

UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE PRODUÇÃO

JOSÉ DE ALMEIDA JÚNIOR

**DETECÇÃO DE *OUTLIERS* COMO SUPORTE PARA CONTROLE ESTATÍSTICO
DO PROCESSO MULTIVARIADO: Um estudo de caso em uma empresa do
setor plástico**

João Pessoa
2013

JOSÉ DE ALMEIDA JÚNIOR

**DETECÇÃO DE *OUTLIER* COMO SUPORTE PARA O CONTROLE ESTATÍSTICO
DO PROCESSO MULTIVARIADO: Um estudo de caso em uma empresa do
setor plástico**

Dissertação submetida para apreciação da banca examinadora do Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal da Paraíba como parte dos requisitos necessários para obtenção do grau de mestre em Engenharia de Produção.

Professor orientador: Dr. Márcio Botelho da Fonseca Lima

Área de Concentração: Tecnologia, Trabalho, e Organização

Subárea: Gestão e tecnologia

**João Pessoa
2013**

A447d Almeida Júnior, José de

Detecção de *outlier* como suporte para o controle estatístico do processo multivariado: um estudo de caso em uma empresa do setor plástico. / José Almeida Júnior. – João Pessoa, 2013.

85f. il.:

Orientador: Prof. Dr. Márcio Botelho da Fonseca Lima

Dissertação (Curso de Pós-Graduação em Engenharia de Produção. Centro de Tecnologia - CT. Universidade Federal da Paraíba - UFPB.

1. Controle estatístico da qualidade 2. Carta T^2 de Hotelling 3. Análise de componentes principais 4. Algoritmo de busca sucessiva 5. Algoritmo BACON I. Título.

BS/CT/UFPB

CDU: 658.5(043)

JOSÉ DE ALMEIDA JÚNIOR

**DETECÇÃO DE *OUTLIER* COMO SUPORTE PARA O CONTROLE ESTATÍSTICO
DO PROCESSO MULTIVARIADO: Um estudo de caso em uma empresa do
setor plástico**

Dissertação julgada e aprovada em ____ de _____ de 2013 como parte dos requisitos necessários para a obtenção do título de Mestre em Engenharia de Produção no Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal da Paraíba.

BANCA EXAMINADORA

Prof. Dr. Márcio Botelho da Fonseca Lima-Orientador
Departamento de Engenharia de Produção – UFPB

Prof. Dr. Luiz Bueno da Silva - Examinador
Departamento de Engenharia de Produção – UFPB

Prof. Dr. Joab de Oliveira Lima – Examinador externo
Departamento de Estatística - UFPB

AGRADECIMENTOS

Esse trabalho de pesquisa foi bastante árduo e ao mesmo tempo prazeroso, pois recebi diversos apoios para que o objetivo fosse alcançado. Por esses apoios que tive, quero deixar o meu reconhecimento:

Primeiramente a Deus por ter dado força e persistência e por proporcionar condições para a realização deste trabalho.

A meus pais, pelo apoio moral, incentivo e participação fundamental em todos os momentos da minha vida.

A meus irmãos e familiares que direta ou indiretamente participaram e contribuíram para a realização do presente trabalho.

Ao Prof. Dr. Márcio Botelho da Fonseca Lima, orientador, pelo incentivo, motivação, presteza nos esclarecimentos, e apoio durante todo o processo acadêmico de dissertação, sempre disposto a ajudar sem medir qualquer esforço para que esse objetivo fosse alcançado.

Um agradecimento especial ao Prof. Dr. Joab de Oliveira Lima do departamento de estatística da UFPB pela disponibilidade, sugestões e orientações em todos os momentos que foi solicitado, atendendo sempre de modo muito gentil e dirimindo nossas dúvidas, contribuindo bastante para que esse trabalho de dissertação fosse concluído.

A coordenação do Programa de Pós graduação em Engenharia de Produção por procurar sempre atender da melhor forma nossas necessidades, nas pessoas da secretária Ana Araújo Silva, da vice coordenadora prof. Dr^a Maria de Lourdes Barreto Gomes e do Coordenador Prof. Dr. Luiz Bueno da Silva, o qual também aproveito a oportunidade para agradecer a disposição em participar da banca de avaliação.

A coordenação de aperfeiçoamento de pessoal de nível superior (CAPES), pelo apoio financeiro que foi de fundamental importância ao longo do processo acadêmico.

RESUMO

O projeto de pesquisa estudado teve o objetivo de aplicar um algoritmo de busca sucessiva para o auxílio à tomada de decisão no controle estatístico do processo multivariado, na fabricação de garrafeiras em uma empresa de produtos plásticos. Além disso, a utilização das técnicas de análise de componentes principais (ACP) e da carta T^2 de Hotelling pode sumarizar parte das informações relevantes desse processo. Produziram-se então dois resultados de considerável importância: os escores dos componentes principais e um gráfico T^2 de Hotelling adaptado, evidenciando a relação entre as dez variáveis analisadas. O algoritmo de busca sucessiva detecta pontos discordantes do restante do agrupamento de dados que, quando se encontram muito distantes ou têm características muito diferentes, são denominados *outliers*. O algoritmo BACON foi utilizado para a detecção de tais ocorrências, o qual parte de um pequeno subconjunto, comprovadamente livre de *outliers*, dos dados originais e vai adicionando novas informações, que também não são *outliers*, a esse subconjunto inicial até que nenhuma informação possa mais ser absorvida. Uma das vantagens da utilização desse algoritmo é que ele combate os fenômenos do mascaramento e do esmagamento que alteram as estimativas da média e da covariância. Os resultados da pesquisa mostraram que, para o conjunto de dados estudados, o algoritmo BACON não detectou nenhum ponto discordante. Uma simulação foi então desenvolvida, utilizando uma distribuição uniforme através da obtenção de números aleatórios dentro de um intervalo para a modificação dos valores da média e do desvio-padrão, a fim de mostrar que tal método é eficaz na detecção desses pontos aberrantes. Para essa simulação, foram alterados aleatoriamente os valores da média e do desvio-padrão de 5% dos dados originais. O resultado dessa simulação mostrou que o algoritmo BACON é perfeitamente aplicável ao caso estudado, sendo indicada a sua utilização em outros processos produtivos que dependam simultaneamente de diversas variáveis.

Palavras - Chave: Controle estatístico da qualidade. Carta T^2 de Hotelling. Análise de componentes principais. Algoritmo de busca sucessiva. Algoritmo BACON.

ABSTRACT

The research project studied, aimed to apply a forward search algorithm to aid decision making in multivariate statistical process control in the manufacture of crates in a company of plastic products. Besides, the use of principal components analysis (PCA) and the Hotelling T square chart can summarize relevant information of this process. Thus, they were produced two results of considerable importance: the scores of the principal components and an adapted Hotelling T square chart, highlighting the relationship between the ten variables analyzed. The forward search algorithm detects discordant points of the data clustering rest that, when are too far away or have very different characteristics, are called outliers. The BACON algorithm was used for the detection of such occurrences, which part of a small subset demonstrably free of the original data outliers and it goes adding new information, which is not outliers, to this initial subset until no information can more be absorbed. One of the advantages of using this algorithm is that it combats the masking and swamping phenomena that alter the mean and covariance estimates. The research results showed that, for the dataset studied, the BACON algorithm did not detected no dissenting point. A simulation was then developed, using a uniform distribution by obtaining random numbers within a range for modifying the mean and standard deviation values, in order to show that this method is effective in detecting these outliers. For this simulation, they were randomly changed 5% of the mean and the standard deviation values of the original data. The result of this simulation showed that the BACON algorithm is perfectly applicable to this case study, being indicated its use in other processes that simultaneously depend on several variables.

Keywords: Statistical quality control. Hotelling T square chart. Principal components analysis. Forward search algorithm. BACON algorithm.

LISTA DE SÍMBOLOS

Amplitude do gráfico de controle de Shewart	R
Base de dados contendo p variáveis	BD
Correlação entre duas variáveis	ρ
Covariância entre as variáveis x_i e x_j .	σ_{ij}
Distância de Mahalanobis	D^2
Estatística de Fisher-Snedecor	F
Estatística de Hotteling	T^2
Estatística qui-quadrado	χ^2
Estatística t de student	t
Fator de Correção para as discrepâncias	c_{np}
Função densidade de probabilidade de uma distribuição normal	$f(x)$
Graus de liberdade	v
Hipótese Alternativa	H_1
Hipótese nula	H_0
Matriz de Covariância do Processo	$\bar{S}$
Matriz de Variância e covariância	S^{-1}
Máximo valor residual que aponta a ocorrência de pontos discordantes	$\max \tau_i$
Média da distribuição normal	μ
Média das amplitudes do gráfico de Shewart	$\bar{R}$
Média das Médias	$\bar{\bar{x}}$
Média das medidas	$\bar{\bar{x}}$
Número de reamostragens dos vetores a serem alterados	B
Número de variáveis	p
Percentual de vetores a serem alterados, $0 \leq k \leq 1$	K
Probabilidade do erro tipo I	α
Probabilidade do erro tipo II	β
Subgrupos amostrais	m
Tamanho da amostra	n
Tamanho do subconjunto básico	r

Taxa de inflacionamento da média, $TM > 1$	TM
Taxa de inflacionamento do desvio padrão, $TD > 1$	TD
Valor tabelado que depende do tamanho da amostra para o cálculo do limite superior de controle do gráfico da amplitude	D_4
Valor tabelado que depende do tamanho da amostra para o cálculo do limite inferior de controle do gráfico da amplitude	D_3
Valor tabelado que depende do tamanho da amostra para o cálculo dos limites de controle do gráfico da média	A_2
Variância da distribuição normal	σ^2
Variância da variável x_i .	σ_{ii}
Variável aleatória contínua	Z
Vetor de média	$\bar{X}$

LISTA DE FIGURAS

Figura 1 - Gráfico de Shewhart sob controle	23
Figura 2 - Gráfico de Shewhart fora de controle	24
Figura 3 - Densidades de duas distribuições normais bivariadas com $\mu_1 = \mu_2 = 0$ e $\sigma_{12} = \sigma_{22} = 2$	27
Figura 4 - Elipse de controle para duas variáveis independentes	28
Figura 5 - Elipse de controle para duas variáveis dependentes	28
Figura 6 - Intervalos de confiança e região de confiança	34
Figura 7 - Direções de variabilidades máximas das componentes principais	36
Figura 8 - Diagrama de dispersão sobreposto à elipse de confiança	40
Figura 9 - Vista lateral	51
Figura 10 - Vista Superior	51
Figura 11 - Gráfico T^2 de Hotelling obtido no software R	57
Figura 12 - Variabilidade acumulada absorvida pelas componentes principais	59
Figura 13 - T^2 usando análise de componentes principais	59
Figura 14 - Novo gráfico T^2 de Hotelling para o modelo de ACP	60

LISTA DE QUADROS

Quadro 1 - Correlação das variáveis originais	58
Quadro 2 - Covariância do subconjunto básico inicial	62
Quadro 3 - Resultado do cálculo do ponto de corte	63

LISTA DE TABELAS

Tabela 1 - Modelo teórico para a tomada de decisão em relação ao controle estatístico de processo	25
Tabela 2 - Resultados do cálculo das distâncias de Mahalanobis	61
Tabela 3 - Vetor de média do subconjunto básico inicial	62
Tabela 4 - Distância de Mahalanobis adicionando o primeiro ponto ao subconjunto básico inicial	63
Tabela 5 - Desempenho do algoritmo BACON com aumento na média e no desvio-padrão em 5% dos vetores do conjunto de dados originais	65
Tabela 6 - Resultado que mostra a ocorrência de <i>outliers</i> com 50% de inflação da média e desvio padrão em 5% dos dados originais	66

SUMÁRIO

CAPÍTULO I - INTRODUÇÃO	13
1.1	DEFINIÇÃO DO TEMA.....13
1.2	JUSTIFICATIVA.....16
1.3	OBJETIVOS.....20
1.3.1	Objetivo Geral20
1.3.2	Objetivos Específicos20
CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA	21
2.1	GRÁFICO DE CONTROLE DE SHEWHART- UNIVARIADO.....22
2.2	DISTRIBUIÇÃO NORMAL MULTIVARIADA.....25
2.3	CONTROLE ESTATÍSTICO DO PROCESSO MULTIVARIADO27
2.4	GRÁFICOS DE CONTROLE MULTIVARIADO T^2 DE HOTELLING.....29
2.5	GRÁFICO DE CONTROLE QUI-QUADRADO (χ^2)32
2.6	CONSTRUÇÃO DA ELIPSE DE CONFIANÇA.....33
2.7	ANÁLISE DE COMPONENTES PRINCIPAIS35
2.8	OUTLIERS.....36
2.8.1	Outliers: caso univariado38
2.8.2	Outliers: caso multivariado39
2.9	O ALGORITMO BACON.....43
CAPÍTULO 3 - PROCEDIMENTOS METODOLÓGICOS	46
3.1	ANÁLISE DE EFICIÊNCIA DO GRÁFICO DE $\bar{X}$ VIA <i>AVERAGE RUN LENGTH</i> (ARL).....47
3.1.1	Erros dos tipos 1 e 2 e poder de detecção do gráfico de X barra47
3.1.2	Escolhas do intervalo de tempo entre amostras (h), do tamanho da amostra (n), de número de amostras (m) e separação das fases I e II ...49
3.2	ESCOLHA DAS VARIÁVEIS DE ESTUDO.....51
3.3	PONTOS FORA DOS LIMITES DE CONTROLE DO GRÁFICO DA MÉDIA NAS 25 PRIMEIRAS AMOSTRAS DE TAMANHO 4.....52
3.4	TESTES DE NORMALIDADE, ESPECIFICAÇÕES E COEFICIENTES DE CAPACIDADE DO PROCESSO.....53

3.5	PROCEDIMENTOS PARA A CONSTRUÇÃO DA ESTATÍSTICA T^2 DE HOTTELING	53
3.6	UTILIZAÇÃO DO ALGORITMO BACON PARA A DETECÇÃO DE OUTLIERS.....	54
CAPÍTULO 4 - ANÁLISE DOS RESULTADOS E DISCUSSÃO		56
4.1	APLICAÇÃO DO ALGORITMO BACON.....	61
4.2	UTILIZAÇÃO DO ALGORITMO BACON COMO SUPORTE AO CEP	64
4.3	APLICAÇÃO DO ALGORITMO BACON PARA UM ESTUDO SIMULADO ..	64
CAPÍTULO 5 - CONCLUSÕES E RECOMENDAÇÕES DE TRABALHOS FUTUROS		67
REFERÊNCIAS.....		70
APÊNDICE A – Pontos for a dos limites de controle nas 25 primeiras amostras de n=4		73
APÊNDICE B – Testes de normalidade, especificações e coeficientes de capacidade do processo		74
APÊNDICE C - Dados originais		75
APÊNDICE D - Rotina para a construção do gráfico T^2		78
APÊNDICE E - Rotina para o cálculo das distâncias de mahalanobis.....		81
APÊNDICE F - Subconjunto básico inicial		82
APÊNDICE G - Resultado das iterações.....		83

CAPÍTULO I - INTRODUÇÃO

1.1 DEFINIÇÃO DO TEMA

As fábricas de produtos plásticos estão a cada dia que passa ganhando mais espaço no cenário brasileiro. O crescimento desse mercado nos últimos anos é consequência da substituição gradativa dos materiais como vidro e metais por produtos plásticos, os quais proporcionam menores custos de produção e uma grande capacidade de serem reutilizados.

Segundo Medeiros (2005), os plásticos são um grupo de materiais artificiais de origem orgânica sintética que adquirem a condição plástica pela aplicação de calor e pressão durante a moldagem. São compostos de resinas sintéticas ou polímeros naturais modificados e possuem considerável resistência mecânica.

De acordo com a Associação Brasileira da Indústria do Plástico (ABIPLAST), em 2102, a indústria de transformação do plástico produziu cerca de R\$ 53,83 bilhões, transformando 6,66 mil toneladas de material plástico, o que resultou em um faturamento de R\$ 56,49 bilhões, aproximadamente 7% superior ao de 2011.

Esses números refletem a quantidade de empresas que se propuseram e se propõem a investir nesse setor que é caracterizado pela sua diversidade de produtos fabricados. Tais empresas produtoras têm uma capacidade de fabricar vários produtos de diversas formas e tamanho tais como: baldes, garrafeiras, sacolas, garrafas, entre diversos outros itens. Dentre os tipos de plásticos utilizados no processo produtivo desse tipo de fábrica, um dos mais utilizados é a resina termoplástica, cujo processo de formação de filme ocorre exclusivamente pela secagem física (evaporação de solventes).

A empresa objeto de estudo deste trabalho produz vários produtos plásticos podendo ser dividida em dois segmentos, sendo o primeiro representado pela linha de utilidades domésticas e o outro pela linha de embalagens. O produto analisado para a realização deste trabalho de pesquisa foi a caixa de refrigerantes, cujo processo de fabricação consiste na injeção de resina termoplástica em um molde específico, que determinará a forma do produto a ser obtido.

O equipamento utilizado para a obtenção desses produtos é a máquina injetora que controla variáveis tais como temperatura e pressão de trabalho, entre

outras. Nesse tipo de máquina a resina é depositada, mas anteriormente ela é misturada em uma betoneira com o objetivo de se obter a cor desejada do produto a ser produzido.

Já dentro da injetora, essa mistura sofre um aquecimento por um determinado período de tempo e logo após é levada ao molde por um cilindro de injeção que proporcionará a obtenção do produto desejado pelo fabricante.

Espera-se o período de resfriamento no molde e, só então, o produto é retirado da máquina. As caixas plásticas obtidas nesse processo servem para armazenamento e transporte das garrafas de refrigerantes. Essas caixas, também conhecidas como garrafeiras plásticas, podem ser empilhadas e podem também ser fornecidas com gravações personalizadas.

Para o controle e monitoramento desse tipo de processo, é necessário utilizar ferramentas que auxiliem na obtenção de produtos em conformidade com as especificações requeridas pela empresa. O Controle Estatístico de Processos (CEP) é uma ferramenta da qualidade utilizada nos processos produtivos com o intuito de se obter informações mais precisas para a prevenção da fabricação de produtos ou serviços defeituosos, auxiliando para que se atinja um aumento de produtividade e redução dos custos de produção. O CEP consiste em métodos para o monitoramento, compreensão e melhoria do desempenho do processo ao longo do tempo (WOODALL, 2000).

O objetivo principal do CEP é reduzir cada vez mais a variabilidade de um processo. Para isso lança-se mão de ferramentas que auxiliem no cumprimento desse objetivo, denominadas ferramentas do controle da qualidade, as quais permitem uma melhor visualização e entendimento da realidade do processo produtivo em termos de variabilidade. Essas ferramentas são as seguintes: Folha de verificação, histograma, gráfico de Pareto, diagrama de causa e efeito, diagrama de dispersão, diagrama de concentração de defeitos e gráficos de controle. Dessas ferramentas, os gráficos de controle são os mecanismos mais populares para o monitoramento de processos.

Os gráficos de controle foram propostos por Shewhart e nesse tipo de gráfico a análise é feita sob a perspectiva univariada, ou seja, supõe-se que há apenas uma variável no processo ou uma característica da qualidade de interesse. Mas, na prática, existem diversas variáveis no processo produtivo que precisam ser analisadas e, para isso, surgiu o controle estatístico de processo multivariado, que

tem a função de monitorar ou controlar simultaneamente duas ou mais características da qualidade.

O controle estatístico de processo multivariado visa dar maior confiabilidade na análise de diversas variáveis que se encontram nos processos produtivos que influenciam o desempenho da fabricação dos produtos e que culmine, consequentemente, com a melhoria do processo de tomada de decisão gerencial.

Os gráficos de controle multivariados são utilizados no monitoramento conjunto de duas ou mais variáveis do processo ou produto. Os métodos mais comuns de controle de qualidade multivariado são aqueles baseados no gráfico T^2 de Hotelling. Esses tipos de gráficos multivariados fornecem informações mais precisas do que aquelas relativas aos gráficos de Shewart, permitindo um poder de monitoramento mais eficaz na detecção de pontos fora da região de controle.

Alguns pontos fora dessa região de controle podem ser classificados como *Outliers*. *Outliers*, ou observações atípicas, são pontos ou observações que se encontram distante da maioria do agrupamento amostral de dados ou que apresentam alguma inconsistência com esse agrupamento. Hair *et al.* (2009) conceituam *outliers* como observações com uma combinação única de características identificáveis como sendo notavelmente diferentes das outras observações.

A detecção de *outliers* pode ser dividida em univariada ou multivariada. A detecção univariada de *outliers* examina a distribuição das observações para cada variável na análise e seleciona como *outliers* os casos que se enquadram nas faixas exteriores da distribuição (HAIR *et al.*, 2009). Os *outliers* multivariados são aquelas observações que apresentam uma posição bem distante das restantes no espaço p -dimensional definido por todas as variáveis.

Os *Outliers* são gerados por variabilidade no processo produtivo. A variabilidade é uma medida do grau de dispersão do resultado em torno do valor médio. Essa variação pode ser uma causa natural do processo que é fruto de uma série de pequenas perturbações e que são admissíveis. Mas também pode ser uma causa especial que é um problema ou modo de operação anormal do processo que pode ser corrigido ou eliminado.

Portanto, é necessário estabelecer métodos para o controle e a redução da variabilidade dos processos que proporcionam um processo estável. Então, para proporcionar essa redução da variabilidade, é necessário detectar observações

extremas, lançando mão de técnicas que estabeleçam uma região de corte, a partir da qual se pode afirmar a presença dessas observações atípicas que fazem com que o processo venha a ficar cada vez mais em descontrole.

Nesse contexto, o presente trabalho apresenta a seguinte questão de pesquisa: **Como se pode detectar a presença de *outliers* (multivariados) no processo produtivo em uma empresa do setor plástico?**

1.2 JUSTIFICATIVA

A maior abertura da economia brasileira, ocorrida principalmente nos anos 1990, proporcionou a entrada de diversas empresas multinacionais no país, o que resultou no aumento da concorrência. Uma consequência da formação desse cenário foi um maior nível de exigência dos clientes em termos de qualidade dos produtos. Os clientes passaram a serem menos tolerantes em relação a falhas nos produtos. Essa menor tolerância do mercado consumidor aumentou ainda mais com o surgimento de órgãos de defesa do consumidor.

No meio empresarial, o advento das parcerias entre empresas, firmadas com o objetivo de melhor gerenciar a cadeia de valor dos produtos, foi um fator adicional que contribuiu para o aumento da exigência de melhores níveis de qualidade.

Paladini (2004) cita cinco abordagens de qualidade (transcendental baseada no produto, baseada no usuário, baseada na produção e baseada no valor). Para esse autor, a qualidade baseada na produção é uma variável precisa e mensurável, oriunda do grau de conformidade do planejado com o executado. Ainda acrescenta que nessa abordagem dá-se maior ênfase a ferramentas estatísticas.

De fato, as exigências dos clientes com relação à qualidade no mercado de produtos e serviços se traduzem em termos dessas abordagens. A abordagem baseada na produção é crítica em muitas relações entre um cliente e um fornecedor, especialmente quando os clientes são empresas. A falta de conformidade pode produzir falhas nos produtos de empresas ou pode impossibilitar o uso de componentes e peças semiacabadas na sua produção. Problemas desse tipo geram perdas no meio empresarial e têm sido cada vez menos tolerados (LIMA *et al.*, 2012-a).

Mesmo quando as falhas em produtos não se propagam para outras empresas da cadeia empresarial ou para o cliente final, os defeitos por falta de

conformidade são uma preocupação, uma vez que geram perdas no processo produtivo.

Quando o foco da qualidade está na conformidade das características obtidas com aquelas planejadas antes da sua fabricação, surge a necessidade de utilização de um controle eficiente dessas características. É fundamental que sejam monitoradas com ferramentas adequadas para aquela situação.

Paladini (2004) lista três tipos básicos de causas de irregularidades no processo:

- Causa especial: é assinalável, imprevisível e, em geral, é também única. É suficientemente grande para produzir perturbações significativas no processo.
- Causas estruturais: têm características semelhantes às causas especiais, se distinguindo destas por ser periódica. Assim como as causas especiais, elas são elimináveis.
- Causas comuns: são relativamente pequenas, mas ocorrem quase sempre e em grande quantidade. É o acúmulo dessas causas ao longo do tempo que dá origem a aleatoriedade das variáveis. Em geral, essas causas não podem ser eliminadas por meio das ferramentas de CEP. A redução dessas causas requer um sacrifício considerável em termos de tempo e recursos, tornando necessária uma análise *off-line*, exigindo talvez a parada da linha de produção, investimentos em novas tecnologias, em máquinas e matérias primas melhores, um ambiente de trabalho apropriado e treinamento intensivo, fazendo com que a qualidade e custos caminhem juntos (SAMOHYL, 2009).

Assim, o objetivo do CEP é identificar as causas especiais e estruturais presentes no processo produtivo para que posteriormente possam ser eliminadas. Dessa forma, a sua utilização ao longo do tempo é capaz de melhorar o processo em termos de qualidade.

A título de exemplo, a chegada de um operador inexperiente, a não observância de procedimentos ou a ausência de procedimentos racionais para uma tarefa ou processo, bem como uma matéria prima fora das especificações e calibração inadequada de instrumentos de medição podem ser consideradas como fontes de variação ocasionada por causas especiais.

Portanto, a estatística mostra-se, cada vez mais, como uma poderosa ferramenta para a análise e avaliação de dados em várias áreas do conhecimento

que auxiliam a tomada de decisão permitindo uma maior e melhor utilização dos recursos/situações que foram submetidas à análise estatística. Para que essa análise seja mais rápida e completa, tal procedimento necessita de *softwares* cada vez mais difundidos, já que os dados precisam ser tratados mais eficientemente para se ter uma conclusão mais robusta sobre o estudo.

O processo produtivo do setor plástico utiliza máquinas e métodos que, quando não bem implantados e geridos, provocam diversas disfunções no processo produtivo, ocasionando uma alta variabilidade nas variáveis medidas dos seus produtos fabricados, e apontando a necessidade premente de detecção das causas dessa variação.

Para auxiliar no estudo dessa variabilidade, utiliza-se de métodos estatísticos com o intuito de se avaliar essas variáveis, seja de maneira isolada, chamada de análise univariada, seja para analisar essas variáveis de forma conjunta, chamada de análise multivariada.

Quando há um estudo que depende de muitas variáveis, geralmente a análise univariada não é satisfatória, pois a correlação entre as variáveis deve ser observada, sendo, portanto, insuficiente a realização de tal estudo fazendo apenas análise univariada das variáveis observadas. Desse modo, é imprescindível a utilização de técnicas estatísticas multivariadas para conhecer a totalidade dessas informações fornecidas pelo conjunto das variáveis analisadas.

Quando as relações entre as variáveis não são percebidas, efeitos desconhecidos entre as variáveis dificultam a interpretação do fenômeno a partir das variáveis consideradas. Esse tipo de técnica vem cada vez mais sendo explorada nos processos produtivos, pois, na prática, a maioria dos cenários de monitoramento e controle do processo envolve diversas variáveis relacionadas.

Diante disso, Montgomery (2009) afirma que, embora a aplicação de gráficos de controle univariados a cada variável individual seja uma solução possível, verifica-se que esse tipo de técnica não é eficaz e pode levar a conclusões errôneas, pois não se leva em conta a correlação entre as variáveis, sendo necessária então a utilização de uma ferramenta de controle mais satisfatória.

Essa ferramenta é o CEP multivariado, que vai facilitar a detecção das disfunções ocorridas no processo produtivo, diminuindo a probabilidade de produção de produtos defeituosos. Para isso, usa-se o gráfico T^2 de Hotelling, que tem uma

região de ação mais satisfatória para a sinalização de descontrole do processo, nos casos em que se trabalha com muitas variáveis simultaneamente.

Esse tipo de gráfico multivariado permite a visualização de pontos fora dos limites de controle que, antes nos gráficos univariados, não poderiam ser visualizados por tratar cada variável separadamente. Esses pontos podem ser classificados como *outliers* e trazer bastante prejuízo à análise realizada através do CEP.

Como os *outliers* são observações muito distantes do restante do intervalo de distribuição, a estatística T^2 deve ser objeto de estudos complementares nesses casos. Portanto, sugere-se a utilização de um algoritmo de busca sucessiva após a construção da estatística T^2 de Hotelling, para apontar com mais precisão a presença dessas observações insatisfatórias, estabelecendo uma região de corte a partir da qual seja possível indicar a presença de *outliers*.

O algoritmo BACON foi utilizado no presente trabalho, pois combate os fenômenos do esmagamento (*Swamping*) e do mascaramento (*Masking*), que interferem nas estimativas da média e da covariância do conjunto de dados analisados fazendo com que o gestor de produção seja induzido a chegar a conclusões errôneas sobre o desempenho do processo produtivo.

A presença dessas observações atípicas traz como consequência o desperdício, que é definido por Brinson (1996) como aquelas atividades que não agregam valor e que resultam em gastos de tempo, dinheiro e recursos sem lucro, além de adicionarem custos desnecessários aos produtos. Para Bornia (1995, p.5), “a produção de itens defeituosos implica em desperdiçar materiais, disponibilidade de mão-de-obra, disponibilidade de equipamentos, movimento e armazenagem de material defeituoso”.

Como o processo de fabricação das empresas do setor plástico trabalha com diversas variáveis simultaneamente para a obtenção de seus produtos, essas variáveis devem analisadas conjuntamente para a verificação de desvios no processo produtivo que impliquem na presença de *outliers* e consequentemente no descontrole do processo que venha a produzir produtos desconformes.

Nesse sentido, o presente trabalho se torna importante para os gestores de produção, pois facilita a tomada de decisão, fazendo com que a empresa obtenha a eliminação ou a diminuição significativa de desperdícios no seu processo produtivo,

ocasionando, conseqüentemente, menores custos de fabricação, tornando-se então mais competitiva frente às demais concorrentes.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Aplicar o algoritmo BACON para a detecção de *outlier* no processo produtivo em uma empresa do setor plástico.

1.3.2 Objetivos Específicos

- a) Construir a estatística T^2 de Hotelling;
- b) Construir o gráfico de Controle Multivariado;
- c) Aplicar análise de Componentes Principais;
- d) Utilizar um algoritmo de busca sucessiva para a detecção de *outliers*;
- e) Verificar a presença de *outliers*.

Nas linhas que se seguem, primeiramente, no capítulo 2, é apresentada a abordagem teórica, englobando noções de Controle Estatístico do Processo (CEP), o CEP multivariado, a Análise de Componentes Principais (ACP) e o gráfico T^2 de Hotelling e as abordagens teóricas acerca de observações atípicas. Em seguida, no capítulo 3, descreve-se a metodologia adotada, um estudo de caso. Logo após, no capítulo 4, são apresentados os resultados da pesquisa e a aplicação do algoritmo BACON e, finalmente, no capítulo 5, apresentam-se as conclusões e as sugestões de trabalhos futuros.

CAPÍTULO 2 - FUNDAMENTAÇÃO TEÓRICA

O (CEP) é uma ferramenta da qualidade que tem o objetivo de melhorar o processo produtivo através do monitoramento, o qual é realizado, principalmente, por gráficos de controle que têm a função de fornecer informações à tomada de decisão para a prevenção e detecção de problemas ocorridos no processo, colaborando assim para a diminuição dos desperdícios e dos custos.

O CEP utiliza várias técnicas estatísticas destinadas à obtenção de um melhor desempenho para o alcance de seus objetivos. Essas técnicas servem de base para que os produtos sejam cada vez mais fabricados com a qualidade requerida tanto pelos clientes quanto pela empresa. Tais técnicas requerem conhecimentos relativamente aprofundados de Estatística por parte do responsável pelo controle da qualidade na empresa, para que se possa obter o melhor resultado possível dessa ferramenta.

O monitoramento do processo e dos produtos fabricados é o principal objetivo do CEP, pois o intuito é o de deixar o processo livre de causas que impliquem em uma má qualidade no produto. Para isso, esse tipo de controle utiliza-se de algumas ferramentas para o auxílio na detecção das principais causas de controle do processo, tais como o gráfico de causa e efeito e o diagrama de Pareto.

Para Samohyl (2009, p.13) “o primordial desse tipo de controle consiste em obter processos capazes e estáveis, os quais valorizam o produto, e não dependem exclusivamente de inspeção para garantir o cumprimento das especificações exigidas”. No intuito de se atingir tal objetivo, é imprescindível a utilização dos gráficos de controle. Os gráficos de controle são ferramentas que auxiliam o monitoramento da variabilidade do processo ao longo do tempo.

Segundo Vieira (2012), o gráfico de controle serve para mostrar a variabilidade do processo, verificar se o processo está em controle, monitorar o processo de tal modo que causas especiais de variação sejam identificadas e corrigidas e verificar se as melhorias feitas determinaram menor variabilidade do processo. Os gráficos de controle devem ser implementados para monitorar as variáveis que influenciam a qualidade dos itens produzidos. Dessa forma, uma determinada variável de qualidade ou característica de qualidade é monitorada por meio de sucessivas amostras espaçadas no tempo (KONRATH, 2002).

2.1 GRÁFICO DE CONTROLE DE SHEWHART- UNIVARIADO

O gráfico de controle foi proposto inicialmente por Walter Shewhart como uma ferramenta para auxiliar na eliminação de variações anormais em processos que podem ser conhecidas como causas especiais ou atribuíveis. De acordo com Costa *et al.* (2004, p. 24) “as causas comuns são as variabilidades naturais do processo que são frutos de uma série de pequenas perturbações contra as quais pouco ou nada se pode fazer”. O mesmo autor ainda afirma que uma causa especial é um problema ou um modo de operação anormal do processo, que pode, portanto, ser corrigido ou eliminado.

Esses gráficos de controle podem ser construídos por variáveis ou por atributos. “Os gráficos de controle por atributos monitoram a variação do número ou da proporção de itens não conformes que ocorrem num processo de produção ao longo do tempo” (VIEIRA, 2012, p. 52). Já os gráficos de controle por variáveis monitoram a variação da média e da amplitude de um característico de qualidade ao longo do tempo. O gráfico por variáveis $\bar{X}$ monitora a centralidade e o da amplitude R monitora a dispersão. Os limites de controle do gráfico de R são calculados da seguinte forma:

$$\begin{aligned} \text{LSC} &= D_4 \bar{R} \\ \text{Linha Central} &= \bar{R} \\ \text{LIC} &= D_3 \bar{R} \end{aligned} \quad [1]$$

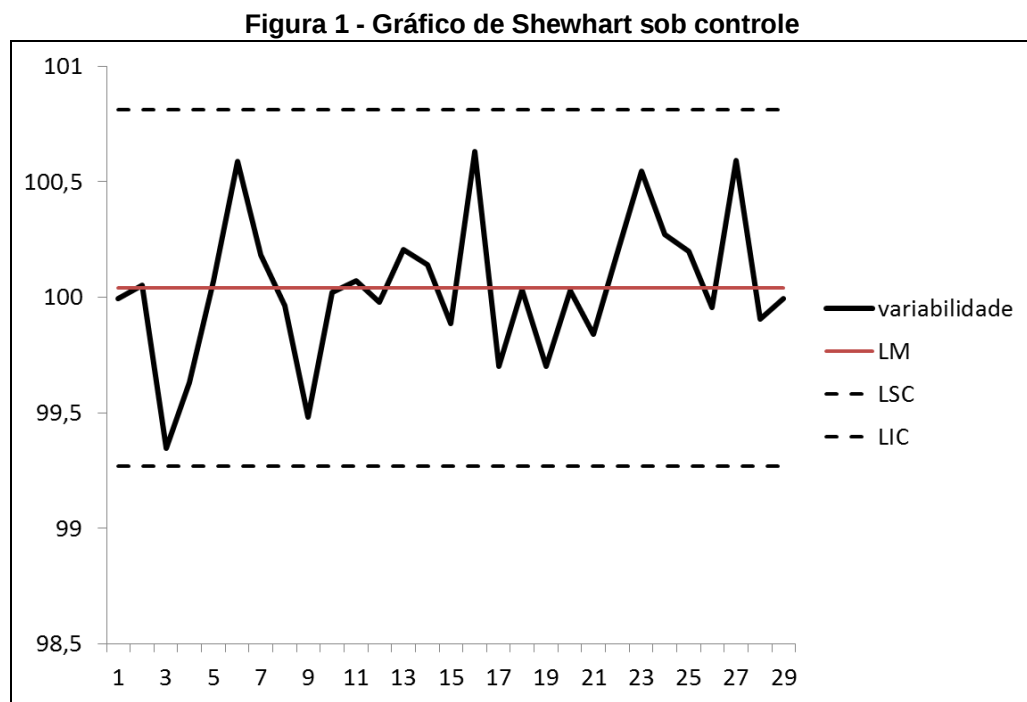
em que $\bar{R}$ é a média das amplitudes; D_4 e D_3 são valores tabelados que dependem do tamanho da amostra.

O cálculo dos limites de controle do gráfico de $\bar{X}$ pode ser obtido de acordo com as seguintes fórmulas:

$$\begin{aligned} \text{LSC} &= \bar{\bar{X}} + A_2 \bar{R} \\ \text{Linha Central} &= \bar{\bar{X}} \end{aligned} \quad [2]$$

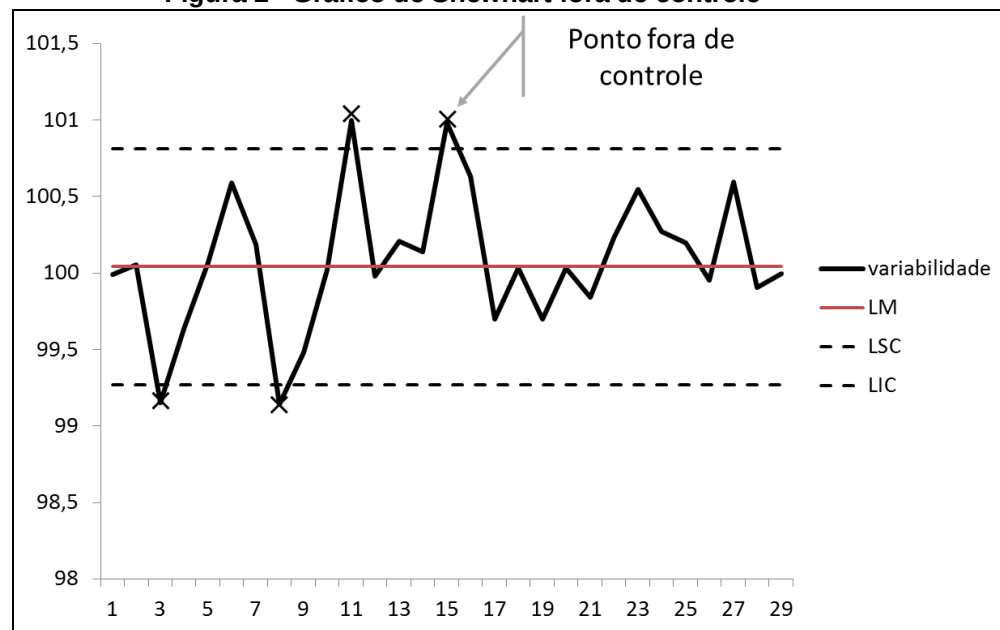
Sendo $\bar{\bar{X}}$ a média das médias amostrais, $\bar{R}$ a média das amplitudes e A_2 um valor tabelado que depende do tamanho da amostra.

A Figura 1 mostra um gráfico de controle por variáveis que está livre de causas especiais. A abcissa (x) mostra a ordem cronológica de retirada das amostras; a ordenada (y) representa os valores observados da característica de qualidade; a linha média (LM) mostra o valor médio da característica de qualidade quando o processo está em controle. E os limites de controle LCS e LCI são as fronteiras segundo as quais os pontos amostrais devem estar entre elas para que se considere o processo estável. Já a Figura 2 apresenta um processo fora de controle sob a influência de causas especiais, podendo-se visualizar vários pontos fora dos limites de controle.



Fonte: Elaboração própria.

Figura 2 - Gráfico de Shewhart fora de controle



Fonte: Elaboração própria.

Os limites do gráfico de controle proposto por Shewart têm abertura de ± 3 desvios-padrão distanciados da média do processo, ou seja, de $\bar{X}$. Esses limites vão definir a região de ação do gráfico. O intuito é que o gráfico alerte o momento de se intervir no processo.

Um ponto nessa região de ação do gráfico pode ser um alarme falso, ou seja, ocorre o risco de se intervir no processo quando ele está em controle. E essa probabilidade de se intervir erroneamente no processo quando o processo está em controle é de 0,27%, implicando que o processo produtivo não deve rejeitar mais que 27 itens em cada 10.000 produzidos.

Costa *et al.* (2004) ainda afirmam que um ponto acima ou abaixo dos limites do gráfico de controle é interpretado como sinal de que o processo deve estar fora de controle e alguma ação corretiva é necessária, ou seja, deve-se proceder a uma investigação para identificar as causas especiais e intervir para eliminá-las.

O modelo de tomada de decisão, em relação ao teste estatístico de verificar se a média do processo está centrada no alvo da meta (chamada de **Hipótese Nula** – H_0) do sistema de produção, pode ser idealizado utilizando-se da teoria estatística conhecida como Teste de Hipótese, que tem a seguinte estrutura esquemática:

Tabela 1 - Modelo teórico para a tomada de decisão em relação ao controle estatístico de processo

Decisão estatística sobre a situação do Processo	Verdadeira situação do processo	
	Está sob Controle (H_0 é verdadeira)	Está fora de Controle (H_0 é falsa)
Fora de Controle (Rejeita H_0)	Decisão Incorreta (Cometeu-se o Erro tipo I, chamado de α)	Decisão Correta ($1 - \alpha$)
Em Controle (Não Rejeita H_0)	Decisão Correta ($1 - \beta$)	Decisão Incorreta (Cometeu-se o Erro tipo II, chamado de β)

Fonte: Elaboração própria.

Se o processo estiver em controle (H_0 verdadeira), α representa o risco (probabilidade) de erroneamente considerar-se o processo fora de controle (alarme falso). Se o processo estiver fora de controle (H_1 verdadeira), β representa o risco (probabilidade) de erroneamente considerar-se o processo em controle (não detecção) (COSTA *et al.*, 2004).

Nesses gráficos propostos por Shewhart, a análise é feita sob a perspectiva univariada, ou seja, supõe-se que há apenas uma variável de estudo em um processo ou uma característica da qualidade de interesse. Mas, na prática, existem diversas variáveis no controle do processo que precisam ser analisadas e, para isso, surgiu o controle estatístico do processo multivariado, que tem a função de monitorar ou controlar simultaneamente duas ou mais características da qualidade.

2.2 DISTRIBUIÇÃO NORMAL MULTIVARIADA

A função densidade de probabilidade de uma distribuição contínua serve para informar a probabilidade de uma variável assumir um valor dentro de um intervalo. A função densidade de probabilidade de uma distribuição normal univariada é dada por

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}$$

[3]

Em que μ e σ^2 são, respectivamente, a média da distribuição normal e a variância. Reescrevendo o termo expoente, tem-se: $(x - \mu)(\sigma^2)^{-1}(x - \mu)$, que é a distância padronizada ao quadrado de x à média μ .

A distribuição Normal multivariada é uma generalização da univariada que pode ser escrita da seguinte forma:

$$f(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp^{-\frac{1}{2}(x-\mu)' \Sigma^{-1}(x-\mu)} \quad [4]$$

Em que p é o número de variáveis e μ' e Σ são, respectivamente, o vetor das médias dos x 's e a matriz de covariância de dimensão $p \times p$. Então, a distância generalizada padronizada ao quadrado entre x e μ' é:

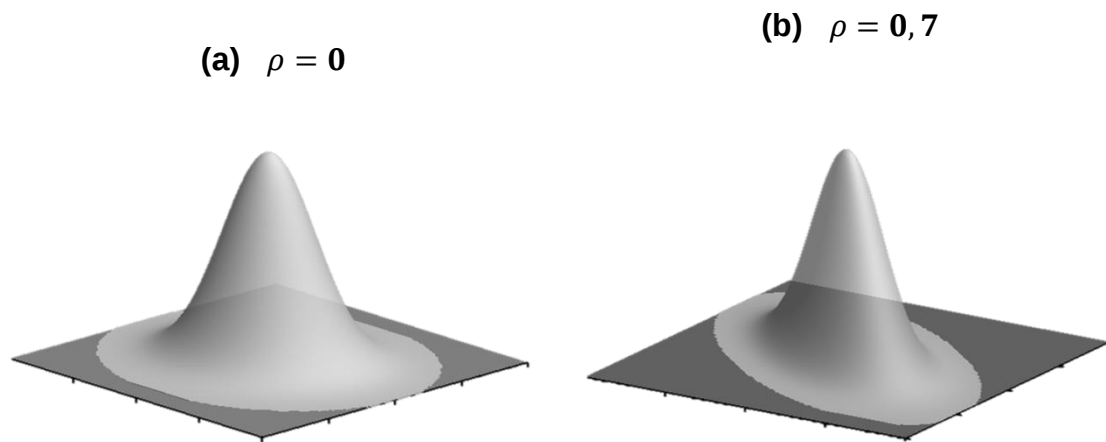
$$(x - \mu')' \Sigma^{-1} (x - \mu')$$

A Figura 3 mostra uma distribuição normal multivariada para duas variáveis. A correlação entre essas duas variáveis é medida por ρ , que é:

$$\rho = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}} \quad [5]$$

Em que σ_{ii} e σ_{ij} são, respectivamente, a variância da variável x_i e a covariância entre as variáveis x_i e x_j com $i \neq j$. Nos casos em que $\rho = 0$, ou seja, não haja correlação entre as variáveis, as curvas de níveis são representadas por círculos como mostra a Figura 3 (a). E nos casos em que $\rho \neq$ zero, as curvas de níveis são representadas por elipse, conforme a Figura 3 (b). Quando $p > 2$, as curvas de níveis são elipsoides para $\rho \neq$ zero e esferas para $\rho = 0$ (COLENGHI, 2008).

Figura 3 - Densidades de duas distribuições normais bivariadas com $\mu_1 = \mu_2 = 0$ e $\sigma_1^2 = \sigma_2^2 = 2$



Fonte: Elaboração própria.

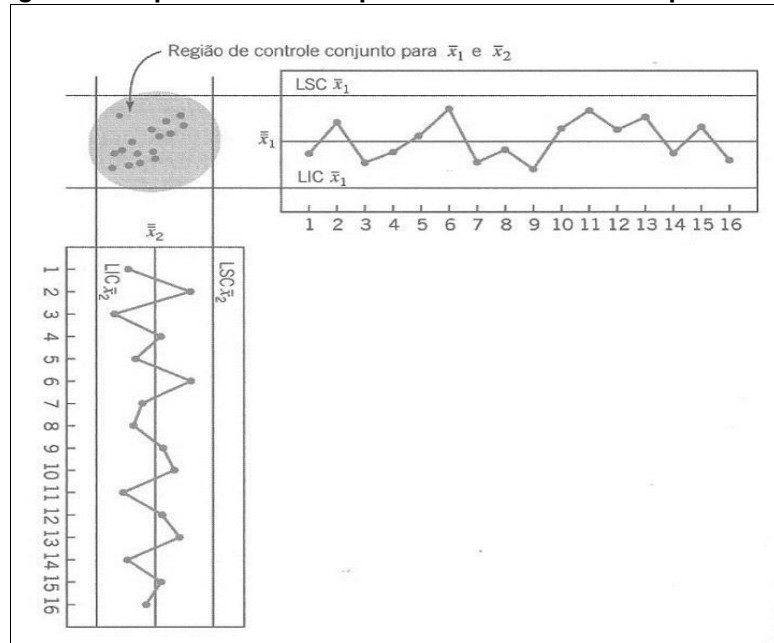
2.3 CONTROLE ESTATÍSTICO DO PROCESSO MULTIVARIADO

O controle estatístico do processo multivariado é análogo ao univariado, no entanto é utilizado para o monitoramento de variáveis correlacionadas. De acordo com Simoglou *et al.* (2000), o controle estatístico de processo multivariado é uma ferramenta utilizada para o acompanhamento global do desempenho do processo de manufatura. O objetivo, segundo Treasure *et al.* (2004), é eliminar frequentemente o grau observado de elevada redundância nos dados, definindo um conjunto reduzido de variáveis estatisticamente não correlacionadas.

Nos gráficos univariados, o processo é considerado em controle se as médias amostrais, por exemplo, de duas variáveis X_1 e X_2 se encontram dentro dos seus limites de controle e que a probabilidade de uma das variáveis exceder os limites de controle de três-sigma é 0,0027. Já nos gráficos multivariados, que é bem representado por uma elipse, a probabilidade de um ponto cair fora dos limites de controle entre as duas variáveis X_1 e X_2 analisadas simultaneamente é $(0,0027) \times (0,0027) = 0,00000729$.

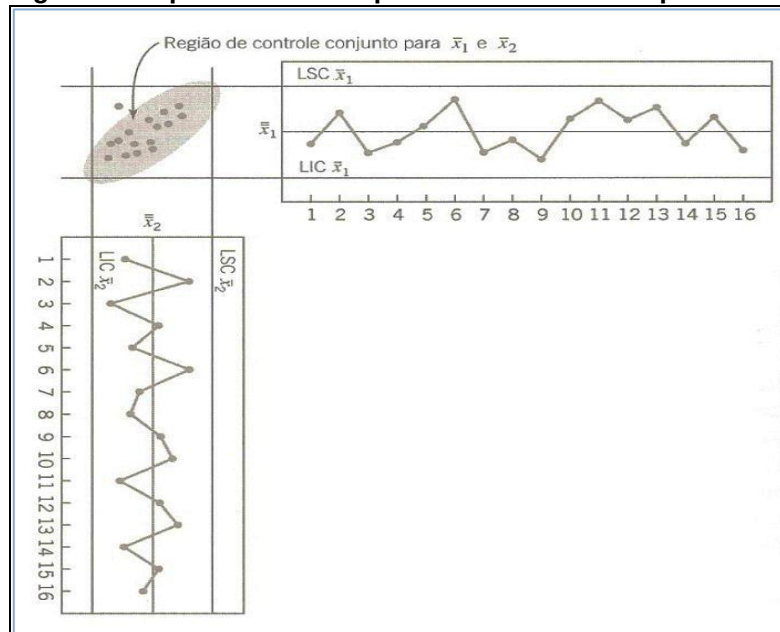
Respectivamente, as Figuras 4 e 5 a seguir comparam o desempenho dos gráficos analisando as regiões de controle conjunto das variáveis independentes e dependentes, na detecção dos pontos fora dos limites de controle.

Figura 4 - Elipse de controle para duas variáveis independentes



Fonte: Montgomery (2009).

Figura 5 - Elipse de controle para duas variáveis dependentes



Fonte: Montgomery (2009).

Uma das grandes preocupações desse tipo de controle é a forma eficaz e eficiente de utilizar a grande quantidade de dados para caracterizar o processo. (CHEN, 2010). Conforme afirma Montgomery (2009), problemas de monitoramento de processo nos quais o interesse recai sobre várias variáveis relacionadas são, geralmente, chamados de problemas de controle de qualidade multivariado (ou de

monitoramento do processo). Esse monitoramento simultâneo é realizado através do gráfico T^2 de Hotelling, que é análogo ao gráfico de Shewhart.

2.4 GRÁFICOS DE CONTROLE MULTIVARIADO T^2 DE HOTELLING

O gráfico T^2 de Hotelling é uma carta de monitoramento para o vetor de médias que permite o controle simultâneo das médias para diversas variáveis interdependentes. Esse tipo de gráfico é também uma ferramenta de detecção de *outliers* multivariados, de mudança nas médias e de outros desvios da distribuição (WILLIAMS *et al.*, 2006).

O gráfico T^2 de Hotelling é um dos gráficos mais utilizados como ferramenta de controle estatístico do processo multivariado, nos casos em que o vetor de médias e a matriz de covariância são desconhecidos, necessitando de serem estimados. Rencher (2002) enumera algumas propriedades importantes da estatística T^2 , quais sejam:

1. Deve-se ter $n - 1 > p$. Caso contrário, S é singular e T^2 não pode ser computado;
2. Em ambos os casos, com uma amostra e duas amostras, o número de graus de liberdade da estatística T^2 será análogo ao teste t - univariado, com $v = n - 1$ para uma amostra e $v = n_1 + n_2 - 2$ para duas amostras;
3. A hipótese alternativa é bilateral, porque o espaço é multidimensional, por isso não se considera a hipótese alternativa unilateral tal como $\mu > \mu_0$. Contudo, embora a hipótese alternativa $H_1 = \mu \neq \mu_0$ seja essencialmente bilateral, a região crítica é unicaudal (Rejeita-se H_0 para grandes valores de T^2). Isso é típico de muitos testes multivariados;
4. No caso univariado, $t_{n-1}^2 = F_{1;n-1}$. A estatística T^2 também pode se convertida para a estatística F como se segue:

$$\frac{v - p + 1}{vp} T_{p,v}^2 = F_{p; v-p+1} \quad [6]$$

A estatística T^2 é derivada da estatística t univariada. A evolução para a T^2 funciona da seguinte forma:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \quad [7]$$

Elevando-se ao quadrado, a t univariada pode ser reescrita da seguinte forma:

$$t^2 = \frac{n(\bar{X} - \mu_0)^2}{S^2} = n(\bar{X} - \mu_0)(S^2)^{-1}(\bar{X} - \mu_0) \quad [8]$$

Substituindo $(\bar{X} - \mu_0)$ por $(\bar{X} - \bar{\bar{X}})$ e S^2 por $\bar{S}$, tem-se a seguinte expressão:

$$T^2 = n(\bar{X} - \bar{\bar{X}})^T \bar{S}^{-1}(\bar{X} - \bar{\bar{X}}) \quad [9]$$

Em que:

$\bar{\bar{X}}$ é a estimativa para o vetor de médias e pode ser calculado da seguinte forma:

$$\bar{\bar{X}} = \frac{1}{m} \sum_{k=1}^m \bar{X}_k \quad [10]$$

Sendo $\bar{S}$ a matriz de covariância do processo, que pode ser obtida da seguinte forma:

$$\bar{S} = \frac{1}{m} \sum_{k=1}^m S_k \quad [11]$$

Em que m são os subgrupos amostrais preliminares de tamanho n retirados do processo.

A distribuição dessa estatística T^2 é dividida em duas fases, 1 e 2. Quanto à primeira fase, ou seja, a fase 1, consiste no uso dos gráficos para o estabelecimento do controle; isto é, para testar se o processo estava, ou não, sob controle quando os m subgrupos preliminares foram extraídos e as estatísticas amostrais $\bar{X}$ e S calculadas. (MONTGOMERY, 2009). Esses limites de controle para a fase 1 são obtidos da seguinte forma:

$$LSC = \frac{p(m-1)(n-1)}{mn - m - p + 1} F_{\alpha; p; mn-m-p+1} \quad [12]$$

$$LIC = 0$$

Em que:

- p : é a quantidade de variáveis;
- m : representa os subgrupos preliminares extraídos;
- n : é o tamanho do subgrupo;
- $F_{\alpha; p; mn-m-p+1}$: é um ponto de uma parte da percentagem superior da distribuição F com p e $mn - m - p + 1$ graus de liberdade, se o tamanho da amostra for $n > 1$ (CHEN ; HSIEH, 2007).

Na fase 2, o gráfico de controle é utilizado para monitorar a produção futura. Isso quer dizer que pontos fora da elipse ou do LSC sugerem que os valores das características de qualidade diferentes são de algum modo as do processo anterior quando estável (JOHNSON; WICHERN, 2007). Esses limites podem ser calculados pelas seguintes equações:

$$\text{LSC} = \frac{p(m+1)(n-1)}{mn-m-p+1} F_{\alpha, p, mn-m-p+1} \quad [13]$$

$$\text{LIC} = 0$$

Para observações individuais, ou seja, para $n = 1$, a equação para a estatística T^2 se torna:

$$T^2 = n(X - \bar{X})^T S^{-1} (X - \bar{X})$$

Os limites de controle para a fase 2 são os seguintes:

$$\text{LSC} = \frac{p(m+1)(m-1)}{m^2 - mp} F_{\alpha, p, m-p} \quad [14]$$

$$\text{LIC} = 0$$

Segundo Montgomery (2009, p. 331), “Aproximações para os limites da fase 1 com base na distribuição F serão, provavelmente, imprecisas”. Diante disso, surge a necessidade de se pesquisar outros tipos de distribuição para o estabelecimento dos limites de controle da fase 1.

2.5 GRÁFICO DE CONTROLE QUI-QUADRADO (χ^2)

O gráfico de controle qui-quadrado é uma alternativa para processos multivariados quando o vetor de médias μ e a matriz de covariância Σ são conhecidos, ou seja, quando o vetor de médias μ e a matriz de covariância Σ são estimados a partir de um grande número de amostras, o cálculo dos limites de controle para as fases 1 e 2 são realizados da seguinte forma:

$$\boxed{LSC = \chi_{\alpha; p}^2} \quad [15]$$

$$LSC = 0$$

Em que $\chi_{\alpha; p}^2$ é o ponto percentual α superior da distribuição qui-quadrado com p graus de liberdade.

Para se chegar ao cálculo desses limites de controle, é necessário observar as equações que servem de base para tal cálculo:

$$\boxed{Z = \frac{(\bar{X} - \mu_0)\sqrt{n}}{\sigma_0}} \quad [16]$$

Sendo Z uma variável aleatória contínua com distribuição normal padronizada $N(0,1)$. Portanto, elevando-se essa expressão ao quadrado pode-se obter a distribuição Qui-quadrado com um grau de liberdade como mostra a equação abaixo:

$$\chi_1^2 = \frac{(\bar{X} - \mu_0)^2 n}{\sigma_0^2} = \chi_1^2 = n(\bar{X} - \mu_0)(\sigma_0^2)^{-1}(\bar{X} - \mu_0)$$

Então, a estatística do teste plotada no gráfico de controle para cada amostra é:

$$\boxed{\chi_p^2 = n(\bar{X} - \mu_0)^T \Sigma^{-1}(\bar{X} - \mu_0)} \quad [17]$$

Essa expressão é a distância padronizada entre um vetor de média $\bar{X}$ e o vetor de média do processo μ_0 .

O gráfico χ^2 sinaliza um alarme como fora de controle no tempo t , se a estatística χ^2 é maior que LSC dado por um $\chi^2_{\alpha;p}$, em que α é o nível de significância apropriado para a realização do teste. (GUH ; SHIUE, 2008).

O procedimento de monitoramento para duas variáveis correlacionadas pode ser mais bem representado por uma elipse de controle por ser mais fácil de detectar causas atribuíveis, ou seja, pontos fora da elipse de controle, como foi mostrado na Figura 5.

Montgomery (2009) destaca duas desvantagens relacionadas à elipse de controle. A primeira é que se perde a sequência temporal dos pontos plotados. A segunda é a dificuldade de se construir a elipse para mais de duas características de qualidade. Para evitar essa dificuldade, deve-se plotar os valores de χ_p^2 calculados para cada amostra em um gráfico de controle com apenas um limite superior $\chi^2_{\alpha;p}$.

De acordo com Chen e Wang (2004), a estatística Qui-quadrado é um dos métodos mais aplicados para superar as dificuldades encontradas na análise univariada, traçando uma estatística que mede os desvios globais das observações multivariadas do seu alvo.

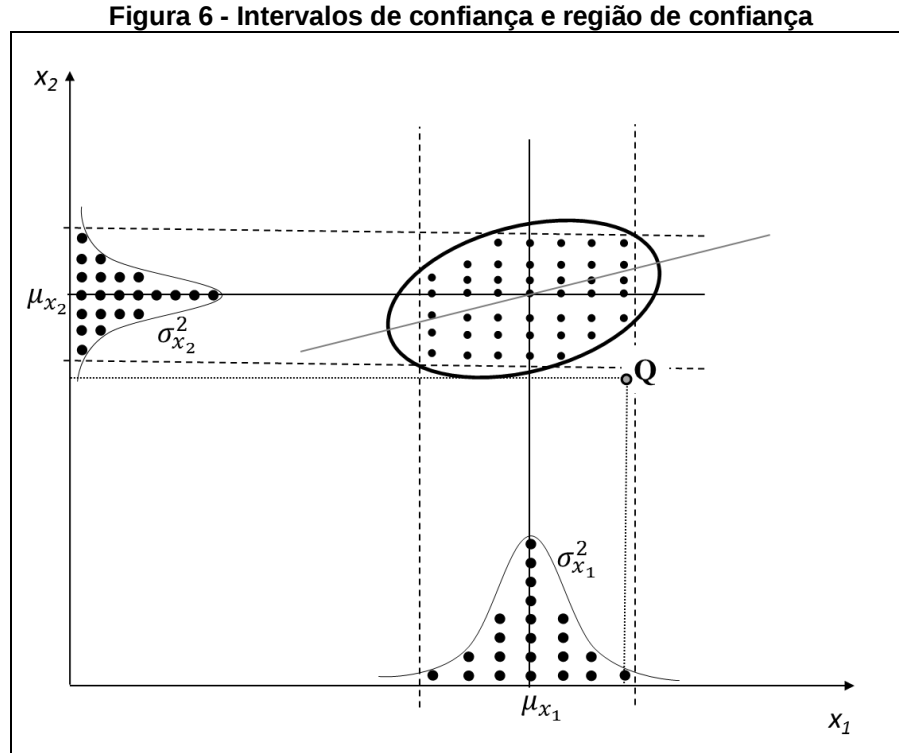
2.6 CONSTRUÇÃO DA ELIPSE DE CONFIANÇA

Para o caso bivariado, é possível construir duas elipses diferentes de confiança. Uma delas é a elipse de confiança para o vetor de médias da população, que apresenta a região em que, com um grau de confiança de $100(1 - \alpha)\%$, estaria esse verdadeiro vetor de médias. Sejam, então, respectivamente, $\bar{X}$ e S as estimativas do vetor de médias e da matriz de variância e covariância de uma amostra aleatória normal bivariada de tamanho n , com média μ e matriz de variância e covariância Σ . A variável $(\bar{X} - \mu)$ tem distribuição normal bivariada com vetor de médias $\mathbf{0}$ (zero) e matriz de variância e covariância $\frac{1}{n}\Sigma$ e é independente de S . Assim, a elipse com $100(1 - \alpha)\%$ de confiança é baseada na estatística T^2 de Hotelling e é calculada como:

$$T^2 = (\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu) = \frac{2(n-1)}{n(n-2)} F_{1-\alpha;2;n-2} \quad [18]$$

Em que $F_{1-\alpha;2;n-2}$ é o percentil $1 - \alpha$ da distribuição F com graus de liberdade 2 e $n - 2$, respectivamente.

A Figura 6 mostra como uma elipse de confiança é construída a partir de duas variáveis normais univariadas.



Fonte: Elaboração própria.

A outra é a elipse de predição que representa a região em que, com um grau de confiança de $(1 - \alpha)\%$, estaria uma nova observação multivariada da população. Ela também é a região que contém uma específica porcentagem da população.

Para esse caso, seja X uma variável aleatória de uma nova observação. Desse modo, a variável $\bar{X} - X$ tem distribuição normal bivariada com vetor de médias $\mathbf{0}$ (zero) e matriz de variância e covariância $(1 + \frac{1}{n})\Sigma$ e é independente de S . Logo, a elipse com $100(1 - \alpha)\%$ de confiança para a nova observação X é dada por:

$$(\bar{X} - \mu)^T S^{-1} (\bar{X} - \mu) = \frac{2(n+1)(n-1)}{n(n-2)} F_{1-\alpha;2;n-2} \quad [19]$$

A família de elipses gerada por diferentes valores da distribuição F tem um centro em comum, que é a média amostral e os eixos maiores e menores também

comuns. Vale ressaltar que a construção das elipses de confiança são visíveis somente para o caso bivariado, isto é, para o plano cartesiano. Quando se considera a avaliação de p características para um determinado produto, é impossível visualizar as regiões de confiabilidade.

2.7 ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (ACP), de acordo com Vicini (2005), é uma técnica matemática da análise multivariada, que possibilita investigações com um grande número de dados disponíveis. Possibilita, também, a identificação das medidas responsáveis pelas maiores variações entre os resultados, sem perdas significativas de informações. Além disso, transforma um conjunto original de variáveis em outro conjunto.

Essa transformação, em outro conjunto de variáveis, ocorre com a menor perda de informação possível, sendo que esta também busca eliminar algumas variáveis originais que possua pouca informação. Essa redução de variáveis só será possível se as p variáveis iniciais não forem independentes e possuírem coeficientes de correlação não nulos, conforme afirma Timm (2002). Na análise de componentes principais, um conjunto de p variáveis correlacionadas é transformado em um conjunto menor de variáveis não correlacionadas.

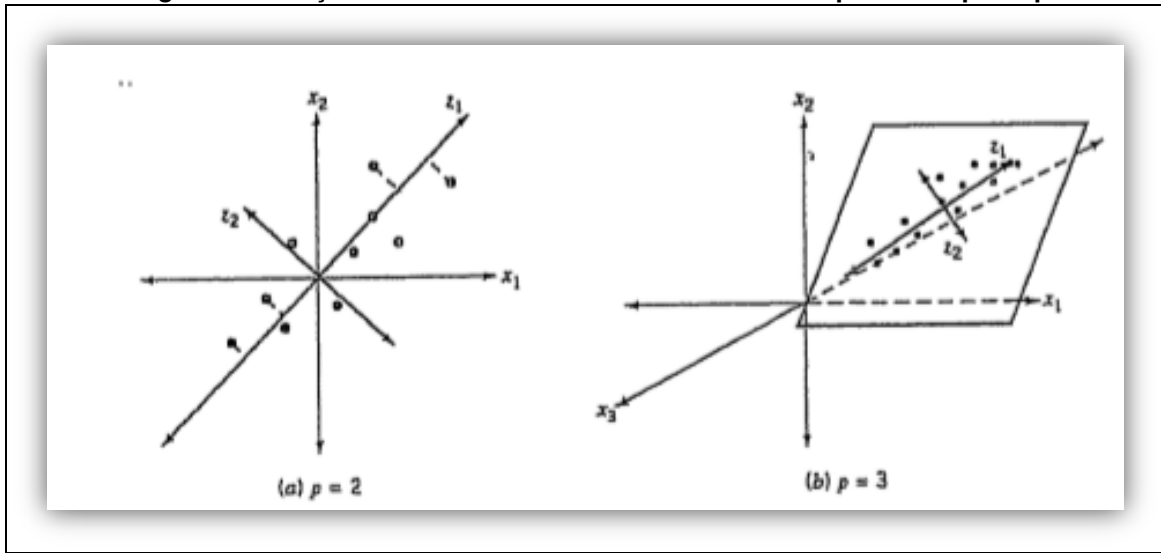
De acordo com Montgomery (2009), as componentes principais de um conjunto de variáveis de um processo $x_1, x_2, \dots, x_p$ são um conjunto particular de combinações lineares dessa variáveis. Por exemplo:

$$\begin{aligned} z_1 &= c_{11}x_1 + c_{12}x_2 + \dots + c_{1p}x_p \\ z_2 &= c_{21}x_1 + c_{22}x_2 + \dots + c_{2p}x_p \\ &\vdots \\ z_p &= c_{p1}x_1 + c_{p2}x_2 + \dots + c_{pp}x_p \end{aligned}$$

em que os c_{ij} são constantes a serem determinadas. O autor supracitado ainda afirma que, geometricamente, as variáveis das componentes principais $z_1, z_2, \dots, z_p$ são os eixos do novo sistema de coordenadas obtido pela rotação do

sistema original (os x 's). Os novos eixos representam as direções de variabilidade máxima. Conforme mostra a Figura 7.

Figura 7 - Direções de variabilidades máximas das componentes principais



Fonte: Montgomery (2009).

Observe-se que, em ambas as situações, com duas e três variáveis originais, a componente principal z_1 responde pela maior parte da variabilidade.

2.8 OUTLIERS

Por definição, *Outliers* são observações atípicas presentes em um conjunto de dados que apresentam um grande afastamento das demais observações ou que possuem alguma característica que difere das demais, tornando-se inconsistente com o restante do agrupamento de dados. A detecção desse tipo de observação é bastante importante para se entender melhor o comportamento dessa distribuição de dados e, conseqüentemente, decidir o que se pode fazer com esses pontos discordantes.

Essas observações discordantes podem ser benéficas ou problemáticas. As benéficas vão servir para indicar certas características de um agrupamento de dados que não seriam descobertas normalmente durante a análise. As problemáticas são aquelas que não representam esse agrupamento de dados e podem comprometer seriamente os testes estatísticos realizados.

Hair *et al.* (2009) dividem os *outliers* em quatro classes, que são baseadas na fonte de suas peculiaridades, a saber:

Classe 1- Surge de um erro de procedimento, como erro na entrada de dados ou uma falha na codificação. Essas observações devem ser identificadas no estágio de limpeza dos dados, mas se permanecerem algumas, elas devem ser eliminadas ou registradas como valores perdidos.

Classe 2 - É aquela que ocorre como o resultado de um evento extraordinário, o que então explica a peculiaridade da observação. Nesse caso, o pesquisador deve decidir se o evento extraordinário se ajusta aos propósitos da pesquisa. Se for o caso, o ponto discordante deve ser mantido na análise. Caso contrário, deve ser eliminado.

Classe 3 - São aquelas observações extraordinárias para os quais o pesquisador não tem explicação. Em tais casos, um padrão único e muito diferente emerge. Apesar de essas observações terem uma maior probabilidade de serem eliminadas, poderão ser mantidas se o pesquisador achar válida a permanência delas.

Classe 4 - Essa última classe contém observações que estão no intervalo usual de valores para cada variável. Essas observações não são particularmente altas ou baixas nas variáveis, mas são únicas em sua combinação de valores entre as variáveis. Em tais situações, o pesquisador deve reter a observação, a não ser que exista uma evidência específica que desconsidere a observação atípica com um membro válido da população.

O estudo de *outliers* pode ser realizado em várias fases. Figueira (1998) classifica as fases da seguinte forma: A fase inicial é a identificação das observações que são potencialmente aberrantes. A identificação de *outliers* consiste na detecção, com métodos subjetivos, das observações surpreendentes. Essa identificação é feita geralmente por análise gráfica, ou, no caso de o número de dados ser pequeno, por observação direta dos mesmos.

Na segunda fase, tem-se como objetivo a eliminação da subjetividade inerente à fase anterior. Pretende-se saber se as observações identificadas como *outliers* potenciais o são, efetivamente. São efetuados testes às observações que se suspeita ser um ponto discordante. Estes tipos de testes dependem do tipo de *outlier* em causa, do seu número, da sua origem, do conhecimento da distribuição subjacente à população de origem das observações, etc.

Na terceira fase é necessário decidir o que fazer com as observações que foram identificadas como discordantes. Se eliminá-las ou tratá-las. Dependendo do que o pesquisador pretende obter com a identificação desses pontos discordante, quaisquer das duas opções são perfeitamente plausíveis de serem escolhidas.

Alguns passos para a detecção de *outliers* são propostos por Johnson e Wichern (2007):

- 1º passo- Faça um gráfico de pontos para cada variável;
- 2º passo- Faça um gráfico de dispersão para cada par de variáveis;
- 3º passo- Calcule os valores padronizados $Z_{jk} = (X_{jk} - \bar{X}_k) / \sqrt{S_{k,k}}$, para $j = 1, 2, \dots, n$ e cada coluna $k = 1, 2, \dots, p$. Examine se esses valores padronizados são para grandes ou pequenos valores. Esse termo “grande” é com relação ao tamanho da amostra e o número de variáveis.
- 4º passo- Calcule as distâncias generalizadas ao quadrado $(X_j - \bar{X})' S^{-1} (X_j - \bar{X})$. Examine essas distâncias para valores grandes. Em uma parcela do Qui-quadrado, estes seriam os pontos mais afastados da origem. Esse termo “grande”, nessa etapa, é medido por um percentual adequado de distribuição do Qui-quadrado com p graus de liberdade.

2.8.1 Outliers: caso univariado

Os *Outliers* em amostras univariadas são analisados quando a preocupação está em avaliar os valores de uma única variável que se encontram bem distantes dos intervalos de distribuição. Para isso, é necessário utilizar bases para designação de uma observação atípica (HAIR *et al.*, 2009). É necessário, primeiramente, converter os valores dos dados em escores padrão com média zero e desvio-padrão 1, para facilitar a comparação entre as variáveis.

Para a identificação dos *outliers*, Raykov e Marcoulides (2008) sugerem duas propriedades: A primeira diz que a magnitude dos seus escores z tem que ser maior do que 3 ou menor do que -3, e a segunda diz que esses escores z são, em certa medida, desconectados dos escores z das observações restantes. Já Hair *et al.* (2009) descrevem regras de identificação dessas observações que são as seguintes.

- Para pequenas amostras (80 observações ou menos), observações atípicas geralmente são definidas como casos com escores padrão de 2,5 ou mais.
- Para amostras maiores, aumente o valor de referência de escores padrão para até 4.
- Se escores padrão não são usados, identifique casos que recaiam fora dos intervalos de desvio-padrão de 2,5 versus 4, dependendo do tamanho da amostra.

Em contrapartida, existem outros métodos de identificação desses pontos anormais, que podem ser identificados através de recursos gráficos os quais incluem gráfico de box-plot, histogramas e outros tipos de teste como o teste t, teste de Dixon, teste de Grubs e teste de Walsh. Muitos modelos têm sido desenvolvidos através desses testes de rejeição de *outlier*.

Rencher (2002) afirma que, para um único *outlier* em uma amostra de $x_1, x_2, \dots, x_n$, muitos testes são baseados no máximo valor residual “studentizado”, que é representado pela fórmula a seguir:

$$\max \tau_i = \max \left| \frac{X_i - \bar{X}}{S} \right| \quad [20]$$

Se esse valor exceder os limites pré-determinados, observa-se a ocorrência de um ponto discordante no agrupamento de dados analisados, evidenciando a presença de *outliers*.

2.8.2 *Outliers*: caso multivariado

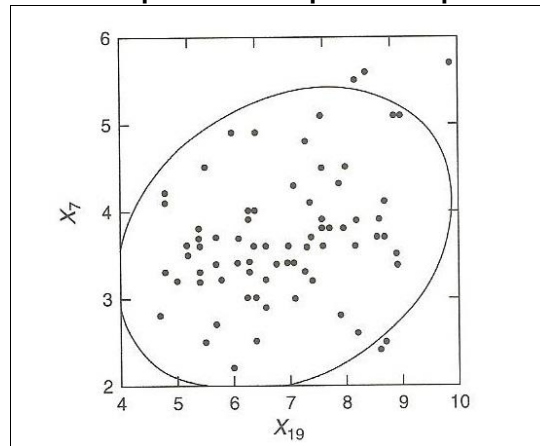
No caso multivariado, a presença de *outliers* é estudada quando há uma preocupação em analisar os valores de mais de uma variável. Um *outlier* multivariado é uma observação com valores em diversas variáveis que não são necessariamente anormais quando cada variável é considerada separadamente, mas são incomuns na sua combinação. (RAYKOV; MARCOULIDES, 2008).

Os casos bivariados de dados são aqueles em que duas variáveis são avaliadas conjuntamente. Nesses casos, é necessário estabelecer uma região de

confiança representada por uma elipse e sobrepor um diagrama de dispersão que vai fornecer os limites a partir dos quais se identificará a presença de *outliers*.

A Figura 8 a seguir exemplifica a detecção de *outlier* no caso bivariado.

Figura 8 - Diagrama de dispersão sobreposto à elipse de confiança



Fonte: Hair et al. (2009).

Uma desvantagem da análise bivariada é que, à medida que o número de variáveis aumenta, aumentam também a quantidade de gráficos de dispersão, ficando muito dispendioso e dificultando a detecção dos *outliers*. Nesse caso, a ideia básica é a utilização de métodos de detecção multivariados, os quais necessitam da utilização de algumas medidas de distâncias que irão avaliar o quão longe uma observação se encontra a partir do seu centro de conjunto de dados (VERDONCK; WOUWE, 2011).

Dentre essas medidas de distância para medir a posição multidimensional de cada observação relativamente a algum ponto comum, a mais utilizada é conhecida como a distância de Mahalanobis. Essa medida representa os valores mais afastados do intervalo de distribuição. O cálculo para a obtenção dessa medida é apresentado da seguinte forma:

$$D^2 = (X - \bar{X})^T S^{-1} (X - \bar{X}) \quad [21]$$

Em que:

X é um ponto dentro do intervalo de distribuição;

$\bar{X}$ é o vetor de média ;

S^{-1} é a matriz de variância e covariância.

Para dados multivariados distribuídos normalmente, os valores de D^2 tem aproximadamente distribuição qui-quadrado com p graus de liberdade (FILZMOSER *et al.*, 2005).

Nos casos de dados multivariados, Rencher (2002) afirma que os problemas na detecção de *outliers* existem por diversas razões. Dentre elas estão:

- Para $p > 2$, os dados não podem ser facilmente plotados para identificação de *outliers*;
- Dados multivariados não podem ser ordenados como podem ser em uma amostra univariada, onde os extremos aparecem prontamente em cada extremidade;
- Um vetor de observação pode ter um grande erro de registro/relação em um de seus componentes ou erros menores em vários componentes;
- Um *outlier* multivariado pode refletir uma influência na média, variância ou na correlação. Um exemplo consistiria de pequenas causas que deslocam a média e a variância, mas que tem um pequeno efeito na correlação.

Outra hipótese de detecção de *outliers* multivariados pode ser realizada através de análise de componentes principais. Esse tipo de análise é uma técnica matemática que tem o objetivo de investigar um grande número de variáveis no intuito de se obter uma melhor informação sobre determinado grupo de dados. Possibilita, também, a identificação das medidas responsáveis pelas maiores variações entre os resultados, sem perdas significativas de informações. Além disso, transforma um conjunto original de variáveis em outro conjunto:

Essa transformação em outro conjunto de variáveis ocorre com a menor perda de informação possível, sendo que tal transformação também busca eliminar algumas variáveis originais que possuam pouca informação. Essa redução de variáveis só será possível se as p variáveis iniciais não forem independentes e possuírem coeficientes de correlação não nulos.

De acordo com Johnson e Wichern (2007), parcelas das componentes principais podem revelar observações suspeitas, bem como proporcionar o controle da suposição de normalidade. As primeiras partes das observações dos componentes principais são poucas vezes normalmente distribuídas, quando estão sendo usadas como entrada para análises adicionais. Já as últimas observações

dos componentes principais podem ajudar a detectar pontos suspeitos, podendo ser *outliers*. Cada observação pode ser expressa como uma combinação linear:

$$Y_i = a_i^T X = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p \quad [22]$$

Em que o vetor desconhecido a_i estabelece a i -ésima combinação linear, para $i=1, 2, \dots, n$.

Johnson e Wichern (2007) ainda resumem essa ideia da seguinte forma:

- Para ajudar a verificar o pressuposto normal, é necessário construir diagramas de dispersão para os pares dos primeiros componentes principais. Além disso, fazer Q-Q *plots* (gráfico que avalia a normalidade multivariada) a partir dos valores de amostras gerados por cada componente principal;
- Construir diagramas de dispersão e plotagem Q-Q *plots* para os últimos poucos componentes principais, pois esses ajudam a identificar as observações suspeitas.

Através da análise de gráficos a duas e a três dimensões dos componentes principais, podem ser detectadas observações *outliers* que adicionam dimensões sem importância aos dados ou escondem singularidades presentes neles (FIGUEIRA, 1998).

Outro método utilizado para a detecção de *outlier* multivariado é proposto por Garret (1989), o qual usa a estatística Qui-quadrado em um processo iterativo de corte até que um *plot* linear aceitável do Qui-quadrado seja obtido.

Para possibilitar esse processo iterativo de corte, foi utilizado o procedimento conhecido como MVT-GAIT, o qual utiliza o IDEAS, que é um ambiente interativo de computação gráfica utilizando-se de gráficos interativos.

Segundo Todorov *et al.* (2010), o procedimento de identificação de *outliers* baseado em $\bar{x}$ (média) e S (covariâncias) se depara com os seguinte problemas:

- *Masking*-(Mascaramento)- É uma ocorrência em *outliers* multivariados que pode distorcer as estimativas clássicas de média e covariância (atraindo $\bar{x}$ e inflando S), de tal forma que eles não fiquem necessariamente

grandes para valores da distância de Mahalanobis, ou seja, o efeito do mascaramento ocorre quando alguns outliers são erroneamente classificados como observações típicas (não-outliers).

- *Swamping*-(esmagamento ou inundação)- É a ocorrência em que valores extremos múltiplos podem distorcer as estimativas clássicas de média e covariância, de tal maneira que as observações que são consistentes com a maioria dos dados pode obter valores de grandes dimensões para a distância de Mahalanobis, ou seja, o efeito da inundação ou esmagamento ocorre quando algumas observações típicas (não-outliers) são erroneamente classificados como outliers.

São encontradas inúmeras publicações sobre estudos relacionados à detecção de outliers, uma vez que esse tema pode ser utilizado em diversas áreas por meio de uma infinidade de aplicações. Tais publicações podem ser verificadas em: Garret (1989), que utiliza um processo iterativo de corte usando um ambiente de computação gráfica; Foss e Zaiane (2010), que aplica um método de separação de classes de variâncias utilizando um novo algoritmo (FASTOUT); Barnard e Aldrich (2000), que usam a técnica de cascos convexos de detecção em tempo real, por meio de um algoritmo que compara os custos com outros algoritmos em termos de outliers falsos.

2.9 O ALGORITMO BACON

O mascaramento e o esmagamento de outliers são ocorrências que distorcem os valores da média e da covariância, prejudicando a análise desses pontos extremos. Para combater esses problemas que ocorre na presença de múltiplos outliers em um conjunto de dados, alguns algoritmos de busca partem de um pequeno subconjunto, comprovadamente livres de outliers, dos dados originais e vão adicionando novas informações, que também não são outliers, a esse subconjunto inicial até que nenhuma informação possa mais ser absorvida. Esses algoritmos são chamados de Algoritmos de Buscas sucessivas.

Billor *et al.* (2000) desenvolveram um desses algoritmos de busca para frente, para detectar outliers multivariados, chamado, numa tradução livre, de Indicador de Outliers através de Blocos Adaptativos Computacionalmente Eficientes ou

simplesmente de algoritmo BACON (*Blocked Adaptive Computationally-Efficient Outlier Nominators*). Existem duas diferentes formas em que o algoritmo BACON tem para gerar o subconjunto inicial básico: uma para dados multivariados e outra para dados de regressão.

O primeiro passo do algoritmo será escolher um subconjunto inicial básico que não contenha *outliers*. Existem duas propostas para se realizar essa tarefa:

PROPOSTA 1: Seleção do subconjunto inicial com base nas distâncias de Mahalanobis, ou seja:

- a) Calcular as distâncias de Mahalanobis para cada uma das observações, isto é:

b)

$$d_i = \sqrt{(X_i - \bar{X})^T S^{-1} (X_i - \bar{X})} \quad i = 1, 2, \dots, n \quad [23]$$

Em que $\bar{X}$ e S são, respectivamente, o vetor de médias e a matriz de variância e covariância das m observações.

- c) Identificar as $m = cp$ observações com os menores de d_i e denominar essas observações como subconjunto inicial básico. O valor de c é um valor inteiro e escolhido pelo usuário, que por padrão é igual a 3.

PROPOSTA 2: Seleção do subconjunto inicial com base nas distâncias a partir das medianas, ou seja:

- a) Para $i = 1, 2, \dots, n$, calcular $\|X_i - X_{Med}\|$, em que X_{Med} é um vetor que contém as medianas de cada uma das p variáveis e $\|\cdot\|$ é a norma Euclidiana;
- b) Identificar as $m = cp$ observações com os menores de $\|X_i - X_{Med}\|$ e denominar essas observações como subconjunto inicial básico.

Em ambas as propostas, no caso de a matriz de variância e covariância envolvida no cálculo for singular, então deve-se aumentar o subconjunto básico em uma unidade até que se tenha uma matriz de posto completo.

De um modo geral, o algoritmo BACON consiste nos seguintes passos:

PASSO 1: Identificar o subconjunto inicial básico $\mathbf{G}$ com tamanho m de observações, usando uma das duas propostas;

PASSO 2: Calcular as distâncias (discrepâncias) d_i para $i = 1, 2, \dots, n$;

PASSO 3: Definir um novo subconjunto G contendo todas as observações com discrepâncias menores que $c_{npr}\chi_{p;\alpha/n}$, em que $\chi_{p;\alpha}$ é o percentil $1 - \alpha$ de uma distribuição de Qui-quadrado com p graus de liberdade; e $c_{npr} = c_{np} + c_{hr}$ é um fator de correção, sendo r = tamanho do subconjunto básico, $h = (n + p + 1)/2$, e os demais cálculos como se seguem:

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{1}{n-h-p} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}$$

[24]

$$c_{hr} = \max\{0; (h-r)/(h+r)\}$$

Observação: quando o tamanho do subconjunto básico r for menor do que h , os elementos da matriz de variância e covariância tendem a ser menores do que realmente deveriam ser. Assim, deve-se entender c_{hr} como um fator de inflação da variância, que é usado para aumentar a variância quando r é menor do que h . Note-se também que, se $r = h$, então c_{npr} se reduz a c_{np} .

PASSO 4: Repetir os **PASSOS** (2) e (3) até que o tamanho do subconjunto básico não se altere mais;

PASSO 5: Nomear (rotular) as observações que ficaram de fora do subconjunto G como *outliers*;

Segundo Billor *et al.* (2000), a Proposta 1 para gerar o subconjunto inicial é menos robusta. Contudo, resultados simulados mostraram que existe um ponto de ruptura empírico em torno de 20% e com um custo computacional mais baixo do que o da Proposta 2. Por outro lado, essa última proposta é mais robusta, apesar de apresentar um ponto de ruptura superior a 40%.

O que se pode afirmar é que o esforço computacional exigido pelo algoritmo BACON é muito pequeno, em especial pelo fato de que esse esforço cresce lentamente com o aumento do tamanho da amostra, característica que torna esse método adequado para examinar grandes massas de dados.

CAPÍTULO 3 - PROCEDIMENTOS METODOLÓGICOS

A empresa objeto deste estudo foi pioneira no estado da Paraíba no ramo de injeção de plásticos. Ela foi fundada no início da década de 70, a partir da percepção da carência na região no que se refere à transformação de plásticos. Sua planta fabril encontra-se instalada em uma área de 56 mil metros quadrados, no distrito industrial de João Pessoa, Paraíba.

Sua gama de produtos compreende as linhas de embalagens e utilidades domésticas, cujo método de obtenção consiste na injeção da resina termoplástica em um molde específico, a qual tomará sua forma. Para tanto, se faz uso de máquinas injetoras que condicionam a obtenção do produto, controlando fatores como temperatura, tempo de permanência no molde, pressão de trabalho, entre outras.

O presente trabalho de pesquisa, sob a ótica dos procedimentos técnicos, é um estudo de caso, pois compreende um estudo profundo de um ou poucos objetos, de maneira que permita um conhecimento mais detalhado da realidade da pesquisa. Segundo Gil (1994), o estudo de caso tem como principais objetivos: explorar situações da vida real cujos limites não são claramente definidos; preservar o caráter unitário do objeto estudado; descrever a situação do contexto em que está sendo feita determinada investigação; e explicar as variáveis causais de determinado fenômeno em situações muito complexas que não possibilitam a utilização de levantamentos e experimentos.

A pesquisa bibliográfica também se faz presente, pois essa pesquisa também foi realizada com a ajuda de materiais publicados, como livros, artigos de periódicos, bem como material disponibilizado na internet. Segundo Rodrigues (2007), a pesquisa bibliográfica recupera o conhecimento científico acumulado sobre um problema.

Este trabalho caracteriza-se também, do ponto de vista da abordagem do problema, como uma pesquisa quantitativa, pois requer o uso de recursos e técnicas estatísticas para a realização da pesquisa. Silva e Menezes (2001) ainda afirmam que a pesquisa quantitativa considera que tudo pode ser quantificável, o que significa traduzir em números opiniões e informações para classificá-las e analisá-las.

Do ponto de vista de seus objetivos, esta pesquisa apresenta um caráter exploratório. Esse tipo de pesquisa é utilizado para realizar um estudo preliminar do principal objetivo da pesquisa que será realizada, ou seja, familiarizar-se com o fenômeno que está sendo investigado, de modo que a pesquisa subsequente possa ser concebida com uma maior compreensão e precisão.

Os instrumentos de coleta de dados que foram utilizados neste presente trabalho consistiram na observação sistemática e na análise documental. A observação sistemática auxilia no processo de análise dos dados amostrais que são colhidos durante a pesquisa de campo e que não são encontrados nos arquivos da empresa, tendo o intuito de robustecer o processo de pesquisa. Para Silva e Menezes (2001), a observação sistemática tem planejamento e realiza-se em condições controladas para responder aos propósitos preestabelecidos.

A análise documental é um método de coleta de dados que busca documentos que podem ser encontrados nos arquivos da empresa e que são referentes ao objeto de pesquisa. Conforme Godoy (1995), a análise documental se constitui em uma fonte estável e rica, tem baixo custo, complementa informações e indica problemas.

A quantidade de amostras retiradas foi em número de 100 (cem), de tamanho igual a quatro, a cada intervalo de tempo regular de 25 minutos. A determinação desses parâmetros depende muito do processo produtivo, dos custos envolvidos e desperdícios de fabricação. Analisando todos esses fatores, determinaram-se tais valores para o tamanho da amostra e para o intervalo entre retiradas, seguindo o esquema apresentado a seguir.

3.1 ANÁLISE DE EFICIÊNCIA DO GRÁFICO DE $\bar{X}$ VIA AVERAGE RUN LENGTH (ARL)

O ARL, ou seja, o comprimento médio da sequência, também pode ser traduzido por número médio de amostras até o alarme (NMA).

3.1.1 Erros dos tipos 1 e 2 e poder de detecção do gráfico de X barra

$$H_0 : \mu = \mu_o$$

[25]

$$H_1: \mu \neq \mu_0$$

[26]

Onde μ_0 é a média do processo isento de causas especiais.

O erro do tipo 1, α , que ocasiona um alarme falso, consiste na probabilidade de que, estando o processo livre de causas especiais, o gráfico de $\bar{X}$ indique uma observação na zona de ação, ou seja, acima do Limite Superior de Controle ($LSC_{\bar{X}}$) ou abaixo do Limite Inferior de Controle ($LIC_{\bar{X}}$). Tal probabilidade pode ser expressa pela seguinte equação:

$$\alpha = \Pr[\bar{X} > LSC_{\bar{X}} \text{ ou } \bar{X} < LIC_{\bar{X}} \mid \mu = \mu_0]$$

[27]

O erro do tipo 2, β , ocorre quando o processo, mesmo estando sujeito à influência de causas especiais, for representado por um Gráfico de $\bar{X}$ exibindo uma observação dentro dos limites de controle. A probabilidade de ocorrência deste tipo de erro é calculada pela expressão seguinte:

$$\beta = \Pr[LIC_{\bar{X}} \leq \bar{X} \leq LSC_{\bar{X}} \mid \mu \neq \mu_0]$$

[28]

O poder de detecção, Pd, do gráfico de $\bar{X}$ é definido como a probabilidade deste gráfico indicar um alarme verdadeiro, ou seja, quando ocorrer uma observação acima ou abaixo dos limites de controle, caso a média do processo tenha se alterado. Portanto, o Pd do gráfico de $\bar{X}$ pode ser calculado por intermédio da seguinte fórmula:

$$Pd = 1 - \beta$$

[29]

Sendo k igual ao número de desvios-padrão a partir da média, através da expressão (27) obtém-se:

$$\alpha = \Pr[Z \leq |K|]$$

[30]

Para o valor usual de $k = 3$, por intermédio da expressão (30), pode-se verificar que o valor de α é igual a 0,0027. O NMA_0 até um alarme falso é $1/\alpha$, ou seja, aproximadamente 370 amostras. Da mesma maneira, o NMA_1 até um alarme verdadeiro é $1/Pd$, pois ambas as distribuições são geométricas de parâmetros α e Pd , respectivamente, sendo suas médias iguais aos inversos desses parâmetros.

3.1.2 Escolhas do intervalo de tempo entre amostras (h), do tamanho da amostra (n), de número de amostras (m) e separação das fases I e II

O intervalo de tempo entre amostras(h) depende de vários fatores, entre os quais se destacam o tempo de produção de cada item, o custo de inspeção, o tempo de inspeção, a estabilidade do processo e o prejuízo de se operar um processo submetido à influência de causas especiais.

Na prática, geralmente o valor de h é escolhido de maneira arbitrária, mas deve-se considerar que valores pequenos de h implicam custos elevados com amostragem e maior incidência de alarmes falsos (COSTA *et al.*, 2004).

O NMA_1 correspondente ao alarme verdadeiro de que o processo mudou depende da intensidade do deslocamento da média do processo: por um lado, no caso em que o processo sofra uma causa especial grave e a média do processo se desloque de uma distancia de vários desvios-padrão, o NMA_1 será pequeno, pois o gráfico detectará rapidamente a presença dessa causa; por outro lado, para pequenos deslocamentos da média, o NMA_1 será grande indicando que esses deslocamentos dificilmente serão detectados rapidamente (SAMOHYL, 2009).

Segundo Montgomery (2009), recentemente o uso do NMA para descrever o desempenho do gráfico de $\bar{X}$ tem sido bastante criticado, pelo fato de a distribuição do NMA ser geométrica. Assim há duas preocupações em relação ao NMA: por um lado, o desvio-padrão do NMA é muito grande; por outro lado, a distribuição geométrica é muito assimétrica, de modo que a média da distribuição ($1/Pd$ ou $1/\alpha$) não é necessariamente um valor “típico” do comprimento da sequência.

No entanto, pode ser útil expressar o NMA em termos do número de unidades individuais amostradas, H, ao invés do número de amostras retiradas para detectar um deslocamento. Assim, se o tamanho da amostra é n e h é constante, a relação entre H e NMA pode ser expressa pela seguinte equação:

$$H = n.NMA$$

[30]

A título de ilustração, pode-se comparar 2 planejamentos do gráfico de $\bar{X}$ para um deslocamento da média de 2 desvios-padrão: $n=4$ com $NMA_1=1,1885$ e $n=9$ com $NMA_1=1,0014$. Verifica-se pela expressão (30) que $H_4=4,7542$ e $H_9=9,0122$.

Assim, um gráfico de $\bar{X}$ com $n=9$ irá requerer que aproximadamente 9 unidades sejam amostradas, ao passo que, se o tamanho da amostra for $n=4$, apenas cerca de 5 unidades serão necessárias, em média. Ainda no que se concerne ao gráfico de $\bar{X}$, o tamanho da amostra geralmente é pequeno, sendo $n=1,2,3...9$. O número mínimo de amostras (m) necessárias à construção de novos gráficos na linha de produção (a primeira fase de implantações de gráficos de controle) é $m=25$.

De fato, a questão da escolha do tamanho da amostra é, sobretudo, de ordem econômico-financeira e não estatística. Amostras maiores são sempre mais precisas, pois têm maior probabilidade de detectar alterações no processo, mas acarretam gastos mais elevados em termos de tempo e recursos além de erros de mensuração mais frequentes. Nesse sentido, na indústria em geral, as mensurações devem ser efetuadas com extremo cuidado e o valor de m deve ser mínimo. Além disso, é raro observar um valor de $n>9$, e o valor de $n=4$ parece ser de uso universal.

Na prática, a implantação e utilização do gráfico de $\bar{X}$ passam por 2 fases de operações: na fase I, a montagem do gráfico, os dados são inicialmente levantados, estatísticas tais como média e desvio-padrão, são posteriormente calculadas. Com esses valores o gráfico de controle é montado.

Na Fase II, o monitoramento do processo, ocorre a experiência de usar o gráfico de $\bar{X}$, quando novos dados são levantados no decorrer do dia e novos pontos são inseridos no gráfico.

Na fase I, os dados são coletados e analisados de uma vez numa análise retrospectiva, construindo limites de controle de teste para determinar se o processo estava em controle estatístico durante o período de tempo onde foram coletados, e para observar se limites de controle confiáveis podem ser estabelecidos para monitorar a produção futura.

Na fase II, geralmente se assume que o processo é razoavelmente estável. De uma maneira geral, as causas especiais que ocorrem na fase II resultam de

pequenos deslocamentos do processo, pois as verdadeiras fontes ruins de variabilidade foram sistematicamente removidas durante a fase I (MONTGOMERY, 2005).

Somente a fase I foi desenvolvida nesta pesquisa, e um grande número de observações (4 grupos de 25 amostras de $n=4$) foram coletadas para se determinar também o valor nominal e os limites de especificação para uma determinada percentagem de itens fora das especificações estabelecidas pela empresa em estudo.

3.2 ESCOLHA DAS VARIÁVEIS DE ESTUDO

As variáveis de estudos deste trabalho são as dimensões da caixa de refrigerantes que se pode ver através das figuras abaixo:

Figura 9 – Vista Lateral



Figura 10 - Vista Superior



Fonte: Elaboração própria.

As características de qualidade que podem ser analisadas conjuntamente são:

- CE – comprimento externo;
- CI – comprimento interno;
- LE – largura externa;
- LI – largura interna;
- AT – altura total;
- AE – altura de empilhamento;
- CF – comprimento de fundo (Módulo de Encaixe do Fundo-Comprimento);
- LF – largura de fundo (Módulo de Encaixe do Fundo – Largura);
- AN – altura do ninho (Altura das Divisórias Internas);
- AF – altura de encaixe do fundo.

Para que tais medições fossem exatas, o instrumento utilizado para a verificação foi o paquímetro, que é usado para medir as dimensões lineares internas e externas bem como a profundidade de uma peça ou produto. Consiste em uma régua graduada com encosto fixo, sobre o qual desliza um cursor que permite uma precisão decimal de leitura através do alinhamento desta escala com uma medida da régua.

3.3 PONTOS FORA DOS LIMITES DE CONTROLE DO GRÁFICO DA MÉDIA NAS 25 PRIMEIRAS AMOSTRAS DE TAMANHO 4

O Apêndice A mostra claramente, em vermelho, a presença de diversos pontos na zona de ação do gráfico da média univariado para as variáveis desta pesquisa. Todavia, tais pontos encontram-se dentro do intervalo dos limites de controle alargados, obtidos através da substituição das estimativas dos desvios-padrão calculadas com base na dispersão amostral, usando $\bar{R}/d^2$, pelas estimativas dos desvios-padrão calculadas com base na variabilidade total das médias amostrais, ou seja, o desvio-padrão amostral das médias, que é devida em parte à variabilidade do nível da série e em parte ao ruído de curto prazo.

Costa *et al.*(2004) afirmam que em processos discretos, especialmente os total ou parcialmente automatizados, poderá haver o fenômeno da autocorrelação. Assim, alguns desses processos podem exibir características de qualidade cujo nível médio varia gradualmente ao longo do tempo (exibindo, portanto, autocorrelação), embora possa haver variação significativa de peça para peça.

Considerando-se ainda o Apêndice A, onde apenas as 25 primeiras amostras apresentaram 28 alarmes falsos, pode-se admitir que o processo de produção das garrafeiras plásticas seja autocorrelacionado. Em suma, embora seja perfeitamente possível a adoção dos gráficos de valores individuais e da amplitude móvel, optou-se pelos gráficos alargados da Média e da Amplitude R, para controlar não apenas o nível médio do processo ao longo do tempo, como também a variabilidade de curto prazo das dimensões das garrafeiras.

3.4 TESTES DE NORMALIDADE, ESPECIFICAÇÕES E COEFICIENTES DE CAPACIDADE DO PROCESSO

Quanto aos testes de normalidade, Samohyl (2009) constata que a simplicidade do teste de Jarque- Bera constitui-se em um grande benefício.

A estatística de Jarque-Bera, que segue a distribuição Qui-Quadrado, rejeita a normalidade da distribuição quando essa estatística for superior a 5,9 (com nível de significância igual a 5% e com número de graus de liberdade igual a 2).

Quando os coeficientes de assimetria (ass) e de curtose excessivo (curt) são muito elevados, a curva em forma de sino que caracteriza a distribuição Normal fica muito irregular fazendo com que a expressão de Jarque Bera se torne maior do que 5,9.

Tal estatística pode ser obtida pela expressão $N*[(ass)^2/6+(curt)^2/24]$, onde N corresponde ao número total de itens amostrados. No Apêndice B, observa-se que todas as variáveis analisadas apresentam estatísticas de Jarque-Bera menores do que 5,9 bem como coeficientes de assimetria e de curtose excessivo próximos de zero, fato que indica que não se pode rejeitar a hipótese nula de Normalidade.

O Apêndice B mostra ainda os limites de especificação superior (LSE) e inferior (LIE), utilizados para se calcular os coeficientes C_p e $C_{p,k}$ de capacidade do processo, todos acima de 1, valores típicos da indústria brasileira.

3.5 PROCEDIMENTOS PARA A CONSTRUÇÃO DA ESTATÍSTICA T^2 DE HOTTELING

O Controle Estatístico do Processo (CEP) é uma metodologia que atua preventivamente sobre o processo produtivo, utilizando a estatística como instrumento básico para avaliar suas alterações, bem como para busca o aperfeiçoamento contínuo da qualidade. Dentro deste contexto, os gráficos de controle são considerados uma das suas ferramentas mais importantes.

Mais especificamente, os gráficos de controle multivariados, como já foram discutidos anteriormente, são preferíveis aos gráficos de controle univariados porque levam em consideração a estrutura de correlação existente entre as várias características que são medidas no mesmo produto. Hotelling (1947) foi um dos

primeiros a analisar variáveis correlacionadas sob uma perspectiva de controle estatístico, utilizando-se de um procedimento multivariado.

Devido à complexidade dos cálculos relacionados à álgebra matricial, a aceitação dos gráficos de controle multivariados tem sido lenta, mas ainda assim o gráfico T^2 é o mais popular no controle multivariado do processo. Este procedimento foi baseado na generalização da estatística t de *Student*.

Para a proposta de trabalho desta dissertação, a construção do gráfico T^2 de Hotelling foi desenvolvida em duas etapas:

FASE I: A partir do banco de dados, contendo 100 amostras de tamanho $n = 4$, obtido na empresa com variáveis selecionadas para o estudo (10 medidas da qualidade), 80% dessas amostras, escolhidas aleatoriamente, foram utilizadas para construir a estatística T^2 de Hotelling (expressão 9) e os respectivos limites da região de 95% de confiança (expressão 12). Essa fase serviu para criar os critérios das medidas de diagnósticos do controle de qualidade multivariado.

FASE II: Uma vez obtidos os limites da região de 95% de confiança, na Fase II, a estatística T^2 de Hotelling foi recalculada para a parte da amostra que foi reservada (20%) para a validação, considerando tanto o vetor de médias quanto a matriz de variância e covariâncias dessa subamostra. A ideia que está por trás dessa partição da amostra é que, se o processo multivariado estiver em controle, então qualquer subamostra dessa população multivariada deverá estar em controle também. Assim, a primeira subamostra (80% de m) serviu para construir os limites da região de confiança, enquanto que a segunda subamostra serviu para testar o controle estatístico do processo (multivariado) propriamente dito, conforme o Apêndice D.

3.6 UTILIZAÇÃO DO ALGORITMO BACON PARA A DETECÇÃO DE OUTLIERS

Usando os dados colhidos no processo produtivo, selecionou-se o subconjunto básico inicial com base nas distâncias de Mahalanobis e seguiram-se os seguintes passos:

PASSO 1: Identificar o subconjunto inicial básico **G** com tamanho m observações, ou seja 30 observações;

PASSO 2: Recalcular as distâncias de Mahalanobis para esse novo conjunto de dados;

PASSO 3: Definir um novo subconjunto G contendo todas as observações com discrepâncias menores que $c_{npr}\chi_{p;\alpha/n}$, em que $\chi_{p;\alpha}$ é o percentil $1 - \alpha$ de uma distribuição de Qui-quadrado com p graus de liberdade; e $c_{npr} = c_{np} + c_{hr}$ é um fator de correção, sendo r = tamanho do subconjunto básico, $h = (n + p + 1)/2$, e os demais cálculos como seguem:

$$c_{np} = 1 + \frac{p+1}{n-p} + \frac{1}{n-h-p} = 1 + \frac{p+1}{n-p} + \frac{2}{n-1-3p}$$

[31]

$$c_{hr} = \max\{0; (h-r)/(h+r)\}$$

PASSO 4: Repetir os **PASSOS** (2) e (3) até que o tamanho do subconjunto básico não se altere mais;

PASSO 5: Nomear (rotular) as observações que ficaram de fora do subconjunto G como *outliers*;

Tanto a construção da estatística T^2 de Hotelling como a utilização do algoritmo de busca sucessiva, BACON, podem ser realizadas através do software R, que é uma linguagem desenvolvida em ambiente computacional para facilitar a obtenção de gráficos e cálculos estatísticos.

CAPÍTULO 4 - ANÁLISE DOS RESULTADOS E DISCUSSÃO

Este capítulo é dedicado à apresentação e discussão dos resultados produzidos de acordo com o que foi apresentado na metodologia, em que a utilização dos dados colhidos na empresa serviu de subsídio para a construção do gráfico de controle multivariado, bem como para a aplicação do algoritmo BACON de busca sucessiva utilizado neste trabalho.

O gráfico de controle multivariado T^2 serve para mostrar o retrato atual de como a empresa se apresentou com relação ao produto produzido, ou seja, diagnostica a situação atual de produção das garrafeiras. Essa ferramenta gráfica indica, entre outras informações, os pontos de alerta do processo de produção. Tais pontos devem ser examinados mais detalhadamente, no sentido de investigar se esses são apenas flutuações aleatórias de medições dos parâmetros da linha produtiva ou são observações extremas que, nesse caso, deveriam gerar ações estratégicas protetivas por parte do analista da qualidade do processo.

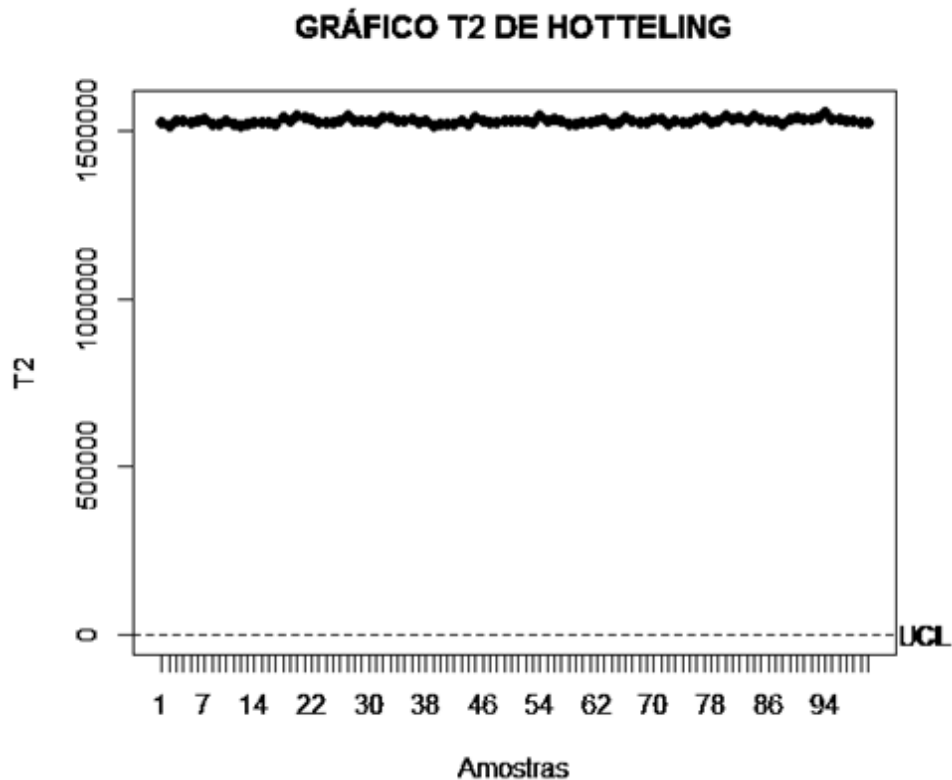
Sob esse último contexto, o presente trabalho propõe a aplicação do algoritmo *BACON* como uma forma bastante eficiente e robusta de identificar os possíveis *outliers* que o sistema de produção está gerando, fazendo com que haja uma distorção das estimativas das médias e variabilidade das medições envolvidas na análise da qualidade.

Portanto, dado que apenas a aplicação da estatística T^2 através da construção do seu gráfico não é suficiente para eliminar essas distorções, o algoritmo BACON será utilizado como ferramenta auxiliar para identificar as observações verdadeiramente discrepantes em relação ao conjunto de dados.

Este capítulo aborda a construção do gráfico de controle multivariado; a utilização da análise de componentes principais, a aplicação do algoritmo BACON e as vantagens de se utilizar esse algoritmo como suporte ao CEP.

Utilizando o software R, foi construído o gráfico de controle multivariado usando a estatísticas T^2 de Hotelling. A rotina usada para a obtenção dessa carta multivariada pode ser visualizada no apêndice C. Após esse procedimento, o gráfico obtido se apresenta na Figura 11 da seguinte forma:

Figura 11 - Gráfico T2 de Hotelling obtido no software R



O resultado da Figura 11 foi alcançado utilizando-se de cem amostras, com tamanho de cada amostra igual a quatro. Fato que resultou em um limite superior de controle no valor de 21.14315 e desvio-padrão igual a $5.839299e-10$. Nesse resultado, observam-se todos os pontos fora dos limites de controle devido à baixa correlação encontrada no grupo de dados amostrais como mostra os seguintes resultados do Quadro 1, pois quanto menor for essa correlação, maiores serão os valores das ordenadas dos pontos do gráfico.

Quadro 1 - Correlação das variáveis originais

Variáveis	AE	AF	AT	NA	CE	CF	CI	LI	LF	LE
AE	1	0,06	0,47	0,24	0,22	0,07	-0,10	-0,04	-0,07	0,18
AF	0,062	1	0,03	0,19	0,02	0,01	-0,21	-0,14	0,21	0,00
AT	0,476	0,03	1	0,16	0,19	0,14	-0,11	-0,12	-0,05	0,05
NA	0,240	0,19	0,16	1	0,04	0,14	-0,14	-0,05	-0,03	0,10
CE	0,220	0,02	0,19	0,04	1	0,09	0,12	-0,01	-0,04	0,07
CF	0,07	0,01	0,14	0,14	0,09	1	0,00	0,13	0,09	0,16
CI	-0,10	-0,21	-0,11	-0,14	0,12	-0,00	1	0,27	0,01	-0,14
LI	-0,04	-0,14	-0,12	-0,05	-0,01	0,13	0,27	1	-0,2	0,26
LF	-0,07	0,02	-0,05	-0,03	-0,04	0,09	0,01	-0,02	1	-0,04
LE	0,18	0,00	-0,05	0,10	0,07	0,16	-0,14	0,26	-0,4	1

Fonte: Elaboração própria.

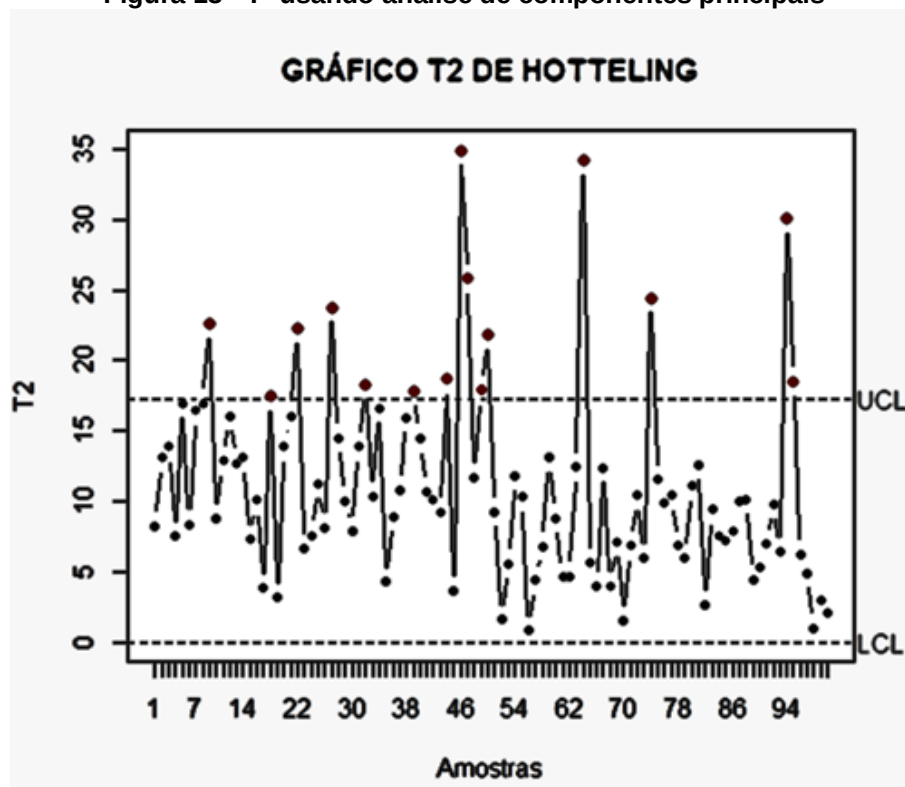
Percebe-se, ao analisar esses resultados, que todas as variáveis possuem pequenas correlações. E que, devido a esse nível baixo de correlação detectado, além do fato de que as observações não sejam independentes, os pontos situam-se completamente fora dos limites de controle do gráfico multivariado.

Um dos métodos utilizados para solucionar tal tipo de problema consiste na utilização da análise de componentes principais, que irá determinar um novo conjunto de variáveis que vai representar de maneira mais eficaz esse grupo amostral. Para a consecução de tal objetivo, é necessária a realização dos cálculos relativos à ACP, com o intuito de se obter esse novo grupo de variáveis cujo resultado pode ser visualizado na Figura 12.

Figura 12 - Variabilidade acumulada absorvida pelas componentes principais

Pela Figura 12, observa-se que mais de 86% das variabilidades do conjunto original dos dados são encontradas nas 7 primeiras componentes principais. Então, essas novas variáveis podem ser utilizadas para representar a estrutura original dos dados.

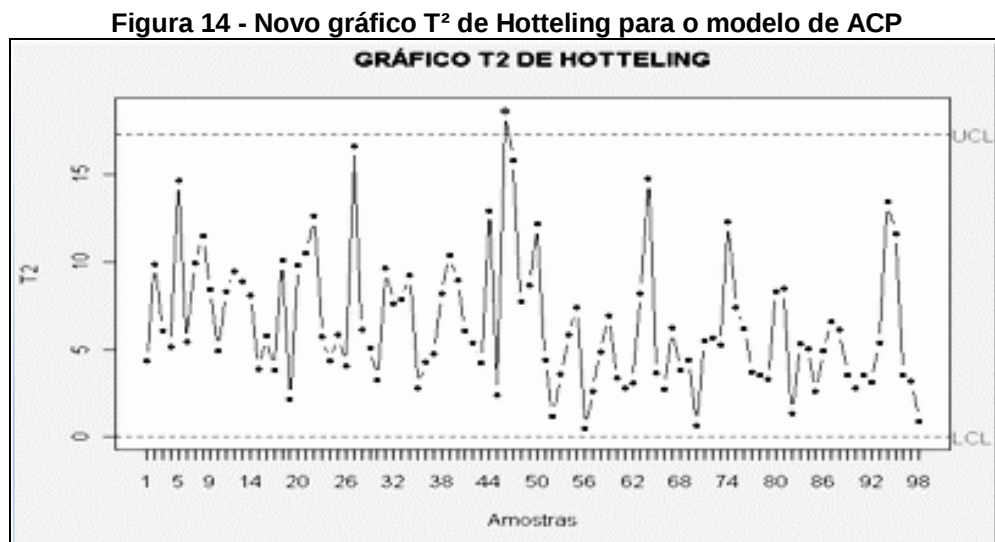
Para a obtenção dessas novas variáveis independentes, foi realizada uma análise em que foram geradas novas componentes a partir da combinação linear das variáveis originais. A Figura 13 mostra o gráfico T^2 de Hotelling obtido a partir da utilização das novas componentes principais.

Figura 13 - T^2 usando análise de componentes principais

A Figura 13 mostra vários pontos fora dos limites de controle, os quais parecem ser alarmes falsos.

Como a produção de garrafeiras consiste em um processo discreto parcialmente automatizado, foram construídos gráficos de controle da média com limites alargados para cada uma das variáveis em estudo (para evitar alarmes falsos), pois, segundo Costa *et al.*(2004), nesse caso ocorre o fenômeno de autocorrelacionamento de cada variável, constatado durante a realização da pesquisa. Também foram construídos gráficos da amplitude habituais e verificou-se que todas as variáveis se encontravam em controle estatístico.

O cálculo da estatística T^2 é feito a partir da matriz de covariância(s) (e variâncias). Os elementos presentes na diagonal principal da matriz representam as variâncias das observações para cada variável. Um meio relevante de reduzir o número de alarmes falsos é substituir o cálculo a partir da variabilidade dentro das amostras por aquele das variâncias das médias amostrais na diagonal principal da matriz de variâncias e de covariâncias. Ao utilizar-se esse procedimento, o novo gráfico obtido para as componentes principais é mostrado na Figura 14.



Na Figura 14, somente a amostra 47 apresentou um alarme de descontrole do processo. Devido ao grande número de amostras preliminares utilizadas e à proximidade desse ponto em relação ao limite superior de controle, pode-se presumir que o processo analisado encontra-se em controle estatístico. É necessário, contudo, utilizar um método mais robusto para diagnosticar se esse

ponto na zona de ação se constitui realmente numa observação atípica. Assim, para realizar esse diagnóstico, foi utilizado o algoritmo BACON, discutido na seção seguinte.

4.1 APLICAÇÃO DO ALGORITMO BACON

Nesta seção, aplicou-se o algoritmo BACON utilizando-se dos dados originais. Para isso, se calculou a distância de Mahalanobis representada pela equação 21 e obtiveram-se os seguintes resultados em destaque, mostrados na Tabela 2.

Tabela 2 - Resultados do cálculo das distâncias de Mahalanobis

[1]	3,190757	4,193056	2,937224	2,458624	5,011099	2,416606	2,94619
[8]	3,917175	4,192628	2,350552	2,72674	2,975305	3,451683	2,580228
[15]	3,375781	2,534331	2,817834	3,550438	1,859557	3,498141	3,076172
[22]	5,110338	3,855263	2,284701	2,856251	2,616241	3,848032	3,465635
[29]	2,387267	2,781031	2,998189	3,419194	2,92309	3,705722	2,124474
[36]	3,067514	2,84978	4,727649	3,595438	2,420288	2,106856	4,535428
[43]	3,954243	2,018061	5,685219	4,195297	3,171527	1,971052	5,267519
[50]	2,434471	2,095551	2,386788	2,166056	2,910859	3,307723	1,997703
[57]	2,28529	3,326652	3,579199	2,204814	2,682053	3,875724	5,631784
[64]	1,830398	3,241353	2,660389	3,385851	1,976373	2,694379	2,635354
[71]	2,353211	3,066074	4,602979	2,435067	3,037851	3,721725	3,014195
[78]	2,410303	2,028213	4,10506	1,653862	2,312025	2,122294	3,910023
[85]	1,970611	2,452155	4,063777	2,357852	2,003461	3,080057	2,077471
[92]	2,374561	4,056002	4,166771	2,646132	2,410995	1,020276	1,41443
[99]	2,230955	2,394147					

Fonte: Elaboração própria.

Diante desses resultados, inicia-se o cálculo do algoritmo BACON pelo primeiro passo que é a escolha de um subconjunto básico inicial livre de observações atípicas. Foram selecionadas as 30 menores distâncias que estão indicadas na Tabela 2, ou seja, foi retirada a quantidade de observações livres de *outliers* apontada pela equação $m = c \times p$.

Como o valor de c é um valor inteiro e escolhido pelo usuário, que por padrão é igual a 3, e p é o número de variáveis, tem-se que $m = c \times p = 30$ observações. Portanto, escolheram-se as 30 menores observações indicadas pelo cálculo da distância de Mahalanobis, mostrado no apêndice E, para compor o subconjunto

básico inicial. Calcularam-se, então, as distâncias de Mahalanobis para os 70 maiores valores restantes, ou seja, aqueles que ficaram de fora do subconjunto básico inicial composto por 30 registros.

Nesse novo cálculo, utilizou-se o vetor de média e a matriz de covariância do subconjunto básico inicial. Esses valores da média e da covariância podem ser visualizados através da Figura 14 e do Quadro 2, respectivamente:

Tabela 3 - Vetor de média do subconjunto básico inicial

VETOR DE MÉDIA DO SUBCONJUNTO BÁSICO INICIAL									
CE	CI	LE	LI	AT	AE	CF	LF	NA	AF
409,515	385,08	306,019	281,73	275,592	269,3133	378,3653	275,3041	137,888	6,275

Quadro 2 - Covariância do subconjunto básico inicial

Variáveis	CE	CI	LE	LI	AT	AE	CF	LF	NA	AF
CE	0,022	0,00	0,00	0,00	0,008	0,01	0,00	0,002	0,01	0,002
CI	0,000	0,06	-0,01	0,02	0,00	0,00	0,00	0,004	-0,02	-0,01
LE	0,009	-0,01	0,05	0,01	0,01	0,01	0,01	0,002	0,04	0,005
LI	0,000	0,02	0,01	0,04	0,00	0,00	0,00	0,003	0,01	0,000
AT	0,008	0,00	0,01	0,002	0,02	0,01	0,00	-0,003	0,02	0,004
AE	0,012	0,00	0,01	0,004	0,01	0,01	0,00	-0,003	0,02	0,003
CF	0,001	0,00	0,01	0,009	0,00	0,00	0,01	0,003	0,007	0,001
LF	0,00	0,00	0,00	0,003	0,00	0,00	0,00	0,015	0,00	0,000
NA	0,013	-0,02	0,04	0,01	0,02	0,02	0,00	-0,007	0,09	0,009
AF	0,002	-0,01	0,00	0,00	0,00	0,00	0,00	-0,002	0,009	0,006

Fonte: Elaboração própria.

De posse dos valores dos vetores da média e da covariância, recalcularam-se as distâncias de Mahalanobis para o subconjunto básico inicial e adicionou-se o menor valor dos pontos discrepantes a esse novo subconjunto que ficou com 31 elementos. O resultado desse novo cálculo pode ser visualizado na Tabela 4 a seguir.

Tabela 4 - Distância de Mahalanobis adicionando o primeiro ponto ao subconjunto básico inicial

[1]	39,20022	65,43254	41,50358	17,84432	1.061,77430	20,26267	30,78552
[8]	76,75220	121,96593	18,15824	58,69140	86,14220	16,79531	27,67661
[15]	17,42306	35,70712	40,69481	103,57532	27,41351	161,83122	47,24597
[22]	19,67719	20,78686	69,78046	29,22015	14,92408	37,06662	35,21048
[29]	47,74419	37,62933	36,56598	22,62826	87,75191	34,91015	81,80052
[36]	56,14648	190,35339	78,34159	39,37793	85,52954	21,16184	37,04760
[43]	44,01910	22,53080	24,38672	18,71886	55,50882	183,39083	41,13080
[50]	33,50997	25,03811	19,45213	23,02994	31,67306	63,00439	12,73336
[57]	32,73367	31,77964	29,23970	15,83097	52,82605	75,07972	18,55721
[64]	30,69310	67,46904	42,98091	72,89881	14,33651	31,42871	

Fonte: Elaboração própria.

Diante desse resultado da Tabela 4, o passo subsequente foi determinar o ponto de corte de acordo com a equação 24. O ponto de corte calculado para o primeiro ponto adicionado mostrou como resultado o valor de 42,31, de acordo com o cálculo obtido que pode ser visualizado no Quadro 3.

Quadro 3 - Resultado do cálculo do ponto de corte

p	α	n	α / n	$\chi_{p; \alpha}$	$cnpr$	Ponto de corte
10	0,1	100	0,001	29,5883	1,43	42,311269

Fonte: Elaboração própria.

Em que:

- p é o grau de liberdade;
- α é o nível de significância;
- n é a quantidade de amostras;
- $\chi_{p; \alpha}$ é o percentil $1 - \alpha$ de uma distribuição de Qui-quadrado com p graus de liberdade;
- $cnpr$ é o fator de correção.

Repetiu-se o procedimento adicionando todas as observações discrepantes ao subconjunto básico inicial até que se obtenha um valor maior que o ponto de corte, ou seja, até que o tamanho do subgrupo não se altere mais. Os elementos que ficarem de fora desse subgrupo serão rotulados como *outliers*.

Os resultados das iterações desse procedimento podem ser observados no Apêndice G, o qual mostra que à medida que se adicionam as observações ao subgrupo básico, o ponto de corte vai diminuindo. Essa diminuição não foi suficiente para que os novos pontos adicionados fossem maiores que o ponto de corte

calculado, evidenciando, assim, que o algoritmo BACON não detectou a presença de *outliers* nos dados amostrais colhidos.

4.2 UTILIZAÇÃO DO ALGORITMO BACON COMO SUPORTE AO CEP

A utilização de um algoritmo de busca sucessiva é crucial para o auxílio ao controle estatístico do processo. Tal ferramenta se torna imprescindível na medida em que a análise dos processos produtivos é feita através da composição de diversas variáveis que precisam ser controladas simultaneamente.

O presente trabalho de pesquisa adotou o algoritmo BACON, que é um algoritmo de busca sucessiva em que parte de um pequeno subconjunto, comprovadamente livres de *outliers*, dos dados originais e vão adicionando novas informações, que também não são *outliers*, a esse subconjunto inicial até que nenhuma informação possa mais ser absorvida.

Uma vantagem dessa ferramenta, é que o algoritmo BACON combate a ocorrência do mascaramento e esmagamento, que são anomalias que atacam as estimativas da média e da covariância, permitindo que o tomador de decisão pense que o processo esteja em controle quando na verdade ele não está, ou, ao contrário, pensar que o processo esteja fora de controle, quando na verdade ele não está.

4.3 APLICAÇÃO DO ALGORITMO BACON PARA UM ESTUDO SIMULADO

Os resultados obtidos mostraram que não há presença de outliers nos dados analisados. Para verificar-se a eficácia da aplicação do algoritmo BACON, foram alterados alguns valores do desvio-padrão e da média em 5% do conjunto de dados originais para verificar se tal método consegue realmente identificar *outliers* na situação estudada. Para isso foi desenvolvido um algoritmo de simulação que tem os seguintes parâmetros de entrada.

k: percentual de vetores a serem alterados, $0 \leq k \leq 1$;

B: Número de reamostragens dos vetores a serem alterados;

TM: Taxa de inflacionamento da média, $TM > 1$;

TD: Taxa de inflacionamento do desvio padrão, $TD > 1$;

BD: Base de dados contendo a p variáveis;

n: é o tamanho da amostra do conjunto de dados (BD);

PASSO 1 (obtem-se a media e dp do vetor): Calcular a média (m_j) e o desvio padrão (dp_j) de cada uma das p variáveis da base de dados; $1 \leq j \leq p$;

PASSO 2 (seleção dos vetores): Selecionar, aleatoriamente, uma amostra de tamanho v ($v = k \times n$) de vetores da base de dados;

PASSO 3 (produzindo os outliers): Para cada vetor selecionado ($S_r, r = 1, 2, \dots, v$), gerar valores novos (inflacionados) e aleatórios para cada uma das suas p medições usando a seguinte regra:

$$x_{rj} = (TM \times m_j - TD \times dp_j) + 2 \times TD \times dp_j \times U_j, \quad 1 \leq j \leq p \text{ e } 1 \leq r \leq v$$

sendo U_j uma distribuição uniforme no intervalo (0;1)

PASSO 4 (substituindo os valores originais): Substituir na base de dados os vetores iniciais pelos vetores com os valores inflacionados;

PASSO 5 (detecção do outliers - BACON): Com a base de dados modificada, aplicar o algoritmo de BACON para verificar se as observações discrepantes geradas serão detectadas.

O resultado da simulação proposta pode ser visualizado na Tabela 5 a seguir:

Tabela 5 - Desempenho do algoritmo BACON com aumento na média e no desvio-padrão em 5% dos vetores do conjunto de dados originais

% inflação da média e do desvio padrão	% de Outliers detectados pelo BACON
50%	60%
60%	80%
70%	80%
80%	100%

Fonte: Elaboração própria.

Ao analisar a Tabela 5, percebe-se que à medida que a média e o desvio-padrão de 5% dos vetores do conjunto original de dados aumenta, há um acréscimo significativo na presença de outliers detectados, mostrando a eficiência do algoritmo BACON em identificar esses pontos discrepantes.

A Tabela 6 apresenta um exemplo dessa simulação que mostra as ocorrências de outliers.

Tabela 6 - Resultado que mostra a ocorrência de *outliers* com 50% de inflação da média e desvio padrão em 5% dos dados originais

[1]	49,20022	28,43254	5597,504	43,84432	32,06177	70,26267	39,78552
[8]	22,7522	39,96593	26,15824	30,6914	3727,142	18,79531	18,67661
[15]	44,42306	10,70712	48,69481	61,57532	36,41351	32,83122	6330,246
[22]	17,67719	16,78686	43,78046	29,22015	24,92408	44,06662	30,21048
[29]	29,74419	32,62933	24,56598	46,62826	13,75191	24,91015	22,80052
[36]	43,14648	62,35339	78,34159	37,37793	51,52954	24,16184	62,0476
[43]	37,0191	51,5308	30,38672	31,71886	23,50882	137,3908	20,1308
[50]	48,1751	14,03811	25,45213	25,02994	54,67306	20,00439	39,73336
[57]	91,73367	87,77964	22,2397	23,83097	30,82605	33,07972	27,60467
[64]	101,6931	20,46904	51,98091	23,89881	50,33651	69,42871	

Fonte: Elaboração própria.

O resultado da Tabela 6 mostra que os pontos 3, 12 e 21 são *outliers*, segundo o algoritmo BACON, demonstrando que tais ocorrências são perfeitamente detectadas por esse algoritmo de busca sucessiva.

CAPÍTULO 5 - CONCLUSÕES E RECOMENDAÇÕES DE TRABALHOS FUTUROS

Neste Capítulo, apresentam-se as considerações finais que mostram a competência do engenheiro de produção para atuar no desenvolvimento de métodos e meios de melhorias ao processo produtivo, implementando modelos estatísticos e matemáticos para se conseguir tal objetivo.

Nesta dissertação, fica evidenciada que a utilização da estatística é uma ferramenta imprescindível para a consecução dos objetivos pretendidos. Tal ferramenta é explicitada na aplicação de um algoritmo de buscas sucessivas BACON para apoio à tomada de decisão no controle estatístico do processo.

Diante disso, essa seção também aponta algumas recomendações para que esse tema seja mais bem estudado e difundido no âmbito da atividade acadêmica da engenharia de produção.

Existem diversos métodos para o controle de processos produtivos que podem e devem ser implementados em qualquer unidade fabril. Para tanto, é necessário um bom planejamento e uma análise aprofundada dos meios e métodos que compõem o sistema produtivo. Para que tais ações sejam efetivamente executadas, é necessário um profissional que lide com os diversos fatores que podem ser encontrados no chão de fábrica.

Esse tipo de profissional é o engenheiro de produção que tem a habilidade de planejar, organizar e executar as atividades de qualquer processo produtivo. Para isso lança mão de métodos que atendam às necessidades primordiais para o atingimento de objetivos como redução de custos e eliminação ou diminuição de desperdícios do processo de fabricação.

A estatística é uma ferramenta muito poderosa de apoio ao engenheiro de produção para a consecução desses objetivos pretendidos pela empresa, que através de métodos estatísticos pode ajudar na tomada de decisão para a eliminação ou minimização de disfunções presentes no sistema fabril.

O presente trabalho de pesquisa teve o objetivo de aplicar um algoritmo de busca sucessiva para ajudar o controle estatístico do processo a detectar pontos discordantes que só com a avaliação do gráfico de controle multivariado não são detectados.

Foi necessário então não abrir mão de um algoritmo que combatesse todos esses tipos de ocorrências, o qual proporciona a detecção de pontos aberrantes combatendo os fenômenos do mascaramento e esmagamento que interferem nas estimativas da média e no desvio-padrão em dados multivariados.

Como resultado desta pesquisa, o conjunto de dados originais utilizados não proporcionou a verificação da presença de *outliers* neste estudo, mostrando que o processo de produção das garrafeiras se encontra em controle estatístico, não necessitando nenhuma intervenção no sentido de provocar aumento de custos ou desperdícios, mas apenas a atenção para o controle dos meios e métodos para que o gerente de produção possa controlá-los e assim evitar que o processo entre em descontrole.

O processo de simulação teve o intuito de mostrar que o algoritmo BACON é robusto na detecção desses pontos discordantes. Esse tipo de algoritmo se mostra o mais indicado para a aplicação em processos fabris que dependam de mais de uma variável, pois leva em conta os fenômenos do mascaramento e esmagamento que mudam as estimativas da média e do desvio-padrão em processos multivariados.

Diante disso, esta pesquisa conseguiu atingir seu objetivo, que foi o de aplicar o algoritmo BACON no processo produtivo das garrafeiras, para verificar a presença de *outliers*, contribuindo para um dos objetivos das empresas de produtos plásticos, que consiste na diminuição dos custos e eliminação dos desperdícios, podendo ser estendido a empresas de outros setores, desde que trabalhem simultaneamente com diversas variáveis.

O presente projeto de pesquisa não esgota o estudo sobre esse tipo de aplicação. Esta pesquisa pode e deve ser desenvolvida e aperfeiçoada, podendo gerar vários trabalhos acadêmicos para melhorar a utilização da ferramenta do controle estatístico do processo, não só em produtos plásticos, mas também em qualquer processo de fabricação que dependa simultaneamente de diversas variáveis.

Sugere-se, para futuros trabalhos, um processo produtivo com uma quantidade maior de variáveis a serem estudadas, para verificar a eficácia do algoritmo Bacon nesse tipo de situação, bem como a tentativa de implementação de outros algoritmos e ferramentas estatísticas multivariadas.

Outra sugestão é a de que, a partir da identificação de *outliers* no processo de fabricação estudado, deve-se tentar identificar qual variável está provocando a

disfunção no processo produtivo. Para isso, pode-se utilizar o teste de comparações múltiplas de Bonferroni, que faz um ajuste estatístico para um nível de significância (α) em relação ao número de testes realizados simultaneamente, para um conjunto de dados (SOUZA; RIGÃO, 2005).

O teste de Haite e Tsui estende essa ideia estabelecendo limites através de intervalos de confiança para a média de cada uma das variáveis.

Uma alternativa a este trabalho de pesquisa é a de que, após a detecção desses pontos discordantes, se aplique a ferramenta gerencial PDCA (*Plan*-planejar, *Do*-fazer, *Check*-averiguar e *Act*-agir), que é um processo de planejamento e controle para a eliminação de custos e desperdícios, o qual irá proporcionar ao gerente de produção estabelecer planos de ação, colocar em prática tais planos, acompanhar se o que foi planejado está sendo executado e padronizar o método ou corrigi-lo, caso não esteja de acordo com o planejado.

Portanto, O presente trabalho de pesquisa pretendeu, além de mostrar a aplicação de um algoritmo de busca sucessiva no controle estatístico do processo multivariado, suscitar a partir dele novas pesquisas no âmbito do CEP, para que tal ferramenta seja mais eficaz na sua aplicação, buscando cada vez mais auxiliar no processo de tomada de decisão, objetivando a redução de fenômenos que aumentem a variabilidade do processo produtivo e provoquem elevação de custos e de desperdícios, gerando produtos não conformes.

REFERÊNCIAS

- ASSOCIAÇÃO BRASILEIRA DE INDÚSTRIA DO PLÁSTICO. **Notícias gerais**. São Paulo. Disponível em: <<http://www.abiplast.org.br>>. Acesso em: 15 jul. 2011.
- BARNARD, J.P; ALDRICH, C. **Detecting outliers in multivariate process data by using convex hulls**. European Symposium on Computer Aided Process Engineering. S. Pierucci, 2000
- BILLOR, N.; HADI, A. S.; VELLEMAN, P. F. BACON: Blocked adaptive computationally-efficient outlier nominators. **Computational Statistics and Data Analysis**, v.34, n.3, p. 279-298, 2000.
- BORNIA, A. C. **Mensuração das perdas dos processos produtivos: uma abordagem metodológica de controle interno**. Tese. Programa de Pós-graduação em Engenharia de Produção. Universidade federal de Santa Catarina, Florianópolis, 1995.
- BRINSON, J. A. **Contabilidade por atividades: uma abordagem de custeio baseado em atividades**. São Paulo: Atlas, 1996.
- CHEN, Y-K.; HSIEH, K-L. Hotelling's T^2 charts with variable sample size and control limit. **European Journal of Operation Research**.v.182, p.1251-1262, 2007.
- CHEN, T. On reducing false alarms in multivariate statistical process control. **Chemical Engineering Research and Design**.v.88, p. 430-436, 2010.
- CHEN, L-H.; WANG, T-Y. Artificial neural networks to classify mean shifts from multivariate χ^2 chart signals. **Computer & industrial engineering**. v.47, p.195-205, 2004.
- COLENGHI, F.K.R. **Estudo comparativo de testes de hipóteses multivariados para o vetor de médias via simulação de Monte Carlo**. Dissertação. Programa de Pós-Graduação em estatística, Universidade federal de Minas Gerais, Belo Horizonte, 2008.
- COSTA, A. F.B.; EPPRECHT, E.K.; CARPINETTI, L.C.R. **Controle Estatístico de Qualidade**. São Paulo: Atlas, 2004.
- FIGUEIRA, M.M.C. Identificação de Outliers. **MILLENIUM**. n.12, 1998.
- FILZMOSER, P, GARRETT, R.G, REIMANN, C. Multivariate outlier detection in exploration geochemistry. **Computers & Geosciences**.v.31, p.579-587, 2005.
- FOSS, A; ZAIANE, O.R. Class separation through variance: a new application of outlier detection. **Know inf. Syst**.v.29, p.565-596, 2011.

GARRET, R.G. The chi-square plot: a tool for multivariate outlier recognition. **Journal of Geochemical Exploration**.v.32, p.319-341,1989.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 4. ed. São Paulo: Atlas, 1994.

GODOY, A. S. Introdução à pesquisa qualitativa e suas possibilidades. **Revista de Administração de Empresas**. São Paulo, v.35, n.2, p. 57-63, 1995.

GUH, R-S.; SHIUE, Y-R. An effective application of decision Tree learning for on-line detections of mean shifts in multivariate control charts. **Computer & industrial engineering**. v.55, p.475-493, 2008.

HAIR, J.R, J.F.; BLACK, W.C.; THATAM; R.L. ANDERSON,R.E; BABIN. B.J. **Análise multivariada de dados**. 6. ed. São Paulo: bookman, 2009.

HOTTELING, H. **Multivariate Quality Control Illustrated by the Testing of Sample Bombsight in Techniques of Statistical Analysis**. New York: McGraw Hill, 1947.

JOHNSON. R.A.;WICHERN. D.W. **Applied multivariate statistical analysis**. 6. ed. New Jersey: Prentice Hall, 2007.

KONRATH, A.C. **Decomposição da estatística do gráfico de controle multivariado T^2 de Hoteling por meio de um algoritmo computacional**. Dissertação. Programa de Pós-Graduação em engenharia de produção e sistemas, Universidade Federal de Santa Catarina, Florianópolis, 2002.

LIMA, Márcio B. F. ; SANTOS, R. L. S. ; SOUZA, E. L. ; SILVA, L. B. Aplicações do controle estatístico multivariado da qualidade: Monitoramento de garrafeiras plásticas numa empresa do Estado da Paraíba. In: ENCONTRO NACIONAL EM ENGENHARIA DE PRODUÇÃO - ENEGEP, 2012. Bento Gonçalves. **Anais...** Bento Gonçalves, 2012-a.

LIMA, Márcio B. F. ; SANTOS, R. L. S. ; SOUZA, E. L. ; SILVA, L. B. Aplicações do controle estatístico multivariado da qualidade: controlando garrafeiras plásticas numa empresa industrial. **Cadernos do IME. Série Estatística**, v. 32, p. 1-14, UERJ, Rio de Janeiro, 2012-b.

MEDEIROS, O.M. **Alternativas de Ultimação de Resíduo Plástico Termofixo**. Tese. Programa de Pós-Graduação em Ciências e engenharia de Materiais, Universidade Federal do Rio grande do Norte, Natal, 2005.

MONTGOMERY, Douglas C. **Statistical Quality Control**. 5. ed. Wiley: Danvers, 2005.

MONTGOMERY, D.C. **Introdução ao controle estatístico de qualidade**. 4. ed. Rio de Janeiro: LTC, 2009.

PALADINI, E.P. **Gestão da qualidade: teoria e prática**. São Paulo: Atlas, 2004.

RAYKOV, T.; MARCOULIDES, G. A. **An introduction to applied multivariate analysis**. New York: Taylor e Francis Group, 2008.

RENCER, ALVIN. C. **Methods of multivariate analysis**. 2. ed. New York: John Wiley & Sons, 2002.

RODRIGUES, W. C. **Metodologia científica**. Rio de Janeiro. Disponível em: <<http://www.faecetec.rj.gov.br/>>. Acesso em: 12 jul. 2011.

SAMOHYL, R. W. **Controle estatístico de qualidade**. Rio de Janeiro: Campus, 2009.

SILVA, E. L.; MENEZES, E. M.; **Metodologia da pesquisa e elaboração de dissertação**. 3 ed. Florianópolis: Atlas, 2001.

SIMOGLU, A.; MARTIN, E. B.; MORRIS, A. J. Multivariate statistical process control of an industrial fluidised-bed reactor. **Control Engineering Practice**. v.8, p.893-909, 2000.

SOUZA, A. M.; RIGÃO, M. H. Identificação de variáveis fora de controle em processos produtivos multivariados. **Revista Produção**, v.15, n.1, p.74-86, 2005.

TIMM, NEIL. H. **Applied multivariate analysis**. New York: Springer-Verlag, 2002.

TODOROV, V.; TEMPL, M.; FILZMOSER, P. Detection of Multivariate outliers in Business Survey Data With Incomplete Information. **Adv. Data. Anal. Classif.** v.5, p.37-56, 2011.

TREASURE, R. J.; KRUGER, U.; COOPER, J. E. Dynamic multivariate statistical process control using subspace identification. **Journl of Process Control**, v.14, p. 279-292, 2004.

VERDONCK; T.; WOUWE, M. V. Detection and correction of outliers in the bivariate chain-ladder method. **Insurance: Mathematics and Economics**. v.49, p. 188-193, 2011.

VICINI, LORENA; SOUZA, A. M. **Análise multivariada da teoria à prática**., Universidade Federal de Santa Maria, 2005. Monografia- Especialização em estatística e modelagem quantitativa

VIEIRA, S. **Estatística para a qualidade**. 2. ed. Rio de Janeiro: Campus, 2012.

WILLIAMS, J. D.; WOODALL, W. H.; BIRCH, J. B.; SULLIVAN, J. H. **On the Distribution of Hotelling's T^2 statistic Based on the Successive Differences Covariance Matrix Estimator**. GE Global Research. VA 24061-0439, 2006.

WOODALL, W. H. Controversies and Contradictions in Statistical Process Control. **Journal of Quality Technology**. v.32, n.4, p.341-350, 2000.

APÊNDICE A – Pontos for a dos limites de controle nas 25 primeiras amostras de n=4

Pontos fora dos limites de controle nas 25 primeiras amostras de n=4

	CE	CI	LE	LI	AT	AE	CF	LF	AN	AF
	409,65	385,5	306,5	282,125	275,15	269,075	378,275	275,4	138,025	6,225
	409,45	385,2	305,695	282,225	275,4	269,1	378,175	275,325	137,95	6,275
	409,275	385,47	305,975	281,6	275,25	268,92	378,3	275	138,2775	6,425
	409,325	385,425	305,45	281,9	275,375	269,175	378,6	275,275	137,675	6,2775
	409,375	385,55	305,65	281,825	275,775	269,425	378,375	275,25	137,95	6,325
	408,95	385,4	305,575	281,175	275,95	269,45	378,2	275,075	138,2	6,35
	409,625	384,825	305,85	281,575	275,6	269,25	378,35	274,95	137,9225	6,425
	409,575	384,425	306,225	281,325	275,725	269,375	378,35	275,2	137,65	6,15
	409,475	385,375	306,35	281,425	275,5375	268,95	378,4	275,605	138,135	6,25
	409,3	385,125	305,7	282,4	275,3	269,25	378,7	275,35	138,3	6,125
	409,5	384,525	306,125	281,625	275,8	269,4	378,46	275,175	138,025	6,375
	409,4	384,55	305,9	281,35	275,725	269,275	378,35	275,15	138,375	6,5
	409,15	385,275	305,95	281,95	275,4	269,6	378,725	275,425	138,475	6,225
	409,35	385,05	305,575	281,5	275,15	268,875	378,5	275,425	138,225	6,225
	409,575	384,5	306,15	281,55	275,575	269,325	378,25	275	138,1	6,5
	409,675	384,625	305,925	281,55	275,5	269,225	378,175	275,675	137,95	6,15
	409,025	385,3	306,225	282,2	275,5	269,225	378,55	275,375	137,9525	6,275
	409,475	385,25	306,2	281,825	275,25	269,125	378,575	275,225	138,175	6,2
	409,025	385,3	305,6	281,8	275,55	269,125	378,175	274,9	137,05	6,175
	409,325	385,3	305,85	282,075	275,575	269,275	378,3	275,375	137,75	6,325
	409,2	384,925	306,3	281,6	275,825	269,35	378,3	275,275	136,95	6,275
	409,675	384,575	305,975	281,4	275,7	269,2	378,3	275,15	137,2775	6,4125
	409,375	385,225	305,675	281,65	275,325	269,025	378,6	275,375	137,425	5,525
	409,4	385,225	305,7	281,375	275,125	269,175	378,425	275,075	137,85	6,15
	409,575	384,575	306,175	281,55	275,7	269,275	378,375	275,4	138,2	6,375
Média	409,389	385,0598	305,8965217	281,7083333	275,5192708	269,2114583	378,3914	275,2397917	137,913125	6,29125
D.P. Intra	0,223409	0,394366	0,277467	0,329853	0,231222438	0,153796	0,235843	0,360207	0,430832	0,195747
2*DPXbar.	0,409033413	0,739326721	0,60676066	0,649358658	0,465901368	0,289294181	0,321422878	0,354112229	0,805077974	0,221683284
DPXbarra	0,204516707	0,36966336	0,30338033	0,324679329	0,232950684	0,144647091	0,160711439	0,177056115	0,402538987	0,110841642
LsC	409,980549	385,3949135	306,3127222	282,2031128	275,7499648	269,4421523	378,7451645	275,7801022	138,559373	6,5848705
LIC	408,797451	384,468251	305,4803212	281,2135538	275,2885768	268,9807643	378,0376355	274,6994812	137,266877	5,9976295
LsC alarg.	410,0025501	386,1687901	306,8066627	282,6823713	276,2181229	269,6453996	378,8735343	275,77096	138,3952593	6,623774926
LIC alarg.	408,7754499	383,9508099	304,9863807	280,7342953	274,8204188	268,7775171	377,9092657	274,7086233	136,705508	5,958725074

APÊNDICE B – Testes de normalidade, especificações e coeficientes de capacidade do processo

Testes de normalidade, especificações e coeficientes de capacidade do processo

	CE	CI	LE	LI	AT	AE	CF	LF	AN	AF
Distorção	-0,552495913	-0,520173121	0,085800172	0,508579395	-0,140036857	-0,448129524	0,560782895	-0,194655674	-1,083088284	0,34402716
Curtose	-0,292975318	-1,295315501	-0,839861359	-0,574859179	-0,874218628	0,074890385	-0,529319602	-0,488303378	0,732102865	-0,802290589
Jarque-Bera	1,361293198	2,875169323	0,70419654	1,365075081	0,842699496	0,80888885	1,602176107	0,390003514	5,228295527	1,117088935
		Amostras de n = 4, m = 25			Especificação		Coef. Capacidade			
σ_s		μ_s	Máximo	Mínimo	LSE	LIE	Cps	Cpi	Cpk	Cp
0,223409	CE	409,389	409,675	408,95	410,481	408,434	1,62929873	1,424890373	1,424890373	1,527094552
0,394366	CI	385,0598	385,55	384,425	386,915	383,301	1,568086498	1,486605505	1,486605505	1,527346002
0,277467	LE	305,8965217	306,5	305,35	307,431	304,607	1,84343635	1,549159286	1,549159286	1,696297818
0,329853	LI	281,7083333	282,4	281,175	282,976	280,57	1,281042026	1,150344076	1,150344076	1,215693051
0,231222438	AT	275,5192708	275,95	275,125	276,437	274,689	1,323010538	1,196929445	1,196929445	1,259969991
0,153796	AE	269,2114583	269,45	268,875	269,9361575	268,5898425	1,570693574	1,347273517	1,347273517	1,458983545
0,235843	CF	378,3914	378,725	378,175	379,239	377,631	1,197972097	1,074726266	1,074726266	1,136349181
0,360207	LF	275,2397917	275,605	274,9	276,9601	273,658	1,591962708	1,46378024	1,46378024	1,527871474
0,430832	AN	137,913125	138,475	136,95	139,77	135,828	1,436658218	1,613254509	1,436658218	1,524956364
0,195747	AF	6,29125	6,5	6,125	6,923	5,555	1,075793414	1,253744204	1,075793414	1,164768809

APÊNDICE C - Dados originais

CE	CI	LE	LI	AT	AE	CF	LF	NA	AF
409,425	385,65	306,06	282,125	275,275	268,925	378,275	275,275	137,7775	6,225
409,3	385,025	306,05	281,475	275,6125	269,3	378,375	275,13	137,85	6,225
409,325	384,825	306,3	281,9	275,325	269,325	378,475	275,375	137,875	6,225
409,325	384,9	306,075	281,8	275,25	269,2	378,3	275,25	137,65	6,25
409,375	385,35	306,125	282,1	275,45	269,2	378,4	275,175	136,975	6,2
409,525	385,075	306,075	281,725	275,35	269,025	378,375	275,475	137,225	6,0875
409,675	385,5	306,425	281,825	275,55	269,175	378,575	275,075	137,575	6,275
409,625	384,875	305,925	281,575	275,375	269,1	378,6	275,25	137,5	6,025
409,55	385,25	306,35	281,7	275,4	269	378,15	275,325	137,15	6,025
409,25	384,925	306,075	281,625	275,525	269,125	378,275	275,175	137,825	6,375
409,675	385,375	305,975	281,8	275,325	269,075	378,325	275,3	137,925	6,475
409,325	385,325	306,275	281,875	275,525	269,275	378,3	275,475	137,525	6,225
409,625	385,375	306,2	281,975	275,65	269,35	378,35	275,225	137,575	5,9775
409,175	385,625	306	281,925	275,325	269,1	377,95	275,225	137,55	6,225
409,625	385,225	306,3	281,95	275,35	269,425	378,425	275,3	138,15	6,225
409,45	385,5	305,95	282,075	275,625	269,15	378,5	275,125	137,825	6,05
409,475	385,55	306,2	282,2	275,45	269,2	378,575	275,225	137,55	6,1
409,425	385,275	306,15	282,025	275,55	269,4	378,425	275,35	137,675	6,175
409,725	385,65	306,05	282,075	275,6	269,275	378,625	275,275	137,625	6,1
409,85	385,475	306,225	282	275,55	269,25	378,375	275,275	137,5	6,225
409,55	385,175	306,1	281,725	275,525	269,25	378,45	275,375	137,5	6,15
409,625	385,425	305,825	281,525	275,25	269,1	378,025	275,225	137,7	6,2
409,55	385,175	305,925	281,8	275,4	269,225	378,15	275,125	137,45	6,05
409,6	385,375	305,875	281,975	275,5	269,375	378,025	275,325	136,775	5,975
409,375	385,525	306,175	281,7	275,25	269,37	378,425	275,05	138,325	6,3
409,15	385,375	305,7	281,525	275,85	269,45	378,325	275,15	138,0475	6,35
409,5	384,6	305,75	281,65	275,675	269,3	378,485	275,2	138	6,3
409,375	384,475	305,85	281,575	275,475	269,2	378,275	275,075	138,25	6,425
409,05	384,775	305,825	281,65	275,625	269,25	378,325	275,175	138,05	6,25
409,25	384,575	305,95	281,475	275,625	269,275	378,225	275,375	137,675	6,45
409,7	384,85	306,2	281,475	275,625	269,375	378,475	275,125	138,2	6,35
409,45	385	305,625	281,4	275,65	269,275	378,35	275,225	137,925	6,375
409,55	384,775	306,125	281,525	275,65	269,25	378,825	275,375	137,65	6,425
409,45	385	306,275	281,375	275,6	269,225	378,175	275,4	137,725	6,3
409	384,575	306,125	281,5	275,5	269,45	378,25	275,35	138,575	6,125
409,225	384,85	306,15	281,7	275,6	269,325	378,45	275,4	137,8	6,375
409,625	385,1	306,475	282,05	275,925	269,55	378,6	275,2	138,125	6,25
409,3	384,875	306,2	281,625	275,525	269,425	378,5	275,3	137,15	6,325
409,425	384,9	306,125	281,65	275,625	269,375	378,4	275,325	137,825	6,3
409,75	385,025	306,1	281,75	275,85	269,325	378,475	275,45	138,1	6,325
409,475	385,225	305,925	281,6	275,75	269,425	378,625	275,575	138,1	6,15

409,1	384,975	306,075	281,925	275,775	269,4	378,625	275,225	138,05	6,575
409,5	385,2	306,1	281,85	275,775	269,3	378,45	275,475	138,075	6,4
409,875	385,375	306,2	281,825	275,7	269,45	378,375	275,4	137,85	6,25
409,5	384,95	306,025	281,525	275,525	269,3	378,6	275,575	137,375	6,325
409,475	385,1	305,8	281,75	275,8	269,425	378,425	275,175	137,65	6,2
409,575	385,425	305,575	281,6	275,75	269,55	378,05	275,15	137,775	6,475
409,525	385,1	305,525	281,525	275,9	269,625	378,3	275,125	137,5	6,3
409,375	385,47	305,41	281,825	275,525	269,2	378,35	275	138,2	6,47
409,45	385,1	306,075	281,775	275,775	269,375	378,55	275,525	137,95	6,375
409,6	385	305,475	281,025	275,725	269,275	378,425	275,45	138,11	6,45
409,275	385,025	305,675	281,875	275,7	269,175	378,65	275,325	138,575	6,425
409,375	385,075	306,125	281,675	275,475	269,05	378,375	275	138,0525	6,175
409,25	385,45	305,95	281,9	275,575	269,175	378,3	275,075	137,4	6,275
409,525	385	305,875	281,475	275,625	269,25	378,55	275,325	137,925	6,3
409,3	385,05	305,775	281,85	275,675	269,225	378,125	275,1	137,575	6,325
409,375	384,65	305,6	281,25	275,675	269,3	378,275	275,35	137,825	6,35
409,25	385,175	306,25	282,4	275,425	269,075	378,6	275,525	137,6	6,2
409,175	384,75	306,05	281,925	275,475	269,2	378,1	275,45	138,15	6,25
409,25	384,875	305,875	281,675	275,35	269,25	378,275	275,45	138,075	6,35
408,925	384,85	306,275	282,15	275,525	269,2	378,3	275,15	137,9	6,25
409,875	385,375	305,9	281,725	275,55	269,45	378,4	275,525	138,15	6,325
409,475	385,55	305,9	281,925	275,55	269,225	378,35	275,225	137,7	6,4
409,65	385,1	306,15	281,675	275,425	269,25	378,525	275,525	138,225	6,4
409,5	385,55	305,675	281,95	275,725	269,275	378,35	275,375	138	5,5075
409,5	385,2	305,95	281,575	275,575	269,3	378,275	275,5	137,8	6,325
409,75	385,3	306,05	282,1	275,45	269,325	378,425	275,55	138,025	6,4
409,55	385,2	306,35	282,05	275,675	269,4	378,425	275,625	137,875	6,225
409,55	385	306,325	282,05	275,5	269,325	378,35	275,475	137,8	6,25
409,55	385,2	306,025	282,225	275,5	269,25	378,4	275,275	137,475	6,2
409,525	385,275	305,9	281,775	275,575	269,325	378,125	275,125	137,45	6,175
409,375	385,025	305,975	281,95	275,5	269,3	377,925	274,95	137,725	6,2
409,65	385,225	305,975	281,625	275,775	269,35	378,675	275,15	137,4	6,075
409,425	385,025	305,925	281,95	275,725	269,3	378,125	275,35	138,05	6,25
409,35	385,125	305,425	281,8	275,625	269,175	378,45	275,075	137,475	6,2825
409,475	385	305,825	281,675	275,675	269,225	378,45	275,275	138,3	6,1
409,45	384,95	305,825	281,65	275,425	269,125	378,45	275,225	138,65	6,325
409,525	385,125	305,9	281,625	275,475	269,25	378,55	276,175	138,05	6,275
409,425	385,1	305,875	281,65	275,625	269,175	378,375	275,075	137,4525	5,525
409,525	384,725	305,7	281,6	275,55	269,2	378,325	275,075	137,8	6,325
409,275	384,7	306,25	281,725	275,375	269,225	378,075	275,1	137,65	6,275
409,25	384,725	306,4	281,7	275,525	269,1	378,35	275,35	137,625	6,275
409,225	384,8	305,575	281,15	275,35	269,025	378,125	275,45	137,55	6,25
409,45	384,75	306,325	281,775	275,6	269,15	378,475	275,125	138,2	6,225
409,075	384,9	306,25	281,9	275,55	269,225	377,725	275,2	137,725	5,425
409,6	385	306,4	282,025	275,65	269,375	378,775	275,4	138,325	6,125

409,45	385,075	305,925	281,65	275,675	269,25	378,45	275,4	137,85	6,275
409,275	385,45	306,1	281,825	275,55	269,275	378,575	275,275	138,025	6,225
409,65	385,05	306,125	281,875	275,675	269,375	378,4	275,5	137,9	6,125
409,725	385,225	306	281,85	275,575	269,2	378,4	275,375	137,65	6,4
409,475	385,175	306,075	282,1	275,575	269,125	378,625	275,45	137,9	6,3
409,775	385,275	306,2	281,975	275,775	269,3	378,575	275,45	137,925	6,275
409,175	385,375	306,45	281,9	275,45	269,3	378,775	275,675	137,8	6,275
409,7	385,175	306,1	281,775	275,625	269,2	378,325	275,35	137,1	6,175
409,575	384,9	305,9	281,65	275,575	269,275	378,425	275,3	137,525	6,25
409,475	385	306,175	282,05	275,425	269,275	378,35	275,3	137,725	6,375
409,55	385,15	305,425	281,575	275,75	269,25	378,475	275,35	137,325	6,3
409,325	384,9	305,975	281,675	275,725	269,25	378,8	275,7	137,575	6,175
409,5	385,425	306,075	281,725	275,625	269,425	378,375	275,25	138,2	6,125
409,625	385,25	306,15	282,05	275,525	269,375	378,2	275,5	138	6,175

APÊNDICE D - Rotina para a construção do gráfico T²

```

ae<-read.table("ae.txt", head=T)
af<-read.table("af.txt", head=T)
at<-read.table("at.txt", head=T)
an<-read.table("an.txt", head=T)
ce<-read.table("ce.txt", head=T)
cf<-read.table("cf.txt", head=T)
ci<-read.table("ci.txt", head=T)
li<-read.table("li.txt", head=T)
lf<-read.table("lf.txt", head=T)
le<-read.table("le.txt", head=T)
dados<-read.table("dados.txt", head=T)
attach(ae)
attach(af)
attach(an)
attach(at)
attach(ce)
attach(cf)
attach(ci)
attach(li)
attach(lf)
attach(le)
ae.t=t(ae)
af.t=t(af)
an.t=t(an)
at.t=t(at)
ce.t=t(ce)
cf.t=t(cf)
ci.t=t(ci)
li.t=t(li)
lf.t=t(lf)
le.t=t(le)
attach(dados)
subgrupos.t=list(ae.t,af.t,an.t,at.t,ce.t,cf.t,ci.t,li.t,lf.t,le.t)
AE.vec=c(AE1,AE2,AE3,AE4,AE5,AE6,AE7,AE8,AE9,AE10,AE11,AE12,AE13,AE14,AE15
,AE16,AE17,AE18,AE19,AE20,AE21,AE22,AE23,AE24,AE25,AE26,AE27,AE28,AE29,AE3
0,AE31,AE32,AE33,AE34,AE35,AE36,AE37,AE38,AE39,AE40,AE41,AE42,AE43,AE44,AE
45,AE46,AE47,AE48,AE49,AE50,AE51,AE52,AE53,AE54,AE55,AE56,AE57,AE58,AE59,A
E60,AE61,AE62,AE63,AE64,AE65,AE66,AE67,AE68,AE69,AE70,AE71,AE72,AE73,AE74,
AE75,AE76,AE77,AE78,AE79,AE80,AE81,AE82,AE83,AE84,AE85,AE86,(AE87,AE88,AE8
9,AE90,AE91,AE92,AE93,AE94,AE95,AE96,AE97,AE98,AE99,AE100)
AF.vec=c(AF1,AF2,AF3,AF4,AF5,AF6,AF7,AF8,AF9,AF10,AF11,AF12,AF13,AF14,AF15,
AF16,AF17,AF18,AF19,AF20,AF21,AF22,AF23,AF24,AF25,AF26,AF27,AF28,AF29,AF30,
AF31,AF32,AF33,AF34,AF35,AF36,AF37,AF38,AF39,AF40,AF41,AF42,AF43,AF44,AF45,
AF46,AF47,AF48,AF49,AF50,AF51,AF52,AF53,AF54,AF55,AF56,AF57,AF58,AF59,AF60,
AF61,AF62,AF63,AF64,AF65,AF66,AF67,AF68,AF69,AF70,AF71,AF72,AF73,AF74,AF75,
AF76,AF77,AF78,AF79,AF80,AF81,AF82,AF83,AF84,AF85,AF86,AF87,AF88,AF89,AF90,
AF91,AF92,AF93,AF94,AF95,AF96,AF97,AF98,AF99,AF100)
AT.vec=c(AT1,AT2,AT3,AT4,AT5,AT6,AT7,AT8,AT9,AT10,AT11,AT12,AT13,AT14,AT15,
AT16,AT17,AT18,AT19,AT20,AT21,AT22,AT23,AT24,AT25,AT26,AT27,AT28,AT29,AT30,
AT31,AT32,AT33,AT34,AT35,AT36,AT37,AT38,AT39,AT40,AT41,AT42,AT43,AT44,AT45,

```

AT46,AT47,AT48,AT4,AT50,AT51,AT52,AT53,AT54,AT55,AT56,AT57,AT58,AT59,AT60,A
T61,AT62,AT63,AT64,AT65,AT66,AT67,AT68,AT69,AT70,AT71,AT72,AT73,AT74,AT75,A
T76,AT77,AT78,AT79,AT80,AT81,AT82,AT83,AT84,AT85,AT86,AT87,AT88,AT89,AT90,A
T91,AT92,AT93,AT94,AT95,AT96,AT97,AT98,AT99,AT100)

AN.vec=c(AN1,AN2,AN3,AN4,AN5,AN6,AN7,AN8,AN9,AN10,AN11,AN12,AN13,AN14,AN
15,AN16,AN17,AN18,AN19,AN20,AN21,AN22,AN23,AN24,AN25,AN26,AN27,AN28,AN29,
AN30,AN31,AN32,AN33,AN34,AN35,AN36,AN37,AN38,AN39,AN40,AN41,AN42,AN43,AN
44,AN45,AN46,AN47,AN48,AN49,AN50,AN51,AN52,AN53,AN54,AN55,AN56,AN57,AN58,
AN59,AN60,AN61,AN62,AN63,AN64,AN65,AN66,AN67,AN68,AN69,AN70,AN71,AN72,AN
73,AN74,AN75,AN76,AN77,AN78,AN79,AN80,AN81,AN82,AN83,AN84,AN85,AN86,AN87,
AN88,AN89,AN90,AN91,AN92,AN93,AN94, AN95,AN96,AN97,AN98,AN99,AN100)

CE.vec=c(CE1,CE2,CE3,CE4,CE5,CE6,CE7,CE8,CE9,CE10,CE11,CE12,CE13,CE14,CE15
,CE16,CE17,CE18,CE19,CE20,CE21,CE22,CE23,CE24,CE25,CE26,CE27,CE28,CE29,CE3
0,CE31,CE32,CE33,CE3,CE35,CE36,CE37,CE38,CE39,CE40,CE41,CE42,CE43,CE44,CE4
5,CE46,CE47,CE48,CE49,CE50,CE5,CE52,CE53,CE54,CE55,CE56,CE57,CE58,CE59,CE6
0,CE61,CE62,CE63,CE64,CE65,CE66,CE67,CE6,CE69,CE70,CE71,CE72,CE73,CE74,CE7
5,CE76,CE77,CE78,CE79,CE80,CE81,CE82,CE83,CE84,CE8,CE86,CE87,CE88,CE89,CE9
0,CE91,CE92,CE93,CE94, CE95,CE96,CE97,CE98,CE99,CE100)

CF.vec=c(CF1,CF2,CF3,CF4,CF5,CF6,CF7,CF8,CF9,CF10,CF11,CF12,CF13,CF14,CF15,
CF16,CF17,CF18,CF19,CF20,CF21,CF22,CF23,CF24,CF25,CF26,CF27,CF28,CF29,CF30,
CF31,CF32,CF33,CF34,CF35,CF36,CF37,CF38,CF39,CF40,CF41,CF42,CF43,CF44,CF45,
CF46,CF47,CF48,CF49,CF50,CF51,CF52,CF53,CF54,CF55,CF56,CF57,CF58,CF59,CF60,
CF61,CF62,CF63,CF64,CF65,CF66,CF67,CF68,CF69,CF70,CF71,CF72,CF73,CF74,CF75,
CF76,CF77,CF78,CF79,CF80,CF81,CF82,CF83,CF84,CF85,CF86,CF87,CF88,CF89,CF90,
CF91,CF92,CF93,CF94, CF95,CF96,CF97,CF98,CF99,CF100)

CI.vec=c(CI1,CI2,CI3,CI4,CI5,CI6,CI7,CI8,CI9,CI10,CI11,CI12,CI13,CI14,CI15,CI16,CI17,CI
18,CI19,CI20,CI21,CI22,CI23,CI24,CI25,CI26,CI27,CI28,CI29,CI30,CI31,CI32,CI33,CI34,CI3
5,CI36,CI37,CI38,CI39,CI40,CI41,CI42,CI43,CI44,CI45,CI46,CI47,CI48,CI49,CI50,CI51,CI52
,CI53,CI54,CI55,CI56,CI57,CI58,CI59,CI60,CI61,CI62,CI63,CI64,CI65,CI66,CI67,CI68,CI69,
CI70,CI71,CI72,CI73,CI74,CI7,CI76,CI77,CI78,CI79,CI80,CI81,CI82,CI83,CI84,CI85,CI86,CI
87,CI88,CI89,CI90,CI91,CI92,CI93,CI94, CI95,CI96,CI97,CI98,CI99,CI100)

LI.vec=c(LI1,LI2,LI3,LI4,LI5,LI6,LI7,LI8,LI9,LI10,LI11,LI12,LI13,LI14,LI15,LI16,LI17,LI18,LI
19,LI20,LI21,LI22,LI23,LI24,LI25,LI26,LI27,LI28,LI29,LI30,LI31,LI32,LI33,LI34,LI35,LI36,LI
37,LI38,LI39,LI40,LI41,LI42,LI43,LI44,LI45,LI46,LI47,LI48,LI49,LI50,LI51,LI52,LI53,LI54,LI
55,LI56,LI57,LI58,LI59,LI60,LI61,LI62,LI63,LI64,LI65,LI66,LI67,LI68,LI69,LI70,LI71,LI72,LI
73,LI74,LI75,LI76,LI77,LI78,LI7,LI80,LI81,LI82,LI83,LI84,LI85,LI86,LI87,LI88,LI89,LI90,LI9
1,LI92,LI93,LI94,LI95,LI96,LI97,LI98,LI99,LI100)

LF.vec=c(LF1,LF2,LF3,LF4,LF5,LF6,LF7,LF8,LF9,LF10,LF11,LF12,LF13,LF14,LF15,LF16
,LF17,LF18,LF19,LF20,LF21,LF22,LF23,LF24,LF25,LF26,LF27,LF28,LF29,LF30,LF31,LF
32,LF33,LF34,LF35,LF36,LF37,LF38,LF39,LF40,LF41,LF42,LF43,LF44,LF45,LF46,LF47,L
F48,LF49,LF50,LF51,LF52,LF53,LF5,LF55,LF56,LF57,LF58,LF59,LF60,LF61,LF62,LF63,
LF64,LF65,LF66,LF67,LF68,LF69,LF70,LF71,LF7,LF73,LF74,LF75,LF76,LF77,LF78,LF79
,LF80,LF81,LF82,LF83,LF84,LF85,LF86,LF87,LF88,LF89,LF9,LF91,LF92,LF93,LF94,
LF95,LF96,LF97,LF98,LF99,LF100)

LE.vec=c(LE1,LE2,LE3,LE4,LE5,LE6,LE7,LE8,LE9,LE10,LE11,LE12,LE13,LE14,LE15,LE
16,LE17,LE18,LE19,LE20,LE21,LE22,LE23,LE24,LE25,LE26,LE27,LE28,LE29,LE30,LE31
,LE32,LE33,LE34,LE35,LE36,LE37,LE38,LE39,LE40,LE41,LE42,LE43,LE44,LE45,LE46,L
E47,LE48,LE49,LE50,LE51,LE52,LE53,LE54,LE55,LE56,LE57,LE58,LE59,LE60,LE61,LE

```
62,LE63,LE64,LE65,LE66,LE67,LE68,LE69,LE70,LE71,LE72,LE73,LE74,LE75,LE76,LE77,  
LE78,LE79,LE80,LE81,LE82,LE83,LE84,LE85,LE86,LE87,LE88,LE89,LE90,LE91,LE92,L  
E93,LE94, LE95,LE96,LE97,LE98,LE99,LE100)  
dados=data.frame(AE.vec,AF.vec,AT.vec,AN.vec,CE.vec,CF.vec,CI.vec,LI.vec,LF.vec,L  
E.vec)  
dados.m=mean(dados)  
dados.m  
mqcc(subgrupos.t,type=c("T2"),dados.m,dados.cov,limits=F,labels=1:98,xlab="Amostr  
as",ylab="T2",title="GRÁFICO T2 DE HOTTELING")
```

APÊNDICE E - Rotina para o cálculo das distâncias de mahalanobis

```
dadoscep<-read.table("dadoscep.txt", head=T)
attach(dadoscep)
dadoscep.m=mean(dadoscep)
dadoscep.m
dadoscep.cov=cov(dadoscep)
dadoscep.cov
mahalanobis(dadoscep,dadoscep.m,dadoscep.cov,Inverted=False)
```

APÊNDICE F - Subconjunto básico inicial

CE	CI	LE	LI	AT	AE	CF	LF	AN	AF
409,5	384,525	306,125	281,625	275,8	269,4	378,46	275,175	138,025	6,375
409,325	385,3	305,85	282,075	275,575	269,275	378,3	275,375	137,75	6,325
409,575	384,575	306,175	281,55	275,7	269,275	378,375	275,4	138,2	6,375
409,475	384,625	305,9	281,275	275,45	269,275	378,175	275,175	137,875	6,3
409,425	385,2	305,8	281,575	275,55	269,1	378,425	275,575	137,7	6,175
409,475	384,55	306,35	281,6	275,6	269,275	378,35	275,35	138,2	6,4
409,625	384,65	306,275	281,625	275,525	269,35	378,425	275,275	138,25	6,25
409,625	384,85	306,275	281,625	275,775	269,525	378,475	275,275	138,325	6,4
409,8	385,025	306,3	281,7	275,7	269,525	378,5	275,35	137,975	6,325
409,275	385,1	305,675	281,65	275,2	269,075	378,125	275,5	137,525	6,175
409,175	385,25	306,275	282,05	275,65	269,275	378,55	275,25	138,125	6,275
409,375	385,325	306,2	281,825	275,475	269,175	378,475	275,25	137,55	6,325
409,3	385,3	305,95	281,925	275,3	268,95	378,3	275,25	137,5	6,25
409,825	385,025	306,175	281,675	275,65	269,5	378,475	275,375	138,175	6,375
409,475	385,225	306,175	281,85	275,75	269,4	378,425	275,175	138,325	6,225
409,575	385,45	306,175	281,975	275,675	269,325	378,45	275,275	138,175	6,325
409,525	385,325	305,85	281,9	275,45	269,35	378,55	275,55	137,625	6,225
409,75	385,125	306,425	282,1	275,75	269,475	378,425	275,3	138	6,225
409,55	385,175	306	281,725	275,575	269,175	378,55	275,35	137,525	6,25
409,575	385,2	305,85	281,425	275,55	269,175	378,325	275,45	137,225	6,25
409,625	385,1	305,875	281,825	275,5	269,425	378,325	275,125	137,875	6,275
409,375	385,2	305,675	281,65	275,4	269,25	378,325	275,15	137,475	6,225
409,6	385,1	305,825	281,625	275,625	269,35	378,075	275,175	138,275	6,4
409,65	385,275	305,75	281,525	275,575	269,275	378,15	275,2	137,55	6,125
409,525	385,125	305,475	281,425	275,625	269,175	378,225	275,175	137,6	6,25
409,525	384,9	305,7	281,675	275,75	269,475	378,275	275,175	137,775	6,375
409,4	385,1	306,125	281,825	275,7	269,35	378,425	275,35	137,9	6,225
409,4	385,125	306,125	281,85	275,75	269,425	378,45	275,35	137,95	6,275
409,5	385,425	306,075	281,725	275,625	269,425	378,375	275,25	138,2	6,125
409,625	385,25	306,15	282,05	275,525	269,375	378,2	275,5	138	6,175

APÊNDICE G - Resultado das iterações

Iteração	Subconjunto básico inicial	Pontos Discrepantes	Ponto a ser adicionado no subgrupo	Ponto de corte	Presença de outlier?
1	31	69	3,2327	42,31	não
2	32	68	2,0746	42,08	não
3	33	67	3,5679	41,71	não
4	34	66	3,7717	41,39	não
5	35	65	3,9996	40,72	não
6	36	64	4,0712	40,33	não
7	37	63	3,8507	39,94	não
8	38	62	4,0292	39,56	não
9	39	61	4,2214	39,19	não
10	40	60	4,1823	38,82	não
11	41	59	4,2692	38,47	não
12	42	58	4,1941	38,12	não
13	43	57	4,1632	37,78	não
14	44	56	4,2776	37,44	não
15	45	55	4,1827	37,11	não
16	46	54	4,3108	36,79	não
17	47	53	4,3671	36,48	não
18	48	52	4,3055	36,17	não
19	49	51	4,2132	35,86	não
20	50	50	4,3977	35,56	não
21	51	49	4,4565	35,27	não
22	52	48	4,2959	34,98	não
23	53	47	4,3798	34,70	não
24	54	46	4,5366	34,43	não

25	55	45	4,6452	34,16	não
26	56	44	4,6147	33,89	não
27	57	43	4,7230	33,63	não
28	58	42	4,6216	33,37	não
29	59	41	4,4750	33,12	não
30	60	40	4,3894	32,87	não
31	61	39	4,3724	32,62	não
32	62	38	4,6509	32,38	não
33	63	37	4,6517	32,15	não
34	64	36	4,4690	31,92	não
35	65	35	4,6444	31,69	não
36	66	34	4,5862	31,46	não
37	67	33	4,7010	31,24	não
38	68	32	4,6557	31,03	não
39	69	31	4,7773	30,81	não
40	70	30	4,6438	30,60	não
41	71	29	4,6891	30,40	não
42	72	28	4,5256	30,19	não
43	73	27	4,8394	29,99	não
44	74	26	4,9090	29,79	não
45	75	25	4,9867	29,60	não
46	76	24	4,9010	29,41	não
47	77	23	4,9853	29,22	não
48	78	22	5,0306	29,03	não
49	79	21	5,0306	28,85	não
50	80	20	4,9706	28,67	não
51	81	19	5,0431	28,67	não
52	82	18	5,0704	28,49	não

53	83	17	5,0671	28,32	não
54	84	16	5,0068	28,15	não
55	85	15	5,1344	27,98	não
56	86	14	5,2104	27,81	não
57	87	13	5,1899	27,64	não
58	88	12	5,1004	27,48	não
59	89	11	5,1313	27,32	não
60	90	10	4,9583	27,16	não
61	91	09	5,2345	27,01	não
62	92	08	5,6123	26,85	não
63	93	07	5,7304	26,70	não
64	94	06	5,7712	26,55	não
65	95	05	5,8156	26,40	não
66	96	04	5,8177	26,11	não
67	97	03	6,7675	26,40	não
68	98	02	8,4478	25,97	não
69	99	01	7,4871	25,83	não
70	100	0	8,3111	25,69	não