com efeitos sazonais por meio dos gráficos CUSUM, bem como o comparativo do desempenho destes gráficos com os gráficos de Shewhart trazem como contribuição:

- Uma generalização da estatística proposta em Rogerson e Yamada (2004) para o monitoramento de séries com distribuição binomial negativa em vez da distribuição Poisson proposta por esses autores.
- A implementação dos resíduos de Anscombe como uma nova estatística de monitoramento para séries de contagem.
- Avaliação das suposições de normalidade e independência das transformações normalizadoras, avaliação em geral omitida em trabalhos publicados na literatura.

1.3 Organização do Trabalho

Este trabalho está estruturado em cinco capítulos, sendo este primeiro o capítulo introdutório. No Capítulo 2 é realizada uma revisão conceitual dos gráficos de controle e das principais medidas de desempenho utilizadas para quantificar a eficácia e rapidez na detecção de alterações do processo. O Capítulo 3 apresenta as diferentes estatísticas que serão utilizadas para o monitoramento da série de interesse por meio dos gráficos de controle CUSUM e Shewhart. No Capítulo 4 são apresentados os ajustes dos modelos lineares generalizados que foram implementados para calcular o valor alvo para a série de interesse no ano 2011 e também é apresentado o algoritmo que será implementado para avaliar a rapidez e eficiência das estatísticas apresentadas no Capítulo 3. O Capítulo 5 apresenta os resultados das simulações avaliando o desempenho das estatísticas para os gráficos de controle tipo Shewhart e CUSUM por meio das medidas de desempenho ARL e MRL. No Capítulo 6 é apresentada uma aplicação em dados reais para monitorar a série do número diário de internações por causas respiratórias por meio dos gráficos tipo Shewhart e CUSUM. Além disso, é apresentada uma proposta de monitoramento baseada nos modelos GARMA. No Capítulo 7 são apresentadas as conclusões.

Capítulo 2

Revisão de Literatura

Em processos de produção é fundamental garantir a qualidade dos produtos e detectar falhas antes que muitas unidades defeituosas sejam produzidas. Nesse contexto, na área de controle de qualidade, os gráficos de controle são ferramentas simples muito utilizadas por atender aos objetivos de monitoramento e sinalização de causas especiais nos processos. Um gráfico de controle consiste basicamente de uma linha que apresenta uma estatística calculada a partir de uma amostra ordenada no tempo e três linhas de referência:

- Linha Central (LC): linha paralela à abscissa que representa o valor alvo do processo que pode ser um valor baseado no passado histórico da estatística monitorada.
- Limite de Controle Superior (LCS): representa o valor máximo aceitável para a variável do processo de controle.
- Limite de Controle Inferior (LIC): essa linha representa o valor mínimo aceitável para a variável do processo de controle.

Se algum valor da estatística estiver acima do LCS ou abaixo do LCI diz-se que o processo está fora de controle. Desse modo, o processo deve ser monitorado continuamente para detectar a ocorrência de causas especiais e implementar ações corretivas para não comprometer a qualidade final dos produtos. Por outro lado, se as observações no gráfico estão dentro dos limites e variam em torno do valor alvo (fixo) de maneira estável e previsível, o processo é considerado sob controle e as observações só possuem em sua variabilidade um componente natural proveniente de causas aleatórias.

Supondo que causas aleatórias não alteram a estatística da variável a ser monitorada, por exemplo, a média μ, e que o valor alvo é μ_0, as seguintes hipóteses de interesse mutuamente exclusivas H_0 e H_1 podem ser definidas como seguem:

$$\begin{aligned} H_0 : \mu &= \mu_0 \text{ (processo sob controle)} \\ H_1 : \mu &\neq \mu_0 \text{ (processo fora de controle).} \end{aligned}$$

No Controle Estatístico do Processo (CEP) um falso alarme ou erro tipo I é medido pela probabilidade (α) de erroneamente considerar-se o processo fora de controle quando ele está sob controle (H_0 verdadeira). Por outro lado, a falta de detecção ou erro tipo II é avaliado pela probabilidade (β) de erroneamente considerar-se o processo sob controle, quando ele está fora de controle (H_1 verdadeira). O alarme verdadeiro é definido como a probabilidade (Pd=1-β) de detecção, ou seja, de rejeitar H_0 quando é falsa (Costa e Carpinetti , 2004).

Para avaliar o desempenho (poder do gráfico) de um gráfico de controle e comparar vários procedimentos, podem ser levados em conta os valores das probabilidades dos erros Tipo I e Tipo II associados às tomadas de decisão. No entanto, é costume recorrer a outros parâmetros relacionados

com a distribuição do número de observações necessárias até um primeiro ponto exceder os limites de controle. Um desses parâmetros e o mais utilizado na literatura é Average Run Lenght (ARL) ou, em português, Comprimento Médio de Corrida.

Neste capítulo discutiremos os gráficos de controle de Shewhart para a média com observações independentes e os gráficos de somas cumulativas (CUSUM). Utilizaremos a notação apresentada nessa referência e manteremos algumas siglas em inglês utilizadas nessa área, como a sigla ARL para Average Run Length.

2.1 Gráfico de Controle Shewhart

Os gráficos de Shewhart destacam-se dentre as ferramentas do CEP devido principalmente à sua simplicidade operacional e à sua efetividade na identificação de problemas no processo (Montgomery , 2009). Considere uma característica de qualidade a ser controlada que tenha distribuição normal com média μ e desvio padrão σ, sendo ambos os valores conhecidos. Se x_1,x_2,...,x_n é uma amostra de tamanho n, então, a média dessa amostra $\overline{x}$ é normalmente distribuída com média μ e desvio padrão $\sigma_{\overline{x}} = \sigma/\sqrt{n}$. Ademais, há uma probabilidade de $1-\omega$ de qualquer média amostral está entre

$$\mu + Z_{1-\omega/2}\sigma_{\overline{x}} = \mu + Z_{1-\omega/2}\frac{\sigma}{\sqrt{n}} \qquad e \qquad \mu - Z_{1-\omega/2}\sigma_{\overline{x}} = \mu - Z_{1-\omega/2}\frac{\sigma}{\sqrt{n}}, \tag{2.1}$$

em que $Z_{1-\omega/2}$ corresponde ao percentil $1-\omega/2$ da distribuição normal padrão ($N(0,1)$).

As expressões apresentadas em (2.1) podem ser usadas como limites de controle, sendo denominadas Limite Inferior de Controle (LIC) e Limite Superior de Controle (LSC). É comum substituir $Z_{1-\alpha/2}$ por 3 de modo que os limites três sigma sejam empregados conforme foi sugerido por Shewhart em 1931. Se a média amostral cai fora desses limites é uma indicação de que a média do processo não é mais igual a μ.

Se os parâmetros μ e σ forem desconhecidos, esses podem ser estimados quando o processo está supostamente sob controle. Suponha m amostras cada uma com n observações da característica da qualidade. Sejam $\overline{x}_1$, $\overline{x}_2$,... , $\overline{x}_m$, as médias de cada uma das amostras. O melhor estimador de μ, a média do processo, é a média amostral, ou seja, a média das médias

$$\overline{\overline{x}} = \frac{(\overline{x}_1 + \overline{x}_2 + \cdots + \overline{x}_m)}{m},$$

sendo que $\overline{\overline{x}}$ deve ser usado como a linha central no gráfico de controle $\overline{x}$. Em muitas aplicações da estatística nos problemas de engenharia de qualidade é conveniente estimar o desvio padrão pelo método da amplitude. Seja $x_1, ...x_n$ uma amostra de tamanho n, a amplitude da amostra é a diferença entre a maior e a menor observação, isto é, $R = x_{max} - x_{min}$. Sejam $R_1, R_2, ..., R_m$ as amplitudes das m amostras. A amplitude média é dada por

$$\overline{R} = \frac{R_1 + ... + R_m}{m}.$$

A variável $W = R/\sigma$ é chamada amplitude relativa. A média de W é uma constante d_2 que depende do tamanho da amostra, isto é $E(W) = d_2$(Montgomery , 2009). Um estimador não viesado do desvio padrão σ de uma distribuição normal é $\widehat{\sigma} = \overline{R}/d_2$. Valores de d_2 para amostras de tamanho $2 \leq n \leq 25$ são apresentados no Apêndice de (**?**). Os limites de controle 3 sigma para o gráfico $\overline{x}$ quando μ e σ são desconhecidos é dado por

$$\begin{aligned} \text{LSC} &= \overline{\overline{x}} + 3\overline{R}/d_2\sqrt{n} \\ \text{Linha central} &= \overline{\overline{x}} \end{aligned} \tag{2.2}$$

$$\text{LIC} = \overline{\overline{x}} - 3\overline{R}/d_2\sqrt{n}.$$

Os gráficos de controle tipo Shewhart são os mais apropriados para controlar a média do processo e detectar a ocorrência de grandes desvios, sendo pouco poderosos para detectar pequenas mudanças. Diz-se que estes gráficos "não possuem memória", pois sua regra de decisão se baseia apenas no exame do último ponto observado. Isto torna os gráficos de Shewhart relativamente insensíveis a pequenas mudanças inferiores a $1,5\sigma$ (desvio padrão) ou menos (Montgomery , 2009). A Figura 2.1 ilustra um exemplo de gráfico do controle de Shewhart.

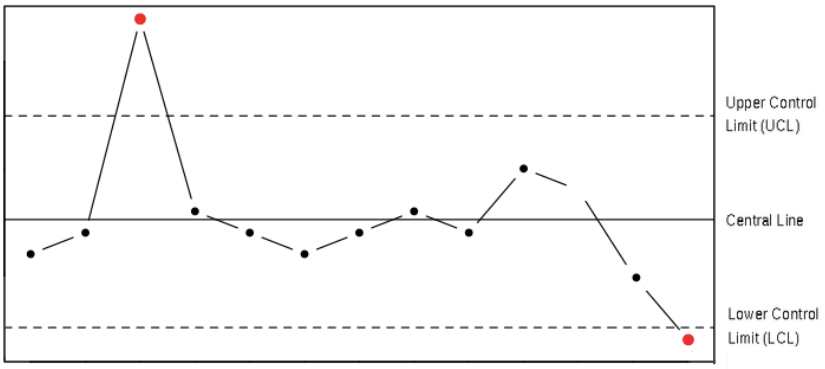


Figura 2.1: *Gráfico de controle de Shewhart*

2.2 Desempenho dos gráficos de controle

O desempenho de um gráfico de controle está intrinsecamente ligado à velocidade que ele detecta a presença de uma causa especial para que a tomada de decisões aconteça o quanto antes e o processo volte a ficar sob controle. Esse desempenho é avaliado geralmente na literatura pela função Característica de Operação (CO) ou Average Run Length (ARL) ou, em português, Comprimento Médio de Corrida. Essa última medida representa o número médio de pontos que devem ser plotados no gráfico antes que um ponto indique a condição de ausência de controle estatístico, podendo este ser falso ou verdadeiro.

2.2.1 Função Característica de Operação (CO)

A efetividade dos gráficos $\overline{x}$ de Shewhart para detectar mudanças ou grandes desvios no processo é descrita pela sua curva característica de operação (CO). Para um gráfico de controle $\overline{x}$ com desvio padrão σ conhecido e constante, se a média desloca-se do valor sob controle μ_0 para outro valor $\mu_1 = \mu_0 + \delta\sigma$, a probabilidade β do gráfico *não* detectar esse deslocamento na primeira amostra subsequente é dada por

$$\beta = P\left\{LCS \leq \overline{x} \leq LCI | \mu = \mu_1 = \mu_0 + \delta\sigma\right\}. \tag{2.3}$$

Para $\overline{x} \sim N(\mu, \sigma^2/n)$ com limites de controle inferior e superior dados por $LIC = \mu_0 - L\sigma/\sqrt{n}$ e $LSC = \mu_0 + L\sigma/\sqrt{n}$ a probabilidade β é expressa como segue:

$$\begin{aligned}
\beta &= \Phi\left[\frac{LSC - (\mu_0 + \delta\sigma)}{\sigma/\sqrt{n}}\right] - \Phi\left[\frac{LIC - (\mu_0 + \delta\sigma)}{\sigma/\sqrt{n}}\right] \\
\beta &= \Phi\left[\frac{\mu_0 + L\sigma/\sqrt{n} - (\mu_0 + \delta\sigma)}{\sigma/\sqrt{n}}\right] - \Phi\left[\frac{\mu_0 - L\sigma/\sqrt{n} - (\mu_0 + \delta\sigma)}{\sigma/\sqrt{n}}\right]
\end{aligned}$$

$$\beta = \Phi(L - \delta\sqrt{n}) - \Phi(-L - \delta\sqrt{n}), \tag{2.4}$$

sendo que Φ denota a função de probabilidade acumulada de uma distribuição normal padrão. A probabilidade do gráfico detectar esse deslocamento na média na primeira amostra subsequente é dada por $1 - \beta$. Para construir a curva característica de operação deve-se plotar β versus a magnitude do deslocamento que se deseja detectar expressa em unidades de desvio padrão para vários tamanhos de amostra n.

Quando o processo encontra-se sob controle, a probabilidade do gráfico apresentar um falso alarme α (erro do tipo I) para um gráfico com ambos os limites superior e inferior é dada por

$$\alpha = P(LCS < \overline{x} < LCI) \tag{2.5}$$

A curva CO para diferentes tamanho de amostra com limites três sigma está ilustrada na Figura 2.2.

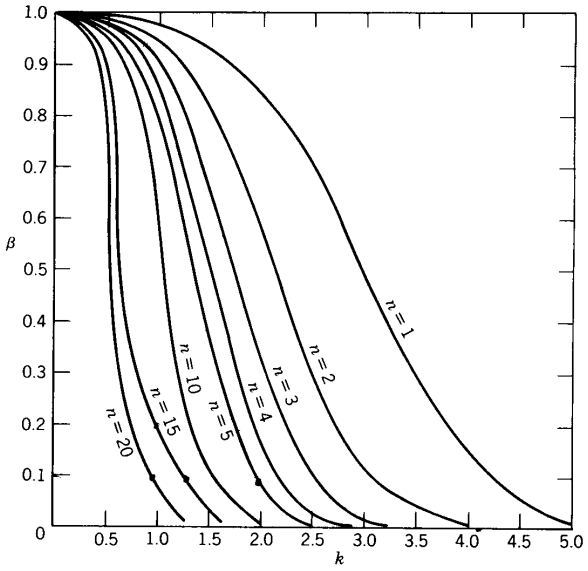


Figura 2.2: *Função Característica de Operação com limites 3 sigma.*
Fonte: Montgomery (2009)

2.2.2 Comprimento Médio da Sequência (ARL)

Se as observações cuja característica está sendo controlada são independentes então, o número de observações necessárias até um primeiro ponto exceder os limites de controle é uma variável aleatória cuja distribuição é geométrica com parâmetro p (Montgomery , 2009). Portanto, a probabilidade de uma mudança ser detectada na r-ésima amostra é simplesmente $1 - p$ vezes a probabilidade da mudança não ser detectada em cada uma das primeiras $r - 1$ amostras, ou seja, $p^{r-1}(1 - p)$.
Em geral o número esperado de amostras necessárias para detectar-se um deslocamento é o comprimento médio da sequência (ARL) ou

$$ARL = \sum_{r=1}^{\infty} rpr - 1(1 - p) = \frac{1}{1 - p}. \tag{2.6}$$

A estatística ARL_0, ou comprimento médio da corrida (average run length), é basicamente o número médio de observações que devem ser plotados no gráfico antes de emitir um sinal fora de controle mesmo quando o processo está sob controle estatístico. Por outro lado, a estatística ARL_1 é o número médio de observações necessárias para emitir um sinal fora de controle quando o processo

já não se encontra sob controle.

$$ARL_0 = \frac{1}{\alpha} \qquad ARL_1 = \frac{1}{1-p}$$

Para os gráficos de Shewhart tradicionalmente os limites de controle são posicionados a uma distância simétrica de três desvios padrão da linha central ($L = 3$ em 2.4). Para dados com distribuição normal e independentes com límites de controle 3σ o valor esperado do ARL_0 é 370. Esse valor corresponde a $\alpha == 0,0027$

A Figura 2.3 apresenta a curva ARL_0 para diferentes tamanhos de n. O ARL é interpretado em termos do número de amostras necessárias para detectar uma mudança.

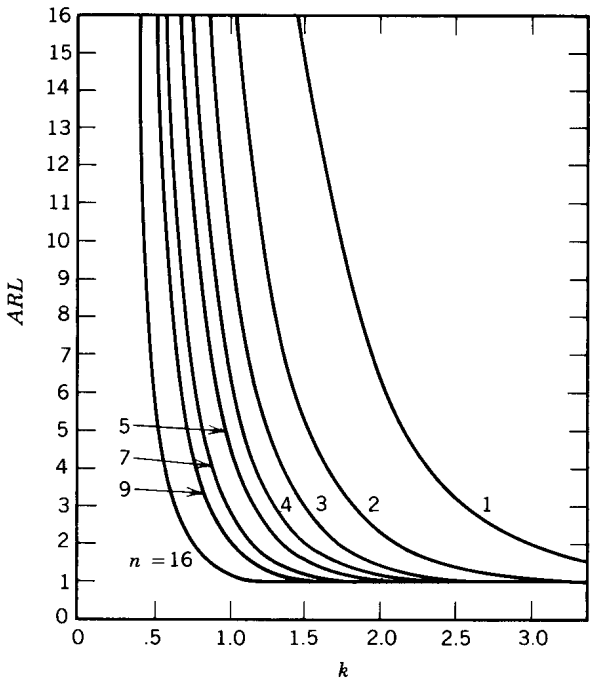


Figura 2.3: *Curva do ARL_0 para gráficos de Shewhart*
Fonte Montgomery (2009)

2.3 Gráfico da Soma Cumulativa tipo CUSUM

Os gráficos de controle tipo CUSUM são propostos por Page em 1954 como alternativas viáveis aos gráficos de Shewhart. Estes novos gráficos detectam com maior rapidez pequenas mudanças, já que incorporam diretamente toda a informação da sequência dos valores da amostra plotando as somas cumulativas dos desvios dos valores da amostra de um valor alvo. Assim, se uma pequena mudança no processo está presente por algumas observações em sequência, esta mudança será percebida por sua soma. Além disso, são particularmente mais eficazes com amostras de tamanho 1 (Montgomery , 2009).

Para m amostras de tamanho n, $n \geq 1$ e $\overline{s}_i$ a estatística de interesse da variável a ser monitorada e D_0 o valor alvo para essa estatística, o CUSUM se define como segue:

$$C_m = \sum_{i=1}^{m}(\overline{s}_i - D_0) = \overline{s}_i - D_0 + \sum_{i=1}^{m-1}(\overline{s}_i - D_0) = (\overline{s}_i - D_0) + C_{m-1}. \tag{2.7}$$

Se o processo permanece sob controle no valor-alvo, D_0, a soma cumulativa definida em (2.7) é um passeio aleatório com média zero. No entanto, se a estatística de interesse se desloca para um

valor superior $D_1 > D_0$, diz-se que uma tendência positiva se desenvolverá na soma cumulativa C_m. Por outro lado, se essa estatística se desloca para baixo para um valor $D_1 < D_0$ uma tendência negativa será observada em C_m. Portanto, se houver uma tendência nos pontos plotados tanto para cima quanto para baixo, podemos considerar esse fato como evidência de que o processo não está sob controle estatístico (Hawkins e Olwell , 1998).

2.3.1 Gráfico CUSUM para observações individuais com distribuição Normal

Se enquanto o processo está sob controle as observações X_i são estatisticamente independentes e seguem uma distribuição normal com média μ_0 e desvio padrão σ, o CUSUM C_m pode ser definido de duas formas equivalentes. A primeira, na escala original, ou seja,

$$C_m = \sum_{i=1}^{m}(X_i - \mu_0). \tag{2.8}$$

Como C_m é a soma de normais independentes com média 0 e variância σ^2, sua distribuição é por conseguinte

$$C_m \sim N(0, m\sigma^2). \tag{2.9}$$

A segunda forma é obtida padronizando-se a variável X_i para ter média 0 e desvio padrão 1. As variáveis padronizadas U_i e a Soma Cumulativa C'_m são:

$$U_i = (X_i - \mu_0)/\sigma \tag{2.10}$$

$$C'_m = \sum_{i=1}^{m} U_i$$

As estatísticas C_m e C'_m contêm a mesma informação, porém, uma vantagem de padronizar X_i é que muitos gráficos CUSUM podem ter os mesmos valores de k e h e as escolhas desses parâmetros não dependem de escala (isto é, não dependem de σ). O parâmetros k e h representam o valor de referência e o limite de controle respectivamente. Quando o processo está sob controle, o CUSUM será um passeio aleatório com média zero e desvio padrão σ. Supondo que em um instante t a distribuição de X muda de $N(\mu_0, \sigma^2)$ para $N(\mu_0 + \delta, \sigma^2)$, então, no instante t a média de X sofre uma mudança de tamanho persistente δ. Escrevendo o CUSUM como

$$C_m = \sum_{i=1}^{m}(X_i - \mu_0) = \sum_{i=1}^{t}(X_i - \mu_0) + \sum_{i=t+1}^{m}(X_i - \mu_0), \tag{2.11}$$

temos que cada elemento da segunda somatória tem uma distribuição $N(\delta, \sigma^2)$. Assim, a segunda somatória

$$C_{m,t+1} = \sum_{i=t+1}^{m}(x_i - \mu_0), \tag{2.12}$$

terá uma distribuição normal com média $(m-t)\delta$ e variância $(m-t)\sigma^2$. Portanto, para um instante $t+1$ a média do CUSUM é $(m-t)\delta$. Isto significa que a partir do ponto (t, C_t) o CUSUM em média vai traçar um caminho centrado em uma linha com inclinação δ.

CUSUM Tabular

O CUSUM tabular permite monitorar a média de um processo acumulando desvios que são maiores que alvo μ_0 com uma estatística C^+, e acumulando desvios de que são menores com a estatística C^-. As estatísticas C^+ e C^- são chamadas CUSUM unilaterais superior e inferior, respectivamente.

Seja X_i a i-ésima observação da característica a ser controlada o CUSUM tabular se define como:

$$\begin{aligned} C_i^+ &= \max[0, x_i - (\mu_0 + K) + C_{i-1}^+], \\ C_i^- &= \min[0, (\mu_0 - K) - x_i + C_{i-1}^-], \\ \text{em que } C_0^+ &= C_0^- = 0. \end{aligned}$$

Se μ_1 é expressa em unidades de desvio padrão como $\mu_1 = \mu_0 + \delta\sigma$, então o valor de referência ou valor de tolerância denotado por K é a metade da magnitude da mudança:

$$K = \frac{\delta\sigma}{2} = \frac{|\mu_1 - \mu_0|}{2}, \tag{2.13}$$

onde δ é o tamanho da mudança em unidades de desvio padrão. Se C_i^+ ou C_i^- excedem o intervalo de decisão H, o processo será considerado fora de controle. Um valor razoável na prática para H é cinco vezes o desvio padrão do processo σ (Montgomery, 2005). Definindo $H = h\sigma$ e $K = k\sigma$, onde σ é o desvio padrão da variável amostral, se $h = 4$ ou $h = 5$ e $k = 1/2$, temos que em geral o CUSUM tem boas propriedades do ARL contra uma mudança de cerca de 1σ na média do processo (Montgomery , 2009). Em geral, os parâmetros H e K são selecionados de modo a fornecer um bom desempenho no ARL.

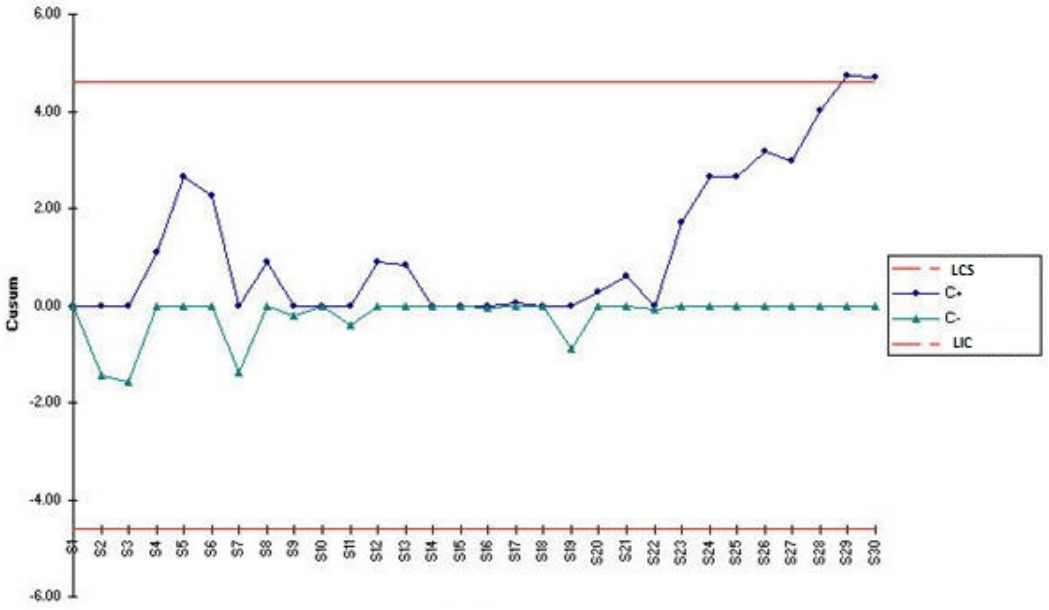


Figura 2.4: *Exemplo - Gráfico CUSUM*

A Tabela 2.1 apresenta o ARL, ou seja, o número de amostras necessárias para detectar uma mudança na média (múltiplo de σ) com $k = 1/2$ e $h = 4$ ou $h = 5$ para observações independentes com distribuição normal quando $\mu = \mu_1$ (Hawkins, 1993a).

δ (múltiplo de σ)	$h = 4$	$h = 5$
0	168	465
0,25	74,2	139
0,5	26,6	38
0,75	13,3	17
1	8,38	10,4
1,5	4,75	5,75
2	3,34	4,01
2,5	2,62	3,11
3	2,19	2,57
4	1,71	2,01

Tabela 2.1: *Valores de ARL para com* $k = 1/2$ *e* $h = 4$ *ou* $h = 5$
Fonte Montgomery (2009)

2.3.2 The Average Run Length (ARL)

Para os gráficos de controle CUSUM a medida ARL é calculada em função das escolhas de k e h. Assim, diferentes técnicas são usadas para calcular o ARL de um CUSUM. Brook e Evans (1975) têm utilizado uma abordagem para o cálculo do ARL que se baseia em modelar transições do estado sob controle para o estado fora de controle utilizando uma cadeia de Markov. Por outro lado, Hawkins (1992) forneceu um procedimento de cálculo de ARL simples mas preciso com base em uma equação de aproximação mas que exige uma tabela de constantes a serem aplicadas e tem uma precisão entre 1% e 3% para o verdadeiro valor do ARL. Woodall e Adams (1993) recomendam a aproximação dada por Siegmund (1985) por causa da sua simplicidade (Montgomery , 2009). Para um CUSUM unitaleral (isto é, C_i^+ ou C_i^-) com parâmetros h e k, a aproximação de Siegmund é

$$ARL = \frac{exp(-2\Delta b) + 2\delta b - 1}{2\delta^2}, \tag{2.14}$$

para $\Delta \neq 0$ onde $\Delta = \delta^* - k$ para o CUSUM unilateral superior C_i^+, $\Delta = -\delta^* - k$ para o CUSUM unilateral inferior C_i^-, $b = h + 1.166$ e $\delta^* = (\mu_1 - \mu_0)/\sigma$. Se $\Delta = 0$, pode-se usar $ARL = b^2$. A quantidade δ^* representa a mudança na média, em unidades de σ, para a qual deve ser calculado o ARL. Portanto, se $\delta^* = 0$ calcularia-se o ARL_0, enquanto se $\delta^* \neq 0$, obteria-se o valor de ARL_1 correspondente a uma mudança do tamanho de δ^*. Alternativamente, métodos numéricos de extração de raízes servem para calcular o ARL. Woodall e Adams (1993) discutem esses métodos.

2.4 CUSUM para distribuições da Família Exponencial

As distribuições de probabilidade que pertencem à família de exponencial são importantes na teoria estatística porque, além de terem boas propriedades em inferência, incluem várias distribuições úteis que são importantes na estatística aplicada. Nesta seção vamos utilizar a parametrização apresentada em Hawkins e Olwell (1998). A função densidade de probabilidade ou função massa de probabilidade para qualquer membro da família exponencial com um único parâmetro θ pode ser escrita como:

$$f(y|\theta) = exp\left\{a(y)b(\theta) + c(y) + d(\theta)\right\}, \tag{2.15}$$

sendo θ o parâmetro da distribuição e Y a correspondente variável aleatória.
Essa formulação é válida para funções discretas e contínuas da família exponencial. A função $b(\theta)$ é chamada de "parâmetro natural". A estatística suficiente minimal para estimar θ se encontra a partir de $a(y)$. A densidade conjunta para uma amostra aleatória de tamanho m de Y é dada por

$$f(y|\theta) = exp\left(\sum_{i=1}^{m} a(y_i)b(\theta) + \sum_{i=1}^{m} c(y_i) + md(\theta)\right). \tag{2.16}$$

Para testar se o parâmetro sob controle θ_0 mudou para um valor fora de controle θ_1 definimos a variável "score" Z_i

$$Z_i = ln\left(\frac{f_{\theta_1}(Y_i)}{f_{\theta_0}(Y_i)}\right) = a(Y_i)\left\{b(\theta_1) - b(\theta_0)\right\} + \left\{d(\theta_1) - d(\theta_0)\right\}. \tag{2.17}$$

O CUSUM, de forma recursiva, pode ser calculado mediante o algoritmo

$$D_i = \max(0, D_{i-1} + Z_i). \tag{2.18}$$

Especificamente,

$$D_i = max\left\{0, D_{i-1} + a(Y_i)\left\{b(\theta_1) - b(\theta_0)\right\} + d(\theta_1) - d(\theta_0)\right\},$$

sinalizando uma mudança quando $D_i > H$. Escrevendo $X_i = a(Y_i)$ definimos

$$k = -\frac{d(\theta_1) - d(\theta_0)}{b(\theta_1) - b(\theta_0)}. \tag{2.19}$$

Para $b(\theta_1) - b(\theta_0) > 0$ o CUSUM pode ser redimensionado dividindo a equação (2.4) por $b((\theta_1) - b(\theta_0))$. Assim, temos:

$$C_i^+ = max(0, C_{i-1}^+ + X_i - k),$$

onde $C_i^+ = D_i/(b(\theta_1) - b(\theta_0))$ e $h^+ = A/b(\theta_1) - b(\theta_0)$, sinalizado uma mudança para $C_i^+ > h^+$. Se $b(\theta_1) - b(\theta_0) < 0$ dividindo a equação (2.4) por $[b(\theta_1) - b(\theta_0)]$ inverte-se o sentido do CUSUM

$$C_i^- = min(0, C_{i-1}^- + X_i - k),$$

onde $C_i^- = D_i/(b(\theta_1) - b(\theta_0))$ e $h^- = A/(\theta_1) - b(\theta_0)$, sinalizando uma mudança para $C_i^- < h^-$. Os parâmetros h^+ e h^- são escolhidos de acordo com o ARL_0 desejado.

2.4.1 Família Normal

Dada a formulação geral para o CUSUM expressa na equação 2.16 é fácil derivar o CUSUM para uma variável com distribuição Normal com média μ e variância σ^2. Para σ fixo e conhecido a densidade da distribuição normal é escrita na forma da família exponencial como:

$$f(y|\mu) = exp\left(\frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2} - ln(\sigma\sqrt{2}\pi) - \frac{\mu^2}{2\sigma^2}\right), \tag{2.20}$$

a partir da qual segue que

$$\begin{aligned} a(y) &= y, \\ b(\mu) &= \frac{\mu}{\sigma^2}, \\ d(\mu) &= \frac{-\mu^2}{2\sigma^2}. \end{aligned}$$

Quando a média μ_0 do processo que está sob controle se desloca para um valor μ_1, com $\mu_1 > \mu_0$, o CUSUM implementado nas equações (2.7) e (2.8) é expresso por:

$$\begin{aligned} C_{i+1}^+ &= max(0, C_i^+ + Y_i - k^+), \\ k^+ &= \frac{-(\mu_1^2 - \mu_0^2)/2\sigma^2}{(\mu_1 - \mu_0)/\sigma^2} = \frac{\mu_0 + \mu_1}{2}. \end{aligned}$$

2.4.2 Família Poisson

A família Poisson também é um membro da família exponencial. Sua função massa de probabilidade, escrita na forma da família exponencial, é dada do seguinte modo

$$f(y|\lambda) = exp\left\{yln\lambda - \lambda - ln\Gamma(y+1)\right\}. \tag{2.21}$$

O CUSUM para monitorar uma mudança na média sob controle λ_0 que desloca-se para um valor λ_1, $\lambda_1 > \lambda_0$, é dado por

$$\begin{aligned} C_{i+1}^{+} &= max(0, C_i^{+} + Y_i - k^{+}), \\ k^{+} &= \frac{\lambda_1 - \lambda_0}{ln(\lambda_1/\lambda_0)} \end{aligned}$$

2.4.3 Famíilia Binomial Negativa

A distribuição binomial negativa tem diferentes parametrizações. Algumas representam o número total de tentativas até obter r sucessos em uma seqüência de ensaios binomiais, outras como o número total de falhas até obter r sucessos em uma seqüência de ensaios binomiais. Optamos por não utilizar qualquer uma destas formas, mas sim representar a binomial negativa como uma distribuição discreta de propósito geral que pode ser usada para modelar valores inteiros. A parametrização da binomial negativa utilizada nos Modelos Lineares Generalizados Hawkins (1992), com média μ e variância $\mu + \mu^2/\phi$, tem como função densidade discreta de probabilidade

$$f(y|\mu,\phi) = \frac{\Gamma(\phi+y)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\mu}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}. \tag{2.22}$$

Fixado o parâmetro de dispersão ϕ, a função massa de probabilidade escrita como um membro da família exponencial é dada por:

$$f(y|\mu,\phi) = exp\left\{y\log\left(\frac{\mu}{\mu+\phi}\right) + \phi\log\left(\frac{\phi}{\mu+\phi}\right) + \log\left(\frac{\Gamma(\phi+y)}{\Gamma(y+1)\Gamma(\phi)}\right)\right\}. \tag{2.23}$$

O CUSUM para monitorar a média μ_0 é dado por

$$\begin{aligned} C_{i+1}^{+} &= \max(0, C_i^{+} + Y_i - k^{+}), \\ k^{+} &= \frac{-\phi log\left\{(\phi+\mu_0)/(\phi+\mu_1)\right\}}{log\left\{(\mu_1(\phi+\mu_0))/\mu_0(\phi+\mu_1)\right\}}. \end{aligned}$$

Capítulo 3

Gráficos de controle para monitorar séries de contagem

Uma primeira abordagem bastante utilizada no Controle Estatístico de Processos para monitorar dados não normais consiste em normalizá-los mediante transformações para posteriormente serem analisados por meio de gráficos de controle para variáveis com distribuição Gaussiana. Outras abordagens baseiam-se no ajuste de modelos lineares generalizados para monitorar os resíduos padronizados do modelo quando se tem um bom ajuste. E finalmente, outro enfoque é desenvolvido para a análise de dados com distribuições não normais que pertencem à família exponencial.
Neste capítulo apresentaremos as cinco estatísticas de monitoramento que serão avaliadas em séries de contagem nesse trabalho:

1. transformação de Rossi *et al.* (1999) para dados com distribuição Poisson;
2. normalização proposta por Jørgensen (1996) para dados com distribuição binomial negativa;
3. resíduos de Anscombe para modelos lineares generalizados;
4. gráficos CUSUM para dados Poisson propostos por Rogerson e Yamada (2004);
5. gráficos CUSUM utilizando a estatística da razão de verossimilhanças (Hohle , 2007).

Os métodos propostos por Rossi *et al.* (1999) e Jørgensen (1996) consistem em transformações simples, dependendo unicamente dos parâmetros das distribuições Poisson e binomial negativa, respectivamente. Já os resíduos de Anscombe estão sendo propostos nesse trabalho para detectar processos fora de controle para dados que seguem modelos da família exponencial. Em modelos lineares generalizados com variáveis respostas que não seguem uma distribuição normal, os resíduos, geralmente, apresentam uma distribuição assimétrica, no entanto os resíduos de Anscombe apresentam uma distribuição mais próxima à distribuição normal estabilizando a variância (Paula, 2012).
Os dois últimos métodos já foram propostos como gráficos de controle CUSUM. O método Rogerson e Yamada (2004) é apresentado para dados com distribuição Poisson e nesse trabalho propomos novos parâmetros para monitorar dados com distribuição binomial negativa. Já o método proposto por Hohle (2007), é baseado na razão de verossimilhanças da distribuição binomial negativa. Além disso, consideramos todos os métodos para dados em que a média varia ao longo do tempo.

Para detectar mudanças na média em séries de contagem, nesse trabalho consideraremos somente uma observação a cada instante de tempo, sendo que a média pode variar ao longo do tempo em função de variáveis explicativas. Como é usual para contagens, a variância da variável muda em função da média, e será modelada utilizando-se distribuição Poisson ou binomial negativa. Por isso, nesse capítulo denotamos a variável resposta correspondente à contagem como X_t.

Este capítulo é dividido em duas seções, sendo que as três primeiras estatísticas propostas são apresentadas na primeira seção e são baseadas em transformações normalizadoras e resíduos padronizados, e na segunda seção são apresentadas as duas últimas que foram propostas diretamente como gráficos CUSUM.

3.1 Transformações normalizadoras

As três estatísticas $Z_{3,t}$, $Z_{4,t}$ e $Z_{5,t}$ propostas nesta seção serão analisadas por meio de um gráfico CUSUM-Gaussiano padronizado como segue

$$C_t = max(0, C_{t-1} + Z_{i,t} - k),$$

sinalizando uma mudança quando $C_t > h$. Os valores dos parâmetros k e h calculados para dados com distribuição normal serão determinados para obter um $ARL_0 = 500$ como é proposto na literatura de vigilância em saúde (Rossi e Marchi , 2010).

3.1.1 Transformação Rossi

Seja $X_1, X_2, ...$ uma sequência de observações independentes com distribuição de Poisson e $n_1, n_2, ...$ os respectivos tamanhos de amostra. Assumimos que X_i tem uma média sob controle $n_i\lambda_0$ e uma média fora de controle $n_i\lambda_1$, $i = 1, 2, ...$ e consideramos somente o caso unilateral $\lambda_1 > \lambda_0$. A primeira transformação para X_t com média sob controle $n_t\lambda_0$ e desvio padrão $\sqrt{n_t\lambda_0}$ baseada na normalidade assintótica está dada por:

$$Z_{1,t} = \frac{X_t - n_t\lambda_0}{\sqrt{n_t\lambda_0}}, \quad t = 1, 2..., \tag{3.1}$$

A segunda transformação é dada por

$$Z_2t = 2(\sqrt{X}_t - \sqrt{n_t\lambda_0}), \quad t = 1, 2..., \tag{3.2}$$

com média sob controle e desvio padrão $\sqrt{n_t\lambda_0}$ e $1/2$, respectivamente. Esta transformação normaliza e estabiliza a variância. Rossi *et al.* (1999) recomendam usar a média das duas transformações anteriores, pois produz valores de ARL mais próximos dos valores ARL esperados. A transformação é expressa como segue

$$Z_{3,t} = 0,5Z_{1t} + 0,5Z_{2t} = \frac{X_t - 3n_t\lambda_{0,t} + 2\sqrt{X_t n_t\lambda_{0,t}}}{2\sqrt{n_t\lambda_{0,t}}}. \tag{3.3}$$

Esta primeira estatística define o gráfico CUSUM-Gaussiano padronizado para dados de contagem provenientes de uma distribuição de Poisson. A distribuição resultante para a transformação na equação (3.3) é aproximadamente normal com média sob controle 0 e desvio padrão 1 (Rossi *et al.* , 1999).

3.1.2 Transformação Jørgensen

A segunda estatística define o gráfico de CUSUM Gaussiano padronizado para dados de contagem provenientes de uma distribuição binomial negativa. Jørgensen (1996) apresenta uma transformação baseada na normalidade assintótica da variável observada, X_t, com média sob controle $\mu_{0,t}$ e variância $\mu_{0,t} + \mu_{0,t}^2/\phi$ respectivamente, e parâmetro de dispersão ϕ fixo. A transformação está dada por

$$Z_{4,t} = \frac{X_t - \mu_{0,t}}{\sqrt{\phi\pi/(1-\pi^2)}}, \tag{3.4}$$

$$\pi = \mu_{0,t}/(\mu_{0,t} + \phi)$$

A distribuição resultante desta transformação é aproximadamente normal com média sob controle igual a 0 e desvio padrão 1 (Jørgensen , 1996).

3.1.3 Resíduos de Anscombe

Para detectar mudanças em séries de tempo, uma estratégia implementada na literatura consiste na análise dos resíduos da série ajustada em Modelos Lineares Generalizados, já que em um bom ajuste do modelo os resíduos são independentes estacionários e apresentam uma distribuição aproximadamente normal. Em modelos lineares generalizados com variáveis respostas que não seguem uma distribuição normal os resíduos apresentam geralmente uma distribuição assimétrica, porém os resíduos de Anscombe normalizam e estabilizam a variância mediante uma função $A(X)$ (McCulloch e Searle , 2001), (McCullagh e Nelder. , 1989) e estão mais próximos da normalidade em MLGs (Paula, 2012). A escolha, desta função $A(X)$ é feita de modo que os resíduos resultantes sejam os mais normais possíveis. Os resíduos de Anscombe são definidos como:

$$r_{A_i} = \frac{\widehat{\phi}^{1/2} A(x) - A(\mu)}{V^{1/2}(\mu) A'(\mu)}. \tag{3.5}$$

Para os MLGs a função $A(.)$ é definida por

$$A(.) = \int_{-\infty}^{\mu} V^{-1/3}(t) dt,$$

em que $\widehat{\phi}$ é a estimativa do parâmetro de dispersão ϕ e $V(t)$ é a função de variância. Por exemplo, para a distribuição de Poisson

$$A(.) = \int_{-\infty}^{\mu} V^{-1/3}(t) dt = \frac{3}{2}\mu^{2/3} \tag{3.6}$$

e os resíduos de Anscombe para dados com distribuição Poisson são dados por

$$r_{Ai} = \frac{\frac{3}{2}(x^2/3 - \mu^2/3)}{\mu^(1/6)}. \tag{3.7}$$

Para a distribuição binomial negativa, $V(t) = \mu + \phi\mu^2$ e os resíduos de Anscombe são dados por

$$r_{Ai} = \frac{3/\alpha \left\{(1+\alpha x)^{2/3} - (1+\alpha\mu)^{2/3}\right\} + 3(x^{2/3} - \mu^{2/3})}{2(\alpha\mu^2 + \mu)^{1/6}}, \tag{3.8}$$

com $\alpha = 1/\phi$. Os resíduos padronizados e studentizados são respectivemente expressos como:

$$A_{s_i} = \frac{r_{Ai}}{\sqrt{\widehat{\phi}(1-h_i)}}$$
$$A_{t_i} = \frac{r_{Ai}}{\sqrt{\widehat{\phi}_i(1-h_i)},}$$

onde $\widehat{\phi}_i$ é uma estimativa de uma etapa de ϕ após excluir a observação i.

Seja X_t uma série temporal para dados de contagem com distribuição binomial negativa, assumindo que X_t tem uma média sob controle $\mu_{0,t}$, os resíduos de Anscombe são calculados como

segue

$$Z_{5,t} = \frac{3/\alpha \left\{(1+\alpha X_t)^{2/3} - (1+\alpha\mu_{0,t})^{2/3}\right\} + 3(X_t^{2/3} - \mu_{0,t}^{2/3})}{2(\alpha\mu_{0,t}^2 + \mu_{0,t})^{1/6}} \tag{3.9}$$

com $\alpha = 1/\phi$. ϕ é parâmetro de dispersão fixo, substituindo ϕ por $\widehat{\phi}$

3.1.4 CUSUM-Binomial Negativa, Rogerson e Yamada

Para dados com distribuição Poisson, Rogerson e Yamada (2004) propõem um gráfico CUSUM com parâmetros k e h que variam ao longo do tempo para detectar mudanças em séries de tempo com efeitos sazonais. Eles discutem os resultados enganosos que seriam obtidos ao implementar simplesmente um CUSUM com parâmetros constantes, pois a média sob controle oscilaria entre períodos.
Seja X_t uma série temporal para dados de contagem com distribuição Poisson, assumindo que X_t tenha uma média sob controle $\lambda_{0,t}$ o CUSUM proposto por eles é dado por:

$$C_t = max[(0, C_{t-1} + c_t(X_t - k_t))]$$

Os valores de k_t estão baseados nos CUSUMs da Família Exponencial, discutidos na revisão da Literatura, para as médias $\lambda_{0,t}$ e $\lambda_{1,t}$ como segue:

$$k = \frac{\lambda_{0,t} - \lambda_{1,t}}{ln(\lambda_{1,t}/\lambda_{0,t})}, \tag{3.10}$$

Por sua vez, o valor de h_t é escolhido para um valor de ARL_0 desejado e c_t é escolhido como a taxa entre h e h_t. Assim, $c_t = h/h_t$, onde h_t é o valor do limiar associado com k_t , o valor desejado de ARL_0 e os valores constantes de $\mu_{0,t}$ e $\mu_{1,t}$. Este parâmetro fará uma contribuição para a sinalização de h. Se, por exemplo, $h > h_t$, então $X_t - k_t$ é aumentado pelo fator h/h_t. Esse método será aplicado neste trabalho para dados de contagem com sobredispersão que seguem uma distribuição Binomial Negativa.
Seja X_t uma série temporal para dados de contagem com distribuição Binomial Negativa, assumindo que X_t tem uma média sob controle $\mu_{0,t}$ o CUSUM proposto é dado por:

$$C_n = max[(0, C_{n-1} + c_t(X_t - k_t))].$$

O parâmetro k_t baseados nos gráficos CUSUM para distribuições da familia exponencial é proposto como segue

$$k_t = \frac{-\phi log\left\{(\phi + \mu_{0,t})/(\phi + \mu_{1,t})\right\}}{log\left\{\mu_{1,t}(\phi + \mu_{0,t})/\mu_{0,t}(\phi + \mu_{1,t})\right\}} \tag{3.11}$$

3.1.5 CUSUM- Hohle

Um método alternativo para monitorar séries de contagem baseia-se na razão de máxima verossimilhança. Para m observações e τ (ponto de mudança, conhecido) é assumido o seguinte modelo:

$$x_t|z_t, \sim \begin{cases} f_{\theta_0}(\cdot|z_t)\ para\ t\ =\ 1\ ,...,\ \tau - 1\ \ (sob\ controle) \\ f_{\theta_1}(\cdot|z_t)\ para\ t\ =\ \tau,\ \tau + 1, ...\ \ (fora\ de\ controle) \end{cases} \tag{3.12}$$

em que z_t denota as covariáveis conhecidas no tempo t e f_θ é por exemplo, a função massa de probabilidade da distribuição Binomial Negativa com média μ que é função de θ e z_t.
O objetivo desse modelo é detectar o ponto de mudança usando as observações $x_1, ..., x_n$ e decidir se uma mudança ocorreu durante $1, ..., n$. Um modo para detectar esse mudança é a utilização do

CUSUM-Máxima verossimilhança (Frisén e Sonesson , 2005). O instante de mudança é dado por:

$$N = min\left\{n \geq 1 : \max_{1\leq\tau\leq n}\left[\sum_{t=\tau}^{n} log\left\{\frac{f_{\theta_0}(x_t|z_t)}{f_{\theta_1}(x_t|z_t)}\right\}\right] \geq h\right\}. \tag{3.13}$$

Para um n específico calcula-se a estatística da razão de verossimilhanças LR ou para testar a hipótese de que todas as observações provêm de uma distribuição sob controle contra a alternativa de que a partir de um ponto τ a distribuição está fora de controle. Maximizar a estatística LR para cada possível ponto $1 \leq \tau \leq n$ significa encontrar o estimador de máxima verossimilhança de τ para obter a localização mais provável do ponto de mudança. Se $LR(t)$ está acima de um limiar pré-determinado h, então existe informação suficiente para identificar que houve uma mudança em τ. Caso contrário nenhuma decisão é tomada e o monitoramento continua no tempo $n+1$. Pré-especificado θ_0 e θ_1 o CUSUM pode ser escrito na forma recursiva:

$$C_0 = 0, \quad C_t = max\left(0, C_{t-1} + log\left\{\frac{f_{\theta_0}(x_t)}{f_{\theta_1}(x_t)}\right\}\right), \quad n \geq 1 \tag{3.14}$$

O primeiro alarme é dado no momento $N = min\{n : C_t \geq h\}$. Este detector é ótimo para a detecção de uma mudança de θ_0 para θ_1 (Hohle , 2007). Para avaliar o desempenho desse método são consideradas as estatísticas ARLs: O comprimento médio da sequências, $ARL_0 = E(N|\tau = \infty)$, ou seja, o tempo de espera até um primeiro falso alarme e $ARL_1 = E(N|\tau = 0)$, que é o tempo esperado até detectar uma mudança, quando esta possibilidade ocorre imediatamente. Lucas (1985) abordou o método de CUSUM baseado na razão de verossimilhança para a distribuição de Poisson com parâmetros constantes $\mu_0 = \theta_0$ e $\mu_1 = \theta_1$.

Capítulo 4

Ajuste de Modelos e Simulação de dados

Neste capítulo apresentam-se os modelos lineares generalizados (MLG) que serão implementados para a análise da série temporal de contagem do número diário de internações por causas respiratórias de pessoas com 65 anos de idade, ou mais, residentes no município de São Paulo. A média da série será modelada com variáveis explicativas e baseado nesse ajuste para o período de janeiro de 2006 a dezembro de 2010 será calculado o valor alvo para a série no ano de 2011, conforme recomenda (Unkel *et al.* , 2012). Ajustado o modelo serão simulados os dados que serão utilizados no capítulo seguinte para avaliar e comparar a eficiência e rapidez das cinco estatísticas para detectar mudanças em séries de contagem, citadas no Capítulo 3. As distribuições consideradas para ajustar esta série serão Poisson e Binomial Negativa por se tratarem de distribuições para dados de contagem. Para estes modelos a população será incluida com coeficiente constante (offset) e a função de ligação adotada será o logaritmo neperiano. Desse modo as conclusões do modelo se referem à taxa de internações por 100 mil habitantes ao longo do tempo. Os dados diários de internações foram obtidos no Sistema de Informação Hospitalar (SIH) da Secretaria de Saúde do Município de São Paulo (PRO-AIM).

Fixando a notação utilizada nesse capítulo, considere:

- y_t contagem observada no instante t;
- g_t função utilizada como offset incluída no modelo como uma variável explicativa com coeficiente igual a 1;
- Pop_t população no instante t;
- T número de observações, sendo que o índice $t = 1, ..., T$;
- μ_t valor esperado de y_t, $E(y_t)$.

4.1 Ajuste dos modelos

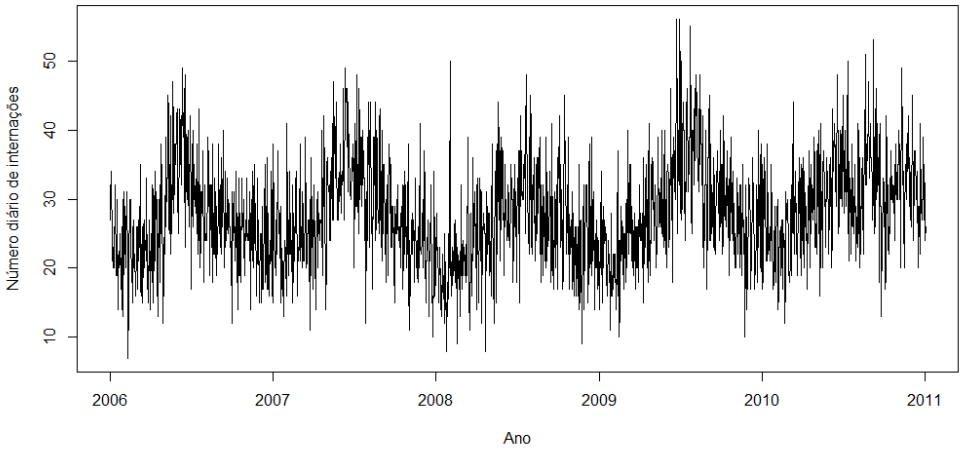


Figura 4.1: *Número diário de internações por causas respiratórias no período de janeiro 2006 a dezembro 2010*

A série acima mencionada apresenta um comportamento sazonal que se assemelha às funções seno e cosseno figura (**??**), de forma que, essas funções foram implementadas para explicar a sazonalidade das internações como é proposto em Serfling (1963). Também foi considerada a variável categórica dia da semana para controlar o efeito destes dias nas internações. Propomos modelo com função de ligação logaritmo neperiano de modo que:

$$ln(\mu_t) = \underset{\sim}{x}\underset{\sim}{\beta} + g_t \tag{4.1}$$

A variável população foi incluída no modelo utilizando-se o logaritmo neperiano ao tamanho da população dividido por 100.000, assim:

$$g_t = log(pop_t/100000), \tag{4.2}$$

assim temos um modelo para a taxa de internações

$$ln(y_t) = ln\left(\frac{\mu_t}{g_t}\right) = ln\left(\frac{\mu_t}{pop_t}100000\right). \tag{4.3}$$

Modelo linear generalizado com distribuição Poisson

Seja y_t o número de internações no instante t. O modelo linear generalizado proposto de Poisson para o período de janeiro 2006 a dezembro 2010 é:

$$\begin{aligned} Y_t &\sim Poisson\left(\mu_{0,t}\right) \\ ln(y_t) = ln\left(\frac{\mu_{0,t}}{g_t}\right) &= \beta_0 + \beta_1 \cos(2\pi t/365) + \beta_2 \sin(2\pi t/365) + \sum_{i=1}^{6} \beta_i dia_i \end{aligned} \tag{4.4}$$

Desse modo a inclusão do termo offset para $g_t = ln(pop_t/100000)$ torna possível modelar a taxa média de internações. O dia_1 corresponde a segunda, dia_2 a terça e assim sucessivamente. Os dados mostraram sobredispersão e o modelo proposto não mostrou um bom ajuste como se vê no gráfico na figura 4.2 do componente do desvio (Paula, 2012) .

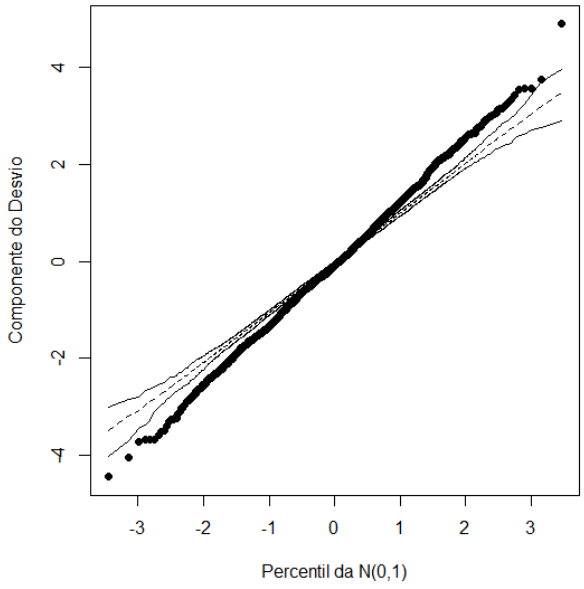


Figura 4.2: *Componente do desvio para o modelo Poisson*

Em consequência deste, ajustou-se um modelo com distribuição binomial negativa, como apresentado na Figura 4.5. O ajuste deste último modelo foi avaliado usando as técnicas de diagnóstico propostas por (Paula, 2012). O gráfico de envelope da Figura 4.3 não apresenta indicações de afastamentos sérios da suposição de distribuição binomial negativa e não há indícios de que a função de ligação utilizada seja inapropriada. Dentre as observações destacadas pelos gráficos de diagnóstico Figura 4.4 nenhuma observação apresenta variação desproporcional. Por conseguinte este modelo será implementado nas análises posteriores para calcular o valor alvo para essa série no período janeiro de 2011 a dezembro de 2011. O modelo binomial negativa ajustado foi:

$$\begin{aligned} Y_t &\sim BinNeg(\mu_{0,t}, \phi) \\ ln(ty_y) = ln\left(\frac{\mu_{0,t}}{g_t}\right) &= \beta_0 + \beta_1 cos(2\pi t/365) + \beta_2 sin(2\pi t/365) + \sum_{i=1}^{6} \beta_i dia_i \end{aligned} \tag{4.5}$$

As estimativas dos parâmetros para o modelo binomial negativo são apresentados na Tabela 4.1:

Tabela 4.1: *Estimativas dos parâmetros para o modelo binomial negativa*

	Estimativa	Erro Padrão	Valor p
Intercepto	1,19	0,014	$<$ 0,001
Cosseno	-0,173	0,007	$<$ 0,001
Seno	-0,046	0,007	$<$ 0,001
Segunda	0,246	0,020	$<$ 0,001
Terça	0,203	0,020	$<$ 0,001
Quarta	0,213	0,020	$<$ 0,001
Quinta	0,195	0,020	$<$ 0,001
Sexta	0,215	0,020	$<$ 0,001
Sábado	0,048	0,021	0.0196
ϕ	70,07	8,24	

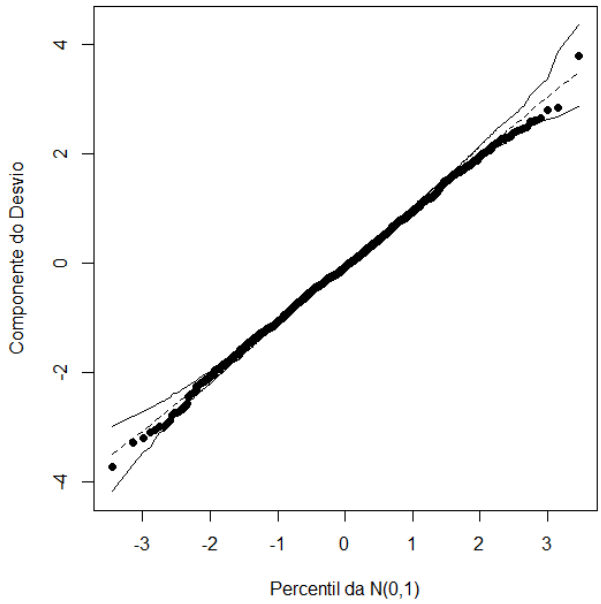


Figura 4.3: *Componente do desvio para o modelo binomial negativa*

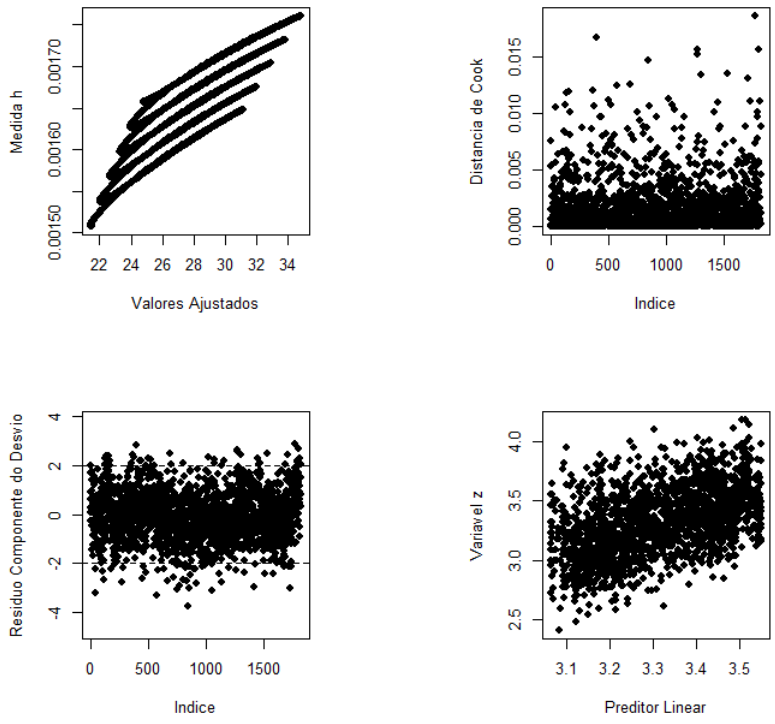


Figura 4.4: *Diagnóstico do modelo binomial negativo proposto por Paula (2012)*

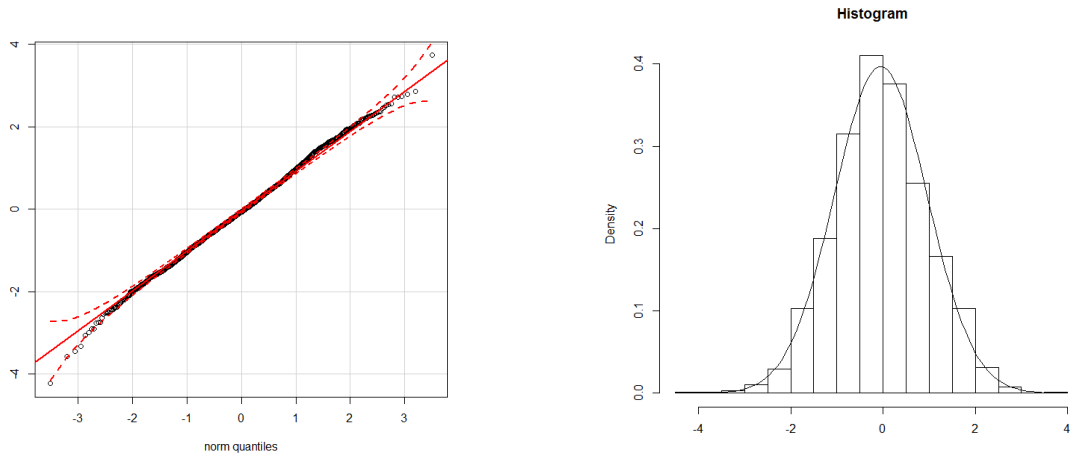


Figura 4.5: *Gráfico de quantis e histograma dos resíduos do modelo binomial negativo*

4.2 Simulação

A eficiência das cinco estatísticas já apresentadas no Capítulo 3 será avaliada por meio de simulações, devido à dificuldade de obter as expressões fechadas das distribuições dos CUSUM para as estatísticas Z_3, Z_4, Z_5 e do método proposto por (Rogerson e Yamada , 2004). Os parâmetros dos gráficos CUSUM para o método proposto por Hohle serão calculados através de cadeias de Markov, já programadas no software R na biblioteca algo.glrnb (Wimmer e Höhle , 2013)
A estimativa da média sob controle $\mu_{0,t}$ da série de interesse para o ano 2011 foi calculada baseada nos coeficientes obtidos do ajuste do modelo binomial Negativo (6.4) para o período de janeiro 2006 a dezembro 2010 como segue:

$$\widehat{\mu_{0,t}} = exp\{1.19 - 0.046\sin(2\pi t/365) - 0.17\cos(2\pi t/365) + log(pop_t/100000) + dms\} \tag{4.6}$$

$$\widehat{\phi} = 70,07$$

A média fora de controle $\mu_{1,t}$ será expressa considerando um aumento na escala log

$$\mu_{1,t} = \exp(\delta)\mu_{0,t}, \tag{4.7}$$

assim as mudanças na média correspondem a efeitos multiplicativos. Os valores de δ considerados nesse trabalho são:

$$\delta = \{1; 1.25; 1.35; 1.5; 1.75; 2\}. \tag{4.8}$$

Quando $\delta = 1$ o processo permanece sob controle e estará sujeito ao erro tipo I. Um aumento na média sob controle μ_0 de, por exemplo, $\delta = 1.25$ representará um aumento percentual no número diário de internações de 25% com respeito a μ_0.

Na Figura 4.6 apresenta-se a série temporal do número diário de internações no período de janeiro de 2006 a dezembro de 2010 e a média ajustada (cor vermelho) do número de internações seguindo o modelo binomial Negativo .

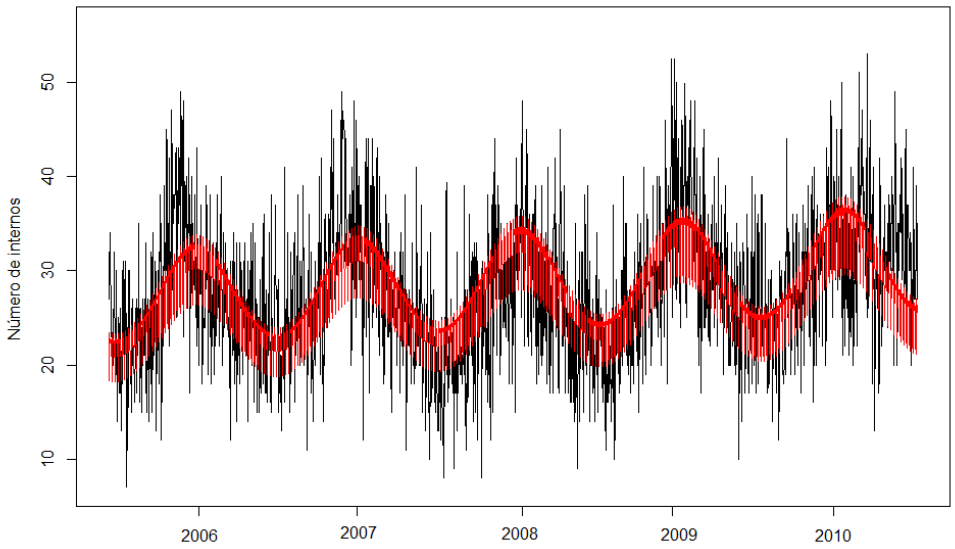


Figura 4.6: *Número diário de internações e número médio ajustado -Binomial Negativa*

Determinação dos limites de controle para os gráficos CUSUM

Baseados nos artigos de vigilância em saúde (Hohle e Paul , 2008) e (Rossi e Marchi , 2010), os gráficos CUSUM que serão implementados para monitorar as cinco estatísticas foram desenvolvidos para detectar uma mudança multiplicativa de tamanho $\delta = 2$, isto é, $\mu_{1,t} = \delta\mu_{0,t}$. Os respectivos limiares k e h foram calculados para obter um ARL_0 de 500 como segue:

i. As estatísticas $Z_{3,t}$ e $Z_{4,t}$ e $Z_{5,t}$ são analisadas por meio dos gráficos tipo CUSUM sob a suposição de que os dados seguem uma distribuição normal, o limite de controle h foi calculado para obter um $ARL_0 = 500$. Na Figura 4.7 mostra-se os valores de ARL_0 em função do limite de controle h.

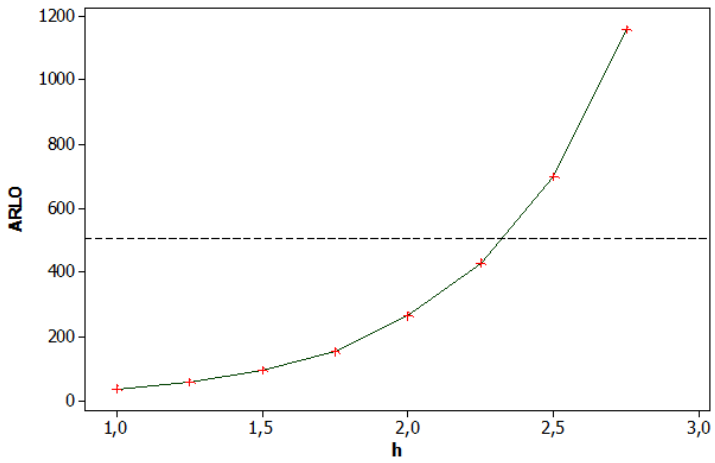


Figura 4.7: *Valor de ARL_0 em função de h / Estatísticas Normalizadoras*

ii. Para a estatística proposta por Rogerson e Yamada que será monitorada por meio do gráfico tipo CUSUM para dados com distribuição binomial negativa, o parâmetro k_t variará ao longo do tempo como foi proposto em Rogerson e Yamada (2004) para dados com distribuição Poisson. Assim, para cada valor de k_t foi calculado o respectivo h_t para se obter um valor de $ARL_0 = 500$. O valor do limiar h "global"foi calculado para se obter um valor de $ARL_0 = 500$

usando um valor de k constante igual a média de k_t para $\Delta = 2$. Na Figura 4.8 mostra-se os valores de ARL_0 em função do limiar h "global".

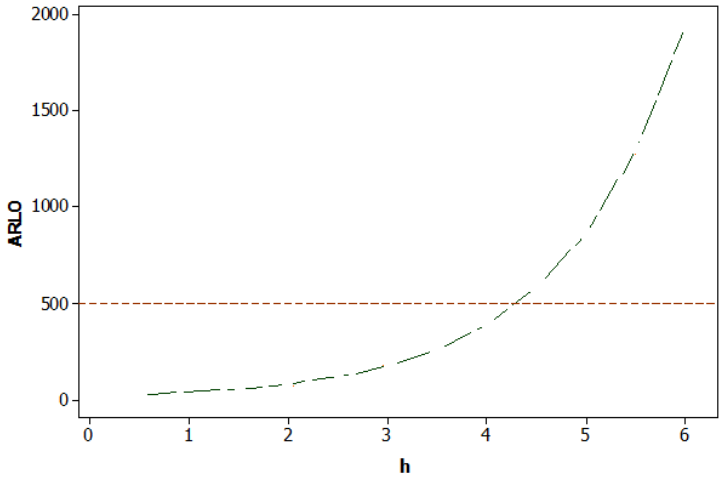


Figura 4.8: *Valor de ARL_0 em função do límite h - Rogerson e Yamada*

iii. Para monitorar a estatística proposta por Hohle (2007) que está baseada na estatística de razão verossimilhanças, o limiar h foi calculado com o pacote surveillance no sofware R (Wimmer e Höhle , 2013) para obter um $ARL_0 = 500$. Na Figura 4.9 mostra-se os valores de ARL_0 em função do h.

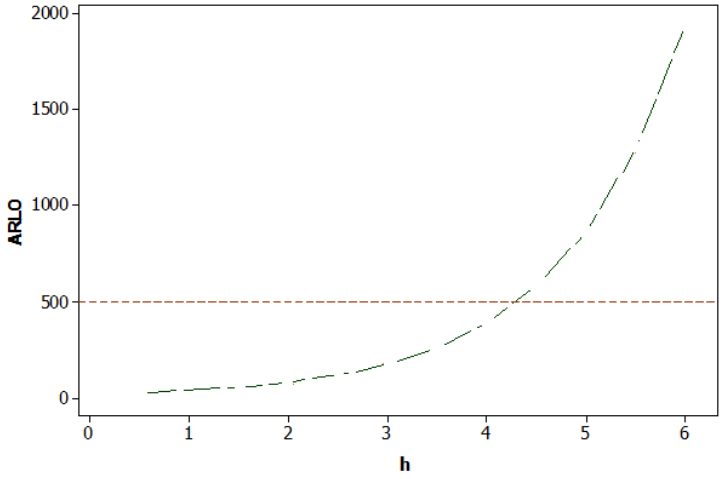


Figura 4.9: *Valor de ARL_0 em função do limite de controle h - LR*

Simulações para monitorar por meio dos gráficos tipo Shewhart

Inicialmente serão simuladas 365 observações 30.000 vezes correspondentes à série do número diário de internações por causas respiratórias no período janeiro 2011 a dezembro 2011. As observações serão simuladas seguindo uma distribuição binomial negativa com média e variância $\delta\mu_{0,t}$ $\delta\mu_{0,t} + \delta\mu_{0,t}^2/\phi$, respectivamente, com parâmetro de dispersão ϕ de $70,07$. Esse algoritmo se repetirá para os diferentes valores de δ mencionados anteriormente.
Para cada série gerada serão calculadas as estatísticas $Z_{3,t}$ e $Z_{4,t}$ e $Z_{5,t}$, apresentadas no Capítulo 3 serão obtidos limiares tal que $ARL_0 = 500$. Também serão avaliadas as suposições de normalidade e independência dessas estatísticas. Por último, serão obtidas as distribuições empíricas do RL (Run Length Distribution) e as estatísticas ARL_0 e ARL_1 para os diferentes valores δ.

Simulações para monitorar por meio dos gráficos tipo CUSUM

A avaliação e comparação da eficiência das cinco estatísticas apresentadas no Capítulo 3 serão analisadas por meio dos gráficos tipo CUSUM. Serão simuladas 365 observações 30.000 vezes com distribuição binomial negativa com média $\delta\mu_{0,t}$ e variância $\delta\mu_{0,t} + \delta\mu_{0,t}^2/\phi$, com parâmetro de dispersão ϕ de $70,07$. Os parâmetros h e k para o gráfico tipo CUSUM foram calculados para obter um valor de $ARL_0 = 500$. Na área de vigilância em saúde (Rossi *et al.* (1999) e Hohle (2007)) usam-se esses valores para o monitoramento de séries de tempo. Em seguida serão calculadas as distribuições empíricas do RL (Run Length Distribution) e as estatísticas ARL_0 e ARL_1 para os diferentes valores δ.

Os valores do ARL_1 serão calculadas com os limiares de controle calculados para obter um $ARL_0 = 500$ porém, será monitorado a série simulada com distribuição binomial negativa com média fora de controle μ_1 com $\mu_1 = \mu_0 \exp(\delta)$ para $\delta \neq 1$.

Capítulo 5

Resultados de Simulação

Neste capítulo apresentaremos o desempenho dos gráficos de controle Shewhart e CUSUM usando as estatísticas apresentadas no Capítulo 3. O desempenho é medido pelo número médio e mediano de instantes até a detecção da mudança na média. Essas medidas serão denotadas respectivamente ARL e MRL (Average and Median Run Length). Para isso utilizaremos os dados simulados da série do número de internações para o ano 2011. Os limiares de controle h e os valores de referência k utilizados inicialmente foram obtidos supondo normalidade das estatísticas, como é usual na literatura, fixando um ARL_0 esperado de 500. Isto é, em média, serão necessárias 500 observações para emitir um sinal fora de controle mesmo quando o processo está sob controle. No entanto, os valores dos ARL_0 obtidos nas simulações para as três estatísticas normalizadoras foram bem menores que 500. Para as estatísticas Rogerson e Yamada e Hohle obtiveram-se valores de ARL_0 próximos a 500. Para comparar e avaliar o desempenho das cinco estatísticas de monitoramento serão modificados os limiares dos gráficos de controle para as três estatísticas normalizadoras, para obter via simulação um valor de $ARL_0 = 500$.

Para os gráficos tipo CUSUM o parâmetro de referência k modificado variará ao longo do tempo dependendo do deslocamento da média sob controle, como proposto em Rossi e Marchi (2010). Assim. temos:

$$k_t = E\left[\frac{Z_{i,t}; \mu_{1,t}}{2}\right], \qquad i = 1, 2, 3 \tag{5.1}$$

O limiar h será calculado para detectar mudanças de tamanho $\delta = 2$ para um k constante igual à média de k_t. Os novos valores do parâmetro h são apresentados na Tabela 5.7.

5.1 Resultados - Gráficos de controle tipo Shewhart

Na Tabela 5.1 são apresentados os valores ARL_0 e ARL_1 e seu respectivo desvio padrão, obtidos via simulação para diferentes valores de δ como resultado da implementação das três estatísticas normalizadoras. O limiar superior de controle utilizado para estes gráficos foi $L = 2,88$ que, supondo normalidade das estatísticas, produziria o valor esperado de ARL_0 igual a 500.

Tabela 5.1: *ARL(EP) para os Gráficos de Shewhart*

Estatística	Valores de δ					
	1	1.25	1.35	1.5	1.75	2
Rossi	168,5(2,35)	10,3(0,10)	6,5(0,05)	3,1(0,04)	1,7(0,01)	1,2(0,01)
Jorgensen	369,1(7,10)	20,3(0,18)	12,1(0,05)	4,7(0,06)	2,1(0,02)	1,4(0,01)
R. Anscombe	267,9(3,30)	31,8(0,22)	16,4(0,07)	6,5(0,08)	2,6(0,03)	1,6(0,01)

A primeira coluna ($\delta = 1$) corresponde aos valores dos ARL_0, isto é, o número de observações que foram necessárias para emitir um sinal fora de controle mesmo quando a média estava sob controle,

as colunas restantes correspondem aos valores dos ARL_1 quando a média sob controle μ_0 muda um fator δ. Para as três estatísticas analisadas obtiveram-se valores de ARL_0 menores que 500. No entanto, destaca-se que a transformação proposta por Jorgensen apresentou um ARL_0 de $369,1$ com grande erro padrão de $7,1$. Para as estatísticas Rossi e R. Anscombe 75% dos sinais fora de controle emitidos, mesmo quando o processo estava sob controle (ou seja falsos alarmes), precisaram no máximo de 228,5 e 355,1 observações respectivamente. Enquanto a estatística Jorgensen precisou no máximo de 355 observações (Figura 5.1). Porém, a comparação dos ARL_1 não é justa. Para comparar o desempenho das estatísticas normalizadoras novos limiares foram obtidos via simulação para obter um ARL_0 de 500. Na Tabela 5.2 apresentam-se os novos valores de L.
Na Tabela 5.3 se apresentam os novos resultados dos ARL depois de modificar os limiares de controle.

Tabela 5.2: *Valores de L ajustados via simulação ARL* $= 500$

Estatística	L
Rossi	3,51
Jorgensen	3,21
Anscombe	3,32

Tabela 5.3: *ARL para os Gráficos de Shewhart* $ARL_0 \approx 500$

Estatística	Valores de δ					
	1	1.25	1.35	1.5	1.75	2
Rossi	489,9(10,08)	28,7(0,35)	13,4(0,28)	5,4(0,06)	2,2(0,02)	1,2(0,01)
Jorgensen	510,81(7,07)	26,1(0,41)	12,1(0,10)	5,1(0,10)	2,1(0,02)	1,3(0,01)
R. Anscombe	481,92(10,28)	31,1(0,42)	14,4(0,12)	6,5(0,08)	2,5(0,03)	1,2(0,01)

Para esta nova simulação as estatísticas Rossi, Jorgensen e os Resíduos de Anscombe apresentaram valores de ARL_0 próximos a 500, assim como os valores do MRL_0, ou seja, 50% de falsos alarmes obtidos para estas estatísticas precisariam respectivamente de 477,8; 511,6; 467,4 observações antes de emitir um sinal fora de controle, como pode ser observado nos boxplots da Figura 5.2. Os valores dos ARL_1 para valores de δ baixos apresentaram valores muito altos para as três estatísticas, confirmando que o gráfico de controle tipo de Shewhart não detecta com rapidez pequenas mudanças na média. Para valores de δ inferiores a $1,5$ seria necessário, em média, um mês para sinalizar esses deslocamentos na média. Para valores de δ maiores que $1,5$ as três estatísticas sinalizam com maior rapidez as alterações na média.

5.2 Resultado - Gráficos de controle tipo CUSUM

Na Tabela 5.4 são apresentados os valores do ARL obtidos por simulação monitorando as séries simuladas por meio dos gráficos de controle tipo CUSUM. Para as três estatísticas normalizadoras se utilizaram os parâmetros h e k que, supondo normalidade, forneceriam um $ARL_0 = 500$. Para os métodos Rogerson e Yamada e Hohle foram implementados os parâmetros que se mostram na Tabela 5.5 para os quais obtém-se um $ARL_0 = 500$.

Tabela 5.4: *ARL para os Gráficos de CUSUM- considerando normalidade*

Estatística	Valores de δ					
	1	1.25	1.35	1.5	1.75	2
Rossi	115,7 (1,4)	7,0 (0,09)	4,2 (0,07)	2,5 (0,03)	1,6 (0,01)	1,2 (0,004)
Jorgensen	267,9 (5,1)	9,1 (0,13)	4,9 (0,07)	2,8 (0,02)	1,7 (0,01)	1,3 (0,003)
Anscombe	340,3 (5,0)	11,1 (0,20)	5,9 (0,08)	3,2 (0,03)	1,9 (0,01)	1,4 (0,001)
Rogerson e Yamada	498,1 (6,2)	9,1 (0,07)	4,9 (0,07)	3,0 (0,04)	1,8 (0,02)	1,3 (0,004)
Hohle	500,3 (2,3)	7,7 (0,07)	4,6 (0,05)	2,8 (0,02)	1,7 (0,01)	1,8 (0,005)

Os valores obtidos do ARL_0 para as estatísticas Rossi, Jogersen R. Anscombe, que deveriam normalizar os dados, são bem menores que o ARL_0 esperado de 500. Porém, os métodos Rogerson e Yamada e Hohle (LR) para dados com distribuição binomial negativa apresentaram valores de ARL_0 próximos de 500. Vale ressaltar que a estatística Rossi apresentou um número maior de falsos alarmes como se vê no boxplot da Figura 5.3. Consequentemente, a comparação dos ARL_1 seria justificável somente para os métodos Hohle e Rogerson e Yamada. No entanto, o interesse deste trabalho é comparar as cinco estatísticas monitoradas mediante este tipo de gráfico de controle. Portanto, para obter ARL_0 próximo a 500 foram modificados via simulação os limiares para os gráficos CUSUM-Gaussiano das três primeiras estatísticas. Nesse procedimento o parâmetro de referência k foi modificado e variou ao longo do tempo dependendo do deslocamento da média sob controle como é proposto em Rossi e Marchi (2010)(Tabela 5.1).

Tabela 5.5: *Valores de h e k para obter um valor teórico de* $ARL_0 = 500$

Estatística	k	h
Rossi	1	2,32
Jorgensen	1	2,32
Anscombe	1	2,32
Rogerson e Yamada	21,4	5,01
Hohle	-	3,96

Na Tabela 5.6 apresentam-se os valores dos ARL obtidos por simulação posteriormente à modificação dos limiares das três estatísticas normalizadoras.

Tabela 5.6: *ARL para os Gráficos CUSUM-* $ARL_0 \approx 500$

Estatística	Valores de δ					
	1	1.25	1.35	1.5	1.75	2
Rossi	492,2(5,2)	18.9(0,21)	8,8(0,11)	3,9(0,08)	1,9(0,01)	1,3(0,006)
Jorgensen	487,9(7,4)	17,5 (0,18)	8,7(0,10)	4,0(0,04)	1,8(0,01)	1,3(0,007)
R. Anscombe	499,1(7,5)	16,2 (0,20)	8,6(0,12)	3,9(0,08)	1,8 (0,02)	1,2(0,006)
Rogerson e Yamada	498,1(6,2)	9,1 (0,07)	5.0(0,07)	3,0(0,04)	1,8(0,02)	1.3(0.004)
Hohle	500,2 (5,3)	7,7 (0,07)	4,6 (0,05)	2,8 (0,02)	1,7 (0,01)	1,3 (0,005)

Os valores dos ARL_0 para as cinco estatísticas estão próximos do valor esperado 500. No entanto, para a estatística Rossi ocorreu um maior número de falsos alarmes antecipados (Figura 5.4). Os métodos Rogerson e Yamada e Hohle apresentam valores de ARL_1 menores para mudanças δ inferiores a $1,5$. Porém, o método Hohle necessitou em média duas observações a menos para sinalizar uma mudança na média de $1,25$. Para valores maiores que $\delta = 1,5$ as cinco estatísticas detectam com igual rapidez esse tipo de mudança.

Tabela 5.7: *Valores de h para um k que varia ao longo do tempo - $ARL_0 = 500$*

Estatística	h
Rossi	1,13
Jorgensen	1,08
Anscombe	0,98

5.3 Avaliação de suposições de normalidade e independência das estatísticas

Para avaliar a normalidade das transformações propostas por Rossi *et al.* (1999) e Jørgensen (1996) e dos Resíduos de Anscombe foram realizados 3 tipos de testes: Kolmogorov-Smirnov(Lilliefors) (Thode , 2002), Shapiro-Wilk (Royston , 1982) e Jarque-Bera (Cromwell *et al.* , 1994) e para avaliar a independência destas transformações se implementou o teste de Ljung-Box para as primeiras 9 defasagens. Por meio deste último teste avaliaremos se as primeiras 9 autocorrelações são nulas (Box e Pierce , 1970) . Foram simuladas 30.000 vezes a série de interesse no período de janeiro 2011 a dezembro 2011 e posteriormente, para cada série simulada se aplicaram os três testes de normalidade e o teste de Ljung-Box. A Tabela 5.8 mostra os resultados obtidos nos testes de normalidade.

Tabela 5.8: *Testes de Normalidade para as séries simuladas*

	Kolmogorov-Smirnov	Shapiro-Wilk	Jarque-Bera
Rossi	83,3%	75,1%	74,1%
Jorgensen	88,7%	86,4%	89,3%
R. Anscombe	72,9%	68,2%	85,6%

As Tabelas 5.8 e 5.9 apresentam a porcentagem de $valores - p$ que foram maiores ou iguais a $0,05$.

Tabela 5.9: *Teste de LjungBox para as séries simuladas*

	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5	Lag 6	Lag 7	Lag 8	Lag 9
Rossi	95.1%	94.9%	95.0%	94.8%	94.8%	94.7%	95.0%	94.6%	94.5%
Jorgensen	95.1%	95.2%	95.0%	95.1%	94.6%	95.3%	94.8%	94.7%	94.5%
R. Anscombe	95.0%	94.1%	94.1%	94.2%	94.3%	94.2%	94.4%	94.1%	94.1%

Para as três transformações normalizadoras comprovou-se via simulação que não vale a suposições de normalidade dos dados. Já que sob normalidade esperaríamos que a porcentagem de teste que rejeitaram a normalidade estivesse em torno de 95%. Porém, as séries transformadas são não correlacionadas como se mostra na Tabela 5.9.

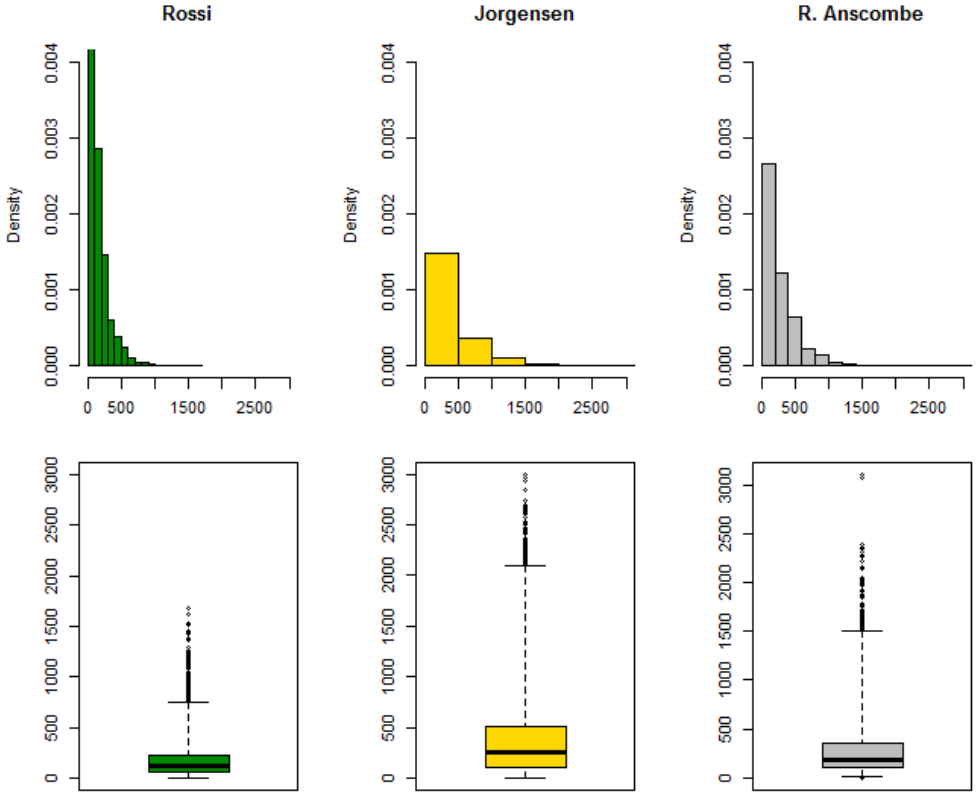


Figura 5.1: *Boxplots do RL- considerando as distribuições assintoticamente normais*

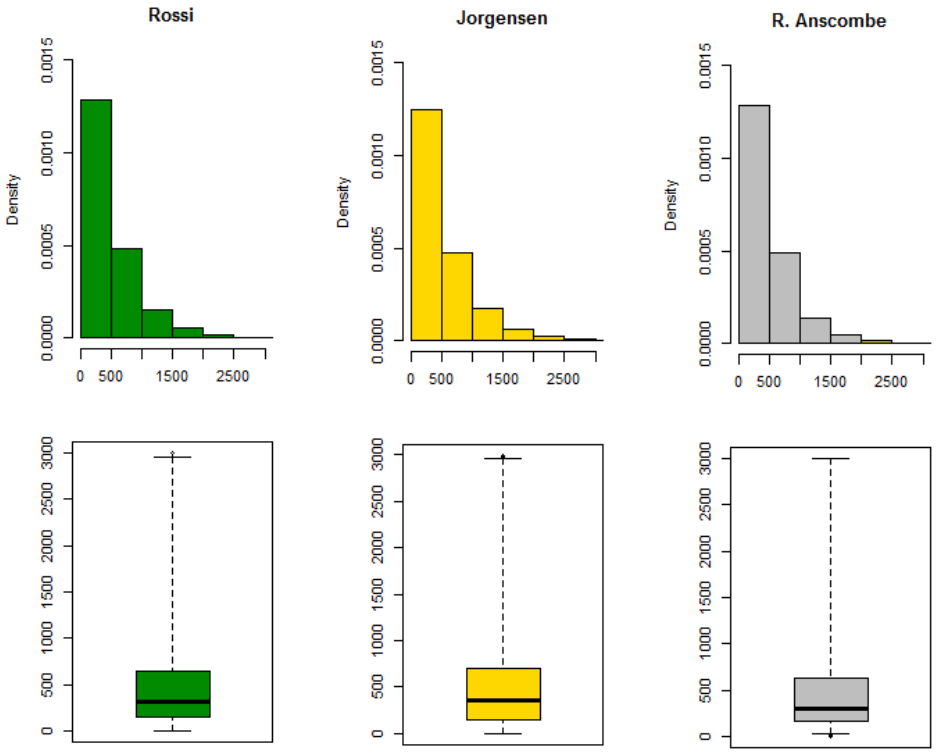


Figura 5.2: *Boxplots do RL- modificando o limiar L para obter* $ARL_0 = 500$

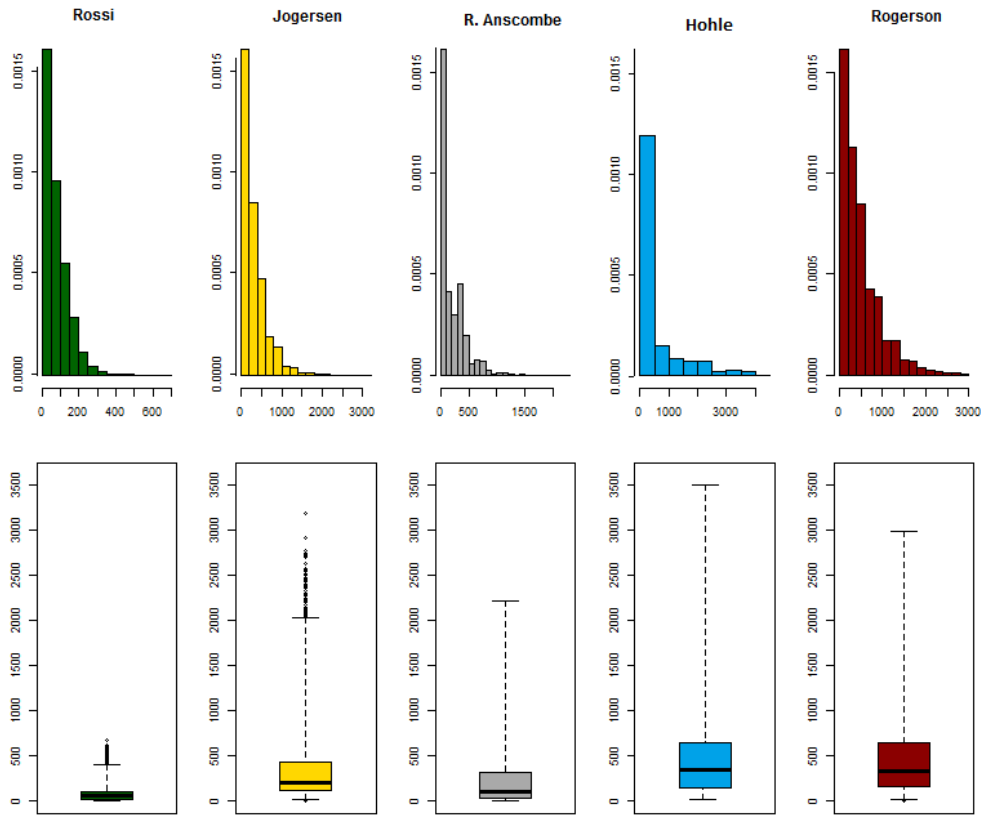


Figura 5.3: *Boxplots do RL- Gráfico CUSUM considerando as distribuições assintoticamente normais*

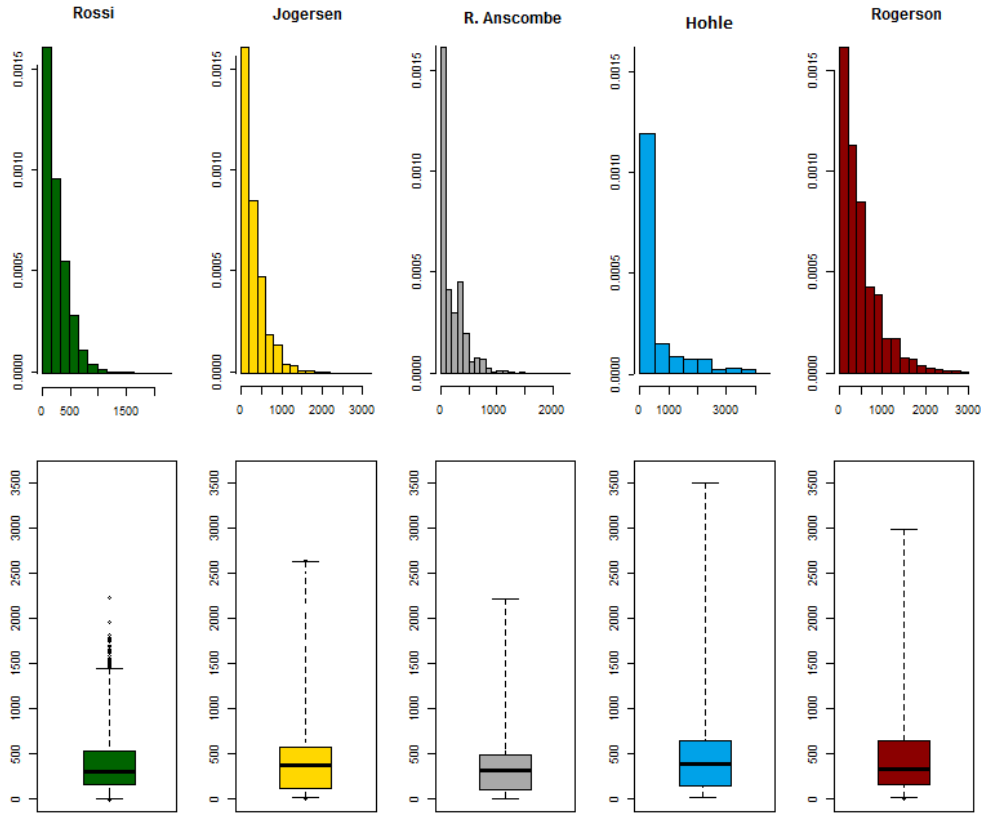


Figura 5.4: *Boxplots do RL- modificando o limiar h e valor de referência k para obter $ARL_0 = 500$*

Capítulo 6

Aplicação em dados reais

Neste capítulo, será monitorada, por meio dos gráficos de controle tipo Shewhart e CUSUM, a série temporal de contagem do número diário de internações por causas respiratórias de pessoas com 65 anos de idade ou mais residentes no município de São Paulo em 2011. Serão implementadas as cinco estatísticas de monitoramento apresentadas no capítulo anterior e avaliadas as suposições de normalidade e independência das estatísticas normalizadoras. Os dados diários de internações foram obtidos no Sistema de Informação Hospitalar (SIH) da Secretaria municipal de Saúde de São Paulo (PRO-AIM).

Análise Descritiva

O gráfico da série do número diário de internações por causas respiratórias é apresentado na Figura 6.1.

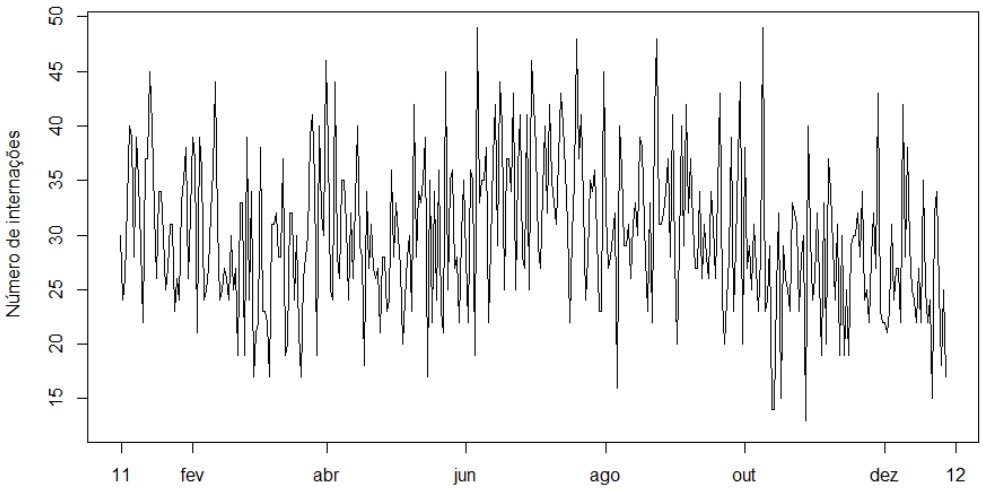


Figura 6.1: *Número diário de internações por causas respiratórias em 2011*

O número médio de internações diária e desvio padrão da série foram 30 e 7,04 respectivamente. Nos dias 11 de outubro(terça) e 30 de outubro (domingo) ocorreram o número máximo e mínimo de internações por causas respiratórias no município de São Paulo no ano 2011 com 49 e 13 internações, respectivamente. Na Tabela 6.1 apresenta-se um resumo da série.

Tabela 6.1: *Estatísticas do nº diário de internações em 2011*

Mínimo	1º Quantil	Mediana	Média	DP	3º Quantil	Máximo
13	25	29	30	7,04	35	49

Na Figura 6.2 é apresentado o gráfico boxplot do número de internações segundo o dia da semana. O número de internações no fim de semana é inferior em comparação ao número de internações de segunda a sexta, o número médio de internações (desvio padrão) para os dias sábado e domingo foram 27.85 (5.88) e 23.87 (5.88) respectivamente. No entanto, o número médio de internações, por exemplo, na segunda foi 32.58 (6.83). Um resumo da série segundo o dia da semana é apresentado na Tabela 6.2.

Tabela 6.2: *Estatísticas do nº diário de internações em 2011 segundo o dia da semana*

Categoria	Mínimo	Média	DP	Máximo
segunda	19	32.58	6.83	49
terça	20	31,08	5.79	49
quarta	15	30.87	6.25	48
quinta	18	32.52	6.61	48
sexta	19	31.19	5.75	42
sábado	14	27.85	5.88	44
domingo	19	23.87	5.88	37

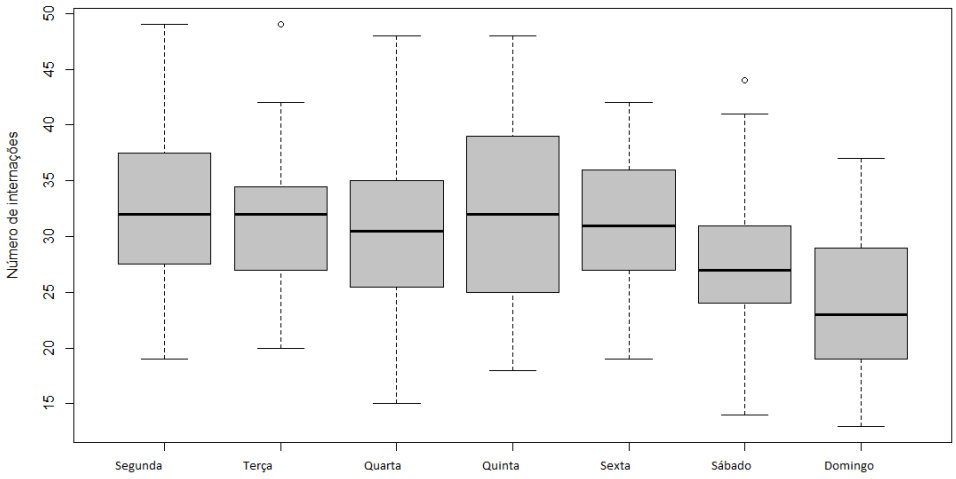


Figura 6.2: *Boxplot do número de internações segundo dia da semana*

Limite superior de confiança - Binomial Negativa

Uma primeira abordagem para monitorar esta série consistiu no ajuste de um modelo linear generalizado com distribuição binomial negativa. A parametrização utilizada é apresentada em Hawkins e Olwell (1998), com função de ligação logaritmo neperiano, modelando a média com as funções seno e cosseno para explicar a sazonalidade, incluindo a variável categórica dia da semana para controlar o efeito destes dias nos dados, e o logaritmo neperiano da população dividido por 100.000 como coeficiente constante (offset) para analisar a taxa de internações ao longo do tempo.

O modelo proposto com função de ligação logaritmo neperiano segue:

$$\ln(\mu_t) = \underset{\sim}{x}\underset{\sim}{\beta} + g_t \tag{6.1}$$

de modo que,

$$g_t = ln\,(pop_t/100000) \tag{6.2}$$

Assim, temos um modelo para a taxa esperada de internações

$$ln(t_x) = ln\left(\frac{\mu_t}{g_t}\right) = ln\left(\frac{\mu_t}{pop_t}100000\right) \tag{6.3}$$

O modelo binomial negativa ajustado para o período de janeiro 2006 a dezembro 2010 foi:

$$\begin{aligned} X_t &\sim BinNeg\left(\mu_t, \phi\right) \\ \ln(t_x) = \ln\left(\frac{\mu_t}{g_t}\right) &= \beta_0 + \beta_1 cos(2\pi t/365) + \beta_2 sin(2\pi t/365) + \beta_3 dom_t + \beta_4 seg_t + \\ &+ \beta_5 ter_t + \beta_6 qui_t + \beta_7 sex_t + \beta_8 sab_t \end{aligned} \tag{6.4}$$

$t = 1, ...,$ sendo t_x a taxa esperada de internações e dom_t e sab_t as variáveis de seg_t a sab_t são indicadoras de cada dia da semana. As estimativas do modelo são apresentadas na Tabela 6.3:

Tabela 6.3: *Estimativas do Modelo Generalizado com distribuição Binomial Negativa*

	Estimativa	Desvio Padrão	Valor p
Intercepto	1,214	0,028	$<$ 0,001
Cosseno	-0,080	0,015	$<$ 0,001
Seno	0,012	0,015	0,430
Domingo	-0,257	0,042	$<$ 0,001
Segunda	0,054	0,040	0,172
Terça	0,007	0,040	0,855
Quinta	0,053	0,040	0,186
Sexta	0,010	0,040	0,802
Sábado	-0,101	0,041	0,013

O limite superior de confiança de 95% para o número de internações no ano 2011 foi calculado baseado no ajuste deste modelo. Na Figura 6.3 é apresentada a série do número de internações, o valor ajustado (vermelho) e o respectivo limiar superior de confiança de 95% (verde). Os meses de janeiro, julho e setembro apresentaram incrementos significativos no número de internações, pois consideramos um incremento significativo aqueles valores que estão acima do limiar superior de confiança. Na Figura 6.4 mostra-se o incremento percentual da série com respeito à média sob controle μ_0 que foi calculada baseada no passado histórico da série no período de janeiro de 2006 a dezembro de 2010 no Capítulo 4. Os dias 3 e 31 de janeiro, 19 de março, 4 de abril e 13 de dezembro tiveram incrementos no número de internações maiores de 75% respeito a μ_0. Porém, só a mudança no mês de janeiro foi identificada na abordagem do intervalo de confiança para a distribuição binomial negativa.

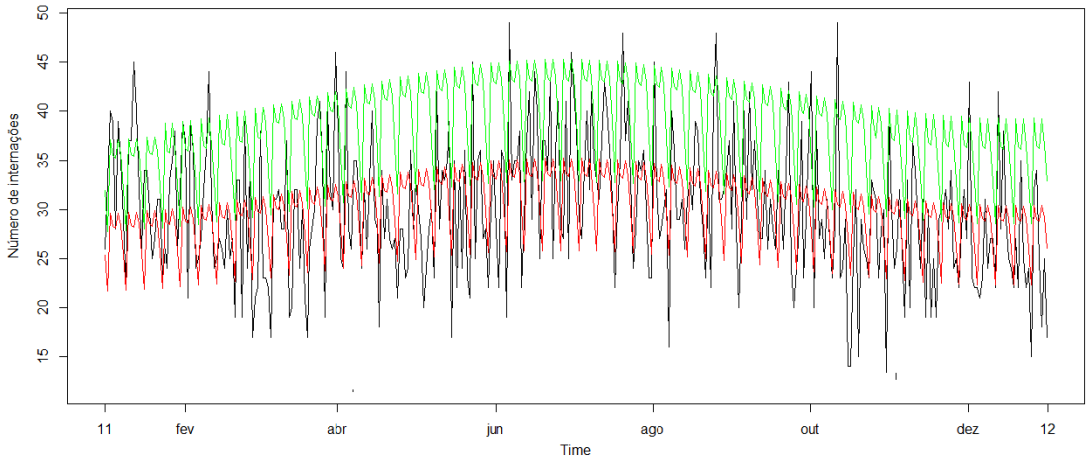


Figura 6.3: *Número de internações, número médio ajustado e Limiar Superior de Confiança de 95% - Binomial Negativa*

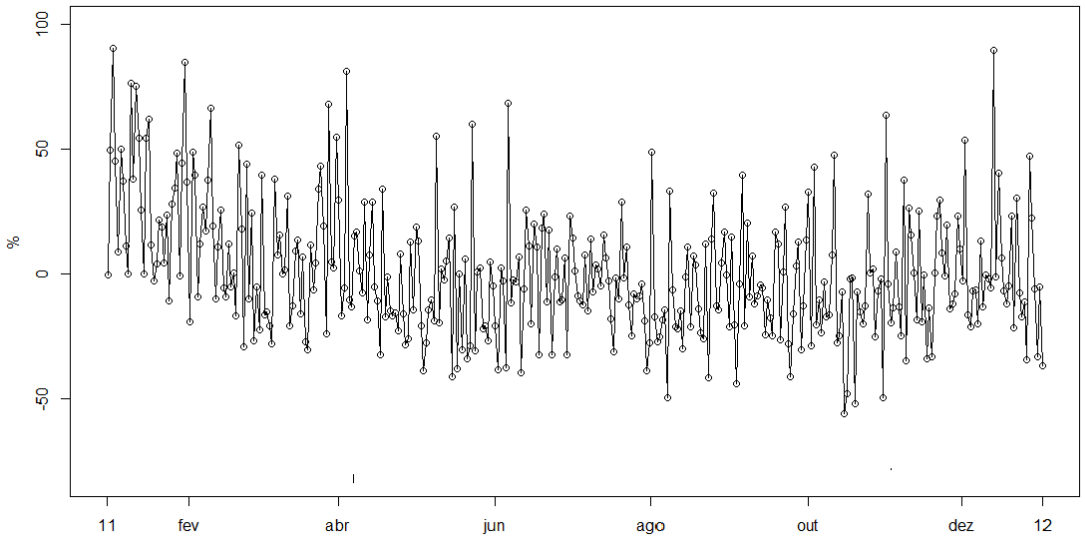


Figura 6.4: *Incremento percentual no número de internações com respeito a* μ_0

Gráficos de controle tipo Shewhart

Uma abordagem utilizada no Controle Estatístico de Processos para monitorar dados não normais consiste em normalizá-los mediante transformações para posteriormente analisá-los por meio de gráficos de controle para variáveis com distribuição normal. Inicialmente, foi implementada esta abordagem para monitorar a série do número de internações por meio do gráfico de controle tipo Shewhart por ser simples e de fácil interpretação. Foram implementadas as três estatísticas normalizadoras apresentadas no Capítulo 3: Rossi, Jorgensen e os resíduos de Anscombe para dados com distribuição binomial negativa. Os limiares de controle utilizados foram obtidos, via simulação, e foram usados os valores de ARL_0 próximos de 500 (Tabela 5.7). Os gráficos de Shewhart para as três estatísticas são apresentados nas Figuras 6.5, 6.6 e 6.7.

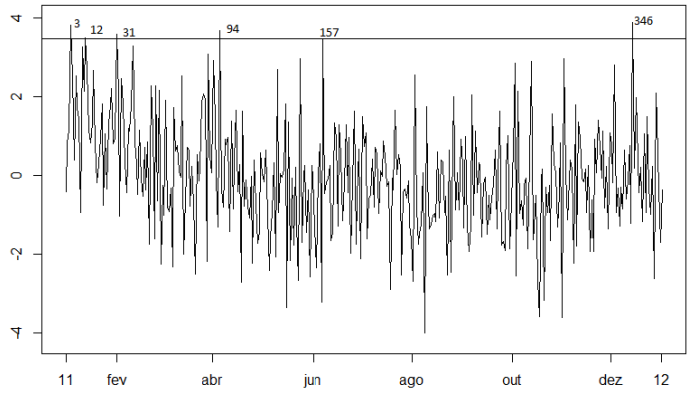


Figura 6.5: *Gráficos de Controle tipo Shewhart- Rossi*

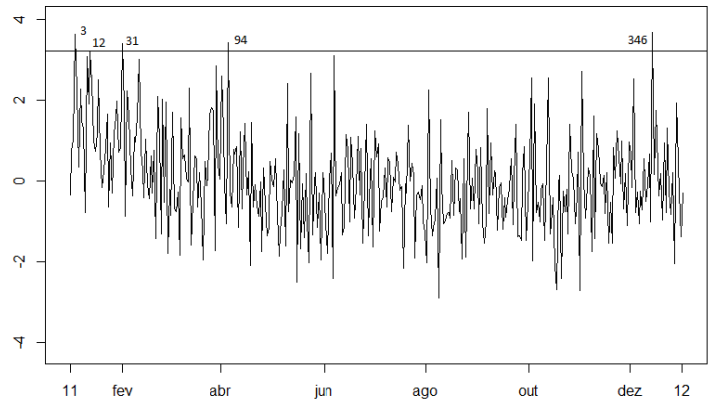


Figura 6.6: *Gráficos de Controle tipo Shewhart - Jorgensen*

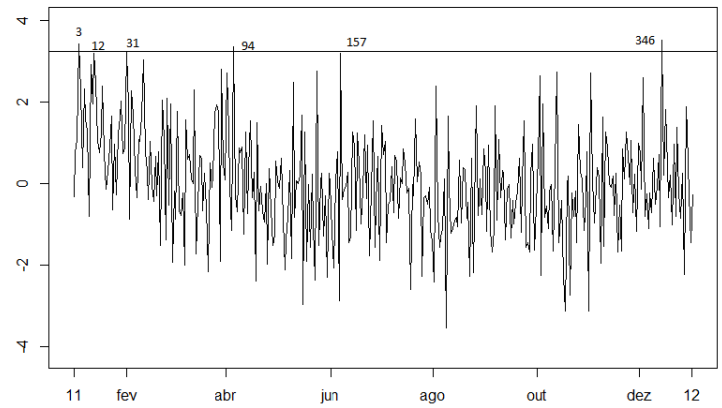


Figura 6.7: *Gráficos de Controle tipo Shewhart - R. Anscombe*

Os três gráficos, simultaneamente, detectaram os dias 3, 12 e 31 de janeiro, 4 de abril, 6 de junho e 19 de dezembro como epidêmicos, esses dias tiveram incrementos maiores que 75% no número de internações com respeito à média sob controle μ_0, portanto esperava-se que fossem detectados por meio destes gráficos de controle já que são poderosos na detecção de grandes mudanças (Montgomery , 2009). O gráfico de Shewhart para a estatística Rossi detecta também os dias 10 de fevereiro

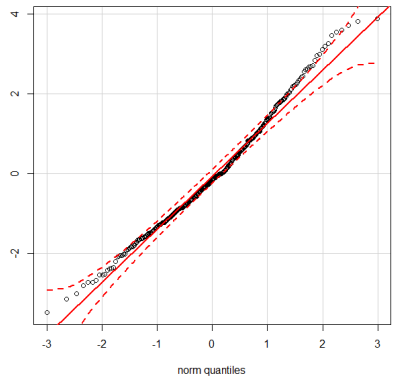
e 26 de março como epidêmicos.

As suposições de normalidade das três estatísticas foram avaliadas por meio dos testes Kolmogorov-Smirnov, Shapiro-Wilk e Jarque- Bera e igualmente por meio dos gráficos quantil-quantil com intervalo de confiança de 95% apresentados na Figura 6.8. A independência destas estatísticas foi avaliada mediante a função de autocorrelação (acf).

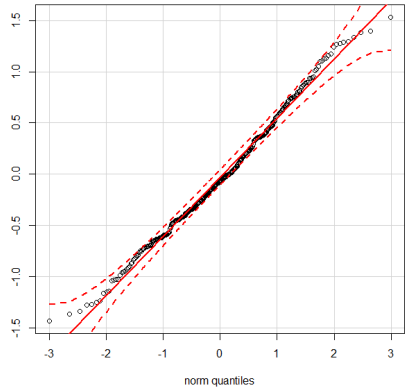
Tabela 6.4: *Nível descritivo os três testes de normalidade*

	Kolmogorov-Smirnov	Shapiro-Wilk	Jarque-Bera
Rossi	0,002	0,001	0,004
Jorgensen	0,033	0,054	0,143
R. Anscombe	0,002	0,001	0,003

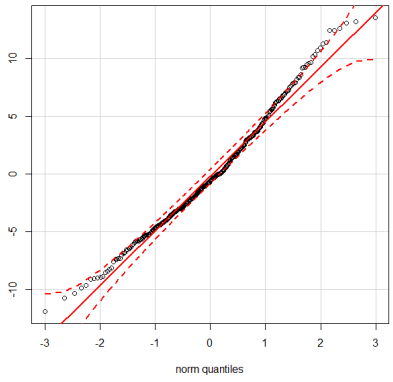
Para as estatísticas Rossi e os Resíduos de Anscombe os três testes rejeitaram a hipótese nula de normalidade. No entanto, para a estatística Jorgensen os testes de Shapiro-Wilks e Jarque-Bera não rejeitaram a hipótese nula.



(a) *quantil-quantil, Rossi*



(b) *quantil-quantil, Jorgensen*



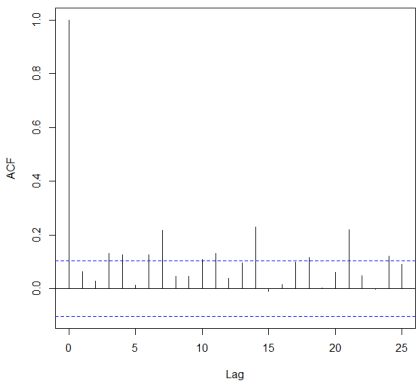
(c) *quantil-quantil, R. Anscombe*

Figura 6.8: *Gráficos quantil-quantil para as Estatísticas Normalizadoras*

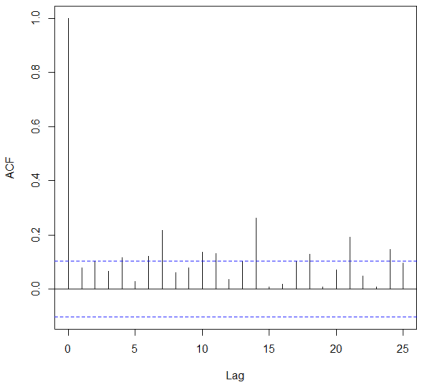
Entretanto o teste Jarque-Bera é pouco poderoso já que é baseado somente na verificação dos desvios de assimetria e curtose. Esse teste compara apenas o terceiro e o quarto momentos da distribuição normal teórica com os mesmos momentos estimados dos dados. Grandes desvios dos momentos amostrais com relação a seus valores esperados são considerados como um indicativo de não-normalidade. Esse procedimento não garante a normalidade no caso de não-rejeição, pois a igualdade dos quatro primeiros momentos é uma condição necessária, mas não suficiente. Já o teste

de Shapiro-Wilk compara uma estimativa do desvio padrão usando combinação linear de estatísticas de ordem com a estimativa usual. Esse último teste apresenta um alto desempenho de poder (Ferreira, 2008). O gráfico quantil quantil para esta estatística apresentou um bom comportamento, conforme Figura 6.8b.

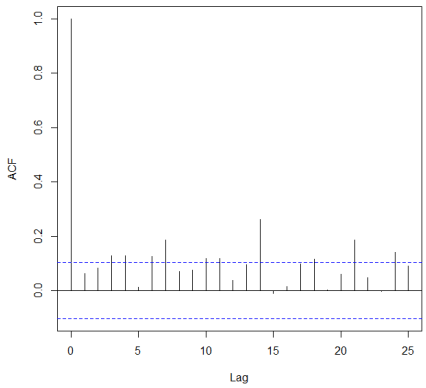
Os gráficos das funções de autocorrelação para as três transformações apresentaram autocorrelação significativa nas defasagens 7 e 14 apesar de ter incluído o dia da semana no modelo Binomial Negativa Generalizado utilizado para a previsão da média sob controle μ_0 para levar em conta que nos finais de semana o número de atendimentos hospitalares é menor do que nos dias de semana.



(a) *Acf-Rossi*



(b) *Acf-Jorgensen*



(c) *Acf-R. Anscombe*

Figura 6.9: *Gráficos acf para as estatísticas normalizadoras*

Gráficos de Controle tipo CUSUM

A série do número de internações por causas respiratórias foi monitorada novamente por meio do gráfico de controle tipo CUSUM implementando as três estatísticas normalizadoras com parâmetros k e h que, supondo normalidade, forneceria um $ARL_0 = 500$. No Capítulo 5 mostrou-se via simulação que utilizando esses parâmetros obtêm-se valores de ARL_0 menores que 500. Neste capítulo, nosso interesse é olhar as consequências de implementar esses parâmetros nos gráficos CUSUM e comparar estes resultados com os gráficos CUSUM com limiar h e valor de referência k para os quais, via simulação, obtiveram-se valores próximos a 500.

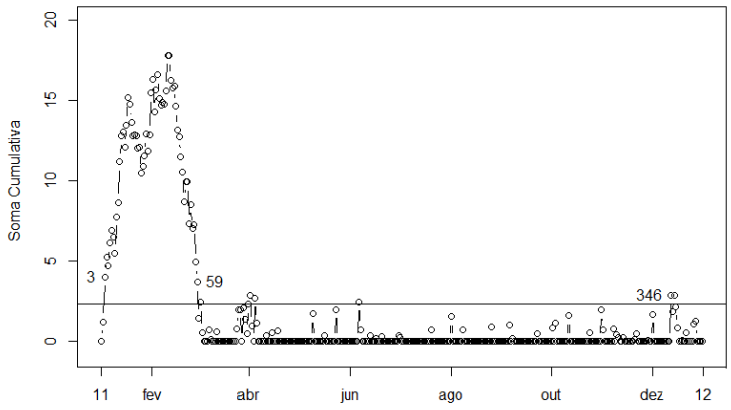


Figura 6.10: *Gráficos CUSUM considerando a distribuição Rossi assintoticamente normal*

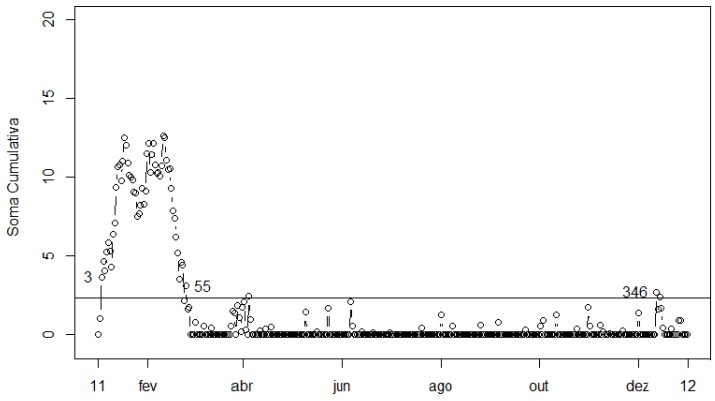


Figura 6.11: *Gráficos CUSUM considerando a distribuição Jorgensen assintoticamente normal*

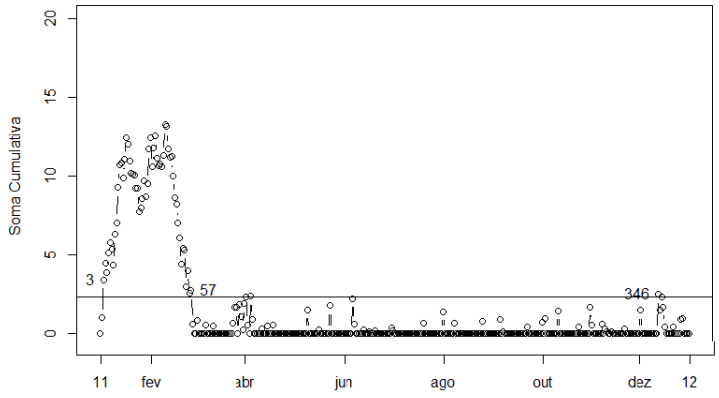


Figura 6.12: *ráficos CUSUM considerando a distribuição R. Anscombe assintoticamente normal*

Como resultado de utilizar estes parâmetros e implementar as estatísticas Rossi, Jorgensen e R. Anscombe no monitoramento da série, no período de janeiro de 2011 a dezembro de 2011, detectou-se para as três estatísticas uma epidemia na série no mês de Janeiro, começando no dia 3 e

finalizando no dia 25 de fevereiro. No entanto, para a estatística Rossi a série volta a estar sob controle no dia 1 de março. Esses gráficos também detectaram o dia 13 de dezembro como dia epidêmico.

Posteriormente, implementando os parâmetros h e k para os quais via simulação se obtiveram valores de ARL_0 próximos de 500 monitorou-se novamente a série utilizando as cinco estatísticas mencionadas no Capítulo 3.

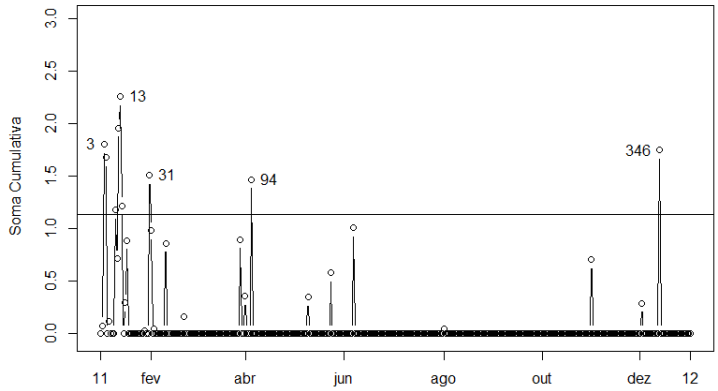


Figura 6.13: *Gráficos de Controle tipo CUSUM - Rossi,* $ARL_0 = 500$

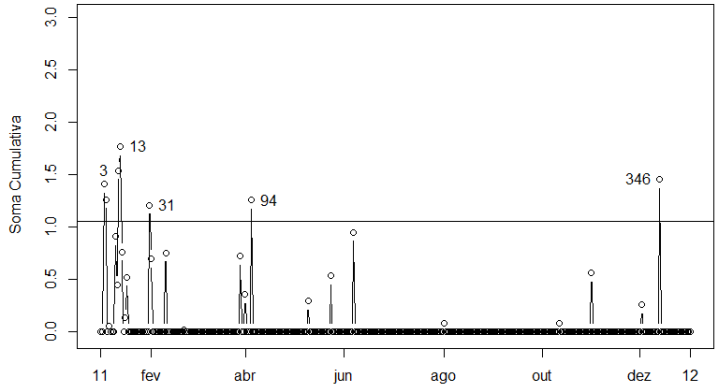


Figura 6.14: *Gráficos de Controle tipo CUSUM - Jorgensen,* $ARL_0 = 500$

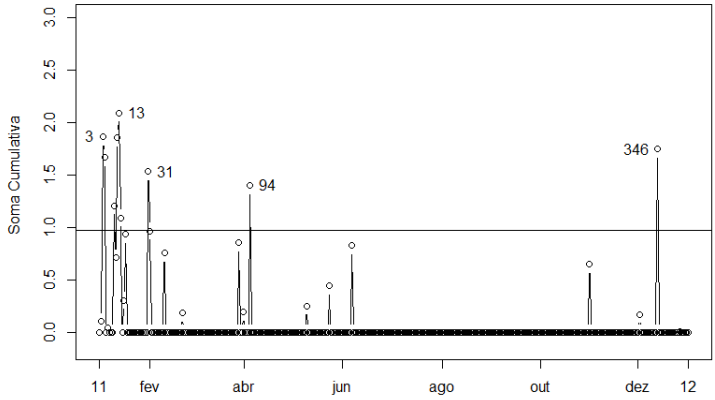


Figura 6.15: *Gráficos de Controle tipo CUSUM - R. Anscombe,* $ARL_0 = 500$

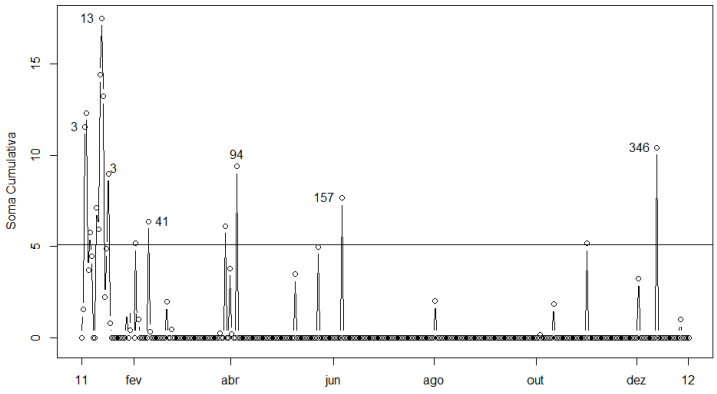


Figura 6.16: *Gráficos de Controle tipo CUSUM - Rogerson-Yamada,* $ARL_0 = 500$

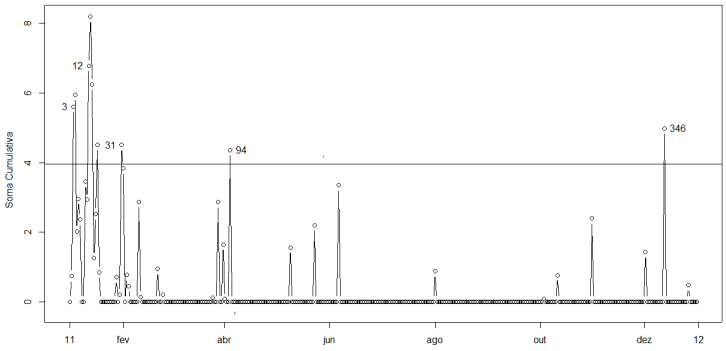


Figura 6.17: *Gráficos de Controle tipo CUSUM - Hohle-Lr,* $ARL_0 = 500$

Implementando estes parâmetros não se detectou o mês de janeiro como mês epidêmico porém, os dias 3, 12 e 31 de janeiro estiveram fora controle. Há muitos dias no inicio do ano sinalizados como epidemia, isto ocorre porque este tipo de gráfico precisa de um histórico de observações que estejam sob controle. As cinco estatísticas detectam como nos gráficos de Shewhart, os dias 4 de abril e 19 de dezembro como dias epidêmicos, pois estes dias tiveram um incremento percentual no número de internações respeito a μ_0 (média sob controle) de 75%. Para a estatística proposta

por Rogerson e Yamada (2004) o dia 30 de março também foi sinalizado como dia epidêmico. Nas Figuras 6.13 até 6.17 são apresentados os gráficos CUSUM para estas estatísticas.

Monitoramento da série ajustando um Modelo GARMA

Neste trabalho é proposta uma abordagem, para o exemplo numérico, baseada no ajuste de um modelo GARMA, uma combinação do modelo ARMA e do modelo linear generalizado que considera além da distribuição Normal, que é utilizada no modelo ARMA, outras distribuições da família exponencial (Benjamin *et al.* , 2003). Deste modo, será possível a inclusão de termos autorregressivos e de médias móveis em um modelo generalizado.

Avaliando as suposições de normalidade e independência dos resíduos estes serão monitorados por meio dos gráficos de controle CUSUM e Shewhart com limites de controle para dados com distribuição Normal. Esta abordagem já foi usada por Kovářík e Klímek (2012) para monitorar dados correlacionados. Eles depois de avaliar as suposições de normalidade e independência propõem monitorar, os resíduos de um modelo ARIMA por meio de gráficos de controle CUSUM e EWMA. Montgomery (2009) também propõe esta abordagem para monitorar dados correlacionados. Com esta nova abordagem, se procurará modelar a correlação persistente no $lag7$ na série de contagem do número de internações por causas respiratórias no ano 2011

Ajustou-se un modelo GARMA(1,0) com distribuição binomial negativa para a série do número de internações por causas respiratórias no período de janeiro de 2011 a dezembro de 2011. Incluindo as funções seno e cosseno no modelo para explicar a sazonalidade das internações como é proposto em Serfling (1963), a variável categórica dia da semana (dsem) para controlar o efeito destes dias nas internações e um termo autorregressivo para modelar a correlação existente na série. A variável população foi incluída como o logaritmo neperiano do tamanho da população dividido por 100.000 com coeficiente igual a 1 utilizando a função offset.

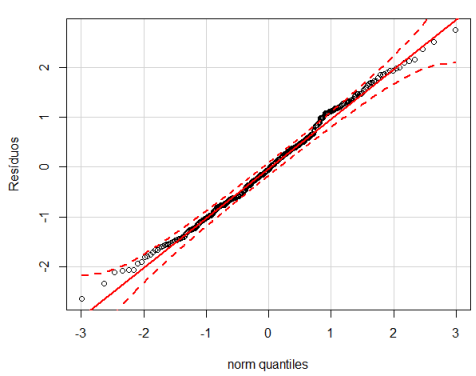
Os resíduos do modelo GARMA (1,0) mostraram um bom comportamento como se vê na Figura 6.18 que apresenta o gráfico normal quantil quantil para os resíduos. Além disso, não há indícios de correlação nos resíduos como se vê no gráfico da função de autocorrelação Figura (6.18b). Com um p-valor de $0,46$ não se rejeitou a hipótese nula de normalidade para o teste de Shapiro-Wilk. O parâmetro de autocorrelação estimado *phi* foi de 0,046.

Tabela 6.5: *Estimativas do modelo GARMA (1,0)*

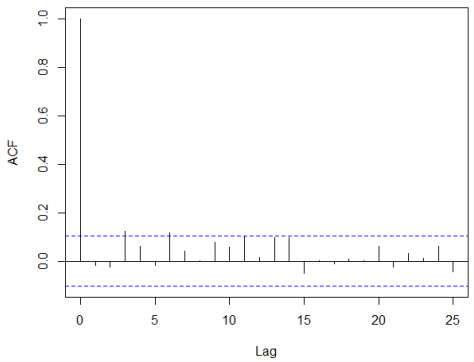
	Estimativa	Erro Padrão	Valor p
Intercepto	3,429	0,028	$<$ 0,001
Seno	0,003	0,016	0,838
Cosseno	-0,080	0,016	0,006
Domingo	-0,257	0,042	0,000
Segunda	0,055	0,039	0,167
Terça	0,007	0,039	0,852
Quinta	0,053	0,039	0,172
Sexta	0,010	0,040	0,793
Sábado	-0,102	0,041	0,012
ϕ	0,046	0,011	0,061
σ	0,009	0,003	0,004

Utilizando os parâmetros h e k que para dados com distribuição normal produziriam um valor esperado de ARL_0 igual a 500, os resíduos foram monitorados por meio dos gráficos de controle tipo CUSUM e para $L = 2,88$, que igualmente produziria um valor de ARL_0 igual a 500 os resíduos

foram monitorados mediante os gráficos tipo Shewhart. Nas Figuras 6.20 e 6.19 são apresentados estes gráficos respectivamente.



(a) *Gráfico quantil quantil dos Resíduos do modelo GARMA (1,0)*



(b) *Função d Autocorrelação dos resíduos do modelo GARMA (1,0)*

Figura 6.18: *Gráfico quantil quantil e autocorrelação dos Resíduos do modelo GARMA (1,0)*

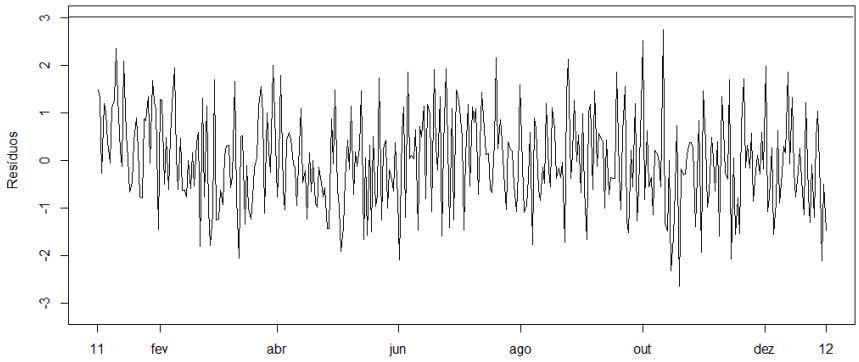


Figura 6.19: *Gráficos de Controle tipo Shewhart - Resíduos*

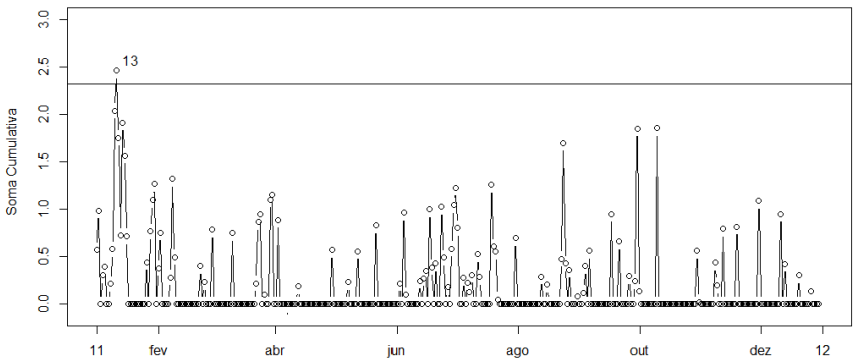


Figura 6.20: *Gráficos de Controle tipo CUSUM- Resíduos*

Implementando esta abordagem não se detectaram epidemias na série de interesse no período de janeiro de 2011 a dezembro de 2011. Em trabalhos futuros será profundada esta abordagem.

Capítulo 7

Conclusões

- O desempenho dos gráficos de controle tipo Shewhart e CUSUM para dados de contagem foi avaliado via simulação utilizando as transformações normalizadoras de Rossi *et al.* (1999) para dados com distribuição Poisson, a transformação proposta por Jørgensen (1996) para dados com distribuição Binomial negativa e os Resíduos de Anscombe (Pierce e Schafer , 1986) para modelos lineares generalizado. Calculando-se os limites de controle para tais transformações, é obtido o valor $ARL_0 = 500$ sob normalidade. Entretanto, utilizando-se esses limites de controle, observa-se um maior número de falso alarmes para as três estatísticas e, consequentemente, valores de ARL_0 menores a 500.

- Modificou-se o parâmetro k para os gráficos de controle CUSUM permitindo que k variasse ao longo do tempo como proposto por Rossi e Marchi (2010) para as três estatísticas normalizadoras Rossi, Jorgensen e os Resíduos de Anscombe. Essa modificação permitiu a obtenção de valores de ARL_0 próximos de 500.

- Avaliou-se as suposições de normalidade das estatísticas Rossi, Jorgensen e Resíduos de Anscombe, em geral omitida em trabalhos publicados na literatura, por meio dos testes Kolmogorov Smirnov, Shapiro-Wilk e Jarque Bera, e a independência por meio do teste Ljung Box, para os quais comprovou-se que as transformações não normalizam os dados, porém são independentes e estacionárias.

- Implementaram-se os gráficos CUSUM propostos por Rogerson e Yamada (2004)para dados Binomial Negativa, mas neste trabalho foi considerado variando ao longo do tempo um parâmetro k_t. Além disso, também foram analisados os gráficos CUSUM baseados na estatística da razão de verossimilhanças propostos por Hohle (2007) para distribuição Binomial Negativa. Esses métodos mostraram via simulação bons resultados para detectar mudanças na média, porém o método proposto por Hohle detectou mais rápido pequenas mudanças. É importante salientar que o método proposto por (Hohle , 2007) é muito mais simples de ser implementado do que o de Rogerson e Yamada (2004), já nesse método é necessário encontrar o valor de h para cada valor de k_t fixado.

- Verificou-se que os gráficos de controle tipo Shewhart apesar de simples e de fácil interpretação não são eficazes para detectar pequenas mudanças na média.

- Todas as estatísticas utilizadas apresentaram autocorrelação significativa no lag 7, apesar da inclusão do dia da semana no modelo linear generalizado implementado para calcular o valor alvo (μ_0).

- Foi proposta uma abordagem de monitoramento de séries de contagem baseada no ajuste do modelo GARMA mostrando bons resultados já que os resíduos foram normais e independentes. Implementando essa abordagem não se encontraram epidemias na série do número diário de internações por causas respiratórias para pessoas com 65 anos ou mais no município de São Paulo.

- A implementação do método de Rogerson e Yamada poderia deixar pronto para uma secretaria de vigilância quais são as bandas de confiança a partir dos quais os valores de internações podem identificar uma epidemia. Entretanto, o método da razão de verossimilhanças apresentou melhores resultados e deveria ser adotado. Dando continuidade a esse trabalho, serão desenvolvidos métodos prospectivos de controle de qualidade para dados de contagem, pois faltam até mesmo resultados para a previsão usando modelos de contagem.

Referências Bibliográficas

Benjamin *et al.* (2003) Michael A Benjamin, Robert A Rigby e D Mikis Stasinopoulos. Generalized autoregressive moving average models. *Journal of the American Statistical Association*, 98 (461):214–223. Citado na pág. 45

Box e Pierce (1970) George EP Box e David A Pierce. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association*, 65(332):1509–1526. Citado na pág. 32

Brook e Evans (1975) D. Brook e D. A. Evans. An approach to the probability distribution of cusum run length. *Biometrika*, 59:539–548. Citado na pág. 12

Costa e Carpinetti (2004) E.K Costa, A.F.B. Epprecht e L.C.R. Carpinetti. *Controle Estatístico de Qualidade*. São Paulo Atlhas, 334p, 2th ed. Citado na pág. 5

Cromwell *et al.* (1994) Jeff B Cromwell, Walter C Labys e Michel Terraza. *Univariate tests for time series models*, volume 99. Sage. Citado na pág. 32

Frisén e Sonesson (2005) M. Frisén e C. Sonesson. *Spatial and Syndromic Surveillance for Public Health*. 2th ed. Citado na pág. 19

Hawkins e Olwell (1998) D. M. Hawkins e D. H. Olwell. *Cumulative Sum Charts and Charting for Quality Improvement*. Springer, New York., 2th ed. Citado na pág. 10, 12, 36

Hawkins (1981) D.M. Hawkins. A CUSUM for a scale parameter. *Journal of Quality Technology*, 13:228–213. Citado na pág. 2

Hawkins (1992) D.M. Hawkins. Evaluation of Averange Run Length of Cumulative Sum Charts for an Arbitrary Data Distribution. *Communications in Statistics -Simulation and Computation*, 21:1001–1020. Citado na pág. 12, 14

Hohle (2007) M Hohle. Surveillance: An R packge for the surveillance of infectious diseases. *Computational Statistics*, 22:571–582. Citado na pág. v, 15, 19, 28, 47

Hohle e Mazick (2009) M. Hohle e A. Mazick. Aberration detection in r illustrated by danish mortality monitoring. *Biosurveillance: A Health Protection Priority*. Citado na pág. 3

Hohle e Mazick (2010) M. Hohle e A. Mazick. Aberration detection in R ilustrated by danish mortality monitoring. *Biosurveillance: Methods and Casa Studies*, páginas 215–237. Citado na pág. 2

Hohle e Paul (2008) M Hohle e M. Paul. Count data regression chart for the monitoring of surveillance time series. *Computational Statistics and Data Analyisis*, 52:4357–4368. Citado na pág. 2, 3, 26

Jørgensen (1996) B Jørgensen. *The Theory of Dispersion Models*. Chapman and Hall, London, 2th ed. Citado na pág. 15, 16, 17, 32, 47

Kovářík e Klímek (2012) Martin Kovářík e Petr Klímek. The usage of time series control charts for financial process analysis. *Journal of Competitiveness*. Citado na pág. 45

Lai (1995) T Lai. Sequential change point detection in quality control and dynamical systems. *Journal of the Royal Statistical Society*, Series B 57:613–658. Citado na pág. 2

Lucas (1985) J.M. Lucas. Counted data CUSUM's. *Technometrics*, 27:129–144. Citado na pág. 2, 19

McCullagh e Nelder. (1989) P. McCullagh e J. A. Nelder. *Generalized Linear Models*. Chapman Hall/CRC, London, 2th ed. Citado na pág. 17

McCulloch e Searle (2001) C. E. McCulloch e S. R. Searle. *Linear and Generalized Linear Mixed Models*. Wiley, New York, 3th ed. Citado na pág. 17

Montgomery (2009) D. C Montgomery. *Introduction to Statistical Process Control*. John Wiley & Sons,Inc, Hoboken, NJ, 6th ed. Citado na pág. ix, 1, 2, 6, 7, 8, 9, 11, 12, 39, 45

Pierce e Schafer (1986) Donald A Pierce e Daniel W Schafer. Residuals in generalized linear models. *Journal of the American Statistical Association*, 81(396):977–986. Citado na pág. 47

Rogerson e Yamada (2004) P. A. Rogerson e I. Yamada. Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report*, 53/Supplement:79–85. Citado na pág. iii, v, 2, 3, 4, 15, 18, 25, 26, 45, 47

Rossi *et al.* (1999) G. Rossi, l. Lampugnani e M. Marchi. An aproximate CUSUM procedure for surveillance of health events. *Statistics in Medicine*, 18:2111–2122. Citado na pág. 3, 15, 16, 28, 32, 47

Rossi e Marchi (2010) S.; Rossi, G.; Del Sarto e M Marchi. Approximate poisson cusum charts for the monitoring of time series with time-varying mean. *Proceedings of the 45th Scientific Meeting of the Italian Statistical Society*, página Padova. Citado na pág. 3, 16, 26, 29, 31, 47

Royston (1982) Patrick Royston. Algorithm as 181: The w test for normality. *Public health reports*, 31:176–180. Citado na pág. 32

Serfling (1963) Robert E Serfling. Methods for current statistical analysis of excess pneumonia-influenza deaths. *Public health reports*, 78(6):494. Citado na pág. 22, 45

Shmueli e Burkom (2010) Galit Shmueli e Howard Burkom. Statistical challenges facing early outbreak detection in biosurveillance. *Technometrics*, 52(1). Citado na pág. 2

Siegmund (1985) D. Siegmund. *Sequential analysis Test and Confidence intervals*. New York: SpringerVerlage, 3th ed. Citado na pág. 12

Skinner e Runger (2003) D. Skinner, K. Montogomery e G. Runger. Process monitoring for multiple count data using generalized linear model-based control charts. *International Journal of Profuction Research 41*, 6:1167–1180. Citado na pág. 2

Thode (2002) Henry C Thode. *Testing for normality*, volume 164. CRC Press. Citado na pág. 32

Unkel *et al.* (2012) Steffen Unkel, C Farrington, Paul H Garthwaite, Chris Robertson e Nick Andrews. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82. Citado na pág. 2, 21

Wimmer e Höhle (2013) Valentin Wimmer e Michael Höhle. The functionalgo. glrnbin the r-packagesurveillance. Citado na pág. 25, 27

Woodall (2006) W. H Woodall. The use of control charts in health-care and public health surveillance. *J. Qual. Technol*, 38:88–103. Citado na pág. 2

Woodall e Adams (1993) W. H. Woodall e B. M. Adams. The statistical design of cusum charts. *Quality Engineering*, 4:559–570. Citado na pág. 12

Woodall e Montgomery (1999) W. H. Woodall e D. C. Montgomery. Research issues and ideas in statistical process control. *Journal of Quality Technology*, 18:2111–2122. Citado na pág. 2

Índice Remissivo