UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE CIÊNCIAS AGRÁRIAS

DEPARTAMENTO DE COMPUTAÇÃO

RAPHAEL MÁRIO DE ALMEIDA ROMUALDO

SISTEMA PARA ESTIMATIVA DE PARÂMETROS DE FUNÇÕES DE PROBABILIDADE PARA VARIÁVEIS CONTÍNUAS COM O MÉTODO DA MÁXIMA VEROSSIMILHANÇA

ALEGRE

2015

RAPHAEL MÁRIO DE ALMEIDA ROMUALDO

SISTEMA PARA ESTIMATIVA DE PARÂMETROS DE FUNÇÕES DE PROBABILIDADE PARA VARIÁVEIS CONTÍNUAS COM O MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Monografia apresentada ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: Prof. M.SC. Clayton Vieira Fraga Filho.

ALEGRE

2015

RAPHAEL MÁRIO DE ALMEIDA ROMUALDO

SISTEMA PARA ESTIMATIVA DE PARÂMETROS DE FUNÇÕES DE PROBABILIDADE PARA VARIÁVEIS CONTÍNUAS COM O MÉTODO DA MÁXIMA VEROSSIMILHANÇA

Monografia apresentada ao Departamento de Computação do Centro de Ciências Agrárias da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Aprovado em 03 de Julho de 2015.

ORIENTADOR

Prof. M. Sc. Clayton Vieira Fraga Filho

Universidade Federal do Espírito Santo
Orientador

“Antes de tudo, sou grato a meu Deus, mediante Jesus Cristo, por todos vocês, porque em todo o mundo está sendo anunciada a fé que vocês têm. ”

Romanos 1:8

AGRADECIMENTOS

Gostaria de agradecer em primeiro lugar a Deus, que me deu força ao longo de todos os anos, e nunca me deixou desistir. E por todas as pessoas que colocou em minha vida para me manter firme na fé.

Agradeço minha família, meus pais e avós que me deram todo apoio para que pudesse estar por conta dos estudos, me dedicando 100% ao curso. E em especial meus avós Mário Batista Pimentel e Rosa Maria de Almeida Pimentel que foram minha inspiração para concluir essa etapa da vida.

Agradeço a Bruna Sales Silva, minha namora, que teve toda paciência e compreensão durante essa jornada.

Deixo um grande abraço para Jakson Laquini, Fernanda Roldi, Thays Ferreira, Rafaella Benedito, Mário Rodrigues, Breno Silva e tantos outros amigos que passaram em minha vida durante esses anos.

Também agradeço todos os professores que tive oportunidade de conhecer, obrigado pelas aulas, pelas dicas, e por me fazer uma pessoa melhor. Irei carregar por toda vida a formação moral e teórica que aprendi ao longo do curso.

Sumário

LISTA DE FIGURAS 8
LISTA DE TABELAS 9
RESUMO 10
1. INTRODUÇÃO 11
1.1 OBJETIVOS 14
1.1.1 Objetivo geral 14
1.1.2 Objetivos específicos 14
2. REVISÃO DE LITERATURA 15
2.1 Função de distribuição de probabilidade 15
2.1.1 Distribuição de probabilidade Weibull 17
2.1.2 Distribuição de probabilidade Normal 19
2.2 Máxima verossimilhança 20
2.3 Dedução da função Weibull pelo método da máxima verossimilhança 21
2.4 Pacote OPTIMX 22
2.5 Algoritmos numéricos 23
2.5.1 Newton-Raphson 23
2.5.2 Broyden-Fletcher-Goldfarb-Shanno (BFGS) 26
2.5.3 Nelder-Mead 27
2.6 Testes não paramétricos 29
2.6.1 Teste de ajustamento Kolmogorov – Smirnov 30
2.6.2 Critério de Informação de Akaike (AIC) 31
2.6.3 Critério de Informação Bayesiano (BIC) 32
2.6.4 Teste de Normalidade Shapiro-Wilk 32
2.6.5 Teste t de Student 32
2.6.6 Teste de Mann-Whitney (Wilcoxon) 33
3. METODOLOGIA 35
3.1 Introdução 35
3.2 Visão geral do sistema desenvolvido 36
3.3 Implementação dos scripts e integração 41
3.4 Integração de novos scripts para funções de distribuição 41
3.5 Execução dos scripts 42
3.6 Integração do Estima com o sistema estatístico R 43

3.7 Apresentação dos dados usados no estudo de caso 45
3.8 Estudo de Caso 49
3.9 Testes usados para validação dos resultados 52
4. RESULTADOS E DISCUSSÃO 55
4.1.1 Validação dos resultados obtidos no Estima 55
4.1.1.1 Validação da função Normal 55
4.1.2 Validação da função Weibull 2p 58
4.1.3 Validação da função Weibull 3P 60
5. CONCLUSAO 64
5.1 Trabalhos Futuros 64
6. REFERÊNCIAS BIBLIOGRÁFICA 65
7. APÊNDICES 70
APÊNDICE A – Dados da função normal usados no estudo de caso 70
APÊNDICE B – Dados da função Weibull2p usados no estudo de caso 71
APÊNDICE C – Script usado para teste t da função Normal no estudo de caso 72
APÊNDICE D – Script usado para teste t da função Weibull2p no estudo de caso 73
APÊNDICE E – Gráficos PDF e CDF da idade 60 74
APÊNDICE F – Gráficos PDF e CDF da idade 72 76
APÊNDICE G – Script usado no SAS para o estudo de caso 78
8. ANEXOS 79
ANEXO A - Dedução da função de máxima verossimilhança para a distribuição: 79

LISTA DE FIGURAS

Figura 1 – Exemplo de uso das funções Weibull, LogGamma, LogNormal, Pearson 5 (3P) e Rayfeigh. 17
Figura 2 – Representação de uma curva gaussiana. 19
Figura 3 – Pseudocódigo do método de Newton-Raphson. 24
Figura 4 – Pseudocódigo do ciclo básico do BFGS. 26
Figura 5 – Exemplo de busca pelo método de Nelder-Mead. 27
Figura 6 - Pseudocódigo do método de Nelder-Mead. 29
Figura 7 – Exemplo de bom e mau ajustamento. 30
Figura 8 – Fluxograma das etapas para construção do sistema proposto. 35
Figura 9 – Diagrama de casos de uso do sistema Estima. 36
Figura 10 – Diagrama de Sequência de um exemplo de utilização do sistema. 37
Figura 11 – Interface gráfica dos resultados dos testes. 38
Figura 12 – Interface gráfica dos resultados dos parâmetros das funções. 39
Figura 13 – Interface gráfica dos gráficos das distribuições. 39
Figura 14 – Script da estatística descritiva. 40
Figura 15 – Interface gráfica das estatísticas realizadas. 41
Figura 16 – Script de configuração das funções. 42
Figura 17 – Exemplo de arquivo csv dos resultados obtidos para função Normal 42
Figura 18 – Diagrama de sequência da execução do script 43
Figura 19 – Diagrama de componentes do rConector. 44
Figura 20 – Diagrama de sequência do uso rConector. 44
Figura 21 – Tela de importação dos dados. 49
Figura 22 – Dados importados para o sistema Estima. 50
Figura 23 – Tela de configuração do Ajuste. 50
Figura 24 – Imagem parcial da tela de exibição de todos os resultados. 52
Figura 25 – Gráfico *Boxplot* dos parâmetros média e desvio. 55
Figura 26 – Gráfico Boxplot dos parâmetros forma e escala. 58

LISTA DE TABELAS

Tabela 1 – Modelos das distribuições de probabilidade gama, log-normal e hiperbólica, com as respectivas funções de densidade de probabilidade (fdp) e a função de distribuição acumulada (f.d.a)..... 16

Tabela 2 – Exemplo de representação dos dados utilizados no estudo de caso. 46

Tabela 3 – Quantidade de árvores por Parcela x Idade. 46

Tabela 4 – Histogramas dos dados referente a cada Idade x Parcela. 47

Tabela 4 – Resultado do teste de normalidade Shapiro Wilk para o parâmetro média da distribuição Normal. 56

Tabela 5 – Resultado do teste de normalidade Shapiro Wilk para o parâmetro Desvio da distribuição Normal 56

Tabela 6 – Resultado do teste f, para verificar homogeneidade dos dados. 57

Tabela 7 – Resultado do teste t de Student. 57

Tabela 8 – Resultados do teste Shapiro Wilk para o parâmetro Escala. 58

Tabela 9 – Resultados do teste Shapiro Wilk para o parâmetro Forma. 59

Tabela 10 – Resultado do teste t de Wilcoxon. 59

Tabela 11 – Gráficos PDF e CDF da idade 36. 60

Tabela 12 – Resultados do teste de KS referente aos gráficos da Tabela 11. 62

Tabela 13 – Resultados do teste de KS referente aos gráficos do Apêndice F. 63

Tabela 14 – Resultados do teste de KS referente aos gráficos do Apêndice G. 63

RESUMO

Em diversas áreas como no desenvolvimento de software, na química, em setores de RH, na engenharia florestal e em inúmeras outras, existe uma diversidade de dados que podem ser mensurados. É de grande utilidade conseguir a partir desse conjunto de dados estimar os parâmetros que, aplicados a uma função de distribuição de probabilidade, são capazes de projetar o comportamento do conjunto, permitindo que se estimem resultados futuros.

Esse trabalho teve como objetivo desenvolver um sistema gratuito para realizar a estimativa de parâmetros das funções de probabilidade Weibull e Normal.

Para encontrar os parâmetros dessas funções foram usados métodos iterativos presente no pacote *optimx* disponível para linguagem R, que receberam a função deduzida pelo método da máxima verossimilhança e tentaram encontrar os parâmetros que explicam o comportamento dos dados. Posteriormente, depois de encontrado os parâmetros, foi usado o teste não paramétrico Kolmogorov-Smirnov para verificar a aderência do modelo aos dados, e consequentemente saber qual modelo melhor expressa aquele conjunto de dados analisado.

Com o objetivo de realizar a estimativa de parâmetros, foi implementado um sistema de informação batizado de “Estima”, que integre as funcionalidades disponíveis no software R com rotinas previamente criadas para ajuste de modelos, junto a uma interface gráfica de uso comum aos usuários.

Para comparar o desempenho dos resultados foi feito um estudo de caso com dados reais e comparado com o resultado do sistema estatístico SAS.

Por meio do estudo de caso realizado, constatou-se que os resultados encontrados pelo Estima são estatisticamente iguais ao SAS, comprovando o desempenho dos resultados obtidos.

Palavras-Chaves: Máxima verossimilhança; Estimativa de parâmetros; Funções probabilidade, Testes não paramétricos.

1. INTRODUÇÃO

As funções de distribuição de probabilidade são importantes para estimar o acontecimento potencial de eventos de ocorrência incerta(GUIMARÃES, 2002). Saber se irá chover, prever uma eventual queda da bolsa de valores, calcular o valor de um terreno dado um período de tempo, são informações que podem levar a tomar decisões, levando ao sucesso ou ao fracasso. Outro exemplo é o processo de previsão de vendas que segundo Correa (2001), é possivelmente o mais importante dentro da função de gestão de demanda. Nestes casos, antecipar-se faz total diferença.

O uso de funções densidade de probabilidade está diretamente ligado à natureza dos dados a que ela se relaciona. Algumas têm boa capacidade de estimação para pequeno número de dados, outras requerem grande série de observações (PUC-RIO, 2002). Existem diversas distribuições de probabilidade, como binomial usada em modelos da distribuição temporal de chuvas intensas em Piracicaba/SP (CRUCIANI, MACHADO, 2002), geométrica, de Poisson, e várias outras para variáveis discretas. Também temos a distribuição de probabilidade Weibull e a Normal usada no ajuste de funções de distribuição de probabilidade à radiação solar (CARGNELUTTI FILHO , 2004), a log-normal usada nas distribuições de probabilidade de velocidade e potência do vento (SANSIGOLO, 2005), entre outras para variáveis contínuas.

De acordo com Bussab (2002), uma distribuição de probabilidade descreve quais chances que uma variável pode assumir em um domínio. Ou seja, é uma função cujo domínio é o valor da variável e a imagem é a probabilidade da variável assumir cada valor do domínio. O conjunto imagem deste tipo de função sempre se restringe ao intervalo entre 0 e 1. Portanto, segundo Peternelli (2006), chama-se de função densidade de probabilidade (f.d.p.) da variável aleatória contínua X, a função f(x) que atenda ás seguintes condições:

$$f(x) \geq 0 \quad para \;\; a < x < b, \tag{1}$$

$$\int_a^b f(x)dx = 1 \tag{2}$$

onde a e b podem ser, respectivamente, $-\infty$ e $+\infty$.

Para a construção de modelos de distribuição de diâmetros, as funções de densidade de probabilidade que mais se destacam são a Beta e a Weibull (LEITE et al., 2013). O sucesso da distribuição Weibull se justifica não só pela sua eficácia, mas principalmente por ser capaz de fazer previsões de acurácia razoável mesmo quando a quantidade de dados disponível é baixa. Ela foi proposta originalmente por Waloddi Weibull que em 1951 lançou um artigo descrevendo a distribuição em detalhes e propondo diversas aplicações (WEIBULL, 1951). Dentre os trabalhos que utilizam essa função, podem ser citados os trabalhos de Catalunha et al. (2002), Binoti *et al.* (2010), Batista (1989).

Toda função de distribuição de probabilidade encontra-se de forma parametrizada. O parâmetro é um termo de amplo uso na matemática, estatística, computação e análise de sistemas. A parametrização de uma distribuição é um processo de abstração que possibilita a flexibilização do seu uso, desse modo, uma mesma distribuição pode ser utilizada para modelar variáveis aleatórias em contextos muito amplos (MACHADO, 2014).

O parâmetro é uma característica numérica da distribuição de probabilidade. Na estatística, a estimativa de parâmetros busca a partir de um conjunto de dados e um modelo estatístico estimar valores para os diferentes parâmetros do modelo. Existem alguns métodos como a máxima verossimilhança, o método dos momentos e o método dos percentis que possibilitam a estimação dos parâmetros de modelos de probabilidade. Embora o método da máxima verossimilhança tenha sido criado antes da década de 80, somente depois passou a ser mais utilizado na econometria, devido ao desenvolvimento dos computadores pessoais de grande potência (MARCELO S. PORTUGAL, 1995).

De acordo com Schmidt (2014), dentre os métodos utilizados para o ajuste da função Weibull, o que tem sido mais utilizado é o método da máxima verossimilhança. Vários estudos já comprovaram a eficiência dele em relação a outros métodos, como método dos mínimos quadrados e método dos momentos (LOPES, 2006). Ele é considerado o melhor por fornecer estimativas mais

confiáveis, precisas e com menor tendenciosidade quando comparado a métodos como o dos momentos e o dos percentis (BAILEY; DELL, 1973; LEITE *et al.*, 2006; SANTOS, 2008; SILVA, 2001).

Embora o método da máxima verossimilhança seja o mais indicado, sua aplicação é bem complexa, exigindo um conhecimento de dedução matemática, integralização e derivação. Na ausência de soluções analíticas é necessário o uso de aproximações na construção da função de verossimilhança, como o uso do método de quasi-máxima verossimilhança, que gera estimadores com mínimo erro quadrático médio ou uso de máxima verossimilhança simulada, que utiliza trajetórias simuladas por discretizações de Euler ou Milstein na avaliação da verossimilhança (LAURINI; HOTTA, 2009).

Devido à dificuldade do uso desses métodos matemáticos, há quem recorra ao apoio de softwares estatísticos para auxiliar nesse processo de estimativa de parâmetros como o SAS (INC.SAS/STAT, 2000), R (R DEVELOPMENT CORE TEAM, 2005) o STATISTICA (AMERICA, 2014). Mas, muitas vezes o usuário possui pouca informação sobre o uso e a teoria do ajuste dos modelos, como também do uso da linguagem de programação ou *script,* usadas nos sistemas citados. O software R (R DEVELOPMENT CORE TEAM, 2005) tem sido amplamente usado para cálculos estatísticos, devido sua ampla variedade de plataformas (UNIX, Windows e MacOS), e recursos disponíveis para análises estatísticas e gráficos. Muitos pesquisadores como Baayen (2001) e Bunn (2014), têm se esforçado para que o software R seja popularizado e que aumente o número de análises a serem realizadas pelo mesmo. Porém, embora todo seu conteúdo seja disponível gratuitamente, é necessário tempo para aprender manipular sua linguagem.

A existência de um sistema de informação com distribuição gratuita para usuários sem experiência prévia de programação de computadores e manipulação de dados, que apresentasse uma interface gráfica comum ao usuário possibilitando a realização de estimativa de parâmetros de funções de distribuição de probabilidade, contribuiria de forma significativa para a área acadêmica e também para profissionais de diversas áreas.

1.1 OBJETIVOS

Os objetivos apresentados nesse trabalho serviram como base para as metas a serem desenvolvidas, buscando ao final do trabalho apresentar o cumprimento ou a justificativa de não cumprimento de cada objetivo. Assim concluindo o desempenho obtido pelo trabalho.

1.1.1 Objetivo geral

Implementar um sistema de informação que permita o usuário realizar a estimativa de parâmetros de funções de probabilidade para variáveis contínuas com o método da máxima verossimilhança.

1.1.2 Objetivos específicos

a. Integrar scripts da função Weibull 2 e 3 parâmetros desenvolvidos por Fraga Filho e Schmidt (2014) para a linguagem R (R DEVELOPMENT CORE TEAM, 2005) com o sistema proposto.

b. Implementar o script para estimativa de parâmetros da função Normal desenvolvido na linguagem R.

c. Integrar os scripts desenvolvidos com o sistema proposto.

d. Validar o sistema proposto comparando seus resultados com o sistema SAS (INC.SAS/STAT, 2000).

2. REVISÃO DE LITERATURA

Nas sessões seguintes serão apresentados os métodos, funções de distribuição, tecnologias e os testes usados para o desenvolvimento deste trabalho, com base em outros trabalhos, livros e artigos.

2.1 Função de distribuição de probabilidade

Segundo Péllico e Brena (1997), a maioria das decisões humanas atuais é tomada a partir da análise ou estudo de apenas uma parte, do todo que envolve o problema. Uma teoria que tem sido utilizada de forma significativa na ciência é a das probabilidades. Por exemplo, para o diagnóstico de uma doença a partir de uma pequena análise de tecido do corpo humano. Outro exemplo, é que a partir da análise de um lote significativo de placas de compensado é possível tirar conclusões a respeito de sua qualidade e, com isso, tomar decisões econômicas ou de interesse científico (SILVA, 2003).

Dentre as diversas funções de distribuição de probabilidade existentes, destacam-se para o ajuste de dados amostrais de variáveis aleatórias discretas: Bernoulli, binomial, binomial negativa, hipergeométrica, geométrica e Poisson. Já as distribuições uniforme, normal, log-normal, gama, valores extremos ou Gumbel, Weibull, exponencial, beta, qui-quadrado, t de Student, F de Snedecor, hiperbólica, entre outras, são utilizadas para o ajuste de dados amostrais de variáveis aleatórias contínuas (CARGNELUTTI FILHO et al., 2004).

De acordo com Bailey e Dell (1973), Cunha (1994), Catalunha et al. (2002), Guimarães (2002), Campos e Leite (2009) e Binoti et al. (2012), os modelos das distribuições de probabilidade gama, log-normal e hiperbólica podem ser descritos de acordo com a Tabela 1:

Tabela 1 – Modelos das distribuições de probabilidade gama, log-normal e hiperbólica, com as respectivas funções de densidade de probabilidade (fdp) e a função de distribuição acumulada (f.d.a).

Distribuição	**Função**	**Modelo**
Gama	f.d.p	$f(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\ (x)^{\alpha-1}\ exp\ \left(-\frac{x}{\beta}\right)$ $0 < x < \infty, \alpha\ > 0$ e $\beta > 0$.
	f.d.a	$F(a) = \int_0^{\infty} u^{\alpha-1}\ exp^{-u}\ du$ $F(x) = \int_0^{x} \frac{(x)^{\alpha-1}\ exp\ \left(-\frac{x}{\beta}\right)}{\beta^{\alpha}\Gamma(\alpha)}\ dx$ $x > 0, \alpha\ > 0$ e $\beta > 0$. α é o parâmetro de forma, β é o parâmetro escala.
Log-normal	f.d.p	$f(x)\ = \frac{exp\ -\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}{x\sigma\sqrt{2\pi}}\ , x \geq 0$
	f.d.a	$F(x) = \frac{1}{x\sigma\sqrt{2\pi}}\int_o^{x} exp\ -\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\ dx$ σ^2 é a variância, σ é o desvio padrão e μ é a média.
Hiperbólica 2 e 3 parâmetros	f.d.p	$f(x)\ = \left\{\frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1}\left(1 - \tanh\ \left(\frac{x}{\beta}\right)^{\gamma}\right)^2\right\}$
	f.d.p	$f(x)\ = \left\{\frac{\gamma}{\beta}\left(\frac{x-\alpha}{\beta}\right)^{\gamma-1}\left(1 - \tanh\ \left(\frac{x-\alpha}{\beta}\right)^{\gamma}\right)^2\right\}$ $x > 0, \beta > 0, \gamma > 0.$ α é o parâmetro de locação, β é o parâmetro de escala e γ é o parâmetro de forma.

Autor: Adaptado de Schmidt (2014)

Segundo Santos (2008), as funções de densidade de probabilidade que mais se sobressaem no meio florestal são a beta e a Weibull.

2.1.1 Distribuição de probabilidade Weibull

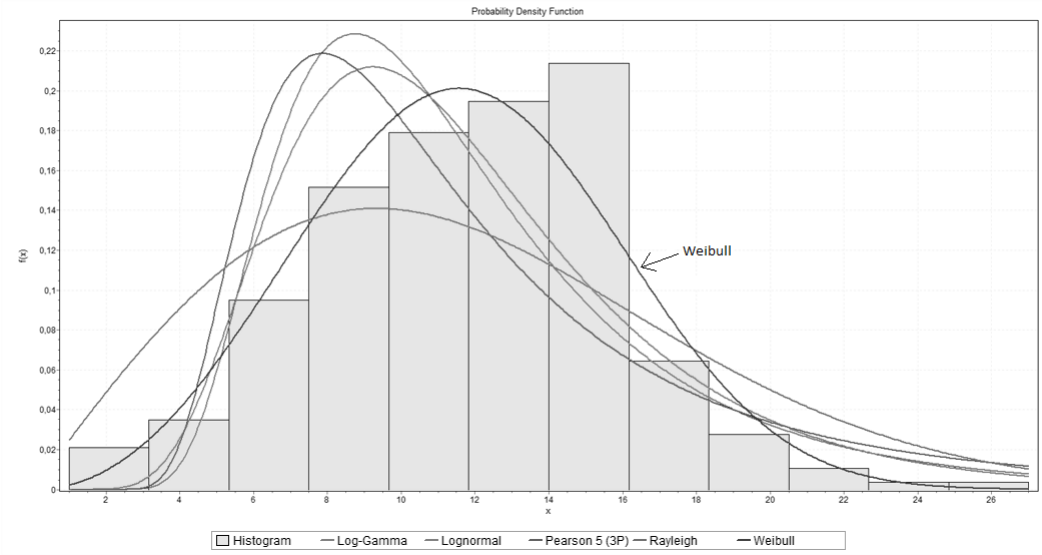
A distribuição de probabilidade Weibull recebeu esse nome devido ao engenheiro e matemático Waloddi Weibull que a desenvolveu em seus estudos sobre a resistência dos materiais no ano de 1939 (BAILEY; DELL, 1973).

Em 1951 Waloddi escreveu um artigo descrevendo a distribuição em detalhes e propondo diversas aplicações da função (WEIBULL, 1951).

Para exemplificar, nas ciências florestais, a função Weibull tem sido usada intensamente, desde que Bailey e Dell (1973) empregou-a como modelo de distribuição diamétrica. Além disso, consideraram a função Weibull matematicamente mais simples e flexível, além de ser, quando comparada a distribuição beta mais fácil de ser aplicada pelo fato de não exigir integração numérica para se calcular o número de indivíduos nas classes (SCHMIDT, 2014).

Ainda exemplificando na área florestal, seja uma base de dados com informações sobre o diâmetro à altura do peito (DAP) usados no trabalho de Schmidt (2014), usando o software EasyFit (MATHWAVE TECHNOLOGIES, 2014) foi gerada a Figura 1.

Figura 1 – Exemplo de uso das funções Weibull, LogGamma, LogNormal, Pearson 5 (3P) e Rayfeigh.



Cada curva da Figura 1 representa uma função de distribuição, e observando a aproximação de curva ao longo das caixas do histograma, a curva da função Weibull demonstra melhor o comportamento dos dados em relação as outras funções usadas para esse conjunto de dados.

A distribuição diamétrica de determinado povoamento ou floresta é facilmente avaliada por meio da medição direta das árvores, com os diâmetros agrupados em classes de amplitudes definidas. No entanto, somente o conhecimento acerca da distribuição diamétrica atual de um povoamento nem sempre é suficiente. Quando se deseja analisar o crescimento e a produção por classe de diâmetro, é necessário empregar modelos de distribuição diamétrica (ARAÚJO JÚNIOR *et al.*, 2010).

De acordo com Batista (1989), as funções de densidade de probabilidade podem ser representadas de diversas formas na distribuição de Weibull, porém nos estudos de distribuição de diâmetros a forma mais comum é a seguinte:

$$f(x) = \left(\frac{c}{b}\right)\left(\frac{x-a}{b}\right)^{c-1} exp\left[-\left(\frac{x-a}{b}\right)^{c}\right] \qquad \text{para } x > a \qquad (3)$$

$$f(x) = 0 \qquad \text{para } x < a \qquad (4)$$

Onde:

- x é a variável aleatória, $a \geq 0$, $b > 0$ e $c > 0$ são os parâmetros forma, escala e locação respectivamente da distribuição.

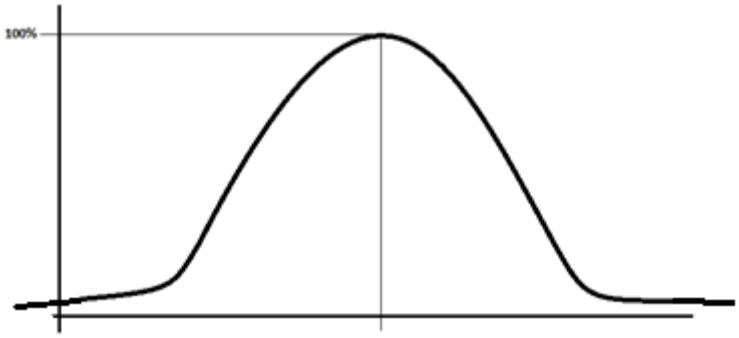
A função de densidade de probabilidade da equação (3) é normalmente chamada de “Weibull - 3 parâmetros” (BATISTA, 1989). O parâmetro a é chamado de parâmetro de posição ou locação, pois controla a posição da curva sobre o eixo das abcissas. Quando ocorre a supressão do parâmetro a, ou seja, quando esse é zero, a distribuição inicia-se na origem surgindo um caso especial, caracterizando a função como “Weibull – 2 parâmetros” (BATISTA, 1989; BINOTI et al., 2010).

2.1.2 Distribuição de probabilidade Normal

No século XVIII, astrônomos e outros cientistas observaram que medidas repetidas de mensurações como a distância à lua variavam quando coletadas em grande número (AFONSO, 2005).

Esta forma gráfica apresentada na figura 2, era associada aos erros de mensuração, daí o nome de “Distribuição normal dos erros” e depois “Distribuição normal”. Também é conhecida por “Distribuição Gaussiana”, em função do modelo matemático desenvolvido por Karl F. Gauss para este comportamento (AFONSO, 2005). A distribuição normal é a função de probabilidade contínua mais utilizada (HASTINGS; PEACOCK, 1975).

Figura 2 – Representação de uma curva gaussiana.



Como é possível visualizar na Figura 2 a curva normal tem forma de sino e é simétrica em relação a média, prolongando-se de $-\infty$ a $+\infty$ (AFONSO, 2005). A forma mais comum de representação é a equação 6:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty \qquad (5)$$

onde:

- $\mu =$ média.
- $\sigma =$ desvio padrão.

2.2 Máxima verossimilhança

O método da máxima verossimilhança, foi proposto inicialmente por Gauss em 1821 e redescoberto por Fisher em 1922. Ele é o método mais usado para a estimação de parâmetros, pois se baseia no princípio de que estando o modelo correto, tudo que os dados tinham para informar sobre os parâmetros deverá estar contido na função de verossimilhança (SILVA & CUNHA, 1994).

A estimação por máxima verossimilhança é um método de estimação para obtenção de estimadores de uso extremamente amplo porque possibilita realizar inferências com propriedades altamente desejáveis. Segundo Neter et. al. (1990), o método da máxima verossimilhança utiliza a distribuição de probabilidade conjunta de uma amostra de observações. Ela é chamada de função de verossimilhança quando uma distribuição de probabilidade conjunta é escrita como uma função de parâmetros, dada uma amostra em particular de observações.

De acordo com Naghettini e Pinto (2007), o método da máxima verossimilhança consiste basicamente em maximizar uma função dos parâmetros da distribuição, conhecida como função de verossimilhança. A partir do equacionamento para a condição de máximo, tem-se um sistema com o mesmo número de equações e incógnitas, cujas soluções produzem os estimadores de máxima verossimilhança.

Portanto em termos formais, a função da verossimilhança é dada pela equação 8:

$$L(\theta_1, \theta_2, \dots, \theta_k) = \prod_{i=1}^{N} f_y\,(y_i, \theta_1, \theta_2, \dots, \theta_k) \qquad (6)$$

Segundo Naghettini e Pinto (2007), essa é uma função dos parâmetros θ_j, exclusivamente. A busca da condição de máximo para a função de verossimilhança resulta no seguinte sistema de *k* equações e *k* incógnitas:

$$\frac{\partial L(\theta_1, \theta_2, \dots, \theta_k)}{\partial \theta_j} = 0;\ j = 1,2, \dots, k \qquad (7)$$

As soluções desse sistema de equações são os estimadores $\hat{\theta}_j$ de máxima verossimilhança. É frequente o emprego da função logaritmo de verossimilhança $ln[L(\theta)]$ em substituição à função de verossimilhança propriamente dita para facilitar a construção do sistema de equações 6. Isso se justifica pelo fato da função logaritmo ser contínua, constante e crescente, e, portanto, maximizar o logaritmo da função é o mesmo que maximizar a função (NAGHETTINI; PINTO, 2007).

2.3 Dedução da função Weibull pelo método da máxima verossimilhança

As amostras utilizadas na distribuição Weibull podem ser completas ou apresentar truncamento (BATISTA, 1989). O truncamento é uma maneira de aperfeiçoar as estimativas. Essa pode ser definida como um subintervalo da distribuição completa, na qual a função de densidade de probabilidade não é a mesma para ambas. Quando se deseja trabalhar com um subintervalo da distribuição completa é necessário desenvolver novas equações com base em uma distribuição de diâmetros truncada (SILVA, 2001).

O truncamento ocorre devido ao processo de seleção inerente ao planejamento do estudo. O truncamento é caracterizado por excluir os indivíduos que não são relevantes para o estudo em questão (FERNANDES, 2010).

Na área florestal, por exemplo, de acordo com Silva (2001) os dados de mensuração dos diâmetros das árvores são ditos como completos quando uma parcela apresenta todas as árvores com os diâmetros mensurados.

Segundo Fernandes (2010), o truncamento à direita ocorre quando apenas são considerados na análise estatística os indivíduos para os quais se observou um determinado acontecimento de interesse durante o período em estudo. Ou seja, quando os indivíduos não são acompanhados a partir do tempo inicial, mas somente após ocorrer determinado evento. O truncamento à esquerda é mais comum e ocorre quando apenas são incluídos na amostra indivíduos que sobrevivem tempo suficiente para que ocorra um determinado acontecimento antes do evento de interesse.

A dedução da função Weibull de 2 parâmetros (2P) e Weibull de 3 parâmetros (3P) completa deduzidas por Schmidt (2014), encontra-se no Anexo A.

2.4 Pacote OPTIMX

O pacote *optimx* é um pacote de funções para o software R, desenvolvido e mantido por John C. Nash e Ravi Varadhan em 1995 (NASH; GROTHENDIECK; VARADHAN, 2014). Segundo Nash (2014), o pacote *optimx* é um envoltório para uma série de rotinas de otimização numérica. O propósito é unificar as chamadas para um número de otimizadores úteis em uma sintaxe comum, e assim permitir a sua comparação de desempenho de uma forma objetiva.

Esta capacidade de comparação facilita a compreensão das chamadas realizadas pelo *optimx* a uma série de outros pacotes. Sintaxes diferentes para chamadas de diferentes rotinas são um grande incômodo para os usuários e uma fonte frequente de erro (NASH, 2014).

Uma das principais ferramentas do pacote é a função *optimx()* que agrupa seis principais métodos de otimização, que usam abordagens diferentes buscando a resolução da função informada como parâmetro:

- Nelder-Mead;
- Broyden-Fletcher-Goldfarb-Shanno (BFGS);
- Conjugate gradient (CG);
- Limited-memory BFGS (L-BFGS-B);
- Simulated annealing (SANN);
- Brent;

No entanto, Nash (2014) afirma que o *optimx* é mais do que simplesmente uma ferramenta para unificar a sintaxe da sequência de chamadas. Seu design incorpora diversos conceitos importantes, dentre eles:

- Uma série de testes iniciais da função do gradiente e do dimensionamento estão disponíveis, antes de iniciar a otimização;

- Depois de completar a otimização, é realizado a verificação das soluções;
- Métodos como *summary(), coef() e print()* estão disponíveis para visualização dos resultados encontrados.

O pacote *optimx*, utiliza métodos matemáticos iterativos conhecidos como algoritmos numéricos, onde é realizada uma busca iterativa para estimar os parâmetros da função.

2.5 Algoritmos numéricos

Algoritmos numéricos são uma área que tenta resolver problemas matemáticos utilizando a computação. Esses problemas podem ser os mais diversos na ciência: física, química, matemática, engenharias, biologia, etc.

Uma área que historicamente impulsionou muito o desenvolvimento dos algoritmos numéricos é a área bélica. A situação mais direta de como esses algoritmos podem ser aplicados é no cálculo da trajetória de um projétil, seja ele um foguete, míssil, morteiro, etc. O grande interesse nessas situações é garantir que o projétil atinja o alvo com maior precisão possível (FILHO, CAMPOS, 2007).

Independentemente da área de aplicação, os algoritmos numéricos são métodos que buscam aproximar o resultado analítico dos problemas estudados. Entretanto, ao se controlar o nível de erro da resposta numérica, esta pode se tornar confiável perante o custo da obtenção da resposta analítica exata. Tendo isso em mente, é importante que o nível de erro seja controlado, inserido em cada etapa, até se obter a resposta final do problema real (FILHO, CAMPOS, 2007).

2.5.1 Newton-Raphson

Este método foi desenvolvido por Isaac Newton e Joseph Raphson, e tem o objetivo de estimar as raízes de uma função. Ele foi otimizado para que fosse possível estimar os parâmetros através da função de verossimilhança. Na Figura 3, é apresentado um pseudocódigo do método de Newton-Raphson, onde se escolhe uma aproximação inicial para o passo x_0. Após isso, calcula-se a

derivada da função nesse ponto e a interseção dela com o eixo das abcissas, buscando encontrar uma aproximação melhor para a raiz.

Figura 3 – Pseudocódigo do método de Newton-Raphson.

Considere a equação $f(x) = 0$;
a aproximação inicial: $x0$; a tolerância : tol

Passo 0: Avaliar f em $x0$: $f0 = f(x0)$;
se $|f0| <$ tol , faça $\overline{x} = x0$. Parar.
caso contrário: k = 0;

Passo 1: Avaliar a derivada de f em $x0$: $dv = f'(x0)$;

Passo 2: Faça: $x1 = x0 - f0/dv$;
Passo 3: Avaliar f em $x1$: $f1 = f(x1)$;
se $|f1| <$ tol , faça $\overline{x} = x1$. Parar.

Passo 4: Se $|x1 - x0| <$ tol , faça $\overline{x} = x1$. Parar.

Passo 5: Faça: $x0 = x1$ e $f0 = f1$;

Passo 6: Se $k = itmax$, parar: excedido o número de iterações.
caso contrário: $k = k + 1$, voltar ao passo 1.

Fonte: Reggiero, 2008.

Repetindo-se o processo, cria-se um método iterativo para encontramos a raiz da função (BURDEN; FAIRES; TASKS, 2008). O critério de parada será quando a valor de f estiver próximo de zero $|f\ (x_k)| < tol$, quando a diferença entre as duas últimas aproximações está próximo de zero $|x_{k+1} - x_k| < tol$ ou quando atingir o número máximo de iterações.

Seja a equação $f\ (x)\ =\ 0$, da qual se conhece a raiz aproximada x_0 e seja δ_0 o erro dessa raiz:

$$f\ (x_0 + \delta_0) = 0 \tag{8}$$

se

$$f(x_0 + \delta_0) = f(x_0) + f'(x).\delta_0 + R_2 = 0 \tag{9}$$

onde R_2 é um infinitésimo de 2ª ordem em δ_0 que pode ser escrito como:

$$R_2 = \frac{\delta_0^2}{2!}.F'(x_0) + \cdots \quad (10)$$

e por conseguinte será pequeno se δ_0 e $f''(x_0)$ não for muito elevado.

Teoricamente o valor $\delta_{n-1} = -\frac{f'(x_{n-1})}{f''(x_{n-1})}$ tem tendência a ser a cada vez menor. O que justifica a possibilidade de ser desprezado. Desprezando R_2 temos que:

$$f(x_0) + f'(x).\delta_0 = 0 \quad (11)$$

ou seja,

$$\delta_0 = -\frac{f(x_0)}{f'(x_0)} \quad (12)$$

Assim um novo valor $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$ mais aproximado da raiz da equação pode ser obtido.

Prosseguindo a iteração, obtém-se uma sequência de valores sucessivamente mais aproximados da raiz.

A fórmula de recorrência é dada por:

$$x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})} \quad (13)$$

As condições de convergência são:

1. x_0 é suficientemente próximo de uma raiz da equação.
2. $f''(x)$ não toma valores excessivamente grandes.
3. $f'(x)$ não é muito próxima de zero.

Como um exemplo de aplicação, seja uma função $f(x) = \mathrm{x}^2 + \mathrm{x} - 6$, $x_0 = 3$ e $\delta = 0{,}020$ a resolução pelo método de Newton-Raphson é dada a seguir:

Onde $f'(x) = 2x + 1$ e $x_n = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$

$x_1 = 3 \quad - \frac{3^2+3-6}{2*3+1} \quad = 2{,}14$	$\delta_1 = \lvert x_1 - x_0 \rvert = 0{,}86$
$x_2 = 2{,}14 \quad - \frac{2{,}14^2+2{,}14-6}{2*2{,}14+1} \quad = 2{,}003$	$\delta_2 = \lvert x_2 - x_1 \rvert = 0{,}137$
$x_3 = 2{,}003 - \frac{2{,}003^2+2{,}003-6}{2*2{,}003+1} = 2{,}00001$	$\delta_3 = \lvert x_3 - x_2 \rvert = 0{,}0029$

Portanto como $\delta_3 = 0{,}0029 \leq 0{,}020$, o resultado é $\cong$ 2,00.

2.5.2 Broyden-Fletcher-Goldfarb-Shanno (BFGS)

O método Broyden Fletcher Goldfarb Shanno é classificado como método quasi-Newton. A ideia por trás dos métodos quasi-Newton é fazer uma aproximação iterativa da inversa da matriz hessiana.

Dada uma função real de várias variáveis, $f: R^n \rightarrow R$ a matriz jacobiana do gradiente é denominada de *matriz hessiana*. Ela é sempre quadrada e seu determinante é denominado *função hessiana* (MASSAGO, 2010).

Na figura 4 é demonstrado um pseudocódigo de um ciclo básico do BFGS exemplificando seu processo de busca pelo parâmetro.

Figura 4 – Pseudocódigo do ciclo básico do BFGS.

```
1  inicio
2      Escolhe x_0;
3      B_0 = I;
4      g_0 = ∇f(x_0);
5      se ∇f(x_0) == 0 então
6          x̄ = x_0;
7          pare;
8      fim
9      enquanto critério de parada não satisfeito faça
10         d_i = B(x_i)^-1 ∇f(x_i);
11         Calcula o escalar t_i através de uma busca linear;
12         x_{i+1} = x_i + t_i d_i;
13         g_{i+1} = ∇f(x_{i+1});
14         Calcula B(x_{i+1})^-1 usando x_{i+1}, x_i, g_{i+1}, g_i e B(x_i)^-1;
15         i = i + 1;
16     fim enquanto
17 fim
```

Fonte: Xavier, 2009.

Para o caso de uma dimensão, $f''(x_{i+1})$ pode ser aproximado por:

$$f''(x_{i+1}) = \frac{f'(x_{i+1}) - f'(x_{i+1} - x_i)}{x_{i+1} - x_i} \qquad (14)$$

Ou seja, podemos nos aproximar do método de Newton sem precisar calcular $f''(x_{i+1})$. Da mesma forma, em dimensões maiores, B é uma aproximação de H (Matriz Hessiana) se:

$$B(x_{i+1} - x_i) = \nabla f(x_{i+1}) - \nabla f(x_i)\ , \qquad (15)$$

onde, ∇f é o gradiente da função.

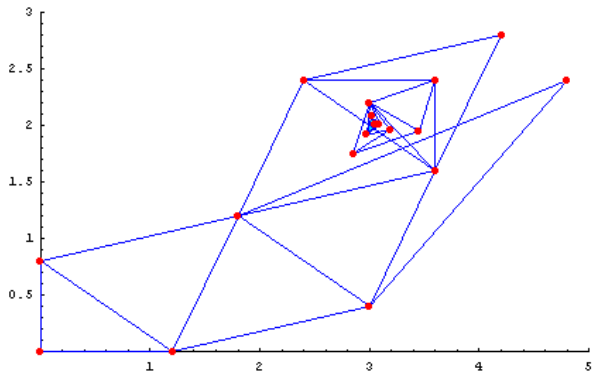
A equação (27) não é suficiente para definir B unicamente. Porém, o método BFGS, assim como outros métodos Quase-Newton, calcula uma aproximação B_{i+1} que satisfazem (27) e utiliza apenas as informações de $\nabla f(x_{i+1}), \nabla f(x_i)$, x_{i+1} e B_i (XAVIER, 2009).

2.5.3 Nelder-Mead

O método foi desenvolvido em meados de 1960 por Nelder-Mead possui esse nome devido à John Nelder & Roger Mead, e consiste de um método numérico para otimização de problemas livres multidimensionais, pertencentes à classe mais geral de algoritmos de busca (OZYMANDIAS, 2009). Na Figura 5 é apresentado um exemplo da iteração final de uma busca usando o método de Nelder-Mead.

O método usa o conceito do método de otimização simplex, que é um padrão de *N* + 1 vértices, onde N é o número de dimensões. Como apresentado na Figura 5, ele se move através da superfície a ser explorada, algumas vezes mudando o tamanho, mas sempre conservando a mesma forma (WRIGHT, 2012).

Figura 5 – Exemplo de busca pelo método de Nelder-Mead.



Fonte: Adaptado de Fullerton (2003).

Segundo Gao e Han (2010), o algoritmo de Nelder-Mead é a busca direta mais utilizada para resolver problemas de otimização sem restrições, ou seja,

$$minimizar\ f\ (x) \tag{16}$$

onde $f: R_n \rightarrow R$ é chamado de função objetivo, e n é a dimensão do problema. Um simplex é uma figura geométrica em n dimensões, representando o fecho convexo de $n + 1$ vértices. O Simplex será denotado com vértices $x_1, x_2, \ldots, x_{n+1}$ por Δ.

De acordo com Gao e Han (2010), o método iterativo de Nelder-Mead gera uma sequência de pontos para aproximar um ponto ótimo da função objetivo da equação 28. A cada iteração, os vértices $\{x_j\}_{j=1}^{n+1}$ do simplex são ordenados de acordo com os valores da função objetivo:

$$f\ (x_1) \leq\ f\ (x_2) \leq \cdots \leq\ f\ (x_{n+1}) \tag{17}$$

Referindo a x_1 como o melhor vértice, e x_{n+1} como o pior vértice. Se vários vértices têm os mesmos valores para f, regras de desempate consistentes, como as indicadas por Lagarias (1998) são necessárias para que o método seja bem definido.

Gao e Han (2010) afirmam que o algoritmo utiliza quatro possíveis operações: reflexão, expansão, contração, e encolhimento, sendo cada um deles associado com um parâmetro escalar: α (reflexão), β (expansão), γ (contração) e δ (encolhimento). Os valores destes parâmetros satisfazem $\alpha > 0,\ \beta > 1, 0 < \gamma < 1\ e\ 0 < \delta < 1$. Na implementação padrão do método de Nelder-Mead, os parâmetros são:

$$\{\alpha, \beta, \gamma, \delta\} = \{1, 2, 1/2, 1/2\} \tag{18}$$

Seja $\bar{x}$ o baricentro dos melhores n vértices. Então $\bar{x}$ é descrito por:

$$\bar{\mathrm{x}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i} \tag{19}$$

Na Figura 6 encontra-se um pseudocódigo para realizar uma iteração do algoritmo de Nelder-Mead segundo a versão apresenta por Lagarias (1998).

Figura 6 - Pseudocódigo do método de Nelder-Mead.

1. **Ordenar**: Avalie f nos $n+1$ vértices de Δ e classifique os vértices de modo que satisfaça (17).

2. **Reflexão:** Calcule o ponto de reflexão x_r para:

$$x_r = \bar{x} + \alpha(\bar{x} - x_{n+1})$$

   Avalie $f_r = f(x_r)$.
   Se $f_1 \leq f_r < f_n$, substitua x_{n+1} por x_r

3. **Expansão**: Se $f_r < f_1$ então calcule a expansão do ponto x_e para:

$$x_e = \bar{x} + \beta(x_r - \bar{x})$$

   Avalie se $f_e = f(x_e)$.
   Se $f_e < f_r$, substitua x_{n+1} por x_e;
   Caso contrário substitua x_{n+1} por x_r.

4. **Contração lado de fora:** Se $f_n \leq f_r < f_{n+1}$, calcule o ponto de contração de fora:

$$x_{oc} = \bar{x} + \gamma(x_r - \bar{x})$$

   Avalie $f_{oc} = f(x_{oc})$.
   Se $f_{oc} \leq f_r$, substitua x_{n+1} por x_{oc};
   Caso contrário, vá para etapa 6.

5. **Contração lado de dentro**: Se $f_r \geq f_{n+1}$, calcule o ponto de contração de dentro x_{ic} para:

$$f_{ic} = \bar{x} - \gamma(x_r - \bar{x})$$

   Avalie $f_{ic} = f(x_{ic})$.
   Se $f_{ic} < f_{n+1}$, substitua f_{n+1} por x_{ic};
   Caso contrário, vá para etapa 6.

6. **Encolhimento**: Para $2 \leq i \leq n+1$, definir $x_i = x_1 + \delta(x_i - x_1)$.

Fonte: Lagarias (1998).

2.6 Testes não paramétricos

A designação “teste não paramétrico” se refere ao fato de não ser necessário especificar a distribuição da população de onde provém a amostra. Os métodos não paramétricos usam procedimentos que são aplicáveis independentemente da distribuição da população. Raramente são exigidas algumas hipóteses como a de simetria ou a de continuidade da distribuição.

Alguns destes métodos se aplicam a dados qualitativos. Outra situação em que os testes não paramétricos são muito úteis é quando a amostra possui uma dimensão muito pequena e não se conhece a distribuição exata da população (COIMBRA, 2005).

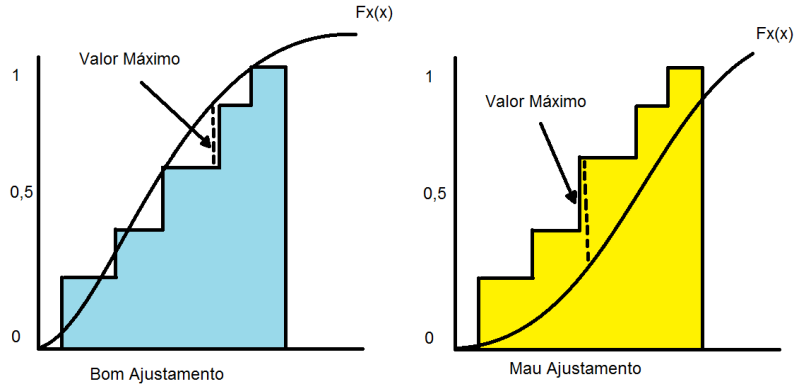
2.6.1 Teste de ajustamento Kolmogorov – Smirnov

O teste de Kolmogorov-Smirnov consistem em averiguar se uma amostra pode ser considerada como proveniente de uma população que possui uma determinada distribuição (COIMBRA, 2005).

Ele é largamente utilizado para análise de aderências de distribuições em estudos climáticos. Ele pode ser usado tanto para dados agrupados quanto para dados individuais. Nos dados agrupados não há restrição quanto ao número nem ao valor das classes. É baseado no módulo da maior diferença entre a probabilidade observada e a estimada, que é comparada com um valor tabelado de acordo com o número de observações da série sob teste. Isto evita o aspecto cumulativo dos erros (CATALUNHAA et al., 2002).

O princípio deste teste baseia-se na comparação da curva da frequência cumulativa dos dados com a função de distribuição teórica em hipótese. Quando as duas curvas se sobrepõem a estatística de teste é calculada através da máxima diferença entre ambas, como mostrado na figura 7 que apresenta dois exemplos, um de bom ajustamento e outros de mal ajustamento.

Figura 7 – Exemplo de bom e mau ajustamento.



Se os dados experimentais se afastam significativamente do que é esperado segundo a distribuição em hipótese, então as curvas obtidas devem

encontrar-se igualmente afastadas, e por um raciocínio análogo, se o ajustamento ao modelo hipotético é admissível, as curvas têm um delineamento próximo (COIMBRA, 2005).

De acordo com Silva e Ferreira (2009), o teste pode ser calculado da seguinte forma:

$$D = \max(D^{+}, D^{-}) \tag{20}$$

Em que

$$D^{+} = \max_{i=1,\ldots,n} \left(\frac{i}{n} - p_{(i)}\right) \tag{21}$$

$$D^{-} = \max_{i=1,\ldots,n} \left(p_{(i)} - \frac{i-1}{n}\right) \tag{22}$$

$$p_{(i)} = \varphi\left(X_{(i)} - \frac{\overline{X}}{S}\right) \tag{23}$$

e φ é a função de distribuição acumulada da distribuição normal padrão, e $\overline{X}\ e\ S$ são respectivamente a média e o desvio padrão dos valores da amostra.

2.6.2 Critério de Informação de Akaike (AIC)

Akaike(1974), mostrou que o viés é dado assintoticamente por p, onde p é o número de parâmetros a serem estimados no modelo, e definiu seu critério de informação como:

$$AIC = -2ln(mv) + 2p \tag{25}$$

Em que: p = número de parâmetros do modelo; mv = valor da máxima verossimilhança.

Segundo Mendonça et al. (2011), quanto menor o valor de AIC melhor o modelo avaliado.

O critério de AIC assim como o teste de Kolmogorov-Smirnov, pode ser usado para avaliar o quão bom foi o ajuste dos dados para dada distribuição, buscando o menor valor de AIC calculado a partir dos parâmetros encontrados para o modelo.

2.6.3 Critério de Informação Bayesiano (BIC)

O Critério de Informação Bayesiano (BIC) é um critério que penaliza a verossimilhança, para que um modelo mais adequado seja selecionado. E Schwarz (1978) propôs a expressão, onde é dado por:

$$BIC = -2ln(mv) + pln(n) \tag{26}$$

Em que: n = número de observações.

De forma análoga ao AIC, quanto menor o valor de BIC melhor o modelo avaliado (MENDONÇA *et al.*, 2011) . Assim o critério de BIC pode ser usado para avaliar o quão bom foi o ajuste dos dados para dada distribuição.

2.6.4 Teste de Normalidade Shapiro-Wilk

O teste Shapiro-Wilk, calcula uma variável estatística (W) que investiga se uma amostra aleatória provém de uma distribuição normal (SCUDINO, 2008). A variável W é calculada da seguinte forma:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \tag{27}$$

sendo,

- x_i os valores ordenados de amostras (x_1 é o menor).

- a_i constantes geradas a partir de meio, variâncias e covariâncias da ordem estatística de uma amostra de tamanho n e uma distribuição normal.

2.6.5 Teste t de Student

O teste t é uma técnica estatística usada quando podemos assumir que duas distribuições possuem a mesma variância.. Esse teste é extremamente usando para comparações de dois grupos independentes (HÉLIO; BITTENCOURT, 2006).

Apresentação das hipóteses:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_a: \mu_1 = \mu_2 \end{cases} \tag{28}$$

A estatística do teste tem a formula apresentada pela equação a seguir.

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{(n_1 + n_2 - 2)} \times \left(\frac{1}{\mathbf{n}_1} + \frac{1}{\mathbf{n}_2}\right)}} \tag{29}$$

As regiões de rejeição e aceitação são verificadas pelo p-valor que apresenta o resultado da estatística, e comparada com uma significância adotada. Através desse valor p é possível comparar com a significância, onde geralmente é usada $\alpha = 0{,}05$. Se o valor de p for menor que a significância, rejeita-se a hipótese. Caso contrário aceita-se (HÉLIO; BITTENCOURT, 2006).

2.6.6 Teste de Mann-Whitney (Wilcoxon)

O teste de Mann-Whitney foi desenvolvido primeiramente por F. Wilcoxon em 1945, para comparar tendências centrais de duas amostras independentes de tamanhos iguais. Em 1947, H.B. Mann e D.R. Whitney generalizaram a técnica para amostras de tamanhos diferentes. O teste de Mann-Whitney (Wilcoxon rank-sum test) é indicado para comparação de dois grupos não pareados para se verificar se pertencem ou não à mesma e cujo os dados não apresentem uma distribuição normal da população. Na verdade, verifica-se se há evidências para acreditar que valores de um grupo A são superiores aos valores do grupo B (ACTION, 2014).

Para realizar o Teste de Wilcoxon Pareado deve-se primeiramente estabelecer as hipóteses:

$$\begin{cases} H_0 : \Delta = 0 \\ H_1 : \Delta \neq 0 \end{cases} \quad \begin{cases} H_0 : \Delta = 0 \\ H_1 : \Delta > 0 \end{cases} \quad \begin{cases} H_0 : \Delta = 0 \\ H_1 : \Delta < 0 \end{cases} \tag{30}$$

Ou seja, estará testando se as populações diferem em localização ou não utilizando a seguinte ideia:

- Caso a hipótese nula seja aceita, temos que a mediana da diferença é nula, ou seja, as populações não diferem em localização. Já, se a hipótese nula for rejeitada, ou seja, se a mediana da diferença não for nula, temos que as populações diferem em localização.

Dessa forma, o teste se torna um Teste de Wilcoxon para uma única amostra, menos a amostra $D_1, ..., D_n$ onde $\theta_0 = 0$ (ACTION, 2014).

3. METODOLOGIA

A metodologia do trabalho descreve detalhadamente os passos, tecnologias e os métodos utilizados para estar alcançando o objetivo geral e os objetivos específicos propostos na sessão 1.1.1 e 1.1.2.

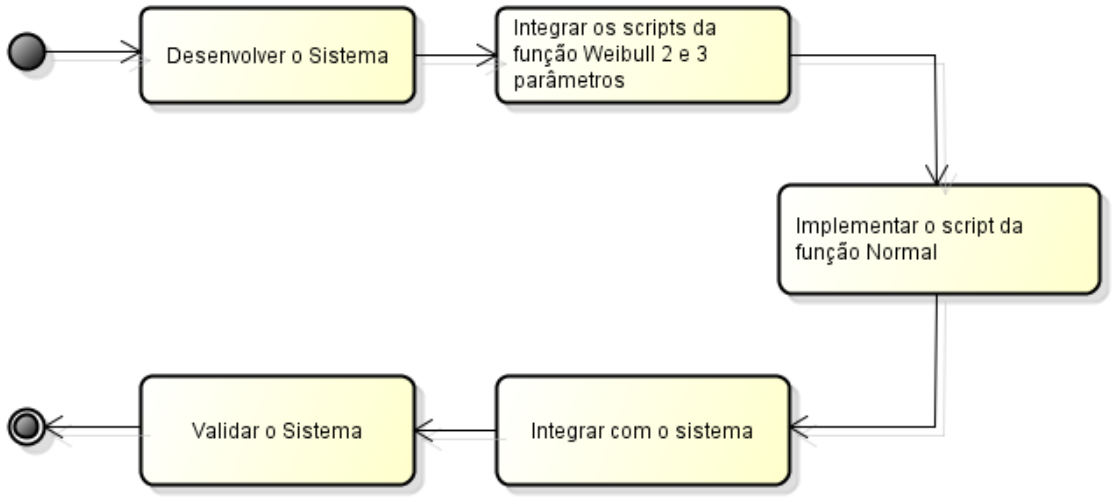
Devido a distribuição Weibull possuir 3 parâmetros, esses parâmetros geram grande flexibilidade no ajuste da forma da curva para se adaptar a situação. Em contraste com a função Normal que possui apenas 2 parâmetros. Isso permite fazer a distribuição positivamente assimétrica, simétrica ou negativamente assimétrica, conforme desejado. Essa flexibilidade é a principal vantagem dessa distribuição (HILLIER; HILLIER, 2014).

Portanto sabendo da flexibilidade que a função Weibull possui, ela foi adotada para realizar o ajuste no sistema proposto. E sabendo que a função normal contrasta em relação a flexibilidade que a função Weibull fornece, também foi adota para realizar o ajuste no sistema proposto, podendo assim visualizar o ajuste com duas funções de distribuição diferentes.

3.1 Introdução

A construção do sistema de informação para realização da estimativa de parâmetros para as funções de probabilidade Weibull e Normal, se dividiu em seis etapas, apresentadas na figura 8.

Figura 8 – Fluxograma das etapas para construção do sistema proposto.



1. Na etapa de desenvolvimento foi implementado um sistema de informação utilizando a linguagem Java versão 1.7 (ORACLE, 2014), tendo como objetivo

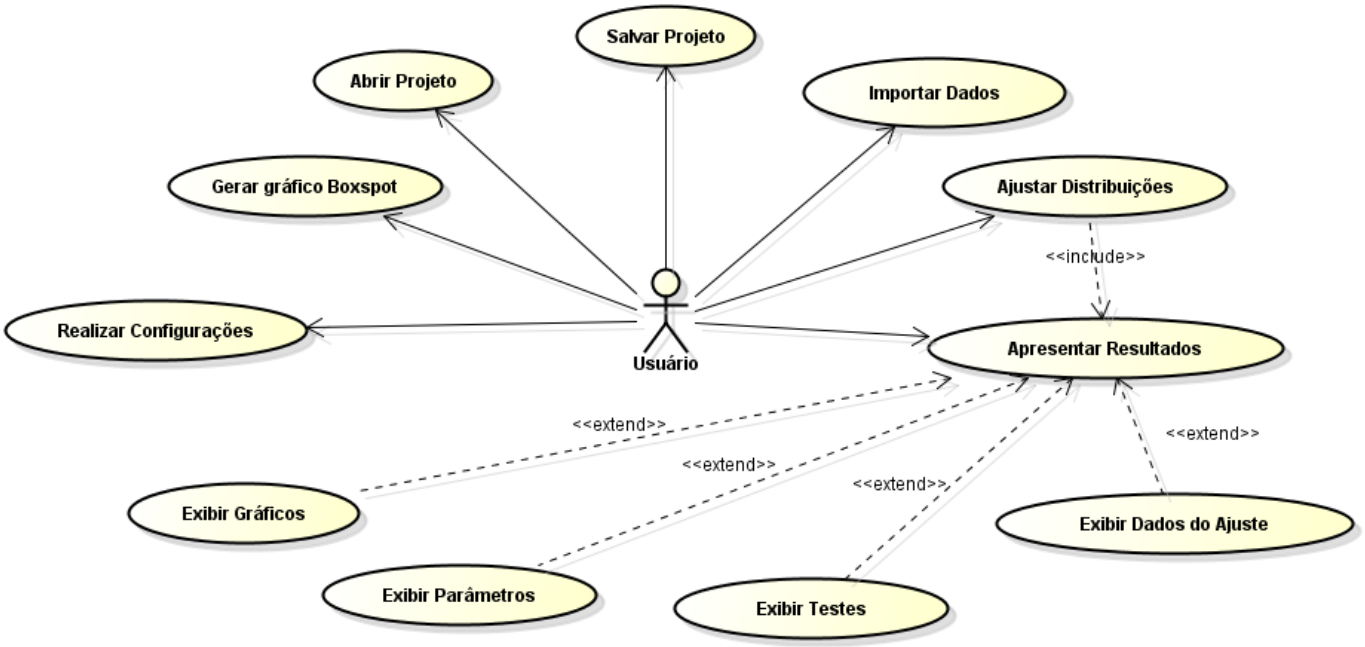
principal realizar a estimativa de parâmetros de funções. A visão geral dos casos de uso do sistema proposto é apresentada na Figura 9;

2. Na etapa de integração dos scripts Weibull 2 e 3 parâmetros foram usados os scripts desenvolvidos por Fraga Filho e Schmidt (2014), adaptados e integrados com o sistema proposto;
3. Na quarta etapa, foi desenvolvido o script para a função Normal;
4. Na quinta etapa foi realizada a integração do script da função Normal com o sistema proposto;
5. Por fim, na etapa de validação, foi realizado um estudo de caso para estimativa de parâmetros das distribuições Weibull 2 e 3 parâmetros e Normal, com dados de diâmetros de 8 áreas florestais em 3 idades diferentes ao sistema desenvolvido e comparado os resultados com o sistema estatístico SAS (SAS Institute Inc, 2000).

3.2 Visão geral do sistema desenvolvido

Nesta seção é apresentada a visão geral do sistema desenvolvido neste trabalho. A Figura 9 apresenta o diagrama de casos de uso do sistema de informação batizado com o nome de Estima.

Figura 9 – Diagrama de casos de uso do sistema Estima.

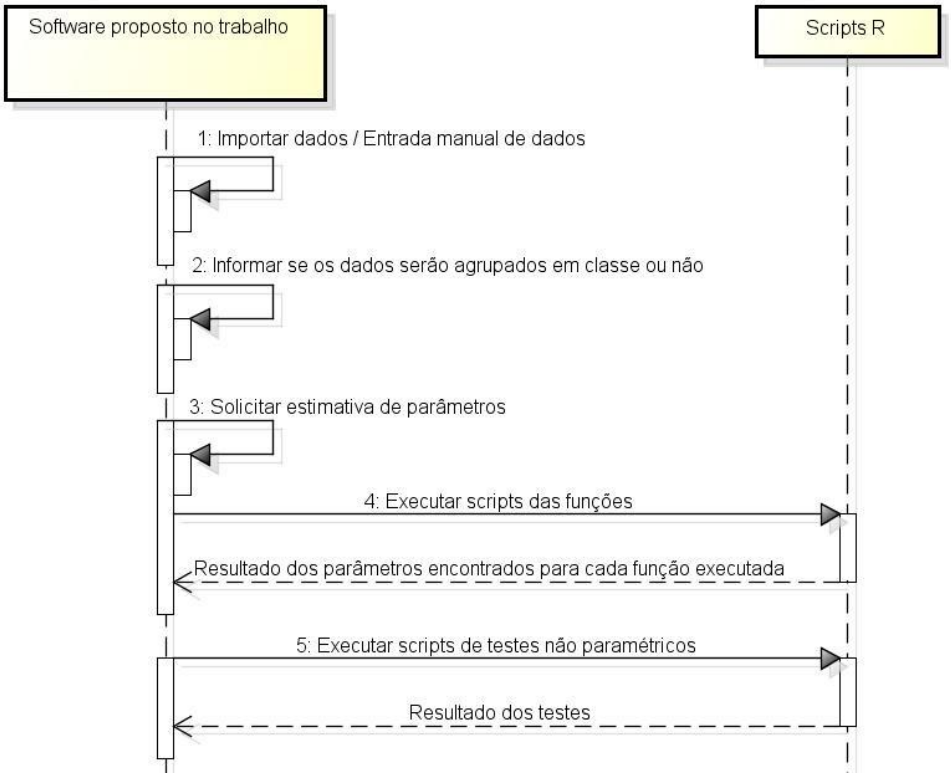


A Figura 9 apresenta 11 funcionalidades, sendo que os resultados englobam quatro delas.

O fluxo geral para realizar o ajuste de distribuição é primeiramente importar um arquivo com os dados nos formatos permitidos. Posteriormente configurar os parâmetros do ajuste e confirmar. O sistema atualmente possui duas funções para ajustes, a Weibull 2 e 3 parâmetros e a Normal. Sendo que a Weibull pode ser completa, trucada à esquerda, trucada à direita, ou ambos. A função Normal implementada e integrada ao software desenvolvido neste trabalho não possui truncamento, só será ajustada na modalidade completa. Após ser realizado o ajuste, serão apresentados os resultados, sendo possível visualizá-los individualmente para cada grupo ou todos os resultados juntos.

Os scripts da estimativa de parâmetros contém a chamada do pacote *optimx* responsável por receber a função de verossimilhança que foi deduzida das funções de probabilidade Weibull e Normal pelo método da máxima verossimilhança. O pacote *optimx* de forma iterativa tenta buscar os parâmetros para aquela função passada. Na Figura 10, é demonstrado o processo de integração dos scripts com o sistema proposto.

Figura 10 – Diagrama de Sequência de um exemplo de utilização do sistema.



Para realizar a estimativa de parâmetros, o usuário deverá ter alguma base de dados presente no sistema, seja fornecida manualmente pelo usuário ou importada de um arquivo externo. Em seguida o usuário terá a opção de utilizar a estimativa usando os dados brutos, ou seja, usando os dados da forma que estão sem nenhuma alteração, ou agrupá-los em classes, onde o usuário deverá informar a amplitude desejada. Realizando essas pré-configurações dos dados, ao solicitar o ajuste, o sistema Estima executará os scripts das distribuições solicitadas pelo usuário em uma sessão do software R (R DEVELOPMENT CORE TEAM, 2005) que irá realizar a estimativa de parâmetros para as funções retornando ao final da execução os parâmetros encontrados para cada função.

Posteriormente, como demostrado na Figura 10, o script dos testes, irá receber os dados e os parâmetros encontrados, sendo aplicado o teste para avaliar qual distribuição se adequa melhor aos dados. O usuário poderá escolher em qual teste se basear, dentre o KS, AIC, BIC e Anderson Darling (AD) para verificar qual distribuição foi melhor de acordo com cada teste, sendo do usuário a responsabilidade de decidir.

Como resultado do teste será retornado o nome da distribuição e os valores estatísticos encontrados de cada teste, conforme é apresentado na Figura 11 com dados para exemplificar.

Figura 11 – Interface gráfica dos resultados dos testes.

Grupo	Função	Método	Estatísti...	p-valor AD	BIC	Estatísti...	p-valor KS	AIC
Normal_3...	Normal	BFGS	6.23	0.0	410.61	0.53	0.0	409.52
Normal_3...	Normal	CG	5.62	0.0	380.09	0.53	0.0	379.0
Normal_3...	Normal	nlm	5.9	0.0	347.48	0.52	0.0	346.39
Normal_3...	Normal	Nelder-M...	0.63	0.64	92.99	0.12	0.9	91.9
Normal_3...	Normal	ucminf	0.51	0.76	92.99	0.12	0.9	91.9
Normal_3...	Normal	nlminb	2.15	0.07	92.99	0.12	0.9	91.9
Normal_3...	Normal	bobyqa	0.75	0.53	92.99	0.12	0.9	91.9
Normal_3...	Normal	newuoa	0.48	0.79	92.99	0.12	0.9	91.9
Normal_3...	Normal	nmkb	0.8	0.49	92.99	0.12	0.9	91.9
Normal_3...	Normal	L-BFGS-B	-	-	92.99	-	-	91.9
Normal_3...	Normal	spg	-	-	92.99	-	-	91.9
Normal_3...	Normal	L-BFGS-B	-	-	92.99	-	-	91.9
Normal_3...	Normal	spg	-	-	92.99	-	-	91.9
Normal_3...	Normal	nmkb	0.8	0.49	92.99	0.12	0.9	91.9
Normal_3...	Normal	newuoa	0.48	0.79	92.99	0.12	0.9	91.9
Normal_3...	Normal	bobyqa	0.75	0.53	92.99	0.12	0.9	91.9
Normal_3...	Normal	nlminb	2.15	0.07	92.99	0.12	0.9	91.9
Normal_3...	Normal	ucminf	0.51	0.76	92.99	0.12	0.9	91.9

Conforme é apresentado na figura 11, os parâmetros encontrados são exibidos para cada distribuição, como por exemplo, para a função Weibull 2P.

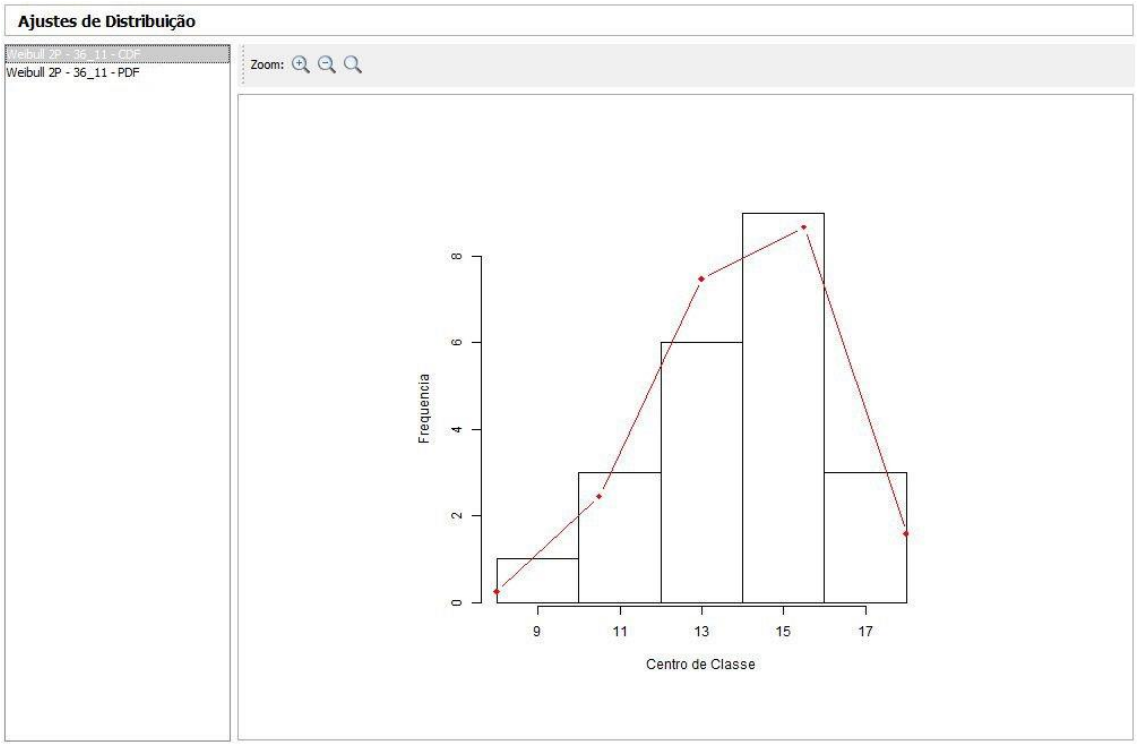
Figura 12 – Interface gráfica dos resultados dos parâmetros das funções.

Parâmetros Encontrados

nome	método	forma	escala
Weibull 2P	nmkb	10.31	16.01
Weibull 2P	newuoa	10.31	16.01
Weibull 2P	bobyqa	10.31	16.01
Weibull 2P	Nelder-Mead	10.31	16.01
Weibull 2P	Nelder-Mead	10.31	16.01
Weibull 2P	bobyqa	10.31	16.01
Weibull 2P	newuoa	10.31	16.01
Weibull 2P	nmkb	10.31	16.01

Os parâmetros encontrados para cada função serão exibidos como na interface gráfica da Figura 12, exibindo o nome da função de distribuição de probabilidade, o método usado do pacote *optimx* e seus parâmetros encontrados. Também é possível visualizar gráficos de densidade e gráficos de distribuição acumulada como na Figura 13 que apresenta o CDF da distribuição Weibull 2p, por exemplo.

Figura 13 – Interface gráfica dos gráficos das distribuições.



A partir da leitura do gráfico, é possível observar o quanto do eixo y (Frequência) pode indicar a probabilidade de uma observação particular (um valor de y) ocorrer dentro do intervalo de x (Classes). Sendo que as barras do gráfico representam um agrupamento de dados em classes, de tal forma que contabilizamos o número de ocorrências em cada classe. Já a linha feita unindo-se os pontos, corresponde aos pontos médios de cada classe.

Além disso, o sistema é capaz de “Realizar estatística descritiva”, ou seja, disponibilizar ao usuário os cálculos de média, mediana, variância, desvio padrão, valor máximo e valor mínimo dos dados analisados. Para realizar estatística descritiva foi implementado um script R, apresentado na Figura 14, contendo todos esses cálculos, bastando apenas executá-lo para obter os resultados.

Figura 14 – Script da estatística descritiva.

```
dadosEstat = summary(dfDados[,colunaEstatistica])
vetNomes = names(dadosEstat)
vetDadosEstat = matrix(dadosEstat)
varianca = var(dfDados[,colunaEstatistica],na.rm=T)
desvioPa = sd(dfDados[,colunaEstatistica],na.rm=T)
coefVariacao = round(desvioPa/vetDadosEstat[4],5)

# Nomes que serão apresentados no Software Estima para as estatísticas
vetNomes = c(vetNomes,"Variancia","Desvio Padrao","Coef. Var")

# A variável vetDadosEstat será resgatada pelo Software Estima e
# os valores serão apresentados respectivamente com os nomes de vetNomes'
vetDadosEstat = c(vetDadosEstat,varianca,desvioPa,coefVariacao)
```

Para adicionar novas estatísticas, basta apenas realizar os cálculos na sessão de cálculos estáticos demonstrada na Figura 14, informar o nome da estatística e adicionar a variável do cálculo no vetor de resultados. Dessa forma, o sistema Estima irá resgatar o vetor de nomes e resultados, e será apresentada a nova estatística junto com as outras, como na Figura 15.

Figura 15 – Interface gráfica das estatísticas realizadas.

Estatística Descritiva

Estatística Descritiva	Valor
Min.	6.2
1st Qu.	13.4
Median	15.6
Mean	15.64
3rd Qu.	17.7
Max.	26.5
Variancia	9.64
Desvio Padrao	3.1
Coef. Var	0.2

3.3 Implementação dos scripts e integração

Embora o software R (R DEVELOPMENT CORE TEAM, 2005) auxilie a manipulação de cálculos matemáticos, ele não possui todas as funções necessárias para realizar a estimativa de parâmetros proposta. Por isso existe a necessidade de implementar scripts na linguagem R (R DEVELOPMENT CORE TEAM, 2005) com essas funções. Esses scripts da função Weibull 2 e 3 parâmetros deduzidos pelo método da máxima verossimilhança desenvolvidas por Fraga Filho e Schmidt (2014) foram integrados com o sistema Estima. A função Normal também foi implementada usando o método da máxima verossimilhança e foi realizada a integração dos mesmos com o sistema.

Conforme apresentado na introdução, o sistema foi implementado pensando em uma forma fácil para inserção de novas funções de distribuição de probabilidade.

3.4 Integração de novos scripts para funções de distribuição

Para incorporar novas funções, é necessário entender o funcionamento da linguagem estatística R, e ter conhecimento do paradigma de programação orientado a objeto. O script deve ser escrito usando Orientação ao Objeto (OO), seguindo os padrões dos scripts WeibullClass.r e NormalClass.r que se

encontram na pasta “Script” na raiz da instalação do sistema. Na pasta “Script” existe um script R com nome configFuncoes.r, que é apresentado na Figura 16.

Figura 16 – Script de configuração das funções.

```
#Importe suas funções aqui source("nomeDoArquivo.r")
listFuncoesNomes = c("Weibull 2P","Weibull 3P","Normal");
listFuncoes = list(weibullFactory,weibull3pFactory,normalFactory);
posParametros = c("2;3","2;3;4","2;3");
```

No script configFuncoes.r é necessário realizar 4 operações. Primeiramente importar os scripts referentes à nova função que se deseja adicionar. Após importado os scripts, lembrando que deve ser importado antes das linhas já existentes, será necessário informar o nome da função. Esse nome deve ser colocado na variável listFuncoesNomes após o último nome. Ele será usado para gerar os resultados da função. Posteriormente informe na variável listFuncoes o nome da função que irá realizar o ajuste, após o último nome. E por último na variável posParametros, a posição onde os parâmetros encontrados e exportado para o csv se encontram. Veja melhor na Figura 17.

Figura 17 – Exemplo de arquivo csv dos resultados obtidos para função Normal

obj.nome	metodo	media	desvioPadrao	KS	KS_p_value	KS_rejeita
Normal	BFGS	117,5972114	1590,707639	0,524912158	0	TRUE
Normal	nlm	73,67129186	997,7341623	0,522168679	0	TRUE

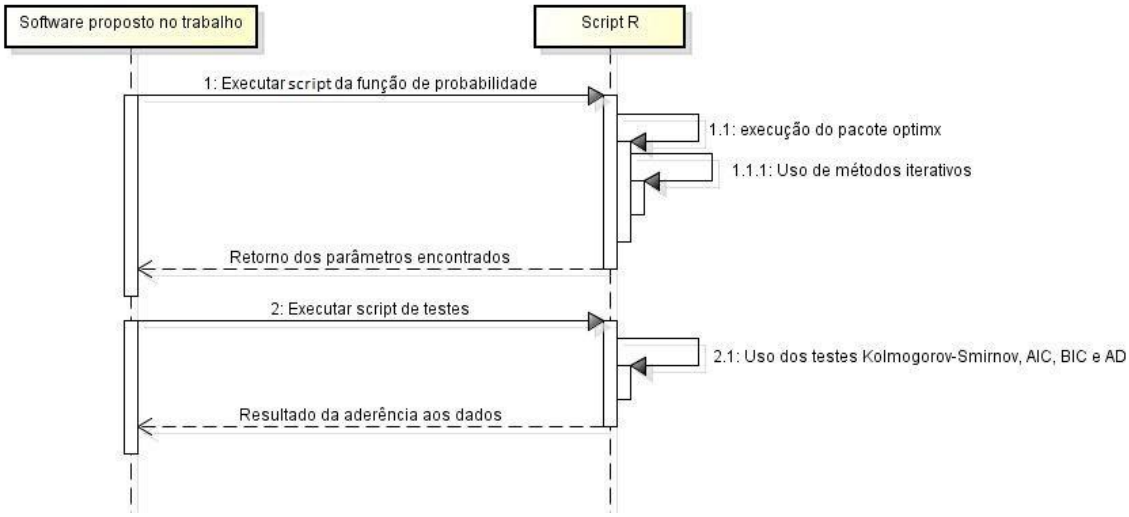
Como é demonstrado na Figura 17, os parâmetros encontram-se na posição 2 e 3 começando do zero. São essas posições que devem ser informadas na variável posParametros. Lembrando que deve ser mantida a ordem de cada nome, função e posição respectivamente.

Realizadas essas configurações, o software irá pegar as funções adicionadas e executa-las.

3.5 Execução dos scripts

A execução dos scripts é apresentada na Figura 18, onde é possível visualizar os passos realizados para execução de um script R.

Figura 18 – Diagrama de sequência da execução do script



Quando o sistema proposto realizar a execução do script de uma função de probabilidade, o script irá executar a função do pacote *optimx* (NASH; GROTHENDIECK; VARADHAN, 2014) responsável por receber a dedução da função de distribuição de probabilidade. Por meio de métodos iterativos presente no pacote *optimx*, cada algoritmo de otimização tentará, a partir de chutes iniciais para a função de máximo verossimilhança, encontrar os parâmetros que expliquem o comportamento dos dados.

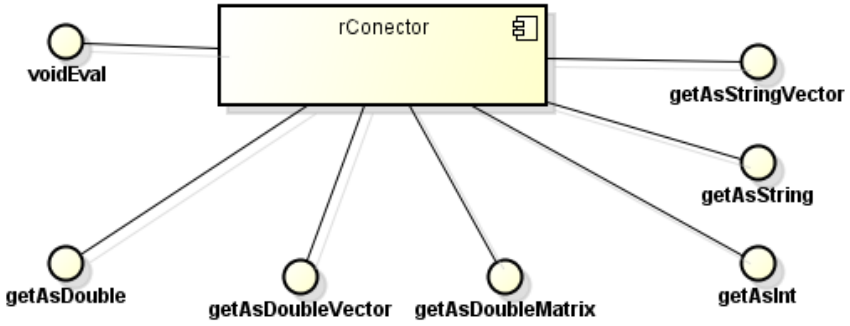
Depois de encontrado os parâmetros para cada um dos algoritmos numéricos executados, o sistema proposto irá executar o script de testes, onde por meio dos testes de Kolmogorov-Smirnov, AIC, BIC e AD, o usuário poderá escolher por qual dos testes se basear, sendo possível verificar qual das distribuições apresentou melhor aderência aos dados.

3.6 Integração do Estima com o sistema estatístico R

Para executar todos os scripts citados anteriormente o sistema proposto necessita comunicar-se com a aplicação R (R DEVELOPMENT CORE TEAM, 2005), sendo necessário um intermediador para realizar essa conexão.

O intermediador que foi usado é um plugin que foi desenvolvido para o software Forest Measure (FRAGA FILHO, *et al.*, 2014), cujo nome é “rConector”. Sua interface pública é apresentada na figura 19.

Figura 19 – Diagrama de componentes do rConector.

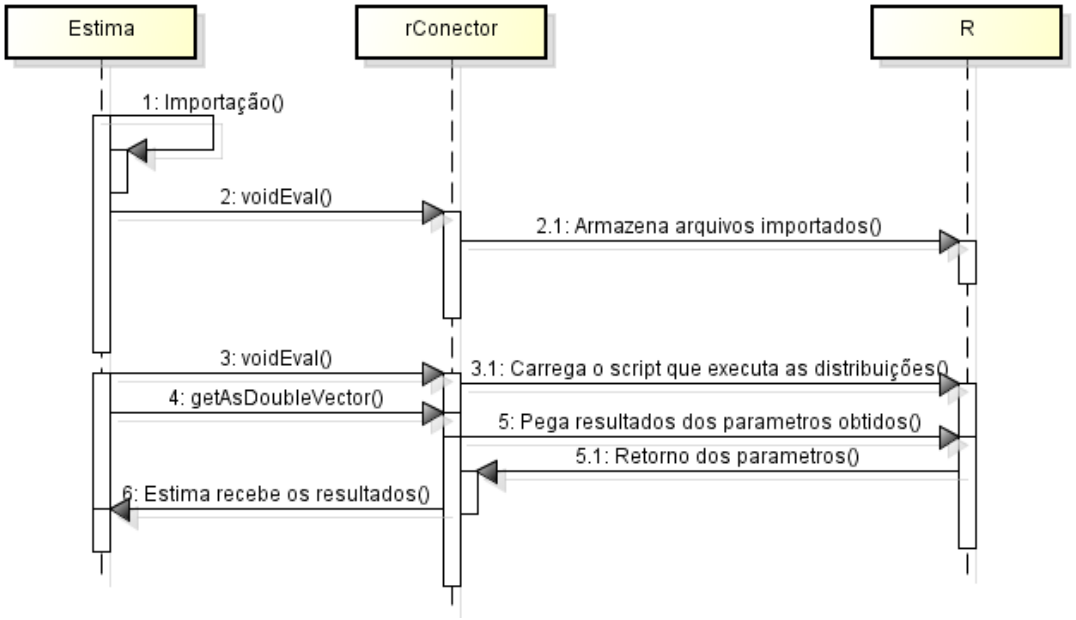


O rConector disponibiliza 7 métodos públicos para manipulação dos dados diretamente no software R, onde por meio do método *voidEval()* é possível executar qualquer comando diretamente na instancia do R, e pelos métodos *get()* recuperar as informações da instancia em execução.

Por meio deste plugin, é possível realizar a execução dos scripts, e retornar as informações sobre os dados em tempo de execução, buscar resultados obtidos, ou seja, manipular o software R (R DEVELOPMENT CORE TEAM, 2005) através do sistema proposto via interface pública do rConector.

Para demonstrar um dos momentos onde o rConector é usado no Estima, o diagrama de sequência apresentado na Figura 20, exemplifica o momento onde é executado o ajuste das distribuições.

Figura 20 – Diagrama de sequência do uso rConector.



Após realizar a importação dos dados, é usado o método sem retorno *voidEval()* onde é enviado os dados para o R, e armazenados em uma variável. Posteriormente é usado novamente o método *voidEval()* para mandar executar o script já criado anteriormente, que contém a rotina para executar o ajuste. Por fim, depois que o script terminar sua execução, o Estima utiliza o método *getAsDoubleVector()* para buscar um vetor com os resultados dos parâmetros.

3.7 Apresentação dos dados usados no estudo de caso

O objetivo da validação, foi comparar os resultados da estimativa de parâmetros realizada no Estima e no SAS (SAS Institute Inc, 2000) versão para estudantes, para as funções Weibull 2 e 3 parâmetros e Normal, buscando avaliar se são estatisticamente iguais.

Para realizar a validação, foi feito um estudo de caso com dados provenientes da área florestal, usados no trabalho de Schmidt (2014). Os dados são referentes a um plantio homogêneo com o híbrido obtido a partir do Eucalyptus urophylla e Eucalyptus grandis, localizado no município de Guaçuí, região sul do Espírito Santo, a 622 metros de altitude, com coordenadas centrais aproximadas no Sistema de Projeção Cartográfica UTM, Datum SAD 69, Zona 24S iguais a: 220.051 m e 7.690.288 m (Schmidt, 2014).

Devido ao fato de ser impossível medir vários diâmetros ao longo do fuste de uma árvore, ou seja, da base até a ponta da árvore, a medida mais utilizada refere-se ao diâmetro com casca à altura do peito, denominado DAP que é realizada na altura de 1.30 metros (DAP Florestal, 2014).

Dentre as várias razões para se usar o DAP, as que mais se destacam é o fato de ser uma medida fácil de ser avaliada quando comparada com outras características das árvores, são medidas mais confiáveis, ou seja, os erros de medição e suas causas são reconhecidos e podem ser limitados a um valor mínimo pela utilização de instrumentos apropriados, pela utilização de métodos de medição adequados e pelos cuidados nas tomadas das medidas (DAP Florestal, 2014).

A base de dados utilizada para realizar o estudo de caso possui 3 informações relativas a árvore em questão. Como apresentado na Tabela 2, é fornecida a idade em que a árvore se encontra (em meses), sua DAP e a qual parcela ela pertence.

Tabela 2 – Exemplo de representação dos dados utilizados no estudo de caso.

IDADE (MESES)	DAP	PARCELA
36	9.8	3
60	10.1	3
60	10.2	4
72		

Os dados apresentam um total de 567 árvores, sendo que estão divididas em três idades diferentes dispostas nas seguintes parcela apresentadas na Tabela 3.

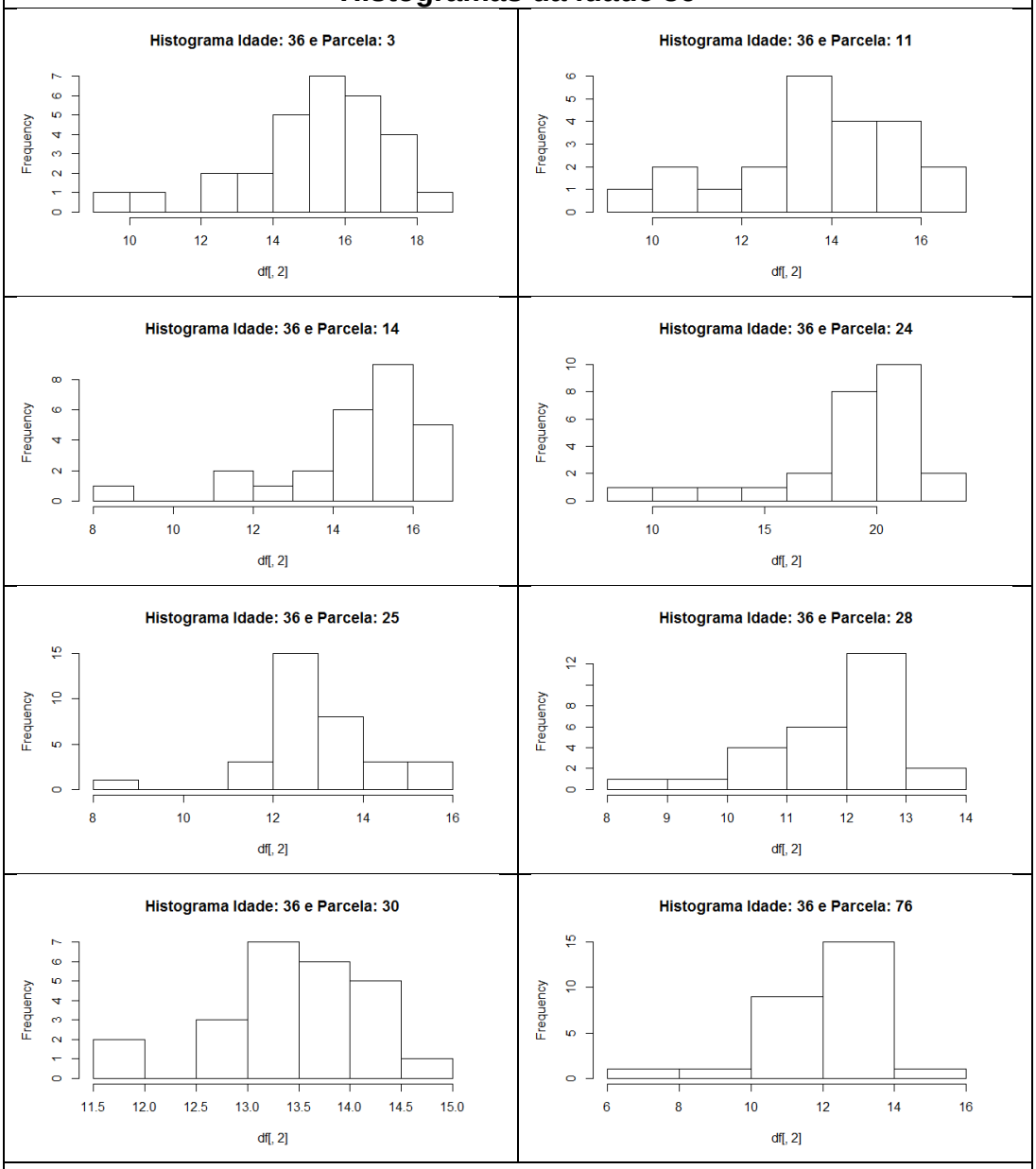
Tabela 3 – Quantidade de árvores por Parcela x Idade.

PARCELA	IDADE		
	36	**60**	**72**
3	29	24	29
11	22	23	23
14	28	28	28
24	34	33	33
25	33	33	33
28	27	27	27
30	24	24	24
76	27	27	26

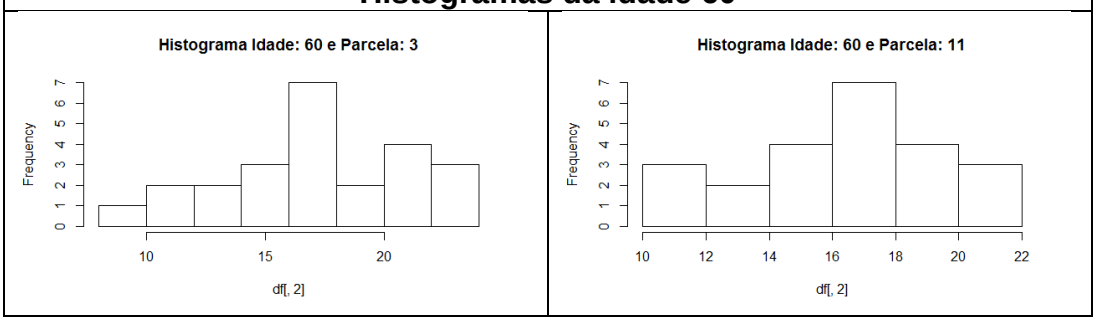
De acordo com a Tabela 4 os histogramas apresentam os dados.

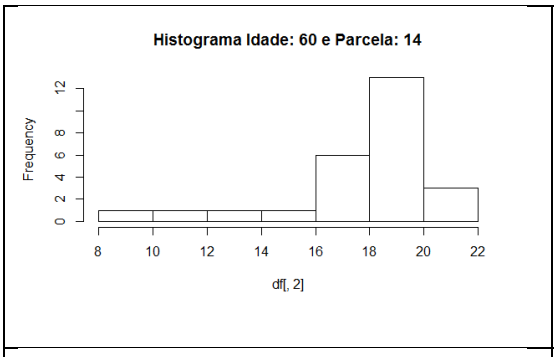
Tabela 4 – Histogramas dos dados referente a cada Idade x Parcela.

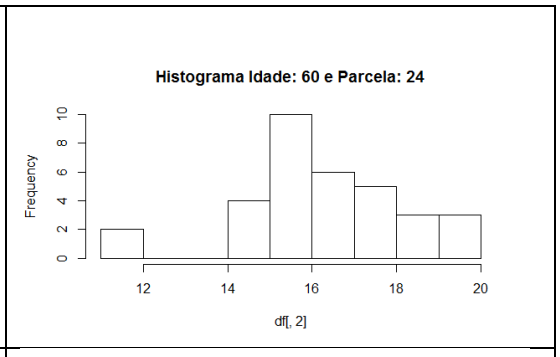
Histogramas da idade 36

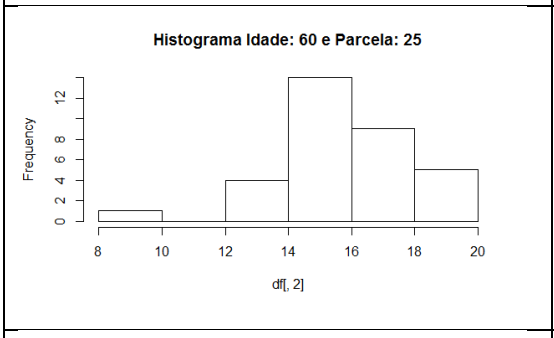


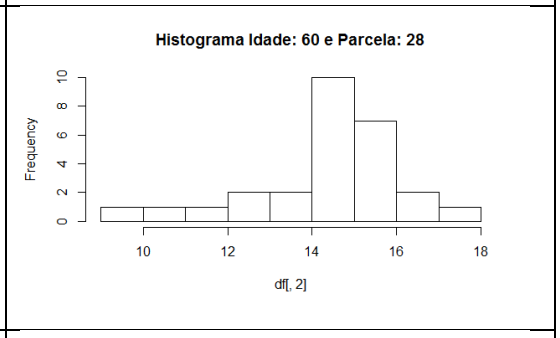
Histogramas da idade 60

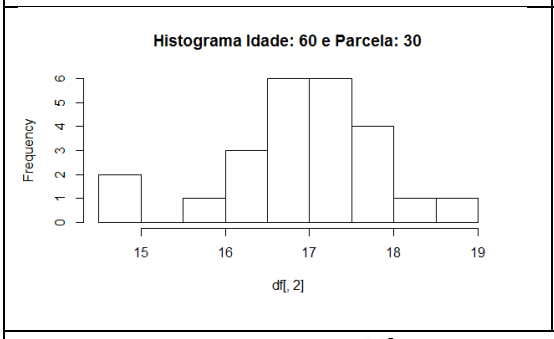


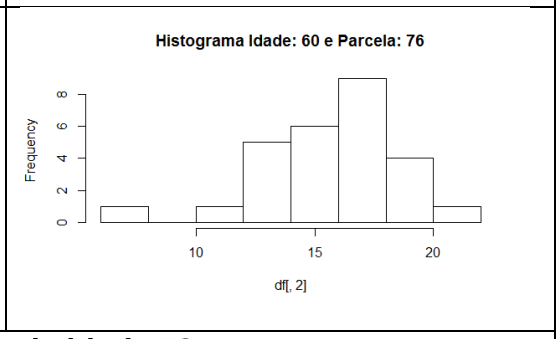




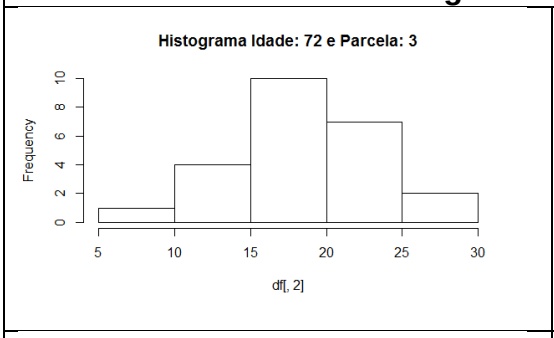


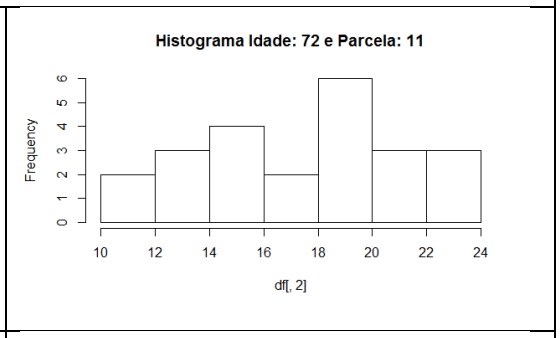


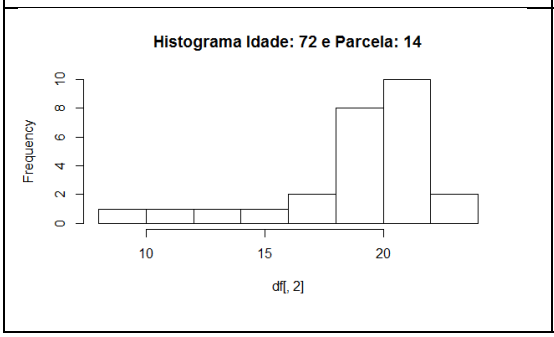


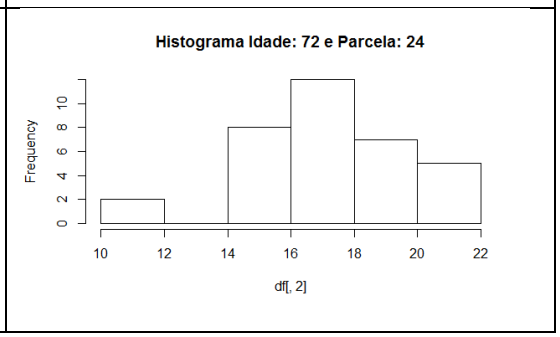


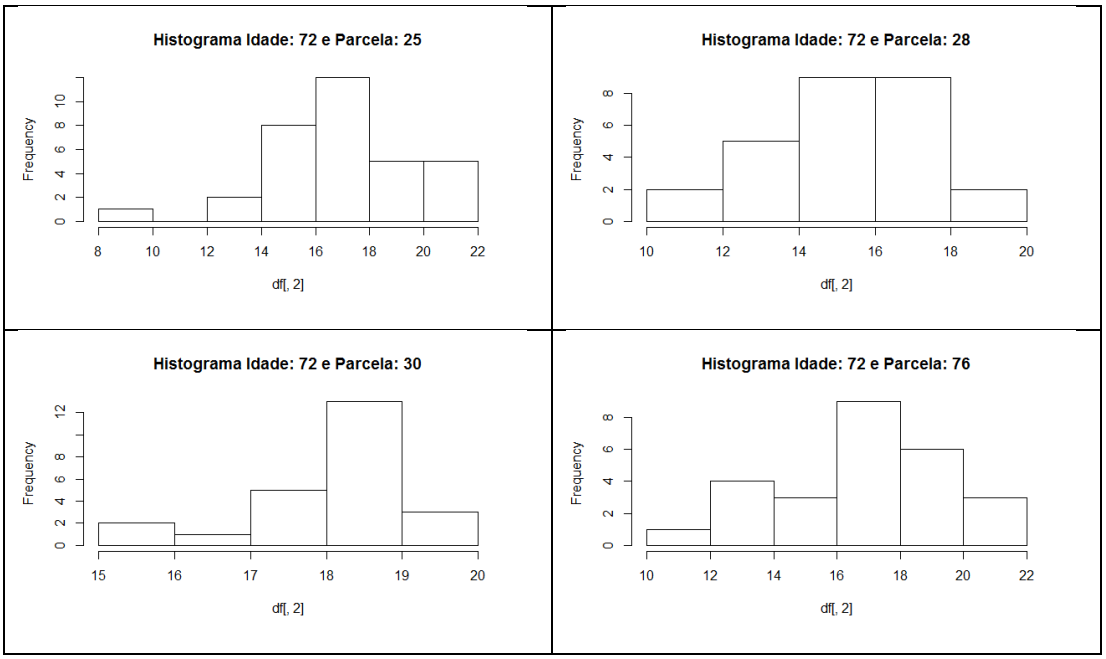
Histogramas da idade 72











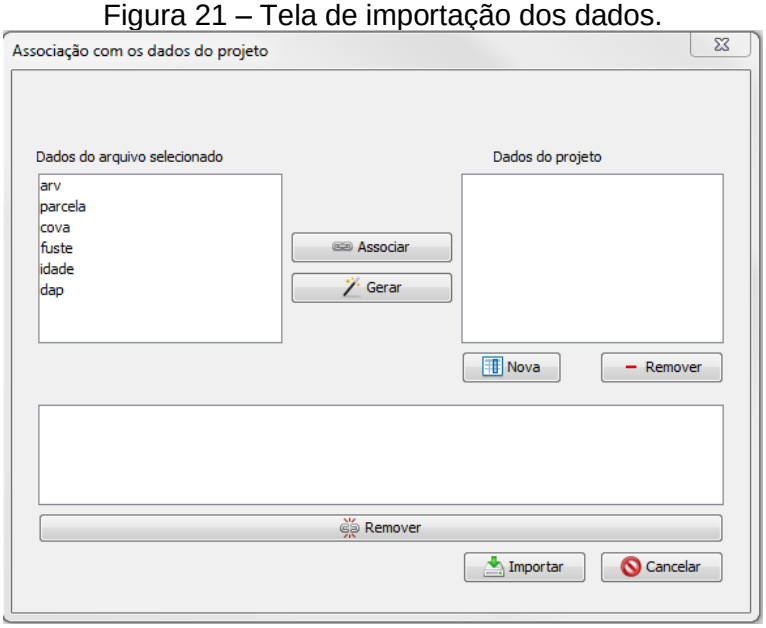
3.8 Estudo de Caso

Com o propósito de avaliar a qualidade dos dados obtidos no software Estima, foi realizado um estudo de caso usando os dados provenientes da área florestal apresentados na sessão 3.7 da metodologia.

Para realizar o estudo de caso primeiramente foi importado os dados para o Estima, usando uma fonte de formato CSV (Comma-separated values).

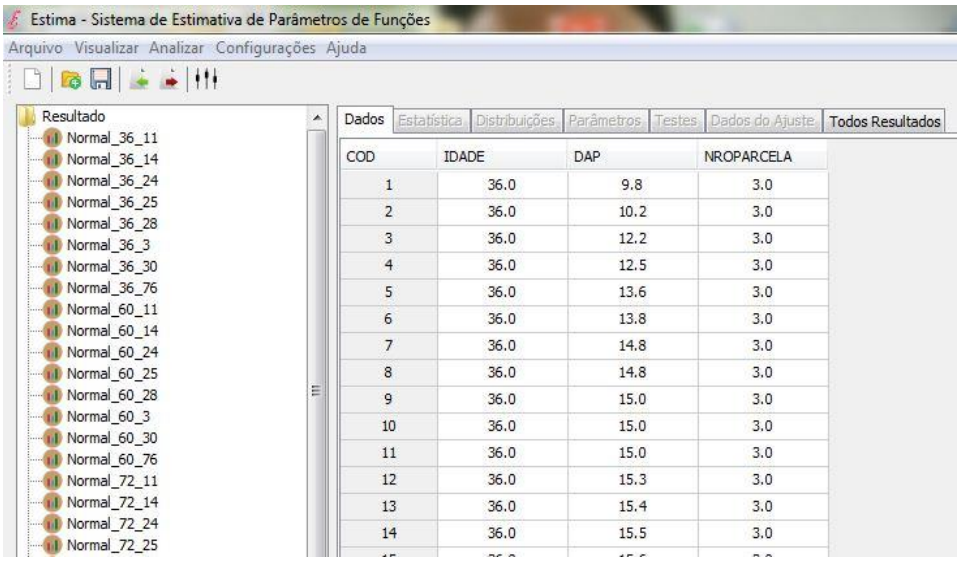
A Figura 21 apresenta a interface gráfica aberta após selecionar o arquivo que deseja importar para que seja informado quais serão as colunas que irão fazer parte do projeto.

Figura 21 – Tela de importação dos dados.



Após importados os dados para o sistema já será possível usar algumas funcionalidades como gerar gráficos *Boxplot*, realizar estatística descritiva dos dados e consequentemente o ajuste dos dados. Com a importação dos dados realizada, a interface principal do sistema irá apresentar os dados, sendo possível dar prosseguimento no ajuste, como apresentado na figura 22.

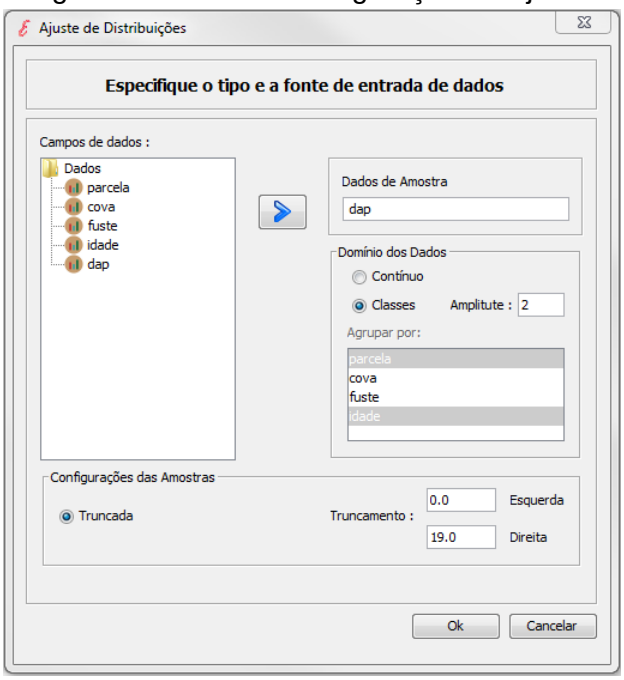
Figura 22 – Dados importados para o sistema Estima.



COD	IDADE	DAP	NROPARCELA
1	36.0	9.8	3.0
2	36.0	10.2	3.0
3	36.0	12.2	3.0
4	36.0	12.5	3.0
5	36.0	13.6	3.0
6	36.0	13.8	3.0
7	36.0	14.8	3.0
8	36.0	14.8	3.0
9	36.0	15.0	3.0
10	36.0	15.0	3.0
11	36.0	15.0	3.0
12	36.0	15.3	3.0
13	36.0	15.4	3.0
14	36.0	15.5	3.0

Para realizar o ajuste, foi necessário configurá-lo primeiramente. A Figura 23 apresenta a tela de configuração do sistema Estima relativa ao ajuste.

Figura 23 – Tela de configuração do Ajuste.



A realização da estimativa de parâmetros, conforme demonstrado na Figura 23, realizou-se a partir da escolha do campo (coluna) da amostra selecionada, ou seja, os dados que serão ajustados. Para o estudo de caso, a variável em questão foi a DAP. Feito isso, foi informado qual seria a forma de ajuste no painel “Domínio dos Dados”, usando agrupamento pelas classes idade e parcela com amplitude igual a 2.

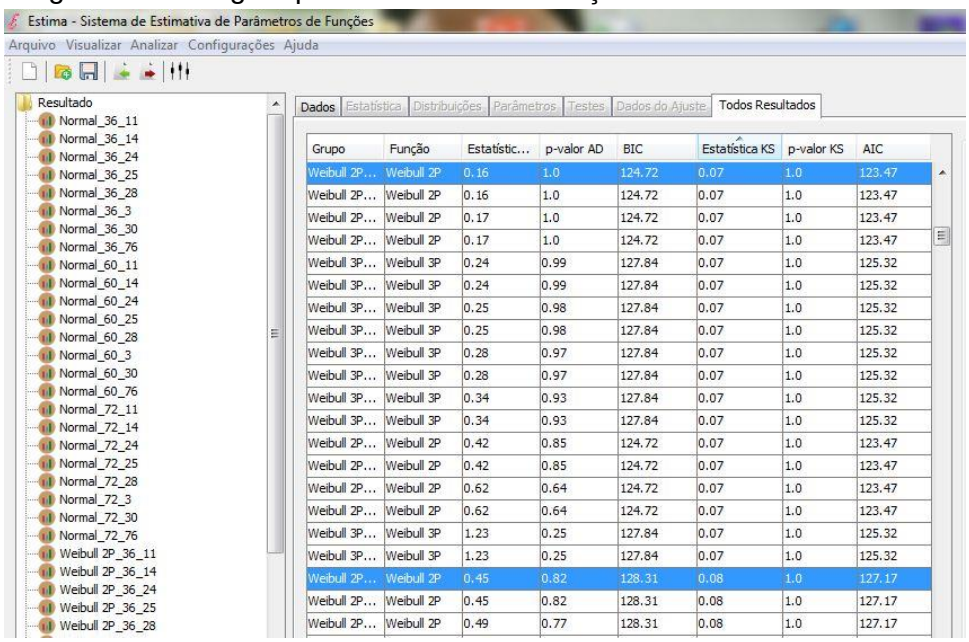
O agrupamento por classes é útil neste caso, devido a quantidade de informações diferenciadas, tornando inviável a construção de uma tabela com uma linha para cada representação de valor. Quando a variável é continua, ou quando é uma variável discreta que pode tomar muitos valores, o uso de agrupamento ou intervalos torna-se necessário para contar quantos elementos existem no intervalo.

Por fim, foi usada a opção de realizar o truncamento das funções, pegando o menor valor da classe para truncamento à esquerda e o maior valor da classe para truncamento à direita.

Vale ressaltar que o sistema possui uma particularidade sobre a função Normal. Atualmente nessa versão, a função Normal não possui truncamento, sendo possível somente realizar o ajuste completo para função Normal.

Após realizar as configurações, foi executado o ajuste, cujos resultados são apresentados na interface gráfica representada na Figura 24 para os casos descritos anteriormente.

Figura 24 – Imagem parcial da tela de exibição de todos os resultados.



Grupo	Função	Estatístic...	p-valor AD	BIC	Estatística KS	p-valor KS	AIC
Weibull 2P...	Weibull 2P	0.16	1.0	124.72	0.07	1.0	123.47
Weibull 2P...	Weibull 2P	0.16	1.0	124.72	0.07	1.0	123.47
Weibull 2P...	Weibull 2P	0.17	1.0	124.72	0.07	1.0	123.47
Weibull 2P...	Weibull 2P	0.17	1.0	124.72	0.07	1.0	123.47
Weibull 3P...	Weibull 3P	0.24	0.99	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.24	0.99	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.25	0.98	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.25	0.98	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.28	0.97	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.28	0.97	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.34	0.93	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	0.34	0.93	127.84	0.07	1.0	125.32
Weibull 2P...	Weibull 2P	0.42	0.85	124.72	0.07	1.0	123.47
Weibull 2P...	Weibull 2P	0.42	0.85	124.72	0.07	1.0	123.47
Weibull 2P...	Weibull 2P	0.62	0.64	124.72	0.07	1.0	123.47
Weibull 2P...	Weibull 2P	0.62	0.64	124.72	0.07	1.0	123.47
Weibull 3P...	Weibull 3P	1.23	0.25	127.84	0.07	1.0	125.32
Weibull 3P...	Weibull 3P	1.23	0.25	127.84	0.07	1.0	125.32
Weibull 2P...	Weibull 2P	0.45	0.82	128.31	0.08	1.0	127.17
Weibull 2P...	Weibull 2P	0.45	0.82	128.31	0.08	1.0	127.17
Weibull 2P...	Weibull 2P	0.49	0.77	128.31	0.08	1.0	127.17

Na Figura 24 é possível visualizar que existe a aba “Distribuição”, “Parâmetros”, “Teste”, “Dados do Ajuste” e “Todos Resultados”. Todas estas abas são pertinentes aos resultados encontrados pela distribuição.

A aba Distribuição exibe os gráficos PDF e CDF individuais para cada grupo, representados por cada nó na árvore de projetos. Já a aba Parâmetros, exibe os parâmetros encontrados para cada determinada distribuição. A aba Teste, exibe os testes pré-configurados na tela de configurações do sistema. A aba de Dados do Ajuste, exibe informações pertinentes aos dados que foram utilizados para realizar os ajustes daquele determinado grupo.

E, por fim, a aba Todos Resultados, onde são exibido todos os resultados encontrados juntos, para que seja possível comparar em relação a todos os grupos. Ao lado direito da Figura 24 é possível visualizar dois gráficos, no qual podem ser selecionados na tabela, e ao clicar na coluna os resultados são ordenados pelo maior ou menor, sendo possível verificar qual função obteve melhor desempenho segundo cada teste.

3.9 Testes usados para validação dos resultados

Para avaliar se os resultados obtidos pelo Estima são estatisticamente consistentes, foi realizada a comparação com os resultados obtidos pelo SAS

por dois testes estatísticos, o teste t de Student e o teste t de Wilcoxon, apresentado nas sessões 2.5.4 e 2.6.5.

Como apresentado nas sessões, pelo teste t, uma vez que formulado o problema e definida as hipóteses e o coeficiente de variação, é possível julgar se a hipótese nula é aceita ou rejeitada.

Inicialmente, para entender os dados, foi gerado o gráfico *Boxplot* e analisado. O gráfico Boxplot é um método alternativo ao histograma para representar os dados. Ele fornece informações sobre as seguintes características do conjunto de dados: locação, dispersão, assimetria, comprimento da cauda e outliers (PETENATE, 2013). E em seguida, para saber se os dados aderem ou não à distribuição em questão foi realizado o teste de normalidade Shapiro-Wilk.

Para saber se os dados aderiram ou não a distribuição foi analisado o valor de p-value retornado do teste. O p-value ou nível descritivo, é a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula. Caso o valor de p-value seja maior que o nível de significância, é possível dizer que os dados aderiram bem a distribuição. Caso contrário, os dados não aderiram bem a distribuição.

Após realizado o teste de normalidade, constatou-se que para a função de distribuição Normal poderia ser usado o teste t de Student, pois os dados aderiram bem a distribuição, ou seja, apresentaram uma distribuição normal. Porém para a distribuição Weibull 2P o teste de normalidade informou que os dados não aderiram bem a distribuição, sendo assim foi usado o teste t de Wilcoxon.

Para avaliar os resultados obtidos da função Weibull 3P pelo Estima, foi usado uma abordagem diferente, devido à versão usada "*SAS University Edition*" não possuir alguns módulos essenciais para estimar os parâmetros da função Weibull com três parâmetros. Portanto para Weibull 3P, foi analisado a qualidade dos resultados apenas pelos gráficos de densidade e probabilidade (PDF) e distribuição acumulada (CDF) gerados pelo Estima, com base no teste de Kolmogorv-Smirnov, sem comparar com o SAS.

3.10 Hipóteses

Uma vez formulado o problema de pesquisa e apresentados os objetivos geral e específicos faz-se necessário desenvolver as hipóteses que servem como base para o alcance dos objetivos.

- **H0:** Os resultados encontrados pelo sistema Estima **não apresentam** diferenças de precisão na estimativa dos parâmetros em relação ao sistema estatístico SAS.
- **H1:** Os resultados encontrados pelo sistema Estima **apresentam** diferenças de precisão na estimativa dos parâmetros em relação ao sistema estatístico SAS.
- $\boldsymbol{\alpha}$ **:** O nível de significância adotado nos testes será de 0,05.

É esperado que através do teste t de Student e Wilcoxon, a hipótese alternativa **H1** seja negada e a hipótese **H0** seja aceita, atestando a eficácia das estimativas realizadas pelo software Estima desenvolvido neste trabalho.

4. RESULTADOS E DISCUSSÃO

Com base na metodologia apresentada, foi desenvolvido o sistema Estima e realizado o estudo de caso buscando verificar a eficiência do sistema comparado com ao SAS. A validação dos resultados obtidos pelo sistema Estima, é apresentado nas sessões seguintes.

4.1.1 Validação dos resultados obtidos no Estima

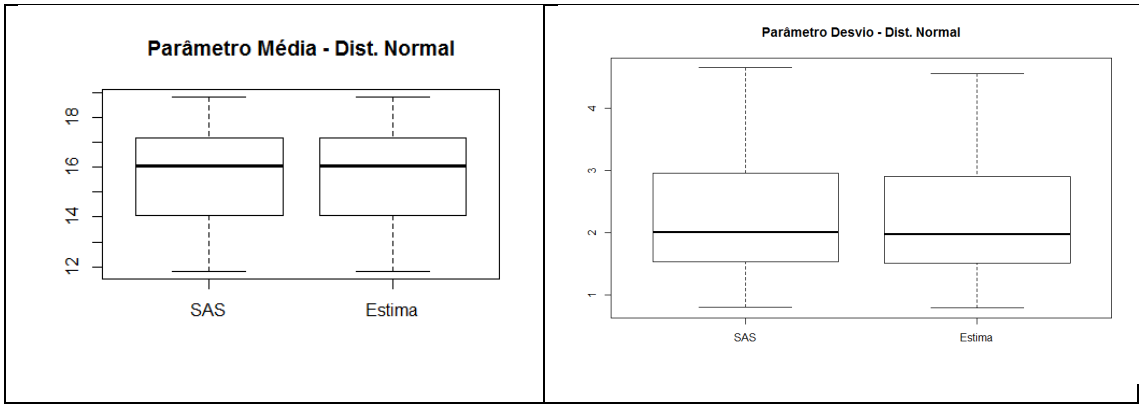
O mesmo procedimento de estimativa dos parâmetros realizado no Estima, foi executado no sistema SAS (SAS Institute Inc, 2000). Foi importado os mesmos dados para o SAS e mandando executar o ajuste dos dados para as funções Weibull 2P e Normal usando agrupamento pelas classes idade e parcela, com amplitude 2 e truncamentos pegando o menor valor de DAP para truncamento à esquerda e o maior valor da classe para truncamento à direita.

A partir do estudo de caso realizado na sessão 3.8 foram gerados resultados dos dois sistemas, e criado um arquivo com os melhores resultados encontrados para a função Normal, Weibull 2P que se encontram nos Apêndices A e B, e a partir deles realizados testes (Apêndices C e D) para verificar se estatisticamente são iguais.

4.1.1.1 Validação da função Normal

Primeiramente, para entender o comportamento dos dados, foi gerado o gráfico *Boxplot* dos parâmetros média e desvio que é apresentado na Figura 25.

Figura 25 – Gráfico *Boxplot* dos parâmetros média e desvio.



Analisando os gráficos da Figura 25, é possível observar que não foi identificada a presença de outliers nos parâmetros média e desvio. Os dados são assimétricos devido ao fato da linha da mediana não estar no centro da caixa e os parâmetros encontrados pelo Estima não variam perceptivamente do SAS.

A validação da função Normal foi realizado em 3 etapas, e na primeira etapa foi verificada a normalidade dos dados para as variáveis desvio padrão e média no SAS, através do teste de normalidade Shapiro-Wilk, com o objetivo de avaliar se o teste t de Student é adequado para comparar os resultados dos parâmetros estimados pelo SAS e o Estima.

Os testes de normalidade são usados para determinar se um conjunto de dados de uma dada variável aleatória é bem modelada por uma distribuição Normal ou não. Nas Tabelas 4 e 5 são apresentados os resultados obtidos pelo teste.

Tabela 4 – Resultado do teste de normalidade Shapiro Wilk para o parâmetro média da distribuição Normal.

IDADE	**SAS**		**ESTIMA**	
	W	p- value	W	p- value
36	0.96899	0.8900	0.96919	0.8916
60	0.96607	0.8655	0.96612	0.8659
72	0.95107	0.7220	0.95111	0.7224

Como pode ser observado na Tabela 4, o valor de *p-value* se manteve superior a α (0,05) em todas as idades, em ambos os sistemas. Portanto para o parâmetro média os dados aderiam a distribuição.

Tabela 5 – Resultado do teste de normalidade Shapiro Wilk para o parâmetro Desvio da distribuição Normal .

IDADE	**SAS**		**ESTIMA**	
	W	p- value	W	p- value
36	0.92001	0.4299	0.92075	0.4359
60	0.98355	0.9783	0.98313	0.9767
72	0.98020	0.9639	0.98152	0.9700

Como pode ser observado na Tabela 5, o valor de p-value se manteve novamente superior a α (0,05) em todas as idades, em ambos os sistemas.

Portanto como o valor de p-value foi maior em ambos os parâmetros, podemos dizer que os dados se aderem bem a distribuição.

Posteriormente, na segunda etapa foi verificada a homogeneidade dos dados, por meio do teste f. O teste f foi usado para detectar a existência ou não de diferenças dos desvios e médias entre o SAS e o Estima, e justificar o uso do teste t de Student ao invés de Wilconxon. Na Tabela 6 é apresentado os resultados obtidos pelo teste f.

Tabela 6 – Resultado do teste f, para verificar homogeneidade dos dados.

IDADE	**MEDIA**		**DESVIO**	
	F	p- value	F	p- value
36	1.001065	0.9989	1.043608	0.9565
60	1.001632	0.9983	1.047504	0.9528
72	0.999858	0.9999	1.047538	0.9527

Portanto, o valor de p-value foi maior que α (0,05), tanto para média, como para o desvio em todas as idades. Então aceitamos a hipótese de que os parâmetros encontrados pelo Estima não diferem significativamente, em relação aos encontrados pelo SAS, podendo assim prosseguir com o teste t de Student ao invés do Wilcoxon.

Por último, é realizado de fato o teste t, e por meio do resultado de p-value no teste foi possível afirmar se a **H0** é aceita ou rejeitada, sendo assim possível afirmar a eficiência do Estima em relação aos resultados encontrados para a função Normal.

Na Tabela 7 é apresentado o resultado do teste t.

Tabela 7 – Resultado do teste t de Student.

IDADE	**MEDIA**		**DESVIO**	
	T	p-value	T	p- value
36	0.0012775	0.9999	0.130810	0.8978
60	0.0023095	0.9982	0.102880	0.9195
72	0.0026439	0.9979	0.099128	0.9224

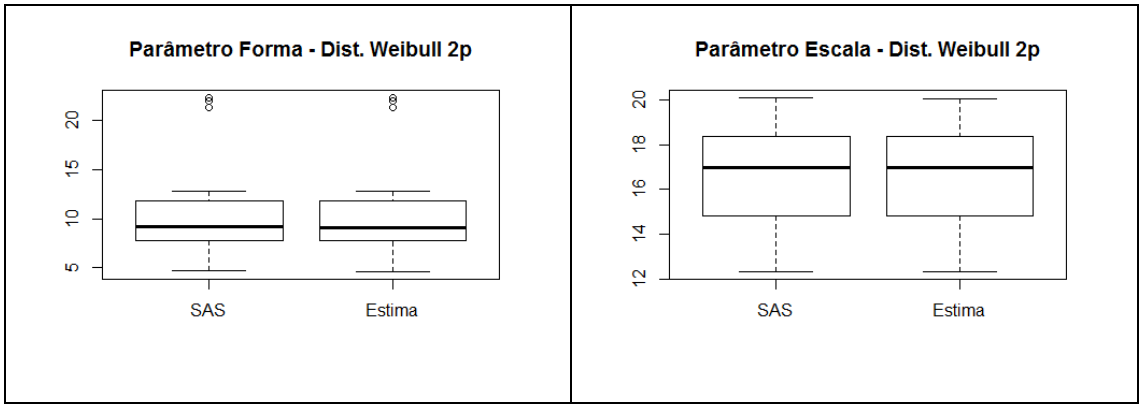
Portanto, analisando o valor de p-value e comparando com α constatamos que os parâmetros da média e do desvio foram maiores que α, permitindo afirmar

que nossa hipótese **H0** foi aceita. Ou seja, os parâmetros obtidos pelo Estima são equivalentes aos encontrados pelos SAS.

4.1.2 Validação da função Weibull 2p

Para realizar o teste na função Weibull 2p primeiramente foi gerado o gráfico boxspot dos parâmetros forma e escala que é apresentado na Figura 26

Figura 26 – Gráfico Boxplot dos parâmetros forma e escala.



Analisando o gráfico do parâmetro forma na Figura 26, leva-nos a suspeitar que os dados podem não ser aderentes a função, devido a presença de outliers do gráfico do parâmetro forma. Em ambos gráficos, os dados são assimétricos e novamente não é perceptível a diferença dos parâmetros encontrados pelo Estima em relação ao SAS.

Em seguida, foi verificada a normalidade dos dados para saber se os mesmos aderem ou não a essa distribuição. A Tabela 8 e 9 apresentam os resultados encontrados pelo teste de Shapiro Wilk.

Tabela 8 – Resultados do teste Shapiro Wilk para o parâmetro Escala.

IDADE	**SAS**		**ESTIMA**	
	W	p- value	W	p- value
36	0.93928	0.6040	0.97141	0.9088
60	0.93714	0.5832	0.93929	0.6041
72	0.93325	0.5461	0.92969	0.5132

Como pode ser observado na Tabela 8, o valor de p-value se manteve superior a α (0,05) em todas as idades, em ambos sistemas. Portanto para o parâmetro escala os dados aderiam a distribuição.

A Tabela 9 apresenta os resultados obtidos pelo teste.

Tabela 9 – Resultados do teste Shapiro Wilk para o parâmetro Forma.

IDADE	SAS		ESTIMA	
	W	p- value	W	p- value
36	0.73238	0.005205	0.73519	0.005594
60	0.75793	0.009991	0.76048	0.010660
72	0.70735	0.002727	0.71112	0.003007

Embora os dados aderiram bem a função para o parâmetro Escala, o parâmetro Forma apresentado na Tabela 9, não aderiu bem a função. Para todos os valores de p-value em todas idades em ambos softwares o valor de p-value se manteve inferior a α, não satisfazendo as condições para aplicar o teste t de Student. Portanto para validar a função Weibull 2p foi usado o teste t de Wilcoxon.

A Tabela 10 apresenta os resultados encontrados pelo teste t de Wilcoxon, para os parâmetros Forma e Escala do Estima x SAS.

Tabela 10 – Resultado do teste t de Wilcoxon.

IDADE	FORMA		ESCALA	
	W	p-value	W	p- value
36	35	0.7984	32	1
60	32	1	32	1
72	34	0.8785	32	1

Como todos os valores encontrados para o p-value foram maiores que o grau de significância $\alpha = 0{,}05$, então pode-se afirmar que a hipótese **H0** é verdadeira. Portanto, os resultados encontrados pelo Estima são equivalentes aos encontrados pelo SAS.

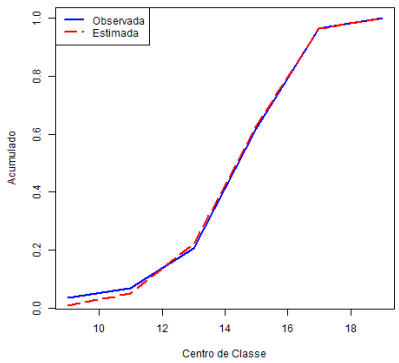
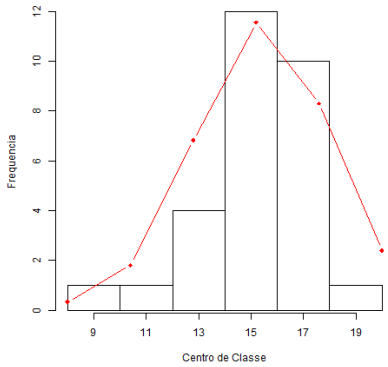
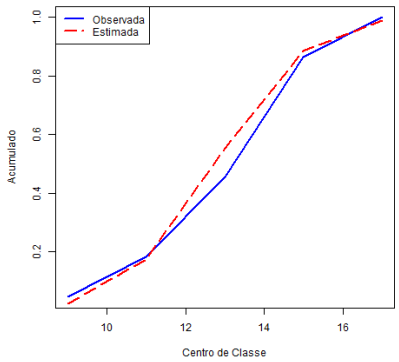
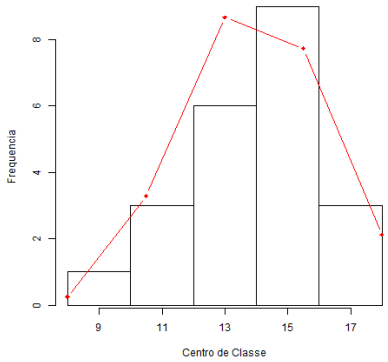
4.1.3 Validação da função Weibull 3P

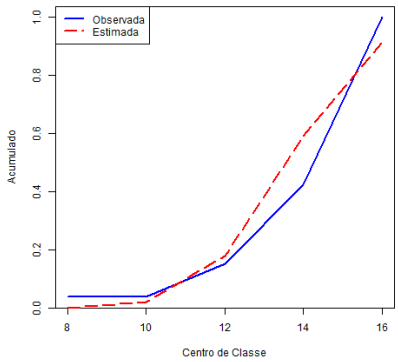
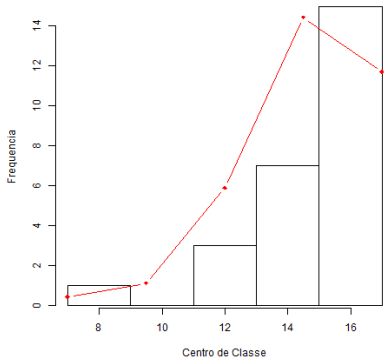
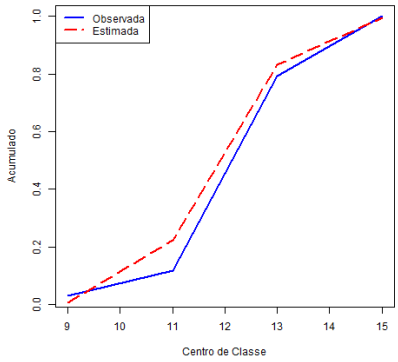
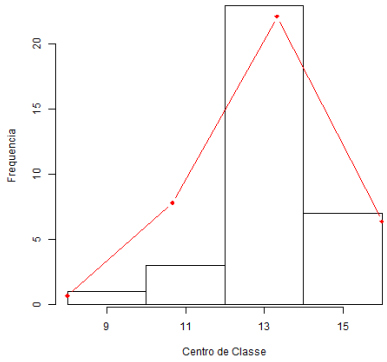
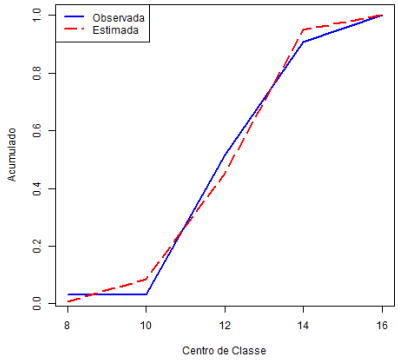
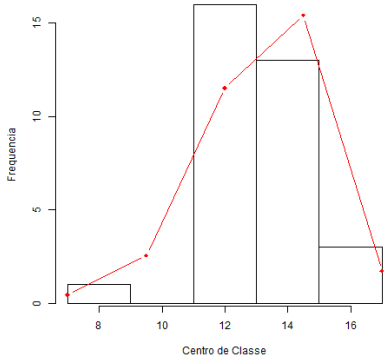
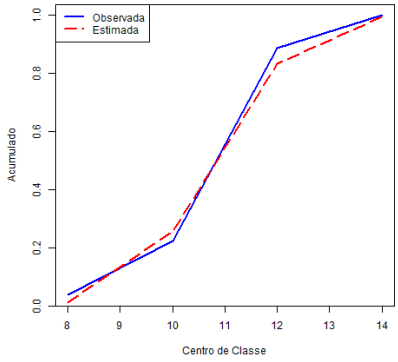
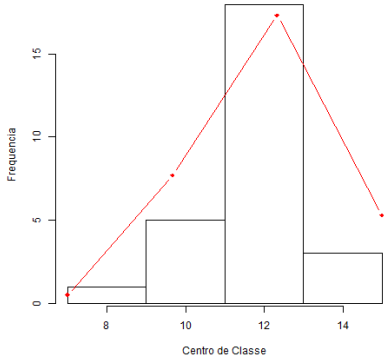
Conforme apresentado na sessão 3.9 da metodologia, para realizar a validação dos resultados encontrados para função de distribuição Weibull 3P no sistema Estima, foi utilizado o teste de Kolmogorov-Smirnov.

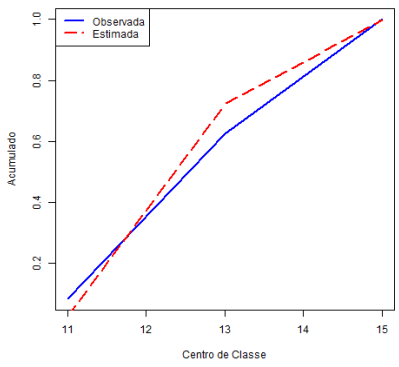
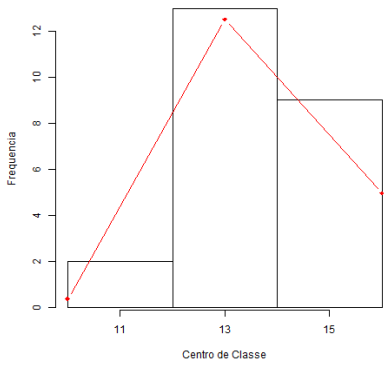
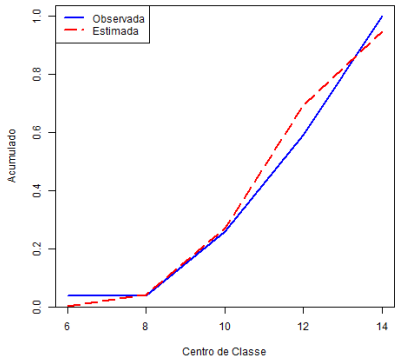
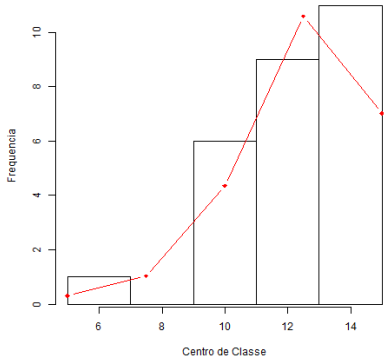
Para avaliar a hipótese **H0,** foi observado o valor de p-value do teste e comparado com o nível de significância. Se o valor-p do teste for menor que o nível de significância apresentado na sessão 3.10, rejeita-se a hipótese **H0** (BIELEFELDT; TORMAN; COSTER, 2012).

A Tabela 11 apresenta os gráficos PDF e CDF encontrados para cada parcela da idade 36.

Tabela 11 – Gráficos PDF e CDF da idade 36.

Idade	**Parcela**	**PDF**	**CDF**
36	3		
36	11		

36	14		
36	24		
36	25		
36	28		

36	30		
36	76		

Com base nos gráficos apresentados na Tabela 11, foi realizado o teste de Kolmogorov-Smirnov e obtido os resultados apresentados na Tabela 12.

Tabela 12 – Resultados do teste de KS referente aos gráficos da Tabela 11.

IDADE	PARCELA	P-VALUE
36	3	0.5191334
36	11	0.9736181
36	14	0.6044404
36	24	0.8105887
36	25	0.2120625
36	28	0.8699978
36	30	0.8080389
36	76	0.7961780

Analisando a Tabela 12, observa-se que para todas as parcelas o valor de p-value > α. Ou seja, podemos afirmar que pelo teste do Kolmogorov-Smirnov para idade 36, a hipótese **H0** foi aceita, atestando a qualidade dos resultados encontrados para idade em questão.

O mesmo processo feito para idade 36, foi realizado para idade 60 e 72. Os gráficos PDF e CDF da idade 60 e 72, estão respectivamente nos Apêndices F e G.

A partir dos gráficos apresentados no Apêndice E, foi realizado o teste de Kolmogorov-Smirnov e obtido os resultados apresentados na Tabela 13.

Tabela 13 – Resultados do teste de KS referente aos gráficos do Apêndice E.

IDADE	PARCELA	P-VALUE
60	3	0.7758537
60	11	0.9952816
60	14	0.5216811
60	24	0.7793905
60	25	0.8740851
60	28	0.6342631
60	30	0.7761374
60	76	0.8200536

Analisando a Tabela 13, observa-se que para todas as parcelas o valor de p-value > α. Ou seja, podemos afirmar que pelo teste do Kolmogorov-Smirnov para idade 60, a hipótese **H0** foi aceita, atestando a qualidade dos resultados encontrados para idade em questão.

A partir dos gráficos apresentados no Apêndice F, foi realizado o teste de Kolmogorov-Smirnov e obtido os resultados apresentados na Tabela 14.

Tabela 14 – Resultados do teste de KS referente aos gráficos do Apêndice F.

IDADE	PARCELA	P-VALUE
72	3	0.9918541
72	11	0.9421885
72	14	0.2808189
72	24	0.7024821
72	25	0.8854151
72	28	0.9781815
72	30	0.9447123
72	76	0.9994400

Analisando a Tabela 14, observa-se que para todas as parcelas o valor de p-value > α. Ou seja, podemos afirmar que pelo teste do Kolmogorov-Smirnov para idade 72, a hipótese **H0** foi aceita, atestando a qualidade dos resultados encontrados para idade em questão.

Após realizar o teste de Kolmogorov-Smirnov para as idades de 36, 60 e 72 é possível afirmar que os resultados encontrados satisfazem o teste, logo os resultados encontrados para função Weibull 3P são satisfatórios.

5. CONCLUSAO

Foi desenvolvido neste trabalho um software livre “Estima”[1] para o auxílio da estimação de parâmetros de função de probabilidade para variáveis contínuas com o método da máxima verossimilhança, onde por meio de uma interface gráfica de uso comum a usuários com pouca experiência em linguagens de programação, poderão estimar parâmetros de funções a partir de seus dados.

Foi desenvolvido e integrado o script da função Normal, e integrados os scripts da função Weibull 2 e 3 parâmetros desenvolvidos por Fraga Filho (2014).

Como constatado pelo estudo de caso, o sistema realizou a estimativa de parâmetros das funções Normal e Weibull 2 e 3 parâmetros de forma estatisticamente eficiente, comparada com o software SAS (INC.SAS/STAT, 2000), atingindo as hipóteses esperadas.

O sistema foi projetado para que sua arquitetura fosse plugável, ou seja, que novos scripts fossem adicionados incrementado suas funcionalidades, sendo assim possível a inserção de novas funções.

O software Estima se mostrou eficiente na estimação de parâmetros de funções para dados florestais, porém recomenda-se a realização de novos estudos de caso para dados de outras áreas.

5.1 Trabalhos Futuros

Espera-se que novas funções como Log-Normal, SB-Johnson, Binomial, entre outras sejam adicionadas ao Estima contribuindo para seu aprimoramento. Espera-se ainda que outros estudos de caso sejam realizados com dados provenientes de outras amostras, preferencialmente de diferentes áreas da ciência.

[1] Disponível em http://sourceforge.net/projects/estima/

6. REFERÊNCIAS BIBLIOGRÁFICA

ACTION, Portal. *Teste de Wilcoxon Pareado - Portal Action*. Disponível em: <http://www.portalaction.com.br/manual-action/533-pareado>. Acesso em: 1 jun. 2015.

AFONSO, Aminta Paiva. *Distribuições de Probabilidade*. Disponível em: <http://matematiques.com.br/download.php?tabela=documentos&id=427>. Acesso em: 18 nov. 2014.

AKAIKE, Hirotugu. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, v. 19, n. 6, p. 716–723, 1974.

AMERICA, STATSOFT SOUTH. *STATISTICA 12 - StatSoft South America | Tornando o Mundo mais Produtivo*. Disponível em: <http://www.statsoft.com.br/pt-BR/componentes/statistica-12.html>. Acesso em: 19 nov. 2014.

ARAÚJO JÚNIOR, Carlos Alberto *et al.* Projeção da distribuição diamétrica de povoamentos de eucalipto em diferentes amplitudes de classe. *Pesquisa Agropecuaria Brasileira*, v. 45, n. 11, p. 1275–1281, 2010.

BAAYEN, R H. A Practical Introduction to Statistics. 2001.

BAILEY, Robert L; DELL, T R. Quantifying diameter distributions with the Weibull function. *Forest Science*, v. 19, n. 2, p. 97–104, 1973.

BATISTA, J L F. *A função Weibull como modelo para a distribuição de diâmetros de espécies arbóreas tropicais*. [S.l.]: ESALQ, 1989.

BIELEFELDT, Vanessa; TORMAN, Leotti; COSTER, Rodrigo. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Rev. HCPA*, v. 32, n. 2, p. 227–234, 2012.

BINOTI, Daniel Henrique Breda *et al.* Uso da função Weibull de três parâmetros em um modelo de distribuição diamétrica para plantios de eucalipto submetidos a desbaste. *Revista Árvore*, v. 34, n. 1, p. 147–156, 2010.

BUNN, Andy. Time Series Analysis in dplR. v. 2, 2014.

BURDEN, Richard L; FAIRES, J Douglas; TASKS, All. *Análise numérica*. [S.l.]: Cengage Learning, 2008.

BUSSAB, Wilton O; MORETTIN, P A. Estatística Básica, Editora Saraiva, 5a. edição. *São Paulo*, 2002.

CARGNELUTTI FILHO, Alberto *et al.* Ajustes de funções de distribuição de probabilidade à radiação solar global no Estado do Rio Grande do Sul. *Pesquisa Agropecuária Brasileira*, v. 39, n. 12, p. 1157–1166, 2004.

CATALUNHA, M J *et al.* Avaliação de cinco funções densidade de probabilidade a séries de precipitação pluvial no Estado de Minas Gerais. *Revista Brasileira de Agrometeorologia*, v. 10, n. 1, p. 153–162, 2002.

COIMBRA, UNIVERSIDADE DE. *Teste não paramétricos*. Disponível em: <http://www.mat.uc.pt/~cmtm/ECwww/TestesNP.pdf>. Acesso em: 19 nov. 2014.

CORRÊA, Henrique L; GIANESI, Irineu G N; CAON, Mauro. Planejamento, programação e controle da produção. *São Paulo: Atlas*, v. 1, 2001.

CRUCIANI, Décio E; MACHADO, Ronalton E; SENTELHAS, Paulo C. Modelos da distribuição temporal de chuvas intensas em Piracicaba , SP. p. 76–82, 2002.

DAP Florestal. Disponível em: <http://www.dapflorestal.com.br/area/faq>. Acesso em: 27 maio 2015.

FERNANDES, Alda Marisa Ribeiro. *Análise de Dados em Modelos Multiestado*. 2010. Universidade do Minho, 2010.

FILHO, Clayton V. Fraga *et al. Forest Measure*. . [S.l: s.n.]. Disponível em: <http://sourceforge.net/projects/forestmeasure/>. Acesso em: 27 nov. 2014. , 2014

FILHO, Clayton V. Fraga; SCHMIDT, Luciane Naimeke. *Weibull 2 e 3p Maximum Likelihood*. Disponível em: <http://sourceforge.net/projects/weibull23p/>. Acesso em: 25 nov. 2014.

FILHO, Frederico Ferreira Campos. Algoritmos Numéricos. 2007.

FULLERTON, California State University. *Nelder-Mead Search for a Minimum*. Disponível em: <http://mathfaculty.fullerton.edu/mathews/n2003/neldermead/NelderMeadMod/Links/NelderMeadMod_lnk_5.html>. Acesso em: 25 nov. 2014.

GAO, Fuchang; HAN, Lixing. Implementing the Nelder-Mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, v. 51, n. 1, p. 259–277, 4 maio 2010. Disponível em: <http://link.springer.com/10.1007/s10589-010-9329-3>. Acesso em: 25 out. 2014.

GUIMARÃES, D P. Uma função hiperbólica de distribuição probabilística de alta flexibilidade. *Embrapa Cerrados. Documentos*, 2002.

HASTINGS, Nicholas Anthony John; PEACOCK, J Brian. *Statistical distributions: A handbook for students and practitioners*. [S.l.]: Butterworth London:, 1975.

HÉLIO, Prof; BITTENCOURT, Radke. Bioestatística. p. 1–63, 2006.

HILLIER, F S; HILLIER, M S. *Introdução à Ciência da Gestão - 4.ed.: Modelagem e Estudos de Caso com Planilhas Eletrônicas*. [S.l: s.n.], 2014. Disponível em: <https://books.google.com.br/books?id=KO6_AwAAQBAJ>.

INC.SAS/STAT, SAS Institute. *Statistical Analysis System*. . [S.l: s.n.]. Disponível em: <http://www.sas.com/en_us/software/enterprise-solutions.html>. Acesso em: 21 nov. 2014. , 2000

LAURINI, Márcio Poletti; HOTTA, Luiz Koodi. *Estimação de Equações Diferenciais Estocásticas Usando Verossimilhança Empírica e Mínimo Contraste Generalizado*. . [S.l: s.n.], 2009.

LEITE, Helio Garcia *et al.* Um modelo de distribuição diamétrica para povoamentos de Tectona grandis submetidos a desbaste. *Revista Árvore*, v. 30, n. 1, p. 89–98, 2006.

LOPES, Prof. Dr. Luis Felipe. Treinamento sistema - s.a.s. 2006.

MACHADO. *Sistema Galileu de Educação Estatística*. Disponível em: <http://www.galileu.esalq.usp.br/mostra_topico.php?cod=303>. Acesso em: 26 nov. 2014.

MARCELO S. PORTUGAL. Notas Introdutórias Sobre o Princípio de Máxima Verossimilhança: Estimação e Teste de Hipóteses. *Nota*, n. 1986, p. 1–24, 1995.

MASSAGO, Sadao. Matriz Hessiana e Aplicações. 2010.

MATHWAVE TECHNOLOGIES. *EasyFit - Distribution Fitting Software - Benefits*. Disponível em: <http://www.mathwave.com/easyfit-distribution-fitting.html>. Acesso em: 19 nov. 2014.

MENDONÇA, Adriano Ribeiro De *et al.* Modelos hiposométricos e de crescimento em altura das árvores dominantes e codominates para Pinus caribaea var. hondurensis Height diameter relationship and growth in height the dominant and codominant trees model to Pinus caribaea var. hondurensis. *Scientia Forestalis*, v. 39, n. 90, p. 151–160, 2011.

NAGHETTINI, Mauro; PINTO, Éber José de Andade. *Hidrologia Estatística*. [S.l: s.n.], 2007.

NASH, John C. *Nonlinear Parameter Optimization Using R Tools*. [S.l.]: Wiley, 2014. Disponível em: <http://books.google.com/books?id=QZFPAwAAQBAJ&pgis=1>. Acesso em: 27 nov. 2014.

NASH, John C.; GROTHENDIECK, Gabor; VARADHAN, Ravi. *Package "optimx".* Disponível em: <http://cran.r-project.org/web/packages/optimx/optimx.pdf>. Acesso em: 19 nov. 2014.

NETER, John; WASSERMAN, William; KUTNER, Michael H. Applied Linear Models, Regression, Analysis of Variance and Experimental Designs. *RD Irwin, Boston, USA*, 1990.

ORACLE, Technology Network. *Java SE Development Kit 7*. Disponível em: <http://www.oracle.com/technetwork/pt/java/javase/downloads/jdk7-downloads-1880260.html>. Acesso em: 7 jul. 2015.

OZYMANDIAS. *Introdução à Pesquisa Operacional/Algoritmo Simplex - Wikiversidade*. Disponível em: <http://pt.wikiversity.org/wiki/Introdução_à_Pesquisa_Operacional/Algoritmo_Simplex>. Acesso em: 19 nov. 2014.

PÉLLICO NETTO, Sylvio; BRENA, Doádi Antonio. Inventário florestal. 1997.

PETENATE, Marcelo. *Box Plot: como analisar e interpretar esse gráfico*. Disponível em: <http://www.escolaedti.com.br/o-que-e-um-box-plot/>. Acesso em: 20 jun. 2015.

PETERNELLI, Luiz Alexandre. Variáveis aleatórias e distribuições de probabilidade. p. 47–67, 2006.

PUC-RIO. *4 Teoria da Probabilidade*. Disponível em: <http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0210463_06_cap_04.pdf>. Acesso em: 19 nov. 2014.

R DEVELOPMENT CORE TEAM, R Development Core. *R: A language and environment for statistical computing*. *R foundation for Statistical Computing*. Vienna, Austria: sn. , 2005

REGGIERO, Márcia A. Gorems. Método de Newton. 2008.

SANSIGOLO, Clóvis Angeli. Distribuições de probabilidade de velocidade e potência do vento. *Revista Brasileira de Meteorologia*, v. 20, n. 2, p. 207–214, 2005.

SANTOS, Fabiano Lourenço Dos. Regulação da produção de floresta de eucalipto submetida a desbaste e destinada a miltiprodutos. 2008.

SCHMIDT, Luciane Naimeke. Avaliação de diferentes formas de ajuste da função Weibull. *TCC*, 2014.

SCHWARZ, Gideon. Estimating the dimension of a model. *The annals of statistics*, v. 6, n. 2, p. 461–464, 1978.

SCUDINO, Patrícia Araujo. *A Utilização de Alguns Testes Estatísticos para Análise da Variabilidade do Preço do Mel nos Municípios de Angra dos Reis e Mangaratiba, Estado do Rio de Janeiro*. 2008. Univesidade Federal Rural do Rio de Janeiro, 2008.

SILVA, A A L. *Emprego de modelos de crescimento e produção em análise econômica de decisões de manejo florestal. 2001. 79 f*. . [S.l.]: Tese (Doutorado em Ciência Florestal)–Universidade Federal de Viçosa, Viçosa, 2001.[Links]. , [S.d.]

SILVA, Cosme Teixeira Da; FERREIRA, Eric Batista. DESEMPENHO DE TESTES DE NORMALIDADE VIA SIMULAÇÃO MONTE CARLO. p. 1–5, 2009.

SILVA, Eduardo Quadros. Nova função densidade de probabilidade aplicável à ciência florestal. 2003.

SILVA, Ulisses; CUNHA, D A. DISTRIBUIÇÕES DIAMÉTRICAS E RELAÇÕES HIPSOMÉTRICAS DE UMA FLORESTA TROPICAL ÚMIDA DE 1000 ha , ESTAÇÃO. 1994.

WEIBULL, Waloddi. *A Statistical Distribution Function of Wide Applicability*. . [S.l: s.n.]. , 1951

WRIGHT, Margaret H. Nelder , Mead , and the Other Simplex Method. v. I, n. October 1924, p. 271–276, 2012.

XAVIER, Carolina Ribeiro. *Comparação de métodos de otimização para o problema de ajuste de histórico em ambientes paralelos*. 2009. UNIVERSIDADE FEDERAL DE JUIZ DE FORA, Juiz de Fora, 2009.

YEH, Kuo-Chen; LAGARIAS, J Clark. Eukaryotic phytochromes: light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proceedings of the National Academy of Sciences*, v. 95, n. 23, p. 13976–13981, 1998.

7. APÊNDICES

APÊNDICE A – Dados da função normal usados no estudo de caso

Idade	Parcela	Media_SAS	Media_Estima	Desvio_SAS	Desvio_Estima
36	3	1,52E+06	1,52E+14	2,01E+06	1,98E+14
36	11	13.75	1,37E+14	1,91E+05	1,87E+14
36	14	1,46E+06	1,46E+14	1,78E+06	1,75E+14
36	24	1,29E+06	1,29E+14	1,18E+06	1,16E+14
36	25	1,30E+06	1,30E+14	1,36E+04	1,34E+14
36	28	1,18E+06	1,18E+14	1,25E+06	1,23E+14
36	30	1,35E+06	1,35E+14	0,8055776	0,788810023974369
36	76	1,21E+05	1,21E+14	1,83E+05	1,79E+14
60	3	1,72E+05	1,72E+14	3,97E+06	3,88E+14
60	11	16,5	1,65E+14	3,17E+06	3,10E+14
60	14	1,77E+06	1,77E+14	3,01E+05	2,96E+14
60	24	1,63E+06	1,63E+13	2,01E+06	1,98E+13
60	25	1,58E+06	1,58E+14	2,24E+06	2,21E+13
60	28	14,4	1,44E+14	1,71E+06	1,68E+14
60	30	1,69E+06	1,69E+14	0,992791	0,9718873485954
60	76	15,7	1,57E+13	2,89E+06	2,84E+14
72	3	1,84E+06	1,84E+14	4,65E+06	4,55E+14
72	11	1,76E+06	1,76E+14	3,79E+06	3,70E+14
72	14	1,88E+06	1,88E+14	3,46E+06	3,39E+14
72	24	1,72E+06	1,72E+14	2,27E+06	2,23E+13
72	25	1,68E+06	1,68E+14	2,63E+05	2,59E+13
72	28	1,53E+06	1,53E+14	1,99E+06	1,96E+14
72	30	1,82E+06	1,82E+14	1,12E+06	1,10E+14
72	76	1,70E+06	1,70E+14	2,58E+06	2,53E+14

APÊNDICE B – Dados da função Weibull2p usados no estudo de caso

Idade	Parcela	Forma_SAS	Forma_Estima	Escala_SAS	Escala_Estima
36	3	1,03E+06	1,03E+14	1,60E+06	1,60E+13
36	11	9,22E+05	9,20E+14	1,45E+06	1,45E+14
36	14	1,24E+06	1,24E+14	1,53E+06	1,53E+13
36	24	1,24E+06	1,24E+14	1,24E+06	1,34E+14
36	25	1,13E+06	1,13E+13	1,36E+06	1,36E+14
36	28	1,27E+06	1,27E+14	1,23E+06	1,23E+14
36	30	2,23E+05	2,23E+13	1,39E+06	1,39E+14
36	76	9,05E+06	8,99E+14	1,28E+06	1,28E+13
60	3	5,29E+06	5,22E+14	1,87E+06	1,87E+14
60	11	6,25E+06	6,26E+14	1,78E+05	1,78E+14
60	14	9,06E+06	9,05E+13	1,88E+06	1,88E+14
60	24	9,33E+06	9,33E+14	1,72E+06	1,72E+14
60	25	8,32E+06	8,32E+14	1,67E+06	1,67E+13
60	28	1,08E+05	1,08E+14	1,51E+06	1,51E+14
60	30	2,13E+06	2,13E+14	1,74E+06	1,74E+14
60	76	7,12E+06	7,09E+14	1,68E+05	1,68E+14
72	3	4,69E+06	4,60E+14	2,01E+06	2,01E+14
72	11	5,45E+06	5,40E+14	1,91E+06	1,90E+14
72	14	8,21E+04	8,21E+14	2,01E+06	2,01E+14
72	24	8,77E+06	8,77E+14	1,81E+06	1,81E+14
72	25	7,69E+06	7,69E+14	1,79E+05	1,79E+14
72	28	9,60E+06	9,59E+14	1,62E+06	1,62E+14
72	30	2,19E+06	2,19E+14	1,87E+06	1,87E+14
72	76	7,81E+06	7,81E+14	1,80E+06	1,80E+14

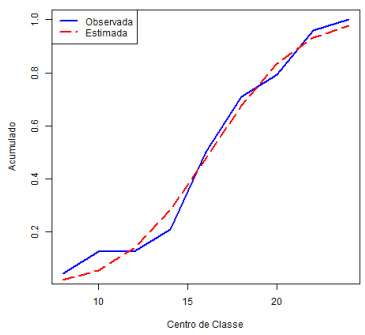
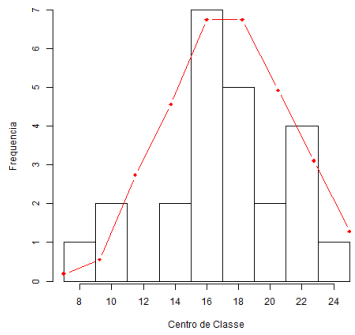
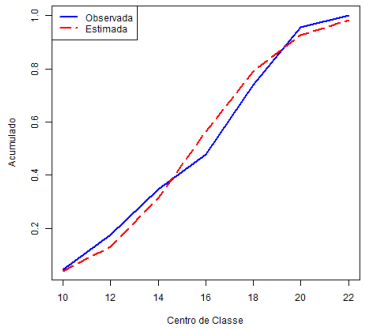
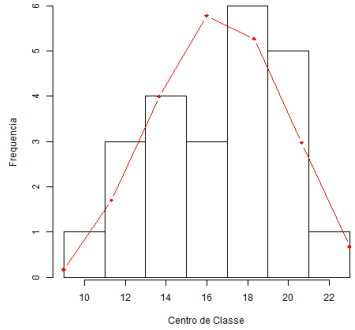
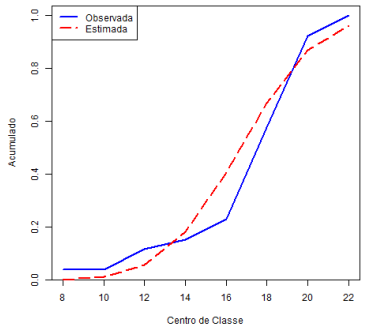
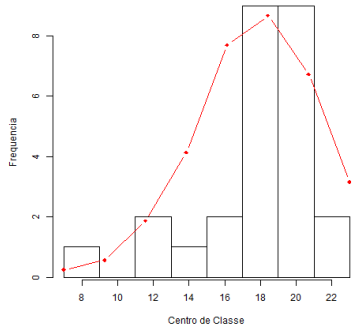
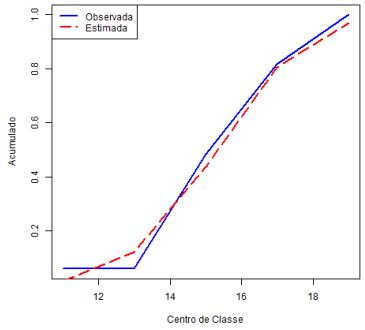
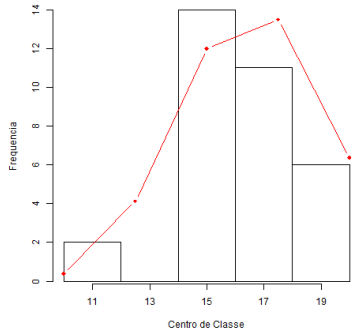
APÊNDICE C – Script usado para teste t da função Normal no estudo de caso

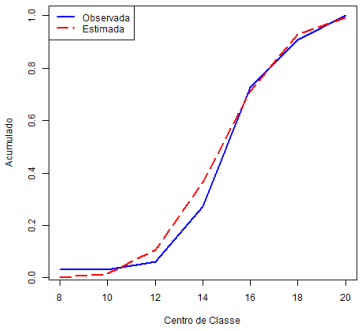
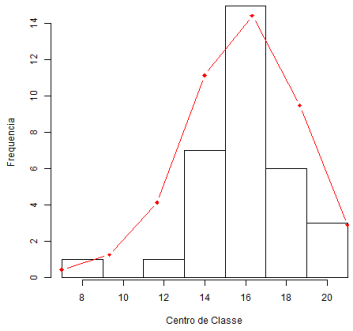
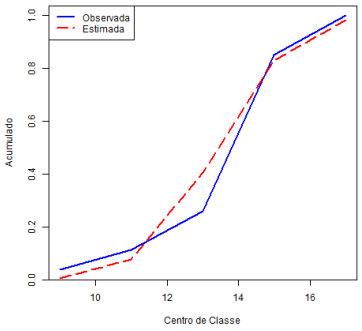
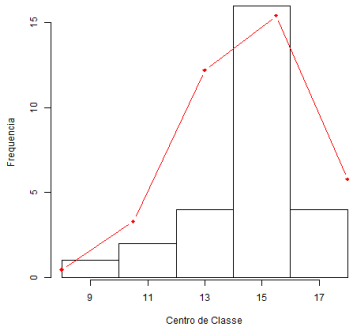
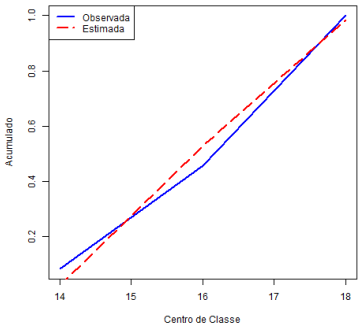
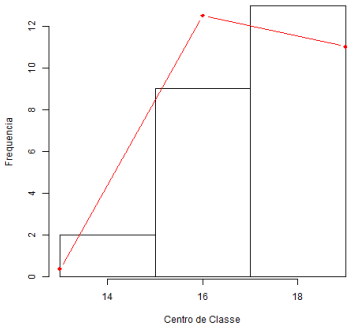
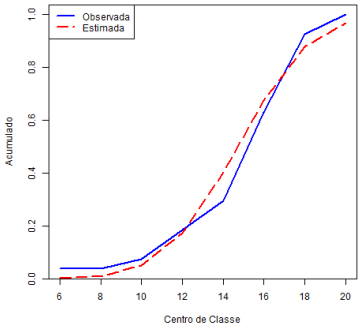
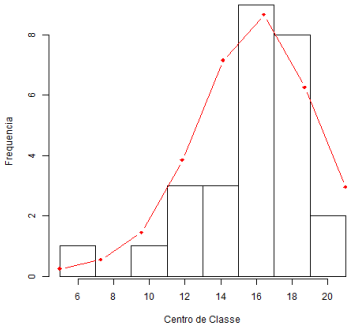
```
if (!require("nortest")) {install.packages("nortest")}

dadosNormal = read.table(file = "dadosNormal.csv", sep = ";", header=T)

#Leia a sequencia a partir da página 25: http://cran.r-
project.org/doc/contrib/Beasley-BioestatisticaUsandoR.pdf
#Tem um bom exemplo, usaremos o shapiro.test e seguiremos na sequencia até o
test t nao pareado
by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Desvio_SAS)
})


by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Desvio_Estima)
})

#Leia sobre variancias homogeneas
by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  var.test(dadosNaIdade$Desvio_SAS, dadosNaIdade$Desvio_Estima)
})

############## PARA O PARAMETRO MEDIA

by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Media_SAS)
})


by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Media_Estima)
})

#Leia sobre variancias homogeneas
by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  var.test(dadosNaIdade$Media_SAS, dadosNaIdade$Media_Estima)
})

#########################

by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  t.test(dadosNaIdade$Media_SAS, dadosNaIdade$Media_Estima, var.equal = T)
})

by(dadosNormal, dadosNormal$Idade, function(dadosNaIdade) {
  t.test(dadosNaIdade$Desvio_SAS, dadosNaIdade$Desvio_Estima, var.equal = T)
})

#Coloque também os gráficos:
Boxplot(dadosNormal$Media_SAS, dadosNormal$Media_Estima, names = c("SAS",
"Estima"), main="Parâmetro Média - Dist. Normal")


Boxplot(dadosNormal$Desvio_SAS, dadosNormal$Desvio_Estima, names = c("SAS",
"Estima"), main="Parâmetro Desvio - Dist. Normal")
```

APÊNDICE D – Script usado para teste t da função Weibull2p no estudo de caso

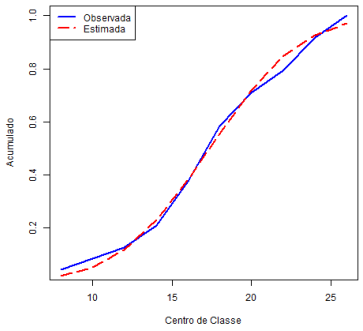
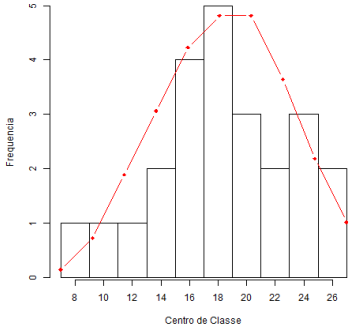
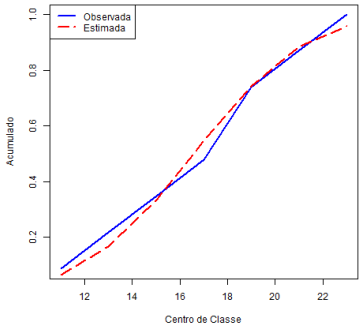
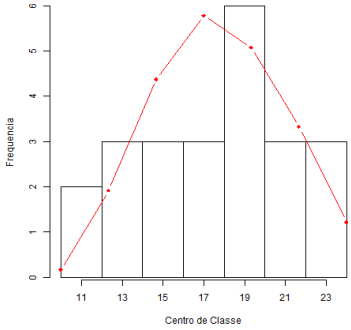
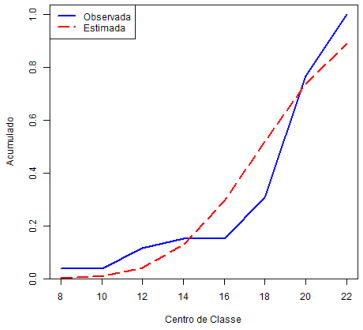
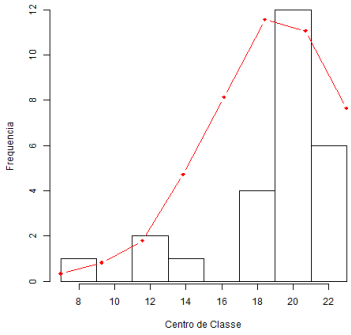
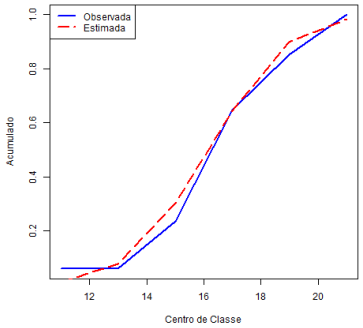
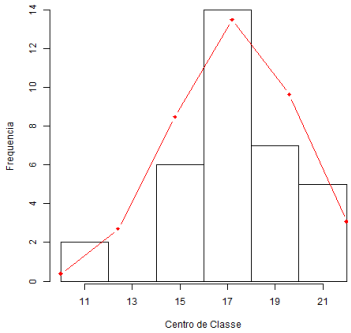
```
dadosW2 = read.table(file = "dadosWeibull2p.csv", sep = ";", header=T)

#Mostrando que os dados não aderem a distribuição Normal pelo teste Shapiro
# e assim justificando o uso do Wilcoxon

by(dadosW2, dadosW2$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Forma_SAS)
})


by(dadosW2, dadosW2$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Escala_SAS)
})

by(dadosW2, dadosW2$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Forma_Estima)
})


by(dadosW2, dadosW2$Idade, function(dadosNaIdade) {
  shapiro.test(dadosNaIdade$Forma_SAS)
})


#Resultados para cada idade, para o parametro forma Weibull2
by(dadosW2, dadosW2$Idade, function(dadosNaIdade) {
  wilcox.test(dadosNaIdade$Forma_SAS, dadosNaIdade$Forma_Estima)
})

#Resultados para cada idade, para o parametro escala Weibull2
by(dadosW2, dadosW2$Idade, function(dadosNaIdade) {
  wilcox.test(dadosNaIdade$Escala_SAS, dadosNaIdade$Escala_Estima)
})


#Coloque também os gráficos:
Boxplot(dadosW2$Forma_SAS, dadosW2$Forma_Estima, names = c("SAS", "Estima"),
main="Parâmetro Forma - Dist. Weibull 2p")


Boxplot(dadosW2$Escala_SAS, dadosW2$Escala_Estima, names = c("SAS", "Estima"),
main="Parâmetro Escala - Dist. Weibull 2p")
```

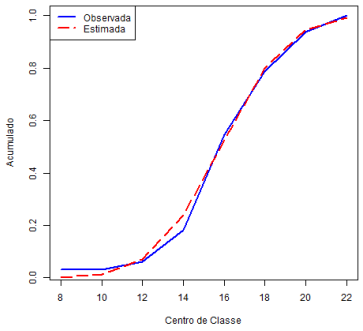
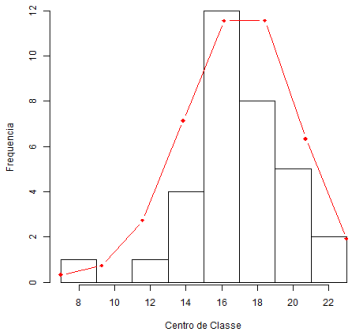
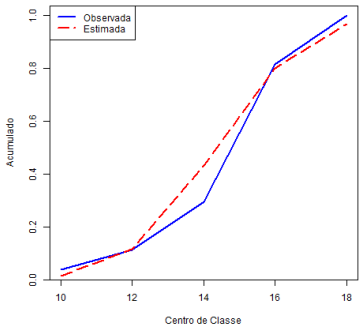
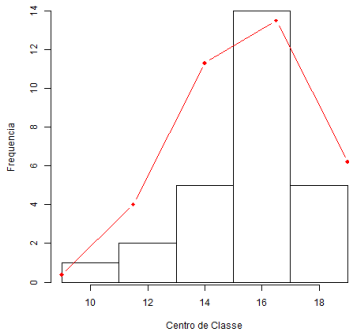
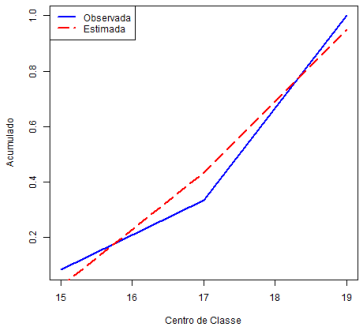
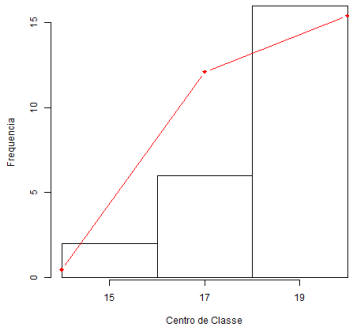
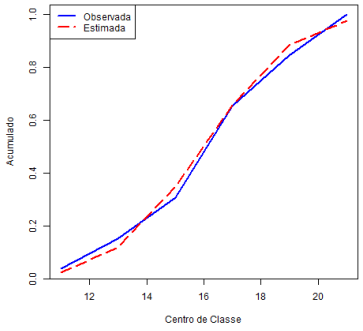
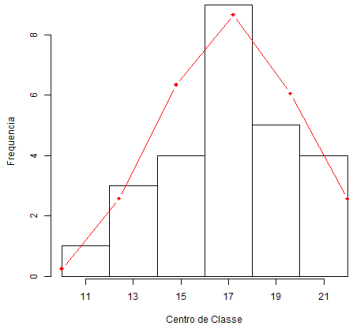
APÊNDICE E – Gráficos PDF e CDF da idade 60

IDADE	PARCELA	PDF	CDF
60	3		
60	11		
60	14		
60	24		

60	25		
60	28		
60	30		
60	76		

APÊNDICE F – Gráficos PDF e CDF da idade 72

IDADE	PARCELA	PDF	CDF
72	3		
72	11		
72	14		
72	24		

72	25		
72	28		
72	30		
72	76		

APÊNDICE G – Script usado no SAS para o estudo de caso

```
proc import datafile="meusdados.csv"
                  out=mydata
                  dbms=dlm
                  replace;
                  delimiter=';';
                  getnames=yes;
run;

title 'Estimação de parâmetros';
ods select ParameterEstimates GoodnessOfFit FitQuantiles MyHist;
proc univariate data=mydata;
   var dap;
   histogram / midpoints=0.2 to 1.8 by 0.2
               normal
               weibull;
run;
```

8. ANEXOS

ANEXO A - Dedução da função de máxima verossimilhança para a distribuição:

1. Dedução da função Weibull de 2 parâmetros (2P)

Nessa seção é apresentado a dedução da função Weibull de 2 parâmetros completa (2p), que foi deduzida por Schmidt (2014):

$$f(x) = \left(\frac{c}{b}\right)\left(\frac{x}{b}\right)^{c-1} exp\left[-\left(\frac{x}{b}\right)^{c}\right]$$

Dedução da função de máxima verossimilhança (L):

É realizado o produto da função de densidade $f(x)$

$$L = \prod f(x) = \left(\frac{c}{b}\right)^{n} \prod \left(\frac{x}{b}\right)^{c-1} exp\left[-\sum \left(\frac{x}{b}\right)^{c}\right]$$

São retirados os parâmetros b do produto e c do somatório

$$L = \frac{c^n}{b^n} \left(\frac{1}{b^{c-1}}\right)^{n} \prod (x^{c-1})\, exp\left[-\left(\frac{1}{b^c}\right)\sum x^c\right]$$

Aplica-se a potência $\left(\frac{1}{b^{c-1}}\right)^{n}$

$$L = \frac{c^n}{b^n} \frac{1}{b^{n(c-1)}} \prod (x^{c-1})\, exp\left[-\left(\frac{1}{b^c}\right)\sum x^c\right]$$

Realiza-se a multiplicação das frações

$$L = \frac{c^n}{b^{nc}} \prod (x^{c-1})\, exp\left[-\left(\frac{1}{b^c}\right)\sum x^c\right]$$

2. Dedução da função Weibull de 3 parâmetros (3P)

Nessa seção é apresentado a dedução da função Weibull de 3 parâmetros completa (3P) deduzida por Schmidt (2014):

$$f(x) = \left(\frac{c}{b}\right)\left(\frac{x-a}{b}\right)^{c-1} exp\left[-\left(\frac{x-a}{b}\right)^{c}\right]$$

Dedução da função de máxima verossimilhança (L):
De forma Análoga a função Weibull 2P,

$$L = \prod f(x) = \left(\frac{c}{b}\right)^{n} \prod \left(\frac{x-a}{b}\right)^{c-1} exp\left[-\sum\left(\frac{x-a}{b}\right)^{c}\right]$$

$$L = \frac{c^n}{b^n}\left(\frac{1}{b^{c-1}}\right)^{n} \prod (x-a)^{c-1} exp\left[-\left(\frac{1}{b^c}\right)\sum (x-a)^c\right]$$

$$L = \frac{c^n}{b^n}\ \frac{1}{b^{n(c-1)}} \prod (x-a)^{c-1} exp\left[-\left(\frac{1}{b^c}\right)\sum (x-a)^c\right]$$

$$L = \frac{c^n}{b^{nc}} \prod (x-a)^{c-1}\ exp\left[-\left(\frac{1}{b^c}\right)\sum (x-a)^c\right]$$

3. Dedução da Função Normal

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} . e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty < x < \infty$$

Dedução da função de máxima verossimilhança (L):

Reescrevendo a formula, temos que:

$$y_i \sim N(\theta, y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(y_i-\mu_i)^2}{2\sigma^2}}$$

Onde $\mu_i = x_i\beta$, podendo ser reescrita como:

$$y_i \sim N(\mu_i, \sigma^2 I) = \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}}$$

Assim, a probabilidade de uma única observação é:

$$L(y_i|x, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}}$$

Segue-se que a probabilidade para o conjunto da amostra é:

$$L(y_i|x, \beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}}$$

Na forma matricial, a verossimilhança é:

$$\frac{1}{(\sqrt{2\pi\sigma^2})^N} . e^{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)}$$

Portanto, o log- verossimilhança é:

$$ln\, L(y|x, \beta, \sigma^2) = \sum_{i=1}^{N} ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} . e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}}\right)$$

$$= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{\left(y_{i-x_i\beta}\right)^2}{\sigma^2}\right]$$