UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA E ESTATÍSTICA

Dennison Célio de Oliveira Carvalho

Gráficos de Controle Bayesianos em Alguns Processos da Família Exponencial

Orientadora: Profa. Dra. Maria Regina Madruga Tavares

Belém
2009

Dennison Célio de Oliveira Carvalho

Gráficos de Controle Bayesianos em Alguns Processos da Família Exponencial

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Matemática e Estatística, da Universidade Federal do Pará, como requisito parcial para a obtenção do grau de Mestre em Matemática e Estatística.

Área de Concentração: Inferência Estatística Bayesiana
Orientadora: Profa. Dra. Maria Regina Madruga Tavares

Belém
2009

FICHA CATALOGRÁFICA

Carvalho, Dennison Célio de Oliveira, 1981 -

Gráficos de controle bayesianos em alguns processos da família exponencial. / Dennison Célio de Oliveira Carvalho; orientadora, Maria Regina Madruga Tavares. — 2009

Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Exatas e Naturais, Programa de Pós-Graduação em Matemática e Estatística, Belém, 2009.

1. Estatística matemática. 2. Controle de processo - Métodos estatísticos. I. Título.

CDD - 22. ed. 519.5

Dennison Célio de Oliveira Carvalho

Gráficos de Controle Bayesianos em Alguns Processos da Família Exponencial

Esta Dissertação foi julgada e aprovada, para a obtenção do grau de Mestre em Matemática e Estatística, no Programa de Pós-Graduação em Matemática e Estatística, da Universidade Federal do Pará.

Belém, 3 de julho de 2009

Prof. Dr. Mauro de Lima Santos
(Coordenador do Programa de Pós-Graduação em Matemática e Estatística - UFPA)

Banca Examinadora

Profa. Dra. Maria Regina Madruga Tavares
Universidade Federal do Pará
Orientadora

Prof. Dr. Joaquim Carlos Barbosa Queiroz
Universidade Federal do Pará
Examinador

Profa. Dra. Terezinha Ferreira de Oliveira
Universidade Federal do Pará
Examinadora

À Deus e minha família.

Agradecimentos

⋆ À Deus todo poderoso por ter me concedido a vida e ter me dado forças para finalizar este trabalho;

⋆ À minha família, em especial à minha mãe, Amélia Maria de Oliveira Carvalho, pelo amor, carinho, apoio e paciência e pelas orações não só neste, mas em todos os meus trabalhos;

⋆ À minha namorada Ana Garcêz, pelo amor, atenção, preocupação, orações e pela força que sempre me passou para que eu nunca desista diante das dificuldades;

⋆ Aos meu amados e verdadeiros amigos Edney Fernandes e Ronaldo Corrêa, por sempre estarem dispostos a me ajudar no que fosse preciso;

⋆ À minha orientadora Profa. Dra. Regina Tavares por ter aceitado me orientar em um difícil momento da minha vida acadêmica;

⋆ À Universidade Federal do Pará (UFPA) pela oportunidade de concluir um curso de pós-graduação;

⋆ À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro destinado à este trabalho;

⋆ Ao Programa de Pós-Graduação em Matemática e Estatística (PPGME), representado pelo Prof. Dr. Mauro Santos;

⋆ A todos os professores que contribuíram de alguma forma na minha formação;

⋆ E finalmente, a todos que de alguma forma contribuíram para a realização deste trabalho.

"Clássico ou Bayesiano,
seja suficiente, seja Estatístico".

Dennison Carvalho.

Resumo

CARVALHO, Dennison Célio de Oliveira. Gráficos de Controle Bayesianos em alguns Processos da Família Exponencial. Dissertação de Mestrado (Programa de Pós-Graduação em Matemática e Estatística-UFPA, Belém - PA, Brasil).

Apresenta-se neste trabalho os conceitos básicos de inferência estatística necessários à construção de gráficos de controle bayesianos, assim como a metodologia proposta por Menzefricke (2002) para construção dos limites de controle baseados em distribuições preditivas *a posteriori*. A partir desta metodologia obteve-se a distribuição preditiva em alguns processos da Família Exponencial, em particular Poisson e Exponencial, e como a distribuição preditiva encontrada não foi analiticamente tratável, foi necessário a utilização de métodos de simulação de Monte Carlo para quantidades preditivas, para obter os limites de controle. A partir destes métodos de simulação obteve-se, também, o comprimento médio da sequência, isto é, o número médio de amostras até que o gráfico sinalize que o processo está fora de controle pela primeira vez. Para a avaliação dos gráficos de controle propostos neste trabalho, assumiu-se hipoteticamente que houve uma mudança na média do processo de magnitude δ, analisando-se assim a probabilidade de que uma obervação qualquer caia na região de rejeição, ou seja, a probabilidade de que o gráfico sinalize, dado que houve esta mudança no processo. Os resultados obtidos na avaliação dos gráficos de controle para processos Poisson e Exponencial mostraram que, para amostras futuras ($\mathbf{y}$) de tamanho $n = 10$, ambos os gráficos apresentaram sensibilidade a pequenas mudanças no processo.

Palavras Chaves: Gráficos de Controle, Distribuição Preditiva e Métodos de Monte Carlo.

Abstract

CARVALHO, Dennison Célio de Oliveira. Bayesian Control Charts in some Processes of the Exponential Family. Master Dissertation (Post Graduation Program in Mathematics and Statistics - UFPA, Belém - PA, Brazil).

It is presented in this paper the basic concepts of statistical inference that are necessary to construct the Bayesian control charts as well as the methodology proposed by Menzefricke (2002) to the construction of the control limits based on posterior predictive distributions. From this methodology it was achieved the predictive distributions in some processes of the Exponential family, in particular Poisson and Exponential, and how the predictive distributions found was not analytic treated, it was necessary the use of methods of simulation of Monte Carlo to predictive distributions to obtain the control limits. From these methods of simulation it is also achieved the average run length that is the average number of samples until the chart shows that the process is out of control for the first time. For the evaluation of the control charts that were proposed in this paper, it was hypothetically assumed that there was a change in the average of the magnitude process δ, analyzing the probability of any observation going to the rejection region, in other words, the probability that the chart shows, because there was this change in the process. The results obtained in the evaluation of the control charts to the processes Poisson and Exponential show that, to future samples (y) of size $n = 10$, both charts present sensitivity to small changes in the process.

Key-words: Control Charts, Anticipating Distribution and Monte Carlo Methods.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

1.1 Aspectos gerais

A busca pela melhoria da qualidade em diversas áreas tem sido a meta de muitos órgãos, empresas, entidades públicas dentre outros, visto que, com a melhoria da qualidade pode-se, por exemplo, aumentar a produtividade acentuando a penetração no mercado com maior lucratividade e forte competitividade. Dentre os diversas métodos estatísticos para o controle e a melhoria da qualidade, destaca-se o controle estatístico de processos (CEP), que tem os gráficos de controle como principal e mais utilizada ferramenta para a melhoria da qualidade.

Walter A. Shewhart, dos *Bell Telephone Laboratories*, desenvolveu em 1924 o conceito estatístico de gráfico de controle, que é considerado, em geral, como o começo formal do controle estatístico da qualidade (Montgomery, 2001). Estes gráficos de controle são construídos marcando-se em ordenadas uma linha central (também chamada de linha média), uma linha inferior e uma linha superior de controle. Na abordagem estatística clássica estes limites são calculados com base na teoria de construção de intervalos de confiança (Casella e Berger, 2002).

A informação que se tem sobre uma quantidade de interesse θ é fundamental na Estatística (Ehlers, 2007). O verdadeiro valor de θ é desconhecido e a idéia é tentar reduzir este desconhecimento. Além disso, a intensidade da incerteza a respeito de θ pode assumir diferentes graus. Do ponto de vista bayesiano, estes diferentes graus de incerteza são representados através de modelos probabilísticos para θ, denominados distribuição *a priori*. Neste contexto, é natural que diferentes pesquisadores possam ter diferentes graus de incerteza sobre θ (especificando modelos distintos). Sendo assim, a construção de gráficos de controle bayesianos apresentam esta característica muito importante que é a capacidade de

incorporar formalmente o conhecimento que o pesquisador tem a respeito da quantidade de interesse θ.

Contudo, dada a importância da informação inicial que se tem sobre θ, Menzefricke (2002) propõe a construção de gráficos de controle para a média μ de um processo com distribuição Normal, assumindo as duas situações: desvio padrão conhecido e desconhecido. Na construção destes gráficos ele baseia-se na obtenção de uma região de credibilidade HPD (*highest posterior density*) para a observação futura do processo, utilizando sua distribuição preditiva *a posteriori*.

1.2 Justificativa e importância do trabalho

A qualidade tornou-se um dos mais importantes fatores de decisão dos consumidores na seleção de produtos e serviços que competem entre si. O fenômeno é geral, independente do fato de o consumidor ser um indivíduo, uma organização industrial, uma loja de varejo, ou um programa militar de defesa. Consequentemente, compreender e melhorar a qualidade é um fator-chave que conduz ao sucesso, ao crescimento e a uma melhor posição de competitividade de um negócio. A melhoria e o emprego bem-sucedido da qualidade como parte integrante da estratégia geral da empresa, produzem retorno substancial sobre o investimento (Montgomery, 2001).

A crise econômica internacional atual que atinge diversos países, inclusive o Brasil, reforça ainda mais a idéia da necessidade do uso de métodos de controle estatístico da qualidade (CEQ), visto que, em tempos de crise ou não, o desperdício de materiais, consumo excessivo de determinados produto, a produção em grandes escalas de produtos possivelmente defeituosos, em indústrias, empresas e no mercado econômico de um modo geral, contribui ainda mais para o agravamento da crise.

No CEQ, os métodos clássicos são amplamente utilizados pelas comunidades estatísticas e por diversas outras áreas, inclusive para construção de gráficos de controle. A construção dos gráficos de controle clássicos é feita a partir de subgrupos racionais, ou seja, são retiradas k amostras de tamanho n de uma determinada população, e posteriormente, calcula-se os limites de controle.

Mesmo quando a característica da qualidade* não tem distribuição normal, na abordagem clássica, costuma-se utilizar os gráficos de controle sob a suposição de normalidade, já que, por exemplo, Burr (1967) comenta que os limites de controle baseados na teoria normal são robustos com relação à hipótese de normalidade e podem ser empregados desde que a população não seja extremamente não-normal.

Na abordagem bayesiana, os gráficos de controle para a média também foram construídos sobre a suposição de que a distribuição da característica da qualidade é normalmente distribuída (Menzefricke, 2002). No entanto, a característica da qualidade pode apresentar outras distribuições de probabilidade. A distribuição de Poisson é bastante útil no CEQ, onde, uma aplicação típica desta distribuição é quando se deseja monitorar o número de defeitos por unidade de produto. Outra distribuição muito importante no CEQ é a distribuição Exponencial, que é amplamente utilizada na área de engenharia da confiabilidade como modelo do tempo de falha de um componente ou sistema (Montgomery, 2001).

Neste contexto, este projeto visa construir gráficos de controle bayesianos em alguns membros da Família Exponencial, em particular Poisson e Exponencial, visto que, em tese, estes gráficos seriam de extrema importância para situações em que a característica da qualidade não seja normalmente distribuída.

1.3 Objetivos

1.3.1 Objetivo geral

Desenvolver gráficos de controle bayesianos em alguns processos da Família Exponencial, baseados em distribuições preditivas.

1.3.2 Objetivos Específicos

Como objetivos específicos têm-se

- Apresentar os principais conceitos da teoria estatística necessária para a construção de gráficos de controle;

* A característica da qualidade pode ser uma variável ou atributo a partir do qual se deseja monitorar um processo. Ela é uma variável quando é medida em uma escala numérica e, um atributo quando classificada como conforme ou não conforme (Montgomery, 2001).

- Estender a teoria bayesiana proposta por Menzefricke (2002) para gráficos de controle, considerando processos não-normais, em particular, processos Poisson e Exponenciais;
- Implementar os gráficos bayesianos propostos a partir de métodos de simulação de Monte Carlo;
- Avaliar os resultados obtidos.

1.4 Estrutura do trabalho

Este trabalho encontra-se dividido em seis capítulos, a saber

- Capítulo 1: Refere-se à introdução do trabalho, onde estão contidos a justificativa e importância, objetivos geral e específicos;
- Capítulo 2: Faz-se uma breve introdução ao controle estatístico da qualidade e mostra-se dois dos principais gráficos de controle clássicos para processos normais e não-normais;
- Capítulo 3: Mostra-se alguns conceitos de inferência estatística de grande importância para a construção de gráficos de controle bayesianos;
- Capítulo 4: Apresenta-se a metodologia para construção dos gráficos de controle bayesianos para a média com desvio padrão conhecido e desconhecido em processos com distribuição normal;
- Capítulo 5: Apresenta-se a construção e avaliação dos gráficos de controle para processos Poisson e Exponenciais;
- Capítulo 6: São apresentadas as conclusões e recomendações para trabalhos futuros.

Capítulo 2

Gráficos de Controle Clássicos

2.1 Controle estatístico da qualidade (CEQ)

Para Deming (2000), qualidade significa atender e, se possível, exceder as expectativas do consumidor. Para Juran (1999), qualidade significa adequação ao uso. Já para Crosby (1995), qualidade significa atender às especificações. O CEP é um dos métodos mais utilizados para a melhoria da qualidade, onde a principal ferramenta é o gráfico de controle. Neste contexto, apresenta-se neste capítulo os conceitos necessários para a construção dos gráficos de controle para processos normais, quando a característica da qualidade é uma variável e, para processos não-normais, quando a característica da qualidade é um atributo.

O controle da qualidade de produtos é tão antigo quanto a própria indústria; durante muito tempo foi realizado sob a forma tradicional denominada "inspeção". Somente a partir de 1920 é que se desenvolveu o controle estatístico da qualidade, cuja aplicação vem crescendo em diversas áreas de interesse. Portanto, controle estatístico da qualidade é um sistema amplo e complexo; abrange todos os setores de uma empresa, em um esforço comum e cooperativo, tendo como objetivo estabelecer, melhorar e assegurar a qualidade de produção, em níveis econômicos, para satisfazer aos desejos dos consumidores (Filho, 1970).

Segundo Montgomery (2001), se um produto deve corresponder às exigências de um cliente, deve, em geral, ser produzido por um processo que seja estável ou replicável. Mais precisamente, o processo deve ser capaz de operar com pequena variabilidade em torno das dimensões-alvo ou nominais das características de qualidade do produto. O controle estatístico do processo (CEP) é uma poderosa coleção de ferramentas de resolução de problemas útil na obtenção da estabilidade do processo e na melhoria da capacidade através da redução da variabilidade.

Os gráficos de controle são construídos a partir da técnica de estimação por intervalos, os intervalos de confiança. A construção destes intervalos na abordagem clássica é feita utilizando-se quantidades pivotais. Uma variável aleatória $Q(\mathbf{X};\theta) = Q(X_1, ..., X_n;\theta)$ é uma quantidade pivotal se a distribuição de $Q(\mathbf{X};\theta)$ é independente de todos os parâmetros (Casella e Berger, 2002). No caso em que se tem uma amostra aleatória $\mathbf{X} = (X_1, ..., X_n)$ i.i.d. (independentes e identicamente distribuídas) de tamanho n da distribuição normal com média μ e variância σ^2 conhecido, encontra-se a partir da quantidade pivotal $(\bar{X} - \mu)/(\sigma/\sqrt{n})$, que tem distribuição $N(0, 1)$ (normal com média 0 e variância 1), o intervalo de confiança para μ, dado pela seguinte expressão:

$$\left\{\mu : \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} \tag{2.1}$$

onde, μ é o parâmetro desconhecido; $\bar{X}$ é a média amostral; σ é o desvio padrão; n é o tamanho da amostra e $z_{\alpha/2}$ é o valor tabelado da normal padrão.

No caso em que a variância σ^2 é desconhecida, utiliza-se S^2 que é o estimador não viesado para a variância σ^2 e a partir da quantidade pivotal $(\bar{X} - \mu)/(S/\sqrt{n})$, que tem distribuição *t-student* com $n - 1$ graus de liberdade, encontra-se o seguinte intervalo de confiança:

$$\left\{\mu : \bar{X} - t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right\} \tag{2.2}$$

Vale ressaltar que tanto na expressão 2.1 quanto na 2.2, μ não é uma variável aleatória. Portanto, se forem construídos m intervalos de confiança baseados em amostras de tamanho n, 95% deles conteriam o verdadeiro valor de μ. Isto é, o intervalo de confiança é uma variável aleatória.

2.2 Gráfico de controle clássico para processos normais

2.2.1 Gráfico de controle para a média

Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra aleatória de uma população normal com média μ e variância $\sigma^2 < \infty$, então $E(\bar{X}) = \mu$, $Var(\bar{X}) = \sigma^2/n$ e ainda $E(S^2) = \sigma^2/n$ (Casella e Berger, 2002).

Supondo que uma característica da qualidade seja normalmente distribuída com média μ e desvio padrão σ conhecidos, se $\mathbf{X} = (X_1, ..., X_n)$ é uma amostra aleatória de tamanho n, então, há uma probabilidade $1 - \alpha$ de qualquer média amostral cair entre

$$\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad e \quad \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{2.3}$$

que podem ser usados como limites superior e inferior em um gráfico de controle para médias amostrais (Montgomery, 2001).

Quando os parâmetros forem desconhecidos, caso muito comum especialmente na fase inicial do controle, é necessário calcular as estimativas dos parâmetros as quais devem se basear em, no mínimo, em $k = 25$ amostras de tamanho $n = 4$ ou $k = 20$ amostras de tamanho $n = 5$ ítens (Filho, 1970).

- Estimativa da média - a estimativa da média é calculada pela média geral, também chamada de média das médias das amostras:

$$\bar{\bar{X}} = \frac{1}{k}(\bar{X}_1 + ... + \bar{X}_k), \tag{2.4}$$

onde $\bar{X}_1$ é a média da primeira amostra, $\bar{X}_2$ a da segunda e assim por diante.

- Estimativa do desvio padrão - o cálculo da estimativa do desvio padrão σ pode ser feito a partir do desvio padrão amostral S. Para a i-ésima amostra, de n ítens, o desvio padrão amostral é

$$S_i = \left[\frac{(X_1^2 + ... + X_n^2 - n\bar{X}^2)}{n}\right]^{1/2}, \tag{2.5}$$

e a estimativa do desvio padrão da população σ é obtido calculando-se preliminarmente o desvio padrão amostral médio dado por

$$\bar{S} = \frac{1}{k}(S_1 + ... + S_k), \tag{2.6}$$

a estimativa de σ é então $\hat{\sigma} = \bar{S}/c_2$, onde c_2 é um fator de correção da estimativa, em função do tamanho da amostra. Para amostras com mais de 25 ítens, $c_2 = 1$.

Substituindo os estimadores de μ e σ na Equação 2.3, os limites para o gráfico de controle para a média são dados por:

$$\begin{aligned} LSC &= \bar{\bar{X}} + 3\frac{\bar{S}}{c_2\sqrt{n}} \\ LC &= \bar{\bar{X}} \\ LIC &= \bar{\bar{X}} - 3\frac{\bar{S}}{c_2\sqrt{n}} \end{aligned} \tag{2.7}$$

Observe que na Equação 2.7, $z_{\alpha/2}$ é substituído pelo valor 3. Montgomery (2001) justifica o uso dos limites de controle três *sigmas* (onde a porcentagem da área dentro do intervalo $\mu \pm 3\sigma$ é de 99,73%) pelo fato de darem bons resultados na prática. Além disso, em muitos casos, a verdadeira distribuição da característica da qualidade não é conhecida o bastante para se calcular os limites de probabilidade exatos. Contudo, mesmo que a característica da qualidade não seja normal, os resultados substituindo-se $z_{\alpha/2}$ por 3 são aproximadamente corretos devido ao teorema central do limite que estabelece que se $X_1, ..., X_n$ são variáveis aleatórias independentes com média μ_i e variância σ_i^2 e se $Y_i = \sum_{i=1}^{n} X_i$, então a distribuição de

$$\frac{Y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \tag{2.8}$$

se aproxima da distribuição normal $N(0; 1)$ à medida que n tende para infinito.

2.3 Gráfico de controle clássico para processos não-normais

2.3.1 Gráfico de controle para o número de defeitos

Existem casos em que o pesquisador está interessado na quantidade de não-conformidades (ou defeitos). Neste caso, utiliza-se o gráfico de controle para não-conformidades. Suponha que os defeitos, ou, não-conformidades de um processo qualquer, ocorram de acordo com a distribuição de Poisson, isto é,

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, ..., \tag{2.9}$$

onde x é o número de defeitos e $\lambda > 0$ é o parâmetro da distribuição de Poisson. Sabe-se que, tanto a média, quanto a variância desta distribuição são o parâmetro λ (Bussab e Morettin, 2003). Portanto, um gráfico de controle para não-conformidades (defeitos), com limites 3 sigmas*, dado que λ é conhecido, ou que um padrão é especificado pela gerência é definido como

$$\begin{aligned} LSC &= \lambda + 3\sqrt{\lambda} \\ LC &= \lambda \\ LIC &= \lambda - 3\sqrt{\lambda} \end{aligned} \tag{2.10}$$

Se não é dado nenhum padrão, então λ pode ser estimado como o número médio de defeitos observado em uma amostra preliminar de unidades de inspeção, por exemplo $\bar{\lambda}$. Neste caso, o gráfico de controle é definido da seguinte maneira

$$\begin{aligned} LSC &= \bar{\lambda} + 3\sqrt{\bar{\lambda}} \\ LC &= \bar{\lambda} \\ LIC &= \bar{\lambda} - 3\sqrt{\bar{\lambda}} \end{aligned} \tag{2.11}$$

* O risco α para os limites 3 sigma não é igualmente alocado acima do LSC e abaixo do LIC, visto que a distribuição de Poisson é assimétrica. Alguns autores recomendam o uso dos limites de probabilidades para este gráfico, particularmente quando λ é pequeno (Montgomery, 2001).

Capítulo 3

O Cenário da Inferência Estatística

3.1 Introdução

O problema fundamental em Inferência Estatística consiste em inferir sobre o valor de um parâmetro θ, associado ao modelo de probabilidade de uma variável aleatória X de interesse, com base na informação trazida por uma amostra aleatória de tamanho n de valores de X, $\mathbf{X} = (X_1, ..., X_n)$.

Se a amostra de tamanho n é grande, então as observações amostrais $\mathbf{x} = (x_1, ..., x_n)$ são uma longa lista de números que podem ser de difícil interpretação (Casella e Berger, 2002). Neste sentido, destaca-se um conceito associado à redução de dados: Suficiência. Esta redução se dá através de uma estatística (função da amostra $\mathbf{x}$) que condensa toda a informação sobre θ contida na amostra.

A idéia é promover um método de redução de dados que não descarte nenhuma informação sobre θ quando alcançada alguma sumarização dos dados. A definição de estatística suficiente é dada abaixo.

Definição 3.1.1. *Uma estatística $T(\boldsymbol{X})$ é uma estatística suficiente para θ se a distribuição condicional da amostra $\boldsymbol{X}$, dada $T(\boldsymbol{X})$, for independente do parâmetro θ. Isto é, $T(\boldsymbol{X})$ é uma estatística suficiente para θ se, e somente se, $P[X_1 = x_1, ..., X_n = x_n | T(X_1, ..., X_n) = T(x_1, ...x_n)]$ for independente de θ, (no caso discreto) ou $f[X_1 = x_1, ..., X_n = x_n | T(X_1, ..., X_n) = T(x_1, ...x_n)]$ for independente de θ, (no caso contínuo).*

A Definição 3.1.1 mostra como se pode verificar se uma estatística $T(\mathbf{X})$ é ou não suficiente. Contudo, não pode ser utilizada para obtenção de uma estatística suficiente. Bickel e Doksum (2001) apresentam um procedimento para a obtenção de estatísticas suficientes, denominado de teorema da fatoração, formalizado a seguir.

Teorema 3.1.1. *Uma estatística $T(\boldsymbol{X})$ é suficiente para θ se, e somente se, existe uma função $g(t|\theta)$ e $h(x)$ tal que,*

$$f(\mathbf{x}|\theta) = g(T(\mathbf{X})|\theta)h(\mathbf{x}), \tag{3.1}$$

para todo $\mathbf{x} \in \chi$ e $\theta \in \Theta$, onde $f(\mathbf{x}|\theta)$ representa a função de verossimilhança.

Muitos modelos estatísticos podem ser considerados como casos especiais de uma família mais geral de distribuições (Bolfarine e Sandoval, 2001). Esta família é denominada Família Exponencial e é definida a seguir.

Definição 3.1.2. *Diz-se que a distribuição da variável aleatória X pertence a Família Exponencial unidimensional de distribuições, se pudermos escrever sua f.p. (função de probabilidade) ou f.p.d. (função densidade de probabilidade) como*

$$f(x|\theta) = e^{c(\theta)T(x)+d(\theta)+S(x)}, \quad x \in A, \tag{3.2}$$

onde c, d são funções reais de θ; T, S são funções reais de x e A, que é o suporte da variável X, não depende de θ.

O próximo teorema estabelece que amostras aleatórias de famílias exponenciais unidimensionais são também membros da Família Exponencial unidimensional (Bolfarine e Sandoval, 2001).

Teorema 3.1.2. *Sejam $X_1, ..., X_n$ uma amostra aleatória de tamanho n da variável aleatória X, com função de densidade (ou de probabilidade) dada por 3.2. Então, a distribuição conjunta de $X_1, ..., X_n$ é dada por:*

$$f(x_1, ..., x_n|\theta) = e^{c^*(\theta)\sum_{i=1}^n T(x_i)+d^*(\theta)+S^*(x)}, \quad \mathbf{x} \in A^n, \tag{3.3}$$

que também é da família exponencial com $T(\mathbf{x}) = \sum_{i=1}^n T(x_i)$, $c^*(\theta) = c(\theta)$, $d^*(\theta) = nd(\theta)$ e $S^*(\mathbf{x}) = \sum_{i=1}^n S(x_i)$.

Observe de 3.3 que considerando

$$h(x_1, ..., x_n) = e^{\sum_{i=1}^{n} S(x_i)} \prod_{i=1}^{n} I_A(x_i), \quad e \quad g_\theta(T) = e^{c(\theta) \sum_{i=1}^{n} T(x_i) + nd(\theta)},$$

temos, pelo critério da fatoração (Teorema 3.1.1), que a estatística $T(\mathbf{X}) = \sum_{i=1}^{n} T(X_i)$ é suficiente para θ.

Entre os membros da família exponencial pode-se destacar os seguintes modelos de probabilidade (Ross, 1977): Bernoulli, Binomial, Geométrica, Poisson, Normal, Gama, Exponencial e t-student. Sendo que, este trabalho limita-se a apresentar apenas dois modelos: Poisson e Exponencial.

i) Distribuição Poisson

Uma variável aleatória X, que assume os valores $0, 1, 2, ...$, é dita ser uma variável aleatória Poisson com parâmetro λ, se para algum $\lambda > 0$,

$$p(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, ... \tag{3.4}$$

Agora, seja $X_1, ..., X_n$ uma amostra aleatória de $X \sim Poi(\lambda)$ a partir do Teorema 3.1.2 e do critério da fatoração, sabe-se que: $c(\lambda) = \ln \lambda$; $T(x) = x$; $d(\lambda) = -\lambda$ e $S(x) = \ln x!$. Portanto, $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ é suficiente para λ.

ii) Distribuições Exponencial

Diz-se que X tem distribuição exponencial com parâmetro λ, se para algum $\lambda > 0$ sua f.d.p. poder ser escrita como

$$f(x|\lambda) = \lambda e^{-\lambda x}, \quad 0 < x < +\infty. \tag{3.5}$$

Agora, seja $X_1, ..., X_n$ uma amostra aleatória de $X \sim Exp(\lambda)$ a partir do Teorema 3.1.2 e do critério da fatoração, sabe-se que: $c(\lambda) = -\lambda$; $T(x) = x$; $d(\lambda) = \ln \lambda$ e $S(x) = 0$. Portanto, $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ é suficiente para λ.

3.2 Aspectos gerais da inferência bayesiana

Nos métodos bayesianos, o grau de incerteza inicial sobre o parâmetro desconhecido $\theta \in \Theta$ deve ser representado por um modelo de probabilidade para θ. Isto é, nos métodos bayesianos θ é uma variável aleatória. Este modelo de probabilidade, que representa a incerteza inicial ou a priori sobre θ é denotado por $h(\theta)$ e chamado de distribuição a priori de θ. Deste modo, $h(\theta)$ indica o grau de credibilidade dado pelo pesquisador ao parâmetro de interesse, onde $h(\theta)$ pode ser uma função de probabilidade, no caso em que θ é discreto, ou uma função densidade de probabilidade no caso contínuo.

Ao realizar inferências sobre o parâmetro não observável θ, os métodos clássicos se baseiam em probabilidades associadas com diferentes amostras $\mathbf{X}$, que poderiam ocorrer para algum valor fixo, mas desconhecido, do parâmetro θ. É o que ocorre quando se fazem inferências com base nas distribuições amostrais de certas estatísticas (Paulino *et al.*, 2003). Já na metodologia bayesiana todo o processo de inferência baseia-se na distribuição *a posteriori* de θ, dada por

$$h(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)h(\theta)}{\int_\Theta f(\mathbf{x}|\theta)h(\theta)d\theta}, \quad \theta \in \Theta, \tag{3.6}$$

onde $f(\mathbf{x}|\theta)$ representa a distribuição conjunta da amostra, a função de verossimilhança. Como o denominador (distribuição preditiva) da Equação 3.6, não depende de θ, este funciona como uma constante normalizadora de $h(\theta|x)$, sendo assim, a Equação 3.6 pode ser escrita como o produto da verossimilhança pela priori, a menos de uma constante,

$$h(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)h(\theta), \quad \theta \in \Theta. \tag{3.7}$$

3.2.1 Distribuições *a priori* conjugadas

Uma das principais dificuldades na abordagem bayesiana é a não trivialidade das soluções analíticas, quando existem, das distribuições *a posteriori*. A simplicidade da operação bayesiana pode ficar garantida na medida em que se impõe a família de distribuições *a priori* $\mathcal{H}$, fechada sob amostragem de (qualquer elemento de) $\mathcal{F} = \{f(x|\theta) : \theta \in \Theta\}$, isto é,

$$h(\theta) \in \mathcal{H} \Rightarrow h(\theta|x) \propto h(\theta) f(x|\theta) \in \mathcal{H}, \tag{3.8}$$

nestas condições, diz-se também que $\mathcal{H}$ é uma família conjugada de $\mathcal{F}$ (Paulino *et al*, 2003).

A seguir, mostra-se as distribuições conjugadas para as distribuições Poisson e Exponencial.

***i*) Distribuição de Poisson e distribuição Gama**

Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra aleatória da distribuição de Poisson com parâmetro θ. Sua função de verossimilhança é dada por:

$$L(\mathbf{x}|\theta) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}; \quad \theta > 0.$$

O núcleo desta verossimilhança (a parte que depende só de θ) caracteriza a família de distribuições Gama. Assim, a distribuição *a priori* para o parâmetro θ é dada pela distribuição Gama com hiperparâmetros a e b, isto é,

$$h(\theta) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}, \quad a, b > 0 \quad e \quad \theta > 0.$$

Então, utilizando a Equação 3.7, temos

$$\begin{aligned} h(\theta|\mathbf{x}) &\propto L(\theta|\mathbf{x})h(\theta) \\ &\propto e^{-n\theta}\theta^{\sum_{i=1}^{n} x_i}\theta^{a-1}e^{-b\theta} \\ &\propto \theta^{a+\sum_{i=1}^{n} x_i - 1}e^{-\theta(b+n)}. \end{aligned}$$

O núcleo de $h(\theta|\mathbf{x})$ corresponde a distribuição $Ga(a + \sum_{i=1}^{n} x_i, b + n)$. Portanto, a distribuição Gama é conjugada para a distribuição Poisson.

***ii*) Distribuição Exponencial e distribuição Gama**

Seja $\mathbf{X} = (X_1, ..., X_n)$ uma amostra aleatória da distribuição Exponencial com parâmetro θ. Sua função de verossimilhança é dada por:

$$L(\mathbf{x}|\theta) = \theta^n e^{-\theta \sum_{i=1}^n x_i}, \quad \theta > 0.$$

O núcleo desta verossimilhança (a parte que depende só de θ) caracteriza a família de distribuições Gama. Assim, a distribuição *a priori* para o parâmetro θ é dada pela distribuição Gama com hiperparâmetros a e b, isto é,

$$h(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta}, \quad a, b > 0 \quad e \quad \theta > 0.$$

Então, utilizando a Equação 3.7, temos

$$\begin{aligned} h(\theta|\mathbf{x}) &\propto L(\theta|\mathbf{x}) h(\theta) \\ &\propto \theta^n e^{-\theta \sum_{i=1}^n x_i} \theta^{a-1} e^{-b\theta} \\ &\propto \theta^{a+n-1} e^{-\theta(b+\sum_{i=1}^n x_i)}. \end{aligned}$$

O núcleo de $h(\theta|\mathbf{x})$ corresponde a distribuição $Ga(a+n, b+\sum_{i=1}^n x_i)$. Portanto, a distribuição Gama é conjugada também para a distribuição Exponencial.

3.2.2 Distribuições *a priori* não-informativa

A primeira idéia de "não informação"a priori que se pode ter é pensar em todos os possíveis valores de θ como igualmente prováveis, isto é, com uma distribuição *a priori* uniforme (Ehlers, 2007). Neste caso, fazendo $h(\theta) \propto k$ para θ variando em um subconjunto da reta significa que nenhum valor particular tem preferência (Bayes, 1763). Porém esta escolha de priori pode trazer algumas dificuldades técnicas,

i) Se o intervalo de variação de θ for ilimitado então a distribuição *a priori* é imprópria, isto é,

$$\int_\Theta h(\theta) d\theta = \infty. \tag{3.9}$$

ii) Se $\phi = g(\theta)$ é uma reparametrização não linear monótona de θ então $h(\phi)$ é não uniforme já que pelo teorema de transformação de variáveis

$$h(\phi) = h(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right| \propto \left| \frac{d\theta}{d\phi} \right|. \tag{3.10}$$

Na prática, como o interesse é a distribuição *a posteriori*, o que realmente importa, independente da impropriedade da distribuição *a priori*, é que a *posteriori* seja própria antes de fazer qualquer inferência.

A classe de prioris não informativas proposta por Jeffreys é invariante a transformações 1 a 1, embora em geral seja imprópria. Antes de defini-la é necessário apresentar a definição da medida de informação de Fisher.

Definição 3.2.1. *Considere uma única observação X com f.p. ou f.d.p. $f(x|\theta)$. A medida de informação esperada de Fisher de θ através de X é definida como*

$$I(\theta) = E\left[-\frac{\partial^2 log f(x|\theta)}{\partial \theta^2} \right]. \tag{3.11}$$

A medida de informação de Fisher $I(\theta)$ fica então definida como

$$I(\theta) = -\frac{\partial^2 log f(x|\theta)}{\partial \theta^2}. \tag{3.12}$$

Definição 3.2.2. *Seja uma observação X com f.p ou f.d.p. $f(x|\theta)$. A priori não-informativa de Jaffreys tem f.p ou f.d.p. dada por*

$$h(\theta) \propto [I(\theta)]^{1/2}. \tag{3.13}$$

Existem outros métodos para a obtenção de prioris não-informativas como por exemplo: método de Bayes-Laplace, Box-Tiao, Berger-Bernardo, dentre outros. Porém, neste trabalho limita-se a apresentação do método mostrado anteriormente.

3.2.3 Distribuição preditiva

Frequentemente as inferências sobre os parâmetros do modelo postulado não são um fim em si, mas antes, um meio visando predizer dados amostrais futuros (Paulino *et al.*, 2003). Formalmente, pretende-se predizer um vetor $\mathbf{Y}$ (ou alguma função dele) com distribuição amostral dependente de θ com base em observações de um vetor aleatório $\mathbf{X}$, com distribuição $f(\mathbf{x}|\theta)$, considerando todo o conhecimento acumulado sobre θ.

Dada a aleatoriedade intrínseca dos dados futuros, afigura-se natural pretender atingir aquele objetivo através de um modelo probabilístico para Y, condicionado nos dados atuais $\mathbf{x}$. Como estes transmitem informação sobre o parâmetro θ, que governa a distribuição amostral de $\mathbf{Y}$ dado $\mathbf{x}$, $f(\mathbf{y}|\mathbf{x},\theta)$, é natural ponderar esta com o conhecimento acumulado sobre θ, quantificado na distribuição *a posteriori* de θ, obtendo-se a chamada distribuição preditiva *a posteriori*, cuja a f.p ou f.d.p. é

$$h(y|x) = \int_{\Theta} f(\mathbf{y}|\mathbf{x},\theta)h(\theta|x)d\theta. \tag{3.14}$$

3.2.4 Intervalos de credibilidade

A estimação por intervalos na abordagem Bayesiana dá-se através da construção dos chamados intervalos de credibilidade. Um intervalo com $100(1-\alpha)\%$ de credibilidade para um parâmetro θ (suposto aqui um escalar) é composto por um par de valores $(\underline{\theta},\overline{\theta})$ de Θ, tais que,

$$P(\underline{\theta} \leq \theta \leq \overline{\theta}|\mathbf{x}) = \int_{\underline{\theta}}^{\overline{\theta}} h(\theta|\mathbf{x})d\theta = 100(1-\alpha)\%. \tag{3.15}$$

Uma forma de construir um intervalo de credibilidade é considerar na distribuição a posteriori abas de igual credibilidade,

$$\int_{-\infty}^{\underline{\theta}} h(\theta|x_1,...,x_n)d\theta = \int_{\overline{\theta}}^{+\infty} h(\theta|x_1,...,x_n)d\theta = \frac{\alpha}{2}. \tag{3.16}$$

Conforme Paulino *et al.* (2003) a Equação 3.16 possui um inconveniente: o intervalo $(\underline{\theta},\overline{\theta})$ não é único, podendo suceder que valores de θ contidos neste intervalo tenham menor

credibilidade que valores de θ não incluídos no mesmo intervalo. Assim, para proceder à escolha de um certo intervalo que atenda ao nível de credibilidade de $100(1-\alpha)\%$ e ao mesmo tempo minimize a respectiva amplitude, os bayesianos preferem trabalhar com intervalos *HPD* (*highest posteriori density*), ou seja, um intervalo $(\underline{\theta}, \overline{\theta})$ tal que

$$(\theta', \theta'') \subset \theta : h(\theta|x_1, ..., x_n) \geq k(\alpha), \tag{3.17}$$

onde $k(\alpha)$ é o maior número real tal que,

$$P(\theta' \leq \theta \leq \theta'') = 100(1-\alpha)\%. \tag{3.18}$$

3.2.5 Aspectos computacionais - Métodos de Monte Carlo

Em muitas situações na inferência bayesiana, o cálculo por via analítica de distribuições *a posteriori* não apresentam soluções analíticas. Em tais situações, aproximações por simulação podem ser aplicadas naturalmente (Chen *et al.*, 2000). Os métodos de Monte Carlo são uma excelente alternativa para solução destes problemas, visto que, baseia-se em simulação estocástica, ou seja, estes métodos simulam valores de números (pseudo) aleatórios de distribuições de probabilidade.

A distribuição *a posteriori* pode ser convenientemente resumida em termos de esperanças de funções particulares do parâmetro θ (Ehlers, 2007), isto é

$$E[g(\theta)|\mathbf{x}] = \int g(\theta)h(\theta|\mathbf{x})d\theta, \tag{3.19}$$

ou distribuições *a posteriori* marginais quando $\boldsymbol{\theta}$ for multidimensional, isto é

$$h(\boldsymbol{\theta}_1|\boldsymbol{x}) = \int h(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}_2$$

onde $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Assim, o problema geral da inferência Bayesiana consiste em calcular tais valores esperados sob a distribuição *a posteriori* de θ.

- **Método de Monte Carlo Simples**

A idéia deste método é escrever a integral que se deseja calcular como um valor esperado (Ehlers, 2007). Supondo que pode-se gerar uma amostra $\theta_1, ..., \theta_m$ da distribuição posterior $h(\theta|\mathbf{x})$, o Método de Monte Carlo Simples consiste em aproximar a integral 3.19 por

$$\widehat{I} = \widehat{E}[g(\theta)|\mathbf{x}] = \overline{g} = \frac{1}{m}\sum_{i=1}^{m} g(\theta_i), \tag{3.20}$$

onde a variância deste estimador pode ser estimada como

$$v = \frac{1}{m^2}\sum_{i=1}^{m}[g(\theta_i) - \overline{g}]^2. \tag{3.21}$$

Uma vez que as gerações são independentes, pela Lei Forte dos Grandes Números (Casella e Berger, 2002) segue que $\widehat{E}[g(\theta)|\mathbf{x}]$ converge quase certamente para $E[g(\theta)|\mathbf{x}]$. Além disso, de acordo com o Teorema Central do Limite, para n grande segue que

$$\frac{\overline{g} - E[g(\theta)|\mathbf{x}]}{\sqrt{v}} \sim N(0,1). \tag{3.22}$$

No caso multivariado a extensão também é direta. Seja $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)'$ um vetor aleatório de dimensão k com função de densidade $h(\boldsymbol{\theta})$. Neste caso os valores gerados serão também vetores $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m$ e o estimador de Monte Carlo fica

$$\widehat{I} = \frac{1}{m}\sum_{i=1}^{m} g(\boldsymbol{\theta}_i).$$

- **Intervalos de credibilidade**

Considere que $(\theta_i, 1 \leq i \leq m)$ é uma amostra aleatória da densidade *a posteriori* univariada $h(\theta|x)$, com função de distribuição $H(\theta|x)$, que pretende ser resumida por um intervalo de credibilidade com $100(1-\alpha)\%$ de credibilidade. A determinação exata deste exige o conhecimento completo da distribuição *a posteriori*, o que nem sempre é possível devido a constante normalizadora (Paulino *et al.*, 2003).

O intervalo de credibilidade central para θ é definido por $R_c(\alpha) = \left(\theta_{\frac{\alpha}{2}}, \theta_{1-\frac{\alpha}{2}}\right)$, cujos

extremos definem os quantis de probabilidade *a posteriori* $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$, respectivamente, de θ, isto é, $H(\theta_\beta|\mathbf{x}) = \beta$.

Uma aproximação via método de Monte Carlo de $R_c(\alpha)$ é obtida ordenando a amostra aleatória e usando os quantis empíricos. Especificamente, representando agora $(\theta_{(i)}, 1 \leq i \leq m)$ a amostra ordenada, a estimativa de Monte Carlo de $R_c(\alpha)$ é definida por

$$\widehat{R}_c(\alpha) = \left(\theta_{\left[m\frac{\alpha}{2}\right]}, \theta_{\left[m1-\frac{\alpha}{2}\right]}\right), \tag{3.23}$$

onde $[m(1-\alpha)]$ é a parte inteira de $m(1-\alpha)$.

O intervalo $R_c(\alpha)$ de $h(\theta|x)$ não é o melhor resumo intervalar de uma distribuição unimodal quando esta não é simétrica, sendo por isso claramente preterido em favor do intervalo *HPD* $R_0(\alpha) = \{\theta : h(\theta|x) \geq k_\alpha\}$ onde k_α é a maior constante para a qual a probabilidade *a posteriori* de $R_c(\alpha)$ é no mínimo $1-\alpha$. Pela sua definição, este intervalo é mais difícil de determinar do que $R_c(\alpha)$, mesmo que se disponha de formas fechadas para as funções densidade e de distribuição *a posteriori* de θ (Paulino *et al.*, 2003).

Chen *et al.* (2000) propõe um procedimento de Monte Carlo para aproximações de $R_0(\alpha)$ com base na amostra ordenada, determinando os intervalos de credibilidade $1-\alpha$ da seguinte forma

$$\widehat{R}_i(\alpha) = (\theta_{(i)}, \theta_{(i+[m(1-\alpha)])}), \quad i = 1, ..., m - [m(1-\alpha)], \tag{3.24}$$

desta forma, a aproximação de Monte Carlo de $R_0(\alpha)$ é definido por $\widehat{R}_0(\alpha) = R_{i0}(\alpha)$ tal que

$$\widehat{R}_0(\alpha) = R_{i0}(\alpha) = [\theta_{(i0+[m(1-\alpha)])} - \theta_{(i0)}] = min[\theta_{(i+[m(1-\alpha)])} - \theta_{(i)}], \tag{3.25}$$

com $1 \leq i \leq m - [m(1-\alpha)]$.

- **Quantidades preditivas**

Paulino *et al.* (2003), propõe um procedimento para estimar a densidade preditiva *a*

posteriori. Dado que $h(y|x) = E_{\theta|x}[f(y|\theta, x)]$, facilmente se obtém a respectiva aproximação de Monte Carlo

$$\widehat{h}(y|x) = \frac{1}{m}\sum_{i=1}^{m} f(y|\theta_i, x), \tag{3.26}$$

com base no valores i.i.d. (independentes e identicamente distribuídos) simulados de $h(\theta|x)$. Para a simulação de Monte Carlo de quantidades associadas com a distribuição preditiva de $h(y|x)$ é necessário obter-se uma amostra aleatória desta distribuição. Isto é possível através do chamado método de composição (Tanner, 1996, Sec. 3.3) caso se saiba amostrar da distribuição amostral de y, obtendo-se então a amostra $\mathbf{Y} = (Y_1, ..., Y_m)$ de $h(y|x)$ do seguinte modo:

1. Retira-se uma amostra de valores i.i.d. de $h(\theta|x)$, $(\theta_1, ..., \theta_m)$;

2. Para cada i, retira-se y_i de $f(y|\theta_i, x)$, $i = 1, ..., m$.

Com base nesta amostra pode-se calcular facilmente aproximações de vários resumos da distribuição preditiva. Por exemplo, estimativas da predição média e do intervalo de predição *HPD* para a observação futura $y \in \Re$ obtém-se pela mesma forma como a média *a posteriori* e os intervalos de θ, da mesma forma mostrada no item anterior.

Capítulo 4

Gráficos de Controle Bayesianos

4.1 Gráficos de controle baseados em distribuições preditivas

Menzefricke (2002) propôs gráficos de controle bayesianos para a média de um processo com distribuição normal. Para a construção destes gráficos de controle ele utilizou a distribuição preditiva de uma amostra futura para encontrar a região de rejeição (limites de controle). Inicialmente, o autor mostra a metodologia utilizada para construir os gráficos de controle baseados em distribuições preditivas e, posteriormente, uma forma de avaliar estes gráficos a partir do "comprimento da sequência", ou seja, estabelece o tamanho ótimo da amostra a ser considerada.

4.1.1 Construção dos gráficos de controle

Suponha que um processo estável está gerando os dados $\mathbf{x}$, com f.d.p. dada por $f(\mathbf{x}|\theta)$, onde θ é o parâmetro de interesse. Em muitas situações, o valor de θ não é exatamente conhecido, e assume-se que a informação *a priori* sobre θ pode ser representada pela distribuição *a priori*, $h(\theta)$. Mais adiante assume-se que uma amostra aleatória $\mathbf{X} = (X_1, ..., X_{n_c})$ de tamanho n_c deste processo está disponível, a função de verossimilhança dos dados amostrais é dada por

$$f(x_1, ..., x_{n_c}|\theta) = \prod_{i=1}^{n_c} f(x_i|\theta). \tag{4.1}$$

A partir da existência de uma estatística suficiente para θ, aqui denotada por $T_c = T_c(\mathbf{X})$, tem-se que a distribuição *a posteriori* de $h(\theta|\mathbf{x})$ depende da amostra $\mathbf{X}$ apenas através de $T_c(\mathbf{X})$. Assim, tem-se

$$h(\theta|\mathbf{x}) = h(\theta|T_c) \propto f(T_c|\theta)h(\theta). \tag{4.2}$$

Dado um particular valor para o parâmetro θ, pode-se obter a verossimilhança de uma amostra futura $\mathbf{Y} = (Y_1, ..., Y_n)$, de tamanho n,

$$f(\mathbf{y}|\theta) = \prod_{i=1}^{n} f(y_i|\theta). \tag{4.3}$$

Agora dado $T = T(\mathbf{Y})$ uma estatística suficiente e $f(T|\theta)$ sua f.p. ou f.d.p., então a distribuição preditiva *a posteriori* de T, dada a amostra inicial, é dada por

$$h(T|T_c) = \int_{\Theta} f(T|\theta)h(\theta|T_c)d\theta. \tag{4.4}$$

A partir desta distribuição preditiva, pode-se agora obter a região de rejeição com credibilidade $1 - \alpha$ que depende do valor observado de T_c. Denotando-se esta região por $R(\alpha, T_c)$, temos que o tamanho desta região de rejeição é dado por

$$\alpha = \int_{R(\alpha,T_c)} h(T|T_c)dT. \tag{4.5}$$

Se $T \in R(\alpha, T_c)$, o gráfico de controle sinalizará e deve-se investigar as possíveis causas para esta mudança. Caso contrário, conclui-se que não houve mudanças no processo.

De acordo com Menzefricke (2002), uma amostra futura de tamanho n é retirada repetidamente de um processo, e o interesse é avaliar o comprimento da sequência, isto é, o número de tais amostras r (é importante ressaltar que r não inclui a amostra quando o gráfico sinaliza), até que o gráfico de controle sinalize pela primeira vez. Dado θ e um processo estável, a distribuição do comprimento da sequência r é geométrica com parâmetro

$$\psi(\theta) = \int_{R(\alpha,T_c)} f(T|\theta)dT \tag{4.6}$$

O valor de θ é desconhecido e a incerteza é descrita pela distribuição a posteriori, $h(\theta|T_c)$. Deste modo, pode-se obter a distribuição preditiva de r, onde a média de r é

$$E(r|\alpha, T_c) = \int_{\Theta} \frac{1}{1-\psi(\theta)} h(\theta|T_c) d\theta. \quad (4.7)$$

A Equação 4.7 é denominada *average run length-ARL* (comprimento médio da sequência), isto é, o número médio de amostras r até que o gráfico de controle sinalize pela primeira vez.

4.1.2 Avaliação dos Gráficos de Controle

Para a avaliação dos gráficos de controle, deve-se investigar a performance da região de rejeição, assumindo que o parâmetro θ é igual a um valor hipotético, θ_a, em particular, assume-se que os dados $\mathbf{y} = (y_1, ..., y_n)$ são gerados pelo seguinte modelo:

$$f(y|\theta_a) = \prod_{i=1}^{n} f(\mathbf{y}|\theta_a). \quad (4.8)$$

Na prática, raramente se conhece o atual valor de θ_a, e a incerteza sobre θ é caracterizada pela distribuição *a posteriori* $h(\theta|T_c)$ dado em 4.2. Assumindo θ_a conhecido, a correspondente distribuição da estatística suficiente T pode ser denotada por $f(T|\theta_a)$, e o tamanho da região de rejeição $R(\alpha, T_c)$ é então

$$\alpha(T_c, \theta_a) = \int_{R(\alpha, T_c)} f(T|\theta_a) dT. \quad (4.9)$$

Além disso, dado o valor do parâmetro θ_a, a distribuição do comprimento da sequência, r_a, é geométrica com parâmetro $\alpha(T_c, \theta_a)$.

Na avaliação do gráfico de controle, deve-se investigar $\alpha(T_c, \theta_a)$ para diferentes valores de θ_a. Quando θ_a é unidimensional, um simples gráfico de $\alpha(T_c, \theta_a)$ *versus* θ_a pode ser usado para avaliar o gráfico de controle.

4.2 Gráfico de controle em processo normais

4.2.1 Caso 1: Média com desvio padrão conhecido

Na seção anterior mostrou-se a metodologia utilizada por Menzefricke (2002), para construção de gráficos de controle baseados em distribuições preditivas. Nesta seção é reproduzido o caso em que o parâmetro de interesse é a média do processo, isto é, $\theta = \mu$, conforme apresentado em Menzefricke (2002).

Quando um processo estável está gerando os dados, a variável resposta é gerada por uma distribuição normal com média μ desconhecida e variância σ^2 conhecida. Assume-se que a informação *a priori* para μ pode ser resumida por uma distribuição normal

$$\mu \sim N\left(m_0; \frac{\sigma^2}{n_0}\right). \tag{4.10}$$

Seja $\mathbf{x} = (x_1, ..., x_{n_c})$ dados do processo estável, a estatística suficiente é a média amostral, $T_c = \bar{x}$. A distribuição de $f(T_c|\theta)$ é então $\bar{x}|\mu \sim N(\mu; \sigma^2/n_c)$, e a distribuição *a posteriori* para μ é calculada por

$$\begin{aligned} h(\mu|\bar{x}) &\propto h(\mu) f(\bar{x}|\mu) \\ &\propto \exp\left[\frac{1}{2\frac{\sigma^2}{n_0}}(\mu - m_0)^2\right] \exp\left[\frac{1}{2\frac{\sigma^2}{n_c}}(\bar{x} - \mu)^2\right] \\ &\propto \exp\left\{\frac{1}{2\sigma^2}\left[\underbrace{n_0(\mu - m_0)^2 + n_c(\bar{x} - \mu)^2}\right]\right\} \end{aligned} \tag{4.11}$$

Agora, utilizando-se a identidade fundamental (Paulino *et al.*, 2003), expressa em 4.12

$$d_1(z - c_1)^2 + d_2(z - c_2)^2 = (d_1 + d_2)(z - c)^2 + \frac{d_1 d_2}{d_1 + d_2}(c_1 - c_2)^2, \tag{4.12}$$

onde $c = \frac{d_1 c_1 + d_2 c_2}{d_1 + d_2}$, temos que a expressão destacada em 4.11 pode ser escrita como

$$n_0(\mu - m_0)^2 + n_c(\bar{x} - \mu)^2 = (n_0 + n_c)\left(\mu - \frac{n_0 m_0 + n_c \bar{x}}{n_0 + n_c}\right)^2 + \frac{n_0 n_c}{n_0 + n_c}(m_0 - \bar{x})^2.$$

Deste modo,

$$h(\mu|\bar{x}) \quad \propto \quad \exp\left[-\frac{1}{2\frac{\sigma^2}{n_0+n_c}}\left(\mu - \frac{n_0 m_0 + n_c\bar{x}}{n_0+n_c}\right)^2\right]. \tag{4.13}$$

O núcleo de 4.13 mostra que,

$$\mu|\bar{x} \sim N\left(\frac{n_0 m_0 + n_c\bar{x}}{n_0+n_c}; \frac{\sigma^2}{n_0+n_c}\right).$$

Fazendo $n_1 = n_0 + n_c$ e $m_1 = \dfrac{n_0 m_0 + n_c\bar{x}}{n_1}$, temos que

$$\mu|\bar{x} \sim N\left(m_1; \frac{\sigma^2}{n_1}\right). \tag{4.14}$$

Supondo que o processo permaneça estável, seja $\mathbf{y} = (y_1, ..., y_n)$ dados futuros deste processo, e $T = \bar{y}$ a estatística suficiente. A distribuição de $f(T|\theta)$ é então $\bar{y}|\mu \sim N(\mu; \sigma^2/n)$, e a distribuição preditiva *a posteriori* para $\bar{y}$ é calculada da seguinte maneira

$$h(\bar{y}|\bar{x}) \quad = \quad \int_{-\infty}^{+\infty} f(\bar{y}|\mu) f(\mu|\bar{x}) d\mu$$

$$\begin{aligned} &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{n}}(\mu-\bar{y})^2\right] \frac{1}{\sqrt{2\pi}\frac{\sigma}{\sqrt{n_1}}} \exp\left[-\frac{1}{2\frac{\sigma^2}{n_1}}(\mu-m_1)^2\right] d\mu \\ &= \int_{-\infty}^{+\infty} \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sigma\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{n(\mu-\bar{y})^2 + n_1(\mu-m_1)^2}\right]\right\} d\mu, \end{aligned} \tag{4.15}$$

agora, utilizando-se a identidade dada por 4.12, temos que a expressão destacada em 4.15 pode ser escrita como

$$n(\mu-\bar{y})^2 + n_1(\mu-m_1)^2 = (n+n_1)\left(\mu - \frac{n\bar{y}+n_1 m_1}{n+n_1}\right)^2 + \frac{nn_1}{n+n_1}(\bar{y}-m_1)^2.$$

Deste modo,

$$\begin{aligned} h(\bar{y}|\bar{x}) &= \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{2\pi}\sigma^2}\int_{-\infty}^{+\infty}\exp\left[-\frac{1}{2\frac{\sigma^2}{n+n_1}}\left(\mu-\frac{n\bar{y}+n_1m_1}{n+n_1}\right)^2\right]\times \\ &\times \exp\left[-\frac{1}{2\sigma^2\frac{(n+n_1)}{nn_1}}\left(\bar{y}-m_1\right)^2\right]d\mu \\ &= \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{2\pi}\sigma^2}\frac{\sqrt{2\pi}\sigma}{\sqrt{n+n_1}}\exp\left[-\frac{1}{2\sigma^2\frac{(n+n_1)}{nn_1}}\left(\bar{y}-m_1\right)^2\right]\times \\ &\times \underbrace{\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma/\sqrt{n+n_1}}\exp\left[-\frac{1}{2\frac{\sigma^2}{n+n_1}}\left(\mu-\frac{n\bar{y}+n_1m_1}{n+n_1}\right)^2\right]d\mu}. \end{aligned} \quad (4.16)$$

Na expressão 4.16, o integrando da integral em destaque corresponde a f.d.p. de uma $N\left(\frac{n\bar{y}+n_1m_1}{n+n_1};\frac{\sigma^2}{n+n_1}\right)$, portanto

$$h(\bar{y}|\bar{x}) = \frac{1}{\sqrt{2\pi}\sigma\frac{(\sqrt{n+n_1})}{\sqrt{nn_1}}}\exp\left[-\frac{1}{2\sigma^2\frac{(n+n_1)}{nn_1}}(\bar{y}-m_1)^2\right]. \quad (4.17)$$

Logo, pode-se concluir a partir de 4.17 que

$$\bar{y}|\bar{x} \sim N\left[m_1;\sigma^2\frac{(n+n_1)}{nn_1}\right], \quad (4.18)$$

que também pode ser escrita como

$$\bar{y}|\bar{x} \sim N\left[m_1;\sigma^2\left(\frac{1}{n}+\frac{1}{n_1}\right)\right]. \quad (4.19)$$

Uma região de aceitação de tamanho $1-\alpha$ é denotada por $A(\alpha,T_c)=(t_1,t_2)$, de forma que

$$A(\alpha,T_c)=\int_{t_1}^{t_2}\frac{1}{\sqrt{2\pi}\sigma\frac{(\sqrt{n+n_1})}{\sqrt{nn_1}}}\exp\left[-\frac{1}{2\sigma^2\frac{(n+n_1)}{nn_1}}(\bar{y}-m_1)^2\right]d\bar{y}, \quad (4.20)$$

agora, fazendo a transformação $z = \frac{(\bar{y}-m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})}$, diz-se que $z \sim N(0;1)$, portanto

$$\begin{aligned} A(\alpha, T_c) &= \int_{\frac{(t_1-m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})}}^{\frac{(t_2-m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})}} \frac{1}{\sqrt{2\pi}\sigma\frac{(\sqrt{n+n_1})}{\sqrt{nn_1}}} \exp\left(-\frac{1}{2}z^2\sigma\right) \frac{(\sqrt{n+n_1})}{\sqrt{nn_1}} dz, \\ A(\alpha, T_c) &= \int_{\frac{(t_1-m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})}}^{\frac{(t_2-m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz. \end{aligned} \tag{4.21}$$

Observe que a Equação 4.21 corresponde a integral da f.d.p. de uma distribuição $N(0;1)$, desta forma pode-se isolar o valor de t_1 e t_2 da seguinte maneira

$$-z_{\alpha/2} = \frac{(t_1 - m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})} \Rightarrow t_1 = m_1 - z_{\alpha/2}\frac{\sigma(\sqrt{n+n_1})}{\sqrt{nn_1}}, \tag{4.22}$$

$$z_{\alpha/2} = \frac{(t_2 - m_1)(\sqrt{nn_1})}{\sigma(\sqrt{n+n_1})} \Rightarrow t_2 = m_1 + z_{\alpha/2}\frac{\sigma(\sqrt{n+n_1})}{\sqrt{nn_1}}, \tag{4.23}$$

onde $z_{\alpha/2}$ é o valor tabelado da Normal padrão.

Agora, escrevendo $m_1 = \bar{x} + \frac{n_0}{n_0+n_c}(n_0 - \bar{x})$ e $\frac{\sigma(\sqrt{n+n_1})}{\sqrt{nn_1}} = \frac{\sigma}{\sqrt{n}}\left(\sqrt{1+\frac{n}{n_0+n_c}}\right)$, temos que os limites do gráfico de controle para a média μ com desvio padrão conhecido é dado por:

$$A(\alpha, T_c) = \left[\bar{x} + \frac{n_0}{n_0+n_c}(n_0 - \bar{x}) \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\left(\sqrt{1+\frac{n}{n_0+n_c}}\right)\right] \tag{4.24}$$

- ***ARL* para o gráfico de controle da média com desvio padrão conhecido**

Dado μ, a distribuição do comprimento da sequência é geométrica com parâmetro

$$\begin{aligned} \psi(\mu) &= 1 - \Phi\left(\frac{t_2-\mu}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{t_1-\mu}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(\frac{m_1-\mu}{\sigma/\sqrt{n}} + z_{1-\alpha/2}\sqrt{1+\frac{n}{n_1}}\right) + \Phi\left(\frac{m_1-\mu}{\sigma/\sqrt{n}} + z_{\alpha/2}\sqrt{1+\frac{n}{n_1}}\right) \end{aligned} \tag{4.25}$$

onde $\Phi(z)$ denota a função de distribuição acumulada da Normal padrão avaliada no ponto z. A média da distribuição preditiva do comprimento da sequência pode ser avaliada por integração numérica unidimensional, por exemplo,

$$\begin{aligned} E(r|\alpha,\bar{x}) &= \int_{\mu} \frac{1}{\psi(\mu)} n\left(\mu|m_1;\frac{\sigma^2}{n_1}\right) d\mu \\ &= \int_{z} \frac{\phi(z)}{1-\Phi\left(u_2-\sqrt{\frac{n}{n_1}}z\right)+\Phi\left(u_1-\sqrt{\frac{n}{n_1}}z\right)} dz, \end{aligned} \tag{4.26}$$

onde $u_1 = z_{\alpha/2}\sqrt{1+n/n_1} < 0$ e $u_2 = z_{1-\alpha/2}\sqrt{1+n/n_1} < 0$.

4.2.2 Caso 2: Média com desvio padrão desconhecido

Neste caso, o parâmetro de interesse é $\boldsymbol{\theta} = (\mu, \sigma^2)$. Quando o processo é estável, a variável resposta é gerada por uma distribuição normal com média μ e variância σ^2. Assume-se que a informação a priori para μ e σ^2 pode ser resumida pela distribuição normal-gama

$$h(\mu|\sigma^2) = N\left(m_0, \frac{\sigma^2}{n_0}\right) \quad e \quad h(\sigma^{-2}) = Ga\left(\frac{\nu_0}{2}, \frac{\nu_0}{2}s_0^2\right). \tag{4.27}$$

Seja $\mathbf{x} = (x_1, ..., x_{n_c})$ dados do processo estável, então $\mathbf{T}_c = (\bar{x}, s_x^2)$ é a estatística conjuntamente suficiente, e a distribuição $f(\mathbf{T}_c|\theta)$ é portanto $f(\bar{x}, s_x^2|\mu, \sigma^2) = f(\bar{x}|\mu, \sigma^2, s_x^2)$ $f(s_x^2|\mu, \sigma^2) = f(\bar{x}|\mu, \sigma^2) f(s_x^2|\sigma^2)$, onde $\bar{x}|\mu, \sigma^2 \sim N\left(\mu, \frac{\sigma^2}{n_c}\right)$ e $s_x^2|\sigma^2 \sim Ga\left[\left(\frac{n_c-1}{2}\right), \left(\frac{n_c-1}{2}\sigma^2\right)\right]$. Deste modo, temos que

$$h(\mu,\sigma^2) = h(\mu|\sigma^2)h(\sigma^{-2})$$

$$\begin{aligned} &= \frac{\sqrt{n_0}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n_0}{2\sigma^2}(\mu-m_0)^2\right] \frac{\left(\frac{\nu_0}{2}s_0^2\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1} \exp\left[-\frac{\nu_0}{2}s_0^2\frac{1}{\sigma^2}\right] \\ h(\mu,\sigma^2) &= \frac{\sqrt{n_0}}{\sqrt{2\pi}\sigma} \frac{\left(\frac{\nu_0}{2}s_0^2\right)^{\frac{\nu_0}{2}}}{\Gamma\left(\frac{\nu_0}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0(\mu-m_0)^2+\nu_0 s_0^2\right]\right\}, \end{aligned} \tag{4.28}$$

que é distribuição *a priori* conjunta para $\boldsymbol{\theta} = (\mu, \sigma^2)$ e a densidade conjunta da estatística suficiente é dada por

$$f(\bar{x}, s_x^2|\mu, \sigma^2) = f(\bar{x}|\mu, \sigma^2) f(s_x^2|\sigma^2)$$

$$\begin{aligned} &= \frac{\sqrt{n_c}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n_c}{2\sigma^2}(\mu - \bar{x})^2\right] \frac{\left(\frac{n_c-1}{2\sigma^2}\right)^{\frac{n_c-1}{2}}}{\Gamma\left(\frac{n_c-1}{2}\right)} \left(s_x^2\right)^{\frac{n_c-1}{2}-1} \exp\left[-\frac{(n_c-1)s_x^2}{2\sigma^2}\right] \\ &= \frac{\sqrt{n_c}}{\sqrt{2\pi}\sigma} \frac{\left(\frac{n_c-1}{2\sigma^2}\right)^{\frac{n_c-1}{2}}}{\Gamma\left(\frac{n_c-1}{2}\right)} \left(s_x^2\right)^{\frac{n_c-1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_c(\mu - \bar{x})^2 + (n_c-1)s_x^2\right]\right\} \quad (4.29) \end{aligned}$$

Agora, a distribuição a posteriori para $\boldsymbol{\theta} = (\mu, \sigma^2)$ é calculada da seguinte maneira:

$$h(\mu, \sigma^{-2}|\mathbf{T}_c) \propto h(\mu, \sigma^2) f(\bar{x}, s_x^2|\mu, \sigma^2) \quad (4.30)$$

$$\begin{aligned} &\propto \frac{1}{\sigma}\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[n_0(\mu - m_0)^2 + \nu_0 s_0^2\right]\right\} \frac{1}{\sigma}\left(\frac{n_c-1}{2\sigma^2}\right)^{\frac{n_c-1}{2}} \times \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left[n_c(\mu - \bar{x})^2 + (n_c-1)s_x^2\right]\right\} \\ &\propto \left(\frac{1}{\sigma^2}\right)^{\frac{n_c+\nu_0}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{n_0(\mu - m_0)^2 + n_c(\mu - \bar{x})^2}\right]\right\} \times \\ &\times \exp\left\{-\frac{1}{2\sigma^2}\left[\nu_0 s_0^2 + (n_c-1)s_x^2\right]\right\}, \quad (4.31) \end{aligned}$$

agora, utilizando-se a identidade dada por 4.12, temos que a expressão destacada em 4.31 pode ser escrita como

$$n_0(\mu - m_0)^2 + n_c(\mu - \bar{x})^2 = (n_c + n_0)\left(\mu - \frac{n_0 m_0 + n_c\bar{x}}{n_0 + n_c}\right)^2 + \frac{n_0 n_c}{n_0 + n_c}(m_0 - \bar{x})^2.$$

Deste modo,

$$h(\mu, \sigma^{-2}|\mathbf{T}_c) \propto \exp\left[-\frac{1}{2\frac{\sigma^2}{n_1}}(\mu - m_1)^2\right]\left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1} \exp\left(-\frac{\nu_1 s_1^2}{2}\frac{1}{\sigma^2}\right) \quad (4.32)$$

onde $n_1 = n_0 + n_c$, $m_1 = \frac{n_0 m_0 + n_c \bar{x}}{n_1}$, $\nu_1 = n_c + \nu_0$ e $\nu_1 s_1^2 = \nu_0 s_0^2 + (n_c - 1)s_x^2 + \frac{n_0 n_c}{n_0 + n_c}(m_0 - \bar{x})^2$.

Logo, pode-se concluir que

$$\mu|\mathbf{T}_c, \sigma^{-2} \sim N\left(m_1; \frac{\sigma^2}{n_1}\right) \quad e \quad \sigma^{-2}|\mathbf{T}_c \sim Ga\left(\frac{\nu_1}{2}; \frac{\nu_1 s_1^2}{2}\right). \tag{4.33}$$

Supondo que o processo permanece estável, pode-se agora construir os limites de controle para uma amostra futura $\mathbf{Y} = (Y_1, ..., Y_n)$, onde $\mathbf{T}_y = (\bar{y}, s_y^2)$ é a estatística conjuntamente suficiente. No entanto, Menzefricke (2002) apresentou o gráfico de controle apenas para a média, considerando apenas $T = \bar{y}$. Neste caso, dado μ e σ^2, a distribuição de $f(T|\theta)$ é então $\bar{y}|\mu, \sigma^2 \sim N(\mu; \sigma^2/n)$, e a distribuição preditiva *a posteriori* de $\bar{y}$ é calculada pela seguinte equação:

$$h(\bar{y}|\bar{x}, s_x^2) = \int_0^{+\infty} \int_{-\infty}^{+\infty} f(\bar{y}|\mu, \sigma^2) h(\mu, \sigma^{-2}|\bar{x}, s_x^2) \tag{4.34}$$

$$\begin{aligned}
&= \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\frac{\sigma^2}{n}}(\mu - \bar{y})^2\right] \frac{\sqrt{n_1}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\frac{\sigma^2}{n_1}}(\mu - m_1)^2\right] \times \\
&\times \frac{\left(\frac{\nu_1}{2} s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1} \exp\left(-\frac{\nu_1 s_1^2}{2}\frac{1}{\sigma^2}\right) d\mu d\sigma^{-2} \\
&= \int_0^{+\infty} \int_{-\infty}^{+\infty} \frac{\sqrt{n n_1}}{\sqrt{2\pi}\sqrt{2\pi}\sigma^2} \exp\left\{-\frac{1}{2\sigma^2}\left[\underbrace{n(\mu - \bar{y})^2 + n_1(\mu - m_1)^2}\right]\right\} \times \\
&\times \frac{\left(\frac{\nu_1}{2} s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1} \exp\left(-\frac{\nu_1 s_1^2}{2}\frac{1}{\sigma^2}\right) d\mu d\sigma^{-2},
\end{aligned} \tag{4.35}$$

utilizando-se a identidade 4.12, temos que a expressão destacada em 4.35 pode ser escrita como

$$n(\mu - \bar{y})^2 + n_1(\mu - m_1)^2 = (n + n_1)\left(\mu - \frac{n\bar{y} + n_1 m_1}{n + n_1}\right)^2 + \frac{n n_1}{n + n_1}(\bar{y} - m_1)^2.$$

Agora, fazendo $n_2 = n + n_1$ e $m_2 = \frac{n\bar{y} + n_1 m_1}{n + n_1}$, a Equação 4.34 pode ser escrita como segue

$$
\begin{aligned}
&= \int_0^{+\infty}\int_{-\infty}^{+\infty} \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{2\pi}\sigma^2} \exp\left[-\frac{n_2}{2\sigma^2}(\mu - m_2)^2 - \frac{nn_1}{2\sigma^2 n_2}(\bar{y} - m_1)^2\right] \times \\
&\times \frac{\left(\frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1} \exp\left(-\frac{\nu_1 s_1^2}{2}\frac{1}{\sigma^2}\right) d\mu d\sigma^{-2} \\
&= \int_0^{+\infty} \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{2\pi}\sigma^2} \exp\left[-\frac{nn_1}{2\sigma^2 n_2}(\bar{y} - m_1)^2\right] \frac{\left(\frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1} \times \\
&\times \exp\left(-\frac{\nu_1 s_1^2}{2}\frac{1}{\sigma^2}\right) \frac{\sqrt{2\pi}\sigma}{\sqrt{n_2}} d\sigma^{-2} \underbrace{\left\{\int_{-\infty}^{+\infty} \frac{\sqrt{n_2}}{\sqrt{2\pi}\sigma} \exp\left[-\frac{n_2}{2\sigma^2}(\mu - m_2)^2\right] d\mu\right\}}, \quad (4.36)
\end{aligned}
$$

a expressão em destaque em 4.36 corresponde à integral da f.d.p. de uma $N\left(m_2; \frac{\sigma^2}{n_2}\right)$, portanto, continuando o cálculo da Equação 4.34, temos que

$$
\begin{aligned}
h(\bar{y}|\bar{x}, s_x^2) &= \int_0^{+\infty} \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{n_2}\sigma} \exp\left[-\frac{nn_1}{2\sigma^2 n_2}(\bar{y} - m_1)^2\right] \frac{\left(\frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1} \times \\
\times \quad & \exp\left(-\frac{\nu_1 s_1^2}{2}\frac{1}{\sigma^2}\right) d\sigma^{-2} \\
= \quad & \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{n_2}} \frac{\left(\frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \int_0^{+\infty} \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1}{2}-1+\frac{1}{2}} \exp\left\{-\left[\frac{nn_1}{2n_2}(\bar{y} - m_1)^2 + \frac{\nu_1}{2}s_1^2\right]\frac{1}{\sigma^2}\right\} d\sigma^{-2} \\
= \quad & \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{n_2}} \frac{\left(\frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \frac{\Gamma\left(\frac{\nu_1+1}{2}\right)}{\left(\frac{nn_1}{2n_2}(\bar{y} - m_1)^2 + \frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1+1}{2}}} \int_0^{+\infty} \frac{\left(\frac{nn_1}{2n_2}(\bar{y} - m_1)^2 + \frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1+1}{2}}}{\Gamma\left(\frac{\nu_1+1}{2}\right)} \times \\
\times \quad & \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_1+1}{2}-1} \exp\left\{-\left[\frac{nn_1}{2n_2}(\bar{y} - m_1)^2 + \frac{\nu_1}{2}s_1^2\right]\frac{1}{\sigma^2}\right\} d\sigma^{-2},
\end{aligned}
$$

observe que o integrando da integral acima corresponde a f.d.p. de uma variável aleatória com distribuição $Ga\left[\frac{\nu_1+1}{2}; \left(\frac{nn_1}{2n_2}(\bar{y} - m_1)^2 + \frac{\nu_1}{2}s_1^2\right)\right]$, portanto

$$
h(\bar{y}|\bar{x}, s_x^2) = \frac{\sqrt{nn_1}}{\sqrt{2\pi}\sqrt{n_2}} \frac{\left(\frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)} \frac{\Gamma\left(\frac{\nu_1+1}{2}\right)}{\left(\frac{nn_1}{2n_2}(\bar{y} - m_1)^2 + \frac{\nu_1}{2}s_1^2\right)^{\frac{\nu_1+1}{2}}}, \quad (4.37)
$$

que evidencia o núcleo de uma distribuição *t-Student*. Logo,

$$\bar{y}|\bar{x}, s_x^2 \sim t\left[\nu_1, m_1, \left(\frac{1}{n}+\frac{1}{n_1}\right)s_1^2\right] \tag{4.38}$$

Uma região de aceitação de tamanho $1-\alpha$ é $A(\alpha, T_c) = (t_1, t_2)$, e de forma análoga os limites do gráfico de controle para a média, com o desvio padrão desconhecido, é dado por

$$A(\alpha, T_c) = \left[m_1 \pm t_{\alpha/2,\nu_1} s_1 \sqrt{\frac{1}{n}+\frac{1}{n_1}}\right]. \tag{4.39}$$

- ***ARL* para o gráfico de controle da média com desvio padrão desconhecido**

Dado μ e σ^2, a distribuição do comprimento da sequência é geométrica com parâmetro

$$\psi(\mu, \sigma^2) = 1 - \Phi\left(\frac{t_2-\mu}{\sigma/\sqrt{n}}\right) + \Phi\left(\frac{t_1-\mu}{\sigma/\sqrt{n}}\right), \tag{4.40}$$

a média da distribuição preditiva do comprimento da sequência pode ser avaliada por integração numérica bi-dimensional, por exemplo,

$$\begin{aligned} E(r|\alpha, T_c) &= \int_\mu \int_{\sigma^{-2}} \frac{1}{\psi(\mu,\sigma^2)} p(\mu|T_c, \sigma^2) p(\sigma^{-2}|T_c) d\mu d\sigma^{-2}, \\ &= \int_w \int_g \frac{n(w|0, g^{-1})\gamma(g|\frac{\upsilon_1}{2}, \frac{\upsilon_1}{2})}{1-\Phi(u_2)+\Phi(u_1)} dw dg \end{aligned} \tag{4.41}$$

onde $g = s_1^2\sigma^{-2}$ e $w = \frac{\sqrt{n_1}(\mu-m_1)}{s_1}$, temos ainda que $u_1 = \left(t_{\alpha/2,\upsilon_1}\sqrt{1+\frac{n}{n_1}} - \sqrt{\frac{n}{n_1}}w\right)\sqrt{g}$ e $u_2 = \left(t_{1-\alpha/2,\upsilon_1}\sqrt{1+\frac{n}{n_1}} - \sqrt{\frac{n}{n_1}}w\right)\sqrt{g}$.

Capítulo 5

Resultados e Avaliação

5.1 Gráfico de controle para processos da Família Exponencial

Os gráficos vistos anteriormente foram construídos sobre a suposição de que os dados são gerados por uma distribuição Normal. Neste Capítulo apresenta-se a construção dos gráficos de controle para uma distribuição discreta e uma contínua, pertencentes a Família Exponencial, que é um conjunto de distribuições muito utilizada nas mais diversas áreas, bem como a avaliação destes, feitos a partir da mesma metodologia mostrada no Capítulo anterior. Os resultados numéricos apresentados neste Capítulo foram obtidos a partir dos Métodos de Monte Carlo mostrados no Capítulo 3 e implementados no aplicativo Matlab 5.3.

5.1.1 Gráfico de Controle para Processos Poisson

Neste caso a variável resposta é gerada por uma distribuição Poisson(λ). Assume-se que a informação *a priori* pode ser resumida pela distribuição conjugada Gama

$$\lambda \sim Ga(a; b), \tag{5.1}$$

onde $a > 0$ e $b > 0$ são os chamados hiperparâmetros da distribuição *a priori*.

Suponha que $\mathbf{x} = (x_1, ..., x_{n_c})$ são dados de um processo Poisson, a estatística suficiente é $T_c = \sum_{i=1}^{n_c} x_i$. A distribuição de $f(T_c|\lambda)$ é então $\sum_{i=1}^{n_c} x_i|\lambda \sim Poi(n_c\lambda)$, e a distribuição *a posteriori* para λ é calculada da seguinte maneira

$$\begin{aligned} h(\lambda|\sum_{i=1}^{n_c} x_i) &\propto h(\lambda) f(\sum_{i=1}^{n_c} x_i|\lambda) \qquad (5.2)\\ &\propto \lambda^{a-1}\exp(-b\lambda)\exp(-n_c\lambda)(n_c\lambda)^{t_c} \\ h(\lambda|\sum_{i=1}^{n_c} x_i) &\propto \lambda^{a+t_c-1}\exp[-(b+n_c)\lambda]. \qquad (5.3) \end{aligned}$$

O núcleo de $h(\lambda|\sum_{i=1}^{n_c} x_i)$ mostra que,

$$\lambda|\sum_{i=1}^{n_c} x_i \sim Ga(a+t_c; b+n_c). \qquad (5.4)$$

Supondo que o processo continue sendo gerado por uma distribuição Poisson(λ), seja $\mathbf{y} = (y_1, ..., y_n)$ dados futuros deste processo, $T = \sum_{i=1}^{n} y_i$ é a estatística suficiente. A distribuição de $f(T|\theta)$ é então $\sum_{i=1}^{n} y_i|\lambda \sim Poi(n\lambda)$, e a distribuição preditiva *a posteriori* para $\sum_{i=1}^{n} y_i$ é calculada como segue

$$\begin{aligned} h\left(\sum_{i=1}^{n} y_i|\sum_{i=1}^{n_c} x_i\right) &= \int_0^{+\infty} \frac{(b+n_c)^{a+t_c}}{\Gamma(a+t_c)}\lambda^{a+t_c-1}\exp[-(b+n_c)\lambda] \times \\ &\times \frac{\exp(-n\lambda)(n\lambda)^t}{t!}d\lambda \\ &= \frac{(b+n_c)^{a+t_c}}{\Gamma(a+t_c)t!}n^t\int_0^{+\infty}\lambda^{a+t_c+t-1}\times \\ &\times \exp[-(b+n_c+n)\lambda]d\lambda \\ &= \frac{(b+n_c)^{a+t_c}}{\Gamma(a+t_c)t!}n^t\frac{\Gamma(a+t_c+t)}{(b+n_c+n)^{a+t_c+t}}\times \\ &\times \int_0^{+\infty}\frac{(b+n_c+n)^{a+t_c+t}}{\Gamma(a+t_c+t)}\lambda^{a+t_c+t-1}\exp[-(b+n_c+n)\lambda]d\lambda, \end{aligned}$$

observe que o integrando da integral acima corresponde a f.d.p. de uma $Ga(a+t_c+t; b+n_c+n)$, portanto, pode-se concluir que

$$h\left(\sum_{i=1}^{n} y_i|\sum_{i=1}^{n_c} x_i\right) = \frac{(b+n_c)^{a+t_c}}{(b+n_c+n)^{a+t_c+t}}\frac{\Gamma(a+t_c+t)}{\Gamma(a+t_c)t!}n^t. \qquad (5.5)$$

i) Intervalo *HPD* para processos Poisson

Como a distribuição de dada pela Equação 5.5 não é analiticamente tratável, pode-se utilizar os métodos de simulação Monte Carlo para quantidades preditivas (subseção 3.2.5) para encontrar uma estimativa de $h\left(\sum_{i=1}^{n} y_i | \sum_{i=1}^{n_c} x_i\right)$ e, posteriormente, utilizar os métodos de simulação Monte Carlo para intervalos de credibilidade (subseção 3.2.5) para construir um intervalo *HPD* para processos Poisson, baseados na distribuição preditiva, que serão os limites dos gráficos de controle para estes processos.

Deste modo, utilizando-se o método das quantidades preditivas retira-se uma amostra i.i.d. de $h(\lambda | \sum_{i=1}^{n_c} x_i)$ e assim, obtém-se uma amostra de $h\left(\sum_{i=1}^{n} y_i | \sum_{i=1}^{n_c} x_i\right)$. A partir desta amostra, pode-se agora encontrar uma estimativa para o intervalo *HPD*, utilizando-se o método mostrado na subseção 3.2.5. Portanto, para $m = 1.000$ amostras simuladas da distribuição preditiva e $\alpha = 5\%$, obteve-se 50 intervalos $\widehat{R}_{ip}(5\%) = (\lambda_{(i)}, \lambda_{(i+[950])}), 1 \leq i \leq 50$, onde o intervalo *HPD* é aquele que apresenta a menor amplitude. A Tabela 5.1 apresenta os resultados das simulações para o *HPD* para processos Poisson, com diferentes tamanhos da primeira amostra ($\mathbf{x}$), e para $a = b = 1$, $n = 10$ e $\alpha = 5\%$. Estes intervalos são os limites de controle para o gráfico em estudo, e serão usados para o cálculo do *ARL*.

Tabela 5.1 *Intervalos HPD (limites de controle) para Processos Poisson Baseados na Distribuição Preditiva a Posteriori para T.*

n_c	$\widehat{R}_p(\alpha)$
5	[2;23]
10	[3;22]
25	[4;19]
30	[4;17]

ii) Comprimento médio da sequência - *ARL*

Dado que o processo permaneça estável, o parâmetro da distribuição do comprimento da sequência, r_p, para processos Poisson é dado por

$$\psi_p(\lambda) = \sum_{\lambda_{(i)}}^{\lambda_{(i+[950])}} f(\sum_{i=1}^{n} y_i | \lambda), \tag{5.6}$$

onde, $\lambda_{(i)}$ e $\lambda_{(i+[950])}$ são, respectivamente, os limites inferior e superior do gráfico de controle para processos Poisson apresentados na Tabela 5.1.

Como foi visto anteriormente, $T = \sum_{i=1}^{n} y_i|\lambda \sim Poi(n\lambda)$, portanto a Equação 5.6 pode ser escrita como

$$\psi_p(\lambda) = \sum_{\lambda_{(i)}}^{\lambda_{(i+[950])}} \frac{\exp(-n\lambda)(n\lambda)^t}{t!}. \tag{5.7}$$

Pode-se então obter o ARL para processos Poisson calculando-se a esperança da distribuição do comprimento da sequência, r_p, da seguinte forma

$$E\left(r_p|\alpha, \sum_{i=1}^{n_c} x_i\right) = \int_{\Lambda} \frac{1}{1-\psi_p(\lambda)} h\left(\lambda|\sum_{i=1}^{n_c} x_i\right) d\lambda, \tag{5.8}$$

sabe-se que, $\lambda|t_c = \sum_{i=1}^{n_c} x_i \sim Ga(a+t_c; b+n_c)$, deste modo a Equação 5.8 pode ser escrita como

$$E\left(r_p|\alpha, t_c\right) = \int_{\Lambda} \frac{1}{1-\psi_p(\lambda)} \frac{(b+n_c)^{a+t_c}}{\Gamma(a+t_c)} \lambda^{a+t_c-1} \exp[-(b+n_c)\lambda] d\lambda, \tag{5.9}$$

que pode ser aproximada pelo método de Monte Simples (Equação 3.20), simulando-se uma amostra aleatória $(\lambda_1, ..., \lambda_m)$ da distribuição *a posteriori* dada em 5.4 obtendo-se

$$\widehat{E}\left(r_p|\alpha, t_c\right) \approx \frac{1}{m} \sum_{i=1}^{m} [1-\psi_p(\lambda_i)]^{-1}. \tag{5.10}$$

A Tabela 5.2 apresenta o ARL obtido a partir de 5.10, para Processos Poisson baseados na distribuição preditiva *a posteriori* para T para diferentes tamanhos amostrais da primeira amostra ($\mathbf{x}$) e para $n = 10$ (tamanho da amostra futura $\mathbf{y}$). Nela, pode-se verificar que, fixado o valor de $n = 10$, quanto maior o tamanho da primeira amostra, menor será o comprimento médio da sequência, ou seja, será necessário retirar, em média, uma quantidade de amostras menor, até que o gráfico sinalize pela primeira vez.

Tabela 5.2 *ARL para Processos Poisson Baseados na Distribuição Preditiva a Posteriori para T.*

n_c	ARL
5	155
10	78
25	26
30	16

iii) Avaliação do gráfico de controle para processos Poisson

Para a avaliação do gráfico de controle para processos Poisson, suponha que a média de $T = \sum_{i=1}^{n} y_i$, isto é, $E(T) = n\lambda$, sofra um acréscimo, δ, e passe a ser $E(T) = n\lambda + \delta$. A Figura 5.1 apresenta a probabilidade de que a estatística suficiente, $T = \sum_{i=1}^{n} y_i$, caia na região de rejeição, dado que houve esta mudança na média do processo. Fixado o valor de $\lambda = 1$ e $n = 10$, pode-se observar a partir desta figura que, a medida que o tamanho da primeira amostra é aumentado, aumenta também a probabilidade do gráfico detectar esta mudança no processo.

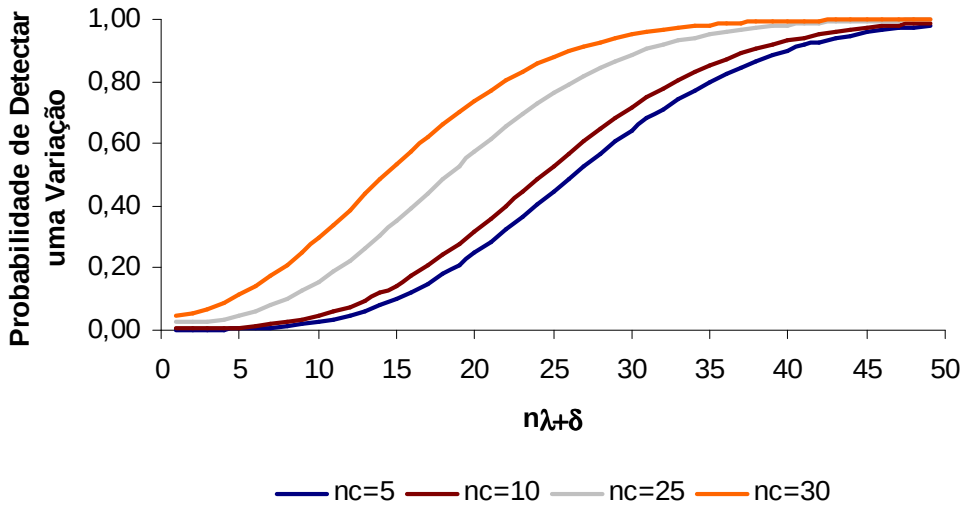

Figura 5.1 *Probabilidade de Rejeição do Gráfico de Controle para Processos Poisson.*

5.1.2 Gráfico de controle para processos Exponenciais

Neste caso a variável resposta é gerada por uma distribuição Exponencial(λ). Assume-se que a informação *a priori* pode ser resumida pela distribuição conjugada Gama

$$\lambda \sim Ga(a; b). \tag{5.11}$$

onde $a > 0$ e $b > 0$ são os chamados hiperparâmetros da distribuição *a priori*.

Suponha que $\mathbf{x} = (x_1, ..., x_{n_c})$ são dados de um processo Exponencial, a estatística suficiente é $T_c = \sum_{i=1}^{n_c} x_i$. A distribuição de $f(T_c|\lambda)$ é então $\sum_{i=1}^{n_c} x_i|\lambda \sim Ga(n_c; \lambda)$, e a distribuição *a posteriori* para λ é calculada da seguinte maneira

$$\begin{aligned} h(\lambda|\sum_{i=1}^{n_c} x_i) &\propto h(\lambda) f(\sum_{i=1}^{n_c} x_i|\lambda) \qquad (5.12)\\ &\propto \lambda^{a-1}\exp(-b\lambda)\lambda^{n_c}\exp(-\lambda t_c) \\ h(\lambda|\sum_{i=1}^{n_c} x_i) &\propto \lambda^{a+n_c-1}\exp[-(b+t_c)\lambda]. \qquad (5.13) \end{aligned}$$

O núcleo de $h(\lambda|\sum_{i=1}^{n_c} x_i)$ mostra que,

$$\lambda|\sum_{i=1}^{n_c} x_i \sim Ga(a+n_c; b+t_c). \tag{5.14}$$

Supondo que o processo continue sendo gerado por uma distribuição Exponencial(λ), seja $\mathbf{y} = (y_1, ..., y_n)$ dados futuros deste processo, $T = \sum_{i=1}^{n} y_i$ é a estatística suficiente. A distribuição de $f(T|\theta)$ é então $\sum_{i=1}^{n} y_i|\lambda \sim Ga(n; \lambda)$, e a distribuição preditiva *a posteriori* para $\sum_{i=1}^{n} y_i$ é calculada como segue

$$
\begin{aligned}
h\left(\sum_{i=1}^{n} y_i|\sum_{i=1}^{n_c} x_i\right) &= \int_0^{+\infty} \frac{(b+t_c)^{a+n_c}}{\Gamma(a+n_c)}\lambda^{a+n_c-1}\exp[-(b+t_c)\lambda]\times \\
&\times \frac{(\lambda)^n}{\Gamma(n)}t^{n-1}\exp(-\lambda t)d\lambda \\
&= \frac{(b+t_c)^{a+n_c}}{\Gamma(a+n_c)}\frac{t^{n-1}}{\Gamma(n)}\int_0^{+\infty}\lambda^{a+n_c+n-1}\exp[-(b+t_c+t)\lambda]d\lambda \\
&= \frac{(b+t_c)^{a+n_c}}{\Gamma(a+n_c)}\frac{t^{n-1}}{\Gamma(n)}\frac{\Gamma(a+n_c+n)}{(b+t_c+t)^{a+n_c+n}}\times \\
&\times \int_0^{+\infty}\frac{(b+t_c+t)^{a+n_c+n}}{\Gamma(a+n_c+n)}\lambda^{a+n_c+n-1}\exp[-(b+t_c+t)\lambda]d\lambda,
\end{aligned}
$$

observe que o integrando da integral acima corresponde a f.d.p. de uma $Ga(a+n_c+n; b+t_c+t)$, portanto, pode-se concluir que

$$
h\left(\sum_{i=1}^{n} y_i|\sum_{i=1}^{n_c} x_i\right) = \frac{(b+t_c)^{a+n_c}}{(b+t_c+t)^{a+n_c+n}}\frac{\Gamma(a+n_c+n)}{\Gamma(n)}t^{n-1}. \tag{5.15}
$$

i) Intervalo *HPD* para processos Exponenciais

Neste caso, também pode-se utilizar os métodos de simulação Monte Carlo para quantidades preditivas (subseção 3.2.5) para encontrar uma estimativa de $h\left(\sum_{i=1}^{n} y_i|\sum_{i=1}^{n_c} x_i\right)$ e, posteriormente, utilizar os métodos de simulação Monte Carlo para intervalos de credibilidade para construir um intervalo *HPD* para processos Exponenciais, baseados na distribuição preditiva, que serão os limites de controle para estes processos.

Deste modo, utilizando-se o método das quantidades preditivas retira-se uma amostra i.i.d. de $h(\lambda|\sum_{i=1}^{n_c} x_i)$ e assim, obtém-se uma amostra de $h\left(\sum_{i=1}^{n} y_i|\sum_{i=1}^{n_c} x_i\right)$. A partir desta amostra, pode-se agora encontrar uma estimativa para o intervalo *HPD*, utilizando-se o método mostrado na subseção 3.2.5. Portanto, para $m = 1.000$ amostras simuladas da distribuição preditiva e $\alpha = 5\%$, obteve-se 50 intervalos $\widehat{R}_{ie}(5\%) = (\lambda_{(i)}, \lambda_{(i+[950])}), 1 \leq i \leq 50$, onde o intervalo *HPD* é aquele que apresenta a menor amplitude. A Tabela 5.3 apresenta os resultados das simulações para o *HPD* para processos Exponenciais, com diferentes tamanhos da primeira amostra ($\mathbf{x}$), e para $a = b = 1$, $n = 10$ e $\alpha = 5\%$. Estes

intervalos são os limites de controle para o gráfico em estudo, e serão usados para o cálculo do *ARL*.

Tabela 5.3 *Intervalos HPD (limites de controle) para Processos Exponenciais Baseados na Distribuição Preditiva a Posteriori para T.*

n_c	$\widehat{R}_e(\gamma)$
5	[3,59;48,09]
10	[3,21;23,27]
25	[3,81;19,15]
30	[4,09;17,56]

ii) Comprimento médio da sequência - *ARL*

Dado que o processo permaneça estável, o parâmetro da distribuição do comprimento da sequência r_e para processos Exponenciais é dado por

$$\psi_e(\lambda) = \int_{\lambda_{(i)}}^{\lambda_{(i+[950])}} f(\sum_{i=1}^{n} y_i|\lambda) d(\sum_{i=1}^{n} y_i), \tag{5.16}$$

onde, $\lambda_{(i)}$ e $\lambda_{(i+[950])}$ são, respectivamente, os limites inferior e superior do gráfico de controle para processos Exponenciais apresentados na Tabela 5.3

Como foi visto anteriormente, sabe-se que $t = \sum_{i=1}^{n} y_i|\lambda \sim Ga(n;\lambda)$, deste modo a Equação 5.16 pode ser escrita como

$$\psi_e(\lambda) = \int_{\lambda_{(i)}}^{\lambda_{(i+[950])}} \frac{(\lambda)^n}{\Gamma(n)} t^{n-1} \exp(-\lambda t) dt. \tag{5.17}$$

Pode-se então obter o *ARL* para processos Exponenciais calculando-se a esperança da distribuição do comprimento da sequência r_e da seguinte forma

$$E\left(r_e|\alpha, \sum_{i=1}^{n_c} x_i\right) = \int_{\Lambda} \frac{1}{1-\psi_e(\lambda)} h\left(\lambda|\sum_{i=1}^{n_c} x_i\right) d\lambda, \tag{5.18}$$

como foi visto anteriormente, $\lambda|t_c = \sum_{i=1}^{n_c} x_i \sim Ga(a+n_c; b+t_c)$, deste modo a Equação 5.18 pode ser escrita como

$$E\left(r_e|\alpha, t_c\right) = \int_{\Lambda} \frac{1}{1-\psi_p(\lambda)} \frac{(b+t_c)^{a+n_c}}{\Gamma(a+n_c)} \lambda^{a+n_c-1} \exp[-(b+t_c)\lambda] d\lambda. \tag{5.19}$$

que pode ser aproximada pelo método de Monte Simples (Equação 3.20), simulando-se uma amostra aleatória $(\lambda_1, ..., \lambda_n)$ da distribuição *a posteriori* 5.14 obtendo-se

$$\widehat{E}\left(r_e|\alpha, t_c\right) \approx \frac{1}{m} \sum_{i=1}^{m} [1-\psi_e(\lambda_i)]^{-1}. \tag{5.20}$$

A Tabela 5.4 apresenta o *ARL* obtido a partir de 5.20, para Processos Exponenciais baseados na distribuição preditiva para diferentes tamanhos amostrais da primeira amostra ($\mathbf{x}$) e para $n = 10$ (tamanho da amostra futura $\mathbf{y}$). Nela, pode-se verificar que, fixado o valor de $n = 10$, quanto maior o tamanho da primeira amostra, menor será o comprimento médio da sequência, ou seja, será necessário retirar, em média, uma quantidade de amostras menor, até que o gráfico sinalize pela primeira vez.

Tabela 5.4 *ARL para Processos Exponenciais Baseados na Distribuição Preditiva a Posteriori para T.*

n_c	*ARL*
5	2.216
10	146
25	44
30	27

***iii*) Avaliação do gráfico de controle para processos Exponenciais**

Para a avaliação do gráfico de controle para processos Exponenciais, suponha que a média de $T = \sum_{i=1}^{n} y_i$, isto é, $E(T) = n/\lambda$, sofra um acréscimo, δ, e passe a ser $E(T) = n/\lambda + \delta$. A Figura 5.2 apresenta a probabilidade de que a estatística suficiente, $T = \sum_{i=1}^{n} y_i$, caia na região de rejeição, dado que houve esta mudança na média do

processo. Fixado o valor de $\lambda = 1$ e $n = 10$, pode-se observar a partir desta figura que, a medida que o tamanho da primeira amostra é aumentado, aumenta também a probabilidade do gráfico detectar esta mudança no processo.

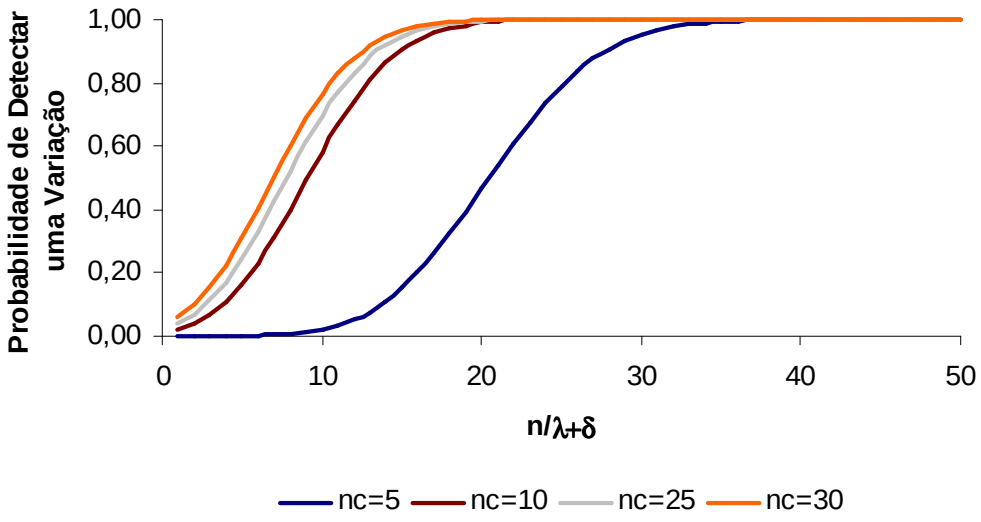

Figura 5.2 *Probabilidade de Rejeição do Gráfico de Controle para Processos Exponenciais.*

Capítulo 6

Conclusões e Recomendações

Este trabalho teve como principal objetivo desenvolver gráficos de controle bayesianos em alguns processos da Família Exponencial, baseados em distribuições preditivas de acordo com a metodologia proposta por Menzefricke (2002). A construção de gráficos de controle, tanto na abordagem bayesiana quanto na clássica, é feita sobre a suposição de que o processo em estudo apresenta distribuição Normal (ou pelo menos aproximadamente). A metodologia proposta neste trabalho mostra a construção de gráficos de controle para processos cuja distribuição pertença a Família Exponencial, mais especificamente, mostrou-se a construção e avaliação dos gráficos de controle baseados em distribuições preditivas para processo cuja distribuição seja Poisson ou Exponencial.

Para a construção e avaliação destes gráficos é necessário que se retire uma primeira amostra $\mathbf{X} = (X_1, ..., X_{n_c})$ de tamanho n_c para encontrar a distribuição *a posteriori* e, posteriormente retira-se uma segunda amostra $\mathbf{Y} = (Y_1, ..., Y_n)$ de tamanho n para encontrar a distribuição preditiva e a partir desta distribuição, construir um intervalo *HPD* para ser utilizado como os limites de controle para os gráficos em questão. Como a distribuição preditiva em ambos os casos, Poisson e Exponencial, não é analiticamente tratável, utilizou-se os métodos de simulação de Monte Carlo para quantidades preditivas e para intervalos de credibilidade, para encontrar os limites de controle para processos Poisson e Exponencial. Assumindo-se hipoteticamente que houve uma mudança na média do processo, verificou-se que ambos os gráficos apresentam sensibilidade para pequenas mudanças no processo, para $n = 10$ e para grandes valores de n_c. Contudo, a principal vantagem do método proposto neste trabalho em relação aos métodos clássicos é a redução de custos, já que são necessárias apenas 2 amostras para fazer o monitoramento do processo, enquanto que nos métodos clássicos costuma-se retirar pelo menos 20 ou 30 amostras para fazer tal monitoramento.

Portanto, os objetivos deste trabalho foram alcançados com êxito, e como recomendações para trabalhos futuros, podem-se destacar:

- Construção de gráficos de controle baseados em distribuições preditivas para outros processos da Família Exponencial;
- Investigar o comportamento dos gráficos apresentados neste trabalho na presença de valores extremos, a partir do ARL e da probabilidade de rejeição;
- Verificar a robustez dos gráficos propostos por Menzefricke (2002) a não normalidade do processo.

Bibliografia

Bayes, T. (1763). **An Essay Towards Solving in the Doctrine of Chances**. Philosophical Transactions of the Royal Society London 53, 370-418.

Bickel, P. J; Doksum, K. A. **Mathematical Statistical: Basic Ideas and Selected Topics**. Prentice-Hall. 2nd ed. ISBN 0-13-850363-X vol.1. New Jersey, 2001.

Bolfarine, H.; Sandoval, M. C. **Introdução à Inferência Estatística**. Rio de Janeiro: Sociedade Brasileira de Matemática, 2001.

Burr, I. J. **Effect of Nonnormality on Constants for $\bar{x}$ and R Charts**. Industrial Quality Control, vol. 23, 1967.

Bussab, W. O.; Morettin, P. A. **Estatística Básica**. São Paulo: Editora Saraiva. 5ª ed., 2003.

Casella, G.; Berger R. L. **Statistical Inference**. Duxbury Advanced Series. 2nd ed. ISBN 0-534-24312-6, 2002;

Chen, M.-H., Shao, Q.-M. e Ibrahim, J. G. **Monte Carlo Methods In Bayesian Computation**. Springer Series in Statistics, New York 2000.

Crosby, P. B.; **Quality Without Tears: The Art of Hassle Free Management.** New York: McGraw-Hill, 1995.

Deming, W. E.; **Out of Crisis.** Cambridge: MIT Press, 2000.

Ehlers, R. S. (2007) **Introdução à Inferência Bayesiana**. Disponível em <http://www.icmc.usp.br/ ehlers/notas/bayes.pdf>. Acesso em: 01/03/2009.

Filho, Ruy C. B. L.; **Controle Estatístico de Qualidade.** 2. Ed. São Paulo: Editora LTC, 1970.

Juran, J. M.; Godfrey, A. B. **Juran's Quality Handbook.** 5. Ed. New York: McGraw-Hill, 1999.

Menzefricke, U. **On the Evaluation of Control Chart limits Based on Preditive Distributions**. Communications in Statistics. Theory and Methods, ISSN 0361-0926, CODEN CSTMDC, 2002.

Montgomery, D. C. **Introdução ao Controle Estatístico da Qualidade**. New York:

John Wiley e Sons, 4ª Ed., 2001.

Paulino, C. D.; Turkman, M. A. A.; Murteira, B. **Estatística Bayesiana**.Fundação Calouste Gulbenkian, ISBN 972-31-1043-1, Lisboa 2003.

Ross, S. M. **Introduction to Probability Models**. Academic Press. 6th ed. ISBN 0-12-598470-7, USA 1977.

Tanner, M. A. (1996). **Tools for Statistical Inference**, 3rd ed. Springer Verlag, New York.