JOSÉ ALFREDO DIAZ ESCOBAR

DISTRIBUIÇÕES DE PROBABILIDADES DO VALOR EXTREMO E TAMANHO AMOSTRAL PARA O MELHORAMENTO GENÉTICO DO QUANTIL MÁXIMO EM PLANTAS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae.*

VIÇOSA
MINAS GERAIS – BRASIL
2016

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

D542d
2016

Diaz Escobar, José Alfredo, 1983-
Distribuição de probabilidades do valor extremo e tamanho amostral para o melhoramento genético do quantil máximo em plantas / José Alfredo Diaz Escobar. - Viçosa, MG, 2016.
ix, 64f. : il. (algumas color.) ; 29 cm.

Inclui anexos.
Orientador: Marcos Deon Vilela de Resende.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Estatística aplicada. 2. Biometria. 3. Teoria dos valores extremos. 4. Produtos agrícolas - Métodos estatísticos. 5. Plantas - Melhoramento genético. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.5

JOSÉ ALFREDO DIAZ ESCOBAR

DISTRIBUIÇÕES DE PROBABILIDADES DO VALOR EXTREMO E TAMANHO AMOSTRAL PARA O MELHORAMENTO GENÉTICO DO QUANTIL MÁXIMO EM PLANTAS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 29 de fevereiro de 2016.

_____	_____
Camila Ferreira Azevedo	Fabyano Fonseca e Silva
(Coorientadora)	

_____

Marcos Deon Vilela de Resende

(Orientador)

ORA, a fé o firme

fundamento das coisas

que se esperam,

e aprova das coisas

não se veem.

Hebreus 11:1

AGRADECIMENTOS

O tempo passa muito rápido:

Quero agradecer a Deus (Jesus) meu amigo fiel, a minha avó, mãe, esposa, filho, irmãs, irmãos, primas, primos, amigos (eles sabem quem são!), irmãos na fé, colegas, professores e outras pessoas que sempre acreditaram e continuam a acreditar em mim, apesar de meus muitos erros.

Hoje pela graças de Deus eu posso dizer que:

MISSÃO CUMPRIDA !!!!!!!!

Sem vocês este logro não seria possível!

Todas as coisas
Foram feitas
Por ele,
E sem ele nada
Do que foi feito
Se fez.

João 1:3

SUMÁRIO

RESUMO

DIAZ ESCOBAR, José Alfredo, M.Sc., Universidade Federal de Viçosa, fevereiro de 2016. **Distribuições de probabilidades do valor extremo e tamanho amostral para o melhoramento genético do quantil máximo em plantas.** Orientador: Marcos Deon Vilela de Resende. Coorientadores: Camila Ferreira Azevedo e Moyses Nascimento.

Dentre os objetivos dos programas de melhoramento genético de plantas de propagação assexuada (como a cana-de-açúcar e o eucalipto) e autógamas encontra-se o de selecionar indivíduos extremos ou segregantes transgressivos. Assim, é conveniente encontrar progênies com distribuições de caudas longas ou mesmo assimétricas, já que elas têm uma maior tendência de gerar indivíduos excepcionais. Os métodos de seleção comumente utilizados no melhoramento dessas espécies enquadram-se na classe BLUP sob os conceitos de média aritmética e média harmônica, os quais não levam em consideração a ocorrência de valores extremos dentro das famílias. Diante do exposto, este trabalho teve como objetivo propor e avaliar uma metodologia estatística para o melhoramento do máximo ou valor extremo das distribuições, e não necessariamente das médias das distribuições. Essa abordagem baseia-se nos quantis superiores da GEV (Distribuição de Valores Extremos Generalizada) dos BLUP´s genotípicos individuais entre e dentro de famílias, como forma de prever o aumento da ocorrência de valores extremos em função do aumento do tamanho da família (seleção de indivíduos extremos dentro de família) e também do número de famílias utilizado para representar uma população (seleção de indivíduos extremos em toda a população). A metodologia consistiu em usar dados simulados e reais, típicos das variáveis consideradas no melhoramento genético (por exemplo, distribuição normal com ampla variabilidade e presença de valores extremos). A partir dessa base de dados, distribuições de valores extremos generalizadas são ajustadas aos máximos de cada família, visando verificar qual a distribuição mais adequada (Gumbel, Fréchet, ou Weibull). Os resultados revelaram que a distribuição Weibull se ajusta melhor à bases de dados com 100 ou mais famílias e mais de 20 indivíduos por família e a distribuição Gumbel se ajusta melhor à bases de dados menores. Uma base de dados experimentais referentes à avaliação de famílias, mediante o uso de uma distribuição de valor extremo para predição do máximo das distribuições dos indivíduos, permite a

previsão do comportamento da eficiência seletiva para os máximos associados a vários tamanhos de famílias e de populações experimentais. Isso possibilita ao melhorista a otimização da experimentação no melhoramento visando a seleção de indivíduos extremos. Para essas previsões, emprega-se o ***período de retorno*** associado à ocorrência de um ***evento raro (nível de retorno)*** típico da distribuição ajustada. No caso, o período de retorno é interpretado como o ***tamanho amostral*** necessário para a ocorrência do nível de retorno do evento raro, interpretado como a ***magnitude do valor extremo***. Simulações estocásticas e reamostragens de dados experimentais indicaram consistentemente que a avaliação de 200 famílias em cada ciclo seletivo maximiza a eficiência do melhoramento visando a seleção de indivíduos extremos. Uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, perfazendo um total de 20000 indivíduos. Segundo a distribuição Weibull, o aumento da eficiência seletiva com o aumento do tamanho de família é em torno de 1,10 quando se passa de 20 para 100 indivíduos por família e de 1,12 quando se passa de 100 para 200 indivíduos e esses números são aproximadamente constantes independentemente do número de famílias avaliadas. Os modelos Gumbel e **Weibull** mostraram-se adequados para analisar as variáveis massa média de colmos (MMC em kg) e teor de Brix (B em %), sendo que a Gumbel mostrou-se adequada apenas nos casos de números de famílias muito pequenos. Assim, recomenda-se a Weibull para inferências práticas. A metodologia é adequada também para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos.

ABSTRACT

DIAZ ESCOBAR, José Alfredo, M.Sc., Universidade Federal de Viçosa, February, 2016**. Probability distributions of extreme value and sample size for the genetic improvement of maximum quantile in plants.** Adviser: Marcos Deon Vilela de Resende. Co-Advisors: Camila Ferreira Azevedo and Moyses Nascimento.

Among the objectives of programs of genetic improvement of asexual propagation of plants (such as sugarcane and eucalyptus) and autogamous is to select extreme or segregating individuals transgressive. It is therefore appropriate to find progenies distributions of long or asymmetrical tails, as they are more likely to generate exceptional individuals. Selective methods commonly used in the improvement of these species fall under the BLUP (Best Linear Unbiased Predictor) class under the concepts of arithmetic mean and harmonic mean, which do not take into account the occurrence of extreme values within families. Given the above, this study aimed to propose and evaluate a statistical methodology to improve the maximum or extreme value distributions, and not necessarily the means of distribution. This approach is based on the upper quantiles of GEV a (Generalized Extremes Values Distribution) of BLUP's individual genotypic between and within families, as a way to predict the increased occurrence of extreme values due to the increase in family size (selection of extreme individuals within family), and also the number of families used to represent a population (selection of extreme individuals in the population). The methodology consisted of using simulated and real data, typical of the variables considered in genetic improvement (eg, normal distribution with wide variability and the presence of extreme values). From this database, generalized extreme value distributions are adjusted to the maximum of each family, in order to ascertain the most appropriate distribution (Gumbel, Fréchet, or Weibull). The results showed that the Weibull distribution best fits the data bases with 100 or more families and more than 20 individuals per family and the Gumbel distribution fits better at smaller databases. A basis of experimental data relating to the evaluation of families, through the use of an extreme value distribution for predicting the maximum of the distribution of individuals, allows a prediction of the selection efficiency behavior to the maximum associated with various families and sizes of experimental populations. This enables the breeder

to optimize the experiment in breeding for the selection of extreme individuals. To these predictions, is employed the return period associated with the occurrence of a rare event (return level) typical of the fitted distribution. In this case, the return period is interpreted as the sample size required for the occurrence of the level of return the rare event, interpreted as the magnitude of the extreme value. Stochastic simulations and experimental data resampling consistently indicated that the evaluation of 200 families in each selection cycle to maximize efficiency improvement in order to select extreme individuals. A good practical option would be the evaluation of 200 families with 100 individuals, a total of 20,000 individuals. According to the Weibull distribution, the increase in selection efficiency with increasing family size is about 1.10 when going from 20 to 100 individuals per family and 1.12 when going from 100 to 200 individuals and these numbers they are approximately constant regardless of the number of families evaluated. The Gumbel and Weibull models have shown to be adequate to analyze the average mass variable stem (MMC kg) and Brix content (B %), and the Gumbel was adequate only in the case of very small families numbers. Thus, it is recommended to Weibull for practical inferences. The methodology is also suitable to classify the families or the progenies ability to generate superior or exceptional individuals and inform the sample sizes to be practiced in every family to capture these individuals.

1. INTRODUÇÃO GERAL

Dentre os objetivos dos programas de melhoramento genético de plantas de propagação assexuada (como a cana-de-açúcar e o eucalipto) e autógamas encontra-se o de selecionar indivíduos extremos ou segregantes transgressivos. Assim, é conveniente encontrar famílias com distribuições de caudas mais longas ou mesmo assimétricas, já que elas têm uma maior tendência de gerar indivíduos excepcionais. Os métodos de seleção comumente utilizados no melhoramento dessas espécies enquadram-se na classe BLUP sob os conceitos de média aritmética e média harmônica (RESENDE e BARBOSA, 2005), os quais não levam em consideração a ocorrência de valores extremos dentro das famílias. Diante do exposto, este trabalho teve como objetivo propor e avaliar uma metodologia estatística para o melhoramento do máximo ou valor extremo das distribuições, e não necessariamente das médias das distribuições. Essa abordagem baseia-se nos quantis superiores da GEV (Distribuição de Valores Extremos Generalizada) dos BLUP´s genotípicos individuais entre e dentro de famílias, como forma de prever o aumento da ocorrência de valores extremos em função do aumento do tamanho da família (seleção de indivíduos extremos dentro de família) e também do número de famílias utilizado para representar uma população (seleção de indivíduos extremos em toda a população).

Os valores extremos, por definição, são poucos, e suas estimações frequentemente são feitas para níveis de um processo que são muito maiores que os observados. Deste modo, o objetivo essencial da teoria dos valores extremos é a extrapolação da informação. Uma vez que não se tem fundamentos empíricos ou físicos para desenvolver uma regra de extrapolação, a teoria assintótica é utilizada para encontrar as distribuições limites dos valores extremos. A teoria de valores extremos (TVE) fornece uma base sólida e uma estrutura para a extrapolação, levando a estimadores naturais para as quantidades correspondentes, como são os quantis extremos; assim a TVE é reconhecida como uma disciplina única na estatística, porque gera técnicas e modelos para descrever o inusitado (raro) ao invés do habitual (COLES, 2001).

As ideias principais da teoria de valores extremos podem ser descritas da seguinte maneira: dado uma amostra de variáveis aleatórias independentes e identicamente distribuídas, pretende-se analisar o comportamento de qualquer

uma das estatísticas de ordem, como o mínimo, o máximo, ou o limiar de um nível crítico, onde as estatísticas de ordem devem ter funções de distribuição ***max-estável***, já que por definição as distribuições do máximo e o mínimo são degeneradas. Portanto, deve-se usar um método que permita aproveitar o argumento assintótico, como o teorema dos Tipos Extremos, para aproximar a distribuição da estatística de ordem por uma dentre as três famílias distribuições interessantes a este propósito, das quais a Gumbel (Tipo I), Fréchet (Tipo II), ou Weibull (Tipo III), que pertencem à classe DEV e que são casos especiais da GEV.

Os três tipos de famílias de distribuições são combinadas numa única família de distribuições com parametrização comum, sendo esta formulação chamada GEV. A GEV tem três parâmetros: μ, parâmetro de localização, σ, parâmetro de escala, e ξ, parâmetro de forma. Utilizar a GEV diminui muito o esforço estatístico e computacional, já que se pode realizar uma inferência de ξ, ou seja, a amostra escolhida determinará qual das famílias é a mais adequada para realizar as análises, sem necessidade de realizar suposições sobre o tipo de DEV que se deve adotar. Além disso, o desconhecimento que se tem do valor de ξ mede a incerteza sobre qual dos três tipos de famílias é a mais apropriada para analisar um conjunto de dados em particular (COLES, 2001).

É importante destacar que um dos campos com maiores aplicações da teoria do valor extremo é no planejamento de estruturas, que devem resistir a algum fenômeno ambiental como o nível do mar, velocidade do vento, nível de um rio ou represa, concentração de contaminantes, chuvas e ondas. Se o fenômeno é muito intenso, a estrutura falhará e desse modo, é necessário projetar está de maneira que a probabilidade de falha em função do evento natural extremo seja menor (COLES, 2001).

1.1. Objetivos

1.1.1. Objetivo Geral

Propor a utilização da teoria de valores extremos no programa de melhoramento genético de espécies de propagação assexuada como a cana-de-açúcar e eucalipto, visando inferir sobre o tamanho amostral adequado para capturar indivíduos extremos, ou seja, aqueles superiores ou excepcionais.

1.1.2. Objetivos Específicos

(i) Comparar o comportamento das diferentes classes de distribuições de probabilidade de valores extremos como ferramenta para inferência sobre o tamanho amostral necessário para a ocorrência de indivíduos extremos em programas de melhoramento genético de plantas.

(ii) Predizer o valor genotípico dos futuros indivíduos extremos em novas amostragens dentro de famílias e dentro de populações formadas por várias famílias.

(iii) Estipular os tamanhos amostrais necessários, em termos de número de indivíduos por famílias e número de famílias necessários para a ocorrência de indivíduos extremos.

(iv) Aferir uma metodologia para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos.

2. REFERENCIAL TEÓRICO

2.1. Teoria clássica de Valores Extremos

2.1.1. Distribuição do Máximo de uma amostra

Seja uma amostra de variáveis aleatórias (v.a.) $X_1, X_2 \ldots\ldots X_n$ independentes e identicamente distribuídas (i.i.d.) com função distribuição acumulada $F_X(x)$ e função densidade de probabilidade $f_X(x)$.

Seja também a variável M_n o máximo de uma amostra aleatória de tamanho *n,* definida por:

$$M_n = max[X_1, X_2 \ldots\ldots X_n] \quad \textbf{(1)}$$

Teoricamente, se a distribuição exata de X_i é conhecida, a função distribuição de M_n pode ser derivada para todos os valores de *n*:

$$\begin{aligned} F_{(M_n)}(x) &= \Pr[M_n \leq x] \\ &= Pr[X_1 \leq x, \ldots\ldots, X_n \leq x] \\ &= Pr[X_1 \leq x] * \ldots\ldots * Pr[X_n \leq x] \\ &= [F_X(x)]^n \end{aligned} \quad \textbf{(2)}$$

Na prática, o resultado anterior não é tão útil, já que a distribuição de X_i é desconhecida, portanto a função de distribuição $F_X(x)$ também. Uma alternativa para resolver esta dificuldade é trabalhar com métodos estatísticos padrão para estimar $[F_X(x)]^n$ com os dados observados, e, por conseguinte substituir a estimativa obtida em (2). Mas, quando se tem pequenas discrepâncias na estimativa de $F_X(x)$, estas podem provocar grandes discrepâncias em $[F_X(x)]^n$ (COLES, 2001).

Uma abordagem alternativa é aceitar $F_X(x)$ como desconhecida e buscar famílias de modelos aproximados para $[F_X(x)]^n$, que podem ser estimados aproveitando exclusivamente os dados extremos. O método anterior é semelhante ao procedimento de aproximação da distribuição das médias amostrais pela distribuição normal, justificado pelo Teorema Central do Limite. Os argumentos da teoria clássica de valores extremos são essencialmente um

análogo à teoria do teorema central do limite (COLES, 2001; BEIRLANT et al., 2004; HARRIS, 2004).

O interesse real está concentrado em encontrar possíveis distribuições limite para os máximos de amostras de variáveis aleatórias independentes e identicamente distribuídas. Seja $F_X(x)$ a função de distribuição base (inicial), y^* seu ponto ínfimo (inicial esquerdo) e x^* seu ponto supremo (final direito), ou seja, $y^* = \acute{i}nf\ \{x: F(x) > 0\} \geq \infty$ e $x^* = sup\ \{x: F(x) < 1\} \leq \infty$. Então

$$max\ (X_1, X_2, \ldots., X_n) \xrightarrow{p} x^*,\ n \to \infty \qquad \textbf{(3)}$$

Tem-se que se $\boldsymbol{x < x^*}$ então $\boldsymbol{F(x) < 1}$ assim $\boldsymbol{[F_X(x)]^n \to 0}$, e quando $\boldsymbol{x \geq x^*}$ então $\boldsymbol{F(x) = 1}$ assim $\boldsymbol{[F_X(x)]^n \to 1}$. Portanto a distribuição de $\boldsymbol{M_n}$ é degenerada e não apresenta valores de interesse prático. Por isso, com o objetivo de conseguir uma distribuição limite não degenerada, uma normalização linear da variável $\boldsymbol{M_n}$ é necessária (DE HAAN e FERREIRA, 2006; COLES, 2001; CASTILLO et al., 2005).

A normalização linear da variável M_n estará definida como:

$$M_n^* = \frac{M_n - b_n}{a_n} \qquad \textbf{(4)}$$

Assim

$$\lim_{n\to\infty} Pr\,[M_n^* \leq x] = \lim_{n\to\infty} Pr\left[\frac{M_n - b_n}{a_n} \leq x\right] = \lim_{n\to\infty} F^n\,(a_n x + b_n) \xrightarrow{d} G(x) \qquad \textbf{(5)}$$

Onde $G(x)$ é uma função distribuição não degenerada para todos os pontos de continuidade de x, assim, todas as funções de distribuição que são obtidas de (4) são chamadas de distribuições de valores extremos; onde $a_n > 0$ e $b_n \in R$ são sequências de constantes escolhidas apropriadamente para estabilizar a locação e a escala da variável M_n^* quando n aumenta e, impedir que tenha os problemas que surgem com a variável M_n (COLES, 2001; CASTILLO et al., 2005; DE HAAN e FERREIRA, 2006).

Logo se devem encontrar as condições necessárias e regulares sobre a distribuição inicial $F_X(x)$ tal que (4) não mude para cada uma das distribuições limite. A classe de distribuições que atende (4) são chamadas de domínio

máximo de atração ou simplesmente domínio de atração de $G(x)$ $\left(D_M(G)\right)$, (DE HAAN e FERREIRA, 2006), ou seja, se (4) é verificado (constantes a_n e b_n foram escolhidas apropriadamente) então as $F_X(x)$ associadas aos máximos tem uma distribuição limite não degenerada, portanto se pode afirmar que $F_X(x) \in D_M(G)$.

Para obter a distribuição limite não degenerada $G(x)$ apresenta-se alguns conceitos e resultados importantes para a formulação do Teorema de Distribuições de Valor Extremo.

2.1.2. Funções Distribuição

2.1.2.1. Funções de Distribuição Empírica

Seja as v.a. $X_1, X_2 \ldots\ldots X_n$ uma amostra simples cuja população tem função distribuição $F_X(x)$. Dada a amostra, a função de distribuição empírica $\widehat{F_n}(x)$ define-se como:

$$\widehat{F_n}(x) = \frac{Número\ de\ elementos\ da\ amostra \leq x}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,x]}(x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\} \qquad \textbf{(6)}$$

Onde $\mathbf{1}_{\{A\}}$ é a função indicadora do evento A.

Logo $\widehat{F_n}(x)$ é uma v.a. e pelo o teorema de Glivenko-Cantelli se conhece que $\widehat{F_n}(x)$ converge uniformemente para $F_X(x)$.

Teorema 1.1. (**Glivenko-Cantelli**) Seja $\widehat{F_n}(x)$ uma função distribuição empírica de uma amostra aleatória simples $X_1, X_2 \ldots\ldots X_n$ de uma função distribuição $F_X(x)$, então

$$sup_x \left|\widehat{F_n}(x) - F_X(x)\right| \to 0 \qquad \textbf{(7)}$$

Com probabilidade 1.

Para um x fixo,

$$\widehat{F_n}(x,\omega) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{x_j(\omega) \leq x\right\} \qquad \textbf{(8)}$$

Assim

$$E(\widehat{F_n}(x,\omega)) = \frac{1}{n}\sum_{i=1}^{n} E\left(\mathbf{1}_{\{x_j(\omega)\leq x\}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} P\left(x_j(\omega) \leq x\right)$$

$$= F_X(x) \quad \textbf{(9)}$$

De tal forma, que para cada x, $\widehat{F_n}(x)$ é um estimador não viesado de $F_X(x)$. Além disso, pela Lei Forte dos Grandes Números, para cada x existe um conjunto nulo A_x tal que

$$lím_{n\to\infty}\, \widehat{F_n}(x,\omega) = F_X(x) \quad \textbf{(10)}$$

sempre que $\omega \neq A_x$.

2.1.2.2. A Função Quantil

O conceito de nível de retorno que apresentara-se mais adiante, está baseado na definição da função quantil, pelo tanto é um conceito que deve-se estudar com muito detalhe.

A função quantil (f.q.) $Q_{(y)}$ é a inversa generalizada de $F_X(x)$:

$$Q_{(y)} = F^{-1}(y) = inf\{x \in R: F(x) \geq y\} \quad \textbf{(11)}$$

Para qualquer $y \in (0,1)$.

Propriedades (BOVIER, 2004; ORTEGA, 2010; AYALA e MONTES, 2010)

1. $Q_{(y)}$ **é crescente**.

Demonstração: Suponha que $y_1 < y_2$ ambos em (0,1), então $F(x) \geq y_2$ implica $F(x) \geq y_1$ assim $\{x: F(x) \geq y_1\} \supset \{x: F(x) \geq y_2\}$, logo:

$$inf\{x: F(x) \geq y_1\} \leq inf\{x: F(x) \geq y_2\}. \quad \textbf{(12)}$$

2. $F^{-1}\big(F(x)\big) \leq x.$

De modo que a consequência decorre:

$$F^{-1}\big(F(x)\big) = inf\{z: F(z) \geq F(x)\} \text{ e } x \in \{z: F(z) \geq F(x)\}. \quad \textbf{(13)}$$

3. $F\big(Q_{(y)}\big) \geq y$.

Demonstração: Se $x_n \in \{x: F(x) \geq y\}$ tal que $x_n \to x_0$ então o limite inf_n $F(x_n) \geq y$, já que $F(x)$ é monótona não decrescente e continua pela direita,

o limite $inf_n\ F(x_n) \leq F(x_0)$. Logo $F(x_0) \geq y$, ou seja, $x_0 \in \{x: F(x) \geq y\}$. Assim $\{x: F(x) \geq y\}$ é fechada e, portanto, contém o seu ínfimo. O que implica que $F\left(Q_{(y)}\right) \geq y$.

4. $\left(Q_{(y)}\right) \leq x$ se e somente se $y \leq F(x)$.

Demonstração: $\left(Q_{(y)}\right) \leq x$ implica $x \in \{z: F(z) \geq y\}$e assim $y \leq F(x)$. Por outro lado, se $y \leq F(x)$ então $x \in \{z: F(z) \geq y\}$ e portanto $\left(Q_{(y)}\right) \leq x$ já que $\left(Q_{(y)}\right)$ é o ínfimo.

5. Seja $X \sim F_X(x)$ e definamos $Y = aX + b$. A $F_Y(y)$ é:

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right) \quad \textbf{(14)}$$

Onde uma transformação de este tipo de uma mudança de localização e escala; tem a como o parâmetro de escala e b o de localização. A função quantil de Y é:

$$Q_Y(p) = aQ_X(p) + b \quad \textbf{(15)}$$

Isto é possível se $F_X(x)$ é continua e estritamente crescente, então Q_X é a inversa de $F_X(x)$. Portanto, uma transformação linear da variável X, produz transformação linear do mesmo tipo na função de quantil.

6. Seja $Y \sim U\ [0,1]$, se F é uma função de distribuição e Q é a função quantil correspondente, então $Q(Y)$ é uma variável aleatória em $[0,1]$, e se distribui como F.

$$P(Q(Y) \leq t) = P\big(Y \leq F(t)\big) = F(t) \quad \textbf{(16)}$$

Lema 1.1 (BOVIER, 2004; ORTEGA, 2010; AYALA e MONTES, 2010) Se $\psi\ (x)$ é uma função não decrescente e contínua à direita, com constantes $a > c$, $c \in R$ e $b \in R$. Se $H(x) = \psi(ax + b) - c$. Então:

1. $\psi\left(\psi^{-1}(x)\right) \geq x$ **.**

Demonstração: Ver propriedade número três da função quantil.

2. Se $\psi^{-1}(x)$ é continua em $\psi(x) \in R$, então $\psi^{-1}\left(\psi(x)\right) = x$.

Demonstração:

Como $\psi^{-1}\left(\psi(x)\right) = inf\{x': \psi(x') \geq \psi(x)\}$, assim $\psi^{-1}\left(\psi(x)\right) \leq x$. Por outro lado, para qualquer $\varepsilon > 0$, $\psi^{-1}(\psi(x) + \varepsilon) = inf\{x': \psi(x') \geq \psi(x) + \varepsilon\}$. Mas $\psi(x')$ só pode ser estritamente maior que $\psi(x)$ se $x' > x$, portanto para

qualquer $y' > \psi(x)$, $\psi^{-1}(y') \geq x$. Assim, se ψ^{-1} é continua em $\psi(x)$, isto implica que $\psi^{-1}\big(\psi(x)\big) = x$.

3. $H^{-1}(x) = a^{-1}(\psi^{-1}(y + c) - b)$

Demonstração: $H^{-1}(y) = inf(x: \psi(ax + b) - c \geq y)$

$$= inf[x: \psi(ax + b) \geq y + c]$$

$$= inf\left[\frac{1}{a}(ax + b) - b: \psi(ax + b) \geq y + c\right]$$

$$= \{a^{-1} inf[(ax + b): \psi(ax + b) \geq y + c] - b\}$$

$$= \{a^{-1}[\psi^{-1}(y + c)] - b\} \quad \textbf{(17)}$$

4. Se $G(x)$ é uma função de distribuição não degenerada, então existem $y_1 < y_2$**,** tais que $G^{-1}(y_1) < G^{-1}(y_2)$**.**

Demonstração: Se $G(x)$ é uma função de distribuição não degenerada, então existem $x_1 < x_2$, tais que $0 < G^{-1}(x_1) \equiv y_1 < G^{-1}(x_2) \equiv y_2 \leq 1$. Mas então $G^{-1}(y_1) \leq x_1$ e $G^{-1}(y_2) = inf\{x: G(x) \geq G(x_2)\}$. Se $G^{-1}(y_2) = x_1$, então para todo $x \geq x_1$, $G(x) \geq G(x_2)$, e dado que $G(x)$ é continua à direita, $G(x_1) = G(x_2)$, que é uma contradição.

2.1.3. Convergência de Funções Monotônicas

Esta seção foi baseada em Bovier (2004), Ortega (2010), Ayala e Montes, (2010).

Para compreender de forma correta os teoremas 1.3 e 1.4 é indispensável entender alguns conceitos de convergência de funções monótonas.

Definição 1.1 Convergência débil de medidas de probabilidade

Para qualquer função H escrevemos:

$$\mathrm{C}(H) = \{x \in R: H \text{ é } finita\ e\ continua\ em\ x\}$$

Uma sucessão de funções não decrescentes $\{H_n, n \geq 1\}$ em $\mathbb{R}$ converge debilmente para H_0 quando $n \to \infty$. O denotamos por:

$$H_n(x) \to H_0(x), \ \forall x \in \mathrm{C}(H_0) \quad \textbf{(18)}$$

Se F_n e F_0 denotam as funções de distribuição de $H_n(x)\ e\ H_0(x)$ respetivamente, então diremos que F_n converge debilmente para F_0 se o

$\lim_{n\to\infty} F_n(x) = F_0(x)$ para qualquer x ponto de continuidade de F_0, o denotamos por $F_n \xrightarrow{\omega} F_0$.

Proposição 1.1 Se $\{H_n, n \geq 1\}$ são funções não decrescentes e $H_n(x) \xrightarrow{\omega} H_0(x)$ então $H_n^{-1}(x) \xrightarrow{\omega} H_0^{-1}(x)$. Ver demonstração em Ortega (2010).

Definição 1.2 Convergência em Probabilidade (Lei) de variáveis aleatórias

Sejam $\{X_n, n \geq 1\}$ e X variáveis aleatórias definidas sobre um mesmo espaço de probabilidade, com distribuições de probabilidade $\{f_n(x), n \geq 1\}$ e $f(x)$ respetivamente, então a sucessão $\{X_n, n \geq 1\}$ converge em lei para X, $X_n \xrightarrow{L} X$ ou $L(X_n) \to L(X)$, se $f_n(x) \xrightarrow{\omega} f(x)$, ou equivalentemente, se $\lim_{n\to\infty} P(X_n \leq x) = P(X \leq x)$ para todo x tal que $P(X = x) = 0$.

2.1.4. Teorema de convergência as classes

Seção baseada em Bovier (2004), Ortega (2010) e Beirlant et al., (2004).

É comum encontrar na literatura que a maioria dos resultados de convergência de variáveis aleatórias é do seguinte tipo: Para uma sucessão de v.a. $\xi_n (n \geq 1)$ e constantes $a_n > 0$ e $b_n \in R$, se pode demonstrar que:

$$\frac{\xi_n - b_n}{a_n} \Longrightarrow Y \qquad \textbf{(19)}$$

Onde Y é uma v.a. não degenerada. Portanto temos

$$P\left[\frac{\xi_n - b_n}{a_n} \leq x\right] \approx P(Y \leq x) = G(x), \qquad \textbf{(20)}$$

Se $y = a_n x + b_n$ temos que:

$$P[\xi_n \leq y] \approx G\left(\frac{y - b_n}{a_n}\right). \qquad \textbf{(21)}$$

Definição 1.3. Duas funções de distribuição, W e H são chamadas "do mesmo tipo" ou pertencem à mesma família se e somente se, existem constantes $a_n > 0$ e $b_n \in R$, de modo que

$$W(x) = H(ax + b), \quad \forall\, x \in R \qquad \textbf{(22)}$$

Em termos de v.a., se $L(X) = W$ e $L(Y) = H$ então

$$Y \xrightarrow{d} \left(\frac{X-b}{a}\right) \quad \textbf{(23)}$$

O resultado e a definição anterior permite aproximar a distribuição de ξ_n por uma família de distribuições com parâmetros de localização e escala. Mas uma pergunta logica seria, são únicas as constantes de normalização a_n e b_n?, a resposta a esta questão se pode encontrar no teorema de convergência para famílias de distribuições: as constantes e a distribuição limite estão determinadas salvo por equivalências assintóticas e, por parâmetros de localização e escala respetivamente (ORTEGA, 2010).

Corolário 1.1. Se $G(x)$ é uma função de distribuição não degenerada, e existem constantes $a > 0$, $\alpha > 0$, e $b \in R$, e $\beta \in R$, de tal modo que para todo $x \in R$,

$$G(ax + b) = G(\alpha x + \beta) \quad \textbf{(24)}$$

assim $a = \alpha$ e $b = \beta$.

Demonstração: seja $H(x) = G(ax + b)$, então pelo o item três (3) do lema 1.1 temos que$H^{-1}(y) = a^{-1}\,(G^{-1}(y) - b)$ mas também se pode escrever como $H^{-1}(y) = \alpha^{-1}\,(G^{-1}(y) - \beta)$, porque $G(ax + b) = G(\alpha x + \beta)$. Pelo o item quatro do lema 1.1 se tem que, existem pelo menos dois valores de y tais que para $G^{-1}(y)$ eles são diferentes, ou seja, existem valores $x_1 < x_2$ de tal modo que

$$a^{-1}(x_i - b) = \alpha^{-1}(x_i - \beta) \quad \textbf{(25)}$$

Logo se tem que $a = \alpha$ e $b = \beta$.

Teorema 1.2 (De Khintchine ou Convergência as Famílias).

Seja F_n $(n \in N)$ funções de distribuição e, $G(x)$ uma função de distribuição não degenerada. Seja uma sequência de constantes $a_n > 0$ e $b_n \in R$ de modo que

$$F_n(a_n x + b_n) \xrightarrow{\omega} G(x) \quad \textbf{(26)}$$

Segundo o corolário 1.1 então se tem que se existem uma sequências de constantes $\alpha_n > 0$ e $\beta_n \in R$, e $G_X(x)$ é uma função de distribuição não degenerada cumpre-se que:

$$F_n(\alpha_n x + \beta_n) \xrightarrow{\omega} G_X(x) \quad \textbf{(27)}$$

se e somente se

$$\frac{\alpha_n}{a_n} \to a \qquad e \qquad \frac{\beta_n - b_n}{a_n} \to b \tag{28}$$

Onde $a_n > 0$, $b_n \in R$ e $n \to \infty$, e $G_X(x) = G(ax + b)$ **(29)**

Este teorema faz que o corolário 1.1 seja preciso, dizendo que as diferentes escolham da escala de sequências a_n, b_n somente podem conduzir às distribuições que são relacionadas por uma transformação. Ver demonstração em Ortega (2010) e Bovier (2004).

Definição 1.4. Seja $X_1, X_2 \ldots . X_n$ v.a.i.i.d. com função de distribuição $F_X(x)$. A função de distribuição $F_X(x)$ é chamada ***max-estável*** se para alguma eleição de constantes $a_n > 0$ e $b_n \in R$ tem-se que

$$\begin{aligned} Pr\left[\frac{M_n - b_n}{a_n} \leq x\right] &= F^n(a_n x + b_n) \\ &= Pr(X \leq x) \\ &= F(x). \end{aligned} \tag{30}$$

A max-estabilidade é uma propriedade satisfeita pelas distribuições as quais quando se pega uma amostra de máximos, leva a uma função de distribuição idêntica das variáveis iniciais, além de uma mudança de locação e escala (COLES, 2001; BEIRLANT et al., 2004).

Outras formulações equivalentes da definição de ***max-estável*** são as seguintes:

Definição 1.5. A função de distribuição não degenerada $G(x)$, é chamada ***max-estável,*** si para todo $n \in N$, existem $a_n > 0$ e $b_n \in R$, tais que, para todo $x \in R$,

$$G^n(a_n x + b_n) = G(x) \tag{31}$$

Proposição 1.1

I. A função de distribuição $G(x)$, é ***max-estável,*** se e somente se, existem funções de distribuição F_n e $a_n > 0$ e $b_n \in R$, tal que, para todo $k \in N$,

$$F_n(a_{nk} x + b_{nk}) \xrightarrow{\omega} G^{1/k}(x) \tag{32}$$

II. $G(x)$ é ***max-estável,*** se e somente se, existe a função de distribuição $F_X(x)$, $a_n > 0$ e $b_n \in R$, de modo que

$$F^n(a_n x + b_n) \xrightarrow{\omega} G(x) \quad \textbf{(33)}$$

2.1.5. Teorema dos Tipos Extremos (DEV)

Nas sessões anteriores definiu-se que as únicas distribuições que podem ser distribuições extremas são as distribuições ***max-estável,*** portanto, as famílias de distribuições limite não degeneradas que se podem encontrar como resultado de 3. chamam-se distribuições de valores extremos.

Teorema 1.3 (Fisher –Tippet - Gnedenko) Qualquer distribuição ***max-estável*** $G(x)$é do tipo valor extremo, ou seja, igual a $G(ax + b)$, e se existem sequências de constantes $a_n > 0$ e $b_n \in R$, tais que

$$Pr\left[\frac{M_n - b_n}{a_n} \leq x\right] \xrightarrow{\omega} G(x), \text{ quando } n \to \infty \quad \textbf{(34)}$$

então $G(x)$ pertence a uma das famílias de distribuições seguintes:

Distribuição Gumbel (Tipo I):

$$G(x) = exp\left\{-exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, -\infty < x < \infty; \quad \textbf{(35)}$$

Distribuição Fréchet (Tipo II):

$$G(x) = \begin{cases} 0, & se\ x \leq b, \\ exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, & se\ x > b, \end{cases} \quad \textbf{(36)}$$

Distribuição Weibull (Tipo III):

$$G(x) = \begin{cases} exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{-\alpha}\right]\right\}, & se\ x \leq b, \\ 1, & se\ x > b, \end{cases} \quad \textbf{(37)}$$

Para parâmetros $a > 0$ e $b \in R$, e no caso das famílias **Fréchet e Weibull** $\alpha > 0$.

Então o teorema indica os máximos da amostra com nova escala $\frac{M_n - b_n}{a_n}$, convergem debilmente para uma nova variável que tem uma distribuição dentro

de umas das famílias apresentadas anteriormente; estas famílias de distribuições são conhecidas em conjunto, como Distribuições de Valores Extremos (DEV), todas as famílias tem parâmetro de escala a e de localização b, mas as famílias Fréchet e Weibull tem um parâmetro de forma α.

É importante destacar que o teorema 1.3 é um análogo para os valores extremos do teorema do limite central, já que quando M_n é estabilizado com sequências adequadas de a_n e b_n, a variável padronizada M_n^* tem uma distribuição limite que é uma das famílias DEV, ou seja, as distribuições de valores extremos são os únicos limites possíveis das distribuições de M_n^*, sem importar a função de distribuição $F_X(x)$ da população (COLES, 2001; BEIRLANT et al., 2004; HARRIS, 2004).

2.1.6. Distribuição de Valores Extremos Generalizada (GEV)

As distribuições de valores extremos comportam-se de formas distintas devido às diversas formas de comportamento da cauda para a $F_X(x)$. Isto se pode fazer por necessidade, já que tem na conta o desempenho do limite da distribuição $G(x)$ em seu ponto final direito x^*, note-se que x^* é finito para as famílias de distribuições Weibull, enquanto tanto x^* é infinito para as famílias de distribuições Gumbel e Fréchet. No entanto, a densidade do $G(x)$ decai exponencial e polinomialmente para as distribuições Gumbel e Fréchet respetivamente, com relação ás taxas relativamente diferentes de decaimento na cauda do $F_X(x)$; portanto em aplicações as três famílias oferecem resultados totalmente diferentes do comportamento do valor extremo (COLES, 2001).

Nos primeiros trabalhos dos valores extremos era frequente assumir uma das três famílias como a mais próxima aos dados observados, e após estimar os parâmetros relevantes da distribuição selecionada, porém este procedimento tem dois pontos fracos: primeiro, para selecionar qual das famílias é a mais adequada para a amostra escolhida, precisa-se trabalhar com uma técnica especializada; segundo, logo que se toma a decisão anterior, todas as inferências feitas presumem que essa seleção é verdadeira, y não tem na conta as incertezas que tal decisão implica, embora essas incertezas possam ser substanciais (COLES, 2001). Para contornar os problemas anteriores e realizar

análises melhores Von Mises (1936 -1954) e Jenkinson (1955), reformularam o teorema 1.3, ou seja, os três tipos de famílias de distribuições foram combinadas numa única família de distribuições com parametrização comum, a reformulação é chamada Distribuições de Valores Extremos Generalizada (GEV), que tem a seguinte função de distribuição:

$$G(x|\mu,\sigma,\xi) = exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, x \in \mathbb{R} \quad \textbf{(38)}$$

Definido no conjunto $\left\{x: 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, e para parâmetros $-\infty \leq \mu \leq \infty$, $\sigma > 0$, e $-\infty \leq \xi \leq \infty$.

A GEV tem três parâmetros: μ, parâmetro de localização, σ, parâmetro de escala, e ξ, parâmetro de forma.

Com esta nova parametrização para $\xi > 0$ tem-se a distribuição Fréchet, se $\xi < 0$ a distribuição obtida é Weibull, e tem-se como limite à família de distribuições Gumbel quando $\xi \to 0$, portanto o parâmetro ξ determina o tipo da distribuição (COLES, 2001; BEIRLANT et al., 2004; DE HAAN e FERREIRA, 2006).

Trabalhar com a GEV diminui a implementação estatística, já que podem-se realizar inferências com relação ao índice de valor extremo ξ (BEIRLANT et al., 2004; DE HAAN e FERREIRA, 2006), ou seja, a amostra escolhida determina qual das famílias é a mais adequada para realizar as análises, sem necessidade de realizar juízos subjetivos iniciais sobre o tipo de DEV adotar. Além disso, a incerteza que se tem do valor de ξ mede a falta de certeza sobre qual dos três tipos de famílias é a mais apropriada para analisar um conjunto de dados em particular (COLES, 2001). Logo o teorema 1.3 pode-se reescrever pelo teorema 1.4.

Teorema 1.4 (Von Mises – Jenkinson) Qualquer distribuição ***max-estável*** $G(x)$ é do tipo valor extremo, ou seja, igual a $G(ax+b)$ e, se existem sequências de constantes $a_n > 0$ e , $b_n \in R$ tais que

$$Pr\left[\frac{M_n - b_n}{a_n} \leq x\right] \xrightarrow{\omega} G(z), \text{ quando } n \to \infty \quad \textbf{(39)}$$

então $G(x)$ pertence a família de distribuições GEV**:**

$$G(x|\mu,\sigma,\xi) = exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, x \in \mathbb{R} \quad \textbf{(40)}$$

Definido no conjunto$\left\{x: 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, e para parâmetros $-\infty \leq \mu \leq \infty$, $\sigma > 0$, e $-\infty \leq \xi \leq \infty$.

Então se pode concluir pelo teorema 1.4 que quando se tem amostras suficientemente grandes é prático aproximar a distribuição M_n^*, através de um elemento da família de distribuições GEV; onde a determinação das constantes de normalização a_n e b_n se obtém pelo cumprimento do teorema 1.2 e as definições 1.4 e 1.5. Assim a função de distribuição limite de M_n também pode ser aproximada por um elemento da família de distribuições GEV; portanto na prática termina sendo irrelevante que os parâmetros da M_n^* sejam diferentes dos de M_n (COLES, 2001; EMBRECHTS et al., 1997).

2.1.7 Método Do Bloco Máximo

2.1.7.1 Modelo Básico

Método criado por razão do comportamento assintótico do máximo normalizado de uma amostra aleatória, ou seja, para uma amostra escolhida aleatoriamente, só existe um "único" máximo, portanto não se poderiam estimar os parâmetros μ, σ e ξ somente com uma observação extrema. Deste modo cria-se o Método Do Bloco Máximo, que consiste em dividir a amostra de tamanho **m,** em **b** blocos de igual tamanho **n** (**n** suficientemente grande), logo obter os valores máximos $(M_{n,1}, M_{n,2}, \ldots., M_{n,b})$ de cada bloco (BEIRLANT et al., 2004; COLES, 2001)**.**

Os blocos são independentes e identicamente distribuídos, mas os elementos de cada bloco podem ser dependentes. Nas seções anteriores foi comprovado que se existe um limite não degenerado, as famílias de valores extremos são as únicas formas restritivas (limites) possíveis para um máximo normalizado de uma amostra aleatória, portanto com a nova amostra $M_{n,1}, M_{n,2}, \ldots., M_{n,b}$ podem ser estimados μ, σ, e ξ mediante o ajuste de GEV (utilizar o teorema 1.4), e logo determinar qual é a família de distribuições mais

adequada para a amostra escolhida (GUMBEL, 2004; EMBRECHTS et al., 1997).

Nas diferentes ciências onde utiliza-se a teoria dos valores extremos e trabalham com o método do bloco máximo, frequentemente os blocos são equivalentes a um período de tempo como anos, semestres, trimestres, meses ou dias. Em todas as situações **n** significa o número de observações no período de tempo determinado, e os máximos dos blocos representam os máximos no período de tempo.

2.1.8 Nível de Retorno

Na teoria de valores extremos a estimação de quantis altos das GEV e DEV, tem muita relevância, já que o quantil x_p, tal que $G(x_p) = 1 - p$ é conhecido como **nível de retorno** o qual está associado ao **período de retorno** $\frac{1}{p}$**.** Portanto se tem-se uma amostra aleatória simples $X_1, X_2 \ldots . X_n$ de $G(x)$. Dado que qualquer $p \in (0,1)$ define-se o quantil p por:

$$Q_{(p)} = x_p = G_x{}^{-1}(p) \qquad \textbf{(41)}$$

Segundo a definição de $G(x)$ na equação (40), a função quantil é:

$$Q(p) = \begin{cases} \mu - \frac{\sigma}{\xi}(1 - (-\log p)^{-\xi}), & para\ \xi \neq 0, \\ \mu - \sigma \log(-\log p), & para\ \xi = 0, \end{cases} \qquad \textbf{(42)}$$

Devido que os quantis permitem aos modelos de probabilidade ser expressados pela escala dos dados, a relação do modelo GEV com seus parâmetros pode-se interpretar melhor em termos das expressões quantis de (42) (COLES, 2001). Em particular, se $y_p = -\log(1 - p)$ de modo que:

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left(1 - y_p{}^{-\xi}\right), & para\ \xi \neq 0, \\ \mu - \sigma \log y_p, & para\ \xi = 0, \end{cases} \qquad \textbf{(43)}$$

Logo com uma análise gráfica utilizando os níveis de retorno (quantis x_p), pode-se determinar qual das famílias de DEV é a mais adequada para realizar

as análises da amostra escolhida. O gráfico é construído com x_p em função de y_p ou $logy_p$. Consequentemente, se o gráfico exibir um comportamento linear, côncavo ou convexo, tem-se suficiente evidencia para afirmar que $\xi = 0$, $\xi < 0$ ou $\xi > 0$ respectivamente (COLES, 2001; BEIRLANT et al., 2004; DE HAAN e FERREIRA, 2006).

Os gráficos dos níveis de retorno são úteis e importantes, eles contém os períodos (escala logarítmica) e níveis de retorno estimados, permitindo apresentar e validar a distribuição pertinente para a análise; devido à escolha da escala o gráfico condensa a cauda da distribuição, logrando que o efeito da extrapolação seja destacado, ou seja, as estimativas de x_p para os maiores $\frac{1}{p}$ podem-se observar no gráfico (KETCHEN et al., 2006; COLES, 2001).

Uma interpretação dos níveis de retorno pode ser a seguinte: dado um nível razoável de precisão, espera-se que o nível de retorno x_p seja superado em média uma vez em $\frac{1}{p}$ períodos de tempo, ou seja, o x_p é superado por o M_n em um período de tempo qualquer com probabilidade $\boldsymbol{p}$ $\left(\boldsymbol{p = 1 - G(x_p)}\right)$ (COLES, 2001; GUMBEL, 2004; REISS e THOMAS, 2007) .

É importante ressaltar que a interpretação está em termos de média, porque se considera que o número de anos até que o M_n supere o x_p como uma variável aleatória, a qual possui uma distribuição Geométrica com parâmetro $\boldsymbol{p}$. Também se deve destacar que a probabilidade $\boldsymbol{p}$, pode ser chamada **probabilidade de retorno**, porque é empregada nas situações onde o objetivo é obter as magnitudes de eventos com probabilidades $\boldsymbol{p}$ de ocorrência (CASTILLO et al., 2005; KOTTEGODA e ROSSO, 2008).

Na figura 1 apresenta-se um exemplo do gráfico de probabilidade da distribuição Gumbel, o qual proporciona a associação dos níveis e períodos de retorno.

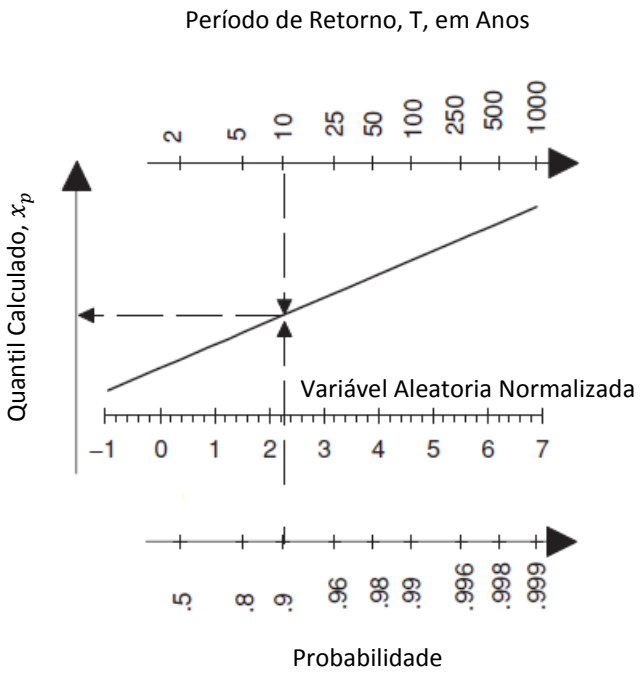


Figura 1. Gráfico de probabilidade Gumbel – Fonte: Kottegoda, N. e Rosso R., (2008)

2.2. Inferências Para a Distribuição de Valores Extremos Generalizada

A GEV proporciona um modelo para a distribuição encontrada com o método do bloco máximo. Segundo o estudado na seção 2.1.6 a distribuição de GEV tem três parâmetros: μ, parâmetro de localização, σ, parâmetro de escala, e ξ, parâmetro de forma. Portanto a estimação de os parâmetros faz-se necessária e importante, para lograr este propósito se podem utilizar diferentes técnicas de estimação paramétricas.

As técnicas estimação paramétricas usuais para a GEV são os Método de Máxima Verossimilhança (MML), Máxima Verossimilhança Generalizada (GMML), Momentos pesados por Probabilidades (PWM) ou também chamado L-Momentos de una Distribuição de Probabilidade, e os Métodos Bayesianos (MB), no entanto, em está seção só explicaram-se o MML e os MB, já que estes são os de maior interesse para o autor.

2.2.1 Método de Máxima Verossimilhança (MML) para a GEV

Seja $g(x)$ a função de densidade de $G(x)$. A função de verossimilhança baseada no vetor de dados $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ é dada por

$$L(\theta; \mathbf{X}) = \prod_{\boldsymbol{i}}^{\boldsymbol{n}} \boldsymbol{g_\theta}(\mathbf{X_i}) \mathbf{1}_{\{1+\xi(X_i-\mu)/\sigma>0\}} \qquad (\mathbf{44})$$

Onde $\mathbf{1}_{\{A\}}$ é a função indicadora do evento A.

então

$$\hat{\theta}_n = \arg \max_{\theta \epsilon \Theta} \ell(\theta; \mathbf{X}) \qquad \mathbf{(45)}$$

No caso $\xi \neq 0$ temos

$$\ell(\xi, \mu, \sigma; \mathbf{X}) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \log\left(1 + \xi\left(\frac{X_i - \mu}{\sigma}\right)\right) - \sum_{i=1}^{m} \left(1 + \xi\left(\frac{X_i - \mu}{\sigma}\right)\right)^{-1/\xi}. \quad (\mathbf{46})$$

No caso $\xi = 0$ temos

$$\ell(0, \mu, \sigma; \mathbf{X}) = -n \log \sigma - \sum_{i=1}^{m} \exp\left\{-\frac{X_i - \mu}{\sigma}\right\} - \sum_{i=1}^{m} \frac{X_i - \mu}{\sigma}. \qquad (\mathbf{47})$$

Onde ℓ é $\log L(\theta; \mathbf{X})$.

Diferenciando esta última função com relação μ e σ e igualando a 0 se tem:

$$n - \sum_{i=1}^{n} \exp\left\{-\frac{X_i - \mu}{\sigma}\right\} = 0. \qquad (\mathbf{48})$$

$$n + \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma} \left(\exp\left\{-\frac{X_i - \mu}{\sigma}\right\} - 1\right) = 0. \qquad (\mathbf{49})$$

Logo não tem-se uma solução explicita para equações (48) e (49), por conseguinte precisa-se de métodos numéricos que obtenham os estimadores desejados. Entretanto deve-se ressaltar que para as GEV o método de máxima verossimilhança apresenta uma dificuldade pratica, as condições de regularidade que se necessitam para que as propriedades assintóticas geralmente associadas com os estimadores de máxima verossimilhança não são

satisfeitas, em razão de que os extremos (suporte) da GEV são uma função dos valores dos parâmetros (COLES, 2001; EMBRECHTS et al., 1997). Portanto deve-se tomar precaução quando se trabalha com o método de máxima verossimilhança.

Deve ser ressaltado que se $\xi > -0.5$ os estimadores de máxima verossimilhança tem as propriedades assintóticas usuais (consistência, eficiência, invariância e normalidade), no caso que $\xi < -1$ se tem pouca probabilidade de obter estimadores de máxima verossimilhança, enquanto que quando $-1 < \xi < -0.5$ é possível obter os estimadores mas não tem as propriedades assintóticas desejadas (SMITH, 1985).

2.2.2 Inferência Bayesiana

Utilizar a inferência bayesiana é uma ótima alternativa para realizar análises de valores extremos, já que a natureza preditiva da analises de valores extremos tem relação com o conceito de predição posterior, podem-se usar outras fontes de informação independentes dos dados disponíveis, ademais os métodos bayesianos não dependem das condições de regularidade requeridas para os MML e PWM (COLES, 2001; BEIRLANT et al., 2004; REISS e THOMAS, 2007).

A $GEV\ (\mu, \sigma, \xi)$ é o modelo adequado para estudar o comportamento do máximo anual Z de um processo qualquer, se a estimativa de θ $(\theta = \mu, \sigma \text{ e } \xi)$ é obtida através de uma análise Bayesiana, o resultado é uma distribuição a posteriori $f(\theta|x)$ e uma função de densidade preditiva que fica definida assim:

$$Pr\{\mathrm{Z} \leq z|x_1, \ldots\ldots, x_n\} = \int_{\Theta} Pr\{\mathrm{Z} \leq z|\theta\} f(\theta|x)\, d\theta$$

$$= 1 - {}^{1}/_{m} \qquad \textbf{(50)}$$

Onde $Pr\{\mathrm{Z} \leq z|\theta\}$ é a GEV avaliada em z, $f(\theta|x) = f(\theta)L(\theta; \mathbf{X})$ e $f(\theta)$= Distribuição a priori. As distribuições a priori mais utilizadas são:

a. Distribuição Normal Trivariada em $\theta' = (\mu, \log\ \sigma, \xi)$ levando à densidade a priori:

$$f(\theta) \propto \frac{1}{\sigma}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta'} - \boldsymbol{\nu})^T\Sigma^{-1}(\boldsymbol{\theta'} - \boldsymbol{\nu})\right\}.$$

Abordagem utilizada por Coles e Powel (1996). Onde $\boldsymbol{\nu}$ = vetor de médias e Σ = matriz de covariâncias.

b. Distribuição Normal Trivariada em $\theta' = (\log\mu, \log\sigma, \xi)$ levando à densidade a priori:

$$f(\theta) \propto \frac{1}{\mu\sigma}\exp\left\{-\frac{1}{2}(\boldsymbol{\theta'} - \boldsymbol{\nu})^T\Sigma^{-1}(\boldsymbol{\theta'} - \boldsymbol{\nu})\right\}.$$

A parametrização log-normal para o parâmetro de localização pode ser útil, se um limite inferior para este parâmetro é necessário. Onde $\boldsymbol{\nu}$ = vetor de médias e Σ = matriz de covariâncias (STEPHENSON e RIBATET, 2006).

Stephenson e Ribatet (2006) usam o método de simulação de Monte Carlo via Cadeias de Markov (MCMC) $\theta_b, \ldots, \theta_n$ e concluem que as funções preditivas podem ser estimadas usando:

$$\frac{1}{n-b+1}\sum_{i=b}^{n}\Pr(\mathrm{Z} \leq z|\theta_i) \qquad (\mathbf{51})$$

É importante destacar que Coles (2001) e Beirlant et al. (2004) também propõem o método de MCMC, para a obtenção das distribuições a posteriori e distribucoes preditivas das GEV. A equação (50) apresenta a distribuição de um máximo anual futuro de um processo, proporcionando a aleatoriedade em observações futuras e a incerteza dos parâmetros estudados. Pelo tanto fornece um valor similar do nível de retorno de m anos, que claramente introduz a incerteza por causa da estimação do modelo (COLES, 2001), ou seja, se supomos que $Pr(\mathrm{Z} > z|x) = p$ ou que $Pr(\mathrm{Z} > z) = p$, tal que z é o nível de retorno correspondente para o período de retorno $\frac{1}{p}$. Para cada valor de z, pode-se obter a estimativa de p utilizando (51) (STEPHENSON e RIBATET, 2014).

2.2.3 Funções do Perfil da Verossimilhança para Intervalos de Confiança

Seja um espaço paramétrico Θ com dimensão n com elementos $\theta = (\theta_1, \ldots., \theta_n)$. Suponha que θ se pode dividir em dois componentes ($\theta^{(1)}$, $\theta^{(2)}$)

e o interesse somente está em $\theta^{(1)}$, de tal forma que $\theta^{(2)}$ são parâmetros de ruído. Para estimar $\theta^{(1)}$ podemos usar o perfil da verossimilhança que se define como:

$$L_p(\theta^{(1)}) = \max_{\theta^{(2)}|\theta^{(1)}} \mathrm{L}(\theta^{(1)}, \theta^{(2)}) \qquad \textbf{(52)}$$

Onde L representa a função de verossimilhança.

Logo se pode definir o perfil da verossimilhança como o desempenho da função de verossimilhança com base em parâmetros estabelecidos com antecedência (SPROTT, 2000), ou seja, o perfil da verossimilhança se obtém maximizando a função de verossimilhança avaliada nos elementos de Θ com $\theta^{(1)}$ fixo (ORTEGA, 2010; CASTILLO et al., 2005).

A avaliação numérica do perfil da verossimilhança para qualquer um dos parâmetros μ, σ e ξ é obtida de uma forma relativamente simples. Se quisermos obter o perfil da verossimilhança de ξ, fixa-se ξ= ξo e maximizamos a log-verossimilhança com relação aos parâmetros restantes. Logo se reproduz isso para um determinado número de valores de ξ=ξo. Os valores associados da log-verossimilhança constituem o perfil da log-verossimilhança para ξ, baseado nestes fatos pode-se obter os intervalos de confiança aproximados (COLES, 2001; CASTILLO et al., 2005).

A metodologia anterior pode-se aplicar quando precisamos empregar a inferência e têm-se combinações de parâmetros. Portanto pode ser utilizada para obter intervalos de confiança de níveis de retorno específicos ou parâmetros. Com referência aos intervalos que se obtêm apoiados no método delta, os intervalos obtidos com o Perfil da Verossimilhança são usualmente assimétricos, revelando a assimetria que existe em estas situações da função de verossimilhança (COLES, 2001; BEIRLANT et al., 2004).

2.2.4 Intervalos de Verossimilhança e Aproximação χ^2

Para conferir que tão verosímil são os diferentes valores de θ pode-se utilizar a função de verossimilhança relativa, que é definida como:

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta}_n)} \qquad \textbf{(53)}$$

Ao conjunto $\theta: \mathbb{R}(\theta) > l$ com $l \in (0,1)$ é chamado região de verossimilhança, no entanto que l é nível de verossimilhança. Quando a dimensão de θ é um (1) estas regiões chamam-se intervalos de verossimilhança.

As regiões de verossimilhança fazem referencia somente à pertinência ou razoabilidade relativa dos diferentes valores do parâmetro θ, e não à incerteza do intervalo (ORTEGA, 2010). Em algumas situações é factível aproximar a probabilidade que estas regiões contêm o verdadeiro valor do parâmetro. Uma maneira de realizar este é com a função Deviance.

$$D_n(\theta) = 2\left(\ell\left(\hat{\theta}_n\right) - \ell(\theta)\right) = -2\log\left(R(\theta)\right) \quad \textbf{(54)}$$

Pelo tanto, baixo condições de regularidade padrão:

$$D_n(\theta) \xrightarrow{\mathrm{d}} \chi^2_{\mathrm{k}}, \quad n \to \infty \quad \textbf{(55)}$$

Em consequência, uma região de confiança de nível aproximado $(1-\alpha)$ está dada por

$$C_\alpha = \{\theta: D(\theta) \leq c_\alpha\} \quad \textbf{(56)}$$

Onde c_α é o quantil $(1-\alpha)$ da distribuição $\chi^2{}_d$**.** Escrevendo (56) em termos da verossimilhança relativa tem-se

$$C_\alpha = \left\{\theta: R(\theta) \geq \exp\left(-\tfrac{1}{2}c_\alpha\right)\right\} \quad \textbf{(57)}$$

Logo com as condições do resultado anterior temos que uma região de verossimilhança com nível $-c_\alpha/2$ tem uma probabilidade de cobertura aproximada de $(1-\alpha)$.

2.2.5 Inferência Para os Níveis de Retorno

A inferência para os níveis de retorno é algo transcendental na análise dos valores extremos, já que nas análises univariada pode-se dizer que é a estimação mais importante e de maior aplicação.

Posteriormente de obter as estimativas dos parâmetros da GEV por qualquer um dos métodos mencionados em 2.2, pelo princípio de invariância dos métodos, podem-se substituir por seus respectivos parâmetros (BEIRLANT et al., 2004): portanto para x_p tem-se:

$$\hat{x}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1 - y_p{}^{-\hat{\xi}}\right), & para\ \xi \neq 0, \\ \hat{\mu} - \hat{\sigma}\log y_p, & para\ \xi = 0, \end{cases} \quad \textbf{(58)}$$

É importante destacar que períodos de retorno longos são associados a valores de p pequenos, os quais sempre são os de maior beneficio ou relevância.

Para obter os intervalos de confiança para os valores estimados de x_p, deve-se reparametrizar o modelo GEV de tal forma que x_p seja um dos parâmetros do modelo. A reparametrização ficaria:

$$\mu = x_p + \frac{\sigma}{\xi}\,(1 - (-\log(1-p))^{-\xi} \quad \textbf{(59)}$$

Substituindo esta expressão por μ na log-verossimilhança logra-se o perfil da log-verossimilhança para x_p, assim se pode obter os intervalos de verossimilhança-confiança, metodologia que foi explicada anteriormente.

Outra forma de obter intervalos de confiança aproximados é utilizando o Método Delta, no qual

$$Var(\hat{x}_p) \approx \nabla x_p^t\, \mathbf{V}\, \nabla x_p \quad \textbf{(60)}$$

onde **V** é a matriz de covariâncias de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ e

$$\nabla x_p^t = \left(\frac{\partial x_p}{\partial \xi}, \frac{\partial x_p}{\partial \mu}, \frac{\partial x_p}{\partial \sigma}\right)$$

$$= \left(\frac{\sigma}{\xi^2}\left(1 - y_p{}^{-\xi}\right) - \frac{\sigma}{\xi}\, y_p{}^{-\xi}\log y_p,\ 1, -\frac{1}{\xi}\left(1 - y_p{}^{-\xi}\right)\right) \quad \textbf{(61)}$$

Avaliando em $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Após utilizar a propriedade de normalidade assintótica do estimador logra-se estimar intervalos de confiança aproximados. Segundo Coles (2001) se $\xi < 0$ é possível realizar inferência sobre o extremo direito da distribuição, que de fato é o valor de x_p quando $p = 0$**.** O estimador de máxima verossimilhança é

$$x_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}} \quad \textbf{(62)}$$

e (62) é certa com

$$\nabla x_0^t = (\sigma\xi^{-2}, 1, -\xi^{-1}) \quad \textbf{(63)}$$

A estimação para a cauda da distribuição, para x no domínio apropriado é,

$$\bar{G}_{\widehat{\theta}}(x) = 1 - \exp\left\{-\left(1 + \hat{\xi}\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^{-1/\hat{\xi}}\right\} \quad \textbf{(64)}$$

Onde $\widehat{\theta} = \hat{\mu}, \hat{\sigma}, \hat{\xi}$ é estimado por MML , PWM, MB ou GMML.

3 REFERÊNCIAS BIBLIOGRÁFICAS

AYALA G., MONTES F. **Teoría de la Probabilidad Departament d'Estadística i Investigació Operativa Universitat de Valéncia**, 109p, 2010.

BEIRLANT J.; GOEGEBEUR Y.; SEGERS J.; TEUGELS J. **Statistics of extremes: theory and applications**. London - Chichester : Wiley, 522p, 2004.

BOVIER, A. **Extreme values of random processes** -Lecture Notes-. Bonn: Institut für Angewandte Mathematik. 97p, 2010.

CASTILLO, E.; HADI, A. S.; BALAKRISHNAN, N.; SARABIA, J. M. **Extreme Value and Related Models with Applications in Engineering and Science.** New Jersey: Wiley, 353p, 2005.

COLES, S. **An introduction to statistical modeling of extreme values.** London Berlin Heidelberg: Springer, 205p, 2001.

COLES, S.; POWELL, E. A. Bayesian methods in extreme value modelling: a review and new developments. **Int. Statist. Rev**., 64, p.119–136, 1996.

EMBRECHTS P.; KLÜPPELBERG, C.; MIKOSCH, T.; **Modelling Extremal Events for Insurance and Finance**. Berlin: Springer, 644p, 1997.

DE HAAN, L.; FERREIRA, A. **Extreme Value Theory: An Introduction.** New York: Springer, 417p, 2006.

FISHER, R. A.; TIPPETT, L.H.C. On the Estimation of the Frequency Distributions of the Largest or Smallest Member of a Sample. **Proceeding of the Cambridge Philosophical Society**, 24, p. 180-190, 1928.

GNEDENKO, B. V. Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. **Annals of Mathematics**, 44, p. 423-453, 1943.

GUMBEL, E. J. **Statistics of Extremes**. New York: Dover Publ, 371p, 2004.

KETCHEN, D.J.; KETCHEN, D. J. JR.; BERGH, D. D. Research Methodology in Strategy and Management, **Emerald,** Group Publishing Limited, Volume 3, 2006.

KOTTEGODA, N.; ROSSO, R. **Applied Statistics For Civil and Environmental Engineers.** Second Edition. Oxford: Blackwell Publishing, 705p, 2008.

KOTZ S.; NADARAJAH S. **Extreme value distributions: theory and applications**., London: Imperial College Press, 185p, 2000.

ORTEGA S. J. **Introducción a la Teoría de Valores Extremos** - Notas de clase- . Guanajuato: CIMAT, 130p, 2010.

REISS, R-D.; THOMAS, M; **Statistical Analysis of Extreme Values With Applications to Insurance, Finance, Hidrology and Other Fields**. Third Edition. Basel: Birkhauser Verlag, 508p, 2007.

RESENDE M.D.V.; BARBOSA M. **Melhoramento Genético de Plantas de Propagação Assexuada**, Colombo: Embrapa Florestas, 121p, 2005.

SPROTT, D.A. **Statistical Inference in Sciense**. New York: Springer, 229p, 2000.

SMITH, R. L. Maximum likelihood estimation in a class of non-regular cases. ***Biometrika*** 72, p. 67-90, 1985.

SMITH, R. L. **Statistics of Extremes, With Applications in Environment, Insurance and Finance**. University of North Carolina, 62p, 2003.

STEPHENSON, A.; RIBATET, M. **A User's Guide to the evdbayes Package** (Version 1.1). 35p, 2006.

DISTRIBUIÇÕES DE PROBABILIDADES DO VALOR EXTREMO E TAMANHO AMOSTRAL PARA O MELHORAMENTO GENÉTICO DO QUANTIL MÁXIMO EM PLANTAS

RESUMO

Dentre os objetivos dos programas de melhoramento genético de plantas de propagação assexuada (como a cana-de-açúcar e o eucalipto) e autógamas encontra-se o de selecionar indivíduos extremos ou segregantes transgressivos. Assim, é conveniente encontrar progênies com distribuições de caudas longas ou mesmo assimétricas, já que elas têm uma maior tendência de gerar indivíduos excepcionais. Diante do exposto, Este trabalho teve como objetivo propor e avaliar uma metodologia estatística para o melhoramento do máximo ou valor extremo das distribuições, e não necessariamente das médias das distribuições. A metodologia consistiu em usar dados simulados e reais, típicos das variáveis consideradas no melhoramento genético (por exemplo, distribuição normal com ampla variabilidade e presença de valores extremos). A partir dessa base de dados, distribuições de valores extremos generalizadas são ajustadas aos máximos de cada família, visando verificar qual a distribuição mais adequada (Gumbel, Fréchet, ou Weibull). Os resultados revelaram que a distribuição Weibull se ajusta melhor à bases de dados com 100 ou mais famílias e mais de 20 individuos por família e a distribuição Gumbel se ajusta melhor à bases de dados menores. Simulações estocásticas e reamostragens de dados experimentais indicaram consistentemente que a avaliação de 200 famílias em cada ciclo seletivo maximiza a eficiência do melhoramento visando a seleção de indivíduos extremos. Uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, perfazendo um total de 20000 indivíduos. Os modelos Gumbel e Weibull mostraram-se adequados para analisar as variáveis massa média de colmos (MMC em kg) e teor de Brix (B em %), sendo que a Gumbel mostrou-se adequada apenas nos casos de números de famílias muito pequenos. Assim, recomenda-se a Weibull para inferências práticas. A metodologia é adequada também para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos.

Palavras Chave: Valores extremos, máximo de famílias, melhoramento genético

ABSTRACT

Among the objectives of programs of genetic improvement of asexual propagation of plants (such as sugarcane and eucalyptus) and autogamous is to select extreme or segregating individuals transgressive. It is therefore appropriate to find progenies distributions of long or asymmetrical tails, as they are more likely to generate exceptional individuals. Given the above, this study aimed to propose and evaluate a statistical methodology to improve the maximum or extreme value distributions, and not necessarily the means of distribution. The methodology consisted of using simulated and real data, typical of the variables considered in genetic improvement (eg, normal distribution with wide variability and the presence of extreme values). From this database, generalized extreme value distributions are adjusted to the maximum of each family, in order to ascertain the most appropriate distribution (Gumbel, Fréchet, or Weibull). The results showed that the Weibull distribution best fits the databases with 100 or more families and more than 20 individuals per family and the Gumbel distribution fits better at smaller databases. Stochastic simulations and experimental data resampling consistently indicated that the evaluation of 200 families in each selection cycle to maximize efficiency improvement in order to select extreme individuals. A good practical option would be the evaluation of 200 families with 100 individuals, a total of 20,000 individuals. The Gumbel and Weibull models have shown to be adequate to analyze the average mass variable stem (MMC kg) and Brix content (B%), and the Gumbel was adequate only in the case of very small families numbers. Thus, it is recommended to Weibull for practical inferences. The methodology is also suitable to classify the families or the progenies ability to generate superior or exceptional individuals and inform the sample sizes to be practiced in every family to capture these individuals.

Keywords: extreme values, maximum of families, genetic improvement

1. INTRODUÇÃO

Segundo Conselho de Informações sobre Biotecnologia (CIB) o Brasil é o maior exportador de açúcar, respondendo por 45% de todo o produto comercializado no mundo, No ano 2014 Brasil foi o segundo maior produtor mundial de etanol, detrás dos Estados Unidos. A área plantada, a produtividade, e a produção de cana-de-açúcar na temporada de 2014/2015 foram de mais de nove milhões de hectares cultivados, 70 t/hectare, e 635 milhões de toneladas respetivamente (CONAB). Demostrando o papel essencial que tem esta atividade agroindustrial na economia nacional.

Na atualidade os subprodutos derivados da cana-de-açúcar como a açúcar, etanol e energia elétrica, são relevantes para a agricultura e a indústria mundial, no entanto, a produtividade média do Brasil na temporada 2014/2015 foi inferior à produtividade média mundial, que segundo Waclawovsky et al. (2010) é de 80 t/hectare, e ficou muito em baixo em relação à produtividade máxima teórica que é de 380 t/hectare. Portanto, é preciso buscar melhorar a cultura da planta, de tal forma que se alcance estar perto de seu potencial genético, ou seja, fazer que os cultivares cada dia se tornem mais produtivos.

Segundo o CIB o Brasil conseguiu aumentar a produtividade da cana-de-açúcar em mais de 50% nos últimos 30 anos, em parte pelos investimentos em pesquisa e na melhora da cultura. Embora, nas atuais circunstancias precisam-se obter iguais ou melhores resultados aos descritos anteriormente, em um período de tempo muito menor. Consequentemente indivíduos excepcionais devem ser selecionados pelos programas de melhoramento genético; no entanto, a maioria dos programas de melhoramento genético utilizam procedimentos de predição que são funções da média aritmética das observações, esta seleção é apropriada por proporcionar a distinção através de a produtividade média, mas sem levar em consideração a homogeneidade das famílias (RESENDE e BARBOSA, 2005).

Pela abordagem das metodologias de BLUP sob média aritmética e sob média harmônica, não é dada muita importância aos valores extremos de uma família. Porém, famílias com grande tendência de gerar indivíduos excepcionais ou superiores são muito desejáveis, para a clonagem de maneira

geral, ou seja, tecnicamente para os programas de melhoramento genético de cana-de-açúcar são mais convenientes famílias com distribuição assimétrica (RESENDE e BARBOSA, 2005).

No melhoramento genético de plantas, uma metodologia estatística que utilize os valores máximos (indivíduos excepcionais e seus valores genotípicos preditos), para a análise de famílias e a escolha dos genitores com base no desempenho da progênie, tem sido pouco ou nada explorada. Portanto, este trabalho teve como objetivo propor e avaliar uma metodologia estatística para o melhoramento do máximo ou valor extremo das distribuições, e não necessariamente das médias das distribuições. Essa abordagem baseia-se nos quantis superiores da GEV (Distribuição de Valores Extremos Generalizada) dos BLUP´s genotípicos individuais entre e dentro de famílias, como forma de prever o aumento da ocorrência de valores extremos em função do aumento do tamanho da família (seleção de indivíduos extremos dentro de família) e também do número de famílias utilizado para representar uma população (seleção de indivíduos extremos em toda a população).

As ideias principais da teoria de valores extremos podem ser descritas da seguinte maneira: dado uma amostra de variáveis aleatórias independentes e identicamente distribuídas, pretende-se analisar o comportamento de qualquer uma das estatísticas de ordem, como o mínimo, o máximo, ou o limiar de um nível crítico, onde as estatísticas de ordem devem ter funções de distribuição ***max-estável***, já que por definição as distribuições do máximo e o mínimo são degeneradas. Portanto, deve-se usar um método que permita aproveitar o argumento assintótico, como o teorema dos Tipos Extremos, para aproximar a distribuição da estatística de ordem por uma dentre as três famílias distribuições interessantes a este propósito, das quais a Gumbel (Tipo I), Fréchet (Tipo II), ou Weibull (Tipo III), que pertencem à classe DEV (Distribuições de Valores Extremos) e que são casos especiais da GEV.

2. MATERIAIS E MÉTODOS

Para estudar o tamanho amostral visando à maximização da eficiência do melhoramento do máximo de uma distribuição foram consideradas duas situações: simulação de dados e um conjunto de dados reais de cana-de-açúcar.

Os processos de simulação e com dados reais foram realizados com o software livre R (R Development Core Team, 2015), os pacotes utilizados foram *extremes* (GILLELAND e KATZ, 2012) o qual tem um conjunto de funções para a realização de análises dos valores extremos de um processo de interesse, utilizando o método do bloco máximo ou excessos ao longo de um limiar elevado; *in2extremes* (GILLELAND e KATZ, 2013) proporciona um conjunto de janelas (interfaces gráficas) que resumem algumas das principais funções do pacote *extremes*. O pacote *evd* (STEPHENSON, 2012) estende funções de simulação, distribuição, quantis e densidades para as distribuições de valores extremos paramétricos uni e multivariadas, também fornece funções de ajuste que calculam estimativa de máxima verossimilhança para os métodos de bloco máximo e limiar uni e bivariadas; para corroborar os resultados obtidos com os métodos bayesianos do pacote *extremes,* utilizou-se o pacote *evdbayes* (STEPHENSON e RIBATET, 2014) que fornece funções para a análise bayesia na de modelos de valores extremos, usando métodos MCMC.

O pacote *truncnorm* (TRAUTMANN et al., 2014) proporciona funções para encontrar as diferentes caraterísticas (densidade, função de distribuição, função quantil, geração aleatória e função do valor esperado) da distribuição normal truncada.

As rotinas computacionais implementadas no software R estão apresentadas no anexo 1.

2.1 Dados simulados

Foram simulados nove cenários com diferentes médias, desvios padrões e valores máximos, conforme quadro a seguir.

Tabela 1 – Cenários de famílias e indivíduos simulados

Núm. de Famílias \ Núm. de Indivíduos	**20**	**100**	**200**
20	C1	C2	C3
100	C4	C5	C6
200	C7	C8	C9

Cada cenário foi simulado cem vezes e os indivíduos nas famílias correspondem a uma distribuição normal truncada para valores positivos com tamanhos amostrais de 20, 100 e 200 indivíduos. Para os cenários C1, C2, e C3 as médias utilizadas foram 5, 10, 15 e 20 de toneladas de produtividade de açúcar por hectare (TPH), os desvios padrões usados para cada uma das médias anteriores foram 1, 2, 3, 4, e 5 TPH. Nos cenários C4, C5, e C6 as médias iniciam em 1,6 e incrementam-se em 1 (uma) unidade até chegar a 20,6 TPH, os desvios padrões são os mesmos dos cenários anteriores. Em referência aos cenários C7, C8, e C9 os valores das médias de TPH iniciaram com 0,3 e foram aumentando 0,3 até 12,0, também se utilizaram os desvios padrões 1, 2, 3, 4, e 5 TPH para cada uma das médias simuladas anteriormente.

Um indivíduo simulado representa a produção de uma parcela experimental. Após simular os cenários C1 até C9, foram encontrados os máximos, médias e desvios padrões das amostras dos máximos. Para encontrar o modelo adequado para analisar as amostras dos máximos, utilizou-se a metodologia GEV e para os intervalos de confiança utilizou-se o método do Perfil de Verossimilhança.

Realizaram-se os testes de Anderson Darling e Kolmogorov, para verificar a bondade dos ajuste modelos encontrados (Gumbel, Weibul ou Frechét); depois de conhecer a distribuição adequada para as análises das amostras, obtiveram-se os níveis de retornos (valor do máximo) associados aos períodos de retorno (números de famílias a serem avaliadas para a obtenção do máximo) 20, 50, 100, 200, 500 e 1000. Deve ser ressaltado que os tamanhos totais das “novas populações” variaram de 400 a 200000 indivíduos. Dessa forma, as eficiências em nível individual, de família e populacional foram determinadas.

O cálculo das eficiências para o melhoramento em cana de açúcar foi feito com relação ao aumento de indivíduos, famílias e a população, desse modo foram utilizadas as seguintes expressões:

$$E_{ind} = \frac{Est_{F.n}}{Est_{F.20}} \; ; \; E_{fam} = \frac{Est_{F.n}}{Est_{20.n}} \; ; \; E_{pop} = \frac{Est_{F.n}}{Est_{20.20}} \qquad \textbf{(65)}$$

Onde:

$E_{ind} = \; Eficiência\ com\ relação\ ao\ aumento\ de\ indivíduos$

$Est_{F.n} = \; Estimativa\ do\ valor\ máximo\ com\ F\ novas\ famílias\ e\ n\ indivíduos$

$F = 20, 50, 100, 200, 500\ e\ 1000;\ n = 20, 100\ e\ 200$

$Est_{F.20} = Estimativa\ do\ valor\ máximo\ com\ F\ novas\ famílias\ e\ 20\ indivíduos$

$F = 20, 50, 100, 200, 500\ e\ 1000$

$E_{fam} = \; Eficiência\ com\ relação\ ao\ aumento\ de\ famílias$

$Est_{20.n} = \; Estimativa\ do\ valor\ máximo\ com\ 20\ novas\ famílias\ e\ n\ indivíduos$

$n = 20, 100\ e\ 200$

$E_{pop} = \; Eficiência\ com\ relação\ ao\ aumento\ da\ população$

$Est_{20.20} = Estimativa\ do\ valor\ máximo\ com\ 20\ novas\ famílias\ e\ 20\ indivíduos$

2.2 Reamostragem em dados reais de cana-de-açúcar

Utilizando dados reais de cana-de-açúcar obtidos no programa de melhoramento conduzido pela UFV em Oratórios – MG, as variáveis analisadas foram massa média de colmos (MMC em kg) e teor de Brix (B em %). Foram tomados subgrupos (por meio de reamostragens) criando-se quatro subpopulações com diferentes médias, desvios padrões e valores máximos. As subpopulações foram compostas por 20, 30, 43 e 63 famílias, especificamente, das sessenta e três famílias foram usadas vinte famílias (COOK,1985) para ajustar o modelo GEV e obter as estimativas dos níveis de retorno associados a vinte, trinta e quarenta e três “novas famílias”. Das quarenta e três “novas famílias” restantes foram escolhidas ao acaso vinte e trinta famílias, para validação das estimativas obtidas anteriormente.

Deve ser ressaltado que o processo descrito anteriormente (validação cruzada), tem como finalidade analisar a idoneidade da extrapolação dos

modelos GEV na prática biométrica, já que a teoria dos valores extremos é muito utilizada em diferentes áreas do conhecimento.

O modelo estatístico utilizado para a predição de indivíduos superiores foi a Distribuição de Valores Extremos Generalizada (GEV) proposto por Von Mises (1936 -1954) e Jenkinson (1955):

$$G(x\backslash\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, x \in \mathbb{R} \qquad \textbf{(66)}$$

Definido no conjunto $\left\{x: 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, os parâmetros μ, σ e ξ são chamados de localização, escala e forma respectivamente, e os valores que podem tomar são $-\infty \leq \mu \leq \infty$, $\sigma > 0$, e $-\infty \leq \xi \leq \infty$. Se $\xi \to 0$ se diz que a distribuição limite é Gumbel (tipo I), se $\xi > 0$ tem-se a distribuição Fréchet (tipo II), e se $\xi < 0$ a distribuição obtinha é Weibull (tipo III) (COLES, 2001). Ás famílias das distribuições anteriores são conhecidas como Distribuições de Valores Extremos (DEV), as quais estão definidas pelo teorema de Fisher – Tippet e Gnedenko (CASTILLO et al., 2005; BOVIER, 2004; COLES, 2001) como:

Distribuição Gumbel:

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{x-b}{a}\right)\right]\right\}, -\infty < x < \infty; \qquad \textbf{(67)}$$

Distribuição Fréchet:

$$G(x) = \begin{cases} 0, & se\ x \leq b, \\ \exp\left\{-\left(\frac{x-b}{a}\right)^{-\alpha}\right\}, & se\ x > b, \end{cases} \qquad \textbf{(68)}$$

Distribuição Weibull:

$$G(x) = \begin{cases} exp\left\{-\left[-\left(\frac{x-b}{a}\right)^{-\alpha}\right]\right\}, & se\ x \leq b, \\ 1, & se\ x > b, \end{cases} \qquad \textbf{(69)}$$

Onde os parâmetros $a > 0$ e $b \in R$, e no caso das famílias Fréchet e Weibull $\alpha > 0$.

A metodologia para identificar uma observação ou valor extremo foi o Método do Bloco Máximo, que em sua forma básica consiste em dividir a amostra de tamanho ***m***, em ***b*** blocos (famílias) de similar tamanho ***n*** (***n*** suficientemente grande), e logo obter os valores máximos de cada família (***b)***, assim com a nova amostra $M_{n,1}, M_{n,2}, \ldots, M_{n,b}$, foram estimados μ, σ, e ξ mediante o ajuste da GEV.

Para estimar os parâmetros das amostras dos máximos utilizaram-se os métodos de Máxima Verossimilhança (MML) e Bayesianos; quando utilizou-se o MML foi necessário analisar o intervalo de confiança de ξ e o teste de razão de verossimilhança, para determinar qual das famílias de distribuições era a adequada para estudar a amostra ***m*** escolhida (GUMBEL, 2004) das respectivas variáveis MMC e B.

O MML é o mais usado por sua generalidade (propriedades assintóticas dos estimadores) e flexibilidade (CARIES, 2011; COLES, 2001), no entanto, na maior parte dos programas de melhoramento de plantas perenes como cana de açúcar, trabalham com conjuntos de dados pequenos, e os métodos Bayesianos proporcionam resultados precisos quando tem-se pequenas amostras, já que proporcionam uma distribuição a posteriori fidedigna para obter as inferências pertinentes, seja para grandes ou pequenas amostras (RESENDE, 2000).

Com a metodologia Bayesiana para a analises dos parâmetros das variáveis MMC e B foram usadas como a priori a distribuição normal trivariada $(\mu = 2,\ \sigma = 0{,}3,\ \xi = 0{,}04)$ e $(\mu = 15,\ \sigma = 2,\ \xi = 0{,}05)$ respetivamente, e utilizaram-se métodos de simulação de Monte Carlo via Cadeias de Markov (MCMC); para o MML também precisou-se de métodos numéricos para obter a estimativa dos parâmetros porque analiticamente não é possível obtê-los, entretanto, deve-se ter cautela quando se trabalha com o MML, já que as condições de regularidade que se precisam para que as propriedades assintóticas, associadas com os estimadores de máxima verossimilhança em algumas circunstâncias não são satisfeitas (COLES, 2001).

A função de densidade preditiva utilizada no método bayesiano para a obtenção dos parâmetros e níveis de retorno foi a seguinte:

$$Pr\{\mathrm{Z} \leq z|x_1, \ldots\ldots, x_n\} = \int_{\Theta} Pr\{\mathrm{Z} \leq z|\theta\} f(\theta|x)\, d\theta \qquad (\mathbf{70})$$

Onde $Pr\{\mathrm{Z} \leq z|\theta\}$ é a GEV avaliada em z, $\theta = \mu, \sigma$ e ξ e $f(\theta|x)$ é a distribuição a posteriori (CASTILLO et al., 2005; DE HAAN e FERREIRA, 2006; STEPHENSON e RIBATET 2014)

Os intervalos de confiança para as estimativas dos parâmetros foram obtidos pelo método do Perfil de Verossimilhança, e na metodologia Bayesiana

utilizou-se o Bootstrap Paramétrico, que emprega o procedimento do cálculo do porcentual a partir da simulação da amostra (percentis do MCMC). Para fazer qualquer tipo de predição (extrapolação), foi necessário constatar que o ajuste realizado do modelo GEV era válido, para isto se utilizaram ferramentas gráficas como: quantil-quantil (QQ), probabilidade (PP), densidade dos dados Vs. densidade do modelo, ademais se realizaram os testes de aderência de Anderson Darling e Kolmogorov.

Após de conhecer a distribuição mais adequada para analisar as variáveis em estudo e conhecer as estimativas dos parâmetros μ, σ, e ξ , se obtiveram os níveis de retornos x_p (quantis) associados aos quatro (20,30,43,63) períodos de retorno $(\frac{1}{p})$ desejados, através da seguinte expressão:

$$\hat{x}_p = \begin{cases} \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\left(1 - y_p^{-\hat{\xi}}\right), & para\ \xi \neq 0, \\ \hat{\mu} - \hat{\sigma}\log y_p, & para\ \xi = 0, \end{cases} \qquad \textbf{(71)}$$

Onde $y_p = -\log\,(1-p)$.

Construíram-se os respectivos gráficos dos níveis de retorno vs. períodos de retorno, já que são muito úteis porque facilitam distinguir os resultados da extrapolação de níveis de retorno em períodos de retorno longos; e também permite apresentar e validar a distribuição pertinente para a análise da amostra dos máximos, quando utiliza-se o método de Máxima Verossimilhança para estimar os parâmetros. Foram obtidos os intervalos de confiança dos níveis de retornos com três metodologias: Perfil de Verossimilhança (Ve. P), Método Delta (Aproximação Normal) e Bootstrap Paramétrico (percentis do MCMC).

- **Perfil de Verossimilhança**

O método do Perfil de Verossimilhança consiste em maximizar a função de verossimilhança com base em parâmetros definidos com anterioridade (SPROTT 2000; CASTILLO et al., 2005), ou seja:

$$L_p\left(\theta^{(1)}\right) = \max_{\theta^{(2)}|\theta^{(1)}} \mathrm{L}\left(\theta^{(1)}, \theta^{(2)}\right) \qquad \textbf{(72)}$$

A avaliação numérica do perfil da verossimilhança para qualquer um dos parâmetros μ, σ, ξ ou os níveis de retorno $\hat{x}_p$ (para $\hat{x}_p$ precisa-se duma nova

parametrização do modelo GEV) é obtida de uma forma relativamente prática, por exemplo se quisermos obter o perfil da verossimilhança de ξ, fixa-se $\xi = \xi_0$ e maximizamos a log-verossimilhança com relação aos parâmetros restantes. Logo se reproduz isso para um determinado número de valores de $\xi = \xi_0$, os respectivos valores maximizados da log-verossimilhança compõem o perfil da log-verossimilhança para ξ; com o procedimento geral para construção de intervalos de confiança, podem-se obter os respectivos intervalos aproximados para cada parâmetro (COLES, 2001; CASTILLO et al., 2005).

- **Método Delta**

O método Delta admite a normalidade aproximada (convergência em distribuição) de $\hat{x}_p$ para ser utilizada na obtenção dos intervalos de confiança de x_p, pelo tanto o método permite aproximar a $Var(\hat{x}_p)$ por meio da seguinte expressão:

$$Var(\hat{x}_p) \approx \nabla x_p^t \, \mathbf{V} \, \nabla x_p \qquad \textbf{(73)}$$

onde **V = Σ** é a matriz de covariâncias de $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$ e

$$\nabla x_p^t = \left(\frac{\partial x_p}{\partial \xi}, \frac{\partial x_p}{\partial \mu}, \frac{\partial x_p}{\partial \sigma}\right)$$

$$= \left(\frac{\sigma}{\xi^2}\left(1 - y_p^{-\xi}\right) - \frac{\sigma}{\xi} y_p^{-\xi} \log y_p,\ 1, -\frac{1}{\xi}\left(1 - y_p^{-\xi}\right)\right) \qquad \textbf{(74)}$$

Avaliadas em $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

- **Bootstrap Paramétrico**

O método de bootstrap paramétrico pressupõe que existe uma distribuição de probabilidade que originou $x_1, x_2 \ldots\ldots x_n$ (amostra aleatória de tamanho n, chamada amostra original), portanto utiliza o modelo de probabilidade pressuposto e as estimativas dos parâmetros obtidas com a amostra original, para gerar um grande número de amostras independentes $x_1^*, x_2^* \ldots\ldots x_B^*$, chamadas amostras bootstrap. θ representa o parâmetro de interesse, e uma

réplica bootstrap $\theta_b^*,\ b = 1,2,\ldots.B$, corresponde ao valor do estimador de máxima verossimilhança $\hat{\theta}$ avaliado em cada uma das B amostras bootstrap.

Após de obter as B amostras bootstrap, também é admissível construir uma distribuição bootstrap para o estimador de máxima verossimilhança $\hat{\theta}$, pelo tanto a distribuição obtida pode-se usar para realizar inferências sobre o parâmetro em estudo. Os intervalos de confiança bootstrap percentil são obtidos pelos respectivos percentis $\frac{\alpha}{2} - ésimo$ e $(1 - \frac{\alpha}{2}) - ésimo$ da distribuição de $\hat{\theta}^*$, representada por $\hat{F}$ (EFRON e TIBSHIRANI, 1983; ATHREYA e FUKUCHI, 1993)

As metodologias apresentadas foram comparadas por meio dos intervalos de confiança e da Eficiência Média ($E_M; indicador\ de\ viés$), em que 100% indica ausência de viés:

$$E_M = \frac{\bar{y}_n}{V.\ Est.} \quad \textbf{(75)}$$

Onde $\bar{y}_n$ é a média dos três ou cinco melhores indivíduos, e $V.\ Est.$ é a estimativa do nível de retorno $\hat{x}_p$.

3. RESULTADOS E DISCUSSÕES

3.1 Dados simulados

O modelo mais frequente para ajustar amostras de 20 famílias independentemente do número de indivíduos é a Gumbel, entretanto, o modelo Weibull também ajusta-se razoavelmente para este tipo de amostras. É importante ressaltar que os níveis de retornos ($\hat{x}_p$) para a variável TPH, obtidos pelo modelo Gumbel são maiores que os conseguidos pelo modelo Weibull, situação que reflete-se ao analisar os respetivos intervalos de confiança (Anexo 4).

As estimativas obtidas de $\hat{x}_p$ para amostras de 20 famílias conseguidas através dos modelos Weibull, sempre ficam muito perto do limite inferior do intervalo de confiança (Anexo 4), circunstância que também é frequente para os modelos Weibull ajustados para amostras de 100 e 200 famílias (Anexo 5). Deve-se destacar que nos três cenários em média o 60% dos modelos Weibull

ajustados com amostras de 20 famílias, o valor estimado do parâmetro ξ é maior que – 0,5 (analises individual dos resultados das simulações), pelo tanto as inferências realizadas são verossímeis, contudo, os modelos que tiveram uma estimativa de ξ diferente á apresentada anteriormente, obtiveram estimativas pontuais e intervalos de confiança de $\hat{x}_p$ similares.

Em geral para as amostras de 100 e 200 famílias independentemente do número de indivíduos (20, 100 e 200) por família, o modelo que melhor ajusta-se aos máximos das amostras estudadas é a Weibull (Anexo 5). Para todos os modelos Weibull a estimativa do parâmetro ξ é maior que – 0,5 (analises individual de cada uma das cem simulações feitas), portanto pode-se ter uma alta confiança nas estimativas pontuais e intervalos de confiança de $\hat{x}_p$ (SMITH, 1985) da variável TPH.

Os resultados preditivos dos modelos Gumbel e Weibull apresentam estabilidade e consistência nas estimativas dos períodos de retorno e seus respectivos intervalos de confiança (Anexo 4 e 5), por tanto na Tabela 2 só encontram-se os prognósticos dos valores máximos de uma simulação dos cenários C7, C8 y C9.

Tabela 2 – Valores do ***máximo*** (obtidos pela distribuição Weibull) para populações compostas por ***nfam*** famílias representadas por ***nind*** indivíduos e eficiências do aumento de nfam (***efic fam***), nind (***efic ind***) e população total (***efic pop***)

nfam	nind	npop	máximo	efic ind	efic fam	efic pop	Média de efic fam*	Média de efic ind**
20	20	400	20.03	1.00	1.00	1.00		1.00
20	100^{+}	2000	22.43	1.12	1.00	1.12		1.10^{+}
20	200	4000	23.37	1.17	1.00	1.17	1.00	1.12
50	20	1000	22.17	1.00	1.11	1.11		
50	100	5000	24.62	1.11	1.10	1.23		
50	200	10000	25.38	1.14	1.09	1.27	1.10	
100	20	2000	23.53	1.00	1.17	1.17		
100	100	10000	25.99	1.10	1.16	1.30		
100	200	20000	26.6	1.13	1.14	1.33	1.16	
200	20	4000	24.73	1.00	1.23	1.23		
200^{++}	100^{++}	20000^{++}	27.18	1.10	1.21	1.36^{++}		
200	200	40000	27.61	1.12	1.18	1.38	1.21	
500	20	10000	26.08	1.00	1.30	1.30		

500	100	50000^{+++}	28.49	1.09	1.27^{+++}	1.42^{+++}	
500^{++++}	200	100000	28.69	1.10	1.23	1.43	1.27^{++++}
1000	20	20000	26.96	1.00	1.35	1.35	
1000	100	100000	29.32	1.09	1.31	1.46	
1000	200	200000	29.34	1.09	1.26	1.46	1.31

$^{+}$ Ótimo para número de indivíduos; $^{++}$ Ótimo para situação prática; $^{+++}$ Ótimo para tamanho de população; $^{++++}$ Ótimo para número de famílias; * Média da eficiência através dos diferentes tamanhos de família; ** Média da eficiência através dos diferentes números de famílias; Média geral do caráter nos vários cenários igual a 6,4 com desvio padrão de 5,5

Os valores extremos ou máximos variaram de 20,3 com 20 famílias e 20 indivíduos por família a 29,34 para 1000 famílias e 200 indivíduos por família. Sendo a média geral igual a 6,4 a clonagem de indivíduos com esses valores máximos proporcionaria ganhos genéticos de elevada ordem. Valores máximos da ordem de 28,4 advém de uma distribuição com média 6,4 e desvio padrão 5,5, truncada em quatro desvios padrão. Valores genotípicos dessa magnitude (acima de 28) ocorrem em experimentos com 500 famílias representadas por 200 indivíduos cada uma (Tabela 2). Nessa situação, a eficiência seletiva é de 1,27 em relação ao uso de apenas 20 famílias, ou seja, ganho genético ou superioridade de 27%. Duplicando-se o número de famílias para 1000, a eficiência se eleva a 1,31, valor esse, não compensatório dado o grande esforço experimental necessário.

Quanto ao tamanho da população total, para obtenção de eficiência de 1,42 torna-se necessária a avaliação de 50000 indivíduos, número esse que pode ser proibitivo em algumas espécies. Uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, perfazendo um total de 20000 indivíduos, número esse corriqueiro no melhoramento de eucalipto, por exemplo. Nesse cenário, a eficiência é de 1,36. Dentre os cenários, a eficiência máxima é de 1,46 e ocorre avaliando-se 1000 famílias com 100 ou 200 indivíduos. Os valores de 1,36 e 1,42 são mais atrativos dado o esforço experimental necessário. O número adequado de família em torno de 200 foi também recomendado em estudo teórico (via avaliação numérica e simulação determinística) reportado por Resende (1995).

O aumento da eficiência seletiva com o aumento do tamanho de família é em torno de 1,10 quando se passa de 20 para 100 indivíduos por família e de 1,12 quando se passa de 100 para 200 indivíduos. Esses números são aproximadamente constantes independentemente do número de famílias avaliadas. Assim a inferência sobre o tamanho adequado de família pode ser

realizado de forma independente ao número de famílias avaliadas e os resultados remetem ao uso de 100 indivíduos por família, não sendo compensatório duplicar esse número, pois o acréscimo seria da ordem de apenas 2%.

O comportamento da eficiência seletiva em função do tamanho da população experimental total é apresentado na Figura 2. A distribuição dos pontos revela um comportamento assintótico da eficiência com grande aproximação à assíntota em 20000 indivíduos.

É importante destacar que nas análises feitas em todos os cenários estudados, nenhuma das amostras simuladas indicaram as famílias de distribuições Fréchet, como as mais adequadas das distribuições para realizar as respetivas análises dos valores extremos, pelo tanto recomenda-se realizar outras pesquisas onde analisem-se os valores extremos de outras variáveis biométricas em diferentes cenários, de tal maneira que possa-se aferir, a aplicabilidade das famílias de distribuições Fréchet na analises dos valores extremos em variáveis utilizadas no melhoramento genético de plantas.

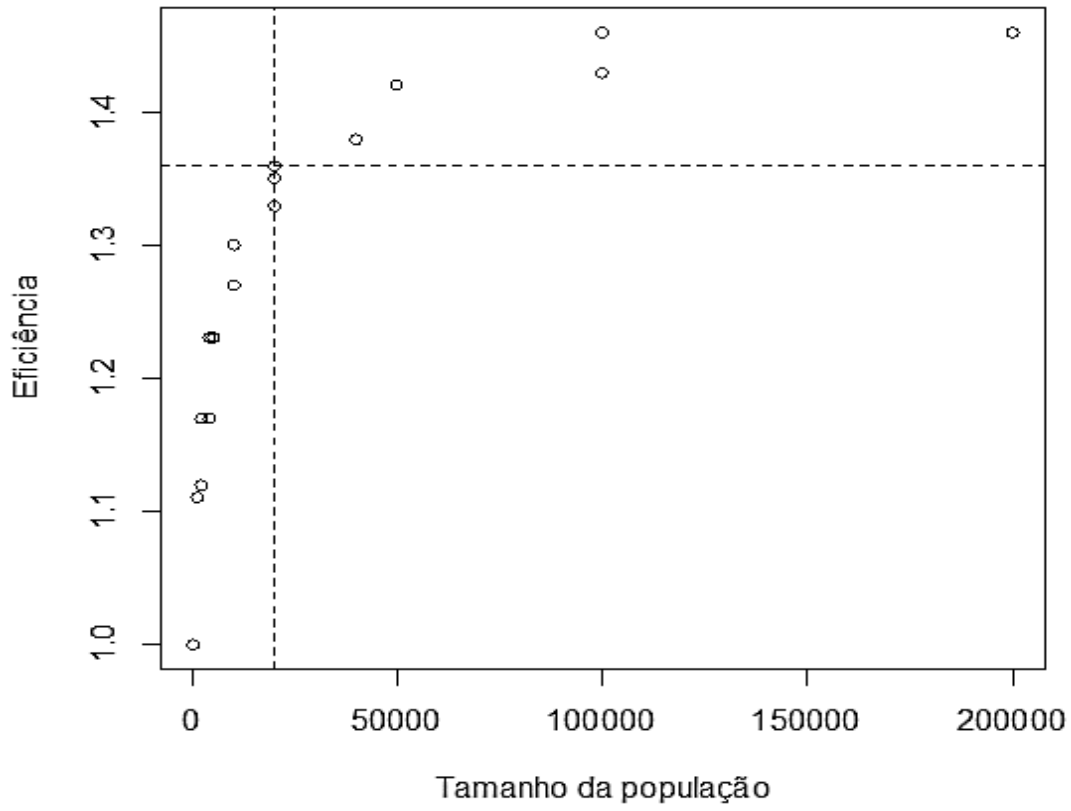


Figura 2. Comportamento da eficiência seletiva em função do tamanho da população experimental total

3.2 Dados Reais

De acordo com o teste da razão de verossimilhança para os modelos ajustados GEV ($\xi = 0$ e $\xi \neq 0$) para a variável MMC o modelo adequado para analisar esta é a Gumbel p-valor = 0,758 (associado à hipótese $\xi = 0$), situação que corrobora-se analisando o respectivo intervalo de confiança de ξ (-0,232 ; 0,290), para a variável B o teste apresenta um p-valor = 0,014 (associado à hipótese $\xi \neq 0$), indicando que a distribuição mais adequada para o analises concernente é diferente da Gumbel, por isso analisou-se o intervalo de confiança correspondente que está entre (-0,636 ; -0,152), determinando que a distribuição mais adequada é a Weibull.

Deve-se ter muita precaução com as estimativas do modelo Weibull porque ξ (-0,593) está entre -1 e -0,5, advertindo que estes não possuem as propriedades assintóticas (consistência, eficiência, invariância e normalidade) dos estimadores (SMITH, 1985), contudo, as estimativas obtidas com os métodos bayesianos e o MML, dos parâmetros μ, σ, ξ e seus respectivos intervalos de confiança dos modelos ajustados para as variáveis MMC e B, apresentam resultados similares (Tabela 3), situação que reflete a flexibilidade da teoria dos valores extremos para ser aplicada no melhoramento genético de plantas, já que até o momento sua maior aplicação se dá em hidrologia, engenharia civil, oceanografia, metalurgia, geoestatística, entre outras importante áreas do conhecimento.

Os diagnósticos dos modelos Gumbel e Weibull ajustados com o MML para as amostras dos máximos de MMC e B, evidenciam que estes ajustam-se razoavelmente ás amostras dos máximos estudadas (Anexos 2 e 3), isto é confirmado pelos testes de Kolmogorov-Smirnov (p-valor = 0,463 e p-valor = 0,659) e Anderson- Darling (p-valor = 0,062 e p-valor = 0,608), pelo tanto é possível analisar os níveis de retorno desejados.

Os modelos Gumbel e Weibull ajustados com os métodos bayesianos, também podem-se considerar admissíveis para realizar as inferências correspondentes (Anexos 2 e 3); é importante advertir que as taxas de aceitação dos parâmetros nos modelos estudados ficaram entre 0,399 e 0,632, intervalo que é razoável para realizar os respectivos analises bayesianos.

Tabela 3 - Estimativas e intervalos de confiança dos parâmetros μ, σ, ξ dos modelos ajustados para variáveis massa média de colmos (MMC) e teor de Brix (B).

Variável	Método	Estimador	Estimativa	Intervalos
MMC	MLE	$\hat{\mu}$	2,175	(2,027 ; 2,332)
		$\hat{\sigma}$	0,316	(0,247 ; 0,455)
		ξ	- 0,042	(-0,232 ; 0,29)$^{+}$
	Bayesiana	$\hat{\mu}$	2,179	(1,979 ; 2,369)
		$\hat{\sigma}$	0,341	(0,242 ; 0,522)
		ξ	- 0,044	(-0,127 ; 0,042)
B	MLE	$\hat{\mu}$	15,612	(15,374 ; 16,607)
		$\hat{\sigma}$	2,112	(1,970 ; 3,612)
		$\hat{\xi}$	- 0,593	(- 0,636 ; - 0,152)*
	Bayesiana	$\hat{\mu}$	15,642	(14,414 ; 16,564)
		$\hat{\sigma}$	2,304	(1,627 ; 3,287)
		$\hat{\xi}$	- 0,587	(- 0,704 ; - 0,477)

Fonte: Elaboração Própria - MMC= massa média de colmos – B=Brix – MLE =Máxima Verossimilhança
$^{+}$Gumbel - *Weibull

Da Tabela 3 pode-se verificar que para 20 novas famílias as estimativas pontuais das toneladas máximas de MMC. ($\hat{x}_p$) obtidas pelo método bayesiano (MB) e o MML, são maiores que o maior indivíduo dessas famílias, portanto seus respectivos intervalos de confiança e HPD também contêm aquele valor.

Com 30 e 43 novas famílias as estimativas pontuais de ambos métodos são menores que o maior indivíduo estudado em essas amostras, não obstante, o intervalo de confiança obtido pela metodologia do perfil de verossimilhança na amostra de 43 famílias abrange o valor do maior indivíduo. Deve ser ressaltado que em ambas amostras os intervalos HPD incluem os valores analisados.

Quando apresentam-se as 63 famílias os intervalos de confiança e HPD obtidos pelos MML e MB respectivamente, estes abrangem o maior indivíduo dessas famílias, entretanto só o MB conseguiu obter uma estimativa pontual satisfatória para o indivíduo em questão.

Em geral para a variável MMC a eficiência média obtida quando utilizou-se a estimativa do MML é elevada, já que para as médias dos três e cinco melhores indivíduos ficou entre 91,5% - 102,9% e 89,1% - 96,2% respectivamente, valores

esses muito próximos do valor esperado que é igual a 1. Por outro lado, quando usou-se o MB a eficiência média apresentou resultados satisfatórios para $\overline{y}_3$, mas não para $\overline{y}_5$ onde que só um valor pode-se considerar admissível (88,0%).

Tabela 4 - Comparação das estimativas pontuais e intervalos de confiança dos níveis de retorno com os valores máximos das amostras, e resultados das acurácias das médias dos três e cinco melhores indivíduos amostrais da variável MMC.

Nov. Famílias	Máx.	$\overline{y}_3$	$\overline{y}_5$	Métodos	Estimativa do máximo para o caráter	Intervalos	Eficiência Média	
							$\overline{y}_3$	$\overline{y}_5$
				Ve. P	**3,114**	**(2,804 ; 3,599)**	91,5%	89,1%
n = 20	**3,04**	2,85	2,77	A. Normal	**3,114**	**(2,732 ; 3,496)**	91,5%	89,1%
				Q. MCMC	**3,349**	**(2,825 ; 4,710)**	85,0%	83,0%
				Ve. P	**3,244**	**(2,900 ; 3,784)**	102,9%	96,2%
n = 30	**3,93**	3,34	3,12	A. Normal	**3,244**	**(2,822 ; 3,667)**	102,9%	96,2%
				Q. MCMC	**3,546**	**(2,916 ; 5,358)**	94,1%	88,0%
				Ve. P	**3,36**	**(3,053 ; 3,952)**	99,3%	92,9%
n = 43	**3,93**	3,34	3,12	A. Normal	**3,36**	**(2,901 ; 3,819)**	99,3%	92,9%
				Q. MCMC	**3,732**	**(2,987 ; 5,978)**	89,4%	83,6%
				Ve. P	**3,482**	**(3,105 ; 4,122)**	100,1%	94,7%
n = 63	**3,93**	3,48	3,30	A. Normal	**3,482**	**(2,984 ; 3,979)**	100,1%	94,7%
				Q. MCMC	**3,943**	**(3,058 ; 6,806)**	88,4%	83,6%

Fonte: Elaboração Própria - Máx.: Valor máximo - Met.: Metodologia - $\overline{y}_3$:Média dos três melhores indivíduos - $\overline{y}_5$:Média dos cinco melhores indivíduos - MLE :Máxima Verossimilhança - Ve. P: Perfil de Verossimilhança - A. Normal: Aproximação Normal – Q.MCMC: Quantis de Monte Carlo via Cadeias de Markov

Com base na Tabela 5 é possível corroborar que as estimativas pontuais obtida pelo MML de B, ficaram abaixo dos maiores indivíduos nas quatro amostras, mas destaca-se que nas amostras de 43 e 63 famílias estas estimativas estão muito perto dos valores analisados. É importante destacar que todos os respectivos intervalos de confiança incluem os valores máximos das amostras; Coles (2001), De Haan e Ferreira (2006) argumentam que os

resultados obtidos pela metodologia da aproximação normal são poucos satisfatórios ao ser comparados com outras metodologias, entretanto, para as variáveis analisadas apresentam um comportamento similar à metodologia do Perfil de Verossimilhança.

As estimativas pontuais e os intervalos HPD conseguidos através do MB são muito razoáveis, uma vez que todos os intervalos abrangem o valor extremo das amostras estudas, e as estimativas pontuais das amostras 30, 43 e 63 ultrapassam o maior indivíduo em cada uma destas (Tabela 5).

Para a variável B a eficiência média obtida é muito boa, já que independentemente da estimativa (MML ou MB) e a média ($\bar{y}_3$ **ou** $\bar{y}_5$) utilizadas para seu cálculo, os resultados estão muito próximos ao valor esperado do 100% (98,5% - 101,5%). Deve-se destacar que para a amostra n = 63 a eficiência média obtida através da estimativa do MML e com $\bar{y}_3$ e $\bar{y}_5$, é igual ao 100%, corroborando a aplicabilidade da teoria de valores extremos no melhoramento genético de plantas.

Tabela 5 - Comparação das estimativas pontuais e intervalos de confiança dos níveis de retorno com os valores máximos das amostras, e resultados das acurácias das médias dos três e cinco melhores indivíduos amostrais da variável B.

Nov. Famílias	Máx.	$\overline{y}_3$	$\overline{y}_5$	Met.	Estimativa	Intervalos	Eficiência Média	
							$\overline{y}_3$	$\overline{y}_5$
				Ve. P	18,562	**(18,316; 20,014)**	101,5%	100,9%
n = 20	**18,97**	18,85	18,72	A.Normal	18,562	**(18,053; 19,071)**	101,5%	100,9%
				Q.MCMC	18,859	**(18,238; 20,041)**	99,9%	99,3%
				Ve. P	18,695	**(18,684; 20,450)**	101,3%	100,7%
n = 30	**18,97**	18,93	18,83	A.Normal	18,695	**(18,181; 19,210)**	101,3%	100,7%
				Q.MCMC	19,011	**(18,404; 20,220)**	99,6%	99,1%
				Ve. P	18,789	**(18,779; 20,818)**	100,8%	100,6%
n = 43	**18,97**	18,94	18,91	A.Normal	18,789	**(18,247 ; 19,33)**	100,8%	100,6%
				Q.MCMC	19,117	**(18,523; 20,361)**	99,1%	98,9%
				Ve. P	18,868	**(18,826; 21,184)**	100,4%	100,3%
n = 63	**18,97**	18,95	18,92	A.Normal	18,868	**(18,285; 19,450)**	100,4%	100,3%
				Q.MCMC	19,208	**(18,627; 20,484)**	98,6%	98,5%

Fonte: Elaboração Própria - Máx.: Valor máximo - Met.: Metodologia - $\overline{y}_3$:Média dos três melhores indivíduos - $\overline{y}_5$:Média dos cinco melhores indivíduos - MLE :Máxima Verossimilhança - Ve. P: Perfil de Verossimilhança - A. Normal: Aproximação Normal – Q.MCMC: Quantis de Monte Carlo via Cadeias de Markov

Com os dados reais, segundo a metodologia Q. MCMC, os comportamentos (Tabelas 4 e 5) da eficiência seletiva em função do aumento do número de famílias avaliadas são apresentados nas Figuras 3 e 4. A projeção da curva para além das 63 famílias experimentais foi realizada via interpolação harmônica usando o decréscimo na taxa de incremento na variável resposta (eficiência) associado à taxa de acréscimo na variável regressora (tamanho amostral ou número de famílias)

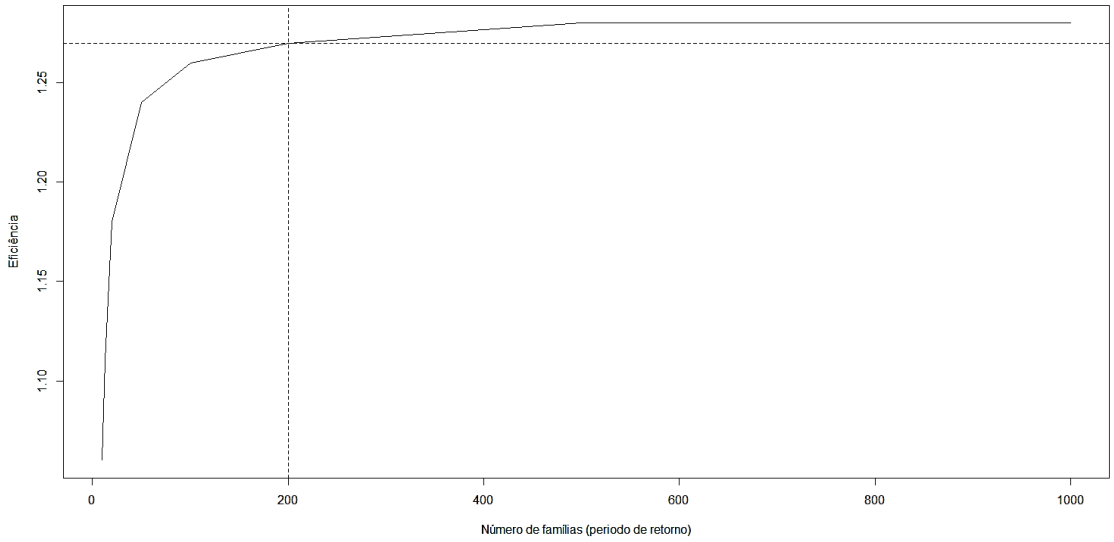


Figura 3 - Comportamentos da eficiência seletiva em função do aumento do número de famílias avaliadas para MMC.

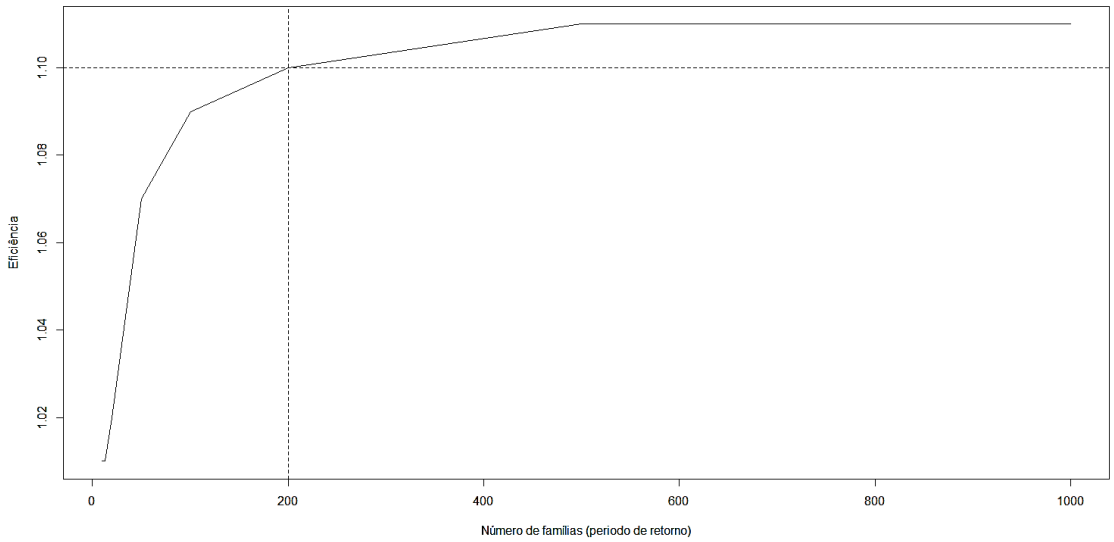


Figura 4 - Comportamentos da eficiência seletiva em função do aumento do número de famílias avaliadas para B.

Verifica-se que a eficiência (em relação ao uso de apenas 20 famílias) atingiu valores assintóticos em torno de 1,29 para a MMC e em torno de 1,10 para a B, associados a 200 famílias. Dessa forma, assim como na simulação

estocástica, pode-se recomendar a avaliação de 200 famílias em cada ciclo seletivo.

3.3 Seleção de famílias pela capacidade de geração de indivíduos superiores

As populações simuladas podem ser tomadas como se fossem diferentes famílias e os resultados da Tabela 2 podem ser usados para aferir uma metodologia para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos. Assim, conforme a Tabela 6, as famílias de códigos 1000 e 500 seriam selecionadas pela maior capacidade de geração de indivíduos extremos. E para recuperar um indivíduo extremo com valor 27, a família 100 demandaria tamanho amostral 200, a família 200 demandaria tamanho amostral 100, a família 500 demandaria tamanho amostral em torno de 50 e a família 1000 demandaria tamanho amostral 20 (Tabela 6).

Tabela 6 - Seleção pela maior capacidade de geração de indivíduos extremos.

Código Familia	nind	Valor máximo
20	20	20.03
20	100	22.43
20	200	23.37
50	20	22.17
50	100	24.62
50	200	25.38
100	20	23.53
100	100	25.99
100	200	26.6
200	20	24.73
200	100	27.18
200	200	27.61
500	20	26.08
500	100	28.49
500	200	28.69
1000	20	26.96
1000	100	29.32
1000	200	29.34

Fonte: Elaboração Própria

A metodologia mostrou-se funcional e é fundamentada da seguinte forma. Uma base de dados experimentais referentes à avaliação de famílias, mediante o uso de uma distribuição de valor extremo para predição do máximo das distribuições dos indivíduos, permite a previsão do comportamento da eficiência seletiva para os máximos associados a vários tamanhos de famílias e de populações experimentais. Isso possibilita ao melhorista a otimização da experimentação no melhoramento visando a seleção de indivíduos extremos. Para essas previsões, emprega-se o ***período de retorno*** associado à ocorrência de um ***nível do evento raro*** típico da distribuição ajustada. No caso, o ***período de retorno*** é interpretado como o ***tamanho amostral*** necessário para a ocorrência de determinado ***nível de retorno*** do evento raro (***valor extremo com sua magnitude***).

4. CONCLUSÕES

Uma base de dados experimentais referentes à avaliação de 200 famílias, mediante o uso da distribuição Weibull para predição do máximo das distribuições dos indivíduos, permite a previsão do comportamento da eficiência seletiva para variáveis tamanhos de famílias e de populações experimentais. Isso possibilita ao melhorista a otimização da experimentação no melhoramento visando a seleção de indivíduos extremos.

A simulação estocástica e a reamostragem de dados experimentais indicaram consistentemente que a avaliação de 200 famílias em cada ciclo seletivo maximiza a eficiência do melhoramento visando a seleção de indivíduos extremos.

Uma boa opção prática seria a avaliação de 200 famílias com 100 indivíduos, perfazendo um total de 20000 indivíduos.

Segundo a distribuição Weibull, o aumento da eficiência seletiva com o aumento do tamanho de família é em torno de 1,10 quando se passa de 20 para 100 indivíduos por família e de 1,12 quando se passa de 100 para 200 indivíduos e esses números são aproximadamente constantes independentemente do número de famílias avaliadas.

A metodologia é adequada também para classificar as famílias ou progênies pela capacidade de geração de indivíduos superiores ou excepcionais e informar os tamanhos amostrais a serem praticados em cada família para capturar esses indivíduos.

Os modelos Gumbel e Weibull mostraram-se adequados para analisar as massa média de colmos (MMC) e teor de Brix (B%), independentemente do número de famílias na amostra (n ≥ 20 famílias); sendo que a Gumbel mostrou-se adequada apenas nos casos de números de famílias muito pequenos. Assim, recomenda-se a Weibull para inferências práticas.

Recomenda-se realizar outras pesquisas onde analisem-se os valores extremos de outras variáveis biométricas em diferentes cenários, de tal maneira que possa-se indagar se as famílias de distribuições Fréchet, tem aplicação na analises dos valores extremos de variáveis utilizadas no melhoramento genético de plantas.

5. REFERÊNCIAS BIBLIOGRÁFICAS

ATHREYA, K.B.; FUKUCHI, J. Bootstrapping Extremes of I.I.D. Random Variables. **NIST SPECIAL** -Extreme Value Theory and Applications– Volume 3, Maryland, Publication 860. p.23-29, 1993.

AYALA G., MONTES F. **Teoría de la Probabilidad Departament d'Estadística i Investigació Operativa Universitat de Valéncia**, 109p, 2010.

BEIRLANT J.; GOEGEBEUR Y.; SEGERS J.; TEUGELS J. **Statistics of extremes: theory and applications**. London - Chichester : Wiley, 522p, 2004.

BOVIER, A. **Extreme values of random processes** -Lecture Notes-. Bonn: Institut für Angewandte Mathematik. 97p, 2010.

CASTILLO, E.; HADI, A. S.; BALAKRISHNAN, N.; SARABIA, J. M. **Extreme Value and Related Models with Applications in Engineering and Science.** New Jersey: Wiley, 353p, 2005.

CAIRES, S. Extreme Value Analysis: Wave Data. **JCOMM Technical**, Report No. 57, p. 1-30, 2011.

CONAB. (2015). Companhia Nacional de Abastecimento. Acompanhamento da Safra Brasileira de cana-de-açúcar. Safra 2014/2015. levantamento Mai/2014 Brasília: Conab, 2015. 14p. Disponível em:< http://www.conab.gov.br/OlalaCMS/uploads/arquivos/14_04_10_09_00_57_boletim_cana_portugues_-_4o_lev_-_13.pdf> Acesso em: 17 de junho 2015.

COOK,N.J. The designer's guide to wind loading on building structures. Part I: Background, damage survey, wind data and structural classification. Butterworth, 1985.

COLES, S. **An introduction to statistical modeling of extreme values.** London Berlin Heidelberg: Springer, 205p, 2001.

DE HAAN, L.; FERREIRA, A. **Extreme Value Theory: An Introduction.** New York: Springer, 417p, 2006.

DE HAAN, L; GOMES, M.I.; E CASTRO, L.C.; ALVES, M.I.F.; PESTANA D. Statistics of Extremes for IID Data. **Extremes,** Volume 11, p. 3-34, 2008.

EMBRECHTS P.; KLÜPPELBERG, C.; MIKOSCH, T.; **Modelling Extremal Events for Insurance and Finance**. Berlin. Springer, 644p, 1997.

EFRON, B.; TIBSHIRANI, R. **An Introduction to the bootstrap**. New York. Chapman and Hall, 430p ,1983.

FISHER, R. A.; TIPPETT, L.H.C. On the Estimation of the Frequency Distributions of the Largest or Smallest Member of a Sample. **Proceeding of the Cambridge Philosophical Society**, 24, p. 180-190, 1928.

GILLELAND, E. (2012). extRemes: Extreme Value Analysis. R package version 2.0-5 http://www.assessment.ucar.edu/toolkit/

GILLELAND, E. (2015). in2extRemes: Into the extRemes Package. R package version 1.0-2 http://www.assessment.ucar.edu/toolkit/

GNEDENKO, B. V. Sur la Distribution Limite du Terme Maximum d'une Série Aléatoire. **Annals of Mathematics**, 44, p. 423-453, 1943.

GUMBEL, E. J. **Statistics of Extremes**. New York: Dover Publ, 371p, 2004.

HARRIS, R.I. Extreme value analysis of epochmaxima-convergence, and choice of asymptote, **JWEIA**, 59, p.1–22, 2004.

KETCHEN, D.J.; KETCHEN, D. J. JR.; BERGH, D. D. Research Methodology in Strategy and Management, **Emerald,** Group Publishing Limited, Volume 3, 2006.

KOTTEGODA, N.; ROSSO, R. **Applied Statistics For Civil and Environmental Engineers.** Second Edition. Oxford: Blackwell Publishing, 705p, 2008.

KINNISON, R.P. **Applied Extreme Value Statistics**. New York: Battelle Press, Columbus. 1985.

KOTZ S.; NADARAJAH S. **Extreme value distributions: theory and applications**., London: Imperial College Press, 185p, 2000.

ORTEGA S. J. **Introducción a la Teoría de Valores Extremos**- Notas de clase-. Guanajuato: CIMAT, 130p, 2010.

R Development Core Team (2015). R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org

REISS, R-D.; THOMAS, M; **Statistical Analysis of Extreme Values With Applications to Insurance, Finance, Hidrology and Other Fields**. Third Edition. Basel: Birkhauser Verlag, 508p, 2007.

RESENDE M.D.V. Delineamento de experimentos de seleção para maximização da acurácia seletiva e do progresso genético. ***Revista Arvore***, v. 19, n.4, p.479-500, 1995.

RESENDE M.D.V.; BARBOSA M. **Melhoramento Genético de Plantas de Propagação Assexuada**, Colombo: Embrapa Florestas, 121p, 2005.

RESENDE, M.D.V., **Inferência bayesiana e simulação estocástica (amostragem de gibbs) na estimação de componentes de variância e de valores genéticos em plantas perenes**. Colombo: Embrapa Florestas 68p, Documentos, 46, 2000.

SPROTT, D.A. **Statistical Inference in Sciense**. New York: Springer, 229p, 2000.

SMITH, R. L. Maximum likelihood estimation in a class of non-regular cases. ***Biometrika*** 72, p. 67-90, 1985.

SMITH, R. L. **Statistics of Extremes, With Applications in Environment, Insurance and Finance**. University of North Carolina, 62p, 2003.

STEPHENSON, A. (2012). Evd: Functions for extreme value distributions. R package version 2.3-0 CRAN.

STEPHENSON, A.; RIBATET, M. (2014). evdbayes: Bayesian Analysis in Extreme Value Theory. R package version 1.1-1 CRAN.

TRAUTMANN, H.; STEUER, D.; MERSMANN O.; BORNKAMP B. (2014). truncnorm: Truncated normal distribution. R package version 1.0-7 CRAN.

WACLAWOVSKY, A. J.; SATO, P.M.; LEMBKE, C.G.; MOORE, P.H.;SOUZA, G.M. Sugarcane for bioenergy production: an assessment of yield and regulation of sucrose content. **Plant Biotechnology Journal**, 8(3), P. 263-276, 2010.

ANEXO 1: **Simulação de Dados cenário C1.**

```
library(extRemes) # Trabalho com o pacote extRemes
library(in2extRemes)
library(evd)
library(truncnorm)


normal <- function (n, k) # simulação normal sd=1,2,3,4,5 #
{
for(i in 1:4)
 {
  m=i*5
  for(i in 1:5)
  {
    s=i*1
    var1<- (sort(rtruncnorm(n, a=0, b=Inf, m, s),decreasing=T))
    arquivo <- stringi::`%stri+%`("E:/VIÇOSA/Resultados_finais10.1/ppa", k)
    arquivo <- stringi::`%stri+%`(arquivo, ".txt")
    write.table(var1, arquivo,append=T, sep=" ",row.names = T,col.names =F, dec=".")

    var2<- data.frame(var1[1])
    arquivo1 <- stringi::`%stri+%`("E:/VIÇOSA/Resultados_finais10.1/ppb", k)
    arquivo1 <- stringi::`%stri+%`(arquivo1, ".txt")
    write.table(var2,arquivo1,append=T, sep=" ",row.names = T,col.names =F, dec=".")
  }
 }
}


DEV <- function (n,k)
{
  for(i in 1:k)
  {
    normal(n, i)#chamar a função normal

    arquivo2 <- stringi::`%stri+%`("ppb",i)
    arquivo2 <- stringi::`%stri+%`(arquivo2, ".txt")
    pepe <- read.table(arquivo2, header = F)

    ordenar <- sort(pepe$V2, decreasing=T)
    maior <- as.character(data.frame(ordenar[1]))
    media <- as.character(mean(pepe$V2))
    desvio <- as.character(sd(pepe$V2))

    resulP <- as.character(fevd(V2, pepe, units="Ton")) #GEV
    reP <- as.character(fevd(V2, pepe, type = "Gumbel" ,units="Ton")) #GUMBEL
    arquivo3 <- stringi::`%stri+%`("E:/VIÇOSA/Resultados_finais10.1/rrr", i)
    arquivo3 <- stringi::`%stri+%`(arquivo3, ".txt")
    write.table(resulP,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(reP,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
```

```
  outros <- fevd(V2, pepe,  units="Ton") # GEV
  gum <- fevd(V2, pepe, type = "Gumbel",units="Ton") # GUMBEL

  print(stringi::`%stri+%`("Arquivo", i))
  print(outros$results$par) #GEV - Location, Scale, Shape
  print(gum$results$par) #GUMBEL - Location, Scale, Shape

  tes <-lr.test(gum, outros) # Likelihood-Ratio Test
  print(tes$p.value)
  vp <- as.character(tes$p.value)
  write.table(vp,arquivo3,append=T, sep=" ",row.names = T,col.names =F)

  if (tes$p.value <= 0.05)
  {
   print(stringi::`%stri+%`("KolmoGEV", i))
   po=outros$results$par[1]
   es=outros$results$par[2]
   fo=outros$results$par[3]
   pepex <- pepe[,2]
   ko<-ks.test(pepex,"pevd",po,es,fo,alternative="two.sided")
   print(ko$p.value)
   kolmo <- as.character(ko$p.value)
   write.table(kolmo,arquivo3,append=T, sep=" ",row.names = T,col.names =F)

   print(stringi::`%stri+%`("I.C.PARAMETROSGEV", i))
   PAR1GEV <- ci(outros, type = "parameter",which.par = 1, xrange = c(1, 50),nint = 100,
method = "proflik", verbose = F)
   PAR2GEV <-ci(outros, type = "parameter",which.par = 2, xrange = c(1, 50),nint = 100,
method = "proflik", verbose = F)
   PAR3GEV <-ci(outros, type = "parameter",which.par = 3, xrange = c(-2, 2),nint = 100,
method = "proflik", verbose = F)

   print(PAR1GEV)
   print(PAR2GEV)
   print(PAR3GEV)

   location <- as.character(PAR1GEV)
   scale <- as.character(PAR2GEV)
   shape <- as.character(PAR3GEV)

   write.table(location,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
   write.table(scale,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
   write.table(shape,arquivo3,append=T, sep=" ",row.names = T,col.names =F)

   print(stringi::`%stri+%`("I.C.RETORNOSGEV", i))
   PR20GEV <- ci(outros,return.period=20, method = "proflik",xrange = c(20,50),verbose=F)
   PR50GEV <- ci(outros,return.period=50, method = "proflik",xrange = c(20,55),verbose=F)
   PR100GEV <- ci(outros,return.period=100, method = "proflik",xrange = c(20,65),verbose=F)
   PR200GEV <- ci(outros,return.period=200, method = "proflik",xrange = c(25,65),verbose=F)
   PR500GEV <- ci(outros,return.period=500, method = "proflik",xrange = c(25,70),verbose=F)
 PR1000GEV <- ci(outros,return.period=1000, method = "proflik",xrange = c(27,58),verbose=F)
```

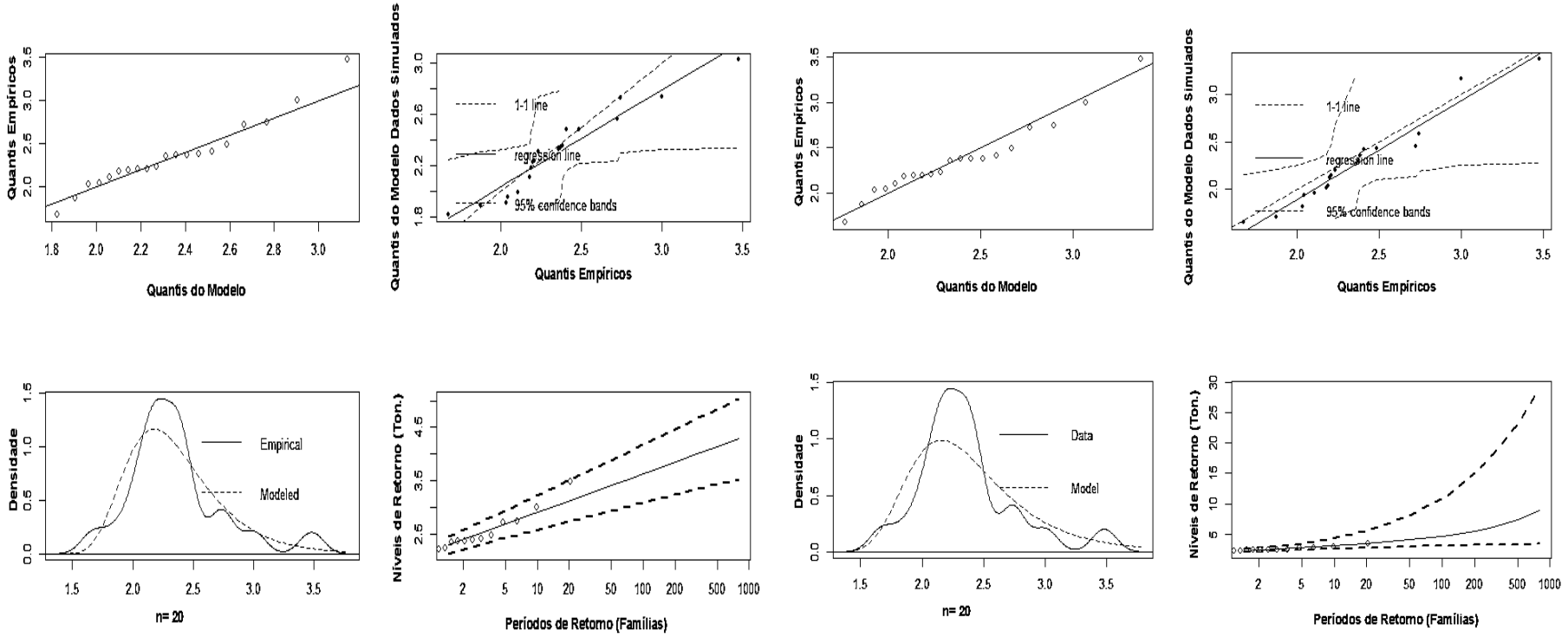
```
    print(PR20GEV)
    print(PR50GEV)
    print(PR100GEV)
    print(PR200GEV)
    print(PR500GEV)
    print(PR1000GEV)

    GEVPR20 <- as.character(PR20GEV)
    GEVPR50 <- as.character(PR50GEV)
    GEVPR100 <- as.character(PR100GEV)
    GEVPR200 <- as.character(PR200GEV)
    GEVPR500 <- as.character(PR500GEV)
    GEVPR1000 <- as.character(PR1000GEV)

    write.table(GEVPR20,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GEVPR50,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GEVPR100,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GEVPR200,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GEVPR500,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GEVPR1000,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(media,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(desvio,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(maior,arquivo3,append=T, sep=" ",row.names = T,col.names =F)

  }
  else
  {
    print(stringi::`%stri+%`("KolmoGUMBEL", i))
    po2= gum$results$par[1]
    es2= gum$results$par[2]
    pepex <- pepe[,2]
    ko1<-ks.test(pepex,"pgumbel",po2,es2,alternative="two.sided")
    print(ko1$p.value)
    kolmo1 <- as.character(ko1$p.value)
    write.table(kolmo1,arquivo3,append=T, sep=" ",row.names = T,col.names =F)

    print(stringi::`%stri+%`("I.C.PARAMETROSGUMBEL", i))
    PAR1GUM <- ci( gum , type = "parameter",which.par = 1, xrange = c(1, 50),nint = 100,
method = "proflik", verbose = F)
    PAR2GUM <-ci( gum , type = "parameter",which.par = 2, xrange = c(1, 50),nint = 100,
method = "proflik", verbose = F)

    print(PAR1GUM)
    print(PAR2GUM)

    GUMlocation <- as.character(PAR1GUM)
    GUMscale <- as.character(PAR2GUM)

    write.table(GUMlocation ,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GUMscale,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
```

```
    print(stringi::`%stri+%`("I.C.RETORNOSGUMBEL", i))
    PR20GUM <- ci(gum,return.period=20, method = "proflik",xrange = c(20,50),verbose=F)
    PR50GUM <-ci(gum,return.period=50, method = "proflik",xrange = c(20,55),verbose=F)
    PR100GUM <-ci(gum,return.period=100, method = "proflik",xrange =c(20,62) ,verbose=F)
    PR200GUM <-ci(gum,return.period=200, method = "proflik",xrange = c(21,67),verbose=F)
    PR500GUM <-ci(gum,return.period=500, method = "proflik",xrange = c(21,72),verbose=F)
   PR1000GUM <-ci(gum,return.period=1000, method = "proflik",xrange = c(22,75),verbose=F)

    print(PR20GUM)
    print(PR50GUM)
    print(PR100GUM)
    print(PR200GUM)
    print(PR500GUM)
    print(PR1000GUM)

    GUMPR20 <- as.character(PR20GUM)
    GUMPR50 <- as.character(PR50GUM)
    GUMPR100 <- as.character(PR100GUM)
    GUMPR200 <- as.character(PR200GUM)
    GUMPR500 <- as.character(PR500GUM)
    GUMPR1000 <- as.character(PR1000GUM)

    write.table(GUMPR20,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GUMPR50,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GUMPR100,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GUMPR200,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GUMPR500,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(GUMPR1000,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(media,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(desvio,arquivo3,append=T, sep=" ",row.names = T,col.names =F)
    write.table(maior,arquivo3,append=T, sep=" ",row.names = T,col.names =F)

   }
  }
}
```

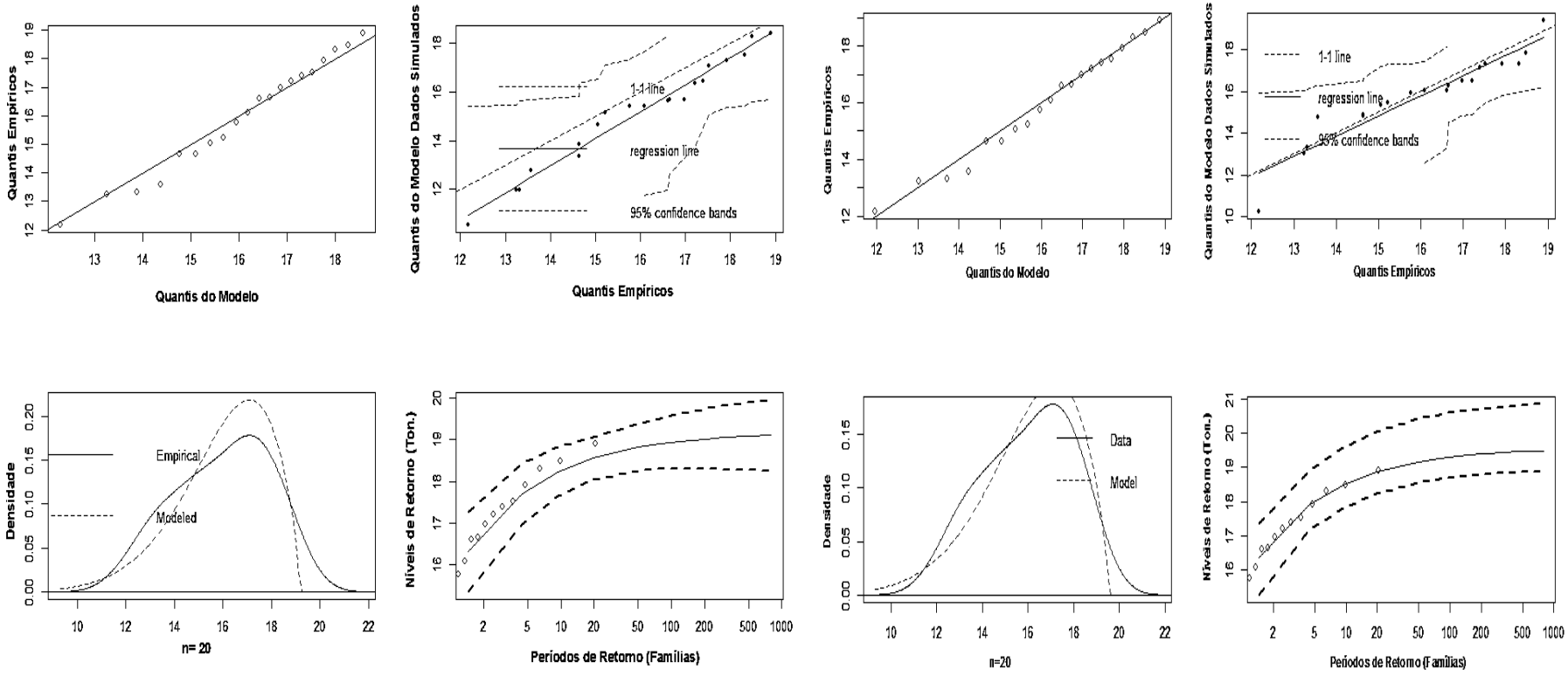
DEV (20,100)

Nota: Para os demais cenários utilizaram-se rotinas muito similares, só mudam os respectivos valores da distribuição normal truncada segundo explica-se em materiais e métodos.

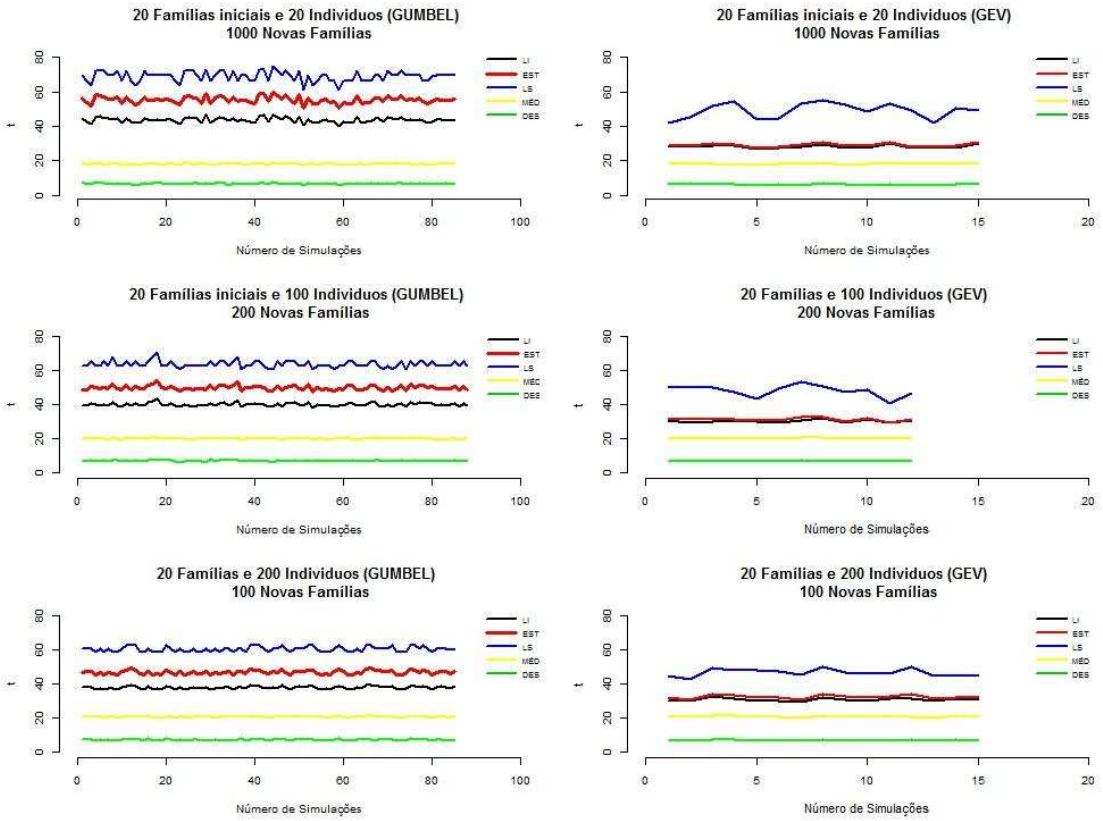
ANEXO 2: Diagnósticos dos modelos Gumbel MML (painel esquerdo) e Gumbel Bayesiana (painel direito) para a variável MMC. Gráficos Quantil-quantil (esquinas superiores esquerdas) – Gráfico de Quantis a partir de uma amostra retirada do modelo Gumbel ajustada contra os quantis dos dados empíricos com bandas do 95% de confiança (esquinas superiores direitas) - Gráficos de densidade de dados empíricos e Gumbel ajustada (esquinas inferiores esquerdas) – Gráficos dos níveis de retorno pontuais e intervalos de confiança do 95% com aproximação normal (esquinas inferiores direitas).



ANEXO 3: Diagnósticos dos modelos GEV MML (painel esquerdo) e GEV Bayesiana (painel direito) para a variável B. Gráficos Quantil-quantil (esquinas superiores esquerdas) – Gráfico de Quantis a partir de uma amostra retirada do modelo GEV ajustada contra os quantis dos dados empíricos com bandas do 95% de confiança (esquinas superiores direitas) - Gráficos de densidade de dados empíricos e GEV ajustada (esquinas inferiores esquerdas) – Gráficos dos níveis de retorno pontuais e intervalos de confiança do 95% com aproximação normal (esquinas inferiores direitas).



ANEXO 4: Intervalos de confiança dos níveis de retorno para 1000, 200 e 100 novas famílias dos modelos ajustados com 20 famílias (20, 100 e 200 indivíduos por família) da variável produtividade de açúcar por hectare. LI: Limite Inferior - EST: Estimativa - LS: Limite Superior - MÉD: Média dos valores máximos das famílias - DES: Desvio padrão dos valores máximos das famílias.



ANEXO 5: Intervalos de confiança dos níveis de retorno para 20, 50 e 100 novas famílias dos modelos ajustados com 100 e 200 famílias (20 indivíduos e 100 indivíduos por família) da variável produtividade de açúcar por hectare. **LI**: Limite Inferior - **EST**: Estimativa - **LS**: Limite Superior - **MÉD**: Média dos valores máximos das famílias - **DES**: Desvio padrão dos valores máximos das famílias.