MENSURAÇÃO DA QUALIDADE DE PRODUTOS E SERVIÇOS DENTRO DE AGRUPAMENTOS DOS PERFIS DE CLIENTES PELA METODOLOGIA SEIS SIGMA, UTILIZANDO SIMULAÇÃO DE MONTE CARLO NA GERAÇÃO DE DADOS FALTANTES

Samir Jorge Guedes Sias Thompson

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Produção, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Produção.

Orientador: Lino Guimarães Marujo

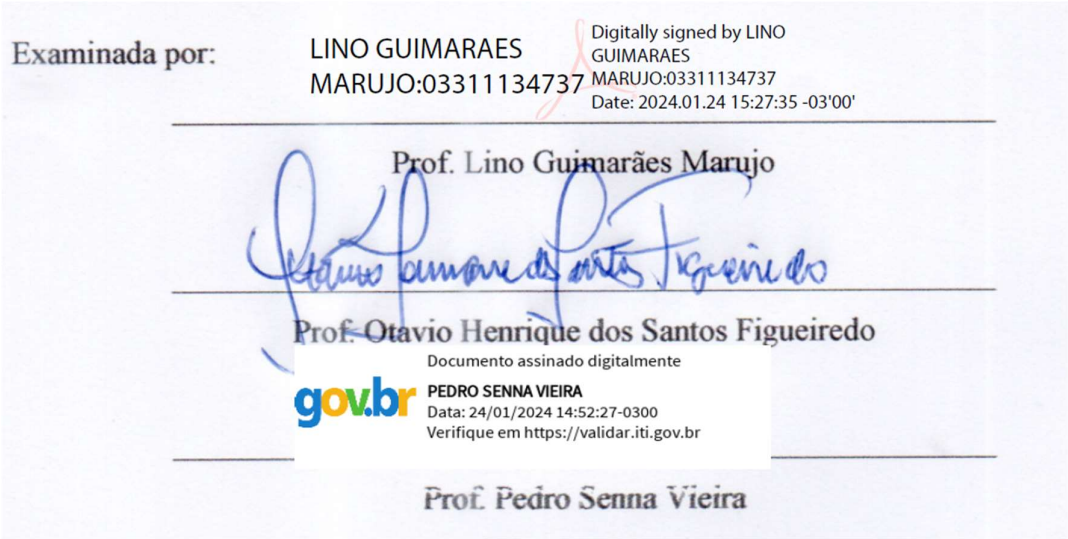
Rio de Janeiro

Dezembro de 2023

MENSURAÇÃO DA QUALIDADE DE PRODUTOS E SERVIÇOS DENTRO DE AGRUPAMENTOS DOS PERFIS DE CLIENTES PELA METODOLOGIA SEIS SIGMA, UTILIZANDO SIMULAÇÃO DE MONTE CARLO NA GERAÇÃO DE DADOS FALTANTES

Samir Jorge Guedes Sias Thompson

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA DE PRODUÇÃO.



RIO DE JANEIRO, RJ - BRASIL

DEZEMBRO DE 2023

Thompson, Samir Jorge Guedes Sias

Mensuração da qualidade de produtos e serviços dentro de agrupamentos dos perfis de clientes pela metodologia seis sigma, utilizando simulação de monte carlo na geração de dados faltantes / Samir Jorge Guedes Sias Thompson. – Rio de Janeiro: UFRJ/COPPE, 2023.

XXII, 132 p.: il.; 29,7 cm.

Orientador: Lino Guimarães Marujo

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Produção, 2023.

Referências Bibliográficas: p. 71-75.

1. Qualidade Produtos. 2. Simulação de Monte Carlo. 3. Seis Sigma. I. Marujo, Lino Guimarães *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Produção. III. Mensuração da qualidade de produtos e serviços dentro de agrupamentos dos perfis de clientes pela metodologia seis sigma, utilizando simulação de monte carlo na geração de dados faltantes.

Dedicatória

Fui tão idiota demorando tanto tempo para defender esta dissertação que infelizmente meu pai veio a falecer em decorrência de um câncer e não me verá com o título de mestre. Sendo assim, dedico esta dissertação ao meu pai, Jorge Luiz Thompson.

Com toda a tecnologia e avanços na medicina, não consigo entender como uma doença desgraçada, como o câncer, não pode ser totalmente curada.

Só desejo que agora ele esteja numa bem melhor que nós aqui na Terra.

Minha dedicatória também vai para minhas mãe e irmã, que seguiram firmes no tratamento com meu pai. Infelizmente perdemos esta batalha.

Deixo aqui o pedaço de uma música que sempre ouvi e sabia o motivo da letra. O guitarrista da banda também perdeu o pai e queria falar com ele após a morte. Escreveu uma carta em forma de música para o pai. "Cedo ou tarde a gente vai se encontrar / Tenho certeza, numa bem melhor."

Agradecimentos

Meus sinceros agradecimentos a todos aqueles que de alguma forma colaboraram para que a conclusão deste trabalho se tornasse possível e colaboraram de alguma forma em minha vida:

À Deus, sem Ele nada seria possível, pois Ele é o caminho, a verdade e a vida.

À minha mãe, Luzia, que sempre fez de tudo por mim, sabendo brigar e fazer elogios na medida certa e na hora certa, mesmo tendo alguns desentendimentos em certos momentos, sei que tudo era para o meu bem e é um exemplo a ser seguido.

Ao meu falecido pai, Jorge Luiz, mais conhecido como Comandante Thompson, que mesmo não podendo estar presente em todos os momentos, pois trabalhava viajando, sempre soube estender a mão, apoiar minhas decisões e, em raras vezes, dar broncas me fazendo tornar uma pessoa melhor.

À minha irmã, Sarah, apesar dos atritos e diferenças, sempre me ajudou e me auxiliou em tudo, desde transporte até a fazer comida. Sem ela, nada seria possível.

À minha amada Ana Luiza, por me encorajar na realização do mestrado. Agradeço toda a paciência que teve comigo nesta jornada, que não foi fácil. Nos momentos difíceis não me deixou entregar os pontos. Amor imenso.

Ao Luís Eduardo, mais conhecido como Duda, que me apresentou e sempre incentivou a realizar este mestrado.

À Academia Kioto *Brazilian Jiu-jitsu* e todos meus companheiros de treino, que há anos me ajudam a controlar meu temperamento e dar mais disciplina, "*Oss*".

Aos companheiros de turma e caminhada neste mestrado. Em especial à Danielle, minha leal colega que em muitas vezes esteve comigo nos estudos e sofrendo com trabalhos para serem entregues. Trabalhos estes que viramos noites ou passamos final de semana na UFRJ para realizar da melhor maneira. Além disso, nossas conversas sobre o mestrado e indignações me ajudaram muito.

Ao professor Lino, meu professor e orientador que acreditou em minha proposta de dissertação, orientou e mostrou toda paciência em momentos que não respondi da melhor forma, porém não me abandonou. Um agradecimento especial, pois, além de professor e orientador foi um amigo.

Agradeço ao meu sentimento de raiva. Acredito que ela seja um dom que me impulsiona sempre quando escuto que não sou capaz de algo ou quando alguma pessoa tenta atrapalhar minha caminhada.

E por último, mas não menos importante, a professora Ligia Gomes Elliot da Fundação CESGRANRIO. Em 2018, fiz o processo para entrar no mestrado desta instituição e ela não me passou com o argumento que ia perder as duas primeiras aulas por questão de viagem de lua de mel. Foi bom pois fui atrás do mestrado que eu realmente queria na UFRJ e passei. Quem perdeu foi a instituição de ter em seu quadro de egressos um excelente aluno e profissional que sou.

Samir Sias Thompson

"*Parte da jornada é o fim.*"

Tony Stark

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MENSURAÇÃO DA QUALIDADE DE PRODUTOS E SERVIÇOS DENTRO DE AGRUPAMENTOS DOS PERFIS DE CLIENTES PELA METODOLOGIA SEIS SIGMA, UTILIZANDO SIMULAÇÃO DE MONTE CARLO NA GERAÇÃO DE DADOS FALTANTES

Samir Jorge Guedes Sias Thompson

Dezembro/2023

Orientador: Lino Guimarães Marujo

Programa: Engenharia de Produção

Este trabalho difunde um fluxograma que serve como guia sobre o estudo da qualidade para produtos e serviços dentro de qualquer empresa. Para a medição desta qualidade a metodologia Seis Sigma é utilizada. Determinadas vezes as empresas não apresentam todos os dados necessários para medir a qualidade, com isto a Simulação de Monte Carlo auxilia para a obtenção destes dados. Ponderando que a existência de diferentes perfis de clientes, a qualidade é mensurada diferenciadamente para cada agrupamento *(cluster*) de perfil. Através de resultados de base real de uma empresa de seguro de vida, é possível afirmar que o fluxograma nortea no estudo dos agrupamentos *(clusters)* e demonstra a possiblidade de estudo da satisfação dos clientes. A satisfação em cada agrupamento (*cluster*) obteve alto índice de qualidade.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MEASURING THE QUALITY OF PRODUCTS AND SERVICES WITHIN THE CLUSTERS OF CUSTOMER PROFILES THROUGH THE SIX SIGMA METHODOLOGY, USING MONTE CARLO'S SIMULATION IN THE GENERATION OF MISSING DATA

Samir Jorge Guedes Sias Thompson

December/2023

Advisor: Lino Guimarães Marujo

Department: Production Engineering

This work disseminates a flowchart that serves as a guide on the study of quality for products and services within any company. To measure this quality, the Six Sigma methodology is used. Sometimes companies do not present all the data necessary to measure quality, so Monte Carlo Simulation helps to obtain this data. Considering the existence of different customer profiles, quality is measured differently for each profile grouping. Through real-base results from a life insurance company, it is possible to state that the flowchart guides the study of groupings and demonstrates the possibility of studying customer satisfaction. Satisfaction in each grouping achieved a high quality index.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Nomenclaturas

CEP – Controle Estatístico do Processo

CQT – *Control Quality Total*

ENCE – Escola Nacional de Ciências Estatísticas

ENIAC – *Electronic Numerical Integrator and Computer*

FDP – Função de Densidade de Probabilidade

GQT – Gerenciamento da Qualidade Total

IID – Independente e Identicamente Distribuída

IOT – *Internet of Things*

ISO – *International Organization for Standardization*

LIC – Linha Inferior de Controle

LIE – Limite Inferior de Especificação

LSC – Linha Superior de Controle

LSE – Limite Superior de Especificação

NPS – *Net Promoter Score*

PNQ – Prêmio Nacional da Qualidade

PPM – Partes por Milhão

SMC – Simulação de Monte Carlo

UFMG – Universidade Federal de Minas Gerais

USP – Universidade de São Paulo

1 Introdução

A qualidade de produtos e serviços tem continuamente o interesse de melhora, centrando-se em preferências e necessidades específicas de cada indivíduo, de acordo com Ramos *et. al* (2021). A importância que Ramos *et. al* (2021) traz no foco da qualidade na área da saúde, exibe que a qualidade no âmbito da saúde evolui ao longo das décadas, agregando novas vertentes e valores inseridos no processo de trabalho.

Para compreender o conceito de qualidade Carvalho e Paladini (2012) informam que é necessário traçar sua evolução ao longo do último século. Antes da Revolução Industrial, para o artesão, o atendimento às necessidades do cliente era a abordagem de qualidade. Além disso, o foco do controle da qualidade era o produto, não o processo. Na Revolução Industrial, a customização foi substituída pela padronização e a produção em larga escala. Nessa época surgiu a função do inspetor, responsável pela qualidade do produto, no entanto, a necessidade do cliente deixou de ser priorizada.

Em 1924, o conceito de controle da qualidade alcançou outro patamar com a criação dos gráficos de controle, difundindo conceitos de estatística à realidade produtiva das empresas, bem como surgindo atividades de análise e solução de problema. Na década de 30, o controle da qualidade continuou evoluindo, com o desenvolvimento do sistema de medidas e do surgimento de técnicas de amostragem e de normas específicas para a área. Na década de 50, foi formulado o primeiro sistema de controle da qualidade total, tratando o conceito de forma sistêmica nas organizações, que influenciaria fortemente o modelo proposto pela Organização Internacional para Padronização (*International Organization for Standardization* – ISO), a série ISO 9000. Foi criado também o primeiro prêmio atribuído à empresa que mais se destacasse na área da qualidade em cada ano. Só na década de 90, surgiu um prêmio similar no Brasil, o Prêmio Nacional da Qualidade – PNQ.

Em 2000, foi feita a terceira revisão da série, ISO 9000:2000 que trouxe novos elementos, passando a adotar uma visão de gestão da qualidade e não mais de garantia, introduzindo elementos de foco no cliente e da gestão por processos e diretrizes. Assim, recupera-se alguns atributos da época artesanal, como a busca da proximidade às demandas do cliente e maior customização, embora agora em massa. Esse resgate da importância dos clientes e a percepção da qualidade como um critério competitivo, trouxe

alguns teóricos da área de estratégia e administração para a área da qualidade, como David Garvin, que em seus trabalhos discutiu o impacto estratégico da qualidade.

Garvin (1984) diz que a qualidade é um fator de decisão de compra pelos clientes, conferindo um importante diferencial competitivo. Propôs oito dimensões da qualidade do produto: desempenho, características, confiabilidade, conformidade, durabilidade, facilidade de manutenção, estética e qualidade percebida. E Batalha (2009) acrescenta informações para estas dimensões e altera algumas da seguinte forma:

1. Desempenho: características operacionais básicas de um produto;
2. Características: itens "extra" adicionados às características operacionais básicas, como pen drive, bancos de couro etc.;
3. Confiabilidade: probabilidade de que um produto opere adequadamente durante um certo período; como por exemplo uma TV deve funcionar sem defeitos por sete anos;
4. Conformidade: grau de aderência a padrões preestabelecidos;
5. Durabilidade: a vida útil do produto em termos de quanto dura antes de ser substituído;
6. Atendimento: aspectos relativos ao serviço associado ao produto, como rapidez, cortesia e facilidade de reparos;
7. Estética: aparência do produto / projeto (*design*);
8. Qualidade percebida: inferências feitas pelos consumidores com base em sua percepção, que é afetada pela marca e reputação;

E amplia as dimensões da qualidade também para os serviços prestados, a saber:

1. Tangíveis: aparência das facilidades físicas, equipamentos, pessoal e comunicação material;
2. Atendimento: nível de atenção dos funcionários com os clientes;
3. Confiabilidade: habilidade de realizar o serviço prometido de forma confiável e acurada;
4. Resposta: atender o cliente e fornecer serviços rápidos;
5. Competência: possuir habilidade para efetuar o serviço;
6. Consistência: ausência de variabilidades no serviço prestado;
7. Cortesia: respeito e afetividade no contato pessoal;
8. Credibilidade: honestidade, tradição, confiança no serviço;

9. Segurança: inexistência de perigo, risco, dúvida;
10. Acesso: proximidade e contato fácil;
11. Comunicação: manter o cliente informado;
12. Conveniência: proximidade e disponibilidade;
13. Velocidade: rapidez para iniciar e executar o serviço;
14. Flexibilidade: capacidade de alterar o serviço prestado;
15. Entender o cliente: conhecer o cliente e suas necessidades.

Segundo Carvalho e Paladini (2012), um programa valioso para a gestão da qualidade surgiu no final da década de 80, na empresa Motorola, chamado Seis Sigma. Contudo, essa ferramenta se popularizou no final do século passado e início do século XXI. Esse programa apresenta várias características dos modelos anteriores, como o embasamento estatístico com ênfase no controle da qualidade e na análise e solução de problemas. No entanto, Seis Sigma promove alinhamento estratégico da qualidade, em especial, com a relação custo-benefício.

Alamsyah e Nurriz (2017) entendiam a evolução da gestão da qualidade e percebido que a qualidade é contemplada sob várias óticas, faz-se necessário dividir os clientes em grupos, uma vez que não é correto generalizar opiniões sobre o conceito de qualidade. É sabido que a clientela possui diferentes perfis, bem como distintas necessidades e expectativas quanto à qualidade do produto. Essa preocupação em agrupar colabora para resgatar a proximidade com o cliente, focando as atividades do projeto, como fundamentais para a satisfação do cliente e para criação de uma qualidade robusta. A satisfação é a ação ou o efeito de satisfazer ou satisfazer-se. É o contentamento, ou seja, o prazer resultante da realização daquilo que se espera ou do se deseja. É cumprir ou corresponder à expectativa e apresentado este pensamento na figura 1, segundo Slack *et al.* (2006).

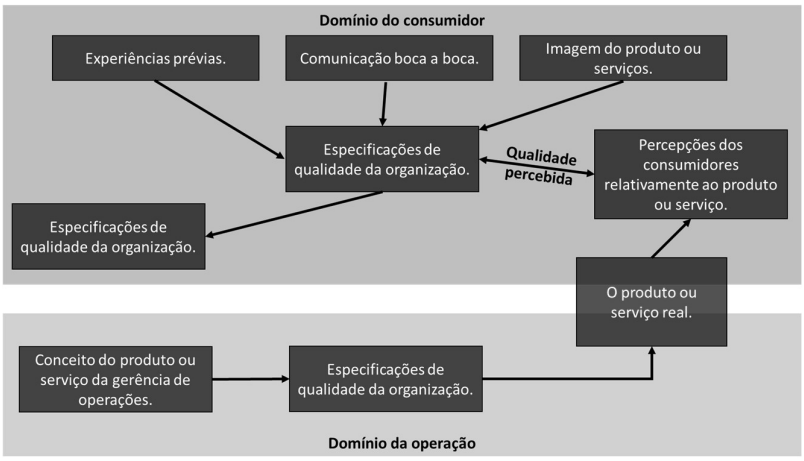


Figura 1 – Perspectivas do consumidor e da operação.
Fonte: Baseado em Slack *et al.* (2006).

Ferreira, Seifert e Venanzi (2020) expõem que as organizações se tornam cada vez mais competitivas com o auxílio da evolução contínua do mercado econômico atual, e assim buscam compreender melhor as exigências de cada cliente em relação a qualidade do produto ou serviço. Atribuem as expectativas do consumidor à medida que a variedade de produtos e serviços são incluídos no mercado. A Internet das Coisas (*Internet of Things* – IOT) é uma tecnologia que auxilia no atingimento de processos e com isso obter mais eficiência nas tarefas focadas na qualidade.

Lui e Petarnella (2020) dialogam sobre qualidade e o desafio na garantia da entrega de serviços de qualidade em cidades inteligentes. Comentam a falta de discussões com maior notoriedade sobre o tema, apontando assim a importância de trabalhos modernos com a temática da qualidade. E, assim como exibido na próxima sessão, este trabalho concilia a importância da qualidade com métodos antes não empregados unidos.

1.1 Objetivos

Como visto em Lui e Petarnella (2020) e Ferreira, Seifert e Venanzi (2020) é necessário a junção de técnicas consolidadas com novas tecnologias na supervisão da qualidade. Sendo assim, esta dissertação tem o objetivo de apresentar um fluxograma na identificação do significado de qualidade pela clientela para qualquer produto ou serviço oferecido, segmentando pelos perfis de clientes. A elaboração deste fluxograma engloba técnicas, salvo melhor juízo aquilo que investiguei, mas não encontrei ainda não

implementadas unidas com foco na percepção da qualidade, assim apontando uma inovação no estudo da qualidade contribuindo para o estado da arte.

A elaboração do projeto proposto consiste nas etapas:

1. Identificar dados não existentes, estimar e verificar a veracidade dos resultados obtidos. Utilizando Simulação de Monte Carlo;
2. Caracterizar e agrupar o perfil dos clientes por meio de informações existentes e simulação dos dados faltantes. Utilizando análise de agrupamentos (*clusters*) por métodos não-hierárquicos;
3. Identificar o foco do cliente, em todos os agrupamentos elencados, mediante o entendimento da qualidade e satisfação com o produto adquirido e/ou o serviço oferecido. Utilizando o método Seis Sigma.

2 Referencial Teórico

A necessidade de serviços públicos municipais e da tecnologia da informação é cada vez mais frequente nas cidades, conforme explanado por Aparecida e Rezende (2021). Estes autores difundiram um modelo de prestação de serviços públicos municipais diretamente plugado por meio de IOT na abrangência da cidade digital estratégica. Modelo este que se conecta diretamente as novas tecnologias que surgiram através dos anos e graças ao avanço tecnológico e necessidades do ser humano, neste caso por meio de IOT. Estas novas tecnologias que surgiram através dos anos são empregadas neste trabalho, mesclando técnicas criadas desde a época da Segunda Grande Guerra Mundial, mesmo assim contemporâneas, com técnicas mais modernas na percepção da qualidade em produtos e serviços.

Esta mescla de técnicas é exposta neste capítulo através de conceitos essenciais e metodologias escolhidas na construção de todas as etapas do fluxograma da percepção da qualidade, mencionado no capítulo anterior. São apresentados aqui os fundamentos dos três pilares deste estudo: a Simulação de Monte Carlo, Análise de Agrupamento (*Cluster*) e identificação da qualidade através do Seis Sigma, por meio dos seus principais conceitos e das discussões trazidas pelos autores referenciados.

2.1 Método de Monte Carlo

Possani (2012) afirma que de acordo com Hammersley e Handscomb (1964) o nome "Monte Carlo" foi criado por John von Neumann, Stanislaw Ulam e Nicholas Metropolis enquanto trabalhavam no projeto Manhattan (projeto voltado a construção da bomba atômica) durante a Segunda Guerra Mundial nos laboratórios da cidade de Los Alamos, Novo México nos Estados Unidos da América. De acordo com Pavani (2019), o nome foi inspirado em um tio de Ulam, que costumava jogar no famoso cassino de Monte Carlo, no Principado de Mônaco. O nome foi formalizado em 1949, por meio do artigo intitulado *"Monte Carlo Method"*, publicado por Metropolis e Ulam (1949).

Segundo Pavani (2019), Stanislaw Ulam jogava paciência, tradicional jogo de cartas, e tentou calcular as probabilidades de ganhar no jogo utilizando análise combinatória. Os cálculos consumiam muito tempo quando finalmente percebeu que uma alternativa mais prática seria efetuar inúmeras jogadas, por exemplo, cem ou mil, e contar quantas vezes cada resultado ocorria. Esta ideia foi adotada por John von Neumann na

implementação de aplicações diretas à difusão de nêutrons em material sujeito a fissão nuclear, como afirmam Carrol e Liang e Lui (2010).

Pavani (2019) cita que Ulam sabia que técnicas de amostragem estatística não eram muito usadas por envolverem cálculos extremamente demorados, tediosos e sujeitos a erros. Entretanto, nessa época, ficou pronto o primeiro computador eletrônico, desenvolvido durante a segunda guerra mundial, o *Electronic Numerical Integrator and Computer* (ENIAC); antes dele eram usados dispositivos mecânicos para fazer cálculos. A versatilidade e rapidez do ENIAC, sem precedentes para a época, impressionaram Ulam, que sugeriu o uso de métodos de amostragem estatística para solucionar o problema adotado por John von Neumann.

Como dito por Gomes (2018), a simulação direta de problemas probabilísticos associado com a difusão das partículas de nêutrons quando submetidos a um processo de fissão nuclear era o que o trabalho decorria. A fissão de um átomo de plutônio enriquecido libera gigantesca quantidade de energia, com isso outro átomo também se divida. É possível a bomba ter duas reações, apresentada na figura 2:

- Reação subcrítica – bomba não explode: não ocorre reação em cadeia. Uma espécie de cadeia de dominós interrompida em alguma parte do caminho.
- Reação supercrítica – bomba explode: quantidade exponencial de energia sendo liberada.

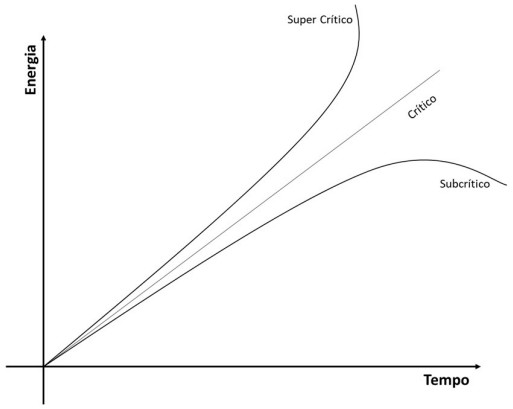


Figura 2 – Gráfico Energia x Tempo explosão bomba atômica.
Fonte: Baseado em Gomes (2018).

Gomes (2018) cita que os cientistas tinham as missões de assegurar que a bomba explodisse e que isso não ocorresse nas mãos deles. Deveriam dividir a quantidade de

plutônio em pedaços pequenos o suficiente de uma forma que não explodisse quando eles não quisessem, mesmo se houvesse um acidente. Deveriam unir os pedacinhos em uma única peça grande, com material suficiente causando uma reação em cadeia no momento da explosão.

Assim sendo, foi concebida a Simulação de Monte Carlo (SMC), aplicando uma variável aleatória, e cada caso, de acordo com a variável, foi calculado um a um por pessoas, e os resultados agrupados por um matemático, modelando matematicamente o comportamento da bomba.

Atualmente, o Método de Monte Carlo pode ser descrito como método de simulação estatística que utiliza sequências de números aleatórios para desenvolver simulações. Em outras palavras, é visto como método numérico universal para resolver problemas por meio de amostragem aleatória, conforme Pavani (2019) cita. A SMC é utilizada em diversos segmentos, como pode ser observado no estudo de Lopes *et. al* (2021) em que para modelar o gerenciamento de estoques em uma farmácia de pequeno porte, utilizam SMC.

Pavani (2019) aponta que Monte Carlo é utilizado rotineiramente em muitos campos de conhecimentos que vão desde simulação de complexos fenômenos físicos a econômicos. Alguns exemplos de aplicação deste método, em diferentes áreas, são:

- Atuária: tábua de expectativa de vida, casamento de passivos/ativos etc.;
- Finanças: séries macroeconômicas, opções futuras, hedge etc.;
- Computação gráfica: redução de artefatos, espalhamento etc.;
- Geologia: caracterização de reservatórios;
- Análise de Projetos: opções reais;
- Jogos: geração de redes (grafos).

O método torna desnecessário escrever as equações diferenciais que descrevem o comportamento de sistemas complexos. A única exigência é que o sistema físico ou matemático seja descrito (modelado) em termos de funções de densidade de distribuição de probabilidade (FDP). Uma vez conhecidas essas distribuições, a Simulação de Monte Carlo pode proceder fazendo as amostragens aleatórias a partir das mesmas. Este processo é repetido inúmeras vezes e o resultado desejado é obtido por meio de técnicas estatísticas (média, desvio padrão etc.) sobre um determinado número de realizações (amostra) que podem chegar a milhões, assim como Pavani (2019) redige.

A figura 3 a seguir ilustra a ideia genérica do método, assumindo que o comportamento do sistema possa ser descrito por apenas uma FDP, segundo Pavani (2019). Utilizamos, também, um rápido e efetivo meio de gerar números aleatórios uniformemente distribuídos dentro do intervalo 0 e 1. Ao término, os resultados desta amostragem aleatória são acumulados e manipulados para produzir o resultado desejado.

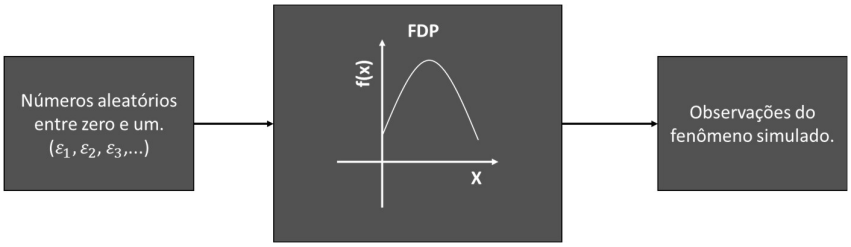


Figura 3 – Ideia genérica do Método de Monte Carlo.
Fonte: Baseado em Pavani (2019).

Na prática, Pavani (2019) afirma que diante de um problema envolvendo incertezas, realizar uma Simulação com Monte Carlo para aproximar sua solução consiste em cinco passos padrões:

i. Modelar o problema definindo uma FDP para representar o comportamento de cada uma das suas incertezas.
ii. Gerar valores pseudo-aleatórios aderentes à FDP de cada incerteza do problema.
iii. Calcular o resultado determinístico substituindo as incertezas pelos valores gerados obtendo, assim, uma observação do problema.
iv. Repetir os passos ii e iii até se obter uma amostra com o tamanho desejado de realizações.
v. Agregar e manipular os resultados da amostra de forma a obter uma estimativa da solução do problema.

Note que este método apenas proporciona uma aproximação da solução, portanto, é fundamental analisar o erro de aproximação, que é $\frac{3\sigma}{\sqrt{N}}$, onde σ é o desvio padrão da amostra e N o tamanho da amostra. Logo, é evidente que quanto maior o tamanho da amostra, menor o erro de aproximação, assim como Pavani (2019) descreve.

Pavani (2019) comenta que a Simulação de Monte Carlo pode ser utilizada para determinar, por exemplo, a probabilidade de ocorrência da soma 7 ao se jogar dois dados comuns de seis faces. Acompanhe o passo a passo a seguir.

i. As incertezas envolvidas são o resultado que cada dado apresentará ao ser jogado. Os valores que podem ser assumidos por cada incerteza são os números inteiros de 1 a 6, com igual probabilidade de ocorrência de cada número, pois são dados idôneos. Portanto, a FDP nos leva a um destes valores.
ii. A geração de dois valores pseudo-aleatórios aderentes à FDP representará o resultado obtido ao se jogar os dois dados, ou seja, dois valores inteiros entre 1 e 6.
iii. O resultado determinístico é dado pela soma do resultado dos dois dados que foram obtidos no passo anterior, ou seja, valores entre 2 e 12. Se este valor é 7 temos um resultado positivo na observação, se não, um negativo.
iv. Repetindo o passo ii e iii conseguiremos obter uma amostra suficientemente grande, com n realizações.
v. Com base na amostra podemos contar quantas vezes tivemos um resultado positivo e dividir pelo tamanho da amostra para obter a probabilidade desta ocorrência.

Pavani (2019) cita que obviamente, esse problema é suficientemente simples para ser resolvido por análise combinatória, mas serve para exemplificar a simulação e tornar evidente que o esforço computacional envolvido está diretamente relacionado ao tamanho da amostra (quantidade de repetição dos passos B e C) que, por sua vez, conforme demonstrado, está diretamente relacionado ao erro de aproximação. Portanto, quanto menor o erro de aproximação desejado, maior o esforço computacional envolvido.

Como exemplo prático a SMC, quando aplicada a problemas reais, requer robusta infraestrutura computacional para alcançar um erro de aproximação satisfatório, o que muitas vezes impede sua utilização, assim como Pavani (2019) explana.

Pavani (2019) aponta que como ocorrido com o lançamento do ENIAC, ambientes de computação na nuvem abrem novas perspectivas de aplicação da SMC em situações que anteriormente apresentariam tempo de resposta inadequado ou demandariam um investimento proibitivo para alcançar resultados em níveis confiáveis.

Observe que cada repetição dos passos ii e iii, apresentados na seção anterior, podem ser executados de forma independente e assíncrona e este processo é o centro do

esforço computacional envolvido. Portanto, é perfeitamente possível paralelizar conjuntos de repetição desses passos para diminuir o tempo total de uma simulação, de acordo com o comentado por Pavani (2019).

Como, a priori, queremos obter os resultados no menor tempo possível, alocaremos todo o esforço computacional disponível e necessário para reduzir o tempo da simulação ao máximo. Por outro lado, imediatamente após o encerramento da simulação não precisaremos mais destes recursos e certamente não queremos pagar por eles. Dessa forma, fica evidente que podemos utilizar a elasticidade (provisionamento e desprovisionamento de recursos computacionais) proporcionada pelo ambiente de computação na nuvem e o seu modelo de cobrança de pagamento pelo uso, para buscar uma simulação ótima, ou seja, que idealmente execute na melhor relação de tempo e custo para os requisitos do negócio, segundo Pavani (2019).

2.1.1 Algoritmo de Metropolis-Hastings (M-H)

Um dos algoritmos mais comuns para a formação das Cadeias de Markov nos métodos MCMC é o de Metropolis-Hastings. Neste algoritmo, a criação de um novo estado, a partir do anterior, é dividida em duas partes: uma proposta, que sugere o próximo candidato aleatório na trajetória da cadeia, e uma aceitação do candidato, que garante que a direção apropriada seja mantida. O primeiro passo do algoritmo de Metropolis-Hastings é a seleção de uma distribuição de proposta de densidade $\pi(P^*|P(n))$, com a qual serão geradas as amostras candidatas P* para o novo estado, dado o atual estado P(n), conforme Herzog (2021) apresenta.

2.1.1.1 Amostrador de Metropolis-Hastings

Santos (2021) comenta que amplamente empregado nas mais diversas áreas, o Metropolis-Hastings é um algoritmo mais geral que permite simular valores de uma Cadeia de Markov que tenha distribuição estacionária f(·). A ideia por trás deles é gerar uma Cadeia de Markov $\{\theta_t; t = 0,1,2, \dots, N\}$ tal que a distribuição estacionária seja a distribuição alvo. Em resumo, o algoritmo deve especificar, para um determinado estado da cadeia, θ_t por exemplo, como gerar o próximo estado, θ_{t+1}. Cabe ressaltar que em todos os algoritmos desta classe, existe um valor candidato Y, gerado a partir de uma

distribuição de proposta $g(\cdot\,|\theta_t)$. Se o valor candidato for aceito, move-se a cadeia para este estado. Caso contrário, a cadeia permanecerá em seu estado atual.

Algoritmo do Amostrador de Metropolis-Hastings:

Entrada:
Uma distribuição alvo f(·);
Uma distribuição geradora de candidatos $g(\cdot\,|\theta)$;
Um inteiro N representando o número de amostras desejadas;
Um valor inicial $\theta^{(0)}$;
Simulação:
para t = 1, ... , N se faz:
1. Gerar $\theta^{*}\sim g\left(\cdot\,\middle|\theta^{(t-1)}\right)$;
2. Gerar U ~ U(0,1);

enquanto $U > \frac{f(\theta^{*})g\left(\theta^{(t-1)}|\theta^{*}\right)}{f\left(\theta^{(t-1)}\right)g\left(\theta^{*}|\theta^{(t-1)}\right)}$ fazer:
repetir (1) e (2);
fim:
Fazer $\theta^{t} = \theta^{*}$;
fim.
Saída:
A cadeia $\left\{\theta^{(t)}|t = 1, \dots, N\right\}$ contendo os valores gerados.

É importante notar que a distribuição geradora de candidatos pode, ou não, depender do estado corrente da cadeia, cabendo ressaltar que, caso a distribuição da proposta atenda às condições de regularidade (irredutibilidade e a periodicidade), a cadeia convergirá para uma distribuição estacionária única π. O algoritmo é projetado para que a distribuição estacionária seja de fato a distribuição alvo, f(·), conforme explanado por Santos (2021).

2.1.2 *Bootstrap*

Chen (2017) afirma que *Bootstrap* é uma abordagem da Simulação de Monte Carlo baseada nos dados disponíveis para estimar a incerteza de uma estatística ou de um estimador. Um recurso poderoso do *bootstrap* é não precisar saber a verdadeira distribuição.

Pegue a mediana como exemplo. Como estimar sua distribuição / incerteza (digamos, variância)? Pode se gerar o mesmo tamanho de dados da mesma distribuição muitas vezes e, em seguida, usamos o histograma / variância dessas novas realizações, assim como Chen (2017) afirma.

Chen (2017) também explana que o *bootstrap* usa uma ideia semelhante, mas agora trata-se os dados originais como a população e a amostra com substituição a partir deles. Um elemento chave aqui é a amostra com substituição. Isso é para imitar o processo de geração de uma amostra IID (Independente e Identicamente Distribuída) - lembre-se de que, quando amostramos com substituição, todos os pontos são IID.

Ye (2020) divulga que *Bootstrap* é um método de reamostragem em que um grande número de amostras do mesmo tamanho é retirado repetidamente, com substituição, de uma única amostra original.

Segundo Ye (2020), *bootstrap* se divide nas seguintes etapas:

1. Decidir quantas amostras de *bootstrap* executar
2. Qual é o tamanho da amostra?
3. Para cada amostra de *bootstrap*:
    - desenhar uma amostra com reposição com o tamanho escolhido
    - calcular a estatística de interesse para essa amostra
4. calcular a média das estatísticas de amostra calculadas

É importante ressaltar que a utilização de *bootstrap* ocorre na carência de determinada informação ou a dúvida na distribuição dos dados, em ambos casos só é possível esta utilização ao se ter certeza a que amostra de estudo é representativa na população.

2.1.2.1 Árvores de Decisão

Duque (2019) exibi que árvore de decisão é modelada através de um conjunto de decisões hierárquicas sobre as variáveis explicativas, organizadas em estrutura de árvore. Os nós das árvores são os critérios de divisão. Cada nó na árvore representa logicamente um subconjunto do espaço de dados definido pela combinação de critérios de divisão nos nós acima dele. O objetivo principal das árvores de classificação é particionar os dados em grupos menores que sejam homogêneos. A homogeneidade significa que cada nó de divisão contém uma proporção maior de uma determinada classe. Os métodos matemáticos Índice de Gini e entropia são os dois procedimentos mais utilizados para a escolha das partições. Em cada nó, é selecionada a variável que melhor promova a separação das classes, de acordo com o critério utilizado.

Karthe (2016) expõe a terminologia básica usada em árvores de decisão:

1. **Nó Raiz**: Representa a população inteira ou amostra, sendo ainda dividido em dois ou mais conjuntos homogêneos.
2. **Divisão**: É o processo de dividir um nó em dois ou mais sub-nós.
3. **Nó de Decisão**: Quando um sub-nó é dividido em sub-nós adicionais.
4. **Folha ou Nó de Término**: Os nós não divididos são chamados Folha ou Nó de Término.
5. **Poda**: O processo de remover sub-nós de um nó de decisão é chamado poda. Podemos dizer que é o processo oposto ao de divisão.
6. **Ramificação/Sub-Árvore**: Uma sub-seção da árvore inteira é chamada de ramificação ou sub-árvore.
7. **Nó pai e nó filho**: Um nó que é dividido em sub-nós é chamado de nó pai. Os sub-nós são os nós filhos do nó pai.

Árvore de decisão é um tipo de algoritmo de aprendizagem supervisionada (com uma variável alvo pré-definida), muito utilizada em problemas de classificação. Ele funciona para ambas as variáveis categóricas e contínuas de entrada e de saída. Na árvore de decisão, dividimos a população ou amostra em dois ou mais conjuntos homogêneos (ou sub-populações) com base nos divisores/diferenciadores mais significativos das variáveis de entrada. Ao contrário dos modelos lineares, eles mapeiam muito bem relações não-lineares. E podem ser adaptados para resolver vários tipos de problema (classificação ou regressão), conforme explanado por Karthe (2016).

2.1.2.2 Método de Agregação (*Bagging*)

Karthe (2016) afirma que Agregação (*Bagging*) é uma técnica usada para reduzir a variância das previsões. Ela combina o resultado de vários classificadores, modelados em diferentes sub-amostras do mesmo conjunto de dados. A figura 4 a seguir deixa mais claro:

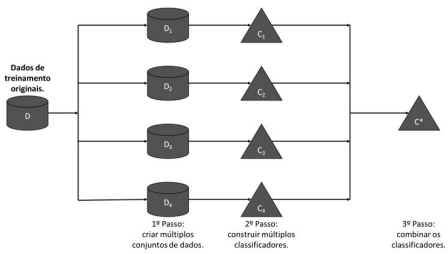


Figura 4 – Etapas do Agregação (*Bagging*).
Fonte: Baseado em Karthe (2016)

As etapas do Agregação (*Bagging*) são as seguintes, segundo Karthe (2016):

1. Criar vários conjuntos de dados:
    - A amostragem é feita com a substituição dos dados originais e a formação de novos conjuntos de dados;
    - Os novos conjuntos de dados podem ter uma fração das colunas e das linhas, que geralmente são hiper-parâmetros em um modelo de Agregação (*Bagging*);
    - Tomando frações de linha e coluna menores que 1 ajuda na montagem de modelos robustos, menos propensos a sobreajuste.
2. Criar múltiplos classificadores
    - Classificadores são construídos em cada conjunto de dados;
    - Em geral, o mesmo classificador é modelado em cada conjunto de dados, e a partir disso as previsões são feitas.
3. Combinar classificadores:
    - As previsões de todos os classificadores são combinadas usando a média, a mediana ou a moda, dependendo do problema;
    - Os valores combinados são geralmente mais robustos do que um único modelo.

Note-se que o número de modelos construídos não são hiper-parâmetros. Um maior número de modelos é geralmente melhor. Ou podem dar um desempenho semelhante ao de números mais baixos. Pode-se mostrar em teoria que a variância das previsões combinadas é reduzida para 1 / n (n: número de classificadores) da variância original, sob algumas premissas, conforme Karthe (2016) expõe.

Segundo Karthe (2016), existem várias implementações de modelos Agregação (*Bagging*). O modelo de floresta aleatória (*random forest*) é uma delas e a discutiremos a seguir. Duque (2019) combina diferentes modelos de aprendizagem para melhorar a previsão da acurácia do modelo. O método diminui a variância da predição dos modelos, através da amostragem com repetição dos dados originais. Assim, florestas aleatórias constroem e combinam múltiplas árvores de decisão para obtenção de uma melhor acurácia.

2.1.2.3 Floresta Aleatória (*Random Forest*)

Segundo Duque (2019), floresta aleatória (*random forest*) é um conjunto de árvores de decisão, treinadas através do método de Agregação (*Bagging*). Para Karthe (2016), é um método de aprendizagem versátil e capaz de executar tarefas de regressão e de classificação. Ele também aplica métodos de redução dimensional, trata valores faltantes, *outliers* e outras etapas essenciais da exploração de dados.

Duque (2019) explana que no método de florestas aleatórias, a criação de uma árvore é feita selecionando aleatoriamente, através de *bootstrap*, um conjunto de registros com repetição. Além disso, em cada nó de divisão se faz a análise apenas com um subconjunto de variáveis escolhidas aleatoriamente, ao invés de se considerar todas as variáveis. São feitas então diferentes árvores repetindo o procedimento, tendo como resposta um conjunto de modelos de árvores. Durante a classificação de um novo registro, cada árvore irá votar e a classe com maior número de votos será a resposta obtida.

Para Karthe (2016) funciona da seguinte maneira. Cada árvore é plantada e cultivada da seguinte forma:

1. Assume que o número de casos no conjunto de treinamento é N. Então, a amostra desses N casos é escolhida aleatoriamente, mas com substituição. Esta amostra será o conjunto de treinamento para o cultivo da árvore.
2. Se houver M variáveis de entrada, um número m < M é especificado de modo que, em cada nó, m variáveis de M sejam selecionadas aleatoriamente. A melhor divisão nestes m é usada para dividir o nó. O valor de m é mantido constante enquanto crescemos a floresta.
3. Cada árvore é cultivada na maior extensão possível e não há poda.
4. Preveja novos dados agregando as previsões das árvores (ou seja, votos majoritários para classificação, média para regressão).

De acordo com Karthe (2016), vantagens do modelo floresta aleatória (*random forest*):

1. Este algoritmo pode resolver os problemas de classificação e de regressão, fazendo uma estimativa decente em ambos.
2. Um dos benefícios da floresta aleatória que me agrada mais é o poder de lidar com dados em grandes volumes e com muitas dimensões. Ele pode lidar com milhares de variáveis de entrada e identificar as variáveis mais significativas, sendo por isso considerado um dos métodos de redução de dimensões. Além

disso, o modelo produz o grau de importância das variáveis, o que pode ser um dado muito útil (em algum conjunto de dados aleatórios).

3. Possui um método eficaz para estimar os dados faltantes e mantém a precisão quando uma grande parte dos dados estão faltando.
4. Possui métodos para equilibrar erros em conjuntos de dados onde as classes são desequilibradas.
5. As capacidades do método anterior podem ser estendidas para dados sem rótulo, levando os agrupamentos (c*lusters*) não supervisionados, visualizações de dados e detecção *outliers*.
6. A floresta aleatória envolve a amostragem dos dados de entrada com substituição chamada como amostragem de '*bootstrap*'. Aqui um terço dos dados não é usado para treinamento e pode ser usado para testes. Estes são chamados de amostras de fora da cesta. O erro estimado nas amostras de fora da cesta é conhecido como erro de fora da cesta. O estudo de estimativas do erro de fora da cesta fornece evidências para mostrar que a estimativa de fora da cesta é tão precisa quanto usar um conjunto de teste do mesmo tamanho que o conjunto de treinamento. Portanto, usar a estimativa de erro de fora da cesta remove a necessidade de ter um conjunto de teste extra.

Desvantagens do modelo floresta aleatória (*random forest*):

1. Enquanto faz um bom trabalho na classificação, já não é tão bom para o problema de regressão, uma vez que não fornece previsões precisas para variáveis contínuas. No caso da regressão, não prevê além do intervalo dos dados de treinamento, e que eles podem sobre-ajustar os conjuntos de dados que tenham muita discrepância ('*noisy*').
2. A floresta aleatória pode ser considerada como uma caixa preta para quem faz modelagem estatística – você tem muito pouco controle sobre o que o modelo faz. Você pode, na melhor das hipóteses, experimentar diferentes parâmetros.

2.2 Análise de Agrupamento (*Cluster*)

Em seu livro, muito utilizado nas faculdades de Estatística como por exemplo na ENCE, UFMG e USP, Mingoti (2007) disserta que a análise de agrupamento (*cluster*) tem como objetivo dividir os elementos da amostra, ou população, em grupos de forma

que os elementos pertencentes a um mesmo grupo sejam similares entre si com respeito às variáveis (características) que neles foram medidas, e os elementos em grupos diferentes sejam heterogêneos em relação a estas mesmas características. Ghosn (2020) reforça esse objetivo em "O objetivo do agrupamento é organizar, agrupar uma coleção de dados em grupos, tal que os dados contidos dentro de um mesmo agrupamento (*cluster*) são mais "parecidos" entre si que em relação aos indivíduos dos demais. A noção de similaridade pode ser expressa de diversas formas, de acordo com a proposta de cada estudo."

Esta técnica é muito popular em mineração de dados, que está relacionada com análise de dados e uso de ferramentas computacionais na busca de padrões em conjunto de dados. E muito utilizada para encontrar padrões de grupos em pesquisas de mercado, segundo Alamsyah e Nurriz (2017).

2.2.1 Medidas de Similaridade

Mingoti (2007) revela que uma questão significativa referente ao critério utilizado ao determinar até que ponto dois elementos do conjunto de dados podem ser considerados como semelhantes ou não. A solução desta questão exige considerar medidas que descrevam a similaridade entre elementos a partir das características que neles foram medidas. Cada componente apresenta informações em p-variáveis inserida num vetor a comparação de diferentes elementos onde será feita através de medidas matemáticas (métricas), que possibilitem a comparação de vetores, como as medidas de distância. Assim o cálculo das distâncias entre os vetores, de observações dos elementos e agrupar aqueles de menor distância se torna plausível de realização.

De acordo com Mingoti (2007), num conjunto de dados de n elementos amostrais, com p-variáveis aleatórias em casa, agrupar estes elementos em g grupos se torna o objetivo. Para casa elemento amostral j tem-se o vetor de medidas X_j:

$$X_j = [\ X_{1j}\ X_{2j}\ ...\ X_{nj}\ ]', j = 1, 2, ..., n$$

onde X_{ij} representa o valor observado da variável *i* medida no elemento *j*. A decisão da medida de similaridade a ser utilizada é o primeiro passo para que se possa realizar o agrupamento dos elementos. Inúmeros medidas existem e em cada uma delas aborda um tipo de análise de agrupamento (*cluster*). Porém, a única medida de similaridade apresentada neste estudo é Distância Euclidiana, logo em seguida é possível ter acesso a

ela. Esta distância é utilizada no método de K-Médias *(K-Means)*, método este utilizado e apresentado mais a frente neste trabalho.

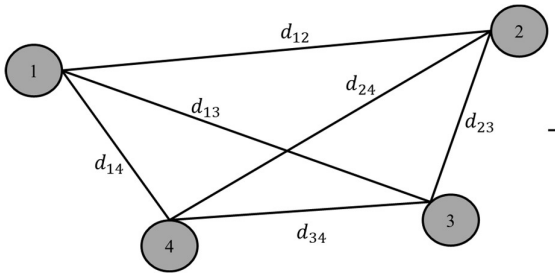
2.2.1.1 Distância Euclidiana

Mingoti (2007) apresenta a Distância Euclidiana entre dois elementos X_l e X_k, $l \neq k$:

$$d(X_l, X_k) = [(X_l - X_k)'(X_l - X_k)]^{1/2} = \left[\sum_{i=1}^{p}(X_{il} - X_{ik})^2\right]^{1/2} \qquad (1)$$

ou seja, dois elementos amostrais são comparados em cada variável pertencente ao vetor de observações.

Uma matriz de dimensão *nxn*, chamada de matriz de distâncias, armazena as distâncias entre os elementos amostrais. A figura 5 mostra esta matriz onde d_{ij} representa a distância do elemento amostral *i* ao elemento amostral *j*, apresentado por Mingoti (2007).



$$\rightarrow D_{4x4} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{21} & 0 & d_{23} & d_{24} \\ d_{31} & d_{32} & 0 & d_{34} \\ d_{41} & d_{42} & d_{43} & 0 \end{bmatrix}$$

Figura 5 – Esquema de armazenamento de distâncias numa matriz 4x4.
Fonte: Baseado em Mingoti (2007).

2.2.2 Agrupamento (*Cluster*) Não Hierárquico

Mingoti (2007) exprime que a análise de agrupamentos (*clusters*) é classifica em duas: técnicas hierárquicas e não hierárquicas. As técnicas hierárquicas são usualmente utilizadas em análises exploratórias dos dados na identificação de possíveis agrupamentos (*clusters*) e o valor provável do número de agrupamentos (*clusters*) g. Nas técnicas não hierárquicas é imprescindível que o valor do número de agrupamentos (*clusters*) já esteja pré-especificado.

Conforme Mingoti (2007), as técnicas não hierárquicas têm como objetivo encontrar diretamente uma partição de n elementos em k agrupamentos (*clusters*), assim

a partição satisfaz dois requisitos básicos: semelhança interna e separação dos agrupamentos (*clusters*) formados. Na busca da "melhor" partição de ordem k, algum critério de qualidade da partição deve ser aplicado. A criação de todas as partições possíveis de ordem k se torna impossível, computacionalmente, e, de acordo com o conhecimento destas partições, decidir a mais apropriada. Portanto, processos que apurem algumas das partições possíveis no objetivo da descoberta da partição "quase ótima" se tornam necessários.

Existem diferenças em certos aspectos nas técnicas não hierárquicas e hierárquicas. Avessa as técnicas hierárquicas aglomerativas, as não hierárquicas necessitam previamente do número de agrupamentos (*clusters*) k almejado. Em cada estágio do agrupamento, novos grupos são formados através da divisão ou junção de grupos já combinados em passos anteriores. OU seja, caso em determinado passo do algoritmo dois elementos foram inseridos no mesmo agrupamento (*cluster*), não obrigatoriamente estarão juntos na partição final. Em consequência disto, não é mais possível a construção de dendogramas, como descrito por Mingoti (2007).

É importante ressaltar que este estudo aborda somente a técnica não hierárquica. Esta decisão é tomada pois a outra técnica não é pertinente para o fluxograma abordado pois tem uma maior capacidade de análise de conjunto de dados de maior porte, ou seja, com um grande número de observações. Esta certeza se torna mais plausível de acordo com Santana e Pontes (2020) que utilizam a técnica não hierárquica na análise agrupamentos (*clusters*) em um sistema de recomendação de produtos baseado em perfis de compra, similar ao fluxograma abordado neste estudo. Outro estudo que auxiliou esta decisão foi o de Ghosn (2020) que utilizou somente esta técnica na construção de um modelo qualitativo-quantitativo de avaliação ordenada de mercados para a internacionalização, através da exportação, de PMEs da indústria alimentícia brasileira, embora, assim como neste estudo, cita a técnica hierárquica porém reconhece a técnica não hierárquica se encaixa melhor e de forma mais direta.

Métodos K-Médias (*K-Means*) e C-Médias Difusas (*Fuzzy C-Means*) são exemplos de métodos não hierárquicos, assim como redes neurais artificiais aplicadas à análise de agrupamentos (*clusters*). Este estudo aborda apenas o método de K-Médias *(K-Means)* pela praticidade e efetividade a ser utilizada no fluxograma abordado. Santana e Pontes (2020), Ghosn (2020) e Gavira-Durón, Gutierrez-Vargas e Cruz-Aké (2021) utilizam K-Médias (*K-Means*) como método para a formação de agrupamentos (*clusters*).

Gavira-Durón, Gutierrez-Vargas e Cruz-Aké (2021) empregaram K-Médias (*K-Means*) em agrupamentos (*clusters*) de dados obtidos através de simulações de Cadeias de Markov, o que aproxima deste estudo na junção das técnicas de agrupamento e simulação de dados.

2.2.2.1 Método K-Médias *(K-Means)*

O método K-Médias *(K-Means)* é provavelmente um dos mais conhecidos e mais utilizados métodos de análise de agrupamento (*cluster*) em problemas práticos, e isto pode ser visto nos estudos, já apresentados anteriormente neste texto, de Santana e Pontes (2020) e Ghosn (2020). E, vale ressaltar também, o estudo que Gavira-Durón, Gutierrez-Vargas e Cruz-Aké (2021) onde aplicaram este no estudo da migração entre agrupamentos (*clusters*) de uma análise que determina a probabilidade de créditos de Hedge migrarem para especulativo e depois para Ponzi, por meio de simulações com cadeias de Markov homogêneas.

Mingoti (2007) apresenta este método explanando que, basicamente, cada elemento amostral é alocado aquele agrupamento (*cluster*) cujo centróide (vetor de médias amostral) é o mais próximo do vetor de valores observados para o respectivo elemento. É formulado por quatro passos:

1. Primeiramente escolhe-se k centróides, chamados de “sementes” ou “protótipos”, para dar início ao processo de partição;
2. Cada elemento do conjunto de dados é, então, comparado com cada centróide inicial, através de uma medida de distância que, em geral, é a distância Euclidiana. O elemento é alocado ao grupo cuja distância é a menor;
3. Após aplicação do passo 2, para cada um dos n elementos amostrais, recalcula-se os valores dos centróides para cada novo grupo formado, e repete-se o passo 2, considerando os centróides destes novos grupos;
4. Os passos 2 e 3 devem ser repetidos até que todos os elementos amostrais estejam bem alocados em seus grupos, ou seja, até que nenhuma realocação de elementos possa ser possível.

No momento da escolha das sementes certos cuidados são fundamentais pois o agrupamento (*cluster*) final é diretamente impactado pelas sementes iniciais. Formas

diferentes para implementar K-Médias *(K-Means)* são aplicadas, computacionalmente, nos programas (*softwares*) estatísticos, assim como Mingoti (2007) aborda. Sendo assim, algumas sugestões para esta escolha são:

1. Uso de técnicas hierárquicas aglomerativas – primeiramente, aplica-se alguma técnica de agrupamento hierárquicas aglomerativas para se obter os g=k grupos iniciais. Em seguida, calcula-se o vetor de médias de cada grupo formado, sendo esses vetores de médias as sementes iniciais usadas no método das K-Médias *(K-Means)*.
2. Escolha aleatória – as k sementes iniciais são aleatoriamente escolhidas dentro do conjunto de dados. Um procedimento amostral que pode ser utilizado para isso é o de amostragem aleatória simples sem reposição. Essa estratégia de escolha das sementes não é eficiente, embora seja de execução simples. Uma forma de melhorar sua eficiência é selecionar m amostras aleatórias constituídas de k sementes, m>1. Desse modo, o procedimento de amostragem aleatória simples e repetido m vezes e, no final, calcula-se o vetor de médias das m sementes selecionadas para cada grupo. Estes vetores constituem os centróides de inicialização do processo de agrupamento das K-Médias *(K-Means)*.
3. Escolha via uma variável aleatória – escolhe-se a variável aleatória de maior variância dentre as p componentes do vetor aleatório X em consideração. Assim, a variável já induz uma partição natural dos dados. Divide-se o domínio da variável em k intervalos. A semente inicial será o centróide de cada intervalo.
4. Observação dos valores discrepantes do conjunto de dados – através de uma análise estatística, busca-se k elementos discrepantes no conjunto de dados. Cada um desses elementos estabelecerá a semente de um conglomerado inicial. A discrepância neste caso é em relação às p-variáveis observadas conjuntamente.
5. Escolha prefixada – as sementes são escolhidas aleatoriamente. É um método não muito recomendável devida a subjetividade. Todavia, pode ser usado em casos nos quais o indivíduo tenha um grande conhecimento do problema estudado ou quer validar uma solução já existente.

6. Os k primeiros valores do banco de dados – a maioria dos programas (*softwares*) estatísticos usa, como *default* na escolha de sementes iniciais, as k primeiras observações do banco de dados, a menos que seja especificado diretamente quais sementes para iniciar o algoritmo serão utilizadas.

Mingoti (2007) termina sua explanação sobre K-Médias *(K-Means)* indicando que a forma de escolha das sementes é de responsabilidade do indivíduo que utiliza este método. Para isto, é importante o correto entendimento do banco de dados para assim satisfazer o melhor requisito de escolha das sementes.

2.3 Qualidade

Como já escrito anteriormente, Ramos *et. al* (2021) declara que a qualidade de produtos e serviços tem continuamente o interesse de melhora, centrando-se em preferências e necessidades específicas de cada indivíduo. Este pensamento sobre qualidade é compartilhado por Lui e Petarnella (2020), Aparecida e Rezende (2021) e Ferreira, Seifert e Venanzi (2020) em seus estudos mais recentes que abordam, cada um da sua forma, a importância da qualidade nos dias de hoje.

Segundo Dean e Bowen (1994), a palavra qualidade é usada para significar coisas diferentes em diferentes estudos, tais como a qualidade do processo interno, uma ou várias dimensões da qualidade do produto, a satisfação do cliente e o desempenho operacional.

Em seu livro, muito utilizado nas faculdades de Estatística como por exemplo na ENCE e USP, Montgomery (2009) disserta que a qualidade pode ser definida de várias maneiras. A maioria dos indivíduos tem uma compreensão conceitual de qualidade como algo relacionado a uma ou mais características desejáveis que um produto ou serviço deva ter, o que como ponto de partida seja bom para uma compreensão conceitual.

A qualidade tornou-se um dos mais importantes fatores de decisão dos consumidores na seleção de produtos e serviços que competem entre si. Este fato é geral, independente do fato de ser um indivíduo, uma organização industrial, uma loja de varejo, ou um programa militar de defesa. Sendo assim, compreender e melhorar a qualidade é um fator-chave que conduz ao sucesso, crescimento e melhor posição de competitividade no negócio. A melhor e o emprego bem-sucedido da qualidade como parte integrante da estratégia geral de uma empresa produzem grande retorno sobre o investimento. A

melhoria da qualidade é a redução da variabilidade nos processos, produtos e serviços. Uma definição alternativa e altamente útil é que a melhoria da qualidade é a redução do desperdício, como Montgomery (2009) aborda.

2.3.1 Controle Estatístico de Processo – CEP

Este método é continuamente utilizado, conforme observamos nos estudos de Santana *et. al* (2019), Fonseca *et. al* (2020) e Soriana, Oprime e Lizarelli (2020). Estes estudos abordam que a exigência por produtos e serviços de qualidade vem crescendo a cada dia, e com isso as empresas têm que se adaptar para conquistar os clientes e melhorar seu desempenho quanto à qualidade e à produtividade, e que o CEP corresponde a um conjunto de ferramentas estatísticas que visam à melhoria e estabilidade dos padrões da qualidade. Santana *et. al* (2019) analisam o processo de confecção em roupas femininas pela a filosofia *slow fashion*, através da utilização do CEP, especificamente de gráficos de controle, e identificam que o processo produtivo está fora de controle estatístico. Soriana, Oprime e Lizarelli (2020) fazem uma revisão de literatura sobre o CEP. Fonseca *et. al* (2020) aplicam a engenharia de métodos junto com o CEP na linha de produção em uma fábrica que produz pão de forma, desta forma visando a qualidade desta linha de produção.

Um produto ou serviço tem exigências esperadas por um cliente, e estas devem ser produzidas em um processo estável ou replicável. O processo deverá operar com mínima variabilidade ao redor das dimensões-alvo ou nominais das características de qualidade do produto ou serviço. Um poderoso conjunto de ferramentas de resolução de problemas útil para estabilidade do processo e melhoria da capacidade através da redução da variabilidade é o controle estatístico do processo (CEP), segundo Montgomery (2009).

Montgomery (2009) afirma que qualquer processo pode ser aplicado pelo CEP. Suas sete principais ferramentas são:

1. Apresentação em histogramas ou ramo-e-folhas
2. Folha de controle
3. Gráfico de Pareto
4. Diagrama de causa-e-efeito
5. Diagrama de concentração de defeito
6. Diagrama de dispersão

7. Gráfico de controle

Apesar dessas ferramentas, geralmente chamadas "as sete ferramentas", sejam uma parte importante do CEP, incluem apenas seus aspectos técnicos. O CEP constrói uma atmosfera onde todos os indivíduos numa organização desejam a melhora continuada na qualidade e na produtividade. Essa atmosfera é melhor próspera quando a gerência se envolve em um processo contínuo de melhoria da qualidade. Estabelecida essa atmosfera, a rotina de aplicação das sete ferramentas se torna habitual da maneira de se fazerem negócios, e a organização foca na obtenção de seus objetivos de melhoria da qualidade.

Das sete ferramentas do controle da qualidade o gráfico de controle de Shewhart é, provavelmente, o mais sofisticado tecnicamente. Seu desenvolvimento ocorreu nos anos 20 pelo Dr. Walter A. Shewhart, do *Bell Telephone Laboratories*. Sua finalidade era compreender os conceitos estatísticos que compõem a base do CEP. Mas antes de apresentar este gráfico é necessário apresentar a teoria da variabilidade de Shewhart, conforme abordado por Montgomery (2009).

Num processo de produção, independentemente de ser bem estruturado ou cuidadosamente mantido seja, certa quantidade de variabilidade inerente ou natural sempre existirá. Variabilidade natural ou "ruído de fundo" é o efeito cumulativo de inúmeras pequenas causas, essencialmente inevitáveis. No sistema do controle estatístico da qualidade, essa variabilidade natural é, comumente, conhecida como "sistema estável de causas aleatórias". É possível afirmar que um processo que opera somente com causas aleatórias da variação está sob controle estatístico. Ou seja, as causas aleatórias são partes inerentes ao processo, segundo Montgomery (2009).

Montgomery (2009) cita que outros tipos de variabilidade podem, casualmente, ocorrerem na saída de um processo. Tal variabilidades são, habitualmente, muito grandes quando comparada com o “ruído de fundo”, representando um nível inaceitável do desempenho do processo. Essas fontes de variabilidade, que não pertencem ao padrão de causas aleatórias, são conhecidas como "causas atribuíveis". Sendo assim, um processo contendo causas atribuíveis está fora de controle.

A figura 6 explana causas de variabilidade aleatórias e atribuíveis. Na figura 6, o processo está sob controle até instante t_0, ou seja, somente as causas de variação aleatórias estão presentes. Com isso, a média e o desvio padrão do processo estão com seus valores sob controle (μ_0 e σ_0, respectivamente). Uma causa atribuível acontece no instante t_1 e o

deslocamento da média do processo para um valor novo $\mu_1 > \mu_0$ é a consequência desta causa atribuível, também observado na figura 8. Uma outra causa atribuível decorre no instante t_2, resultando em $\mu = \mu_0$, neste momento o desvio padrão do processo se deslocou para um valor maior $\sigma_1 > \sigma_0$. Também transcorre uma causa atribuível no instante t_3, resultando em valores fora de controle para a média e para o desvio padrão. Em decorrente a causas atribuíveis desde o instante t_1, este processo está fora de controle, como Montgomery (2009) aborda.

A maioria da produção estará entre os limites inferior (LIE) e superior (LSE), quando um processo está sob controle. Já o processo está for de controle quando uma proporção maior de saída do processo fica fora dessas especificações, de acordo com Montgomery (2009).

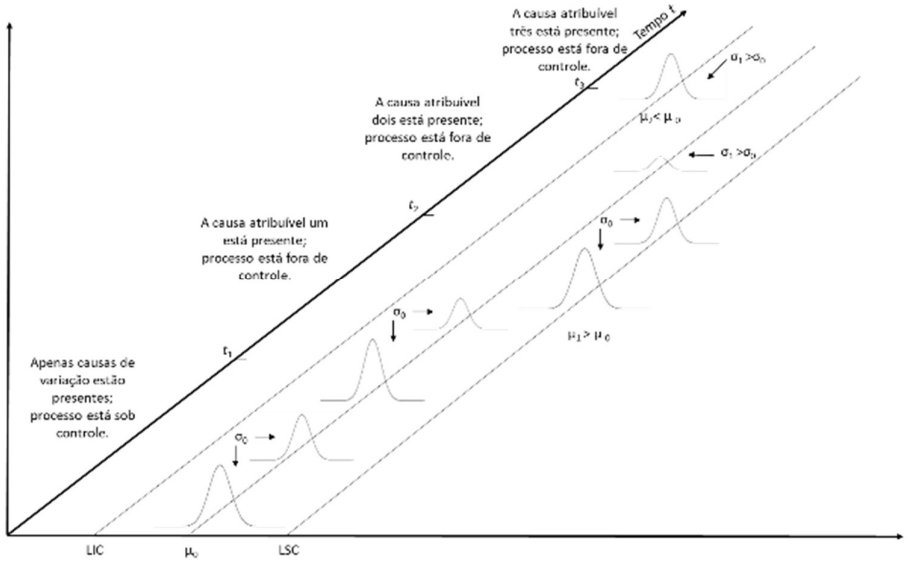


Figura 6 – Causas atribuíveis e aleatórias da variabilidade.
Fonte: Baseado em Montgomery (2009).

A figura 7 apresenta o gráfico de uma característica da qualidade medida ou calcula de uma amostra pelo número da amostra ou no tempo, um típico exemplo de gráfico de controle. A linha central do gráfico representa o valor médio da característica da qualidade que corresponde ao estado sob controle, ou seja, somente causas aleatórias são presentes. A linha horizontal superior é chamada de limite superior de controle (LSC) e a inferior é a linha inferior de controle (LIC). Esses limites de controle servem para que averiguar se o processo está sob controle, caso todos os pontos amostrais estejam dentro deles. Quando o processo é considerado sob controle, os pontos estão dentro dos limites de controle, não existindo a necessidade de ação. Quando um ponto está fora dos limites

de controle, existe a evidência de que o processo esteja fora de controle. Neste caso, uma investigação e uma ação corretiva são necessárias para descobrir e acabar com a causa ou causas atribuíveis responsáveis por esse comportamento, como Montgomery (2009) descreve.

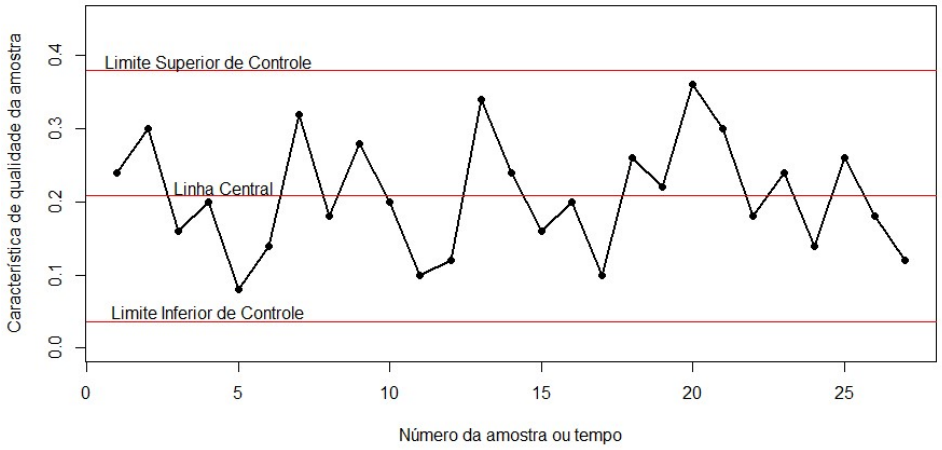


Figura 7 – Exemplo de gráfico de controle.
Fonte: Baseado em Montgomery (2009).

Montgomery (2009) deixa claro que mesmo que todos os pontos estejam entre os limites de controle, caso seu comportamento esteja de maneira sistemática ou não-aleatória, isso pode ser um indício que o processo está fora de controle. Por exemplo, se, dos últimos 20 pontos marcados, 18 estiverem acima da linha central, mas abaixo do limite superior de controle e apenas 2 estiverem abaixo da linha central, mas acima do limite inferior de controle, são motivos para suspeitar que algo esteja errado. Se o processo está sob controle, todos os pontos devem ter um padrão estritamente aleatório.

Montgomery (2009) apresenta um modelo geral para o gráfico de controle. Considere a estatística amostral w que mede alguma característica da qualidade de interesse, e considera a média de w como μ_w e o desvio padrão σ_w. Sendo assim, a linha central e os limites superior e inferior de controle são:

$$LSC = \mu_w + L\sigma_w \qquad (2)$$

$$\text{Linha central} = \mu_w \qquad (3)$$

$$LIC = \mu_w - L\sigma_w \qquad (4)$$

onde L é a "distância" dos limites de controle à linha central, manifesta-se em unidades de desvio padrão. Essa teoria geral dos gráficos de controle foi proposta, primeiramente, pelo Dr. Walter S. Shewhart, e os gráficos de controle desenvolvidos segundo esses princípios são, geralmente, chamados de gráficos de controle de Shewhart.

2.3.2 Modelo Seis Sigma

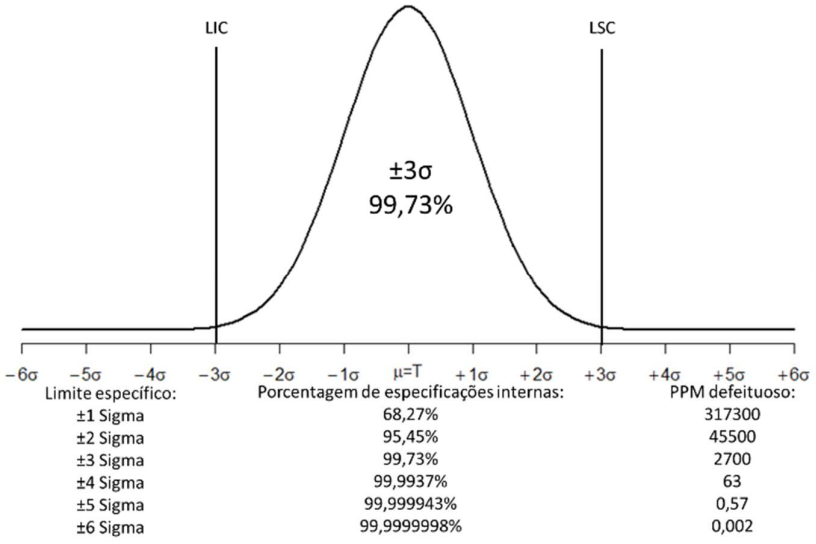
O modelo de gestão da qualidade abordado neste estudo é o Seis Sigma, pois segundo Carvalho e Paladini (2012) diferentemente de outros programas de qualidade, as empresas que utilizam o Seis Sigma divulgam cifras milionárias de ganhos obtidos com sua implementação. O sucesso dos programas Seis Sigma não pode ser explicado apenas pela utilização exaustiva de ferramentas estatísticas, mas também pela harmoniosa integração do gerenciamento por processo e por diretrizes, mantendo o foco nos clientes, nos processos críticos e nos resultados da empresa. Este modelo tem sido chamado como a qualidade para o século XXI, como explica Montgomery (2009).

Segundo Silva *et al.* (2020), a competitividade entre as empresas requer a busca por alternativas não só na garantia por produtos melhores aos clientes, mas no que diz respeito ao processo produtivo sem desperdícios e elevados custos. O estudo apresenta a implementação da metodologia Seis Sigma como meio de obter resultados melhores na qualidade do processo produtivo e como uma mudança organizacional para com a resolução de problemas em uma indústria alimentícia.

Tem-se também Barbosa *et al.* (2020), que abordam um estudo que visa a redução da variabilidade dos recursos restritivos, diminuição de perdas, ineficiências e desperdícios, otimização do trade-off custo-qualidade dos produtos e processos, além da consolidação do conceito de Indústria 4.0, com embasamento da metodologia Seis Sigma. Este estudo e o Silva *et al.* (2020) comprovam a contemporaneidade deste modelo.

Montgomery (2009) aborda que produtos de alta tecnologia com muitos componentes complexos têm, tradicionalmente, muitas possibilidades para falhas e defeitos. Tendo isso em vista, a Motorola desenvolveu seu programa a seis sigma, na década de 80, como solução a esta demanda. O programa se concentra na redução da variabilidade nas principais características de qualidade do produto, no grau em que falhas e defeitos são improváveis.

É possível averiguar que a figura 8 apresentada uma distribuição de probabilidade normal como modelo para uma característica da qualidade, tendo limites de especificação em três desvios padrão de cada lado da média. Nessa condição, a probabilidade de gerar um produto dentro dessas especificações é de 0,9973, correspondendo a 2700 partes por milhão (ppm) de defeituosos. Isto é conhecido como desempenho de qualidade três sigma, o que primeira vista configura algo bom. Porém, suponha um produto consista em um conjunto de 100 componentes ou partes e que todas essas 100 partes devem ser não defeituosas para que o produto exerça corretamente sua função. A probabilidade de uma unidade específica do produto ser não defeituosa é 0,9973 x 0,9973 x ... x 0,9973 = $(0,9973)^{100}$ = 0,7631. Ou seja, cerca de 23,7% dos produtos produzidos sob a qualidade três sigma serão defeituosos. Este contexto não é satisfatório, pois vários produtos de alta tecnologia são feitos de milhares de componentes, como por exemplo um carro, de acordo com Montgomery (2009).

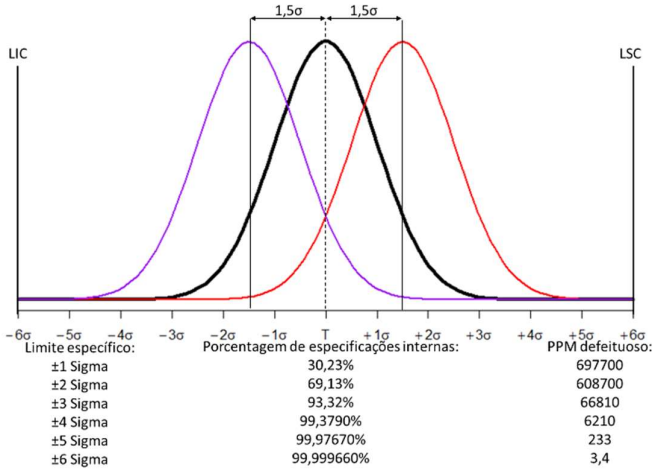


Limite específico:	Porcentagem de especificações internas:	PPM defeituoso:
±1 Sigma	68,27%	317300
±2 Sigma	95,45%	45500
±3 Sigma	99,73%	2700
±4 Sigma	99,9937%	63
±5 Sigma	99,999943%	0,57
±6 Sigma	99,9999998%	0,002

Figura 8 – Distribuição normal centrada no alvo (T).
Fonte: Baseado em Montgomery (2009).

Montgomery (2009) mostra que o conceito seis sigma da Motorola é reduzir a variabilidade no processo de modo que os limites de especificação estejam a seis desvios padrão da média. Sendo assim, como mostrado na figura 8, terá apenas duas partes por bilhão de defeituosos. Sob a qualidade seis sigma, a probabilidade de uma unidade específica do produto seja não defeituosa é de 0,9999998, ou 0,2 partes por milhão, uma conjuntura muito melhor.

Conforme Montgomery (2009), no início de desenvolvimento do conceito seis sigma fez-se uma suposição que quando o processo alcançasse o nível de qualidade seis sigma, a média do processo estaria ainda sujeita a impactos que poderiam fazer com que ela mudasse em até 1,5 desvio padrão para longe do alvo, a figura 9 apresenta esta situação. O processo seis sigma produziria cerca de 3,4 ppm de defeituosos na pela figura abaixo.



Limite específico:	Porcentagem de especificações internas:	PPM defeituoso:
±1 Sigma	30,23%	697700
±2 Sigma	69,13%	608700
±3 Sigma	93,32%	66810
±4 Sigma	99,3790%	6210
±5 Sigma	99,97670%	233
±6 Sigma	99,999660%	3,4

Figura 9 – Distribuição normal com média deslocada de 1,5σ do alvo.
Fonte: Baseado em Montgomery (2009).

Um dos idealizadores deste "programa", Michel Harry, define o Seis Sigma como uma estratégia que não deve estar encapsulada na área de qualidade, deve ser de conhecimento de toda uma organização, da manufatura e engenharia à área de serviço, é o que Carvalho e Paladini (2012) descrevem.

Também para Carvalho e Paladini (2012) a utilização estruturada do pensamento estatístico é um dos elementos mais marcantes deste programa. O uso intensivo de ferramentas estatísticas e a sistemática análise da variabilidade são as marcas registradas deste programa, que lhe conferiu o nome Seis Sigma, significando, em linguagem estatística, seis desvios padrão. Essa é uma métrica de capacidade que implica um processo praticamente isento de erros, ou seja, com 3,4 defeitos por milhão de oportunidades.

Não é apenas o pensamento estatístico e a redução da variabilidade que caracterizam este programa. O Seis Sigma proporciona um alinhamento estratégico, utilizando indicadores de desempenho alinhados aos resultados da organização e prioridades estratégicas como objetivos dos projetos de melhoria, como explanado por Carvalho e Paladini (2012).

Em resumo, para Carvalho e Paladini (2012), o modelo de Gestão da Qualidade Seis Sigma é um método gerencial disciplinado, que provêm de uma abordagem sistêmica e a utilização intensiva do pensamento estatístico, cujo objetivo é reduzir eficientemente a variabilidade dos processos críticos e aumentar a lucratividade das empresas, por meio da otimização de produtos e processos, buscando satisfação de clientes e consumidores.

Percebe-se que orientar o negócio para a meta de capacidade Seis Sigma é um esforço de melhoria contínua, que exige o monitoramento das mudanças do mercado e agilidade para mudanças.

2.3.2.1 Modelo DMAIC

Conforme descrito por Carvalho e Paladini (2012), intitulado como a qualidade para o século XXI com metodologia de implementação denominada DMAIC (definir, medir, analisar, aperfeiçoar, controlar). O sucesso do modelo leva em conta o uso exaustivo de ferramentas estatísticas, bem como a robusta integração do alinhamento estratégico com indicadores de desempenho alinhados aos resultados da organização, além do gerenciamento por processos e diretrizes com foco nos clientes. Denominado Seis Sigma, pois em linguagem estatística, significa seis desvios padrão, sendo uma métrica que implica em um processo praticamente isento de erros. Dessa forma, recorrendo ao modelo Seis Sigma, as empresas ganham competitividade ao reduzir suas taxas de defeitos e ao aumentar a lucratividade, por meio da otimização de produtos e processos, buscando satisfação de clientes e consumidores.

Chiroli *et al.* (2020) descrevem que devido à crescente competitividade entre empresas, estas têm novos desafios relacionados à disputa de mercado, visto que esta situação não envolve apenas organizações fabris, como engloba organizações prestadoras de serviço. Por este motivo, as organizações têm a preocupação maior com a qualidade do serviço prestado e a satisfação do cliente, oferecendo aos clientes agilidade e qualidade na entrega do serviço. Sendo assim, o Seis Sigma auxilia a organização para reduzir a variabilidade, aumentando a satisfação do cliente e na qualidade do serviço. Com essas informações, o estudo propôs a implementação parcial da metodologia DMAIC em uma unidade de saúde localizada na região dos Campos Gerais, no estado do Paraná.

Segundo Carvalho e Paladini (2012), a metodologia DMAIC contempla cinco fases: definição, medição, análise, aperfeiçoamento e controle, a saber:

a) Fase 1: “D” definição (*define*) – definir as prioridades e as necessidades do cliente, traduzindo em características críticas para a qualidade. Esta etapa é fundamental, pois a voz e a visão do cliente são levadas para dentro da organização. O objetivo do projeto tem de estar relacionado a um. Assim, os processos críticos, especialmente, com resultados ruins, como reclamações de clientes, problemas funcionais, problemas trabalhistas, altos custos de mão de obra e baixa qualidade de suprimentos são identificados. Em seguida, é realizada uma análise custo-benefício do projeto, de modo a ter uma visão clara do retorno que a atividade trará à empresa.

b) Fase 2: “M” medição (*measure*) – medir e executar os processos relacionados às características críticas para a qualidade, definindo as entradas e as saídas, ou seja, as relações Y= f(X) são estabelecidas. O sistema de medição deve ser adequado para atender às necessidades do processo. Em seguida, há a coleta dos dados do processo por meio de um sistema que produza amostras representativas e aleatórias. Por fim, os índices de capacidade do processo a curto e longo prazos são determinados.

c) Fase 3: “A” análise (*analyze*) – analisar os dados coletados após a identificação das principais causas, utilizando, além das ferramentas tradicionais da qualidade, as ferramentas estatísticas de modo a identificar as causas óbvias, bem como as causas não óbvias. Os processos são analisados levando em conta sua variabilidade, descobrindo as causas geradoras dos defeitos e as fontes de variações nos processos.

d) Fase 4: “I” aperfeiçoamento (*improve*) – aperfeiçoar o processo existente mediante à transformação dos dados e à eliminação das causas dos defeitos. Evidentemente podem ocorrer modificações técnicas do processo e das atividades, atuando sobre as causas-raiz. Nesta fase, existe a oportunidade de uso dos conceitos de produção enxuta (*lean*), agregando ao sistema Seis Sigma uma grande possibilidade de melhorias vitais do processo, realizando a quantificação dos efeitos nas características críticas para a qualidade, ou seja, nas metas financeiras e de desempenho.

e) Fase 5: “C” controle (*control*) – controlar a manutenção das melhorias realizadas, por meio de um sistema de medição e controle estabelecido e validado para medir continuamente o processo de modo a garantir que a capacidade do processo seja mantida. O controle estatístico tem por objetivo conhecer a estabilidade do processo estudado, monitorando seus parâmetros ao longo do tempo. O monitoramento dos processos críticos é fundamental não só para manter a capacidade do processo

estabelecida, mas para indicar melhorias futuras. Nesta fase, é elaborada a documentação, além do monitoramento das novas condições do processo por meio de métodos estatísticos de controle de processo. A capacidade do processo é reavaliada para garantir que os ganhos sejam mantidos. Dependendo dos resultados desta reavaliação, pode ser necessário rever uma ou mais fases do processo. A implementação correta do programa Seis Sigma permite criar uma linguagem comum entre as diversas áreas de uma empresa, compartilhando sucessos e fracassos, fazendo com que uma unidade aprenda com a experiência de outra.

Portanto, conforme Carvalho e Paladini (2012), para estudar a capacidade do processo é preciso conhecer as suas especificações. Geralmente, quando se trata de uma indústria, as especificações são fornecidas pela área de produção e essas especificações são alteradas somente quando há um novo projeto. Já nas prestadoras de serviços, nem sempre as especificações são definidas a priori. Nesses casos, é de fundamental importância fazer pesquisa com os clientes e a avaliação comparativa (*benchmarking*) com os concorrentes para uma adequada definição das especificações. Vale destacar que as especificações se tornam obsoletas, ou seja, alterações significativas nos padrões tecnológicos e/ou de comportamento dos mercados, seguidas de melhoria do desempenho dos concorrentes, possuem impacto direto nos padrões de exigência do consumidor e, portanto, alteram as especificações. Indicadores que eram aceitáveis pelo cliente, pois não havia desempenho melhor, passa rapidamente a ser inaceitável, demandando ações drásticas das concorrentes para se manterem competitivas. Dessa forma, percebe-se que orientar o negócio para a meta de capacidade Seis Sigma é um esforço de melhoria contínua, que exige o monitoramento das mudanças do mercado e agilidade para mudanças.

2.3.2.2 Índice de Capacidade Sigma

Para gerenciar a qualidade segundo uma expectativa de redução da variabilidade dos processos, se requer das empresas a admissão sistemática de técnicas de controle estatístico e estudo dos índices de capacidade, de acordo com Carvalho e Paladini (2012).

Carvalho e Paladini (2012) dissertam que o controle estatístico tem por objetivo conhecer a estabilidade do processo estudado, monitorando seus parâmetros ao longo do

tempo. Ou seja, é como se o processo fosse um velho conhecido daqueles que se pode antever o comportamento.

No estudo dos índices de capacidade do processo, o propósito é determinar se um processo estável, ou seja, cujo comportamento seja conhecido, é capaz de produzir itens ou prestar o serviço de acordo com as especificações predeterminadas pelo cliente, como Carvalho e Paladini (2012) exibem.

Para estudar a capacidade do processo é preciso entender suas características. Ao se tratar de uma empresa de manufatura, a maioria das especificações é fornecida pelos responsáveis de engenharia das áreas de produção e essas especificações são alteradas apenas quando há um novo projeto, de acordo com Carvalho e Paladini (2012).

É importante ressaltar que o índice de capacidade é relacionado às especificações atribuídas, segundo Carvalho e Paladini (2012). Conclusões errôneas quanto à capacidade do processo podem ocorrer devido a erros na definição da especificação.

Chiroli *et al.* (2020) afirmam que após o cálculo do índice de capacidade sigma uma meta foi definida, um plano de ação desenvolvido e apresentado como proposta de melhoria à direção do hospital, podendo ser aceito ou adaptado de acordo com as necessidades identificadas pela organização.

Em conformidade com Carvalho e Paladini (2012), das métricas tradicionais mais utilizadas pelas empresas e que mais se assemelham ao índice de capacidade Seis Sigma são o Cpk e Ppk. Para se realizar cálculo desses índices de capacidade, é necessária que a hipótese de normalidade da distribuição seja válida para os dados e que o processo esteja sob controle estatístico (estável).

$$Cpk = min\left(\frac{LSE - \mu}{3\sigma}; \frac{\mu - LIE}{3\sigma}\right) \qquad (5)$$

$$Ppk = min\left(\frac{LSE - \bar{x}}{3s}; \frac{x - LIE}{3s}\right) \qquad (6)$$

onde:

- μ é a média do processo;
- $\bar{x}$ é a estimativa da média;
- σ é o desvio padrão do processo;
- s é a estimativa do desvio padrão;
- LSE é o limite superior de especificação;
- LIE é o limite inferior de especificação.

As figuras 10 e 11 demonstram a análise realizada por estes índices, quando a média do processo é equidistante dos limites de especificação. Diante disso, se o índice de capacidade for igual a 1, é possível comprovar que os limites de especificação distam três desvios-padrão da média, como afirmado por Carvalho e Paladini (2012). Se a média não estiver equidistante dos limites de especificação, adota-se a pior capacidade, ou seja, aquela calculada utilizando-se o limite de especificação mais próximo.

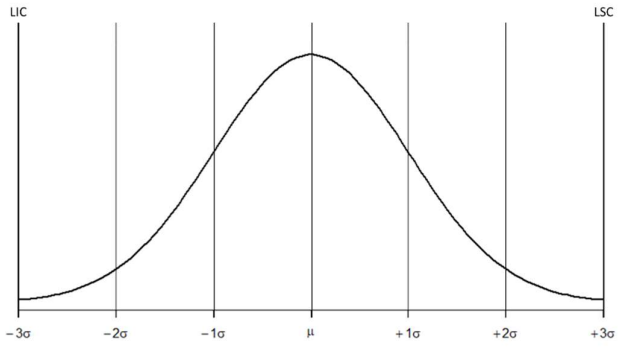


Figura 100 – Índice Cpk igual a 1.
Fonte: Baseado em Carvalho e Paladini (2012)

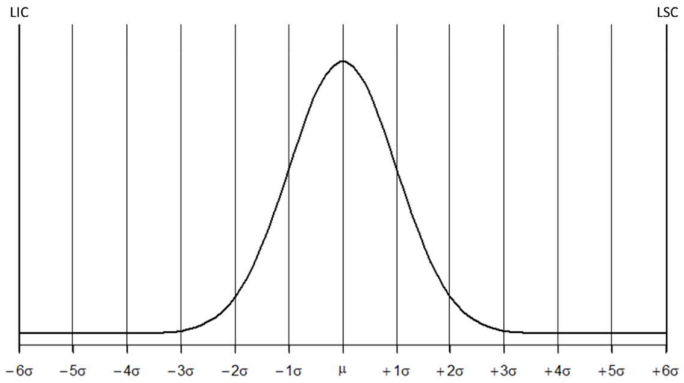


Figura 11 – Índice Cpk igual a 1.
Fonte: Baseado em Carvalho e Paladini (2012)

Carvalho e Paladini (2012) destacam que é importante ressaltar que em relação às métricas tradicionais o índice de capacidade Seis Sigma externa certas diferenças. O ponto importante após este é identificar se com este resultado pode-se concluir que o processo é capaz. Nos padrões Seis Sigma, a resposta seria "não", pois o índice Cpk precisaria ser igual a 2, conforme ilustra a figura 13. Verifica-se, então, que no padrão Seis Sigma, um processo considerado capaz é aquele cuja média esteja à distância de 6 desvios-padrão dos limites de especificação.

O índice para determinar a capacidade Seis Sigma é trivial, pois mede a distância da média à especificação mais próxima (LIE ou LSE) em quantidade de desvios-padrão (sigmas), usufruindo a normal reduzida (z), como dito por Carvalho e Paladini (2012). Abaixo a fórmula que exprime o índice de capacidade Seis Sigma.

$$z = \frac{x - \mu}{\sigma} \qquad (7) \rightarrow \begin{cases} z_i = \dfrac{LIE - \mu}{\sigma} = \dfrac{(\mu - 6\sigma) - \mu}{\sigma} = -6 & (8) \\ z_s = \dfrac{LSE - \mu}{\sigma} = \dfrac{(\mu + 6\sigma) - \mu}{\sigma} = 6 & (9) \end{cases}$$

Sendo, P(x<LIE) = P(z<-6) = 1,25 parte por bilhão P(x>LSE) = P(z>+6) = 1,25 parte por bilhão.

Comparando o índice de capacidade Seis Sigma com o Cpk e Ppk é observado que:

$$3Cpks = \frac{LSE - \mu}{3\sigma} Cpks = \frac{LSE - \mu}{\sigma} \qquad (10)$$

Ou seja, 3 Cpks = Zs e, de forma similar, Zi = 3 Cpki.

Carvalho e Paladini (2012) comentam que em vista de tudo visto até o momento, um processo Seis Sigma é aquele que gera 1,25 parte defeituosa por bilhão. Mas é preciso questionar se Seis Sigma representa 1,25 parte defeituosa ou 3,4 partes por milhão. Para obter uma explicação para este questionamento é importante pensar:

1. Qual a probabilidade correspondente, na tabela da distribuição normal padrão, para z = 6,0 ;
2. Na tabela normal padrão com a probabilidade 3,4 partes por milhão, qual é o z correspondente.

Esses pensamentos encaminham a probabilidade de 1,25 parte por bilhão para o item 1 e um z de 4,5 para o item 2. Mas estes resultados necessitam de uma explicação consistente. Posto isso, segundo Harry (1998), é complicado um processo sempre se preservar centralizado, já que a longo prazo inúmeros fatores ocasionam seu deslocamento (*shift*), para cima ou para baixo do valor-alvo da especificação, habitualmente, não superior a 1,5 desvio-padrão, conforme ilustra a figura 12. Desta forma, a capacidade obtida, analisando os dados do processo, é a de longo prazo (Z_{LP}).

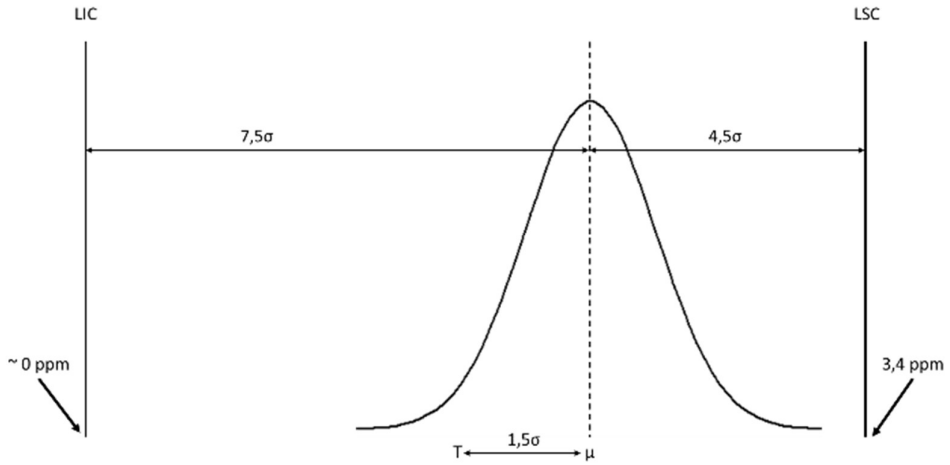


Figura 12 – Capacidade de longo prazo.
Fonte: Baseado em Carvalho e Paladini (2012)

Como Carvalho e Paladini (2012) citam, a capacidade potencial do processo, denominada de curto prazo (Z_{CP}), se desconta o deslocamento (Z_D=1,5), ou seja, o índice de capacidade é obtido da seguinte forma:

$$Z_{CP} = Z_{LP} + 1,5 \quad (11)$$

Deste modo, caso um processo tenha capacidade Seis Sigma, significa que sua capacidade potencial (Z_{CP}) é 6-sigma. Mas como esse processo se deslocou no decorrer do tempo, gera 3,4 partes por milhão de defeitos que correspondem à capacidade de longo prazo (Z_{LP} = 4,5), como apresentado abaixo na tabela 1:

Tabela 1 – Capacidade de longo prazo.
Fonte: Baseado em Carvalho e Paladini (2012)

Z_{CP}	Z_{LP}	PPM	
6	4,5	3,4	Classe Mundial
5	3,5	233	
4	2,5	6.210	Média da indústria
3	1,2	66.807	
2	0,5	308.537	Não competitiva

Quanto maior o valor de sigma, menor a probabilidade do processo gerar defeitos. Desse modo, quanto maior o sigma, maior a confiança dos clientes e menores os custos de não conformidades.

Carvalho e Paladini (2012) apresentam no detalhamento da capacidade para atributos é indispensável a definição de certos conceitos básicos, quais sejam: defeito, defeituoso,

unidade, oportunidade, defeito por unidade e defeitos por milhão de oportunidades, conforme tabela 2 abaixo:

Tabela 2 – Dados atributos.
Fonte: Baseado em Carvalho e Paladini (2012)

Defeito	Qualquer não conformidade às especificações.
Defeituoso	Unidade que apresenta um ou mais defeitos.
Unidade	Saída do processo que se´ra avaliada segundo a presença de defeitos.
Oportunidade	Formas que o processo tem de se desviar do que é específico para cada unidade, gerando não conformidade.
Defeitos por unidade - DPU	Número de defeito / Número de unidades.
Defeitos por oportunidade - DPO	Número defeito / (Número oportunidades * Número unidades)
Defeitos por milhão de oportunidade - DPMO	(Número defeito / (Número oportunidades * Número unidades)) * 10^6

Para Carvalho e Paladini (2012), o cálculo do índice de capacidade deve ser realizado em duas oportunidades durante o projeto. O primeiro cálculo é executado na fase de medição do DMAIC. Nessa fase, o responsável seleciona uma ou mais características de Controle para Qualidade (*Control to Quality* – CTQ), mapeia o respectivo processo, realiza as medidas necessárias e a capacidade do processo a curto e longo prazos. Neste momento o objetivo é calcular a capacidade antes de intervir no processo. O cálculo da capacidade é refeito na fase de controle do DMAIC. O objetivo de reavaliar a capacidade do processo é verificar os ganhos obtidos com as melhorias implementadas pela equipe do projeto Seis Sigma, conforme ilustra a figura 13. Dependendo dos resultados dessa reavaliação, pode ser necessário rever uma ou mais fases precedentes do processo.

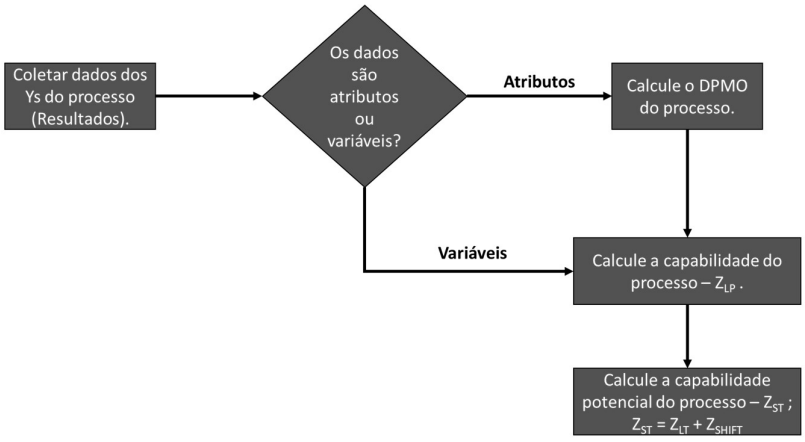


Figura 13 – Fluxograma do estudo de capacidade Seis Sigma.
Fonte: Baseado em Carvalho e Paladini (2012)

Carvalho e Paladini (2012) abordam que determinados autores tratam a medida de capacidade como z de avaliação comparativa (*benchmarking*), já que funciona como referencial de comparação da situação do processo antes e depois do projeto Seis Sigma, bem como entre processos em uma organização.

2.4 Teste de Shapiro Wilk

O teste de Shapiro-Wilk é aplicado para verificação da hipótese de normalidade da distribuição os dados, um dos requisitos para poder se calcular o índice de capacidade seis sigma.

Conforme Costa (2019), Shapiro-Wilk testa a hipótese nula da distribuição normal de uma amostra y_1, y_2, ⋯, y_n, retirada de uma população. Para calcular o valor da estatística W, dada a amostra aleatória, de tamanho n, é necessário seguir as etapas:

1. Obter uma amostra ordenada: $y_1 \leq y_2 \leq \cdots \leq y_n$;
2. Calcular a soma de quadrado dos desvios:

$$s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} \qquad (12)$$

3. Uma vez que se tem o número de amostras (n) coletadas, tem-se que:
   - se n é par, $n = 2k$, calcule:

$$b = \sum_{i=1}^{n} a_{n-i+1}(y_{n-i+1} - y_i) \qquad (13)$$

em que os valores de a_i são obtidos na tabela de coeficientes de Shapiro-Wilk.

   - se n é ímpar, $n = 2k+1$, o cálculo é exatamente como no item anterior, uma vez que $a_{k+1}=0$ quando $n = 2k+1$. Assim, determina-se:

$$b = a_n(y_n - y_1) + \cdots + a_{k+2}(y_{k+2} - y_k) \qquad (14)$$

em que o valor de y_{k+1}, que é a mediana, não entra no cálculo de b.

4. Calcule a estatística de teste:

$$W = \frac{b^2}{S^2} \qquad (15)$$

Para a tomada de decisão a respeito da normalidade dos dados, compara-se o valor calculado de W com o valor tabelado $W_{n;\alpha}$, obtido da Tabela de Probabilidades de Shapiro.

Se o valor calculado *W* for menor que o tabelado, rejeita-se a hipótese de normalidade ao nível α de significância, de acordo com Costa (2019).

Costa (2019) apresenta as hipóteses são como H_0:Os dados têm distribuição normal x H_1:Os dados não têm distribuição normal.

3 Metodologia

Este capítulo aborda, em detalhes, o fluxograma da percepção da qualidade. Vale ressaltar, mais uma vez, que esta dissertação engloba as técnicas, a seguir, ainda não implementadas unidas com foco na percepção da qualidade, esta forma aponta uma inovação no estudo da qualidade contribuindo para o estado da arte. Este fluxograma é o caminho a seguir para o estudo da qualidade para qualquer produto ou serviço. A figura 14, abaixo, é a representação do fluxograma:

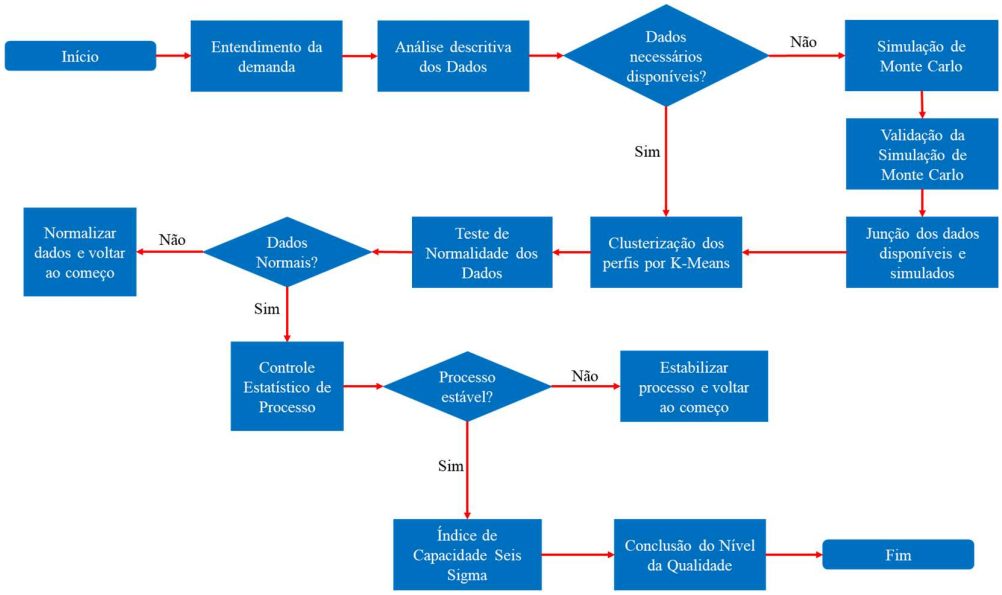


Figura 14 – Fluxograma da percepção da qualidade
Fonte: Elaboração própria.

É apresentada a abordagem de aplicação da Simulação de Monte Carlo através do amostrador de *Metropolis-Hastings* e floresta aleatória (*random forest*) para dados faltantes na base de dados, assim como usado por Herzog (2021) e Duque (2019), respectivamente. Dados estes importantes para a etapa de agrupamento, relatada através do método de K-Médias *(K-Means)*, assim como usado por Santana e Pontes (2020), Ghosn (2020) e Gavira-Durón, Gutierrez-Vargas e Cruz-Aké (2021).

Também é declarado o índice de capacidade Seis Sigma, assim como usado por Barbosa *et al.* (2020) e Silva *et al.* (2020). Porém, antes é importante dissertar sobre o teste de Shapiro-Wilk que verifica se os dados seguem uma distribuição normal, assim como usado por Costa (2019). Assim como, é realizada a verificação de um controle estatístico de processo estável. Dois requisitos necessários para a validação do cálculo do índice de capacidade Seis Sigma.

O programa (*software*) estatístico utilizado na implantação de todas as técnicas apresentadas nesta dissertação é o R versão 4.2.2.

3.1 Entendimento da demanda

O primeiro passo na percepção da qualidade em qualquer produto ou serviço é o correto entendimento da necessidade desta percepção. Compreender de forma correta a usabilidade do produto ou serviço junto ao demandante do estudo da percepção é de extrema importância. O correto compreendimento evitará que em alguma etapa do fluxograma se constate que não é a exigência necessária da percepção do produto ou serviço requisitada pelo demandante, ou seja, gerando retrabalho no estudo.

3.2 Análise descritiva dos Dados

Nesta etapa é importante analisar os dados disponibilizados pelo demandante para o estudo e assim conhecer melhor quais serão utilizadas ou não.

Um detalhamento minucioso de todo banco de dados será realizado através da análise de todas suas variáveis. Esta análise examina o número de registros do banco, tipo das variáveis (quantitativas ou qualitativas) assim como o que elas representam e se todas as variáveis contêm dados. Após esta importante análise serão selecionadas quais as corretas variáveis utilizadas no estudo, descartando as sem impacto – de acordo com o entendimento da demanda.

3.3 Verificação dos Dados Necessários Disponíveis

Caso exista a falta de dados em uma variável importante para o estudo, a Simulação de Monte Carlo é necessária. Do contrário, ou seja, todos os dados necessários estão disponíveis, a próxima etapa do fluxograma será o agrupamento dos perfis por K-Médias *(K-Means)*.

3.3.1 Simulação de Monte Carlo – SMC

Para a realizar a simulação de dados faltantes é utilizado dois métodos vertentes da Simulação de Monte Carlo, o Amostrador de Metropolis-Hastings, como Herzog

(2021) e Santos (2021) utilizaram, e o floresta aleatória (*random forest*), confirme apresentado por Karthe (2016) e Duque (2019).

Como ambos os métodos são renomados e de importante aplicabilidade a decisão da melhor simulação será aquela que obter os dados normalizados, ao término será analisada a melhor simulação para os dados.

3.3.1.1 Amostrador de Metropolis-Hastings

Utilizando o programa (*software*) estatístico R e importando a base de dados para dentro dele, Geyer e Johnson (2020) expõem a ideia para este método através da função "metrop()" dentro a biblioteca "mcmc". Este comando gera os dados faltantes.

3.3.1.2 Floresta Aleatória (*Random Forest*)

Também empregando o programa (*software*) estatístico R e a base de dados já importada para dentro dele, Zhao (2013), para aplicar o método de floresta aleatória (*random forest*) é necessário seguir os seguintes passos:

1. Carregar da biblioteca com o comando "library(randomForest)";
2. Treinar o modelo utilizando com o comando "randomForest()" informando a variável dependente e as independentes;
3. Recuperar uma das árvores geradas através do comando "getTree()";
4. Identificar os centros das classes, commando "classCenter()";
5. Combinar alguns modelos, ou seja, rodar com "randomForest()" outros modelos no intuito de criar uma matriz os combinando para auxiliar o próximo passo;
6. Predizer as classes da matriz citada no item anterior, com o comando "predicted";
7. Mostrando visualmente que floresta aleatória (*random forest*) é capaz de separar conjunto de dados que não são separáveis linearmente e assim ter o resultado a simulação.

3.3.2 Validação da Simulação de Monte Carlo

A validação será realizada comparando o resultado do Amostrador de Metropolis-Hastings e floresta aleatória (*random forest*). Esta validação se dará utilizando o Teste de Shapiro Wilk, segundo Costa (2019), no intuito de verificar a normalidade dos dados das simulações.

Caso as duas simulações apresentem normalidade dos dados, a escolha da simulação a ser utilizada será a que apresentar o melhor resultado no Teste de Shapiro Wilk.

3.3.3 Junção dos dados disponíveis e simulados

Após a validação e escolha da melhor simulação, os dados simulados são unidos aos dados disponíveis.

3.4 Agrupamento (*Cluster*) dos perfis por K-Médias *(K-Means)*

Também operando o programa (*software*) estatístico R e a base de dados já importada para dentro dele, Zhao (2013) apresenta o método de K-Médias *(K-Means)*. Expõe a ideia para este método através da função "kmeans()". Para esta função é necessário inserir o número de centróides, este número será testado até se chegar no melhor para o agrupamento.

Os perfis dos clientes serão agrupados neste momento no intuito de estudar cada perfil de forma personalizada.

3.5 Teste de Normalidade dos Dados

No intuito de facilitar os cálculos, temos a utilização do p-valor, para a interpretação do teste de hipótese. Quando o p-valor é menor que o nível α de significância rejeita-se a hipótese nula. Porém, quando o p-valor é maior que o nível α de significância não se rejeita a hipótese nula. Segundo Cano e Mogueza e Redchuk (2015), o comando "shapiro.test {stats}" no programa (*software*) estatístico R informa diretamente este p-valor.

Este teste é realizado em todos os dados de cada agrupamento (*cluster*) de perfil, pois somente com dados normalizados é possível realizar a etapa 3.6 Controle Estatístico de Processo.

3.5.1 Normalizar dados e voltar ao começo

Caso, através do Teste de Shapiro Wilk, é declarado que algum dado não seja normalizado então é necessário normalizar este dado e voltar todo o processo.

Segundo Bussab e Morettin (2017), o procedimento de Box-Cox automaticamente identifica uma transformação de uma família de transformações de potência Y para normalizar uma variável:

$$Y^* = Y^{\lambda} \qquad (16)$$

λ é um parâmetro a ser estimado a partir dos dados, exemplo:

$$\lambda = 2 \Rightarrow Y^* = Y^2$$

$$\lambda = 0{,}5 \Rightarrow Y^* = \sqrt{Y}$$

$$\lambda = 0 \Rightarrow Y^* = \ln Y$$

3.6 Controle Estatístico de Processo

É transmitido nas duas próximas sessões como realizar a verificação de um processo estatístico estável e não estável.

Vale ressaltar que não será realizada uma análise de *outliers* na base de dados pois estes prováveis *outliers* descobertos podem ser dados fora dos limites de análise de controle, portanto daria um viés no momento de estudo do controle.

3.6.1 Controle Estatístico de Processo Estável

Para apuração do processo é necessária a importação de uma biblioteca especial, a "qcc", que trabalha com o controle estatístico de processo, conforme Cano e Corcoba e Mogueza (2015) explicam em seu trabalho.

Com o comando "qcc()" é possível ao mesmo tempo obter o gráfico de Shewhart, a linha central, os limites superior e inferior.

Se os dados estiverem dentro dos limites superior e inferior, não exibem um padrão no gráfico (dados aleatorizados) e apresentarem normalidade este processo está sob controle, ou seja, é estável.

3.6.2 Controle Estatístico de Processo Não Estável

Se os dados estiverem não estiverem dentro dos limites superior e inferior ou exibem um padrão no gráfico (dados não aleatorizados) e apresentarem não normalidade este processo não está sob controle, ou seja, não é estável.

3.6.2.1 Estabilizar processo e voltar ao começo

Caso o processo seja não estável é imprescindível que o cliente reveja o processo do produto ou serviço, o modelo mais indicado é o DMAIC, já mencionado nesta dissertação.

Refeito o processo então o cliente coleta novamente os dados, volta no começo do fluxograma e testa se o processo se tornou estável para finalmente prosseguir.

3.7 Índice de Capacidade Seis Sigma

Cano e Mogueza e Redchuk (2015) expressam o índice de capacidade seis sigma através do comando “process.capability()”.

O Cpk apresentado dará o norte para análise do nível da qualidade.

3.8 Conclusão do Nível da Qualidade

Através do Cpk é possível averiguar o nível da qualidade e concluir se o produto ou serviço demonstra qualidade para cada perfil de público.

4 Resultados da Qualidade do Serviço – Seguradora de Vida

No intuito de validar o fluxograma utilizando dados reais, uma empresa de seguro de vida e previdência concedeu, de forma espontânea, suas informações. Esta seguradora é especializada em vida e previdência com mais de 185 anos de atuação ininterrupta no Brasil.

O conjunto de dados concedido se refere a pesquisa de recomendação para os produtos referentes ao seguro de vida. Ao se perguntar a recomendação para o cliente é possível medir o grau de satisfação. No final de cada ligação para a central de atendimento do cliente o atendente pergunta "Numa escala de 0 a 10, qual é a probabilidade de você recomendar a 'nome da seguradora' a um amigo ou colega?".

Uma metodologia já conhecida e muito utilizada no mercado de trabalho para o tipo de pergunta realizada é a Pontuação de Chances de Promoção (*Net Promoter Score* – NPS), porém a seguradora quer um estudo mais aprofundado sobre a qualidade de seu produto utilizando dados dos clientes junto com a resposta fornecida pela pergunta de recomendação.

4.1 Entendimento da demanda

A seguradora quer medir a qualidade do produto de seguro de vida oferecido, desde a informação requisitada pelo cliente até a solicitação do beneficiário do seguro, ou seja, averiguar a jornada do cliente.

Caso o indicador de qualidade seja bom, a seguradora manterá todo formato atual. Caso contrário, ou seja, apresente uma performance ruim na qualidade, tomará novas estratégias na venda ou melhorará a jornada do cliente.

4.2 Análise descritiva dos Dados

O banco de dados disponibilizado abrange pesquisas de recomendação somente do produto de seguro de vida no período de todos os meses de 2021 com 2.667 registros únicos e 9 variáveis. Foram enviadas as variáveis que impactam no cálculo atuarial do valor do sinistro, valor este que é pago ao beneficiário quando o cliente morre:

- **CPF** – variável qualitativa, utilizada na verificação de não duplicidade no banco do cliente.

- **Resposta** – variável quantitativa, valor respondida na pergunta de recomendação pelo cliente:
    - Mínimo: 0;
    - Máximo: 10;
- **Idade** – variável quantitativa, idade do cliente no momento do contato com a central de atendimento:
    - Mínimo: 19;
    - Máximo: 97;
- **Sexo** – variável qualitativa, sexo do cliente:
    - Masculino;
    - Feminino;
- **Estado Civil** – variável qualitativa, estado civil do cliente:
    - Casado;
    - Divorciado;
    - Separado Judicialmente;
    - Solteiro;
    - União Estável;
    - Viúvo;
- **Escolaridade** – variável qualitativa, escolaridade do cliente:
    - Fundamental (1º Grau) Completo;
    - Médio (2º Grau) Completo;
    - Superior Completo;
    - Especialização/Residência;
- **Cobertura contratada** – variável qualitativa, cobertura do seguro de vida contratada pelo cliente:
    - Morte;
    - Morte Acidental;
- **Motivo Ligação** – variável qualitativa, motivo da ligação do cliente:
    - Informação;
    - Reclamação;
    - Reclamação Grave;
    - Reclamação Gravíssima;
    - Reclamação Simples;
    - Solicitação.

4.3 Verificação dos Dados Necessários Disponíveis

Na etapa anterior foram verificadas que as variáveis Estado Civil e Escolaridade contêm dados faltantes. Sendo assim, a Simulação de Monte Carlo é necessária na simulação destes dados faltantes.

4.3.1 Simulação de Monte Carlo – SMC

Para a realizar a simulação de dados faltantes das variáveis Estado Civil e Escolaridade é utilizado dois métodos vertentes da Simulação de Monte Carlo, o Amostrador de Metropolis-Hastings e o modelo floresta aleatória (*random forest*).

4.3.1.1 Amostrador de Metropolis-Hastings

Utilizando o programa (*software*) estatístico R, conforme descrito na metodologia exposta na seção 3.3.1.1, verifica-se que não é possível rodar uma simulação dos dados faltantes pois não apresentam uma cadeia de Markov, impossibilitando qualquer cálculo.

4.3.1.2 Floresta Aleatória (*Random Forest*)

De acordo com a metodologia delineada na seção 3.3.1.2, foram obtidos os dados incompletos referentes ao Estado Civil e à Escolaridade por meio do programa (*software*) estatístico R.

4.3.2 Validação da Simulação de Monte Carlo

O Amostrador de Metropolis-Hastings demonstrou incapacidade em fornecer os dados faltantes, enquanto o método de floresta aleatória (*random forest*) foi bem-sucedido em obter informações incompletas relacionadas ao Estado Civil e à Escolaridade.

Os dados ausentes, gerados pela técnica de floresta aleatória (*random forest*), foram submetidos ao teste de Shapiro-Wilk para avaliar a normalidade da distribuição. Para o Estado Civil, o p-valor foi calculado como 0,062839, indicando que, a um nível de significância de α igual a 0,05, não se rejeita a hipótese nula, sugerindo que os dados apresentam uma distribuição normal. No que diz respeito à Escolaridade, o p-valor foi calculado como 0,1359, mostrando que, a um nível de significância de α igual a 0,05, não se rejeita a hipótese nula, indicando uma distribuição normal.

Pode-se inferir, portanto, que os dados faltantes gerados por meio da técnica de floresta aleatória (*random forest*) para Estado Civil e Escolaridade são considerados válidos e podem ser integrados ao banco de dados existente.

4.3.3 Junção dos dados disponíveis e simulados

Após a verificação conduzida no item 4.3.2, os dados sintetizados utilizando a técnica de floresta aleatória (*random forest*) para Estado Civil e Escolaridade foram devidamente incorporados ao banco de dados existente. Como resultado, obteve-se um

banco de dados livre de lacunas, preparado para dar continuidade ao fluxograma da percepção da qualidade de serviço apresentado nesta dissertação.

4.4 Agrupamento (*Cluster*) dos perfis por K-Médias *(K-Means)*

A partir do banco de dados completo, foram examinados diferentes números de centróides, variando de 2 a 15. Após uma análise descritiva dos dados e a avaliação gráfica conforme representada na figura 15 abaixo, foi identificado que o número ótimo de agrupamentos ocorre com um centróide igual a 14. Em outras palavras, os clientes são separados em 14 agrupamentos (*clusters*) distintos de acordo com seus perfis.

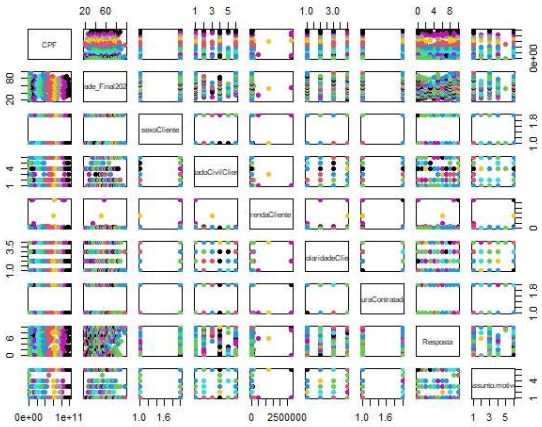


Figura 15 – Agrupamentos (*Clusters*) dos perfis de clientes da seguradora
Fonte: Elaboração própria.

A distribuição do tamanho de cada agrupamento (*cluster*) é descrita na tabela 3:

Tabela 3 – Quantidade de clientes por agrupamento (*cluster*).
Fonte: Elaboração própria.

Agrupamento (*Cluster*)	Quantidade de Clientes
1	125
2	133
3	196
4	301
5	130
6	205
7	145
8	156
9	259
10	142
11	384
12	197
13	124
14	170

4.5 Teste de Normalidade dos Dados

Utilizando o teste de Shapiro-Wilk e considerando um nível de significância de α igual a 0,05, todas as variáveis em todos os agrupamentos (*clusters*) demonstraram um p-valor superior a 0,05. Esse resultado sugere a não rejeição da hipótese nula, indicando que todas os dados exibem uma distribuição normal.

4.6 Controle Estatístico de Processo

Com base na nota de recomendação em cada agrupamento, todos os gráficos de controle demonstram um padrão estável de Controle Estatístico de Processo, já que todos os dados estão contidos dentro dos limites superior e inferior, além de exibirem um padrão de aleatoriedade. Abaixo os gráficos de controle estatístico de processo dos agrupamentos (*clusters*) 1, 6 e 10, para demonstração do resultado:

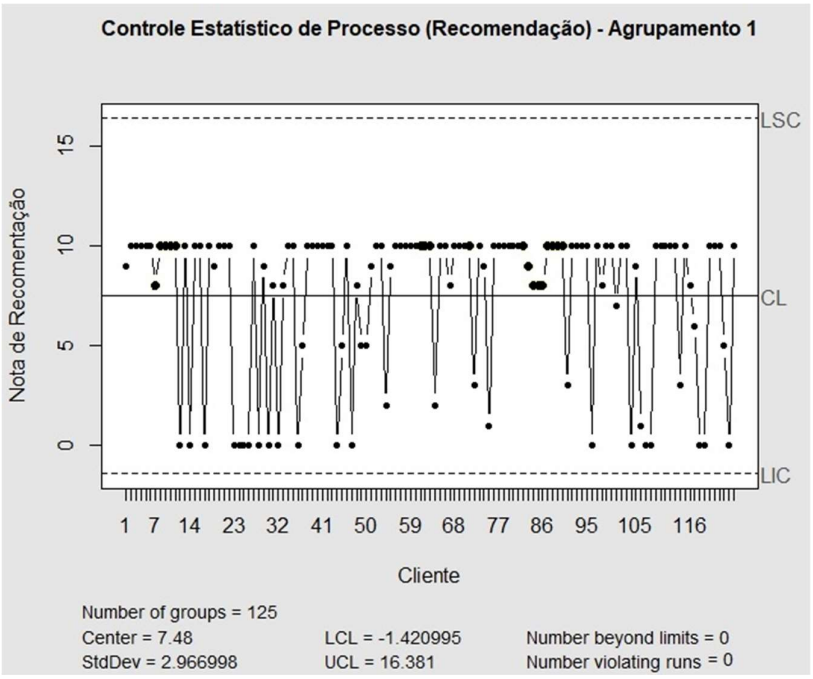


Figura 16 – Gráfico de Controle Estatístico de Processo do Agrupamento (*Cluster*) 1
Fonte: Elaboração própria.

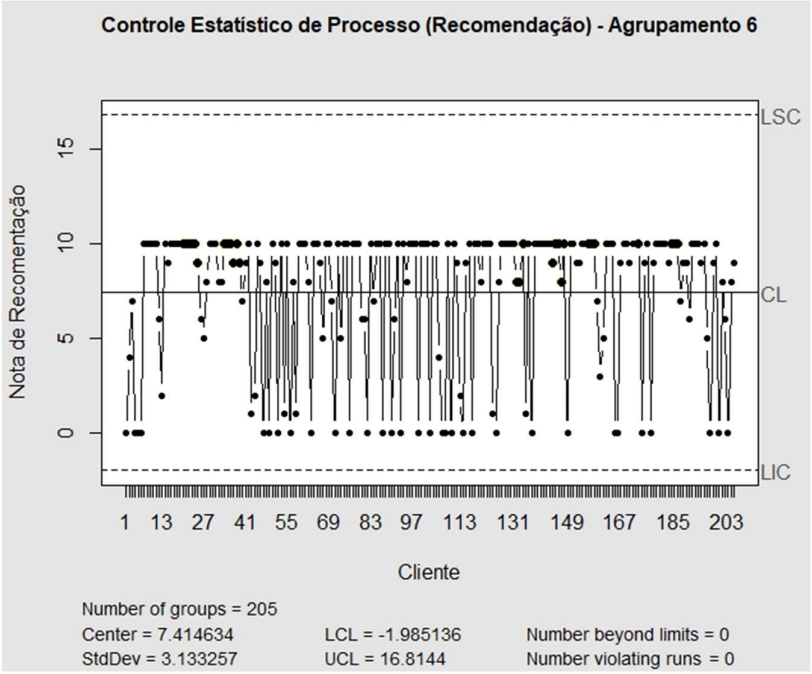


Figura 17 – Gráfico de Controle Estatístico de Processo do Agrupamento (*Cluster*) 6
Fonte: Elaboração própria.

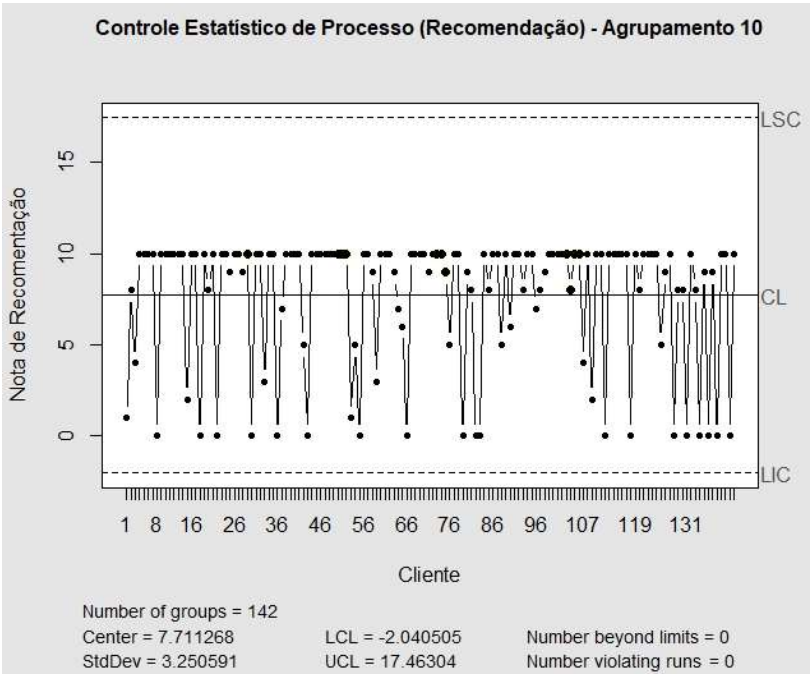


Figura 18 – Gráfico de Controle Estatístico de Processo do Agrupamento (*Cluster*) 10
Fonte: Elaboração própria.

As notas de satisfação de cada agrupamento (*cluster*) foram submetidas ao teste de Shapiro-Wilk. Em todas as instâncias de teste, ao considerar um nível de significância

de α igual a 0,05, os p-valores obtidos foram superiores ao nível de significância estipulado. Em outras palavras, a não rejeição da hipótese nula foi observada em todos os testes. Assim, é possível inferir que todas as notas de satisfação nos diferentes agrupamentos (*clusters*) apresentam uma distribuição normal.

Tabela 4 – p-valor do Teste de Shapiro-Wilk para cada nota de recomendação em nos 14 agrupamentos.
Fonte: Elaboração própria.

Agrupamento (*Cluster*)	p-valor
1	0,972287
2	0,281080
3	0,817159
4	0,261196
5	0,404254
6	0,478935
7	0,123087
8	0,163119
9	0,322843
10	0,132057
11	0,804338
12	0,595956
13	0,674232
14	0,374774

Estes processos podem ter o cálculo realizado para o índice de capacidade seis sigma de cada agrupamento (*cluster*), pois apresenta um processo de controle estável e dados distribuídos normalmente em cada.

Vale ressaltar que não foi realizada a análise *outliers* na base de dados pois estes prováveis *outliers* descobertos podem ser dados fora dos limites, portanto daria um viés no momento de estudo do controle.

4.7 Índice de Capacidade Seis Sigma

Todos os agrupamentos identificados atendem aos critérios exigidos para o cálculo do índice de capacidade Seis Sigma.

Abaixo estão as saídas do comando conforme delineado na seção 3.7 para os agrupamentos (*clusters*) 1, 6 e 10, apresentando uma ilustração dos resultados:

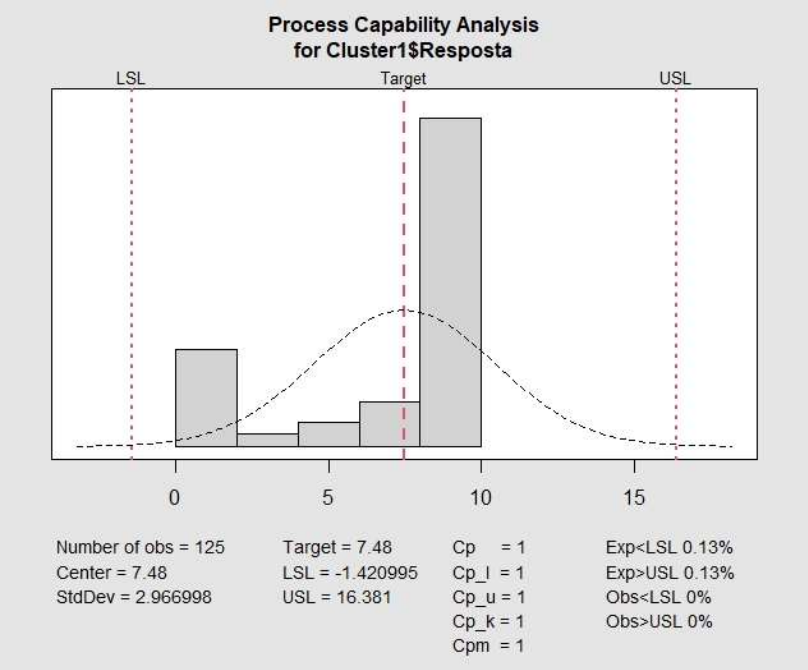


Figura 19 – Índice de Capacidade Seis Sigma - Recomendação Agrupamento (*Cluster*) 1
Fonte: Elaboração própria.

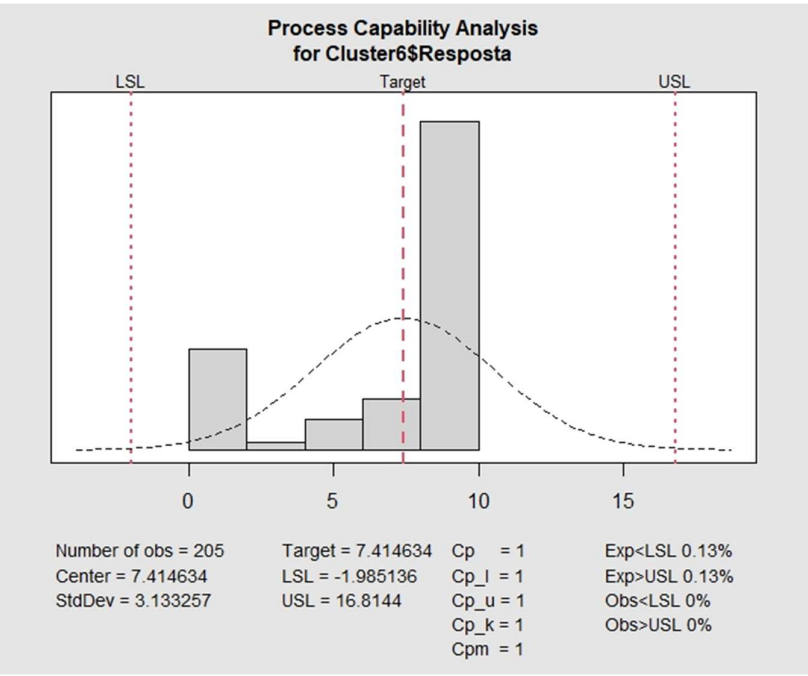


Figura 20 – Índice de Capacidade Seis Sigma - Recomendação Agrupamento (*Cluster*) 6
Fonte: Elaboração própria.

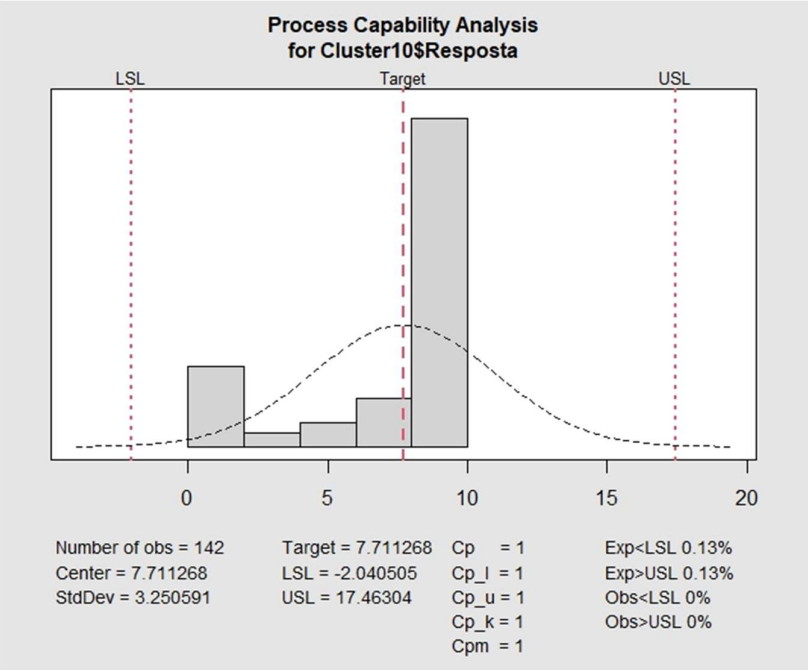


Figura 21 – Índice de Capacidade Seis Sigma - Recomendação Agrupamento (*Cluster*) 10
Fonte: Elaboração própria.

4.8 Conclusão do Nível da Qualidade

Os valores de Cpk para os 14 índices de capacidade Seis Sigma são todos iguais ou superiores a 1. Portanto, é factível afirmar que a variável "Satisfação do Cliente" possui uma taxa de 3,4 defeitos por milhão (ppm), indicando um elevado padrão de qualidade em todos os perfis dos clientes agrupados.

5 Considerações Finais

Esta dissertação discorre e corrobora a validação de um fluxograma destinado à identificação da percepção de qualidade pelos clientes, aplicável a uma ampla gama de produtos ou serviços, com a capacidade de segmentação baseada nos diversos perfis de clientes. Sua concepção abarca a implementação de técnicas até então não aplicadas, enfocando especificamente a compreensão da percepção de qualidade, representando, assim, uma contribuição inovadora para o domínio do estudo da qualidade, contribuindo significativamente para o estado da arte neste campo.

Cada objetivo delineado para o desenvolvimento deste projeto foi integralmente cumprido:

1. A fim de estimar e validar a consistência dos dados ausentes, foram empregadas técnicas de Simulação de Monte Carlo através do Amostrador de Metropolis-Hastings e da Floresta Aleatória (*Random Forest*);
2. Os perfis dos clientes foram agrupados com base nas informações existentes e na simulação dos dados ausentes, utilizando a técnica de K-Médias (*K-Means*);
3. A compreensão da qualidade e da satisfação relacionadas aos produtos adquiridos e/ou serviços oferecidos nos diferentes agrupamentos foi conduzida por meio do método Seis Sigma.

Esta dissertação assegura que a utilização do fluxograma proposto poderá direcionar uma empresa na compreensão da concepção de qualidade percebida pelos clientes, independentemente do tipo de produto ou serviço ofertado, permitindo a segmentação com base nos perfis dos clientes e, consequentemente, possibilitando o estudo personalizado de cada perfil em particular.

Como parte dos meus planos futuros de estudo, pretendo empregar o fluxograma em diversas áreas, incluindo a linha de montagem de peças de motocicletas. Atualmente, encontro-me em processo de negociação para obter os dados necessários junto a uma empresa indiana do setor de motocicletas.

Referências Bibliográficas

ALAMSYAH, A., NURRIZ, B., "Monte Carlo Simulation and Clustering for Customer Segmentation in Business Organization". *3rd International Conference on Science and Technology – Computer (ICST)*, Yogyakarta, Indonesia, July 2017.

APARECIDA, Danieli; REZENDE, Denis Alcides. **Modelo de prestação de serviços públicos municipais conectados por meio da internet das coisas no contexto da cidade digital estratégica**, Revista Contribuciones a las Ciencias Sociales, (Vol 1, Nº 6 junio 2021, pp. 15-28).

BARBOSA, F. A.; HERNÁNDEZ VERGARA, W. R.; YAMANARI, J. S.; SANTOS, K. B. Proposição de um modelo para aprimoramento do sistema de gestão da qualidade. Sistemas & Gestão, [S. l.], v. 14, n. 4, p. 435–447, 2020.

BATALHA, M. O., **Introdução à Engenharia de Produção**. 1 ed. Rio de Janeiro, Elsevier, 2008.

BUSSAB, W.O.; MORETTIN, P.A. Estatística Básica. São Paulo: Editora Saraiva, 2017 (9ª Edição).

CANO, Emilio L.; CORCOBA, Mariano Prieto; MOGUERZA, Javier M., **Quality Control with R An ISO Standards Approach**. Madrid, Springer, 2015.

CANO, Emilio L.; MOGUERZA, Javier M.; REDCHUK, Andrés, **Six Sigma with R Statistical Engineering for Process Improvement**. Madrid, Springer, 2012.

CARROLL, Raymond J., LIANG, Faming, LIU, **Chuanhai, Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples**, Chichester, John Wiley & Sons, 2010.

CARVALHO, M. M., PALADINI, E. P., **Gestão da qualidade. Teoria e casos.** 2 ed. Rio de Janeiro, Elsevier, 2012.

CHEN, Yen-Chi, **Monte Carlo Simultions and Bootstrap**, Washington, University of Washington, 2017.

CHIROLI, D.; LUIZ, L.; DONIN, M.; TYBUSZEUSKY, J. PROPOSTA DE MELHORIA BASEADA NA METODOLOGIA DMAIC EM UMA UNIDADE DE PRONTO ATENDIMENTO DE SAÚDE. The Journal of Engineering and Exact Sciences, Viçosa/MG, BR, v. 6, n. 1, p. 0029–0035, 2020

COSTA, Silvano Cesar da. *Teste de Shapiro-Wilk*. Disponível em: http://www.uel.br/projetos/experimental/pages/arquivos/Shapiro.html . Acesso em: 20 nov. 2021, 22:45:31. 2019.

DEAN, J. W., BOWEN, D. E., "Management Theory and Total Quality: Improving Research and Practice through Theory Development", **The Academy of Management Review** v. 19, no. 3, p. 392-418, 1994.

DOS SANTOS, Sergio Rodrigo Quadros; CUNHA, Ana Paula Martins do Amaral; RIBEIRO-NETO, Germano Gondim. AVALIAÇÃO DE DADOS DE PRECIPITAÇÃO PARA O MONITORAMENTO DO PADRÃO ESPAÇO-TEMPORAL DA SECA NO NORDESTE DO BRASIL. Revista Brasileira de Climatologia, [S.l.], v. 25, aug. 2019.

DUQUE, Maria Clara Machado de Almeida, *Validação De Testes De Produção De Poços De Petróleo Baseada Em Mineração De Dados*. Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2019.

FERREIRA, Denise Lilian Luz; SEIFERT, Amanda Antunes; VENANZI, Délvio. **Conectividade de processos na supply chain via tecnologias da internet das coisas (IOT) e softwares na empresa ABC.** South American Development Society Journal, [S.l.], v. 6, n. 16, p. 1, abr. 2020.

FONSECA, L.; SANTOS, A.; FERREIRA, L.; REIS, A.; PIZETTA, L. (2020). Aplicação integrada do controle estatístico de processo e engenharia de métodos em uma indústria alimentícia. Exacta, 18(1), 244-268.

GARVIN, D. A., "What does product quality really mean?", **Sloan Management Review**, p. 25-43, 1984.

GAVIRA-DURÓN, Nora; GUTIERREZ-VARGAS, Octavio; CRUZ-AKÉ, Salvador. 2021. "Markov Chain K-Means Cluster Models and Their Use for Companies' Credit Quality and Default Probability Estimation" Mathematics 9, no. 8: 879.

GEYER, Charles J., JOHNSON, Leif T., **Package 'mcmc'**. Massachusetts, MIT, 2020.

GHOSN, Luciana Habib Abi, *Modelo De Ordenação De Mercados Para Internacionalização De Pequenas E Médias Empresas Brasileiras Da Indústria De Alimentos*. Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2020.

GOMES, Fabrício Maciel, **Modelagem e Simulação de Sistemas**, São Paulo, USP, 2018.

HAMMERSLEY, J. M., HANDSCOMB, D. C., **Monte Carlo Methods**. 1ed. Londres, Methuen, 1964.

HARRY, M.J. "Six Sigma: A breakthrough Strategy for Profitability". Quality Progress, p. 60-64, maio, 1998.

HERZOG, Iasmin Louzada, *Uso De Redes Neurais Evolutivas Na Solução De Problemas Inversos De Transientes Hidráulicos*. Dissertação de M.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 2021.

KARTHE. *Tree Based Algorithms: A Complete Tutorial from Scratch (in R & Python)*. Disponível em: https://www.analyticsvidhya.com/blog/2016/04/tree-based-algorithms-complete-tutorial-scratch-in-python/ . Acesso em: 24 jul. 2021, 19:56:21. 2016.

LOPES, R. C.; MENDES, A. C. A.; LUNKES, R. J.; COSTA, G. D. Utilização da simulação de Monte Carlo na gestão de estoques para empresas farmacêuticas. REVISTA AMBIENTE CONTÁBIL - Universidade Federal do Rio Grande do Norte - ISSN 2176-9036, [S. l.], v. 11, n. 2, p. 1–18, 2019.

LUI, M. L. C.; PETARNELLA, L. **As cidades inteligentes e os desafios para a implantação da garantia da qualidade de serviços**. R. Tecnol. Soc., Curitiba, v. 16, n. 39, p. 182-198, jan/mar. 2020.

METROPOLIS, Nicholas. ULAM, Stanislaw. The Monte Carlo Method. Journal of the American Statitiscal Association. v. 44, n. 247, Sep. 1949, p. 335-41.

MINGOTI, S. A., **Análise de dados através de estatística multivariada: uma abordagem aplicada.** 1 ed. Belo Horizonte, Editora UFMG, 2007.

MONTGOMERY, Douglas C., **Introduction to Statistical Quality Control**, 6 ed. Arizona, John Wiley & Sons, 2009.

PAVANI, Ana Maria Beltran. Sistema Maxwell. PUC-Rio – Certificação Digital. Disponível em: https://www.maxwell.vrac.puc-rio.br/ . Acesso em: 29 jun. 2021, 11:08:24.

POSSANI, R. G., *Re-Engenharia Do Software Scms Para Uma Linguagem Orientada A Objetos (Java) Para Uso Em Construções De Phantoms Segmentados*. Dissertação de M.Sc., IPEN, São Paulo, SP, Brasil, 2012.

RAMOS, M.; BRANDÃO, A. L.; GRAEVER, L. .; CAMPOS, C. E. A. **Melhoria contínua da qualidade: uma análise pela perspectiva dos profissionais das equipes de atenção primária à saúde do município do Rio de Janeiro.** Revista Brasileira de Medicina de Família e Comunidade, Rio de Janeiro, v. 16, n. 43, p. 2736, 2021.

SANTANA, Nathaly Silva de *et al*. CONTROLE ESTATÍSTICO DA QUALIDADE: UMA APLICAÇÃO EM UMA INDÚSTRIA TÊXTIL. Revista Latino-Americana de Inovação e Engenharia de Produção, [S.l.], v. 7, n. 12, p. 47-56, dez. 2019.

SANTANA, Roniel Venâncio; PONTES, Heráclito Lopes Jaguaribe. Aplicação da Clusterização por K-means para Criação de Sistema de Recomendação de Produtos baseado em Perfis de Compra. Navus - Revista de Gestão e Tecnologia, [S.l.], v. 10, p. 01-14, july 2020. ISSN 2237-4558.

SANTOS, Débora Cristine dos, *Amostrador de Gibbs aproximado usando Computação Bayesiana Aproximada e regressão quantílica via redes neurais artificiais*. Dissertação de M.Sc., EST/UnB, Rio de Janeiro, RJ, Brasil, 2021.

SILVA, M. M.; CAMPAROTTI, C. E. S.; ENAMI, L. M.; GUEDES, K.; REIS, B. L.; ORDENO, T. de S. B. Aplicação da metodologia seis sigma para melhoria contínua da qualidade em uma indústria alimentícia. Revista Produção Online, [S. l.], v. 20, n. 2, p. 546–574, 2020.

SLACK, N., CHAMBERS, S., HARLAND, C., HARRISON, A., JOHNSTON, R., **Administração da produção Edição Compacta**, São Paulo, Atlas, 2006.

SORIANO, Fabiano Rodrigues; OPRIME, Pedro Carlos; LIZARELLI, Fabiane Letícia. Os fatores que devem ser considerados para uma efetiva implantação do controle estatístico de processo (CEP): uma revisão de literatura. Revista Gestão da Produção Operações e Sistemas, [S.l.], v. 15, n. 1, p. 71, mar. 2020.

YE, Leihua. *A Practical Guide to Bootstrap in R*. Disponível em: https://towardsdatascience.com/a-practical-guide-to-bootstrap-with-r-examples-bd975ec6dcea . Acesso em: 06 jul. 2021, 22:18:22. 2020.

ZHAO, Yanchang, **R and Data Mining: Examples and Case Studies**. Amsterdam, Elsevier, 2013.

Apêndice – Códigos no Programa *(Software)* Estatístico R

```
################################
## Diretório da Base de Dados ##
################################
setwd("G:\\DIRETÓRIO LOCAL\\Banco de dados")
getwd()
```

```
###################
## Base de Dados ##
###################
library(openxlsx)
Base = openxlsx::read.xlsx("Base Pesquisa Satisfação.xlsx" ,
                           sheet = "Base Final (5)")
```

```
########################
## Análises de Dados ##
########################
dim(Base)
str(Base)
summary(Base)
cbind(table(Base$sexoCliente, useNA = "always"))
cbind(table(Base$estadoCivilCliente, useNA = "always"))
cbind(table(Base$escolaridadeCliente, useNA = "always"))
cbind(table(Base$coberturaContratada_v02, useNA = "always"))
cbind(table(Base$Assunto.motivo, useNA = "always"))
```

```
########################################
## Amostrador de Metropolis-Hastings ##
########################################
library(mcmc)
mcmc::metrop(Base$estadoCivilCliente)
mcmc::metrop(Base$escolaridadeCliente)
```

```
########################################
## Floresta Aleatória (Random Forest) ##
########################################
library(randomForest)

RFestadoCivilCliente = randomForest(Resposta ~ Base$estadoCivilCliente ,
                    data = Base , prox = TRUE, na.action=na.fail)
RFestadoCivilCliente
summary(RFestadoCivilCliente)
getTree(RFestadoCivilCliente)
classCenter(RFestadoCivilCliente)
RFestadoCivilCliente$predicted
RFestadoCivilCliente$forest

RFescolaridadeCliente = randomForest(Resposta ~ Base$escolaridadeCliente ,
                          data = Base , prox = TRUE, na.action=na.fail)
RFescolaridadeCliente
summary(RFescolaridadeCliente)
getTree(RFescolaridadeCliente)
classCenter(RFescolaridadeCliente)
RFescolaridadeCliente$predicted
RFescolaridadeCliente$forest




###########################################
## Validação da Simulação de Monte Carlo ##
###########################################
shapiro.test(RFestadoCivilCliente$predicted)
shapiro.test(RFescolaridadeCliente$predicted)




#################################################################
## Agrupamento (Cluster) dos perfis por K-Médias (K-Means) ##
#################################################################
(clusters1 = kmeans(Base , centers = 2))
clusters1$size
plot(Base , col = clusters1$cluster, pch = 19)

(clusters2 = kmeans(Base , centers = 3))
clusters2$size
plot(Base , col = clusters2$cluster, pch = 19)

(clusters3 = kmeans(Base , centers = 4))
```

```
clusters3$size
plot(Base , col = clusters3$cluster, pch = 19)

(clusters4 = kmeans(Base , centers = 5))
clusters4$size
plot(Base , col = clusters4$cluster, pch = 19)

(clusters5 = kmeans(Base , centers = 6))
clusters5$size
plot(Base , col = clusters5$cluster, pch = 19)

(clusters6 = kmeans(Base , centers = 7))
clusters6$size
plot(Base , col = clusters6$cluster, pch = 19)

(clusters7 = kmeans(Base , centers = 8))
clusters7$size
plot(Base , col = clusters7$cluster, pch = 19)

(clusters8 = kmeans(Base , centers = 9))
clusters8$size
plot(Base , col = clusters8$cluster, pch = 19)

(clusters9 = kmeans(Base , centers = 10))
clusters9$size
plot(Base , col = clusters9$cluster, pch = 19)

(clusters10 = kmeans(Base , centers = 11))
clusters10$size
plot(Base , col = clusters10$cluster, pch = 19)

(clusters11 = kmeans(Base , centers = 12))
clusters11$size
plot(Base , col = clusters11$cluster, pch = 19)

(clusters12 = kmeans(Base , centers = 13))
clusters12$size
plot(Base , col = clusters12$cluster, pch = 19)

(clusters13 = kmeans(Base , centers = 14))
clusters13$size
plot(Base , col = clusters13$cluster, pch = 19)

(clusters14 = kmeans(Base , centers = 15))
clusters14$size
plot(Base , col = clusters14$cluster, pch = 19)



BaseComCluster = data.frame(Base , clusters13$cluster)
```

```
str(BaseComCluster)

library(dplyr)
Cluster1 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==1)
str(Cluster1)
shapiro.test(Cluster1$Resposta)$p.value

Cluster2 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==2)
str(Cluster2)
shapiro.test(Cluster2$Resposta)$p.value

Cluster3 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==3)
str(Cluster3)
shapiro.test(Cluster3$Resposta)$p.value

Cluster4 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==4)
str(Cluster4)
shapiro.test(Cluster4$Resposta)$p.value

Cluster5 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==5)
str(Cluster5)
shapiro.test(Cluster5$Resposta)$p.value

Cluster6 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==6)
str(Cluster6)
shapiro.test(Cluster6$Resposta)$p.value

Cluster7 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==7)
str(Cluster7)
shapiro.test(Cluster7$Resposta)$p.value

Cluster8 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==8)
str(Cluster8)
shapiro.test(Cluster8$Resposta)$p.value

Cluster9 = BaseComCluster %>% dplyr::filter(BaseComCluster$clusters13.cluster==9)
str(Cluster9)
shapiro.test(Cluster9$Resposta)$p.value

Cluster10 = BaseComCluster %>%
dplyr::filter(BaseComCluster$clusters13.cluster==10)
str(Cluster10)
shapiro.test(Cluster10$Resposta)$p.value

Cluster11 = BaseComCluster %>%
dplyr::filter(BaseComCluster$clusters13.cluster==11)
str(Cluster11)
shapiro.test(Cluster11$Resposta)$p.value
```

```
Cluster12 = BaseComCluster %>%
dplyr::filter(BaseComCluster$clusters13.cluster==12)
str(Cluster12)
shapiro.test(Cluster12$Resposta)$p.value

Cluster13 = BaseComCluster %>%
dplyr::filter(BaseComCluster$clusters13.cluster==13)
str(Cluster13)
shapiro.test(Cluster13$Resposta)$p.value

Cluster14 = BaseComCluster %>%
dplyr::filter(BaseComCluster$clusters13.cluster==14)
str(Cluster14)
shapiro.test(Cluster14$Resposta)$p.value




########################################
## Controle Estatístico de Processo ##
########################################
library(qcc)
CEPcluster1 = qcc::qcc(data = Cluster1$Resposta, type = "xbar.one" , plot = TRUE,
            title = "Controle Estatístico de Processo (Recomendação) - Agrupamento 1",
            xlab = "Cliente" , ylab = "Nota de Recomentação",
            label.limits = c("LIC", "LSC"))
CEPcluster2 = qcc::qcc(data = Cluster2$Resposta, type = "xbar.one" , plot = TRUE,
            title = "Controle Estatístico de Processo (Recomendação) - Agrupamento 2",
            xlab = "Cliente" , ylab = "Nota de Recomentação",
            label.limits = c("LIC", "LSC"))
CEPcluster3 = qcc::qcc(data = Cluster3$Resposta, type = "xbar.one" , plot = TRUE,
            title = "Controle Estatístico de Processo (Recomendação) - Agrupamento 3",
            xlab = "Cliente" , ylab = "Nota de Recomentação",
            label.limits = c("LIC", "LSC"))
CEPcluster4 = qcc::qcc(data = Cluster4$Resposta, type = "xbar.one" , plot = TRUE,
                title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
4",
                xlab = "Cliente" , ylab = "Nota de Recomentação",
                label.limits = c("LIC", "LSC"))
CEPcluster5 = qcc::qcc(data = Cluster5$Resposta, type = "xbar.one" , plot = TRUE,
                title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
5",
                xlab = "Cliente" , ylab = "Nota de Recomentação",
                label.limits = c("LIC", "LSC"))
CEPcluster6 = qcc::qcc(data = Cluster6$Resposta, type = "xbar.one" , plot = TRUE,
                title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
6",
                xlab = "Cliente" , ylab = "Nota de Recomentação",
                label.limits = c("LIC", "LSC"))
CEPcluster7 = qcc::qcc(data = Cluster7$Resposta, type = "xbar.one" , plot = TRUE,
```

```
                title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
7",
                xlab = "Cliente" , ylab = "Nota de Recomentação",
                label.limits = c("LIC", "LSC"))
CEPcluster8 = qcc::qcc(data = Cluster8$Resposta, type = "xbar.one" , plot = TRUE,
                title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
8",
                xlab = "Cliente" , ylab = "Nota de Recomentação",
                label.limits = c("LIC", "LSC"))
CEPcluster9 = qcc::qcc(data = Cluster9$Resposta, type = "xbar.one" , plot = TRUE,
                title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
9",
                xlab = "Cliente" , ylab = "Nota de Recomentação",
                label.limits = c("LIC", "LSC"))
CEPcluster10 = qcc::qcc(data = Cluster10$Resposta, type = "xbar.one" , plot = TRUE,
                 title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
10",
                 xlab = "Cliente" , ylab = "Nota de Recomentação",
                 label.limits = c("LIC", "LSC"))
CEPcluster11 = qcc::qcc(data = Cluster11$Resposta, type = "xbar.one" , plot = TRUE,
                 title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
11",
                 xlab = "Cliente" , ylab = "Nota de Recomentação",
                 label.limits = c("LIC", "LSC"))
CEPcluster12 = qcc::qcc(data = Cluster12$Resposta, type = "xbar.one" , plot = TRUE,
                 title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
12",
                 xlab = "Cliente" , ylab = "Nota de Recomentação",
                 label.limits = c("LIC", "LSC"))
CEPcluster13 = qcc::qcc(data = Cluster13$Resposta, type = "xbar.one" , plot = TRUE,
                 title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
13",
                 xlab = "Cliente" , ylab = "Nota de Recomentação",
                 label.limits = c("LIC", "LSC"))
CEPcluster14 = qcc::qcc(data = Cluster14$Resposta, type = "xbar.one" , plot = TRUE,
                 title = "Controle Estatístico de Processo (Recomendação) - Agrupamento
14",
                 xlab = "Cliente" , ylab = "Nota de Recomentação",
                 label.limits = c("LIC", "LSC"))




#####################################
## Índice de Capacidade Seis-Sigma ##
#####################################
IndCapSeisSig_Cluster1  = qcc:process.capability(object = CEPcluster1,
```

```
                                spec.limits =
c(CEPcluster1$limits[1],CEPcluster1$limits[2]))
IndCapSeisSig_Cluster2  = qcc:process.capability(object = CEPcluster2,
                                spec.limits =
c(CEPcluster2$limits[1],CEPcluster2$limits[2]))
IndCapSeisSig_Cluster3  = qcc:process.capability(object = CEPcluster3,
                                  spec.limits =
c(CEPcluster3$limits[1],CEPcluster3$limits[2]))
IndCapSeisSig_Cluster4  = qcc:process.capability(object = CEPcluster4,
                                  spec.limits =
c(CEPcluster4$limits[1],CEPcluster4$limits[2]))
IndCapSeisSig_Cluster5  = qcc:process.capability(object = CEPcluster5,
                                  spec.limits =
c(CEPcluster5$limits[1],CEPcluster5$limits[2]))
IndCapSeisSig_Cluster6  = qcc:process.capability(object = CEPcluster6,
                                  spec.limits =
c(CEPcluster6$limits[1],CEPcluster6$limits[2]))
IndCapSeisSig_Cluster7  = qcc:process.capability(object = CEPcluster7,
                                  spec.limits =
c(CEPcluster7$limits[1],CEPcluster7$limits[2]))
IndCapSeisSig_Cluster8  = qcc:process.capability(object = CEPcluster8,
                                  spec.limits =
c(CEPcluster8$limits[1],CEPcluster8$limits[2]))
IndCapSeisSig_Cluster9  = qcc:process.capability(object = CEPcluster9,
                                  spec.limits =
c(CEPcluster9$limits[1],CEPcluster9$limits[2]))
IndCapSeisSig_Cluster10  = qcc:process.capability(object = CEPcluster10,
                                  spec.limits =
c(CEPcluster10$limits[1],CEPcluster10$limits[2]))
IndCapSeisSig_Cluster11  = qcc:process.capability(object = CEPcluster11,
                                  spec.limits =
c(CEPcluster11$limits[1],CEPcluster11$limits[2]))
IndCapSeisSig_Cluster12  = qcc:process.capability(object = CEPcluster12,
                                  spec.limits =
c(CEPcluster12$limits[1],CEPcluster12$limits[2]))
IndCapSeisSig_Cluster13  = qcc:process.capability(object = CEPcluster13,
                                  spec.limits =
c(CEPcluster13$limits[1],CEPcluster13$limits[2]))
IndCapSeisSig_Cluster14  = qcc:process.capability(object = CEPcluster14,
                                  spec.limits =
c(CEPcluster14$limits[1],CEPcluster14$limits[2]))
```