ESTIMADORES *LS*, *DRLS* E τ NO MODELO DE REGRESSÃO LINEAR – ESTUDO COMPARATIVO POR SIMULAÇÃO

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Sra *Elisa Maria Caetano dos Santos* e aprovada pela Comissão Julgadora

Campinas, 13 de novembro de 1989

Prof. Dr.Oscar H. Bustos
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito para obtenção do Título de Mestre em Estatística.

AGRADECIMENTOS

Ao meu orientador e amigo, Professor Doutor Oscar Humberto Bustos, pelo exemplo de seriedade, dedicação e companheirismo.

Ao meu marido, Professor José Carlos da Rocha Castelar Pinheiro, que contribuiu com muitas idéias e discussões e teve participação decisiva na elaboração dos gráficos.

A Maria Helena Simões Caetano dos Santos – Mestre em Letras, UFF – pelo cuidadoso trabalho de revisão do texto.

Ao amigo Nelson Leal Filho pelo valioso apoio técnico durante a operacionalização do Estudo Monte Carlo.

A Luisa La Croix pela oportunidade de realizar este trabalho.

Ao Professor Djalma G.C. Pessoa pelo apoio e acolhida durante a elaboração.

A Lais Ventura Santos pelo excelente trabalho de edição.

Aos professores e funcionários do Instituto de Matemática, Estatística e Ciência da Computação - IMECC.

Ao Instituto de Matemática Pura e Aplicada - IMPA pelo uso de suas instalações e equipamentos.

Aos meus pais.

R E S U M O

O modelo de regressão linear é, sem dúvida, uma das técnicas mais empregadas em análise estatística e o método de mínimos quadrados, o mais popular na estimação dos parâmetros dos modelos. No entanto estes estimadores podem ser desastrosos na presença de um único *outlier*.

Um dos procedimentos de estimação mais utilizados na prática, consiste em, a partir de técnicas de diagnóstico, detectar e rejeitar os *outliers* e, então, refazer o ajuste por mínimos quadrados (least squares - LS) com os dados que restaram. Neste estudo tais estimadores recebem a designação genérica de $DRLS$-$DETECÇÃO$+$REJEIÇÃO$+LS.

O objetivo deste trabalho é avaliar alguns dos estimadores $DRLS$ mais utilizados, comparando-os entre si através de um estudo Monte Carlo.

Considera-se ainda um estimador robusto, o τ-estimador (Yohai e Zamar, 1986), sendo este o primeiro estudo de simulação desenvolvido para o mesmo.

ÍNDICE

CAPÍTULO 1

MOTIVAÇÃO – *OUTLIERS* NO MODELO DE REGRESSÃO LINEAR

Observa-se, freqüentemente, que as hipóteses necessárias à aplicação de um determinado modelo probabilístico não se verificam integralmente e que portanto, quase sempre, o que se pode esperar, no dia a dia, são flutuações em torno de tais hipóteses.

Neste trabalho, a ocorrência de *outliers* é considerada, no contexto de regressão linear, como manifestação de violações das hipóteses do modelo, segundo situações bastante freqüentes na prática.

O modelo de regressão linear é uma das técnicas mais empregadas em análise estatística e o método de mínimos quadrados, o mais popular na estimação dos parâmetros do modelo. No entanto esses estimadores podem ser desastrosos na presença de um único *outlier*.

Um dos procedimentos mais utilizados na prática, consiste em, a partir de técnicas de diagnóstico (baseadas nos resíduos do ajuste e ou na diagonal da matriz de projeção), identificar e rejeitar os *outliers* e, então, refazer o ajuste por mínimos quadrados (daqui por diante representado por LS de Least Squares) com os dados que restaram.

O objetivo deste trabalho é avaliar alguns destes procedimentos que neste estudo recebem a designação genérica de $DRLS - Detection + Rejection + LS$, comparando-os entre si através de um estudo Monte Carlo, generalizando, para o modelo de regressão linear, algumas das análises que foram aplicadas ao modelo de posição em Hampel (1985).

Observa-se também o comportamento do τ-estimador, sugerido por Yohai e Zamar (1986), que procura combinar eficiência sob normalidade e alto ponto de ruptura.

Na seção 1.1 apresenta-se o conceito de *outlier*. A seção 1.2 trata de como e com que freqüência os *outliers* ocorrem, enquanto que em 1.3 aborda-se a questão de como proceder

na presença de *outliers*.

1.1 O que é *outlier*

Os *outliers* são observações que, na opinião do analista de dados, parecem inconsistentes em relação ao restante do conjunto. Ressalte-se que, intuitivamente, a fidedignidade de uma observação está intimamente relacionada com o comportamento das outras, obtidas sob as mesmas condições.

A preocupação com *outliers* é antiga. Em Bernoulli (1777) há indicação de que retirar os *outliers* da amostra e trabalhar com o restante era atitude comum há 200 anos. A primeira tentativa de desenvolver métodos estatísticos objetivos para tratar com *outliers* data de 1850. Era de esperar-se que, depois de tanto tempo, se tivesse chegado a uma definição objetiva do conceito de *outlier* e que se dispusesse de métodos consagrados para tratar com eles. No entanto isto não acontece. Beckman e Cook (1983) citam duas definições de *outliers* extraídas de Edgeworth (1887) e de Grubbs (1969), para mostrar como o conceito se mantém vago, como há oitenta anos atrás.

Discordant observations may be defined as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined. (Edgeworth, 1887).

An outlying observation, or outlier, is one that appears to deviate markedly from the others members of the sample in which it occurs. (Grubbs, 1969).

Historicamente, os métodos *objetivos* para tratar com *outliers* eram empregados somente depois de uma identificação visual, quando já se suspeitava de alguma observação. No entanto, hoje, com a disponibilidade dos computadores, podem-se *manusear* grandes conjuntos de dados, inclusive em contextos mais complexos como de análise de regressão, análise multivariada e séries temporais, onde a inspeção visual muitas vezes é insuficiente.

Surge assim a necessidade de criarem-se e adotarem-se critérios objetivos para trabalhar-se na presença de *outliers*. Muitos desses critérios estão baseados em modelos assumidos a priori, na etapa de modelagem do problema. Isto significa que, além de critérios objetivos para tratar com *outliers*, necessita-se de um conjunto de técnicas para assegurar a adequabilidade do modelo selecionado para explicar o fenômeno.

Um exemplo muito interessante a respeito do comportamento do observador, diante dos dados coletados e do modelo selecionado para explicar o fenômeno, foi extraído de Bustos (1988 pág. 6).

Por muitos anos, os cientistas da NASA erraram ao rejeitar, automaticamente, dados que significassem quedas nos níveis de ozônio da camada atmoférica superiores a 30%. Só depois que os ingleses anunciaram em 1985 um deficit na camada de ozônio, sobre a Antartida, é que foram recuperados os registros da NASA, comprovando-se o que os dados do satélite já haviam revelado.

Isso vem mostrar a necessidade de se aprofundar o estudo de técnicas robustas que permitam tratar os dados com mais confiabilidade, sobre um espaço mais amplo de modelos probabilísticos, do que o usualmente permitido em estatística clássica.

1.2 Como e com que freqüência os *outliers* ocorrem

Como já foi dito, anteriormente, não se pode dissociar o conceito de *outlier* da distribuição que se supõe adequada para modelar os dados. Isto significa que a identificação de determinada observação como *outlier* pode mudar bastante, de acordo com o modelo probabilístico que se adote.

O tipo mais freqüente de *outliers* observado na prática são os chamados erros grosseiros. Eles podem ser gerados por erros de transcrição, troca de dois valores ou grupo de valores em um desenho estruturado, falha de equipamento ou por observação de um membro de uma população diferente. Embora sejam potencialmente desastrosos, geralmente, podem ser evitados, empregando-se alguns cuidados.

Ressalte-se que muitas vezes, antes do estatístico ser consultado, os dados passam por algum tipo de crítica, fazendo com que o percentual de erros grosseiros e de outros tipos de *outliers* registrados na literatura, seja, talvez, um pouco menor que o originalmente encontrado na população. Mesmo assim a freqüência de erros grosseiros varia bastante. A seção 1.2 do Capítulo 1 de Hampel et alii (1986) reune vários exemplos reais em que foram registradas as freqüências de erros.

Tal percentual varia, aproximadamente, entre 0,01% e 20% relativos, respectivamente, aos dados levantados no censo americano em 1950, previamente criticados, e aos famosos

dados de Venus Babylonian, encontrados em Huber (1974).

Uma grande variedade de exemplos aplicados à indústria, ciência e agricultura são encontrados em Daniel e Wood (1971). A freqüência correspondente aos exemplos relativos a regressão varia entre 0% e 19%, embora 0% e 1% sejam mais típicos. Discutindo vários tipos de dados de indústria em seu curso, em Berkeley, em 1986, Cuthbert Daniel considerou freqüências entre 1% a 10% e 20% como usuais e citou como excepcional um conjunto com 3000 observações, aparentemente sem erros.

Mais detalhes e particularidades a respeito de *outliers* no modelo de regressão linear são deixados para o Capítulo 3, seção 3.1.

1.3 Como proceder na presença de *outliers*

Três atitudes básicas podem ser tomadas diante de *outliers*: ignorá-los, rejeitá-los ou acomodá-los.

A primeira, obviamente, é a mais simples de todas e a mais desastrosa. As outras duas, embora pareçam antagônicas, têm a identificação do *outlier* como elo, já que um método de acomodação pode produzir um método de identificação e vice-versa.

A rejeição pode basear-se em critérios subjetivos, como análise gráfica, e/ou critérios mais objetivos, como os testes de discrepância ou discordância. Como qualquer teste estatístico, este também trabalha com duas hipóteses: a hipótese nula, que será conservada caso não haja evidência suficiente para indicar sua rejeição, e a hipótese alternativa, que será assumida no caso de se rejeitar a primeira. Para um teste de discordância de *outliers*, a hipótese nula expressa a ausência de *outliers* através de um modelo probabilístico básico responsável pela geração de todos os dados. Por sua vez, a alternativa expressa a forma pela qual o modelo original pode ser modificado, para explicar a presença de *outliers*. No Capítulo 3, seção 3.3, encontram-se as informações a respeito dos modelos utilizados neste estudo para a contaminação do modelo de regressão linear.

Os métodos de acomodação de *outliers*, por sua vez, procuram diminuir os efeitos dessas observações sobre os resultados finais, empregando os chamados métodos de estimação robusta.

Robustez é um conceito de grande importância em inferência estatística, de forma

global e não apenas no que diz respeito ao estudo de *outliers*. Nos últimos anos, muito esforço tem sido dedicado a produzir procedimentos que sejam insensíveis a pequenos afastamentos das hipóteses assumidas a respeito dos mecanismos de geração dos dados. Com base neste princípio foram e estão sendo criados métodos robustos para estimação e teste com propriedades estatísticas desejáveis sob uma gama de distribuições numa certa vizinhança.

Ressalte-se que, embora se admita que a adoção de procedimentos do tipo *rejeição de outliers seguido de estimação por mínimos quadrados* resulte em estimação robusta (num certo sentido), não é a esse tipo que se faz referência nestes dois últimos parágrafos.

Hampel (1985) ressalta que o tratamento de *outliers* pode ser abordado em três níveis distintos. No nível mais baixo, o problema é isolado e tratado com a aplicação direta de um teste de hipótese, onde o nível de significância fornece proteção contra declarar que uma observação é *outlier*, quando não o é. De acordo com Hawkins (1980) e Beckman e Cook (1983) isso ainda é um paradigma dominante na literatura, pois, na verdade, o que se gostaria de controlar é a probabilidade de aceitar um ponto como *bom*, quando na verdade é *ruim*. No segundo nível, o tratamento de *outliers* faz parte de um procedimento global. O principal objetivo é acomodar os *outliers*, minimizando seus efeitos. Este é precisamente um dos propósitos da estatística robusta. O segundo objetivo é identificar *outliers* a fim de conhecê-los melhor, procurar sua origem e aceitá-los, como uma possível indicação de inadequabilidade do modelo suposto inicialmente.

Hampel destaca ainda a análise de resíduos provenientes de ajuste robusto como uma forma de identificação de *outliers*.

E, finalmente, o terceiro nível, o mais amplo de todos, onde os *outliers* são discutidos e interpretados da forma mais completa possível dentro do contexto de análise de dados. Recorre-se não somente a procedimentos estatísticos formais, mas também a outras fontes de informação. Neste nível, a experiência em análise de dados é mais importante do que os resultados de qualquer teste, mas ainda é necessário utilizar métodos de acomodação e identificar observações discrepantes como no nível 2. No decorrer do trabalho, estão apresentadas em detalhes algumas aplicações de testes de discordância. Também estão descritos e exemplificados procedimentos utilizados na acomodação de *outliers*.

CAPÍTULO 2

ESTIMADORES DO TIPO *REJEITA+MÉDIA* – ESTUDO DE HAMPEL PARA O MODELO DE POSIÇÃO

Neste capítulo apresentam-se as principais idéias e resultados descritos em Hampel (1985).

Segundo Bustos (1988), *trata-se de um trabalho pioneiro sobre a relação entre outliers e robustez. É um dos primeiros a ressaltar a utilidade do conceito de ponto de ruptura (breakdown point) tanto na teoria como na prática.*

No passado, métodos de rejeição de *outliers* eram investigados, sem medir os impactos nas fases seguintes de estimação e aplicação de testes. Além disso, embora rejeição de *outliers* seguida de mínimos quadrados seja o procedimento robusto mais antigo e mais difundido, até pouco tempo nenhuma comparação tinha sido feita com outros estimadores robustos.

Neste sentido, Hampel (1985) trata da estimação da média populacional, supondo o modelo normal, na presença de *outliers*.

Analisa, a partir de um estudo Monte Carlo, o comportamento de estimadores do tipo *rejeita+média* e explica os resultados obtidos, utilizando o conceito de ponto de ruptura. Tais estimadores são obtidos em dois estágios: primeiro, aplica-se um teste de discordância aos dados com o objetivo de identificar e rejeitar as observações consideradas *ruins* (limpar a amostra), e depois, obtém-se a média aritmética dos que restaram. Alguns procedimentos deste tipo são comparados entre si e com outros estimadores robustos. Hampel (1985) explica ainda o efeito de *mascaramento*, utilizando o conceito de ponto de ruptura e fornece fórmulas para o cálculo deste último para as seis regras de rejeição consideradas.

2.1 Características Básicas do Estudo

Foram considerados trinta e dois procedimentos do tipo *rejeita+média* criados a partir de seis tipos básicos de regras de rejeição. As variações foram obtidas, fixando-se diferentes valores críticos para os correspondentes testes de discrepância e por versões *com um passo (one Step)* e iterativas. As versões de *um passo* rejeitam no máximo um *outlier*, enquanto as iterativas se aplicam cada vez à observação mais distante, de forma sucessiva, até que cesse a rejeição. Os testes selecionados se aplicam a rejeição de *outliers* à direita e à esquerda.

Os estimadores utilizados estão descritos na seção seguinte.

Como em Andrews et alii (1972), foram selecionadas 640 a 1000 amostras de 20 observações das seguintes distribuições:

$N = N(0,1)$

5% $3N$ (1 *outlier* perto dos dados *bons*).

5%$10N$ (1 *outlier* distante)

10%$10N$ (considerado um exemplo bastante real de dados de baixa qualidade sem crítica).

$T3$ (caudas levemente alongadas, exemplo de dados de alta qualidade).

25%$10N$ (contaminação excessiva dados de baixa qualidade)

Cauchy (contaminação pesada).

onde $x\%yN$ significa que $x\%$ das observações provém de uma distribuição $N(0,y)$ enquanto o restante de uma $N(0,1)$.

A escolha dessas distribuições contempla várias situações de contaminação em torno da $N(0,1)$.

2.2 Os Estimadores

2.2.1 - Procedimentos Robustos

a) 50% ou mediana

b) $H15$ ou *Huber proposal-2* com $k = 1.5$ (Huber, 1964)

Tem por objetivo estimar, simultaneamente, os parâmetros de locação e escala, resolvendo o seguinte sistema de equações:

$$\begin{cases} \sum_{j=1}^{n} \psi[(x_j - T)/S] = 0 \\ \sum_{j=1}^{n} \chi[(x_j - T)/S] = 0 \end{cases}$$

onde

$$\psi(t) = \begin{cases} t & |t| \leq k \\ k(\text{ sinal de } t) & |t| > k \end{cases}$$

e $$\chi(t) = \psi^2(t) - \beta(k)$$

onde $\beta(k) = E_\phi \psi^2 = \int \psi^2(x) d\phi(x)$ onde ϕ é a função de distribuição da $N(0,1)$.

Os valores iniciais para T e S utilizados em Andrews et alii (1972) foram $T_0 = \text{med}_i(x_i)$, mediana das observações, e $S_0 = [(Q_3 - Q_1)/2] \times 0,6745$ e $0,6745 \cong [\phi^{-1}(3/4) - \phi^{-1}(1/4)]/2$ onde ϕ é a função de distribuição da $N(0,1)$.

O estimador T resultante é $H15$.

A função de influência desse estimador é ilustrada no gráfico 1 apresentado no anexo 1. Observe-se que a influência de valores grandes em módulo é limitada.

c) $25A$ (M-estimador)(Hampel, 1974)

O objetivo é encontrar um T que satisfaça a

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T}{S}\right) = 0,$$

onde $S = 1,483\,\mathrm{MAD}(x_i) = 1,483\,\mathrm{med}_i\{|x_i - \mathrm{med}_j(x_j)|\}$ e MAD abreviatura de MEDIAN ABSOLUTE DEVIATION

$$\psi(t;a,b,c) = \begin{cases} t, & |t| \leq a \\ a \text{ sinal } (t), & a < |t| \leq b \\ a(c \text{ sinal } (t) - t)/c - b, & b < |t| \leq c \\ 0, & |t| > c \end{cases}$$

A equação é resolvida, iterativamente, usando $T_0 = \mathrm{med}_i(x_i)$ como estimativa inicial de T. Para o estimador $25A$, $a = 2,5$, $b = 4,5$ e $c = 9,5$.

Observações:

1 - Hampel et alii (1986) recomendam o uso de $S = 1,483\,\mathrm{MAD}(x_i)$ como estimativa inicial de escala para os M-estimadores. S, assim definido, tem ponto de ruptura de 50%. Além disso, estudos de simulação têm mostrado a superioridade dos M-estimadores resultantes, sobre muitos outros (Andrews et alli, 1972, pg. 239). Ressalte-se, ainda, a facilidade de cálculo comparado à versão com estimação simultânea (como em $H15$). O gráfico 2 do anexo 2 ilustra o comportamento da função de influência de 25A. Observe que a curva decresce diminuindo gradativamente o peso de pontos cada vez mais distantes, até atingir a rejeição completa (peso zero).

2 - As denominações $H15$ e $25A$ são as usadas no Estudo de Princeton em Andrews et alii (1972).

2.2.2 - Procedimentos do tipo *rejeita+média*

Como já foi apresentado na seção 2.1, a construção desses estimadores é feita a partir de uma regra de rejeição. Esta por sua vez, resulta da aplicação de um teste de discordância, cuja hipótese nula é de que não há *outliers* na amostra ou seja:

$H_0 =$ As observações são realizações de variáveis aleatórias $X_1, \ldots, X_n$ independentes e identicamente distribuídas com distribuição $N(\mu, \sigma)$ com μ e σ desconhecidos.

Por outro lado a hipótese alternativa, H_1, pode ser escrita como:

$H_1 =$ Existe X_i com distribuição G diferente de $N(\mu, \sigma)$ que é o *outlier*.

Se H_0 é rejeitada, elimina-se a observação mais afastada.

Para os procedimentos *de um passo*, aplica-se o teste apenas uma vez. Nas versões iterativas, o teste é repetido até que cesse a rejeição de H_0. O estimador da média é então obtido, tomando-se a média aritmética das observações que restaram.

Apresentam-se, a seguir, descrições resumidas das regras de identificação de *outliers* empregadas neste estudo. Mais detalhes sobre essas e muitas outras podem ser obtidos em Barnett e Lewis (1984).

a) Quarto Momento ou Curtose[1] - $X24, X25, X26 e X27$ (teste $N15$ em Barnett e Lewis (1984) pág. 175). O teste rejeita a hipótese nula para valores altos da estatística apresentada a seguir.

$$T_{N15} = \sum_{i=1}^{n} (x_i - \overline{x})^4 / (nS^4)$$

onde $S^2 = \sum_{i=1}^{n} (x_i - \overline{x})^2 / (n-1)$.

É usado, geralmente, para detectar qualquer tipo de afastamento da hipótese de normalidade dos dados.

Os valores críticos para os níveis de significância de 5% e 1% podem ser encontrados em Barnett e Lewis (1984) tabela XVb, página 392.

b) Amplitude studentizada - X34, X35 e X36. (Teste N6 em Barnett e Lewis (1984) pág. 171). Utilizado para a detecção de um par de *outliers*: os extremos superior e inferior de uma amostra.

Seja $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ a amostra ordenada. Então define-se a estatística de teste por:

$$T_{N6} = \frac{x_{(n)} - x_{(1)}}{S}$$

onde

$$S = \left[\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1} \right]^{1/2}.$$

[1] A notação que se segue é utilizada mais tarde na Tabela 1, para representar os diferentes testes que se obtêm ao empregarem-se diferentes níveis de significância.

Rejeita-se a hipótese nula quando T_{N6} assume valores superiores aos valores críticos, fixado um certo nível de significância α. Esses valores críticos podem ser encontrados em Barnett e Lewis (1984), Tabela XIIa, página 386.

c) Máximo Resíduo Studentizado - X39, X40 e X41. (Teste N2 em Barnett e Lewis (1984) pág. 168). Teste bilateral para a detecção de um único *outlier* em uma amostra.

A estatística de teste é dada por

$$T_{N2} = \max\left(\frac{x_{(n)} - \overline{x}}{S}, \frac{\overline{x} - x_{(1)}}{S}\right)$$

onde

$$S = \left(\frac{\sum(x_i - \overline{x})^2}{n}\right)^{1/2}$$

A hipótese nula é rejeitada quando T_{N_2} toma valores maiores que um certo k crítico. Estes valores podem ser encontrados em Barnett e Lewis (1984), Tabela VIIIb, página 377.

d) *Huber Skipped Mean* - X84, X42, X43, X44 e X45.

Seja

$$\psi(x) = \begin{cases} x & se \quad |x| \leq k \\ 0, & c.c. \end{cases}$$

k é o valor crítico.

Seja $t = \text{med}_i\, x_i$, $s = 1,483\,\text{MAD}(x_i)$ e m o número de observações em $I = (t - ks; t + ks)$.

O estimador é dado por

$$t + s\sum \psi\left[(x_i - t)/s\right]/m.$$

Então a versão *um passo*, para um dado k crítico, é simplesmente a média aritmética de todas as observações em I. Observe que a expressão *um passo* empregada aqui não significa a rejeição de no máximo um *outlier* e sim a aplicação de apenas um passo do algoritmo para obtenção do M-estimador.

X46 é uma versão iterativa do estimador *Huber skipped mean*, dado pela solução T de

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T}{s}\right) = 0$$

onde

$$\psi(x) = \begin{cases} x, & |x| \leq k \\ 0, & cc \end{cases}$$

e k é um valor crítico.

X47 , também iterativo, é dado pela solução do sistema de equações

$$\begin{cases} \sum \psi[(x_i - T)/S] = 0 \\ \sum \psi^2[(x_i - T)/S] = n\beta \end{cases}$$

onde $\beta = E_\phi \psi^2$ e ψ é a mesma empregada em X46.

Os valores críticos adequados aos estimadores *Huber skip* podem ser obtidos em Schweingruber (1980).

e) Regra de Dixon - X50 e X51.

Teste bilateral baseado em razões de diferenças entre observações. Pode ser usado em casos em que se deseja evitar o cálculo de s. Para esse teste a estatística muda com o tamanho da amostra. Para $n = 20$ utiliza-se

$$T = \max\left\{\frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}};\ \frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}}\right\}$$

onde $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ é a amostra ordenada.

Os valores críticos para este teste podem ser encontrados em Grubbs (1969), tabela 2, página 7.

f) Estatística de Shapiro-Wilk (Teste N17 em Barnett e Lewis (1984) pág. 177).

É um teste bilateral que tem por objetivo detectar quaisquer afastamentos da hipótese nula, isto é, normalidade dos dados. Rejeita H_0 para valores pequenos da estatística de teste, W, onde

$$W - \text{ Estatística de Shapiro-Wilk } =$$

$$\left[\sum_{i=1}^{[n/2]} a_{n,n-i+1}(x_{(n-i+1)} - x_{(i)})\right]^2 / S^2$$

onde $[\frac{n}{2}]$ denota a parte inteira de $\frac{n}{2}$, $a_{n,j}$ são constantes tabeladas em Shapiro-Wilk (1965) e $S^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

Em Shapiro-Wilk (1965), mostra-se que, sob normalidade, a menos de uma constante, o numerador e denominador de W estão estimando σ^2. Para afastamentos da normal, em geral, esses estimadores não estimam a mesma coisa. Os valores críticos empregados neste teste podem ser encontrados em Barnett e Lewis (1984), tabela XVIIa, pág. 394.

2.3 Os Pontos de Ruptura dos Estimadores

2.3.1 - Conceito

O ponto de ruptura δ^* de um estimador é a menor fração de contaminação de dados capaz de fazer o estimador assumir valores arbitrariamente distantes do verdadeiro valor do parâmetro.

É importante saber quanta contaminação o estimador suporta, mantendo sua capacidade de fornecer alguma informação relevante. Por exemplo, suponha que o verdadeiro valor do parâmetro θ, sob a distribuição F, é 3. Considere todas as distribuições G, a uma certa distância de F. Evidentemente, o estimador torna-se muito precário quando as distribuições neste conjunto podem fornecer estimativas arbitrariamente distantes de 3. Hampel (1968, 1971) definiu formalmente o ponto de ruptura de um estimador, generalizando a idéia de Hodges (1967).

Recentemente, com o crescente interesse sobre o tema, surgiram definições de ponto de ruptura para amostras finitas.

Apresenta-se aqui a definição extraída de Donoho e Huber (1983). Mas outras podem ser encontradas em Huber (1984) e Hampel et alii (1986).

– Seja $X = (x_1, \ldots, x_n)$ uma amostra de n elementos.

– Seja $T = \{T_n\}$ $n = 1, 2, \ldots$ um estimador que toma valores em um espaço euclideano e $T_n(X)$ o valor de T_n quando aplicado à amostra X.

– Seja ε a proporção de dados contaminados da amostra (por substituição, adição, modificação, etc.).

– Seja $b(\varepsilon; X, T_n) = \sup |T(X') - T(X)|$ o maior vício que pode ser causado pela contaminação ε da amostra. O supremo é obtido considerando-se o conjunto de todas as amostras X', ε-contaminadas.

Então, define-se o ponto de ruptura de T_n como,

$$\delta^* = \varepsilon^*(X, T_n) = \inf\{\varepsilon : b(\varepsilon; X, T_n) = \infty\}.$$

Em Hampel et alii (1986) o ponto de ruptura é definido como sendo a *maior* contaminação possível *antes* de *quebrar* o estimador, ou seja

$$\left[\varepsilon^*(X, T_n) - \frac{1}{n}\right]$$

geralmente, o ponto de ruptura não depende da amostra $(x_1, \ldots, x_n)$ e apenas ligeiramente de seu tamanho n.

Um exemplo bastante interessante a respeito do uso do ponto de ruptura foi extraído de Donoho e Huber (1983) e apresentado em seguida.

O estudo de Princeton (Andrews et alii (1972)) mostra que pode haver bastante diferença entre um ponto de ruptura de 25% ou 50%. Acidentalmente, o estudo incluiu dois M-estimadores para média populacional, $D15$ e $P15$, cujas propriedades assintóticas coincidiam para todas as distribuições simétricas. No entanto $P15$ teve um desempenho

indiscutivelmente superior a $D15$, embora só diferissem a respeito do estimador de escala ($\frac{Q_3-Q_1}{2}$ para $D15$, com ponto de ruptura de 25%, e a mediana dos desvios absolutos para $P15$, com ponto de ruptura de 50%). Obviamente, o estimador de escala com o ponto de ruptura mais alto se comportou melhor diante de contaminações assimétricas, do tipo que ocorrem em amostras finitas de distribuições de caudas longas.

Sob circunstâncias normais, $\delta^* \leq 50\%$, já que, com mais de 50% de contaminação dos dados, é difícil saber quem é *outlier* ou não, sem nenhuma informação adicional.

Na verdade, existem muitos estimadores robustos, como a mediana, por exemplo, que possuem $\delta^* = 50\%$. Por outro lado a média aritmética com $\delta^* = 0$ perde completamente suas qualidades com apenas um ponto aberrante. Quanto aos estimadores do tipo *rejeita+média*, para que eles não quebrem é necessário que todos os *outliers* distantes sejam rejeitados antes do cálculo da média. Hampel diz que geralmente a identificação/rejeição de *outliers* depende da capacidade da regra de detectar o primeiro ponto. A partir daí a rejeição se tornaria sucessivamente mais fácil.

2.3.2 - Cálculo

Apresentam-se a seguir os cálculos dos pontos de ruptura para os seis tipos de regra empregados neste estudo.

1) Quarto Momento ou Curtose

Considere-se o caso menos favorável de contaminação para calcular o ponto de ruptura da curtose.

$$(1-\delta) \text{ ——— } x \text{ ——— } \delta$$

Seja $(1-\delta) \geq \frac{1}{2}$ uma fração de dados *bons*, isto é, normais, em torno de zero e um ponto de massa δ de *outliers* a uma distância x.

- Cálculo da média:

$$(1-\delta)0 + x\delta = x\delta$$

- Cálculo da variância:

$$(1-\delta)(0-x\delta)^2 + (x-x\delta)^2\delta = (1-\delta)(x\delta)^2 + (1-\delta)^2x^2\delta$$
$$= (1-\delta)x^2\delta[\delta + (1-\delta)] = (1-\delta)x^2\delta$$

- Cálculo do quarto momento:

$$(1-\delta)x^4\delta^4 + (x - x\delta)^4\delta = (1-\delta)x^4\delta^4 + (1-\delta)^4x^4\delta$$
$$= (1-\delta)x^4\delta[(1-\delta)^3 + \delta^3] = (1-\delta)\delta x^4[1 - \delta^3 + 3\delta^2 - 3\delta + \delta^3]$$
$$= (1-\delta)\delta x^4[1 + 3\delta^2 - 3\delta]$$

- Cálculo da estatística de teste:

$$T = \frac{\delta(1-\delta)x^4(1 - 3\delta + 3\delta^2)}{\delta^2(1-\delta)^2x^4}$$
$$= \frac{1 - 3\delta + 3\delta^2}{\delta(1-\delta)} = \frac{1}{\delta(1-\delta)} - \frac{3}{(1-\delta)} + \frac{3\delta}{(1-\delta)}$$
$$= \frac{1}{\delta(1-\delta)} - \frac{3(1-\delta)}{(1-\delta)} = \frac{1}{\delta(1-\delta)} - 3$$
$$\frac{\partial T}{\partial \delta} = \frac{2\delta - 1}{\delta^2(1-\delta)^2}$$

Para $0 \leq \delta \leq \frac{1}{2}$ $\frac{\partial T}{\partial \delta} < 0$ e T decresce em δ. Se $T < k$ os *outliers* não podem ser rejeitados

$$T < k \Leftrightarrow \delta > \delta^*.$$

O intervalo $(0, \delta^*)$ representa a região de rejeição, mostrando que δ^* é a maior fração de contaminação que o estimador suporta. A partir desse ponto como $T < k$ a regra não é capaz de detectar nenhum *outlier* (supondo uma situação parecida com a pior configuração possível, isto é, um grupo de *outliers* bem próximos, localizado a uma distância razoável das observações consideradas *boas*). Então o ponto de ruptura da curtose é tal que

$$\frac{1}{\delta^*(1-\delta^*)} - 3 = k \ ,$$

onde k é o valor crítico do teste de discordância.

Por exemplo: o valor crítico k correspondente ao nível de significância $\alpha = 1\%$ é aproximadamente igual a 5, portanto $\delta^* \cong 14,7\%$. Isto significa que no máximo 14,7% de *outliers* perto uns dos outros e a uma distância suficiente seriam detectados. Uma proporção maior de contaminação desse tipo ficaria mascarada. Hampel (1985) chama

atenção para o fato de o conceito de ponto de ruptura esclarecer completamente o fenômeno de mascaramento citato na literatura.

2) Máximo Resíduo Studentizado

Considere-se a mesma configuração utilizada anteriormente.

- cálculo da média: $(1-\delta)0+\delta x = x\delta$

- cálculo de resíduo: $(0-x\delta)$ e $(x-x\delta)$

- variância $= \delta(1-\delta)x^2$.

A estatística de teste pode ser obtida fazendo-se

$$T = \frac{x(1-\delta)}{[\delta(1-\delta)]^{1/2}x} = \left[\frac{(1-\delta)}{\delta}\right]^{1/2}$$

Como T é uma função decrescente em δ para $0 \leq \delta \leq \frac{1}{2}$, tem-se $T > k \Leftrightarrow \delta < \delta^*$. Na prática, a estatística T utilizada emprega no cálculo da variância o denominador $n-1$ e não n. Dessa forma o valor crítico é dado por

$$k = \left[\frac{(n-1)(1-\delta^*)}{n\delta^*}\right]^{1/2} \Rightarrow$$

$$\Rightarrow k^2 = \frac{(n-1)(1-\delta^*)}{n\delta^*}$$

$$\Rightarrow \frac{nk^2}{(n-1)} = \frac{(1-\delta^*)}{\delta^*}$$

$$\Rightarrow \frac{nk^2}{(n-1)} = \frac{1}{\delta^*} - 1.$$

Enquanto o ponto de ruptura é dado por

$$\delta^* = \left[\frac{nk^2}{n-1} + 1\right]^{-1}.$$

Como

$$\delta^* \leq \frac{1}{2} \Rightarrow k \geq \sqrt{\frac{n-1}{n}} \left(\frac{nk^2}{n-1} + 1 \right)^{-1} \leq \frac{1}{2}$$

$$\Rightarrow \left(\frac{nk^2}{n-1} + \frac{n-1}{n-1} \right)^{-1} \leq \frac{1}{2}$$

$$\Rightarrow \frac{n-1}{nk^2+n-1} \leq \frac{1}{2}$$

$$\Rightarrow 2(n-1) \leq nk^2 + n - 1$$

$$\Rightarrow 2n - 2 - n + 1 \leq nk^2$$

$$\Rightarrow \frac{n-1}{n} \leq k^2$$

$$\Rightarrow k \geq \sqrt{\frac{n-1}{n}}$$

Para $k = 3$, normalmente usado nos testes de discrepância, δ^* é aproximadamente 10%.

3) Amplitude Studentizada

Considerando-se, mais uma vez, a distribuição menos favorável

$(1-\delta)$ ——— x ——— δ

obtém-se

$$T = \frac{x}{\left[\frac{\delta(1-\delta)n}{n-1}\right]^{1/2} x} = \left[\frac{n-1}{\delta(1-\delta)n}\right]^{1/2}$$

$$\ell n(T) = \frac{1}{2}(\ell n(n-1) - \ell n(\delta) - \ell n(1-\delta) - \ell n(n))$$

$$\frac{\partial}{\partial \delta} \ell n(T) = -\frac{1}{2\delta} + \frac{1}{2(1-\delta)} = \frac{-(1-\delta)+\delta}{2\delta(1-\delta)}$$

$$= \frac{2\delta - 1}{2\delta(1-\delta)} \leq 0, \quad \delta \in [0, 1/2]$$

Como T é decrescente em δ no intervalo $[0, 1/2]$, $T > k \Leftrightarrow \delta < \delta^*$, então

$$k = \left[\frac{n-1}{\delta^*(1-\delta^*)n}\right]^{1/2}$$

onde δ^* é o ponto de ruptura. Resolvendo a equação em δ^* obtém-se

$$\delta^* = \frac{1}{2} - \left[\frac{1}{4} - \frac{(n-1)}{nk^2}\right]^{1/2}$$

Como $\delta^* \leq \frac{1}{2} \Rightarrow k \geq 2\sqrt{\frac{n-1}{n}}$.

Geralmente utiliza-se $k = 5$ como valor crítico, o que resulta em $\delta^* \cong 5\%$.

4) Regra de Dixon

Quando $n = 20$, esta regra está baseada na estatística

$$T = \max\left\{\frac{x^{(3)} - x^{(1)}}{x^{(n-2)} - x^{(1)}}; \frac{x^{(n)} - x^{(n-2)}}{x^{(n)} - x^{(3)}}\right\}.$$

Se $T > k$, rejeita-se $x^{(1)}$ ou $x^{(n)}$ dependendo de qual fração é maior. Esta regra tolera até 2 *outliers* do mesmo lado da distribuição, quando $n = 20$, e quebra com 3, o que significa $10\% < \delta^* < 15\%$.

A fim de obter um valor mais exato para δ^*, pode-se dividir a observação $x^{(3)}$ em duas partes. A primeira com peso α junto a $x^{(1)}$ e $x^{(2)}$ e a outra com peso $(1-\alpha)$ junto às demais observações. Neste caso tem-se

$$T = \frac{(1-\alpha)x}{x} = 1 - \alpha$$

A região de rejeição é da forma

$$T \geq k \Leftrightarrow 1 - \alpha \geq k \Leftrightarrow \alpha \leq 1 - k$$

Portanto o maior valor de α para o qual o teste é capaz de rejeitar a hipótese nula é $\alpha^* = 1 - k$. Desta forma calcula-se o ponto de ruptura para $n = 20$ fazendo:

$$\delta^* = \frac{2 + \alpha^*}{20} = \frac{2 + 1 - k}{20}$$

Os valores usuais para o valor crítico são $k = 0,535$ e $k = 0,45$ produzindo $\delta^* = 12,3\%$ e $\delta^* = 12,7\%$, respectivamente.

5) Huber skipped mean

Suponha-se que os dados estão dispostos em dois grupos,conforme foram apresentados anteriormente e que a contaminação δ é menor ou igual a 50%. É fácil ver que a mediana t não é afetada pela contaminação δ, por conseguinte, $S =$ mediana $|x_i - t|/0,6745$ também não, limitando $t \pm kS$ à região em torno dos *bons* dados. Para que esse estimador fosse afetado pela contaminação, seria necessário um percentual superior a 50% de dados *ruins*. Assim, o ponto de ruptura desses estimadores (*one step*) é de 50%. As versões iterativas ($X46$ e $X47$) ainda não estão com seus respectivos pontos de ruptura calculados. $X46$ aparentemente possui $\delta^* = \frac{1}{2}$, enquanto o de $X47$ ainda é desconhecido.

6) Estatística de Shapiro-Wilk

Segundo Hampel, a pior configuração, visando o cálculo do ponto de ruptura, pode ser obtida criando-se um ponto de massa de tamanho $\frac{n-m}{n}$ em zero (onde n é o tamanho da amostra e $m \leq \frac{n}{2}$, o número de observações contaminadas) e distribuindo-se, simetricamente, os m pontos contaminados em torno dos a_i. Essa configuração maximiza W, que rejeita H_0 para valores pequenos, entre todas as configurações que contêm um ponto de massa de tamanho $(n-m)/n$. Hampel afirma que, para $n = 20$ e $k \cong 0.9$, obtém-se $\delta^* \cong 45\%$, um excelente resultado, que só é superado pelos estimadores do tipo *Huber-skip* com escala mais robusta. Para obter-se mais eficiência sob normalidade, basta ser mais rigoroso com k. Com $k = 0.5$ o ponto de ruptura cai para 35%.

2.4 Conclusões

A seguir apresenta-se uma análise dos procedimentos do tipo *rejeita+média*, segundo suas respectivas variâncias Monte Carlo, sob as várias distribuições mencionadas em 2.2 e ainda a comparação destes com dois estimadores robustos, $H15$ e $25A$, que possuem aproximadamente a mesma variância sob normalidade. Assim é possível considerar a eficiência dos estimadores sob normalidade e as mudanças que sofrem sob contaminações em torno da normal. O comportamento dos estimadores será descrito para cada distribuição separadamente. A tabela 1 a seguir, contém o valor da variância Monte Carlo, multiplicado por vinte, de vinte e um procedimentos do tipo *rejeita+média* versão iterativa, sob as sete distribuições descritas anteriormente. Apenas para *Huber skip* são apresentados os resultados da aplicação da versão *(um passo)*. A última coluna fornece o valor do ponto de ruptura.

Tabela 1 - Variâncias Monte Carlo e Pontos de Rupturas

Estimador Código	k	x	Variâncias segundo as distribuições selecionadas N	5%3N	5%10N	10%10N	T3	25%10N	Cauchy	Pontos de ruptura δ^*
Estimadores para comparação										
50%			1.498	1.52	1.56	1.80	1.82	2.5	2.9	50%
H15			1.036	1.16	1.21	1.50	1.71	4.0	5.7	26%
25A			1.046	1.16	1.13	1.26	1.67	2.1	3.7	50%
Quarto-Momento										
X24	5	1%	1.007	1.20	1.14	1.34	1.023	11.4	7.2	14.7%
X25	4	5%	1.021	1.18	1.13	1.28	1.97	3.8	7.2	14.7%
X26	3.5	10%	1.065	1.17	1.16	1.26	1.93	2.7	4.7	19%
X27	3	50%	1.136	1.24	1.22	1.30	1.90	1.9	4.2	21%
Amplitude Studentizado										
X34	4.79	1%	1.004	1.33	4.23	9.93	2.78	28.0	1.000	4.3%
X35	4.49	5%	1.014	1.23	1.20	5.80	2.27	25.2	1.000	4.9%
X36	4.32	10%	1.038	1.19	1.16	4.76	2.21	22.6	28.9	5.4%
Maximo Resíduo Studentizado										
X39	3.03	1%	1.003	1.20	1.14	2.39	2.05	17.9	10.5	9.4%
X40	2.71	5%	1.020	1.17	1.12	1.30	1.98	11.4	7.5	11.5%
X41	2.56	10%	1.046	1.16	1.13	1.28	1.95	7.6	6.0	12.7%
Huber-Skipped Mean										
X84	3.50	10%	1.044	1.20	1.14	1.31	1.85	2.4	4.2	50%
X42	3.03	20%	1.088	1.20	1.16	1.29	1.80	2.1	3.8	50%
X43	2.71	30%	1.137	1.23	1.18	1.30	1.74	1.9	3.5	50%
X44	2.35	50%	1.245	1.29	1.24	1.32	1.75	1.8	3.3	50%
X45	2.00	65%	1.391	1.41	1.35	1.39	1.80	1.7	3.1	50%
X46	2.71	I	1.136	1.23	1.19	1.31	1.78	1.9	3.7	50%(?)
X47	2.71	IP	1.341	1.39	1.37	1.44	1.84	1.8	3.0	?
Regra de Dixon										
X50	.535	2%	1.009	1.21	1.15	1.35	2.09	16.5	9.7	(12.3%)
X51	.450	10%	1.041	1.20	1.15	1.29	2.05	13.3	7.7	(12.7%)
Estatística de Shapiro-Wilk										
X54	.905	5%	1.074	1.27	1.15	1.33	1.85	2.0	4.0	45.2%
X55	.920	10%	1.270	1.30	1.24	1.40	1.84	2.1	3.5	48%

Fonte: Hampel (1985)

1) Normal

Observe-se que, a medida que aumenta o nível de significância dos testes, aumenta também a perda de eficiência sob normalidade. Anscombe (1960) chamou essa perda de *prêmio de seguro*, fazendo uma analogia com o que deve ser pago para que se possa usufruir de determinados benefícios, caso ocorra um sinistro (afastamentos da normal). Observando a coluna de δ^*, pode-se notar que a tal perda está associado um ganho de robustez, já que cresce o valor do ponto de ruptura. Com exceção do *Huber-skipped mean*, parece que, sob um mesmo nível de significância, os procedimentos mais robustos perdem mais sob normalidade. Veja a tabela a seguir.

Tabela 2: Variâncias Monte Carlo e pontos de ruptura de alguns estimadores do tipo *rejeita+média*, com $\alpha = 10\%$

Estimadores	Variância sob normalidade	$\delta^*(\%)$
X36	1,038	5,4
X51	1,041	12,7
X84	1,044	50,0 $\rightarrow$ (exceção)
X41	1,046	12,7
X26	1,065	19,0
X55	1,270	48,0

Fonte: Hampel (1985).

As maiores perdas sob normalidade estão associadas aos seguintes estimadores: mediana (50% de perda); Huber-skip X45 e X47 ($\cong$ 40%); $X55$, Shapiro- Wilk (27%) e $X44$, Huber-skip com $\alpha = 50\%$ (25% de perda). Os demais estimadores têm um comportamento razoavelmente parecido.

É interessante notar que há diferença entre as variâncias de $X46$ e $X47$ sob as outras distribuições, embora esses estimadores só difiram na estimação da escala. E mais, $X46$ (Huber-skip versão iterativa) tem comportamento similar a $X43$ (Huber-skip - one step) sob todas as distribuições.

Diferentes estimadores do tipo *rejeita+média*, com mesma perda sob normalidade, possuem variâncias Monte Carlo semelhantes enquanto seus pontos de rupturas não são atingidos pela contaminação dos dados.

2) 5%3N

Todas as variâncias estão próximas, com algumas exceções (mediana, X45 e X47). A perda relativa, comparada com $H15$, está entre 1% e 10% (raramente superior a 10%). Isto indica que, nesse caso de contaminação nos flancos ou próxima, com apenas um *outlier* freqüentemente escondido entre os *bons* dados, regras de rejeição são praticamente tão boas quanto os estimadores de Huber, representados aqui pelo $H15$. Observe-se o gráfico 3 do anexo 2 retirado de Hampel (1977). Isto já não acontece para contaminações mais pesadas. Talvez, o caso de um *outlier* moderado, seja igualmente fácil para todos os estimadores razoavelmente robustos.

Outro fato a se observar é o comportamento de X34, a amplitude studentizada de menor ponto de ruptura, $\delta^* = 4,3\%$. Como o *outlier* nunca está tão longe, isto é, 5% 3N está longe da pior configuração, ele ainda pode ser detectado em muitos casos. Mas isto não impede que $X34$ já comece a demonstrar sua fragilidade.

3) 5%10N

Esta distribuição é um caso típico de um único *outlier* distante.

$X34$ demonstra, claramente, sua incapacidade de rejeitar um único ponto distante.

Observe que $X35$ com $\delta^* = 4,9\%$ ainda se comporta bem. Aparentemente, $X35$ não se defrontou com a pior configuração.

4) T3

Esta distribuição pode ser considerada como um exemplo de dados de alta qualidade, sem erros grosseiros. Observe-se que nesta situação não faz sentido rejeitar observações. Mas pode-se ainda estar interessado no comportamento dos procedimentos do tipo *rejeita+média*, no que diz respeito à sua eficiência, e adotá-lo como medida de segurança antes de aplicar a média.

Enquanto as curvas de Huber (12A e 25A) são quase ótimas, os procedimentos do tipo *rejeita+média*, geralmente, têm variância 10% a 15% mais altas. Observe-se mais uma vez

o gráfico 3 no anexo 2.

A variância mais elevada ocorre para os estimadores $X34$, $X35$ e $X36$ que baseiam a rejeição na amplitude studentizada. Tal perda de eficiência é ainda suportável e comparável à perda sofrida pelos estimadores de Huber sob contaminações distantes, da ordem de 10% da amostra. O autor embora não inclua na Tabela 1 dados a respeito dos procedimentos *de um passo*, que rejeitam no máximo uma observação, revela que se surpreendeu com o fato de se comportarem tão bem quanto os métodos iterativos. Alerta ainda a respeito de que, embora os procedimentos do tipo *rejeita+média* sejam bons sob $T3$ (uma distribuição de perfil suave de caudas alongadas), deve-se chamar atenção para os danos que sofrem esses estimadores em casos de contaminações menos suaves e perto dos pontos de rejeição. Todas as regras de rejeição nestes casos tornam-se bastante instáveis. Uma pequena alteração nas observações pode causar grandes mudanças na estimativa final. Se os *outliers* estão concentrados perto dos pontos de rejeição, isto é, a partir do qual se rejeita a observação, os procedimentos do tipo *rejeita+média* podem ser arbitrariamente piores que outros estimadores.

5) 10%10N

O caso de dois *outliers* em vinte observações é um exemplo típico de dados que não passaram por nenhuma crítica anterior.

Previsivelmente os estimadores $X34$, $X35$ e $X36$ são muito afetados por esses *outliers*. $X39$ (baseado no máximo resíduo studentizado, uma das mais populares regras de rejeição) começa a dar sinais de problemas para detectar 2 *outliers* em 20 com seu $\delta^* = 9,4\%$.

Todos os outros estimadores, incluindo a regra de Dixon e os outros resíduos studentizados, estão com suas respectivas variâncias em torno da relativa ao $25A$. A mais afastada, depois das já citadas, é a estatística de Shapiro-Wilk.

6) Cauchy

A distribuição de Cauchy é um exemplo de contaminação pesada. Foi incluída, principalmente, por ser bastante conhecida e usada em outros estudos e não por qualquer propriedade específica.

A tabela abaixo mostra o quanto, percentualmente, cada variância excede a variância de $25A$.

Tabela 3: Perda de eficiência com relação ao 25A

Estimadores ($\alpha = 10\%$)	$[\frac{\text{Variância do estimador}}{\text{Variância do } 25A} - 1] \times 100$	$\delta^*(\%)$
Huber-skip	13,5	50,0
Quarto momento	27,0	19,0
Resíduo Studentizado	62,0	12,7
Regra de Dixon	108,0	12,7
Amplitude Studentizada	681,0	5,4

Fonte: Hampel (1985)

A última coluna, contendo o ponto de ruptura, mostra como a piora do estimador é acompanhada do decréscimo do seu δ^*. A exceção é Shapiro-Wilk $\alpha = 10\%$, que se comporta melhor que o $25A$ e tem $\delta^* = 48\%$. Parece que estimadores capazes de rejeitar 10% da amostra podem impedir o descontrole da variância sob a distribuição de Cauchy, para $n = 20$. Ressalte-se que todos os procedimentos *de um passo* que podem rejeitar um ponto, incluindo o melhor na classe dos baseados na amplitude studentizada $X36$ (e ainda *X33 um passo*), têm variância Monte Carlo em torno de 30, enquanto os outros, baseados na amplitude e com $\delta^* < 5\%$, têm variância Monte Carlo *infinita*.

7) 25%10N

Esta distribuição contém 5 *outliers* em 20 observações e é provavelmente um caso raro. Entretanto pode, eventualmente, ocorrer com dados *sujos* ou em casos em que uma certa medição é difícil de ser executada.

O grupo dos estimadores baseados no quarto momento: $X27$, $X26 - X25$ e $X24$ pode rejeitar seguramente 4,3 e 2 *outliers*, respectivamente.

Os cinco estimadores: $X24, X40, X41, X50, X51$ que podem rejeitar 2, mas não 3 *outliers*, têm variância na mesma região, entre aqueles de δ^* alto e os de δ^* baixo.

Finalmente, observe-se que $X27$, aparentemente, não esteve diante do caso menos favorável de cinco *outliers* bem juntos e muito distantes.

Se os *outliers* estão longe dos *bons dados* (como nos casos de 10 desvios padrões), então os procedimentos do tipo *rejeita+média*, que não *quebraram* ainda, são tão bons quanto o melhor M- estimador, sempre comparando a mesma perda sob normalidade.

Os pontos de ruptura fornecem uma surpreendentemente rica e precisa descrição do comportamento dos procedimentos de estimação. Tão logo sejam atingidos pelo percentual de contaminação da amostra, observa-se a *quebra* do procedimento, com reflexo imediato na qualidade da estimativa: tolerável, se os *outliers* estão próximos do resto da massa e desastroso, se estão distantes.

Os pontos de ruptura substituem e explicam completamente o *mascaramento*, um conceito muito empregado na literatura sobre *outliers*. O *mascaramento* ocorre quando os *outliers* estão dispostos de tal forma que é impossível detectar qualquer um deles. Por exemplo: o teste de discordância em que se baseia $X40$ não é capaz de identificar 3 *outliers* em 20, próximos entre si e afastados dos *bons* dados. Neste caso os *outliers* ficam mascarados. Observe-se que o ponto de ruptura explica bem tal efeito: o mascaramento só pode ocorrer, quando a contaminação dos dados for superior ao ponto de ruptura da regra.

Outra questão abordada por Hampel diz respeito à potência dos testes de discordância. Enquanto o ponto de ruptura não é atingido, a potência do teste converge a 1, quando os *outliers* vão para infinito. Se o ponto de ruptura é ultrapassado, isto não ocorre necessariamente.

Embora o ponto de ruptura seja obtido para os casos menos favoráveis de cada regra, o estudo Monte Carlo mostrou que, sob as contaminações escolhidas (as quais, conforme, mencionado em 2.2, ilustram diferentes situações encontradas na prática), a variância do estimador explode, quando o ponto de ruptura é ultrapassado.

CAPÍTULO 3

ESTIMADORES *DRLS*
– ESTUDO MONTE CARLO PARA O MODELO DE REGRESSÃO –

Generalizando para o modelo de regressão linear o tipo de análise apresentado no capítulo anterior, foram selecionadas quatro técnicas de diagnóstico em regressão: $DFFITS_i$, LD_i, $COVRATIO_i$ e *cápsula normal*. Aplicados em duas versões: iterativa e em *um passo* para detecção de outliers e combinados com *LS*, estes diagnósticos produzem estimadores representados daqui em diante por *DRLS - Detecção+Rejeição+LS*. Tais estimadores são avaliados através de um estudo Monte Carlo e comparados entre si. Consideram-se ainda os τ-estimadores propostos por Yohai e Zamar (1986). Estes estimadores combinam eficiência sob normalidade e alto ponto de ruptura.

Na seção 3.1 apresentam-se as propriedades e limitações dos estimadores de mínimos quadrados. A seção 3.2 contém uma descrição detalhada dos estimadores selecionados, enquanto em 3.3 encontram-se as informações a respeito do estudo Monte Carlo. A apresentação dos resultados e a análise são deixadas para o próximo capítulo.

3.1 Estimadores de Mínimos Quadrados e suas Limitações

Considere-se o modelo de regressão linear múltipla:

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{3.1}$$

onde: $\underset{\sim}{Y}$ é o vetor $n \times 1$ de valores da variável resposta (ou dependente);

$X = [x_{ij}]$ $\quad 1 \leq i \leq n$ e $1 \leq j \leq p$ é uma matriz $n \times p$, de posto p, de preditores (fatores, regressores ou variáveis explicativas) incluindo possivelmente um preditor constante;

$\underset{\sim}{x}_i = (x_{i1}, \ldots, x_{ip-1}, x_{ip})'$ é a i- ésima linha da matriz X, escrita como vetor coluna de dimensão $p \times 1$;

$\underset{\sim}{\beta}$ é um vetor $p \times 1$ de coeficientes desconhecidos (parâmetros do modelo);

$\underset{\sim}{\varepsilon}$ é um vetor $n \times 1$ de variáveis independentes com média zero e variância desconhecida σ^2.

O estimador de mínimos quadrados de $\underset{\sim}{\beta}$ é a estatística que minimiza a função

$$\Gamma(\underset{\sim}{\beta}) = \sum_{i=1}^{n} (Y_i - \underset{\sim}{x}_i' \underset{\sim}{\beta})^2 = (\underset{\sim}{Y} - X\underset{\sim}{\beta})'(\underset{\sim}{Y} - X\underset{\sim}{\beta})$$

Este estimador é ótimo sob as condições do teorema de Gauss-Markov (Scheffé, 1959, p. 14).

Teorema de Gauss-Markov. *Sob as hipóteses*

1) $E\varepsilon_i = 0, \quad i = 1, \ldots, n$

2) $\mathrm{Cov}(\varepsilon_1, \ldots, \varepsilon_n) = \sigma^2 I$

toda função estimável $b' \underset{\sim}{\beta}$ *tem um único estimador linear de variância mínima entre todos os estimadores lineares não viciados. Ele é dado por* $b' \underset{\sim}{\hat{\beta}}$, *onde* $\underset{\sim}{\hat{\beta}}$ *é o estimador de mínimos quadrados. Se os erros são normalmente distribuídos, este estimador tem variância mínima entre todos os estimadores não viciados.*

Hampel et alii (1986) fazem algumas observações descritas a seguir:

1) Linearidade é uma restrição muito forte: muitos estimadores de máxima verossimilhança (por ex., considerando-se erros com distribuição logística, t-student ou Cauchy) não são lineares;

2) Rejeição de *outliers* é uma operação não linear;

3) O estimador de mínimos quadrados é ótimo na classe de todos os não viciados somente se os erros são normalmente distribuídos. Portanto a restrição a estimadores lineares pode ser justificada apenas por normalidade ou simplicidade;

4) O modelo normal nunca é exatamente verdadeiro e os procedimentos (estimadores e testes) de mínimos quadrados perdem eficiência, drasticamente, diante de pequenos afastamentos da hipótese de normalidade dos erros (Huber, 1973 e 1977; Hampel, 1973, 1978a e 1980; Schrader e Hettmansperger, 1980; Ronchetti, 1982a e 1982b). Assim, dever-se-iam preferir procedimentos somente aproximadamente ótimos sob normalidade, mas ainda *bem comportados* em sua vizinhança.

A grande popularidade alcançada pelo método de mínimos quadrados deveu-se principalmente à sua facilidade de aplicação, numa época em que não existiam computadores. Mesmo hoje, está implementado na maioria dos pacotes estatísticos, por tradição, e, também, por sua velocidade de processamento. No entanto, recentemente, observou-se que, muitas vezes, os dados reais não satisfaziam completamente às hipóteses clássicas, comprometendo a qualidade da análise estatística (p.e. Student, 1927; Pearson, 1931; Box, 1953 e Tukey, 1960).

Como ilustração, observem-se os efeitos de *outliers* no modelo de regressão linear simples $y_i = \beta_1 x_i + \beta_2 + \epsilon_i$, um caso especial de (3.1) com $p = 2$, pois $x_{i1} = x_i$ e $x_{i1} = 1$.

O gráfico 4 no anexo 2, mostra os 5 pontos e a reta ajustada por mínimos quadrados. Observe-se no gráfico 5 o que acontece, quando um ponto é deslocado no eixo dos y. Tal deslocamento pode ser causado por um erro de transcrição, digitação, etc. (consulte-se a seção 1.2 sobre a origem dos erros) e mudar completamente o resultado anterior.

Observe-se agora no gráfico 7 como um único *outlier* pode alterar radicalmente a reta ajustada. Este *outlier*, na direção do x, (x_1, y_1), é chamado de ponto de alavanca (*leverage point*). É como se (x_1, y_1) exercesse uma forte atração sobre a reta ajustada. Basta observar a definição do método de ajuste para entender este fato. O afastamento de x_1 faz com que o resíduo r_1, obtido a partir do ajuste mostrado no gráfico 6, seja muito grande, contribuindo para o aumento de $\sum_{i=1}^{5} r_i^2$. Isso faz com que a reta mostrada no

gráfico 6 não possa ser a escolhida como aquela que minimiza a soma dos quadrados dos desvios. A reta que possui esta propriedade é mostrada no gráfico 7, deixando bem claro o impacto de um *outlier* sobre o estimador LS.

Embora estes exemplos sejam construídos, é importante ter em mente que tal situação se repete, freqüentemente, no dia a dia de um analista de dados. Portanto a preocupação dos usuários de análise de regressão com observações desproporcionalmente influentes reflete-se em esforços para evitar que o ajuste do modelo seja baseado em umas poucas observações, como no gráfico 7, pois, na verdade, deve estar baseado na maioria dos dados.

Os *outliers* parecem não oferecer muito perigo em um problema de dimensão dois, como no exemplo anteriormente apresentado, pois são facilmente identificados com um simples gráfico de resíduos. No entanto, quando se trata de dimensões maiores, a identificação torna-se bem mais complexa. Ressalte-se ainda que é razoável admitir que, com o aumento da dimensão, cresça também a possibilidade de aparecerem *outliers*. Além disso, geralmente, para detectar pontos influentes não é suficiente observar cada variável separadamente ou mesmo todas as combinações duas a duas e respectivos gráficos (Rousseeuw e Leroy, 1987).

A definição de *outlier* no modelo de regressão é mais delicada que no modelo de posição, tratado no Capítulo 2. Não há mais porque os *outliers* se apresentarem como extremos das amostras de X ou Y.

Em geral, diz-se que (X_k, Y_k) é um ponto de alavanca, quando X_k está afastado da maior parte das observações X_i da amostra. Rousseeuw e Leroy (1987) ponderam *que isso não significa um outlier na regressão, pois nessa definição não se está levando em conta o* Y_k.

Quando (X_k, Y_k) está próximo da reta de regressão determinada pela maioria dos dados e estimada por mínimos quadrados, tal ponto pode ser considerado um ponto de alavanca bom, como no gráfico 8. Portanto dizer que (X_k, Y_k) é um ponto de alavanca refere-se somente ao poder que ele tem de afetar os estimadores dos coeficientes da regressão $\underset{\sim}{\hat{\beta}}$ devido ao afastamento de X_k. Mas isto não acontece, necessariamente, pois o ponto pode

estar exatamente em cima da reta de regressão.

Já Velleman (Chatterjee e Hadi, 1986; Discussão, pág. 413) sugere que se considere como influente qualquer ponto extremo o suficiente para afetar a correlação e as estatísticas t ou F, chamando a atenção para as conseqüências de se afirmar que um ponto de alavanca é *bom* ou não, só por este estar em cima da reta ajustada.

Existem muitas técnicas utilizadas para detecção de *outliers* em regressão. Combinadas, muitas vezes, com análises gráficas, elas se baseiam, principalmente, em resíduos de mínimos quadrados, matriz de projeção (matriz chapéu), elipsóides de confiança e medidas de influência parcial.

A aplicação de tais técnicas, isoladamente, e, muitas vezes, de combinações delas, fazem parte de uma fase da análise que deve ocorrer depois de um (primeiro) ajuste por LS. Nesta fase do trabalho, chamada de diagnóstico em regressão, criada para checar as hipóteses assumidas anteriormente, às vezes, é possível corrigir e/ou remover os *outliers*, permitindo a reaplicação de LS, sob condições mais favoráveis.

Quando a massa de dados possui apenas um *outlier*, muitos destes métodos são capazes de detectá-lo. No entanto, quando existem vários, sua identificação é bem mais difícil (podem, p.e., envolver o cálculo de estatísticas para um número muito grande de subconjuntos da amostra).

Chatterjee e Hadi (1986) ressaltam que uma observação pode afetar diferentemente os diversos resultados obtidos em uma regressão e sugerem que o analista tente escolher o instrumento de diagnóstico que vai usar de acordo com o que se deseja saber, tentando, por exemplo, hierarquizar algumas perguntas que devem ser respondidas. Entre elas estão: qual a influência da i-ésima observação no vetor de parâmetros estimados? E na variância estimada de $\hat{\beta}$?...

Outra forma de tratar com *outliers* no modelo de regressão é através da aplicação de técnicas de regressão robusta, baseadas em estimadores que são menos afetados pelos *outliers*. Rousseeuw e Leroy (1987) ressaltam que esta técnica não ignora os *outliers*, como muitos pensam. Pelo contrário, através da análise dos resíduos obtidos por um ajuste robusto é possível identificar *outliers*. Um procedimento robusto tenta acomodar a maioria dos dados. Desta forma os pontos *ruins*, afastados do padrão estabelecido pelos

bons dados, ficam evidentes, pois possuem resíduos grandes em relação ao ajuste robusto. Portanto diagnóstico e regressão robusta têm, na verdade, o mesmo objetivo, sendo apenas aplicados em momentos diferentes. Usando técnicas de diagnósticos, tenta-se, em um primeiro momento, rejeitar todos os *outliers*, para depois ajustar os *bons* dados por *LS*. Já com estimação robusta dos parâmetros, ajusta-se um modelo à maioria dos dados e, então, identificam-se os *outliers* como aqueles pontos associados aos maiores resíduos deste ajuste. A identificação de *outliers* através de um ajuste robusto não será considerada neste trabalho.

Observe-se que, nesta fase do trabalho, há necessidade de se estabelecer critério para rejeição de observações no modelo de regressão linear. Por analogia com o que foi empregado em Hampel (1985), serão aplicadas duas versões dos procedimentos selecionados. Uma corresponde à versão iterativa, onde se rejeita no máximo uma observação de cada vez (a que corresponde à maior estatística superior ao valor crítico), repetindo-se o procedimento até que cesse a rejeição, para depois aplicar *LS* ao que se supõe serem os *bons* dados. Esta versão é denominada daqui em diante por *um de cada vez (one at a time)*. A outra corresponde à versão *um passo* descrita no Capítulo 2, modificada, onde se rejeitam, de uma só vez, todos os pontos considerados *ruins* pela regra de rejeição e aplica-se *LS* ao restante. Esta versão, para o modelo de regressão, é denominada daqui em diante *todos de uma vez (several at a time)*.

Na seção seguinte apresentam-se as definições dos estimadores.

Observação: Depois do esforço para modelar-se um fenômeno, investindo na estimação dos parâmetros e analisando a limitação das técnicas geralmente empregadas, é importante lembrar que por trás de tudo isto está a hipótese de que o modelo escolhido é razoável. Mas, durante a análise, é preciso estar atento aos possíveis sinais de problemas na modelagem emitidos pelos dados e partir para alteração do modelo original. Por exemplo, a existência de *outliers* pode ser uma indicação de que o modelo não é adequado ou que pode ser melhorado com a inclusão de mais um termo ou interação; ou ainda indicar a necessidade de se efetuar uma transformação na variável resposta.

3.2 Os Estimadores Considerados no Estudo Monte Carlo

Apresentam-se a seguir as descrições dos estimadores empregados neste estudo. São considerados três tipos de estimadores:

1) estimador clássico de mínimos quadrados (LS);

2) estimadores $DRLS$ (resultantes da aplicação de mínimos quadrados aos dados originais, seguida de uma fase de diagnóstico para detecção/rejeição de *outliers* e, finalmente, reaplicação de mínimos quadrados aos dados que restaram;

3) τ-estimador.

3.2.1 Estimador clássico - LS

Considerando o modelo

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

dado em (3.1), suponha que $X'X$ seja inversível, então:

. O estimador LS para $\underset{\sim}{\beta}$ é obtido por:

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y}$$

. O vetor de valores ajustados é dados por:

$$\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}} = X(X'X)^{-1}X'\underset{\sim}{Y} = H\underset{\sim}{Y}$$

onde $H = X(X'X)^{-1}X' = [h_{ik}]$ é a matriz de projeção ortogonal no espaço gerado pelas colunas da matriz X. H também é chamada de matriz *Hat* ou *chapéu*, pois $\underset{\sim}{\hat{Y}} = H\underset{\sim}{Y}$.

. O vetor de resíduos do ajuste é $\underset{\sim}{r} = (r_1, \ldots, r_n)'$ obtido através da diferença entre $\underset{\sim}{Y}$ e $\underset{\sim}{\hat{Y}}$, então

$$\underset{\sim}{r} = \underset{\sim}{Y} - H\underset{\sim}{Y} = (I - H)\underset{\sim}{Y}$$

. O estimador de σ^2 é S^2 definido por

$$S^2 = r'r/(n-p)$$

. $\text{Cov}(\underset{\sim}{r}) = \sigma^2(I-H)$

. $\text{Var}(r_i) = \sigma^2(1-h_{ii})$

3.2.2 Notação complementar e algumas definições

. X_{-i} denota a matriz X sem a i-ésima linha.

. $\underset{\sim}{Y}_{-i}$ denota o vetor $\underset{\sim}{Y}$ sem a i-ésima componente

$$\underset{\sim}{Y}_{-i} = X_{-i}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_{-i}$$

onde $\underset{\sim}{\varepsilon}_{-i}$ é o vetor $\underset{\sim}{\varepsilon}$ sem a i-ésima componente.

. $H_{-i} = X_{-i}(X'_{-i}X_{-i})^{-1}X_{-i}$

. $\underset{\sim}{\hat{\beta}}_{-i}$ é o estimador LS para $\underset{\sim}{\beta}$ baseado em $\underset{\sim}{Y}_{-i}$, isto é,

$$\underset{\sim}{\hat{\beta}}_{-i} = (X'_{-i}X_{-i})^{-1}X'_{-i}\underset{\sim}{Y}_{-i}.$$

. r_{-i} vetor de resíduos, baseado em $\underset{\sim}{\hat{\beta}}_{-i}$ e $\underset{\sim}{Y}_{-i}$, isto é,

$$\underset{\sim}{r}_{-i} = \underset{\sim}{Y}_{-i} - X_{-i}\underset{\sim}{\hat{\beta}}_{-i}$$

. S^2_{-i} é o estimador de σ^2 definido por

$$S^2_{-i} = \underset{\sim}{r}'_{-i}\underset{\sim}{r}_{-i}/(n-p-1) \qquad \text{ou}$$

$$S^2_{-i} = \{(n-p)S^2 - [r_i^2/(1-h_{ii})]\}/(n-p-1) \quad (\text{ Myers, 1986 p. 339 })$$

. O resíduo studentizado internamente (Weisberg, 1980 - 2nd 1985 - pág. 113) ou resíduo padronizado (Belsley, Kuh e Welsch, 1980 pág. 19) é definido por

$$t_i = r_i/[S^2(1-h_{ii})]^{1/2}$$

. O resíduo studentizado externamente (Weisberg, 1980 - 2nd 1985 - pág. 116) ou RSTUDENT (Belsley, Kuh e Welsch, 1980 pág. 20) ou *Jackniffed* é definido por

$$t(i) = r_i/[S^2_{-i}(1 - h_{ii})]^{1/2}.$$

3.2.3 Estimadores Robustos

3.2.3.1 Estimadores *DRLS*

Como já foi descrito anteriormente, estes estimadores resultam da aplicação de mínimos quadrados aos dados que passaram pela fase de diagnóstico. Muitos diagnósticos são baseados nos resíduos obtidos a partir do ajuste por mínimos quadrados. Entretanto este método, por sua própria definição, procura evitar grandes resíduos e pode muitas vezes gerar um ajuste ruim para a maioria dos dados. Observem-se os gráficos 6 e 7 do anexo 2. Os resíduos, neste caso, não revelariam o verdadeiro *outlier*, pelo contrário, o ponto de menor resíduo seria justamente o *outlier*.

Outro tipo de diagnóstico muito utilizado mede o impacto que uma observação tem no ajuste, calculando a diferença entre os valores da estatística com e sem a i-ésima observação. É possível generalizar este princípio para detecção de vários *outliers*, tentando medir algo como influência simultânea. Entretanto não fica tão óbvio qual subconjunto das observações deve ser rejeitado. Às vezes, alguns pontos são conjuntamente influentes, mas individualmente, não. Além disto exige um esforço computacionalmente considerável em virtude do grande número de subconjuntos que teriam de ser levados em conta.

Além destas medidas, utilizam-se ainda os elementos da matriz H, na detecção de *outliers*. Sendo $\hat{\underset{\sim}{Y}} = H\underset{\sim}{Y}$, os elementos da diagonal, h_{ii}, são de especial interesse, pois são as derivadas de $\hat{Y}_i$ em relação a Y_i. Portanto medem o efeito da i-ésima observação em sua própria predição. Como H está baseada apenas nas variáveis explicativas, esta medida não é capaz de detectar *outliers* na direção do Y.

A escolha dos diagnósticos que são considerados aqui foi feita levando em conta vários aspectos, entre eles estão os seguintes: popularidade; disponibilidade em pacotes de uso difundido (SAS, BMDP, SPSS, etc.) e/ou alguma facilidade de cálculo através de bibliotecas

como IMSL e NAG (bibliotecas de programas, que foram usadas, sempre que possível, na construção das diversas sub-rotinas empregadas durante o desenvolvimento do trabalho); sensibilidade a *outliers* em X e em Y e alguns aspectos abordados no trabalho de Chatterjee e Hadi (resenha sobre o assunto, com discussão, 1986).

Assim foram selecionados os quatro diagnósticos apresentados a seguir:

1) $DFITS_i$ (Belsley, Kuh e Welsch, 1980).

Originalmente, os autores o chamaram *DIFFIT*. Depois, a introdução da escala o transformou em *DFFITS*. Recentemente, o Professor Welsch tentou transformá-lo em *DFITS*, segundo ele próprio, sem sucesso. Aqui, para simplificar a notação, utiliza-se a última forma.

É definido por

$$\frac{|\underset{\sim}{x}_i'(\underset{\sim}{\hat{\beta}} - \underset{\sim}{\hat{\beta}}_{-i})|}{S_{-i}\sqrt{h_{ii}}} = \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} \frac{|r_i|}{S_{-i}\sqrt{1-h_{ii}}}$$

onde σ é estimado por S_{-i}.

Tal estatística surgiu da padronização do i-ésimo componente de $\underset{\sim}{\hat{Y}} - \underset{\sim}{\hat{Y}}_{-i}$. Portanto $DFITS_i$ mede o impacto na predição, quando a i-ésima observação é retirada da amostra. O ponto crítico utilizado freqüentemente para a identificação de *outliers* é $2(p/n)^{1/2}$. Isto significa que para a versão *um de cada vez (one at a time)* (iterativa) dos estimadores *DRLS* baseados em *DFITS* entre as observações com $DFITS_i > 2(p/n)^{1/2}$ é rejeitada apenas uma, isto é, aquela cujo $DFITS_i$ é maior. Depois, ao restante dos dados, aplica-se o estimador de mínimos quadrados e o processo se repete até que cesse a rejeição. Já para a versão *todos de uma vez* (em um passo) , todos os pontos com $DFITS_i$ superiores ao valor crítico são rejeitados de uma só vez.

Chatterjee e Hadi (1986) afirmam que $DFITS_i$ pode ser calculado da seguinte forma:

$$DFITS_i = |t(i)|\sqrt{h_{ii}/(1-h_{ii})}$$

onde $t(i)$ é o resíduo studentizado externamente (definido em 3.2.2).

O cálculo de $DFITS_i$ para o estudo Monte Carlo é efetuado a partir de uma sub-rotina construída com Fortran IV utilizando essa última fórmula. Maiores detalhes sobre os algoritmos empregados podem ser encontrados em Bustos, Caetano e Franco (1989).

Em Chatterjee e Hadi, 1986 - Discussões, o Professor Cook alerta para o fato de que embora $DFITS_i$ deva ser vista como uma medida de influência simultânea para $\underset{\sim}{\hat{\beta}}$ e $\hat{\sigma}^2$ não é suficientemente sensível a mudanças de escala. Para ressaltar esse fenômeno mostra dois exemplos:

a) Todos os pontos em cima de uma reta com exceção de um. (gráfico 9 - Anexo 2). Neste caso, $DFITS_i$ aponta o ponto B como o mais influente, dando mais importância ao impacto sofrido pela variância estimada, pois $\hat{\sigma}^2$ sem o ponto B passa a valer zero, embora a supressão de A mude consideravelmente os coeficientes do ajuste.

b) Regressão linear simples passando pela origem com um *outlier* em $X = 0$. (gráfico 10 - Anexo 2).

Reescrevendo $DFITS_i$, para facilitar o entendimento da discussão que se segue, obtém-se: (Cook, Pena e Weisberg, 1984)

$$DFITS_i = (n - p - 1)b_i h_{ii}/[(1 - b_i)(1 - h_{ii})]$$

onde $b_i = t_i^2/n - p$ e t_i é o resíduo studentizado internamente definido em 3.2.2.

A observação A é realmente um *outlier*, no entanto não pode influenciar $\underset{\sim}{\hat{\beta}}$ pois $h_A = 0$. O ponto A altera apenas $\hat{\sigma}^2$. $DFITS_A = 0$, embora a estatística tente medir a influência em $\underset{\sim}{\hat{\beta}}$ e $\hat{\sigma}^2$ simultaneamente, não é capaz de identificar A como *outlier*.

Tais exemplos podem ser usados para ajudar a caracterizar situações em que se devem ou não utilizar certos diagnósticos, mas também como motivação para o estudo que se propõe aqui, no qual se tentará inferir, através do estudo Monte Carlo, sobre o comportamento dos estimadores selecionados em situações que, acredita-se, acontecem na prática.

2) Distância entre Log-Verossimilhança - LD_i, (log-likelihood distance), (Cook e Weisberg, 1982)

$$LD_i(\underset{\sim}{\beta}, \sigma^2) = 2[L(\underset{\sim}{\hat{\beta}}, S^2) - L(\underset{\sim}{\hat{\beta}}_{-i}, S^2_{-i})]$$

$$= n \log \left[\frac{(n-1)}{n} \frac{n-p-1}{t^2(i)+n-p-1} \right]$$

$$+ \frac{t^2(i)(n-1)}{(1-h_{ii})(n-p-1)} - 1$$

onde $t(i)$ é o resíduo studentizado externamente.

Tal estatística mede a influência da i-ésima observação, levando em conta o modelo probabilístico adotado. Toma a diferença entre o logaritmo da função de verossimilhança no ponto $(\underset{\sim}{\hat{\beta}}, S^2)$ (substituindo $\underset{\sim}{\beta}$ e σ^2 pelo estimador de máxima verossimilhança), considerando todas as observações e o logaritmo da função de verossimilhança no ponto $\underset{\sim}{\hat{\beta}}_{-i}, S^2_{-i}$ (substituindo $\underset{\sim}{\beta}$ e σ^2 pelo estimador de máxima verossimilhança), retirando a i-ésima observação. O ponto crítico utilizado, freqüentememente, para detecção de *outliers* é baseado na χ^2_p, adotando-se geralmente $\alpha = 0,05$.

O uso de LD_i é sujerido pelo Professor Cook em Chatterjee e Hadi, 1986 - Discussão.

Observa-se que LD_i identifica o ponto A, no exemplo b apresentado anteriormente, como o mais influente.

Uma sub-rotina construída em Fortran IV para o cálculo de LD_i é apresentada em detalhe em Bustos, Caetano e Franco (1989).

3) Razão de covariâncias - *COVRATIO* (Besley, Kuh e Welsch, (1980))

$$\text{COVRATIO}_i = \frac{\det[S^2_{-i}(X'_{-i}X_{-i})^{-1}]}{\det[S^2(X'X)^{-1}]}$$

$$= \left(\frac{S^2_{-i}}{S^2} \right)^p \frac{1}{1-h_{ii}} = \left(\frac{n-p-t_i^2}{n-p-1} \right)^p \frac{1}{1-h_{ii}}$$

onde t_i é o resíduo studentizado internamente.

Esta estística leva em conta a precisão com que se estimam os parâmetros do modelo. *COVRATIO* mede a influência da i-ésima observação na variância de $\underset{\sim}{\hat{\beta}}$, comparando a

matriz de covariância estimada com todos os dados, $S^2(X'X)^{-1}$ e a matriz de covariância estimada que resulta quando a i-ésima coluna é retirada, $S^2_{-i}(X'_{-i}X_{-i})^{-1}$. Utiliza para tanto a razão de seus determinantes. Já que essas duas matrizes diferem somente pela inclusão da i-ésima linha na soma de quadrados e produtos cruzados, valores desta razão próximos de 1 podem indicar que as duas matrizes estão próximas ou que a matriz de covariância é insensível à retirada da linha i. Como ferramenta de diagnóstico há interesse em identificar as observações que produzem valores de *COVRATIO* afastados de 1. Geralmente, investigam-se pontos com a seguinte característica:

$$|COVRATIO_i - 1| > \frac{3p}{n}.$$

A sub-rotina que calcula COVRATIO $_i$ foi desenvolvida em FORTRAN IV e é apresentada com detalhes em Bustos, Caetano e Franco (1989).

4) Cápsulas normais

Em algumas ocasiões, a identificação de incompatibilidades entre modelo escolhido e os dados pode ser feita através de gráficos de diferentes tipos de resíduos do ajuste. Eles podem indicar, por exemplo, a necessidade de inclusão de um termo quadrático ou ainda a não homogeneidade da variância.

Aqui apresenta-se um desses gráficos, a cápsula normal, com o objetivo de identificar *outliers*. Ele está baseado em $t(i)$, definido na seção 3.2.2. onde

$$t(i) = \frac{r_i}{S_{-i}\sqrt{1-h_{ii}}}.$$

Como S_{-i} não depende de r_i Cook e Weisberg (1982, p. 20) chamaram $t(i)$ de resíduo studentizado externamente.

Observe-se que $t(i)$ tem distribuição t-student $(n-p-1)$ (Atkinson, 1985) e pode ser aproximado por uma normal, para gerar um gráfico similar a um $Q-Q$ *plot*. Esses gráficos tornaram possível verificar se os valores de $t(i)$ podem ser oriundos de uma amostra aleatória de uma distribuição normal. Neste caso, o gráfico deveria ser uma linha reta, fora flutuações resultantes da amostragem e leves distorções resultantes de considerar uma distribuição t-student como normal.

No entanto é difícil saber, sem estabelecer um referencial, se os gráficos estão suficientemente retos ou se as pequenas irregularidades existentes são causadas por algum agente ou são apenas flutuações aleatórias.

Uma forma de estabelecer padrões poderia ser implementada através do uso contínuo da técnica por parte do analista de dados que, ao final de um *certo* período, estaria *apto*, com algum risco, a reconhecer formas razoáveis para os gráficos dos $t(i)$.

Com o objetivo de contornar tal problema, utilizou-se a simulação para criar uma espécie de região razoável de flutuação, chamada de cápsula. Ressalte-se que, embora esses gráficos possam ser utilizados para detectar outros tipos de afastamento da normalidade, são aplicados aqui com o objetivo de identificar e rejeitar *outliers*.

A construção da cápsula consiste em:

1) ajustar o modelo por mínimos quadrados e calcular os $t(i)$, utilizando a matriz X, fixa, de variáveis explicativas e uma amostra gerada a partir de uma distribuição $N(0,1)$.
2) ordenar os $t(i)$ $\quad i = 1, \ldots, n$ gerados no item anterior, a fim de obter estimativas das estatísticas de ordem dos resíduos. O vetor ordenado é denotado por

$$\underset{\sim}{r}^*(R) = (r_1^*(R), \ldots, r_n^*(R))$$

3) repetir este procedimento um número fixo de vezes, NR, freqüentemente $NR = 19$, e calcular, para cada $i = 1, \ldots, n$

$$u^*(i) = \max(r_i^*(R):\ R = 1, \ldots, NR),$$
$$\ell^*(i) = \min(r_i^*(R):\ R = 1, \ldots, , NR)$$

evidentemente:

$$u^*(1) \leq \cdots \leq u^*(n)$$
$$\ell^*(1) \leq \cdots \leq \ell^*(n) \qquad e$$

$$u^*(i) \geq \ell^*(i), \quad \forall i = 1, \ldots, n$$

Os vetores

$$\ell^* = (\ell^*(1), \ldots, \ell^*(n)) \qquad e \qquad u^* = (u^*(1), \ldots, u^*(n))$$

são chamados, respectivamente de piso e teto da cápsula normal.

A simulação é repetida 19 vezes[1] para dar uma *chance* em vinte de que o maior $t(i)$ fique acima da cápsula (Atkinson, 1985 pág. 36). Maiores detalhes a respeito da construção da cápsula, aplicações e uma sub-rotina de cálculo podem ser encontrados em Bustos e Orgambide, (1988).

A aplicação da cápsula normal como técnica de diagnóstico consiste em identificar, entre os pontos que estão fora da cápsula, aquele que está mais afastado e rejeitá-lo. A seguir, aplicar mínimos quadrados ao restante dos dados e repetir o processo até que cesse a rejeição. A versão *todos de uma vez (several at a time)* tem por objetivo rejeitar todos os pontos que estiverem fora da cápsula de uma só vez e, então, obter o estimador LS com os dados que ficaram,enquanto a versão *um de cada vez (one at a time)* rejeita apenas um ponto de cada vez, o mais afastado.

É importante notar que todos estes diagnósticos são caracterizados por alguma função dos resíduos do ajuste por mínimos quadrados, pelos elementos da diagonal da matriz, e/ou elementos da matriz *catcher*.[2] Os resíduos são empregados para avaliar o ajuste, e as matrizes, para investigar a existência de algum *outlier* no espaço das variáveis explicativas. Infelizmente, quando o conjunto de dados analisados possui mais de um *outlier*, tais medidas já não são seguras, pois estão sujeitas ao fenômeno de mascaramento. Esse efeito é causado pela ação conjunta de dois ou mais *outliers*. Isto poderia ser explicado através dos baixos pontos de ruptura das técnicas de diagnóstico, em analogia com o que foi apresentado no Capítulo 2.

3.2.3.2 τ-estimador

Antes de definir o τ-estimador apresenta-se um breve histórico sobre a evolução da estimação robusta dos parâmetros da regressão linear.

Suponha-se que $(y_1, y_2, \ldots, y_n)$ sejam os valores observados do vetor $\underset{\sim}{Y}$. Conforme

[1] Neste trabalho utilizou-se $NR = 99$, de tal forma a obter-se uma menor probabilidade de rejeição, sob a hipótese de que nenhum *outlier* está presente na amostra, ao teste de discordância a partir da cápsula normal.

[2] matriz catcher é definida por $c = (X'X)^{-1}X'$

apresentado na seção 3.1, o método de mínimos quadrados procura o mínimo em $\underset{\sim}{\beta}$ de

$$(1) \qquad \sum_{i=1}^{n}(y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta})^2 = \sum_{i=1}^{n} r_i^2(\underset{\sim}{\beta})$$

que pode ser escrito da seguinte forma

$$\min_{\underset{\sim}{\beta}} \sum \rho(y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta}) =$$

$$= \min_{\underset{\sim}{\beta}} \sum \rho(r_i(\underset{\sim}{\beta}))$$

onde ρ é a função quadrática dos resíduos.

Por diferenciação e levando em conta um estimador de escala, obtém-se:

$$\sum_{i=1}^{n} \psi(r_i/\hat{\sigma})\underset{\sim}{x}_i = \underset{\sim}{0}$$

onde $\psi = \frac{d\rho}{dt}$ e $\underset{\sim}{x}_i$ é um vetor linha de variáveis explicativas.

Uma forma de robustecer o método de estimação poderia ser através da substituição de ρ quadrática por uma outra função, com algumas propriedades básicas. ρ deve ser convexa, não monótona e possuir derivadas limitadas de ordens suficientemente altas. $\psi(t)$, por sua vez, deve ser contínua e limitada.

Os estimadores gerados desta forma são chamados de M-estimadores ou estimadores do tipo máxima verossimilhança, pois em um certo sentido generaliza a idéia subjacente à definição de estimador de máxima verossimilhança.

O surgimento dos M-estimadores (Huber, 1973) constituiu-se em um passo importante em direção à estimação robusta. Muitas pesquisas e esforços têm sido concentrados na construção de funções ρ e ψ tal que os M-estimadores resultantes sejam, ao mesmo tempo, os mais robustos possíveis e eficientes sob normalidade dos erros.

Huber propôs o uso de ψ definida por

$$\psi(t) = \begin{cases} t & se \quad |t| < b \\ b \text{ sinal } (t) & se \quad |t| \geq b, \end{cases}$$

onde b é uma constante, limitando a influência de observações com valores superiores a b, ou inferiores a $-b$. O gráfico desta função é apresentado no Anexo 2 (gráfico 1).

Hampel (1974) definiu uma função ψ que proteje o ajuste contra observações bastante afastadas,

$$\psi(t) = \begin{cases} t & se \quad |t| < a \\ a \text{ sinal } (t) & se \quad a \leq |t| < b \\ \{(c - |t|/(c-b)\}a \text{ sinal } (t) & se \quad b \leq |t| \leq c \\ 0 & c.c. \end{cases}$$

gerando o chamado M-estimador *three-part redescending*.

Ao longo do tempo, alguns aperfeiçoamentos foram introduzidos nos M- estimadores, com o objetivo de aumentar o seu ponto de ruptura, que é zero no modelo de regressão, conseqüência imediata da sua vulnerabilidade aos pontos de alavanca. Assim surgiram os M-estimadores generalizados ou GM-estimadores que dão uma ponderação aos $\underset{\sim}{x}_i$, com o objetivo de diminuir o peso dos pontos de alavanca e têm ponto de ruptura como função de p (número de parâmetros a estimar), isto é, quando p cresce, o ponto de ruptura decresce para zero (Maronna, Bustos e Yohai, 1979). Yohai (1985) observou, ainda, que os GM-estimadores têm baixa eficiência na presença de pontos de alavanca *bons*.

Muitos outros estimadores foram propostos, entre eles os R-estimadores, baseados nos postos dos resíduos do ajuste (Adichie (1967), Jureckova (1971) e Jaeckel (1972)) e os L-estimadores, baseados em combinações lineares de estatísticas de ordem (veja Rousseeuw e Leroy, 1987 para maiores detalhes).

O primeiro estimador de regressão com ponto de ruptura máximo foi apresentado por Siegel (1982). Ele está baseado em todos os subconjuntos de p pontos que se podem formar e definido por

$$\hat{\beta}_j = \underset{i_1}{\text{med}}(\ldots(\underset{i_{p-1}}{\text{med}}(\underset{i_p}{\text{med}}\beta_j(i_1,\ldots,i_p)))\ldots)$$

Rousseeuw (1984) apresentou o LMS (least median squares) como resultado da substituição do operador de somatório em (1) por mediana, obtendo-se:

$$\underset{\underset{\sim}{\beta}}{\text{med}}(y_i - \underset{\sim}{x}_i'\underset{\sim}{\beta})^2$$

Embora o ponto de ruptura do LMS seja de 50%, sua eficiência sob normalidade dos erros, é baixa. Rousseeuw e Leroy (1987) mostram como LMS pode ser usado como uma boa ferramenta para identificação de múltiplos *outliers*, além de ressaltarem a sua adequabilidade como valor inicial para o cálculo de M-estimadores. Sem essa preocupação, corre-se o risco de terminar-se o processo de estimação com um mínimo local que não corresponde à solução robusta esperada.

Embora o algoritmo utilizado para a obtenção do LMS esteja implementado no PROGRESS (Program for Robust Regression), deixamos para outro trabalho a extração das sub-rotinas necessárias para sua inclusão no estudo Monte Carlo.

Finalmente, com o objetivo de combinar eficiência, sob normalidade e alto ponto de ruptura, Yohai e Zamar (1986) definiram o τ-estimador descrito a seguir.

Seja ρ uma função real satisfazendo as seguintes propriedades:

i) $\rho(0) = 0$

ii) $\rho(-u) = \rho(u)$

iii) $0 \leq u \leq v \Rightarrow \rho(u) \leq \rho(v)$

iv) ρ é contínua

v) Seja $a = \sup \rho(u)$, então $0 < a < \infty$

vi) Se $\rho(u) < a$ e $0 \leq u < v$ então $\rho(u) < \rho(v)$

Huber (1981) define o M-estimador de escala de uma amostra $\underset{\sim}{u} = (u_1, \ldots, u_n)$, denotado como a solução da equação $s_n(\underset{\sim}{u})$ por

$$\frac{1}{2}\sum_{i=1}^{n} \rho\left(\frac{u_i}{s_n(\underset{\sim}{u})}\right) = b$$

onde b é convenientemente definido como

$$b = E_\phi(\rho(\underset{\sim}{u})),$$

onde ϕ é a função de distribuição da $N(0,1)$. Sejam ρ_1 e ρ_2 duas funções com as propriedades descritas no parágrafo anterior e seja s_n o M-estimador de escala baseado em ρ_1.

Então, dada uma amostra $\underset{\sim}{u} = (u_1, \ldots, u_n)$ o estimador de escala τ_n é definido por

$$\tau_n^2(\underset{\sim}{u}) = s_n^2(\underset{\sim}{u}) \frac{1}{n} \sum_{i=1}^{n} \rho_2 \left(\frac{u_i}{s_n(\underset{\sim}{u})} \right)$$

E o τ-estimador é definido pelo valor $\underset{\sim}{\hat{\beta}}$ tal que

$$\tau_n(r(\underset{\sim}{\hat{\beta}})) = \min_{\beta} \tau_n(r(\underset{\sim}{\beta}))$$

onde $r(\underset{\sim}{\beta}) = (r_1(\underset{\sim}{\beta}), \ldots, r_n(\underset{\sim}{\beta}))$, isto é, por minimização de um novo estimador de escala, aplicado aos resíduos.

Assintoticamente, um τ-estimador se comporta como um M-estimador, associado a uma função ρ_0, que é uma média ponderada entre as duas funções ρ_1 e ρ_2, empregadas na sua construção.

Um algoritmo para a obtenção do τ-estimador e uma sub-rotina em Fortran IV para o seu cálculo podem ser encontrados em Bustos (1989).

De acordo com Yohai e Zamar, 1986, pág. 17, utilizando-se um estimador inicial para $\underset{\sim}{\beta}$, $T_{0,n}$, com ponto de ruptura máximo e determinadas funções ρ_1 e ρ_2, obtém-se $T_{1,n}$ o estimador de $\underset{\sim}{\beta}$ no passo seguinte com o mesmo ponto de ruptura de 50% e eficiência de 95% para erros normais. Os autores recomendam ainda que dois candidatos para estimadores iniciais de $\underset{\sim}{\beta}$ são os estimadores de Siegel e uma variante do LMS de Rousseeuw e Leroy, 1984. No entanto, como não se dispõe, no momento, de software destes estimadores adequado a um estudo Monte Carlo, optou-se pelo uso do estimador L_1, para estimar $\underset{\sim}{\beta}$, e, assim, iniciar o processo de obtenção do τ-estimador.

O estimador L_1 é definido por

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} |r_i|$$

Embora a substituição da função quadrática de (1) por módulo resulte em ganho de robustez, em termos de ponto de ruptura, L_1 não é nada melhor que LS, pois L_1 permanece vulnerável a pontos de alavanca.

As funções ρ_1 e ρ_2, escolhidas para esse estudo estão definidas a seguir:

$$\rho_1 = \rho_2 = \rho$$

onde

$$\rho(t) = \begin{cases} \frac{t^2}{2}\left(1 - \frac{t^2}{c^2} + \frac{t^4}{3c^4}\right) & \text{se} \quad |t| \leq c \\ \frac{c^2}{6} & \text{se} \quad |t| > c \end{cases}$$

A ψ correspondente está definida por

$$\psi(t) = \begin{cases} t\left(1 - \frac{t^2}{c^2}\right)^2 & \text{se} \quad |t| < c \\ 0 & \text{se} \quad |t| \geq c \end{cases}$$

3.3 O estudo Monte Carlo

Com o objetivo de comparar os resultados do estudo apresentado em Rousseeuw e Leroy, 1987 (Cap. 5), com os obtidos aqui, optou-se por considerar o mesmo modelo básico (situação normal) e as mesmas variações em torno do modelo básico, incluindo-se apenas uma variação adicional. A seguir apresentam-se as descrições dos modelos empregados.

Modelo Básico

$$y_i = x_{i,1} + \ldots, + x_{i,p-1} + x_{i,p} + \varepsilon_i$$

1) situação normal

Todas as amostras são geradas de forma que $\varepsilon_i \sim N(0,1)$ e $x_{i,j} \sim N(0,100)$ para $j = 1, \ldots, p$ (se não há intercepto) e para $j = 1, \ldots, p-1$ (se há intercepto e então $x_{i,p} = 1$).

2) modelo para introdução de *outliers* na direção Y contaminando a média dos erros.

Neste caso, as amostras têm 80% das observações geradas como na situação normal e 20% contaminadas fazendo-se $\varepsilon_i \sim N(10,1)$.

3) modelo para introdução de *outliers* na direção X, contaminando a primeira variável explicativa.

Estas amostras são geradas fazendo-se 80% dos casos de acordo com a situação normal e o restante contaminando $x_{i,1}$ com uma $N(100, 100)$

4) modelo para introdução de *outliers* na direção Y, contaminando a variância dos erros.

Neste caso as amostras têm 80% dos casos gerados a partir da situação normal e o restante fazendo-se $\varepsilon_i \sim N(0, 100)$.

Os gráficos 11 a 14, no Anexo 2, ilustram estas quatro situações para o caso de $p = 2$, com intercepto e $n = 50$.

A tabela A mostra o número de modelos considerados no estudo Monte Carlo, para cada par (n, p).

Tabela A:

	p^*	1	2	8
n				
	20	4	8	-
	50	4	8	8

* p é o número de coeficientes a estimar, já incluindo o intercepto, quando existir.

Para cada um destes modelos são gerados 200 amostras. Uma descrição detalhada a respeito das sub-rotinas de geração de números aleatórios utilizadas é apresentada, no apêndice.

Devido ao grande número de cruzamentos entre modelos e estimadores, estabeleceu-se uma notação, para auxiliar a apresentação e análise dos resultados.

A notação procurou combinar de forma compacta as informações a respeito do procedimento de estimação e do modelo associado.

A notação obedece ao seguinte esquema básico:

$X\ X\ X\ X\ X\ X$	$X\ X$	X	X	X
(1)	(2)	(3)	(4)	(5)

onde:

(1) representa o estimador utilizado, tendo-se adotado a seguinte convenção:

LS – estimador de mínimos quadrados

L1 – L_1

TAU – τ-estimadores

COV – COVRATIO $_i + LS$, versão *um de cada vez (one at a time)*

LIK – LD $_i + LS$, versão *um de cada vez (one at a time)*

DF – DFITS $_i + LS$, versão *um de cada vez (one at a time)*

CNLS – Cápsula normal $+LS$, versão *um de cada vez (one at a time)*

STCOV – COVRATIO + LS, versão *todos de uma vez (several at a time)*

STLIK – LD $_i + LS$, versão *todos de uma vez (several at a time)*

STDF – DFITS $_i + LS$, versão *todos de uma vez (several at a time)*

STCNLS – Cápsula normal + LS, versão *todos de uma vez (several at a time).*

(2) indica o tamanho da amostra (n)

(3) indica o valor de p

(4) indicador de utilização de intercepto no modelo, isto é, 0 (zero) se não há intercepto e 1 (um) caso contrário.

(5) indica o tipo de contaminação do modelo. Assume os valores:

0 (zero), se não há contaminação.

1 (um), se há contaminação da média dos erros

2 (dois), se há contaminação na variância dos erros

3 (três), se há contaminação na primeira variável explicativa.

A seguir apresentam-se dois exemplos que ilustram a sua aplicação:

1) LS20100 traduz-se em:

LS iniciais do estimador de mínimos quadrados

20 tamanho da amostra

1 $p = 1$, um parâmetro a estimar

0 modelo sem intercepto

0 modelo sem contaminação.

2) TAU50211 traduz-se em:

TAU τ-estimador

50 tamanho da amostra

2 $p = 2$, dois parâmetros a estimar

1 modelo com intercepto

1 contaminação na média dos erros.

Para a confecção dos gráficos utilizados na análise dos resultados foi necessário compactar ainda mais a representação dos modelos. Utilizou- se a seguinte convenção:

A0 - 20100	C2 - 20212	F0 - 50210	H2 - 50812
A1 - 20101	C3 - 20213	F1 - 50211	H3 - 50813
A2 - 20102	D0 - 50100	F2 - 50212	
A3 - 20103	D1 - 50101	F3 - 50213	
B0 - 20200	D2 - 50102	G0 - 50800	
B1 - 20201	D3 - 50103	G1 - 50801	
B2 - 20202	E0 - 50200	G2 - 50802	
B3 - 20203	E1 - 50201	G3 - 50803	
C0 - 20210	E2 - 50202	H0 - 50810	
C1 - 20211	E3 - 50203	H1 - 50811	

CAPÍTULO 4

RESULTADOS E CONCLUSÕES

Os resultados do estudo Monte Carlo são apresentados em 32 tabelas (uma para cada modelo) e 191 gráficos, incluídos, respectivamente, nos Anexos 1 e 2.

Vale a pena lembrar que muitos gráficos apresentados no Anexo 2, embora não tenham sido mencionados, especificamente, na análise dos resultados, foram anaxados ao trabalho por terem sido considerados úteis à melhor compreensão do estudo em questão.

As tabelas contêm, nesta ordem, as seguintes informações por estimador e parâmetro ($\underset{\sim}{\beta}$ e σ):

1) a média do j-ésimo estimador

$$\overline{\beta}_j = \frac{1}{200} \sum_{k=1}^{200} \hat{\beta}_j^{(k)}$$

onde $\hat{\beta}_j^{(k)}$ representa a k-ésima estimativa de β_j

2) a variância Monte Carlo do j-ésimo estimador

$$\hat{\text{Var}}(\hat{\beta}_j) = \frac{1}{200} \sum_{k=1}^{200} \left(\hat{\beta}_j^{(k)} - \overline{\beta}_j\right)^2$$

3) o erro quadrático médio Monte Carlo do j-ésimo estimador

$$\hat{EQM}(\hat{\beta}_j) = \frac{1}{200} \sum_{k=1}^{200} \left(\hat{\beta}_j^{(k)} - \beta_j\right)^2$$

que pode ser decomposto em duas parcelas, isto é, o vício ao quadrado e variância.

$$\hat{EQM}(\hat{\beta}_j) = (\overline{\beta}_j - \beta_j)^2 + \frac{1}{200}\sum_{k=1}^{200}\left(\hat{\beta}_j^{(k)} - \overline{\beta}_j\right)^2$$

4) um estimador robusto do EQM

$$\hat{EQMR}(\hat{\beta}_j) = [MED(\hat{\beta}_j) - \beta_j]^2 + [1{,}483 \;\; MAD(\hat{\beta}_j)]^2$$

onde $MED(\hat{\beta}_j)$ é a mediana dos valores de $\hat{\beta}_j$ e MAD é o MEDIAN ABSOLUTE DEVIATION já mencionado anteriormente.

Observações:

1) Esta última estatística não foi calculada para os modelos com 8 parâmetros.
2) A motivação para o cálculo de um estimador robusto do EQM surgiu com a análise do desempenho do τ-estimador, apresentada em detalhe na seção 4.3. Nesse momento, tentando entender melhor uma perda de eficiência de 4500% em relação ao LS no modelo 20100, verificou-se um fato bastante interessante. Enquanto 96% das estimativas estavam concentradas em torno de 1, o verdadeiro valor do parâmetro, os 4% restantes eram suficientes para provocar a distorção do EQM clássico, prejudicando a avaliação do τ-estimador. Usando o estimador robusto de EQM obtém-se uma perda de eficiência de 13% em relação ao LS, um número bem mais adequado à distribuição apresentada no gráfico 179. Diante disto optou-se por incluir o $\hat{EQM}$ robusto na análise de resultados de todos os estimadores e modelos (exceto os de 8 parâmetros). No entanto esta ferramenta deve ser utilizada com muita cautela, pois em algumas situações ela pode levar a interpretações equivocadas. Observe-se, por exemplo, o que ocorre com a COVRATIO versão *todos de uma vez (several at a time)* sob o modelo 20103 (gráfico 119). A perda de eficiência com base no EQM clássico é de 90.000%, enquanto na versão robusta cai para 6485%, um decréscimo expressivo! No entanto, analisando-se o gráfico correspondente, verifica-se a existência de dois grupos bastante distintos de estimativas: um com 55% das observações, concentrado em torno do verdadeiro valor do parâmetro e outro com os 45% restantes em torno de zero. Como

o estimador robusto de EQM possui ponto de ruptura 50%, ele retrata o comportamento do grupo maior, neste caso, os *bons* dados. Obviamente, essa situação é muito diferente daquela para a qual, inicialmente, foi empregado o EQM robusto. No caso da COVRATIO (versão *several at a time)* o EQM robusto indica artificialmente um melhor desempenho do estimador quando comparado com o EQM clássico.

Os gráficos são basicamente de três tipos: histogramas por modelo de $\hat{\beta}_1$, $\hat{\beta}_p$ quando há intercepto no modelo, σ e N (tamanho da amostra utilizada no cálculo de $\hat{\beta}$); estimativas por tamanho de amostra, para estimadores baseados em regras de rejeição; e gráficos de linhas dos coeficientes de variação associados aos diversos modelos por estimador ($\hat{\beta}_1, \hat{\sigma}$ e $\hat{\beta}_p$).

O coeficiente de variação do estimador é definido como sendo

$$\hat{cv}(\hat{\beta}_j) = \frac{\sqrt{E\hat{Q}M(\hat{\beta}_j)}}{\beta_j}$$

A escolha desta medida de variabilidade deve-se principalmente a sua adimensionalidade, que permite uma interpretação em termos de percentual de variabilidade. Outra razão é que o uso da raiz quadrada, face à magnitude dos valores de $E\hat{Q}M$, permite uma melhor visualização gráfica dos resultados.

A análise dos resultados é apresentada por estimador e com bastante detalhe para os modelos 20100, 20101, 20102 e 20103. A seção 4.1 aborda o LS. A seção 4.2 trata do desempenho dos estimadores $DRLS$, enquanto em 4.3 observa-se o comportamento do τ-estimador. Finalmente, a seção 4.4 apresenta as limitações do estudo e alguns comentários a respeito do LMS.

4.1 LS – Análise sob os diversos modelos

a) Como se pode observar no gráfico 15, referente a $\hat{\beta}_1$, há um padrão de comportamento que se repete ao longo de cada classe de modelos.

b) A contaminação na direção dos X é, claramente, a mais danosa na estimação de $\hat{\beta}_1$ com coeficientes de variação (daqui em diante cv) em torno de 100%.

c) Para os outros dois tipos de contaminação o cv do estimador fica em torno de 10% e 6%, para $n = 20$ e $n = 50$, respectivamente.

d) Como já se esperava, o estimador de β_1 se comportou bastante bem, quando não há contaminação.

e) Observando-se o gráfico 16, referente a $\hat{\sigma}$, nota-se que o padrão apresentado pelo gráfico 15 se repete, embora a magnitude dos cv_s seja bastante distinta. Para as contaminações do tipo 1 e 2, os cv_s situam-se em torno de 350%, enquanto na contaminação do terceiro tipo observam-se valores em torno de 900%.

f) É evidente a fragilidade de $\hat{\sigma}$ diante das três contaminações consideradas.

g) Observando-se o gráfico 17 referente a $\hat{\beta}_p$ (o intercepto), nota-se que, em geral, os cv_s assumem valores bem superiores àqueles associados a $\hat{\beta}_1$, mesmo nos modelos sem contaminação. Com exceção do modelo 20213, com aproximadamente 260% de cv, todos os outros modelos de final ímpar (contaminação na média dos erros e na variável independente) estão com cv_s na faixa de 200%.

h) Aparentemente, os $\hat{\beta}_p$ para modelos com $n = 20$ estão com cvs maiores do que para $n = 50$.

i) Os modelos com contaminação na variância dos erros estão com cv na faixa de 60% a 100%.

j) O gráfico 15A apresenta a comparação entre os cv_s clássico e robusto[1], por modelos, para $\hat{\beta}$. Observe-se que praticamente não há diferença entre as curvas, indicando que para o LS as medidas são equivalentes.

[1] o cv robusto associado a $\hat{\beta}_j$ é dado por $\hat{cvR}(\hat{\beta}_j) = \frac{\sqrt{EQMR(\hat{\beta}_j)}}{\hat{\beta}_j}$.

Análise detalhada de LS sob os modelos 20100, 20101, 20102 e 20103.

a) Os gráficos 18 e 19 relativos, respectivamente, a $\hat{\beta}$ e $\hat{\sigma}$ mostram a concentração das estimativas em torno de 1 no modelo sem contaminação, ao longo das 200 repetições do estudo Monte Carlo.

b) Os gráficos 20 a 23, associados a 20101 e 20102, mostram que embora as estimativas de β_1 ainda estejam em torno de 1, as de σ perdem completamente o sentido com $E\hat{Q}M$ de 12.

c) Os gráficos 24 e 25 referentes ao modelo 20103 demonstram a inadequabilidade de LS nessa situação, com as estimativas de β_1 concentradas em torno de zero e as de σ entre 5 e 15. Os EQM estimados são, respectivamente, 0.92 e 75.0, para $\hat{\beta}$ e $\hat{\sigma}$.

4.2 Estimadores *DRLS*

4.2.1 Estimadores *DRLS* - versão *um de cada vez (one at a time)*

4.2.1.1 COVRATIO+LS - Análise sob os diversos modelos.

a) Há perda de eficiência sob os modelos sem contaminação. Essa perda é apresentada para $\hat{\beta}_1$ e $\hat{\sigma}$ na tabela a seguir. Aparentemente, na estimação de β_1, a perda diminui quando n cresce, exceto para o modelo 50100. Observando-se $\hat{\sigma}$, nota-se o aumento da perda de eficiência, quando o número de parâmetros é oito.

	Perda de eficiência da COVRATIO	
Modelos	$\hat{\beta}_1(\%)$	$\hat{\sigma}(\%)$
20100	394	106
20200	385	176
20210	324	128
50100	371	131
50200	129	109
50210	45	159
50800	68	646
50810	112	633

b) Observando-se o gráfico 26, para $\hat{\beta}_1$, nota-se a dificuldade do estimador diante dos modelos 50803 e 50813. Aparentemente, o aumento do número de parâmetros tem influência nesse comportamento.

c) Os modelos com tamanho de amostra 50 (exceto os citados anteriormente) têm geralmente um nível de cv menor.

d) Observa-se que esse procedimento é sempre melhor que LS para estimar β_1 nos modelos com contaminação na variância dos erros e na direção de X, com cv_s de no máximo 20%. Com relação a contaminação na média dos erros, este procedimento é bastante bom em um grande número de amostras, gerando distribuições empíricas bastante concentradas em torno de 1, exceto por algumas estimativas. No entanto estes poucos pontos fazem com que o EQM assuma valores elevados. Neste caso, a análise com base em $EQMR$, gráfico 26A, é mais valiosa, mostrando que $COVRATIO$ é, em geral, superior a LS. Ressalte-se que, sob essa contaminação, a COVRATIO tem muita dificuldade em estimar o intercepto.

e) No que diz respeito a $\hat{\sigma}$, gráfico 27, a COVRATIO tem sempre cv_s mais baixos exceto nos modelos 50803 e 50813. A faixa de variação para os modelos com no máximo 2 parâmetros é de 16% a 325%.

f) O estimador de σ é sempre bastante razoável sob os modelos com contaminação na variância dos erros com cv_s variando entre 17% e 93%.

g) Ainda com relação a σ, nos modelos com contaminação na direção de X e com no máximo dois parâmetros a estimar, é, sem dúvida, melhor que LS, com cv_s variando entre 14% e 199%. Já os modelos com 8 parâmetros, 50803 e 50813, é pior que LS. Comparando-se os cv_s clássico e robusto, observa-se que este último apresenta valores sempre menores, que, neste caso, indicam a presença de alguns *outliers* na distribuição empírica relativa as 200 repetições Monte Carlo.

h) A análise do EQM mostra que nos casos de contaminação na média dos erros $\hat{\sigma}$ apresenta um desempenho inferior ao observado nos outros modelos, embora ainda

seja melhor que o LS (exceto em 50801 e 50811, nos quais LS é melhor). Comparando EQM e $EQMR$, nota-se que nos modelos 20101, 50101 e 50201, $EQMR$ é bem menor que EQM, já em 20201, 20211 e 50211 ocorre o contrário gráfico 27A. Investigando-se a origem deste resultado, verificou-se que em geral as distribições empíricas relativas a $\hat{\sigma}$ são bimodais e sob os modelos 20101, 50101 e 50201 mais de 50% das estimativas estão em torno do verdadeiro valor do parâmetro, enquanto nos três outros modelos, menos de 50% das estimativas estão distantes de 1. Diante disto, $EQMR$ com ponto de ruptura de 50% resiste à influência da minoria, que, por exemplo, em 20201 são os *bons* dados (ou seja, aqueles que estão em torno de 1). Em muitos destes casos, não são apenas 3% ou 4% das estimativas que estão distorcidas mas 40%. Daí a preocupação com o uso de $EQMR$, abordada no início deste capítulo.

i) Observando-se o gráfico 28 com os coeficientes de variação referentes ao $\hat{\beta}_p$ (intercepto), nota-se a superioridade da COVRATIO sobre o LS nos modelos com contaminação na variância dos erros e na direção de X.

Análise detalhada de COVRATIO sob os modelos 20100, 20101, 20102 e 20103.

a) Uma das principais características da $COVRATIO$ é o seu alto índice de rejeição. Em todos os modelos considerados em mais de 50% das amostras do estudo Monte Carlo atingiu-se o percentual máximo permitido de rejeição. Isso pode explicar a grande perda de eficiência sob todos os outros modelos sem contaminação já que, também nesses casos, a $COVRATIO$ estima os parâmetros com amostras bem menores do que deveria, embora nem sempre a rejeição atinja índices tão elevados.

b) Nota-se, claramente, nos gráficos 34 a 37, referentes a $\hat{\beta}$ e $\hat{\sigma}$ no modelo 20101, a incapacidade da regra de rejeição sob essa contaminação. Ressalte-se que na estimação de σ surgem dois grupos de resultados: um em que a regra, aparentemente, identificou corretamente os *outliers* e outro em que isso não ocorreu.

c) O comportamento da $COVRATIO$ sob contaminação na variância dos erros (gráficos 39 a 42) é bastante bom, tanto na estimação de β_1 quanto na de σ. Com relação a σ, observa-se um pequeno grupo de estimativas bem afastado do verdadeiro valor do parâmetro, contribuindo para um EQM estimado de 0,40. O estimador robusto do EQM assume valor 0,12 neste caso, o menor entre os estimadores baseados em regras de rejeição.

d) A contaminação na direção X é aquela onde $COVRATIO$ tem o melhor desempenho. A análise dos gráficos 44 a 47 indica que, a exceção de uma amostra das 200 geradas, o comportamento do estimador é tão bom quanto no modelo sem contaminação. Neste caso o estimador robusto de EQM vale 0.07 enquanto o clássico assume valor 0,45.

4.2.1.2 $LD_i + LS$ ou LIK

Análise sob os diversos modelos

a) Superpondo-se os gráficos 49 a 51 referentes a $\hat{\beta}_1$, $\hat{\beta}_p$ e $\hat{\sigma}$ com os gráficos 15 a 17, constata-se a semelhança de comportamento com o LS puro.

b) Apenas nos modelos com contaminação na variância dos erros LD_i ou LIK apresenta superioridade.

Análise detalhada de LIK sob os modelos 20100, 20101, 20102 e 20103.

a) A observação dos gráficos 52 e 53, relativos a 20100 (modelo sem contaminação), demonstra mais uma vez a forte semelhança com LS puro.

b) A causa fica bastante clara ao observar-se o tamanho das amostras utilizadas para o cálculo dos estimadores. PRATICAMENTE NÃO HÁ REJEIÇÃO, o que reduz bastante a perda de eficiência sob modelos sem contaminação (gráfico 54).

c) Já para o modelo 20101 a não rejeição implica maior dispersão para $\hat{\beta}_1$ e $\hat{\sigma}$. Comparando-se com a COVRATIO observa-se que LIK é quase sempre melhor na estimação de β_1 e, geralmente, menos eficiente para estimar σ.

d) Como se pode observar nos gráficos 58, 59 e 60 referentes ao modelo 20102, os estimadores de β e σ diante da contaminação na variância dos erros se comportam razoavelmente bem. Isso, aparentemente, se deve à maior capacidade, por parte desta regra, de identificação de *outliers* gerados sob este tipo de contaminação. Em relação a COVRATIO o comportamento de LIK sob esta classe de modelos é, globalmente, pior. Ressalte-se a dificuldade de LIK com a estimação de σ.

e) A contaminação na direção X é tão danosa aqui quanto no LS. Como quase não há rejeição de observações por parte da regra, os dois procedimentos de estimação, praticamente, coincidem (gráficos 61 a 63).

4.2.1.3 $DFITS + LS$

Análise sob os diversos modelos

a) Os gráficos 64 a 84 são referentes a $DFITS$. Superpondo-se os gráficos 15, 26 e 64 verifica-se a superioridade desta regra na estimação de β sobre LS e $COVRATIO$, em todos os modelos com contaminação, exceto nos contaminados na direção X. A eficiência de $DFITS$ com relação a $COVRATIO$ (isto é, $E\hat{Q}M(COVRATIO)/E\hat{Q}M(DFITS)$), para todos os modelos com até dois parâmetros a estimar, é mostrada no gráfico 190.

b) Nitidamente, $DFITS$ tem dificuldades para estimar β e σ diante de contaminação na direção X. Sob esta classe de modelos $DFITS$, muitas vezes, produziu dois grupos distintos de estimativas de β e σ: um em torno do verdadeiro valor do parâmetro e outro longe. Observe-se como este fato influencia os valores de $EQMR$ apresentados

junto com os de EQM nos gráficos 64A e 65A. Observe-se ainda neste gráfico como, em geral, não há diferença entre $EQMR$ e EQM sob os outros modelos.

c) Observe-se que no caso do modelo 50103 (D3) o desempenho do estimador é bom, contrariando o que seria de se esperar para este tipo de contaminação. Aparentemente, a capacidade de identificação de *outliers* por parte da regra está associada à razão entre tamanho da amostra e número de parâmetros a estimar, que é máxima sob $D3$. Isso também se verifica no caso da $COVRATIO$.

d) Com relação a σ pode-se observar que $DFITS$ tem um desempenho melhor que LS sob todos os modelos contaminados. Com relação a $COVRATIO$ não se pode afirmar o mesmo. De fato, apenas no caso de contaminação na média dos erros $DFITS$ é superior. Nos outros modelos o comportamento se inverte (gráfico 191).

e) O intercepto é mais bem estimado por $DFITS+LS$ nos modelos com contaminação na média dos erros e pela $COVRATIO + LS$ nos modelos com contaminação na direção de X. Nos modelos com contaminação na variância dos erros, os procedimentos são equivalentes e bastante razoáveis.

Análise detalhada de $DFITS$ sob os modelos 20100, 20101, 20102 e 20103

a) A perda de eficiência de $DFITS$ sob os modelos sem contaminação é menor que o da $COVRATIO$. Isso se deve ao menor índice de rejeição de observações, ou seja, amostras maiores para efetuar a estimação. Observe-se no gráfico 69 como a maioria das amostras tem 19 observações, seguida de 18, 17 e 20. O gráfico 67 mostra um estimador de β_1 bastante bom. Já o referente a σ (gráfico 68) parece um tanto disperso mas ainda razoável.

b) O gráfico 70 mostra o comportamento curioso de $\hat{\beta}$, sob a contaminação na média dos erros. Uma massa concentrada em torno de 1, o verdadeiro valor do parâmetro, e uma estimativa abaixo de zero. Para investigar o fato gerou-se o gráfico 71 que mostra as estimativas de β segundo o tamanho das amostras utilizadas. Observe-se

que aquele ponto estranho de $\hat{\beta}$ deve ter sido resultado de uma rejeição de observações que começou errada e, sofrendo o efeito de mascaramento, rejeitou pontos *bons* e só parou porque há uma limitação (neste caso o tamanho da amostra é de no mínimo 11 observações).

c) No caso de σ, o estimador já não é tão bem comportado. Há nitidamente dois conjuntos de estimativas, um bom e outro ruim. Na maioria das amostras o estimador deixa a desejar. Talvez tais amostras se aproximem de configurações ruins para o $DFITS$. Neste caso, o grupo de estimativas em torno de 1 tem 88 observações e, obviamente, o outro 112, resultando em um $EQMR$ de 8,44 contra um EQM de 4,47.

d) Observe-se no gráfico 73 os dois grupos de $\hat{\sigma}$. Note-se que, a partir de um certo ponto, quanto menos se rejeita pior é a estimação. Observe-se também que, quando $n = 11$ o $DFITS$ substima σ praticamente todas as vezes, e esta situação vai mudando gradativamente com o tamanho da amostra, atingindo um ponto de equilíbrio quando $n = 15$ e $n = 16$, mostrando que nestes casos a regra rejeita exatamente os *outliers* (exceto em 4 casos).

e) Como se pode observar no gráfico 75 e 76, este estimador de β_1 é bastante bom sob o modelo 20102.

f) Os gráficos 77 e 79 apresentam assimetrias opostas que indicam a correlação entre estimativas ruins de σ e amostras maiores (contendo os *outliers*). Observando-se o gráfico 78 verifica-se que, como ocorre em 20101, quanto menos se rejeita mais aumenta a variabilidade e pior são as estimativas de σ.

g) O conjunto de gráficos 80 a 84 descreve o comportamento de $DFITS$ sob o modelo 20103. Observe-se que na maioria dos casos o ajuste é adequado para β e σ. Em alguns poucos casos (nove) as estimativas são desastrosas. Elas coincidem exatamente com as amostras para as quais a regra $DFITS$ não foi capaz de detectar os *outliers* (amostras de tamanho 19 e 20). Neste caso os EQM_s robustos de ambos os estimadores são menores que um vigésimo dos valores assumidos pelos EQM_s clássicos.

4.2.1.4 Cápsula normal+LS ou CNLS

a) O gráfico 85 mostra que $\hat{\beta}_1$ baseado na cápsula normal é bastante bom, exceto nos modelos com contaminação em X, onde se revela incapaz de rejeitar os *outliers* e equipara-se ao LS.

b) Na estimação de β a cápsula normal apresenta, nas contaminações 1 e 2, desempenho global entre $DFITS$ e $COVRATIO$, sendo este último o procedimento menos eficiente.

c) Na estimação de σ os aspectos gerais são os mesmos, ou seja, a cápsula normal tem seu melhor desempenho sob a contaminação do tipo 2, seguida to tipo 1 e finalmente a do tipo 3 comparável ao LS.

d) Na contaminação do tipo 2 seu desempenho é comparável a $DFITS$ e inferior a $COVRATIO$, na do tipo 1 o seu desempenho se alterna em relação a $COVRATIO$ e é inferior a $DFITS$.

e) O gráfico 85A apresenta a comparação entre os cv_s robusto e clássico para os diferentes modelos. Observe-se que os resultados são praticamente os mesmos.

Análise detalhada de $CNLS$ sob os modelos 20100, 20101, 20102 e 20103

a) Pode-se observar que $CNLS$ rejeita pouco sob o modelo 20100, o que garante pouca perda de eficiência (gráficos 87 a 91).

b) No modelo 20101 as estimativas de β são razoáveis, enquanto em relação a σ observa-se a formação de dois grupos de estimativas: o maior deles, em torno do verdadeiro valor do parâmetro e o outro, em torno de seis. Examinando-se o gráfico 95, de estimativas por tamanho de amostra, verifica-se que o grupo mais afastado provém de amostras menores nas quais a fase de rejeição deve ter sofrido os efeitos do fenômeno de mascaramento. Observe-se que o grupo de estimativas fidedignas corresponde, principalmente, às amostras com 15 e 16 elementos.

c) No modelo 20102 o estimador de β baseado em $CNLS$ comporta-se bem com uma rejeição variada de observações ao longo das amostras (gráficos 97 e 98). Por outro lado, o estimador de σ (gráficos 99 e 100) tem um desempenho bastante inferior em relação a $\hat{\beta}$, apresentando dispersão acentuada e atingindo valores muito afastados do verdadeiro valor do parâmetro.

d) O modelo 20103 contaminado na direção de X tem suas estimativas, tanto de β quanto de σ, completamente distorcidas, indicando a incapacidade da regra em lidar com este tipo de contaminação (gráficos 102 a 105).

e) Isso fica evidente através da análise do gráfico 106 no qual se verifica a alta freqüência de amostras de tamanho máximo, nas quais, obviamente, não houve rejeição alguma. Observa-se ainda no gráfico 105 que mesmo nas amostras menores onde houve rejeição de 9 observações não foi possível rejeitar os verdadeiros *outliers*, provavelmente, devido ao fenômeno de mascaramento provocado por eles mesmos.

4.2.2 Estimadores *DRLS* - versão *todos de uma vez (several at a time)*

Análise Global

a) Em geral, os procedimentos na versão *todos de uma vez (several at a time)* têm desempenho inferior ou igual aos da versão *um de cada vez (one at a time)*, principalmente no que diz respeito a estimação de σ. A única exceção diz respeito a COVRATIO sob contaminação na média dos erros. Como já foi dito, esta é a classe de modelos para a qual $COVRATIO$ tem seu pior desempenho. Aparentemente, a regra baseada em $COVRATIO$ tem dificuldade em identificar os verdadeiros *outliers*. Somando-se isto ao seu alto índice de rejeição, obtém-se um estimador ruim. Então, ao adotar-se a versão *todos de uma vez (several at a time)*, cuja taxa de rejeição é sempre menor, verifica-se uma menor perda de eficiência.

b) A contaminação na direção X é nitidamente a mais danosa. Neste caso, esta versão não consegue estimar razoavelmente nem mesmo β. Para os modelos 20103, 20203, 20213 e 50103 o estimador de β_1 baseado em $COVRATIO$ parece ser o único que consegue identificar alguns *outliers*, mas não o suficiente para torná-lo adequado sob esta contaminação.

c) Para 50203, 50213, 50803 e 50813 as regras parecem equivalentes no sentido de não conseguirem rejeitar os *outliers* e, conseqüentemente, produzem estimativas ruins para β_1.

d) Observando-se as regras de rejeição sob modelos sem contaminação, nota-se que todas se comportam bem na estimação de β_1, apresentando inclusive coeficientes de variação menores que os encontrados na versão *um de cada vez (one at a time)*. Isto se deve ao índice de rejeição de observações que é menor quando se tenta rejeitar todos os *outliers* de uma só vez.

e) Em relação a contaminação na média dos erros, $COVRATIO$ produz o pior estimador de β_1, enquanto $DFITS$ é o melhor desta classe. O mesmo acontecia na versão *um de cada vez (one at a time)*.

f) Já na contaminação da variância dos erros, $DFITS$ é o mais adequado para estimar β seguido por LD_i. O lugar do menos eficiente é ocupado alternadamente por LS, $COVRATIO$ e $CNLS$.

g) Em geral, nos modelos com $n = 50$, nota-se um decréscimo dos coeficientes de variação, exceto quando o número de parâmetros a estimar é oito.

h) A estimação de σ nos modelos com contaminação em X são os casos mais problemáticos. Em geral $DFITS$ é o estimador menos ruim, seguido de $COVRATIO$.

i) Na contaminação do tipo 2, o melhor desempenho é de $COVRATIO$. Apenas nos modelos com 8 parâmetros, $DFITS$ o supera por pouco. O pior nesta classe de modelos é o LS.

j) Na estimação de σ com contaminação na média dos erros os melhores estimadores são *COVRATIO* (20101 e 20201), *DFITS* (20211, 50211, 50801 e 50811) e *CNLS* (50101 e 50201).

k) Ainda na estimação de σ, a perda de eficiência nos modelos sem contaminação é pequena e ainda menor que a observada na versão *um de cada vez (one at a time).* Mais uma vez isto se deve ao menor índice de rejeição de observações, quando a identificação de *outliers* é feita de uma só vez.

Análise detalhada por estimador dos modelos 20100, 20102, 20102 e 20103.

COVRATIO versão *todos de uma vez (several at a time)* ou STCOV

a) Pode-se ver que a *COVRATIO* versão *todos de uma vez (several at a time)* conserva algumas de suas características observadas na versão apresentada anteriormente.

b) Entretanto observa-se uma degeneração global desta versão comparada com a anterior.

c) Nos modelos com contaminação em X, onde antes verificava-se sua melhor performance, houve piora considerável para a estimação de β e σ. Observam-se nitidamente dois grupos de estimativas mostrando exatamente as amostras nas quais a regra não conseguiu identificar os *outliers* (gráficos 119 e 120).

d) Nos modelos 20101 e 20102 a estimação de σ piorou consideravelmente.

LD_i versão *todos de uma vez (several at a time)* ou STLIK

a) No que diz respeito a $\underset{\sim}{\beta}$, as versões *um de cada vez (one at a time)* e *todos de uma vez (several at a time)* são bem parecidas.

b) Globalmente, *STLIK* (versão *several at a time*), por rejeitar menos, ainda é pior que *LIK* (versão *one at a time*), principalmente nos modelos com contaminação na variância dos erros onde *LIK* se comportava bem.

c) Observe-se, por exemplo, no modelo 20102 como as estimativas de σ estão mais dispersas e distantes do verdadeiro valor do parâmetro.

d) Comparando-se os gráficos do tamanho das amostras utilizadas na obtenção das estimativas, fica evidente que a rejeição empregada em LIK é bem mais adequada que a de $STLIK$.

e) Já em 20101 e 20103 os tamanhos utilizados diferem muito pouco, produzindo estimativas igualmente ruins para $\underset{\sim}{\beta}$ e σ nas duas versões.

DFITS **versão** *todos de uma vez (several at a time)* **ou STDF**

a) A perda de eficiência de $STDF$ nos modelos sem contaminação é menor que em $DFITS$ versão *um de cada vez (one at a time)*. Isto se deve ao menor índice de rejeição de *outliers* empregado em $STDF$.

b) Em todos os outros modelos, $DFITS$ tem desempenho superior a $STDF$, tanto em relação a $\underset{\sim}{\beta}$, quanto a σ e ao intercepto.

c) Observe-se nos gráficos 74 e 145 que em 20101 a rejeição é freqüentemente menor. Comparando-se as estimativas de σ, nota-se que agora já nem existe mais um grupo bem estimado como na versão *um de cada vez (one at a time)* (gráficos 72 e 144).

e) A deterioração do estimador de σ também fica clara em 20102, ao compararem-se os gráficos 77 e 147.

f) Em 20103 a situação se agrava de tal forma que nem as estimativas de β suportam a contaminação e vão concentrar-se em torno de zero, deixando alguns poucos pontos (seis apenas) em torno de 1. Em conseqüência, $\hat{\sigma}$ perde completamente o sentido (gráfico 150).

Cápsula normal - versão *todos de uma vez (several at a time)* ou *STCNLS*

a) Em geral *CNLS* estima melhor $\underset{\sim}{\beta}$ e σ que *STCNLS*, exceto para os modelos com contaminação em X, onde são, praticamente, iguais e em 50101, onde *STCNLS* é pouco melhor.

b) Sem dúvida a contaminação em X é intratável por *STCNLS*, como em todas as outras regras na versão *todos de uma vez (several at a time).*

c) Como nos outros estimadores deste tipo, o menor índice de rejeição de *outliers* prejudica o estimador, quando há qualquer contaminação.

d) Observe-se em 20101 e 20102 que as estimativas de σ pioraram com o emprego do *STCNLS* em relação a *CNLS*.

f) Já em 20103 permaneceram com o mesmo vício.

4.3 τ-estimador

Como já foi dito na seção 3.2.3.2, obtém-se o τ-estimador a partir de uma estimativa inicial para β. Para tanto, optou-se pela utilização do L_1. Nesta seção, apresentam-se os resultados obtidos através do Estudo Monte Carlo para estes dois estimadores segundo os seguintes modelos: 20100, 20101, 20102, 20103, 50100, 50101, 50102, 50103, 20203, 50210, 50211, 50212 e 50213. Ressalte-se que o L_1 foi incluído apenas por ser o escolhido como estimativa inicial do τ, não sendo o propósito deste estudo avaliar seu desempenho.

a) Como era de se esperar, o τ-estimador revela-se desastroso na contaminação em X, pois o L_1, estimativa inicial, apresenta ponto de ruptura zero para estes modelos.

b) O algoritmo empregado para o cálculo do τ-estimador (Bustos, 1989) faz uso de sub-rotinas que envolvem algoritmos iterativos de convergência. Tais sub-rotinas são interrompidas, se a convergência não se verifica em um número pré-determinado de

iterações. Com isso, nem sempre foi possível contar com 200 repetições, como para os demais estimadores. Apresenta-se a seguir o número final de estimativas por modelo.

Modelo	Número final de estimativas
20100	198
20101	199
20102	196
20103	182
50100	200
50101	200
50102	199
50103	199
20203	185
50210	200
50211	196
50212	195
50213	189

Observe-se que há maior inicidência de perda de estimativas nos modelos com contaminação em X, onde a estimativa inicial dada por L_1 fica, completamente, distorcida.

c) Algumas vezes, embora o algoritmo tenha chegado a uma solução, esta é de natureza local, levando a estimativas, completamente, deturpadas. Isto faz com que o EQM amostral, que é uma medida não robusta, apresente valores, extremamente, elevados, comprometendo a análise comparativa com os outros estimadores. Neste sentido optou-se por incluir uma medida robusta do EQM, como definido no início deste capítulo. Observando-se o gráfico 177 é possível verificar o impacto desta mudança, passando o τ- estimador a figurar entre os melhores estimadores, a exceção da contaminação em X.

d) Observe-se no gráfico 179, referente ao modelo 20100, o problema da convergência das estimativas. Comparando-o com o gráfico 169 referente a L_1 sob o mesmo modelo, verifica-se a semelhança de comportamento, a menos é claro, nos casos já mencionados.

e) A perda de eficiência do τ-estimador sob normalidade, com relação ao LS, é de 4515% considerando-se o estimador clássico de EQM e de 13%, utilizando-se a versão robusta.

f) Com relação à estimação de σ no modelo 20100 (gráfico 180) não se observa o mesmo comportamento do estimador de β, no que diz respeito à convergência. A perda de eficiência em relação ao LS é de aproximadamente 150%.

g) Nos modelos 20101 e 20102 (gráficos 181 a 185) os estimadores de β e σ apresentam o mesmo padrão de comportamento verificado no modelo anterior. Ressalte-se que isto se explica pelo bom desempenho de L_1 sob estas contaminações, como se observa nos gráficos 171 a 174.

h) Observe-se ainda que, sob as contaminações na média e variância dos erros, o τ-estimador tem desempenho igual ou superior ao dos melhores estimadores baseados em regras de rejeição ($DFITS$ e $COVRATIO$, *one at a time*), tanto para β quanto para σ.

4.4 Comentários finais

O trabalho apresentado envolveu um significativo esforço computacional. Isto foi determinante para que não se considerassem outros níveis de contaminação além dos 20% empregados. Sem dúvida seria muito interessante trabalhar com uma gradação indo de níveis baixos de contaminação como 1% até níveis mais elevados que 20%. Um estudo deste tipo poderia ser muito útil na determinação empírica dos pontos de ruptura dos estimadores. Além disso, poder-se-iam considerar ainda outras situações para os erros como t-student e Cauchy, a exemplo de Hampel, 1985. Outras combinações entre tamanho de amostra e número de parâmetros no modelo deveriam ser analisadas, pois parece haver associação entre estes valores e o ponto de ruptura dos estimadores, sob as diferentes contaminações, como o próprio estudo realizado sugere.

Com relação ao estimador LMS definido na seção 3.2.3.2, para o qual um estudo Monte Carlo análogo[1] foi desenvolvido por Rousseeuw e Leroy, 1987, observa-se que:

a) Sob o modelo sem contaminação, LMS apresenta perda de eficiência superior a todos os estimadores baseados em regras de rejeição. Esta perda, em relação a β_1, varia entre 65% (*DFITS one at a time*) e 659% (LIK). Em relação a σ, LMS é em geral menos eficiente que os estimadores baseados em regras de rejeição versão *several at a time* e mais eficiente que os na versão *one at a time*. Isto se explica através dos índices de rejeição verificados em cada classe de estimadores (sempre maior na versão *one at a time*) e pelo fato de que no estudo de Rousseeuw e Leroy se utiliza um estimador de σ robusto (cf p. 208, do referido trabalho).

b) Para β_1, sob contaminação na média dos erros, LMS só é superado em eficiência por $DFITS$ versão *one at a time* apresentando uma perda de 22%. Com relação a σ, LMS é sempre superior.

c) Sob contaminação em X, LMS apresenta desempenho, incomparavelmente, superior aos outros estimadores tanto para $\underset{\sim}{\beta}$ quanto para σ. Ressalte-se que para $n = 50$ e $p = 8$ todas as regras de rejeição consideradas ultrapassaram os respectivos pontos de ruptura, enquanto LMS não.

[1] Os modelos comparáveis são 50800, 50801 e 50803. O estimador LS apresenta resultados, praticamente, idênticos nos dois estudos Monte Carlo.

BIBLIOGRAFIA

Adichie, J.N. (1967), Estimation of regression coefficients based on rank tests, *Ann. Math. Stat.*, **38**, 894-904.

Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972), *Robust Estimates of Location: Survey and Advances,* Princeton University Press, Princeton, N.J.

Anscombe, F.J. (1960), Rejection of Outliers, *Technometrics*, **2**, 123-147.

Atkinson, A.C. (1985), *Plots, Transformations, and Regression*, Clarendon Press, Oxford.

Atkinson, A.C. (1986), Masking unmasked, *Biometrika*, **73**, 533- 541.

Barnett, V., and Lewis, T. (1978), *Outliers in Statistical Data*, John Wiley & Sons, New York, 2nd edition: 1984.

Beckman, R.J., and Cook, R.D. (1983), Outlier...s, *Technometrics*, **25**, 119-149.

Besley, D.A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics*, John Wiley & Sons, New York.

Bernoulli, D. (1777), The Most Probable choice Between Several Discrepant Observations and the Formation therefrom of Most Likely Induction, in Allen, C.G. (1961), *Biometrica*, **48**, 3-13.

Box, G.E.P. (1953), Non-normality and tests on variances, *Biometrika*, **40**, 318-335.

Bustos, O.H. (1988), Outliers y Robustez, *Informes de Matemática*, série B-044/88, Instituto de matemática Pura e Aplicada - IMPA, Rio de Janeiro, Brasil.

Bustos, O.H. e Orgambide, A.C.F. (1988), CNORT: Uma sub-rotina para generacion de capsulas Normales Usando el IMSL, *Informes de Matemática*, série B-046/88, Instituto de Matemática Pura e Aplicada - IMPA, Rio de Janeiro, Brasil.

Bustos, O.H. (1989), TAUBI: Uma sub-rotina para calcular o τ-estimador biquadrado de regressão usando *NAG* (em fase final).

Bustos, O.H., Caetano, E., e Franco, G. (1989), Resultados de um Estudo Monte- Carlo comparando estimadores robustos e regras de rejeição de *outliers* no modelo de regressão (em fase final).

Chatterjee, S., and Hadi, A.S. (1986), Influential observations, high leverage points, and outliers in linear regression (with discussion), *Statist.Sci.*, **1**, 379-416.

Collet, D., And Lewis, T. (1976), The subjective nature of outlier rejection procedures, *Applied Statistics*, **24**, 228-237.

Cook, R.D. (1977), Detection of influential observation in linear regression, *Technometrics*, **19**, 15-18.

Cook, R.D. (1979), Influential observations in regression. *J. Am. Stat. Assoc.*, **74**, 169-174.

Cook, R.D., Pena, D. and Weisberg, S. (1984). The likelihood displacement: a unifying principle for influence measures. MRC Technical Summary Report 2751, Univ.Wisconsin, Madison.

Cook, R.D and Weisberg, S. (1982), *Residuals and Influence in Regression*, Chapman & Hall, London.

Daniel, C., and Wood, F.S. (1971), *Fitting Equations to Data*, John Wiley & Sons, New York.

Dempster, A.P. and Gasko-Green, M. (1981), New tools for residual analysis, *Ann. Stat.*, **9**, 945-959.

Dixon, W.J., and Tukey, J.W. (1968), Approximate behavior of the distribution of winsorized t (Trimming/Winsorization 2), *Technometrics*, **10**, 83-93.

Donoho, D.L., and Huber, P.J. (1983), The notion of breakdown point, in *A Festschrift for Reich Lehmann*, edited by P. Bickel, K. Doksum, and J.L. Hodges, Jr., Wadsworth, Belmont. CA.

Draper, N.R., and Smith, H. (1966), *Applied Regression Analysis*, John Wiley & Sons, New York.

Edgeworth, F.Y. (1887), On observations relating to several quantities, *Hermathena*, **5**, 279-285.

Ferguson, T.s. (1961), On the Rejection of Outliers, in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (vol 1)*, Berkeley, Calif.: University of California Press, 253-287.

Grubbs, F.E. (1969), Procedures for detecting outlying observations in samples, *Technometrics*, **11**, 1-21.

Hampel, F.R. (1968), Contributions to the Theory of Robust Estimation, unpublished Ph.D. thesis, University of California, Berkeley.

Hampel, F.R. (1971), A general qualitative definiton of robustness, *Ann. Math. Stat.* **42**, 1887-1896.

Hampel, F.R. (1973). Robust estimation: A condensed partial survey. Z. Wahrsch.verw.geb. **27**, 87-104

Hampel, F.R. (1974), The influence curve and its role in robust estimation, *J. Am. Stat. Assoc.*, **69**, 383-393.

Hampel, F.R. (1977), Rejection Rules and Robust Estimates of Location: An Analysis of Some Monte Carlo Result, in *Preceedings, of the 1974 European Meeting of Statisticians and Seventh Prague Conference on Information Theory Statistical Decisions Functions, and Random Process*, (vol. A), 187-194.

Hampel, F.R. (1978a). Modern trends in the theory of robustness. Math. Operationsforschung Statist. Ser Statist. **9**, 425-442.

Hampel, F.R. (1978b), Optimally bounding the gross-error sensitivity and the influence of position in factor space, in *Proceedings of the Statistical Computing Section of the American Statistical Association*, ASA, Washington, D.C., 59-64.

Hampel, F.R. (1980). Robuste Schätzungen: Ein anwendungsorientierter Überblick. BiomJ. **22**, 3-21.

Hampel, R.F. (1985), The breakdown points of the mean combined with some rejection rules, *Technometrics*, **27**, 95-107.

Hampel, R.F., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, John Wiley & Sons, New York.

Hawkins, D.M. (1980), *Identification of Outliers*, Chapman & Hall, London.

Hawkins, D.M., Bradu, D., and Kass, G.V. (1984), Location of several outliers in multiple regression data using elemental sets, *Technometrics*, **26**, 197-208.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1983), *Understanding Robust and Exploratory Data Analysis*, John Wiley & Sons, New York.

Hoaglin, D.C., Mosteller, F., and Tukey, J.W. (1985), *Exploring Data Tables, Trends, and Shapes*, John Wiley & Sons, New York.

Hodges, J.L., Jr. (1967), Efficiency in normal samples and tolerance of extreme values for some estimates of location, *Proc. Fifth Berkeley Symp. Math. Stat. Probab.*, **1**, 163-168.

Huber, P.J. (1964), Robust estimation of a location parameter, *Ann. Math. Stat.* **35**, 73-101.

Huber, P.J. (1973), Robust regression: Asymptotics, conjectures and Monte Carlo, *Ann. Stat*, **1**, 799-821.

Huber, P.J. (1974), Early Cuneiform Evidence for the Planet Venus, paper presented at the anual meeting of the *American Association for the Advancement of Science*, San Francisco.

Huber, P.J. (1977). Robust methods of estimation of regression coefficients. Math. Operationsforschung Statist. Ser.Statist. **8**, 41-53.

Huber, P.J. (1981), *Robust Statistics*, John Wiley & Sons, New York.

Huber, P.J. (1984), Finite sample breakdown of M and P-estimators. Ann.Statist. **12**, 119-126.

Jaeckel, L.A. (1972), Estimating regression coefficients by minimizing the dispersion of residuals, *Ann. Math. Stat.*, **43**, 1449-1458.

Jureckova, J. (1971), Nonparametric estimate of regression coefficients, *Ann. Math. Stat.*, **42**, 1328-1338.

Leroy, A., and Rousseeuw, P.J. (1984), PROGRESS: A Program for Robust Regression Analysis, Technical Report 201, Center for Statistics and O.R., University of Brussels, Belgium.

Maronna, R.A., Bustos, O., and Yohai, V. (1979), Bias- and efficiency- robustness of general M-estimators for regression with random carriers, in *Smoothing Techniques for Curve Estimation*, edited by T. Gasser and M. Rosenblatt, Springer Verlag, New York, pp. 91-116.

Montgomery, D.C., and Peck, A.E. (1982), *Introduction to Linear Regression Analysis*, John Wiley & Sons, New York.

Mosteller, F., and Tukey, J.W. (1977), *Data Analysis and Regression*, Addison-Wesley, Reading, **MA**.

Myers, R.H. (1986), Classical and Modern Regression with Applications, Duxbury Press, Boston.

Pearson, E.S. (1931), The analysis of variance in cases of non-normal variation, *Biometrika*, **23**, 114-133.

Prescott, P. (1980), A review of some robust data analysis and multiple outlier detection procedures, Bias, **7**, 141-158.

Ronchetti, E. (1982a), Robust alternatives to the F-test for the linear model. In Probability an Statistical Inference, W. Grossmann, C. Pflug, and W. Wertz (eds). Reidel, Dortrecht, pp. 329-342.

Ronchetti, E. (1982b), Robust Testing in Linear Models: The Infinitesimal Approach, Ph.D. Thesis, **ETH** Zürich.

Rousseeuw, P.J. (1984), Least median of squares regression, *J. Am. Stat. Assoc.*, **79**, 871-880.

Rousseeuw, P.J., and Yohai, V. (1984), Robust regression by means of *S*- estimators, in *Robust and Nonlinear Time Series Analysis*, edited by J. Franke, W. Härdle, and R. D. Martin, Lecture Notes in Statistics Nº 26, Springer Verlag, New York. pp.256-272.

Rousseeuw, P.J. and Leroy, A. (1987), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley.

Schrader, R.M. and Hettmansperger, T.P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. Biometrika **67**, 93-101.

Schweingruber, M. (1980), Das Monte Carlo Verhalten Einiger Verwerfungsregeen, unpublished diploma thesis, Eidgenössische Technische Hochschule Zurick, Fachgruppe für Statistik.

Shapiro, S.S. and Wilk, M.B. (1965), An Analysis of Variance Test for Normality, *Biometrika*, **51**, 591-611.

Siegel, A.F. (1982), Robust regression using repeated medians, *Biometrika*, **69**, 242-244.

Student (1927), Errors of routine analysis, *Biometrika*, **19**, 151- 164.

Tukey, J.W. (1960), A survey of sampling from contaiminated distributions, in *Contributions to Probability and Statistics*, edited by I. Olkin, Stanford University Press, Stanford, **CA**.

Tukey, J.W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.

Velleman, P.F. and Hoaglin, D.C. (1981), *Applications, Basics, and Computing of Exploratory Data Analysis*, Duxbury Press, Boston.

Weisberg, S. (1980), *Applied Linear Regression*, John Wiley & Sons, New York (2nd 1985).

Yohai, V. J. (1985), High breakdown-point and high efficiency robust estimates for regression, to appear in *Ann. Stat.*.

Yohai, V. and Zamar, R. (1986), High breakdown-pont estimates of regression by mean of the minimization of an efficient scale. Technical Report Nº 84, Department of Statistics, University of Washington, Seattle.

APÊNDICE

Algoritmos de Geração das Amostras do Estudo Monte Carlo

Neste apêndice são apresentadas as sub-rotinas empregadas na geração das amostras utilizadas no estudo Monte Carlo descrito na seção 3.3, incluindo-se uma listagem dos programas em linguagem FORTRAN 77.

As amostras foram geradas para os trinta e dois modelos mencionados na Tabela A, página 46, segundo as quatro situações de contaminação consideradas, a saber:

1) situação normal (sem contaminação);

2) contaminação na média dos erros;

3) contaminação na média da primeira variável explicativa;

4) contaminação na variância dos erros.

Para cada modelo foram geradas duzentas repetições.

A geração de cada modelo é feita a partir de um programa principal, GENMOD. FOR, que emprega dois subprogramas, a saber: FUNCTION RANDOM e SUBROUTINE SNCRAN. A listagem anexada ao trabalho contém, em forma de comentários, as informações necessárias a seu uso.

```
C PROGRAMA GENMOD.FOR
C PROPOSITO: GERAR DIVERSOS MODELOS DE REGRESSAO LINEAR.
C
C VARIAVEIS:
C  IX,IY,IZ : TRES INTEIROS NO INTERVALO 1 - 30000 UTILIZADOS PARA
C             GERAR NUMEROS PSEUDO-ALEATORIOS. (INPUT/OUTPUT)
C  NOBS   : INTEIRO, NUMERO DE CASOS A CONSIDERAR, NOBS <= 100. (INPUT)
C  NCOREG: INTEIRO, NUMERO DE COEFICIENTES DE REGRESSAO A ESTIMAR,
C           NCOREG <= 20. (INPUT)
C  INTERC: INTEIRO COM VALORES ZERO OU UM,
C           = 0 SE O MODELO A GERAR TEM INTERCEPTO,
C           = 1 SE O MODELO A GERAR NAO TEM INTERCEPTO.
C           (INPUT)
C           O NUMERO DE VARIAVEIS EXPLICATIVAS A GERAR SERA':
C           NEXPL = NCOREG - INTERC
C  INDSIT: INTEIRO COM VALORES 0,1,2, OU 3,
C           = 0 SE A SITUACAO A CONSIDERAR E' A "NORMAL",
C           = 1 SE A SITUACAO A CONSIDERAR E' A DE CONTAMINACAO
C               NA MEDIA DOS ERROS,
C           = 2 SE A SITUACAO A CONSIDERAR E' A DE CONTAMINACAO
C               NA VARIANCIA DOS ERROS,
C           = 3 SE A SITUACAO A CONSIDERAR E' A DE CONTAMINACAO
C               NA VARIAVEL X1
C           (INPUT)
C  EXPSIG: REAL, VALOR DO DESVIO PADRAO COMUM A TODAS AS VARIAVEIS
C           EXPLICATIVAS (INPUT)
C  NREPL : INTEIRO, NUMERO DE REPETICOES A REALIZAR. (INPUT)
C  DELTA : REAL DE DUPLA PRECISAO COM O VALOR DA PROPORCAO DE
C           CONTAMINACAO QUE SE QUER CONSIDERAR.
C           DELTA = 0.0D0 NA SITUACAO INDSIT=0.
C           (INPUT)
C  X1MU   : REAL, VALOR DA MEDIA POPULACIONAL DA VARIAVEL EXPLICATIVA
C           CUJOS VALORES AMOSTRAIS ESTAO NA PRIMEIRA COLUNA DE X.
C           X1MU = 0.0 NAS SITUACOES INDSIT=0,1,2.
C           (INPUT)
C  EMU    : REAL, VALOR DA MEDIA DOS ERROS.
C           EMU = 0.0 NAS SITUACOES INDSIT=0,2,3.
C           (INPUT)
C  ESIG   : REAL, DESVIO PADRAO DOS ERROS.
C           ESIG = 1.0 NAS SITUACOES INDSIT=0,1,3.
C           (INPUT)
C  X      : MATRIZ REAL DE DIMENSAO 100 X 20 COM OS VALORES AMOSTRAIS
C           DAS VARIAVEIS EXPLICATIVAS.
C           (OUTPUT)
C  Y      : VETOR REAL DE DIMENSAO 100 COM OS VALORES AMOSTRAIS DA
C           VARIAVEL RESPOSTA.
C           (OUTPUT)
C
C FUNCAO E SUB-ROTINA REQUERIDAS:
C  FUNCTION RANDOM
C  SUBROUTINE SNCRAN
C
C OBSERVACOES:
C
C  O ARQUIVO DE ENTRADA DEVE TER A SEGUINTE ESTRUTURA:
C
```

```
C    LINHA 1    : IX,IY,IZ
C    LINHA 2    : NOBS, NCOREG, INTERC
C    LINHA 3    : INDSIT, EXPSIG
C    LINHA 4    : NREPL
C    LINHA 5    : DELTA
C    LINHA 6    : X1MU, EMU, ESIG
C
C
C  O ARQUIVO DE SAIDA (GENMOD.RES) TEM A SEGUINTE ESTRUTURA:
C
C    LINHA 1    : REPETICAO 1-ESIMA
C    LINHA 2    : X(1,1) ... X(1,NCOREG) Y(1)
C    LINHA 3    : X(2,1) ... X(2,NCOREG) Y(2)
C     ...                      ...
C    L. NOBS+1 : X(NOBS,1) ... X(NOBS,NCOREG) Y(NOBS)
C                (OS VALORES ANTERIORES SAO OS OBTIDOS NA PRIMEIRA
C                 REPETICAO)
C    L. NOBS+2 : REPETICAO 2-ESIMA
C       ...              ...
C    NO ARQUIVO DE SAIDA AS VARIAVEIS SAO ESCRITAS EM FORMATO
C    F10.3. NO DE ENTRADA PODEM ESTAR EM FORMATO LIVRE.
C
C-----------------------------------------------------------------------
C
      IMPLICIT REAL*4 (A-C,E-H,O-Z)
      IMPLICIT REAL*8 (D)
      REAL X(100,20), Y(100), R(19), X1(1), EPS(1)
      INTEGER INDOUT(100)
      DATA ZERO /0.0/, UM /1.0/, IUNO /1/
      DATA INDOUT /100*0/
C
C LEITURA DE VARIAVEIS
C
      READ(*,*) IX,IY,IZ
      READ(*,*) NOBS, NCOREG, INTERC
      READ(*,*) INDSIT, EXPSIG
      READ(*,*) NREPL
      READ(*,*) DELTA
      READ(*,*) X1MU, EMU, ESIG
C
      IF (INDSIT.EQ.0) GOTO 10
      IF (INDSIT-2) 2,4,6
    2 ESIG=1.0
      X1MU=0.0
      GOTO 20
    4 EMU=0.0
      X1MU=0.0
      GOTO 20
    6 EMU=0.0
      ESIG=1.0
      GOTO 20
   10 DELTA=0.0D0
      EMU=0.0
      ESIG=1.0
      X1MU=0.0
C
```

```
C CHAMA-SE A FUNCTION RANDOM 100 VEZES PARA "ESTABILIZAR" A
C GERACAO DE PSEUDO-ALEATORIOS
C
   20 DO 25 IANT=1,100
   25 H=RANDOM(IX,IY,IZ)
C
      NEXPL=NCOREG
      IF (INTERC.EQ.0) GOTO 30
      NEXPL=NEXPL-1
      DO 27 IOBS=1,NOBS
   27 X(IOBS,NCOREG)=UNO
C
C GERA O VETOR INDOUT
C  INDOUT(IOBS) = 0 INDICA QUE A OBSERVACAO IOBS NAO ESTA' CONTAMINADA
C               = 1 CASO CONTRARIO
C
   30 IF (DELTA.LE.1.0D-07) GOTO 35
      DNOBS=DBLE(NOBS)*DELTA
      DM=DNOBS+1.0D-07
      M=IDINT(DM)
      IF (M.EQ.0) GOTO 35
      IX0=IX
      IY0=IY
      IZ0=IZ
      DK=1.0D0/DELTA
      DO 32 IOBS=1,M
      I0=IOBS-1
      DI0=DBLE(I0)*DK
      H=RANDOM(IX0,IY0,IZ0)
      DJ=H*DK+1.0D0
      IJ=IDINT(DI0+DJ)
   32 INDOUT(IJ)=1
   35 OPEN (7,FILE='GENMOD.RES',STATUS='NEW')
C
C INICIO DO LACO DAS REPETICOES
C
      DO 40 IREPL=1,NREPL
      WRITE(7,1000) IREPL
 1000 FORMAT(' REPETICAO ',I4,'-ESIMA ')
      DO 41 IOBS=1,NOBS
      YVAL=0.0
C
C GERA O VALOR DA PRIMEIRA VARIAVEL EXPLICATIVA
C
      CALL SNCRAN(IX,IY,IZ,IUNO,X1,UNO,EXPSIG,NOUTI)
      YVAL=YVAL+X1(1)
      IF (INDOUT(IOBS).EQ.0) GOTO 42
      X1(1)=X1(1)+X1MU
   42 X(IOBS,1)=X1(1)
C
C GERA OS VALORES DAS DEMAIS VARIAVEIS EXPLICATIVAS
C
      NR=NEXPL-1
      CALL SNCRAN(IX,IY,IZ,NR,R,UNO,EXPSIG,NOUTI)
      DO 43 J=2,NEXPL
      J1=J-1
```

```
      X(IOBS,J)=R(J1)
   43 YVAL=YVAL+R(J1)
C
C GERA O VALOR DO ERRO
C
      CALL SNCRAN(IX,IY,IZ,IUNO,EPS,ZERO,UNO,NOUTI)
      IF (INDOUT(IOBS).EQ.0) GOTO 44
      EPS(1)=EMU+ESIG*EPS(1)
   44 YVAL=YVAL+EPS(1)
      IF (INTERC.EQ.0) GOTO 45
      YVAL=YVAL+UNO
   45 Y(IOBS)=YVAL
   41 CONTINUE
C
C GUARDA OS VALORES GERADOS NA PRESENTE REPETICAO
C
      DO 46 IOBS=1,NOBS
   46 WRITE(7,'(21(1X,F10.3,1X,:))') (X(IOBS,J),J=1,NCOREG),Y(IOBS)
C
   40 CONTINUE
      CLOSE(7,STATUS='KEEP')
      STOP
      END
C
C
            SUBROUTINE SNCRAN(IX,IY,IZ, NR, R, ETA, SIGMA, NOUTI)
C
C PROPOSITO: GERAR UM VETOR R DE DIMENSAO NR CONTENDO UMA AMOSTRA ALEAT
C            DA FUNCAO DE DISTRIBUICAO DE PROBABILIDADE DEFINIDA POR
C                 P(X<=Y)=(1-ETA)*PHIG(Y)+ETA*PHIG(Y/SIGMA)
C            ONDE PHIG E' A FUNCAO DE DISTRIBUICAO DA N(0,1),
C            ETA, E SIGMA SAO PARAMETROS DESCRITOS ABAIXO
C
C USO:
C       CALL SNCRAN(IX,IY,IZ, NR, R, ETA, SIGMA, NOUTI)
C
C ARGUMENTOS
C  IX,IY,IZ - (INPUT/OUTPUT) TRES INTEIROS NO INTERVALO  1 - 30000
C          USADOS NA GERACAO DE V.A. COM DISTRIBUICAO UNIFORME EM [0,1]
C  NR     - (INPUT) NUMERO DE OBSERVACOES A SEREM GERADAS
C  R      - (OUTPUT) VETOR DE DIMENSAO NR CONTENDO AS OBSERVACOES GERADA
C  ETA    - (INPUT)  PROPORCAO DE OUTLIERS
C           DEVE-SE TER  0 <= ETA <= 1,
C           SE ETA=0 SAO GERADAS  N(0,1) "PURAS"
C           SE ETA=1 ARE GERADAS  N(0,SIGMA) "PURAS"
C  SIGMA - (INPUT) DESVIO PADRAO DA NORMAL CONTAMINANTE
C  NOUTI - (OUTPUT) NUMERO DE OBSERVACOES CONTAMINADAS
C
C  FUNCAO REQUERIDA - FUNCTION RANDOM
C
C NOTA: A FUNCAO RANDOM (GERADOR DE V.A. UNIFORMES EM [0,1])
C       E' UTILIZADA COMO PONTO DE PARTIDA PARA O GERADOR NORMAL
C       ATRAVES DO METODO POLAR
C
C -----------------------------------------------------------------
C
```

```
C
      IMPLICIT REAL*4(A - H,O - Z)
      REAL*4 R(NR), AUX(2)
      LOGICAL GCASE
C
      DATA GCASE / .TRUE. /
C
      NOUTI = 0
C
      DO 70 J = 1, NR
        IF (GCASE) GO TO 40
        R(J) = S
        IF (ETA .LT. 1.) GO TO 20
        R(J) = SIGMA * R(J)
        GO TO 30
   20   IF (ETA .LE. 0.) GO TO 30
        IX0=IX
        IY0=IY
        IZ0=IZ
        AUXV=RANDOM(IX0,IY0,IZ0)
        IF (AUXV .GT. ETA) GO TO 30
        R(J) = SIGMA * R(J)
        NOUTI = NOUTI + 1
   30   GCASE = .TRUE.
        GO TO 70
   40   AUX(1)=RANDOM(IX,IY,IZ)
        AUX(2)=RANDOM(IX,IY,IZ)
        U1 = 2. * AUX(1) - 1.
        U2 = 2. * AUX(2) - 1.
        S = U1 * U1 + U2 * U2
        IF (S .GT. 1.) GO TO 40
        S = SQRT(-2.*ALOG(S)/S)
        R(J) = U1 * S
        S = U2 * S
        IF (ETA .LT. 1.) GO TO 50
        R(J) = SIGMA * R(J)
        GO TO 60
   50   IF (ETA .LE. 0.) GO TO 60
        IX0=IX
        IY0=IY
        IZ0=IZ
        AUXV=RANDOM(IX0,IY0,IZ0)
        IF (AUXV .GT. ETA) GO TO 60
        R(J) = SIGMA * R(J)
        NOUTI = NOUTI + 1
   60   GCASE = .FALSE.
   70 CONTINUE
C
      RETURN
      END
C
C
            REAL FUNCTION RANDOM(IX,IY,IZ)
C
C PROPOSITO : GERAR NUMEROS PSEUDO-ALEATORIOS SEGUNDO A DISTRIBUICAO
C             UNIFORME EM [0,1]
```

```
C
C USO :  U = RANDOM(IX,IY,IZ)
C
C ARGUMENTOS :
C    IX,IY,IZ : (INPUT/OUTPUT) TRES INTEIROS NO INTERVALO  1 - 30000,
C               PREFERIVELMENTE ESCOLHIDOS DE FORMA ALEATORIA
C
C REFERENCIA : WICHMANN,B.A. E  HILL,I.D. (1982)
C              AN EFFICIENT AND PORTABLE PSEUDO-RANDOM NUMBER GENERATOR
C              ALGORITHM AS 183 - APPLIED STATISTICS 31 (2) - 188 TO 19
C
C-------------------------------------------------------------------------
C
C
      IX=171*MOD(IX,177)-2*(IX/177)
      IY=172*MOD(IY,176)-35*(IY/176)
      IZ=170*MOD(IZ,178)-63*(IZ/178)
      IF (IX.LT.0) IX=IX+30269
         IF (IY.LT.0) IY=IY+30307
            IF (IZ.LT.0) IZ=IZ+30323
            RANDOM=AMOD(FLOAT(IX)/30269.0+FLOAT(IY)/30307.0+
     *                  FLOAT(IZ)/30323.0,1.0)
            IF (RANDOM.GT.0.0) RETURN
                RANDOM=DMOD(DBLE(FLOAT(IX))/30269.0D0+DBLE(FLOAT(IY))/
     *                  30307.0D0+DBLE(FLOAT(IZ))/30323.0D0,1.0D0)
                IF (RANDOM.GE.1.0) RANDOM=.999999
C
                   RETURN
      END
```

A N E X O 1

T A B E L A S

Resultados do Estudo Monte Carlo - Modelo 20100

($n = 20$, $p = 1$, sem contaminação,sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99734	0.99913	1.00746	0.99947	0.99773	0.99933	0.99771	0.99788	0.99760	0.99722	0.99738
β	0.00051	0.00091	0.02394	0.00257	0.00055	0.00163	0.00054	0.00107	0.00056	0.00077	0.00054
	0.00052	0.00092	0.02400	0.00257	0.00056	0.00163	0.00054	0.00107	0.00056	0.00078	0.00054
	0.00053	0.00077	0.00060	0.00190	0.00060	0.00109	0.00055	0.00096	0.00060	0.00066	0.00055
	0.97707	0.90419	0.95393	0.98280	0.96425	0.83809	0.97453	0.91070	0.96496	0.88324	0.97311
σ	0.02594	0.06615	0.06679	0.05435	0.02784	0.03905	0.03027	0.03432	0.02760	0.02875	0.02836
	0.02647	0.07532	0.06891	0.05465	0.02912	0.06617	0.03092	0.04229	0.02882	0.04238	0.02908
	0.02595	0.07919	0.05918	0.06331	0.02860	0.06286	0.02772	0.04717	0.02792	0.04237	0.02727

Resultados do Estudo Monte Carlo - Modelo 20101

($n = 20$, $p = 1$, contaminação na média dos erros, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99648	0.99565	0.99625	1.00825	0.99544	0.99060	0.99690	0.98360	0.99366	0.99795	1.00096
β	0.01316	0.00196	0.02128	0.07840	0.01483	0.00940	0.00797	0.04013	**0.01483**	0.00478	0.01073
	0.01318	0.00198	0.02142	0.07847	0.01485	0.00949	0.00798	0.04040	**0.01487**	0.00478	0.01074
	0.01208	0.00136	0.00187	0.00737	0.01236	0.00211	0.00156	0.05618	0.01245	0.00410	0.01098
	4.58168	1.25633	1.37030	2.63424	4.50880	2.43319	2.11110	2.81154	4.52637	3.57394	3.47649
σ	0.08855	0.12216	0.45720	4.34783	0.21457	2.41084	4.32117	0.65099	0.14242	0.58610	0.29681
	12.91698	0.18786	0.59432	7.01858	12.52627	4.46488	5.55622	3.93268	12.57771	7.21129	6.42984
	13.00440	0.16491	0.18592	0.78440	12.87810	8.44341	0.10235	3.94198	12.87810	7.78065	6.29942

Resultados do Estudo Monte Carlo - Modelo 20102

($n = 20$, $p = 1$, contaminação na variância dos erros, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99767	0.99798	1.00408	0.99478	0.99732	0.99411	0.99430	1.00825	1.00463	0.99695	0.99571
β	0.01070	0.00128	0.01293	0.00867	0.00229	0.00210	0.00504	0.00977	0.00554	0.00226	0.00910
	0.01071	0.00129	0.01294	0.00870	0.00230	0.00214	0.00508	0.00984	0.00556	0.00227	0.00912
	0.01046	0.00108	0.00198	0.00265	0.00112	0.00188	0.00160	0.00853	0.00368	0.00198	0.00491
	4.32282	1.17837	1.64422	1.14946	1.63077	1.91013	2.02540	2.01528	2.82290	2.70554	3.05708
σ	1.84810	0.10262	22.58569	0.37550	1.37898	2.14775	2.41915	0.74571	1.22353	1.79091	2.06398
	12.88922	0.13444	23.00070	0.39784	1.77684	2.97609	3.47059	1.77650	4.54649	4.69978	6.29557
	12.86960	0.11466	0.16946	0.12028	0.20568	0.65805	0.53708	1.47089	3.66465	4.09414	5.23632

Resultados do Estudo Monte Carlo - Modelo 20103

($n = 20$, $p = 1$, contaminação na direção de X, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.04140	0.02772	-0.11909	0.99076	0.04252	0.95377	0.04162	0.51745	0.04209	0.12738	0.04136
β	0.00204	0.00312	4.88856	0.00641	0.00228	0.04209	0.00220	0.24034	0.00224	0.02652	0.00217
	0.92095	0.94845	6.14091	0.00649	0.91904	0.04423	0.92070	0.47319	0.91982	0.78799	0.92116
	0.91651	0.93764	0.91514	0.00197	0.91530	0.00206	0.91611	0.03490	0.91518	0.81267	0.91667
	9.53480	8.68333	9.71900	1.00061	9.44847	1.13851	9.53876	5.05437	9.45900	8.48085	9.52863
σ	2.73136	5.71140	58.65979	0.45423	2.97511	2.14925	2.82780	19.69237	2.89857	3.93279	2.76416
	75.57417	64.74491	134.68018	0.45423	74.35181	2.16843	75.73824	36.13030	74.45333	59.89590	75.50177
	74.72540	63.12930	71.48780	0.06902	74.13720	0.08754	74.69660	1.25972	74.13720	60.15450	74.75720

Resultados do Estudo Monte Carlo - Modelo 20200

($n = 20,\ p = 2$,sem contaminação, sem intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99682	0.99389	0.99698	0.99553	0.99714	0.99685	0.99698	0.99594	0.99716
β_1	0.00055	0.00268	0.00055	0.00157	0.00059	0.00086	0.00055	0.00083	0.00059
	0.00056	0.00272	0.00056	0.00159	0.00060	0.00087	0.00056	0.00085	0.00059
	0.00041	0.00121	0.00041	0.00111	0.00043	0.00069	0.00041	0.00079	0.00041
	0.99854	0.99828	0.99867	0.99821	0.99849	1.00023	0.99867	0.99783	0.99854
β_2	0.00052	0.00225	0.00054	0.00170	0.00054	0.00086	0.00054	0.00079	0.00054
	0.00053	0.00225	0.00054	0.00170	0.00054	0.00086	0.00054	0.00080	0.00054
	0.00048	0.00141	0.00052	0.00125	0.00056	0.00091	0.00052	0.00096	0.00055
	0.97655	1.01906	0.97110	0.74904	0.97690	0.93406	0.97111	0.84733	0.97577
σ	0.01965	0.05552	0.02008	0.03040	0.02141	0.02498	0.02008	0.02319	0.02076
	0.02020	0.05589	0.02091	0.10193	0.02194	0.02933	0.02091	0.04650	0.02135
	0.02301	0.05386	0.02300	0.10813	0.02490	0.03253	0.02300	0.05111	0.02518

Resultados do Estudo Monte Carlo - Modelo 20201

($n = 20$, $p = 2$,contaminação na média dos erros, sem intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.98700	0.96794	0.98442	0.99754	0.99273	0.98190	0.98442	0.99759	0.99585
β_1	0.01217	0.04529	0.01239	0.00289	0.00551	0.02953	0.01239	0.00566	0.01002
	0.01234	0.04631	0.01263	0.00290	0.00557	0.02985	0.01263	0.00566	0.01003
	0.01254	0.00962	0.01223	0.00173	0.00125	0.04682	0.01223	0.00403	0.00703
	1.00491	1.00465	1.00225	0.99956	1.00503	1.00802	1.00167	1.00931	1.00711
β_2	0.01181	0.04036	0.01265	0.00279	0.00553	0.03148	0.01275	0.00595	0.00915
	0.01183	0.04038	0.01266	0.00279	0.00556	0.03155	0.01275	0.00604	0.00920
	0.01088	0.01060	0.01206	0.00237	0.00143	0.03513	0.01219	0.00451	0.00966
	4.59189	3.23587	4.54869	1.96211	1.99399	3.27684	4.55349	3.47090	3.61810
σ	0.07958	5.15330	0.12299	2.35336	3.76918	0.32295	0.11078	0.66459	0.38325
	12.98122	10.15239	12.71621	3.27902	4.75720	5.50694	12.73808	6.81451	7.23771
	13.05780	15.13020	12.99800	0.44515	0.06475	5.84719	12.99800	7.38567	7.82645

Resultados do Estudo Monte Carlo - Modelo 20202

($n = 20$, $p = 2$, contaminação na variância dos erros, sem intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.98591	0.99357	1.00262	0.99441	0.99427	0.99475	0.99719	0.99777	0.99714
β_1	0.01120	0.00788	0.00290	0.00229	0.00467	0.01377	0.00541	0.00349	0.00825
	0.01140	0.00792	0.00290	0.00232	0.00470	0.01379	0.00542	0.00349	0.00826
	0.01008	0.00205	0.00161	0.00136	0.00241	0.00719	0.00361	0.00217	0.00432
	0.98389	0.99125	0.99585	0.99364	0.99102	0.98801	0.99563	0.99166	0.98960
β_2	0.01392	0.01061	0.00418	0.00339	0.00716	0.01124	0.00667	0.00485	0.01086
	0.01418	0.01069	0.00419	0.00343	0.00724	0.01139	0.00669	0.00491	0.01097
	0.00817	0.00248	0.00148	0.00223	0.00217	0.00837	0.00309	0.00273	0.00573
	4.39610	1.34278	1.90344	1.44717	2.15870	2.19154	2.92690	2.52941	2.97130
σ	1.99240	0.75565	1.89924	1.85769	2.45869	0.90215	1.38837	1.75689	1.80775
	13.52589	0.87314	2.71544	2.05765	3.80127	2.32191	5.10132	4.09600	5.69378
	13.5979	0.13014	0.52469	0.23044	1.31119	1.82817	4.20119	3.31648	4.69111

Resultados do Estudo Monte Carlo - Modelo 20203

($n = 20$, $p = 2$,contaminação na direção de X,sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.03810	0.02743	0.04836	0.98175	0.03791	0.72702	0.04403	0.24526	0.03791	0.07005	0.03895
β_1	0.00228	0.00346	0.01665	0.01291	0.00235	0.18539	0.00659	0.17077	0.00235	0.00352	0.00232
	0.92753	0.94936	0.92226	0.01324	0.92797	0.25990	0.92047	0.74041	0.92797	0.86833	0.92594
	0.92353	0.93979	0.92198	0.00110	0.92360	0.00577	0.92158	0.91172	0.92360	0.86604	0.92332
	0.98615	0.97819	0.98845	1.01502	0.98558	0.97419	0.98415	0.97761	0.98558	0.99376	0.98817
β_2	0.05125	0.08451	0.22481	0.04447	0.05623	0.05781	0.05308	0.08911	0.05623	0.08019	0.05472
	0.05144	0.08498	0.22495	0.04470	0.05644	0.05847	0.05333	0.08962	0.05644	0.08023	0.05486
	0.04708	0.08167	0.07021	0.00137	0.05413	0.00356	0.05052	0.04018	0.05413	0.05994	0.05238
	9.65771	8.65773	8.88045	1.09295	9.57421	2.80368	9.60399	7.43404	9.57421	8.42598	9.62861
σ	2.60411	5.83091	6.29990	0.60092	2.85390	11.92166	3.52691	15.52425	2.85390	2.50149	2.71712
	77.56003	64.47175	68.40137	0.60956	76.37102	15.17493	77.55557	56.92112	76.37017	57.64662	77.17000
	76.05350	64.30900	70.46700	0.04207	76.12110	0.20269	75.85530	67.20030	76.12110	57.37500	75.10820

Resultados do Estudo Monte Carlo - Modelo 20210

($n = 20$, $p = 2$, sem contaminação, com intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99772	0.99931	0.99739	0.99935	0.99732	0.99899	0.99752	0.99802	0.99796
β_1	0.00056	0.00242	0.00058	0.00212	0.00061	0.00092	0.00058	0.00075	0.00060
	0.00057	0.00242	0.00058	0.00212	0.00062	0.00092	0.00058	0.00076	0.00060
	0.00057	0.00125	0.00060	0.00149	0.00065	0.00107	0.00061	0.00072	0.00064
	1.01192	0.99810	1.00993	0.98491	1.01297	0.99112	1.01046	1.00012	1.01340
constante	0.04893	0.15071	0.05113	0.09318	0.05448	0.05994	0.05148	0.05982	0.05121
	0.04907	0.15071	0.05123	0.09341	0.05465	0.06002	0.05159	0.05982	0.05139
	0.04806	0.07575	0.04900	0.07067	0.04303	0.06473	0.04990	0.06079	0.04322
	0.97799	0.97238	0.96790	0.71175	0.97799	0.92116	0.96884	0.84753	0.97726
σ	0.02785	0.06390	0.03114	0.07189	0.03188	0.03646	0.03078	0.03172	0.02954
	0.02833	0.06466	0.03217	0.15498	0.03237	0.04268	0.03175	0.05497	0.03006
	0.02940	0.07681	0.03200	0.17317	0.02954	0.05248	0.03139	0.06125	0.02836

Resultados do Estudo Monte Carlo - Modelo 20211

($n = 20$, $p = 2$,contaminação na média dos erros, com intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00220	1.01752	1.00205	1.00254	0.99986	0.99215	1.00205	1.00342	1.00208
β_1	0.01203	0.04543	0.01305	0.00562	0.00824	0.03336	0.01305	0.00937	0.01698
	0.01204	0.04573	0.01306	0.00562	0.00824	0.03342	0.01306	0.00938	0.01698
	0.01065	0.01518	0.01071	0.00236	0.00225	0.03621	0.01071	0.00616	0.01674
	3.01688	2.77086	3.01444	1.20020	2.20444	2.58391	3.01443	2.25219	2.70180
constante	0.09561	1.91988	0.09879	0.59847	2.30674	0.30578	0.09879	0.28132	0.32132
	4.16341	5.05583	4.15675	0.63855	3.75741	2.81456	4.15675	1.84930	3.21743
	4.20376	6.40965	4.20137	0.14090	0.99467	2.85719	4.20137	1.76165	2.05528
	4.21932	3.69119	4.21115	1.04244	2.83417	3.81390	4.21115	3.33516	4.44286
σ	0.08303	3.29969	0.09143	1.13696	4.43119	0.25872	0.09143	0.46431	0.26223
	10.44703	10.54218	10.40289	1.13876	7.70538	8.17676	10.40289	5.91729	12.11552
	10.55060	14.04930	10.54740	0.18937	1.04818	8.41245	10.54740	6.18272	12.40250

Resultados do Estudo Monte Carlo - Modelo 20212

($n = 20$, $p = 2$,com contaminação na variância dos erros, com intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99559	0.99331	0.99856	0.99730	0.99049	0.99548	1.00174	0.99312	1.00124
β_1	0.01099	0.00421	0.00337	0.00275	0.00983	0.00997	0.00609	0.00257	0.02578
	0.01101	0.00426	0.00337	0.00276	0.00992	0.00999	0.00609	0.00262	0.02578
	0.00958	0.00237	0.00184	0.00180	0.00265	0.00719	0.00435	0.00161	0.01385
	1.08555	0.96663	1.04794	0.97768	0.98456	1.00030	1.07881	0.99365	1.02321
constante	0.97127	0.18479	0.37956	0.12266	0.86237	0.49874	0.56576	0.31781	1.01979
	0.97859	0.18591	0.38186	0.12316	0.86261	0.49874	0.57197	0.31785	1.02032
	1.10680	0.12034	0.23089	0.09391	0.28199	0.40566	0.70247	0.33971	0.84847
	4.33305	1.08404	2.04821	0.70776	2.37778	2.14360	2.95910	1.93736	3.57764
σ	1.88465	0.15785	2.09542	0.07532	4.23833	0.83468	1.51047	0.70358	3.09787
	12.99386	0.16491	3.19416	0.16073	6.13662	2.14251	5.34855	1.58223	9.74208
	12.78220	0.08942	0.75041	0.19499	0.84807	1.74396	4.13728	1.37702	7.51212

Resultados do Estudo Monte Carlo - Modelo 20213

($n = 20$, $p = 2$,com contaminação na direção de X, com intercepto)

Parâmetros	LS	ONE AT A TIME				SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.05122	0.98140	0.05088	0.62443	0.05678	0.14527	0.05088	0.07053	0.05174
β_1	0.00244	0.02073	0.00253	0.22217	0.00706	0.10259	0.00253	0.00301	0.00244
	0.90263	0.02108	0.90337	0.36322	0.89673	0.83315	0.90337	0.86782	0.90164
	0.90058	0.00169	0.89823	0.01963	0.89750	0.91922	0.89823	0.86680	0.89823
	-0.27278	0.97041	-0.26827	1.00830	- 0.29009	-0.05179	-0.26827	0.41609	-0.27708
constante	5.56110	0.35547	5.56683	3.83263	5.85233	6.85885	5.56683	6.64156	5.54101
	7.18108	0.35634	7.17533	3.83269	7.51809	7.96510	7.17533	6.98251	7.17194
	7.72616	0.08594	7.72616	0.39728	7.50661	7.42589	7.72616	7.07866	7.60942
	9.45149	1.12931	9.42821	3.14637	9.45708	8.10823	9.42821	8.14590	9.48005
σ	2.86966	1.37631	2.92786	11.40243	3.48376	10.10579	2.92786	2.84073	2.87266
	74.29683	1.39303	73.96264	16.00934	75.00598	60.63276	73.96264	53.90463	74.78387
	75.57270	0.07070	75.12940	0.57015	74.84450	64.31440	75.12940	54.22940	75.28100

Resultados do Estudo Monte Carlo - Modelo 50103

($n = 50$, $p = 1$, contaminação na direção de X, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.04351	0.03974	0.04501	1.00043	0.04351	1.00179	0.04268	0.21001	0.04351	0.09031	0.04315
β	0.00085	0.00139	0.00154	0.00065	0.00085	0.00081	0.00117	0.10471	0.00085	0.00140	0.00096
	0.91572	0.92350	0.91355	0.00065	0.91572	0.00082	0.91763	0.72880	0.91572	0.82893	0.91653
	0.91494	0.92485	0.91457	0.00053	0.91494	0.00075	0.91482	0.85119	0.91494	0.83365	0.91511
	9.67241	9.33918	9.57239	0.97997	9.67241	0.85335	9.80379	7.96485	9.67241	8.83958	9.65182
σ	0.87034	2.02937	2.10666	0.02537	0.87034	0.02105	1.96127	8.52960	0.87034	0.78342	1.13212
	76.08097	71.57126	75.59259	0.02577	76.08097	0.04255	79.46804	57.03871	76.08097	62.24246	75.98610
	75.42920	70.36610	75.05860	0.02655	75.42920	0.04543	75.54280	63.03860	75.42920	62.53320	76.04740

Resultados do Estudo Monte Carlo - Modelo 50102

($n = 50$, $p = 1$, contaminação na variância dos erros, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99895	1.00065	1.00323	1.00033	1.00022	1.00141	1.00269	1.00138	1.00061	1.00207	1.00580
β	0.00418	0.00042	0.00919	0.00115	0.00239	0.00060	0.00057	0.00561	0.00342	0.00089	0.00642
	0.00418	0.00042	0.00920	0.00115	0.00239	0.00061	0.00058	0.00562	0.00342	0.00089	0.00646
	0.00412	0.00042	0.00089	0.00078	0.00136	0.00051	0.00031	0.00374	0.00302	0.00089	0.00481
	4.49777	1.24163	1.27677	1.03549	3.07203	2.47352	1.34278	2.27589	3.73717	3.17523	2.93097
σ	0.99332	0.04358	0.05410	0.03266	1.78192	1.65368	0.98003	0.41220	0.89755	1.04189	0.88225
	13.22774	0.10196	0.13070	0.03392	6.07524	3.82494	1.09753	2.04011	8.38968	5.77352	4.61090
	13.13760	0.09563	0.11850	0.03317	6.16400	3.80599	0.04672	1.72711	8.28353	5.24590	3.99513

Resultados do Estudo Monte Carlo - Modelo 50101

($n = 50$, $p = 1$, contaminação na média dos erros, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99766	1.00123	1.00275	1.00297	0.99766	1.00244	0.99608	0.99070	0.99765	1.00362	0.99780
β	0.00445	0.00049	0.00167	0.01728	0.00445	0.00083	0.01257	0.02049	0.00445	0.00146	0.00261
	0.00446	0.00050	0.00168	0.01729	0.00446	0.00084	0.01259	0.02058	0.00446	0.00147	0.00261
	0.00511	0.00050	0.00058	0.00160	0.00511	0.00076	0.01413	0.02807	0.00511	0.00165	0.00304
	4.57380	1.34680	1.34473	2.18205	4.57380	3.16197	6.21090	3.06250	4.57380	3.76248	2.95976
σ	0.02324	0.05051	0.04159	3.67825	0.02324	0.80675	0.48743	0.24096	0.02324	0.15500	0.07601
	12.79529	0.17078	0.16043	5.07549	12.79529	5.48088	27.64090	4.49485	12.79529	7.78631	3.91665
	12.76730	0.13828	0.15790	0.11665	12.76730	6.21739	28.11040	4.44680	12.76730	7.90706	3.92996

Resultados do Estudo Monte Carlo - Modelo 50100

($n = 50$, $p = 1$, sem contaminação, sem intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME				SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	C.N. + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00067	1.00026	1.00079	1.00089	1.00067	1.00130	1.00068	1.00099	1.00067	1.00014	1.00081
β	0.00021	0.00030	0.00022	0.00099	0.00021	0.00065	0.00022	0.00033	0.00021	0.00027	0.00022
	0.00021	0.00030	0.00022	0.00099	0.00021	0.00065	0.00022	0.00033	0.00021	0.00027	0.00022
	0.00018	0.00027	0.00018	0.00072	0.00018	0.00047	0.00019	0.00029	0.00018	0.00024	0.00019
σ	0.99454	0.98255	0.99507	0.99146	0.99454	0.86312	0.99196	0.92969	0.99454	0.91115	0.98283
	0.01018	0.02995	0.02789	0.02359	0.01018	0.01724	0.01454	**0.01404**	0.01018	0.01314	0.01187
	0.01021	0.03025	0.02791	0.02367	0.01021	0.03598	0.01461	0.01898	0.01021	0.02104	0.01217
	0.01003	0.02809	0.02490	0.02436	0.01003	0.03216	0.01245	0.01896	0.01003	0.02176	0.01134

Resultados do Estudo Monte Carlo - Modelo 50200

($n = 50$, $p = 2$, sem contaminação, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99872	0.99908	0.99872	0.99913	0.99708	0.99872	0.99916	0.99846
β_1	0.00020	0.00039	0.00020	0.00059	0.00028	0.00020	0.00027	0.00021
	0.00020	0.00039	0.00020	0.00059	0.00029	0.00020	0.00027	0.00021
	0.00017	0.00025	0.00017	0.00054	0.00031	0.00017	0.00026	0.00018
	0.99989	1.00021	0.99989	1.00063	1.00075	0.99989	1.00086	0.99961
β_2	0.00019	0.00036	0.00019	0.00052	0.00029	0.00019	0.00027	0.00020
	0.00019	0.00036	0.00019	0.00052	0.00029	0.00019	0.00027	0.00020
	0.00018	0.00037	0.00018	0.00047	0.00026	0.00018	0.00028	0.00020
	0.99641	0.98059	0.99641	0.82230	0.94250	0.99641	0.88653	0.98733
σ	0.00977	0.02006	0.00977	0.01648	0.01260	0.00977	0.01036	0.01076
	0.00978	0.02044	0.00978	0.04806	0.01591	0.00978	0.02324	0.01092
	0.00848	0.02232	0.00848	0.04370	0.01545	0.00848	0.02385	0.00918

Resultados do Estudo Monte Carlo - Modelo 50201

($n = 20$, $p = 2$,contaminação na média dos erros, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99688	0.99708	0.99688	1.00075	0.99421	0.99688	1.00101	0.99726
β_1	0.00383	0.00283	0.00383	0.00084	0.01381	0.00383	0.00185	0.00248
	0.00384	0.00284	0.00384	0.00084	0.01385	0.00384	0.00185	0.00249
	0.00444	0.00068	0.00444	0.00043	0.01830	0.00444	0.00152	0.00224
	1.00710	0.99998	1.00710	1.00004	1.01599	1.00710	1.00077	1.00253
β_2	0.00486	0.00166	0.00486	0.00072	0.01465	0.00486	0.00182	0.00251
	0.00491	0.00166	0.00491	0.00072	0.01491	0.00491	0.00182	0.00251
	0.00516	0.00054	0.00516	0.00068	0.01717	0.00516	0.00135	0.00194
	4.56940	1.45431	4.56940	1.88036	3.34371	4.56940	3.54329	2.95721
σ	0.02870	2.03117	0.02870	1.51797	0.16351	0.02870	0.20176	0.11037
	12.76930	2.23756	12.76930	2.30803	5.65649	12.76930	6.67009	3.94102
	12.71750	0.03728	12.71750	0.55881	5.75537	12.71750	6.83698	3.90551

Resultados do Estudo Monte Carlo - Modelo 50202

($n = 20$, $p = 2$,contaminação na variância dos erros, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
β_1	0.99688	0.99746	0.99095	0.99854	0.98841	0.99048	0.99065	0.99518
	0.00411	0.00062	0.00367	0.00070	0.00440	0.00389	0.00110	0.00452
	0.00412	0.00063	0.00367	0.00070	0.00454	0.00389	0.00111	0.00454
	0.00374	0.00052	0.00327	0.00045	0.00436	0.00377	0.00101	0.00423
β_2	0.99862	0.99998	0.99745	1.00051	0.99421	0.99502	0.99802	1.00126
	0.00429	0.00055	0.00328	0.00064	0.00361	0.00369	0.00103	0.00464
	0.00429	0.00055	0.00329	0.00064	0.00364	0.00371	0.00104	0.00464
	0.00389	0.00050	0.00273	0.00054	0.00275	0.00329	0.00103	0.00374
σ	4.49400	1.02759	3.68153	1.35830	2.34452	4.00320	2.69565	2.88914
	0.94468	0.03853	1.73726	0.85645	0.42082	1.08329	0.73760	0.73746
	13.15269	0.03929	8.92786	0.98483	2.22857	10.10248	3.61283	4.30629
	12.53110	0.03435	9.57156	0.13678	1.86237	10.49120	2.58971	3.95066

Resultados do Estudo Monte Carlo - Modelo 50203

($n = 20$, $p = 2$,contaminação na direção de X, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.04369	0.96661	0.04369	0.86735	0.06233	0.04369	0.06946	0.04346
β_1	0.00065	0.02792	0.00065	0.10723	0.00961	0.00065	0.00095	0.00070
	0.91518	0.02904	0.91518	0.12482	0.88884	0.91518	0.86686	0.91566
	0.91039	0.00052	0.91039	0.00087	0.91924	0.91039	0.86343	0.91036
	0.99658	0.99879	0.99658	1.00482	0.99666	0.99658	0.99922	0.99377
β_2	0.02064	0.00150	0.02064	0.00576	0.02684	0.02064	0.02514	0.02204
	0.02066	0.00151	0.02066	0.00578	0.02685	0.02066	0.02514	0.02207
	0.01851	0.00055	0.01851	0.00084	0.02328	0.01851	0.01968	0.01688
	9.75030	1.33469	9.75030	1.79072	9.21943	9.75030	8.59075	9.73932
σ	1.05881	3.85232	1.05881	6.24068	1.69274	1.05881	1.06825	1.15252
	77.62660	3.96434	77.62660	6.86592	69.25180	77.62660	58.82430	77.52829
	78.32690	0.02904	78.32690	0.06679	68.08090	78.32700	58.00140	77.79900

Resultados do Estudo Monte Carlo - Modelo 50210

($n = 50$, $p = 2$, sem contaminação, com intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME			SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00102	1.00080	1.00109	1.00140	1.00102	1.00058	1.00186	1.00102	1.00137	1.00102
β	0.00022	0.00030	0.00022	0.00032	0.00022	0.00062	0.00028	0.00022	0.00026	0.00023
	0.00022	0.00030	0.00022	0.00032	0.00022	0.00062	0.00028	0.00022	0.00026	0.00024
	0.00021	0.00027	0.00020	0.00037	0.00021	0.00048	0.00025	0.00021	0.00024	0.00021
	0.99522	0.98900	0.99255	0.99161	0.99522	0.98917	0.98802	0.99522	0.98799	0.99117
constante	0.01895	0.02815	0.01909	0.02371	0.01895	0.02731	0.02144	0.01895	0.02333	0.01948
	0.01897	0.02827	0.02004	0.02378	0.01897	0.02743	0.02158	0.01897	0.02347	0.01956
	0.01930	0.02959	0.01950	0.02302	0.01930	0.02395	0.01674	0.01930	0.02513	0.02071
	0.99519	0.98389	0.98403	0.93400	0.99519	0.80028	0.92883	0.99519	0.88521	0.98879
σ	0.01039	0.03140	0.02318	0.02266	0.01039	0.02606	0.01336	0.01039	0.01337	0.01183
	0.01042	0.03166	0.02344	0.02701	0.01042	0.06595	0.01842	0.01042	0.02655	0.01196
	0.01015	0.03303	0.02261	0.02471	0.01015	0.06283	0.02060	0.01015	0.02794	0.01130

Resultados do Estudo Monte Carlo - Modelo 50211

($n = 50$, $p = 2$, contaminação na média dos erros, com intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME			SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00141	1.00129	1.00174	0.99902	1.00141	1.00376	0.99846	1.00141	1.00524	0.98974
β	0.00335	0.00056	0.00182	0.00301	0.00335	0.00110	0.00769	0.00335	0.00260	0.00478
	0.00335	0.00057	0.00182	0.00301	0.00335	0.00112	0.00769	0.00335	0.00263	0.00489
	0.00289	0.00055	0.00096	0.00172	0.00289	0.00091	0.00900	0.00289	0.00208	0.00457
	2.98584	1.32728	1.04663	2.29573	2.98584	1.13451	2.68659	2.98583	2.40224	2.27368
constante	0.02623	0.03966	0.16905	1.50978	0.02623	0.27963	0.07762	0.02623	0.09651	0.11504
	3.96978	0.14677	0.17213	3.18871	3.96978	0.20772	2.92219	3.96978	2.06277	1.73729
	4.01545	0.14091	0.09204	5.80465	4.01545	0.04730	2.94716	4.01545	2.04116	1.72728
	4.16070	1.36270	1.36072	2.88853	4.16070	1.03432	3.87233	4.16070	3.58450	3.95681
σ	0.02144	0.05288	0.23310	3.18460	0.02144	0.75395	0.07075	0.02144	0.10908	0.10568
	10.01146	0.18444	0.36322	6.75113	10.01146	0.75513	8.32102	10.01146	6.78873	8.84842
	10.10150	0.16115	0.15479	10.96280	10.10150	0.06231	8.31422	10.10150	6.79591	8.57236

Resultados do Estudo Monte Carlo - Modelo 50212

($n = 50$, $p = 2$, contaminação na variância dos erros, com intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME			SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99838	1.00109	0.99375	1.00277	0.99960	1.00158	1.00234	0.99918	1.00200	1.00643
β	0.00426	0.00044	0.00891	0.00057	0.00300	0.00072	0.00335	0.00370	0.00082	0.01614
	0.00426	0.00044	0.00895	0.00058	0.00305	0.00072	0.00335	0.00370	0.00083	0.01618
	0.00334	0.00044	0.00126	0.00054	0.00219	0.00069	0.00268	0.00285	0.00085	0.00983
	0.97849	0.97995	0.99325	0.98474	1.01444	0.98618	0.99184	1.00239	0.99885	1.04106
constante	0.40817	0.04531	0.15699	0.03550	0.28862	0.04224	0.15389	0.32984	0.13191	0.61970
	0.40863	0.04571	0.15704	0.03573	0.28882	0.04243	0.15396	0.32984	0.13191	0.62138
	0.35237	0.05361	0.11600	0.03097	0.21239	0.03820	0.11418	0.24820	0.13560	0.57401
	4.50042	1.25244	1.33641	0.98070	3.56093	0.82571	2.30629	3.90047	2.17435	3.55357
σ	0.99455	0.04356	0.07345	0.03059	1.58982	0.03577	0.41152	0.95864	0.35229	1.34184
	13.24753	0.10729	0.18662	0.03096	8.14817	0.06615	2.11792	9.37137	1.73139	7.86259
	13.34150	0.10430	0.13456	0.02842	7.80937	0.06511	1.92520	9.02350	1.62788	6.62960

Resultados do Estudo Monte Carlo - Modelo 50213

($n = 50$, $p = 2$, contaminação na direção de X, com intercepto)

Parâmetros	LS	L1	TAU	ONE AT A TIME			SEVERAL AT A TIME			
				COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.05473	0.04881	0.04701	0.97883	0.05473	0.75541	0.05621	0.05473	0.06658	0.05440
β	0.00091	0.00155	0.00732	0.01965	0.00091	0.17444	0.00801	0.00091	0.00126	0.00096
	0.89445	0.90632	0.91551	0.02010	0.89445	0.23427	0.89875	0.89445	0.87254	0.89512
	0.89077	0.90504	0.89747	0.00037	0.89077	0.00253	0.91772	0.89077	0.87065	0.89156
	-0.26221	-0.08990	-0.23787	0.95602	-0.26221	0.89070	-0.33234	-0.26221	0.28070	-0.20350
constante	2.01012	3.30611	2.33432	0.07662	2.01012	0.87973	2.48462	2.01012	2.27119	2.21446
	3.60330	4.49398	3.86665	0.07855	3.60330	0.89167	4.25975	3.60330	2.78858	3.66287
	3.44499	5.39081	4.03562	0.03153	3.44499	0.07457	4.19224	3.44499	3.20207	3.81553
	9.61280	9.27668	9.42813	1.18447	9.61280	2.64801	9.16311	9.61280	8.49281	9.57366
σ	8.86670	2.09714	1.74594	2.49316	0.86670	10.18094	1.52207	0.86670	0.91776	0.95399
	75.04701	70.60054	72.77937	2.52719	75.04701	12.89089	68.15846	75.04701	57.05099	74.46164
	74.1559	69.5455	72.5356	0.02590	74.15590	0.08040	68.48580	74.15590	56.25300	75.11410

Resultados do Estudo Monte Carlo - Modelo 50800

($n = 50$, $p = 8$, sem contaminação, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00004	0.99955	1.00004	1.00081	0.99972	1.00004	1.00010	0.99995
β_1	0.00019	0.00032	0.00019	0.00076	0.00030	0.00019	0.00030	0.00020
	0.00019	0.00032	0.00019	0.00076	0.00030	0.00019	0.00030	0.00020
	1.00011	1.00033	1.00011	1.00029	0.99987	1.00011	1.00026	1.00060
β_2	0.00026	0.00044	0.00026	0.00067	0.00038	0.00026	0.00034	0.00027
	0.00026	0.00044	0.00026	0.00067	0.00038	0.00026	0.00034	0.00027
	0.99918	0.99888	0.99918	0.99972	0.99893	0.99918	0.99979	0.99888
β_3	0.00025	0.00045	0.00025	0.00077	0.00035	0.00025	0.00035	0.00027
	0.00025	0.00045	0.00025	0.00077	0.00035	0.00025	0.00035	0.00027
	1.00016	1.00029	1.00016	1.00149	1.00013	1.00016	1.00061	1.00017
β_4	0.00026	0.00049	0.00026	0.00077	0.00037	0.00026	0.00035	0.00026
	0.00026	0.00049	0.00026	0.00077	0.00037	0.00026	0.00035	0.00026
	1.00038	1.00014	1.00038	1.00314	1.00217	1.00038	1.00130	1.00063
β_5	0.00019	0.00036	0.00019	0.00081	0.00032	0.00019	0.00029	0.00020
	0.00019	0.00036	0.00019	0.00082	0.00032	0.00019	0.00029	0.00020
	1.00140	1.00178	1.00140	1.00310	1.00150	1.00140	1.00137	1.00121
β_6	0.00018	0.00037	0.00018	0.00069	0.00028	0.00018	0.00027	0.00018
	0.00018	0.00038	0.00018	0.00070	0.00028	0.00018	0.00027	0.00018
	0.99892	0.99773	0.99892	0.99979	0.99897	0.99892	0.99889	0.99906
β_7	0.00025	0.00039	0.00025	0.00087	0.00033	0.00025	0.00034	0.00025
	0.00025	0.00039	0.00025	0.00087	0.00034	0.00025	0.00034	0.00025
	1.00007	1.00023	1.00007	0.99899	1.00097	1.00007	1.00037	0.99991
β_8	0.00028	0.00045	0.00028	0.00078	0.00043	0.00028	0.00040	0.00028
	0.00028	0.00045	0.00028	0.00078	0.00043	0.00028	0.00040	0.00028
	0.99453	1.24957	0.99453	0.60630	0.98213	0.99453	0.83186	0.98975
σ	0.01446	0.04589	0.01446	0.04742	0.02063	0.01446	0.01506	0.01486
	0.01449	0.10817	0.01449	0.20242	0.02063	0.01449	0.04333	0.01496

Resultados do Estudo Monte Carlo - Modelo 50801

($n = 50$, $p = 8$, contaminação na média dos erros, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99860	1.00058	0.99860	1.00321	1.00621	0.99860	0.99938	1.00039
β_1	0.00479	0.01511	0.00479	0.00112	0.01054	0.00479	0.00370	0.00465
	0.00479	0.01511	0.00479	0.00113	0.01058	0.00479	0.00370	0.00465
	0.99822	0.99545	0.99822	1.00230	0.99344	0.99822	1.00104	0.99694
β_2	0.00446	0.01391	0.00446	0.00114	0.01189	0.00446	0.00346	0.00545
	0.00446	0.01393	0.00446	0.00114	0.01194	0.00446	0.00346	0.00545
	0.99515	0.99716	0.99515	1.00064	0.99130	0.99515	0.99902	1.00277
β_3	0.00506	0.01188	0.00506	0.00103	0.01043	0.00506	0.00372	0.00539
	0.00508	0.01189	0.00508	0.00103	0.01051	0.00508	0.00372	0.00540
	0.99856	0.99931	0.99856	1.00045	1.00118	0.99856	0.99913	0.99846
β_4	0.00445	0.01308	0.00445	0.00105	0.01050	0.00445	0.00302	0.00396
	0.00445	0.01308	0.00445	0.00105	0.01050	0.00445	0.00302	0.00396
	1.00687	1.01258	1.00687	0.99842	1.01475	1.00687	1.00607	1.00233
β_5	0.00528	0.01420	0.00528	0.00140	0.01160	0.00528	0.00365	0.00438
	0.00533	0.01435	0.00533	0.00141	0.01182	0.00533	0.00369	0.00438
	1.00673	1.00761	1.00673	1.00316	1.00165	1.00673	1.00323	1.00456
β_6	0.00469	0.01171	0.00469	0.00101	0.00931	0.00469	0.00275	0.00445
	0.00473	0.01171	0.00473	0.00102	0.00931	0.00473	0.00276	0.00447
	0.99056	0.98924	0.99056	0.99908	0.98961	0.99056	0.99537	0.99305
β_7	0.00515	0.01437	0.00515	0.00125	0.01252	0.00515	0.00376	0.00505
	0.00524	0.01448	0.00524	0.00125	0.01263	0.00524	0.00378	0.00510
	0.99606	0.99852	0.99606	1.00044	1.00048	0.99606	0.99866	1.00023
β_8	0.00468	0.01363	0.00468	0.00141	0.01154	0.00468	0.00340	0.00490
	0.00469	0.01363	0.00469	0.00141	0.01154	0.00469	0.00340	0.00490
	4.60714	5.21178	4.60714	0.57740	3.97200	4.60714	3.09644	3.37751
σ	0.06117	2.80015	0.06117	0.09954	0.19372	0.06117	0.28315	0.24791
	13.07264	20.53927	13.07264	0.27816	9.02650	13.07264	4.67823	5.90045

Resultados do Estudo Monte Carlo - Modelo 50802

($n = 50$, $p = 8$,contaminação na variância dos erros, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99621	0.99907	0.99803	1.00259	0.99605	0.99780	0.99626	0.99484
β_1	0.00496	0.00120	0.00419	0.00089	0.00343	0.00430	0.00139	0.00373
	0.00497	0.00120	0.00419	0.00090	0.00345	0.00430	0.00141	0.00376
	0.99963	1.00105	0.99971	1.00172	1.00552	0.99890	0.99632	1.00014
β_2	0.00543	0.00140	0.00506	0.00118	0.00456	0.00503	0.00165	0.00527
	0.00543	0.00140	0.00506	0.00118	0.00459	0.00503	0.00166	0.00527
	0.99348	0.99892	0.99578	1.00042	0.99378	0.99515	0.99872	0.99284
β_3	0.00492	0.00177	0.00469	0.00091	0.00427	0.00461	0.00166	0.00494
	0.00496	0.00177	0.00470	0.00091	0.00431	0.00463	0.00166	0.00499
	1.00273	0.99868	1.00323	0.99910	1.00999	1.00228	1.00400	1.00753
β_4	0.00523	0.00129	0.00433	0.00113	0.00463	0.00438	0.00138	0.00430
	0.00524	0.00129	0.00434	0.00113	0.00473	0.00438	0.00139	0.00436
	0.99774	0.99937	0.99815	0.99909	0.99581	0.99960	0.99940	0.99793
β_5	0.00453	0.00096	0.00399	0.00113	0.00339	0.00411	0.00119	0.00416
	0.00454	0.00096	0.00399	0.00113	0.00340	0.00411	0.00119	0.00416
	1.00192	1.00218	1.00176	1.00339	1.00468	1.00185	1.00174	1.00152
β_6	0.00402	0.00137	0.00379	0.00095	0.00428	0.00373	0.00139	0.00338
	0.00402	0.00137	0.00379	0.00096	0.00430	0.00373	0.00139	0.00338
	0.99526	0.99921	0.99570	1.00033	0.99554	0.99632	0.99982	0.99797
β_7	0.00482	0.00105	0.00448	0.00107	0.00409	0.00467	0.00140	0.00491
	0.00485	0.00105	0.00450	0.00107	0.00411	0.00468	0.00140	0.00491
	0.99896	0.99799	0.99716	0.99868	0.99685	0.99677	0.99855	0.99909
β_8	0.00455	0.00129	0.00376	0.00110	0.00452	0.00385	0.00172	0.00394
	0.00455	0.00129	0.00377	0.00110	0.00453	0.00386	0.00173	0.00394
	4.43458	1.57477	4.04396	0.61566	2.61899	4.12507	2.17551	3.07112
σ	0.97149	0.44902	1.26035	0.06989	0.38713	1.09691	0.30471	0.84524
	12.76783	0.77938	10.52606	0.21760	3.00827	10.86299	1.68653	5.13479

Resultados do Estudo Monte Carlo - Modelo 50803

($n = 50$, $p = 8$, contaminação na direção de X, sem intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.04514	0.14290	0.04514	0.13791	0.04851	0.04514	0.05077	0.04656
β_1	0.00093	0.07546	0.00093	0.07190	0.00154	0.00093	0.00134	0.00111
	0.91268	0.81009	0.91268	0.81510	0.90686	0.91268	0.90237	0.91015
	1.00357	1.00855	1.00357	1.02280	1.00425	1.00357	1.00493	1.00130
β_2	0.02550	0.04173	0.02550	0.05625	0.03891	0.02550	0.03550	0.02978
	0.02551	0.04181	0.02551	0.05677	0.03893	0.02551	0.03552	0.02978
	1.00743	1.00618	1.00743	0.98352	0.99959	1.00743	0.99484	1.00965
β_3	0.02344	0.03704	0.02344	0.05800	0.03195	0.02344	0.03077	0.02732
	0.02349	0.03708	0.02349	0.05827	0.03195	0.02349	0.03080	0.02741
	1.01957	1.01212	1.01957	1.04787	1.02333	1.01957	1.03558	1.01618
β_4	0.01942	0.02906	0.01942	0.06554	0.02602	0.01942	0.02548	0.02270
	0.01980	0.02920	0.01980	0.06783	0.02656	0.01980	0.02674	0.02296
	0.97454	0.98295	0.97454	0.98534	0.98351	0.97454	0.97427	0.97772
β_5	0.02331	0.03861	0.02331	0.05641	0.03261	0.02331	0.02919	0.02592
	0.02396	0.03890	0.02396	0.05663	0.03289	0.02396	0.02985	0.02641
	0.99073	0.98697	0.99073	1.00132	0.98309	0.99073	0.98445	0.99231
β_6	0.02349	0.04087	0.02349	0.05778	0.03263	0.02349	0.03401	0.02435
	0.02357	0.04104	0.02357	0.05778	0.03291	0.02357	0.03425	0.02441
	1.00889	1.00706	1.00889	0.99565	1.00511	1.00889	1.00598	1.00961
β_7	0.02475	0.04032	0.02475	0.05553	0.03223	0.02475	0.03187	0.02829
	0.02483	0.04037	0.02483	0.05554	0.03225	0.02483	0.03190	0.02839
	0.99411	0.98139	0.99411	0.99052	0.99168	0.99411	0.99760	0.99763
β_8	0.02204	0.03858	0.02204	0.06015	0.03332	0.02204	0.02738	0.02467
	0.02207	0.03893	0.02207	0.06024	0.03339	0.02207	0.02739	0.02468
	9.80497	11.11689	9.80497	5.60904	9.62409	9.80497	8.15601	9.75976
σ	1.12836	12.90698	1.12836	6.37275	1.59644	1.12836	1.14062	1.44013
	78.65590	115.25836	78.65590	27.61604	75.97135	78.65590	52.34906	78.17353

Resultados do Estudo Monte Carlo - Modelo 50810

($n = 50$, $p = 8$,sem contaminação, com intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00159	1.00182	1.00159	1.00206	1.00072	1.00159	1.00078	1.00147
β_1	0.00024	0.00051	0.00024	0.00084	0.00035	0.00024	0.00035	0.00024
	0.00024	0.00051	0.00024	0.00084	0.00035	0.00024	0.00035	0.00025
	1.00002	0.99831	1.00002	0.99777	0.99926	1.00002	0.99895	1.00017
β_2	0.00024	0.00036	0.00024	0.00082	0.00033	0.00024	0.00029	0.00024
	0.00024	0.00037	0.00024	0.00082	0.00033	0.00024	0.00029	0.00024
	1.00153	1.00239	1.00153	1.00460	1.00246	1.00153	1.00277	1.00207
β_3	0.00024	0.00050	0.00024	0.00083	0.00038	0.00024	0.00035	0.00026
	0.00025	0.00051	0.00025	0.00085	0.00039	0.00025	0.00036	0.00027
	0.99983	0.99912	0.99983	0.99858	0.99881	0.99983	1.00030	0.99959
β_4	0.00025	0.00047	0.00025	0.00087	0.00036	0.00024	0.00034	0.00027
	0.00025	0.00047	0.00025	0.00088	0.00036	0.00025	0.00034	0.00027
	0.99979	0.99887	0.99979	1.00190	1.00054	0.99980	1.00002	0.99983
β_5	0.00027	0.00043	0.00027	0.00092	0.00035	0.00025	0.00036	0.00028
	0.00027	0.00043	0.00027	0.00092	0.00035	0.00027	0.00036	0.00028
	0.99964	0.99815	0.99964	0.99952	0.99842	0.99964	0.99867	0.99960
β_6	0.00023	0.00040	0.00023	0.00083	0.00034	0.00027	0.00031	0.00025
	0.00023	0.00041	0.00023	0.00083	0.00034	0.00023	0.00031	0.00025
	0.99925	0.99962	0.99925	0.99923	0.99952	0.99925	0.99889	0.99955
β_7	0.00022	0.00043	0.00022	0.00075	0.00035	0.00022	0.00028	0.00023
	0.00022	0.00043	0.00022	0.00075	0.00035	0.00022	0.00028	0.00023
	0.96809	0.96184	0.96809	0.96483	0.96296	0.96809	0.97135	0.96725
β_8	0.02363	0.03724	0.02363	0.06985	0.02758	0.02363	0.02799	0.02401
	0.02465	0.03869	0.02465	0.07109	0.02896	0.02465	0.02881	0.02508
	0.98831	1.20955	0.98831	0.56025	0.96890	0.98831	0.82094	0.98518
σ	0.01154	0.04171	0.01154	0.04740	0.01560	0.01154	0.01229	0.01244
	0.01168	0.08562	0.01168	0.24078	0.01657	0.01168	0.04435	0.01266

Resultados do Estudo Monte Carlo - Modelo 50811

($n = 50$, $p = 8$, contaminação na média dos erros, com intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.99350	0.99033	0.99350	1.00069	0.99214	0.99350	0.99612	0.99183
β_1	0.00487	0.00845	0.00487	0.00161	0.00803	0.00487	0.00599	0.00831
	0.00491	0.00854	0.00491	0.00161	0.00809	0.00491	0.00600	0.00838
	0.99450	0.99908	0.99450	1.00108	0.99861	0.99450	0.99731	0.98629
β_2	0.00408	0.00813	0.00408	0.00164	0.00679	0.00408	0.00467	0.00831
	0.00411	0.00813	0.00411	0.00164	0.00680	0.00411	0.00468	0.00850
	1.00034	0.99631	1.00034	1.00206	0.99488	1.00034	0.99767	1.00398
β_3	0.00420	0.00719	0.00420	0.00188	0.00774	0.00420	0.00495	0.00795
	0.00420	0.00720	0.00420	0.00189	0.00776	0.00420	0.00496	0.00797
	0.99479	0.99241	0.99479	0.99898	0.99475	0.99479	0.99897	0.99428
β_4	0.00394	0.00811	0.00394	0.00168	0.00716	0.00394	0.00481	0.00648
	0.00396	0.00816	0.00396	0.00168	0.00719	0.00396	0.00481	0.00652
	0.99944	1.00249	0.99944	1.00157	1.00274	0.99944	1.00133	0.99833
β_5	0.00446	0.00680	0.00446	0.00182	0.00724	0.00446	0.00527	0.00651
	0.00446	0.00690	0.00446	0.00182	0.00725	0.00446	0.00527	0.00652
	1.00181	1.00469	1.00181	0.99949	1.00200	1.00181	1.00606	1.00645
β_6	0.00403	0.00710	0.00403	0.00134	0.00650	0.00403	0.00449	0.00769
	0.00404	0.00712	0.00404	0.00134	0.00650	0.00404	0.00452	0.00773
	0.99558	0.99647	0.99558	0.99263	0.99223	0.99558	0.99454	1.00986
β_7	0.00495	0.00972	0.00495	0.00198	0.00911	0.00495	0.00621	0.00730
	0.00497	0.00973	0.00497	0.00203	0.00917	0.00497	0.00624	0.00739
	2.99271	3.49191	2.99271	1.09538	2.70563	2.99271	2.30873	2.35820
β_8	0.07542	0.73944	0.07542	0.31612	0.18012	0.07542	0.21464	0.35894
	4.04630	6.94904	4.04630	0.32522	3.08928	4.04630	1.92742	2.20365
	4.14099	5.38084	4.14099	0.71150	4.07744	4.14099	3.25306	4.19645
σ	0.04603	0.56920	0.04603	0.53025	0.12623	0.04602	0.12341	0.20201
	9.91186	19.76098	9.91186	0.61348	9.59088	9.91186	5.19967	10.41929

Resultados do Estudo Monte Carlo - Modelo 50812

($n = 50$, $p = 8$,contaminação na variância dos erros, com intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COV. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	1.00953	1.00253	1.00705	1.00144	1.00275	1.00703	1.00166	1.00208
β_1	0.00438	0.00089	0.00364	0.00113	0.00301	0.00384	0.00121	0.01118
	0.00447	0.00090	0.00369	0.00113	0.00302	0.00389	0.00121	0.01119
	0.99905	1.00118	1.00146	0.99904	1.00193	1.00031	0.99970	0.99173
β_2	0.00483	0.00086	0.00378	0.00136	0.00273	0.00412	0.00136	0.01140
	0.00483	0.00086	0.00378	0.00137	0.00273	0.00412	0.00136	0.01147
	1.00720	1.00033	1.00496	1.00217	1.00311	1.00606	1.00315	1.01152
β_3	0.00515	0.00082	0.00474	0.00118	0.00324	0.00493	0.00135	0.01168
	0.00520	0.00082	0.00477	0.00119	0.00325	0.00497	0.00137	0.01182
	0.99217	0.99954	0.98856	0.99929	0.99164	0.98880	0.99735	0.99552
β_4	0.00447	0.00079	0.00402	0.00118	0.00293	0.00426	0.00108	0.01042
	0.00453	0.00079	0.00416	0.00118	0.00300	0.00439	0.00108	0.01044
	0.99264	0.99880	0.99331	1.00035	1.00163	0.99317	0.99843	0.98951
β_5	0.00448	0.00099	0.00416	0.00129	0.00375	0.00463	0.00126	0.01187
	0.00454	0.00099	0.00420	0.00129	0.00375	0.00467	0.00126	0.01198
	1.00232	0.99946	1.00261	0.99836	0.99795	1.00163	0.99865	1.00328
β_6	0.00442	0.00071	0.00382	0.00118	0.00285	0.00393	0.00116	0.01117
	0.00443	0.00071	0.00382	0.00119	0.00285	0.00394	0.00116	0.01118
	0.99736	0.99811	1.00023	0.99660	0.99653	0.99988	0.99890	1.00896
β_7	0.00538	0.00114	0.00508	0.00126	0.00340	0.00514	0.00150	0.00954
	0.00539	0.00115	0.00508	0.00127	0.00341	0.00514	0.00150	0.00962
	0.89522	0.94767	0.94386	0.98248	0.96456	0.94610	0.97143	1.02632
β_8	0.45032	0.07239	0.41819	0.08299	0.27797	0.43182	0.15494	0.69770
	0.46130	0.07512	0.42134	0.08329	0.27923	0.43473	0.15575	0.69840
	4.34363	1.44232	3.87361	0.54777	2.46137	3.97703	2.07274	3.33836
σ	0.76542	0.17223	1.14452	0.04837	0.41781	0.92802	0.28087	1.12782
	11.94526	0.36788	9.40214	0.25288	2.55342	9.79073	1.43165	6.59573

Resultados do Estudo Monte Carlo - Modelo 50813

($n = 50$, $p = 8$,contaminação na direção de X, com intercepto)

Parâmetros	LS	ONE AT A TIME			SEVERAL AT A TIME			
		COV. + LS	L. D. + LS	DFITS + LS	COU. + LS	L. D. + LS	DFITS + LS	C. N. + LS
	0.05094	0.11810	0.05094	0.13663	0.05038	0.05094	0.05298	0.05086
β_1	0.00129	0.05240	0.00129	0.07305	0.00204	0.00129	0.00185	0.00135
	0.90200	0.83014	0.90200	0.81846	0.90382	0.90200	0.89870	0.90222
	1.00230	0.99862	1.00230	0.97034	1.00162	1.00230	1.00550	1.00368
β_2	0.02828	0.03937	0.02828	0.07482	0.03760	0.02828	0.03463	0.02918
	0.02828	0.03937	0.02828	0.07570	0.03761	0.02828	0.03466	0.02920
	0.99483	0.98822	0.99483	0.99206	0.98473	0.99483	0.99519	0.99600
β_3	0.02173	0.03549	0.02173	0.06277	0.03029	0.02173	0.02809	0.02142
	0.02176	0.03563	0.02176	0.06283	0.03053	0.02176	0.02811	0.02144
	1.00525	0.99542	1.00525	0.97425	1.00477	1.00525	0.99589	1.00594
β_4	0.02365	0.03448	0.02365	0.05790	0.03104	0.02365	0.03066	0.02387
	0.02368	0.03450	0.02368	0.05856	0.03106	0.02368	0.03068	0.02390
	0.99437	0.99549	0.99437	0.95943	0.99467	0.99437	0.98222	0.99223
β_5	0.02440	0.03527	0.02440	0.06949	0.03079	0.02440	0.03024	0.02544
	0.02443	0.03529	0.02443	0.07114	0.03082	0.02443	0.03056	0.02550
	0.97442	0.96720	0.97442	0.94622	0.97108	0.97442	0.97407	0.97479
β_6	0.02563	0.04299	0.02586	0.06170	0.03677	0.02563	0.03234	0.02732
	0.02628	0.04407	0.02628	0.06460	0.03760	0.02628	0.03302	0.02796
	0.99779	0.99580	0.99779	0.99861	0.99501	0.99779	1.00944	0.99744
β_7	0.02357	0.03536	0.02357	0.07419	0.03364	0.02357	0.03360	0.02401
	0.02358	0.03538	0.02358	0.07419	0.03367	0.02358	0.03369	0.02401
	-0.18850	-0.04350	-0.18850	0.26701	-0.19971	-0.18850	0.06185	-0.17290
constante	2.58253	2.92560	2.58253	6.05023	2.89356	2.58253	3.20412	2.72990
	3.99506	4.01449	3.99506	6.58750	4.33287	3.99506	4.08425	4.10559
	9.66322	11.27773	9.66322	5.28705	9.50000	9.66322	8.05211	9.62796
σ	1.21108	9.94803	1.21108	6.19346	1.78138	1.21108	1.45627	1.20849
	76.26244	115.58039	76.26244	24.57228	74.03148	76.26244	51.18854	75.65017

A N E X O 2

G R Á F I C O S

GRÁFICO 1

FUNÇÃO DE INFLUÊNCIA

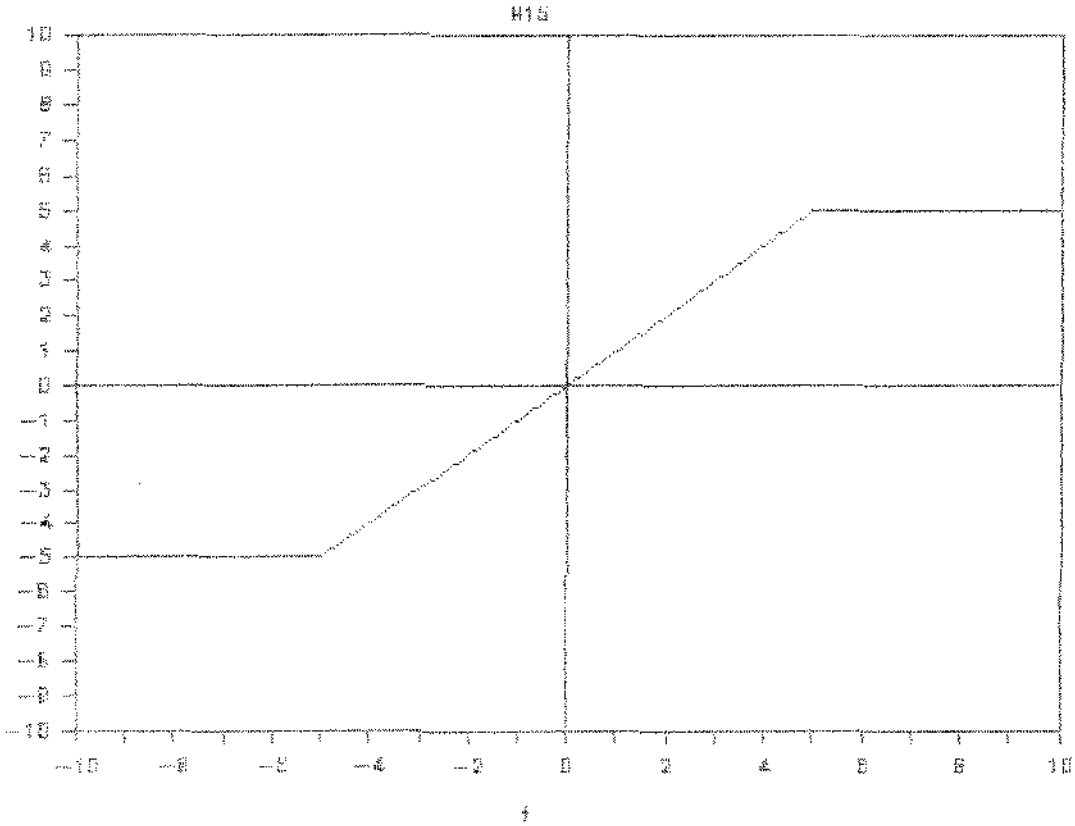


GRÁFICO 2

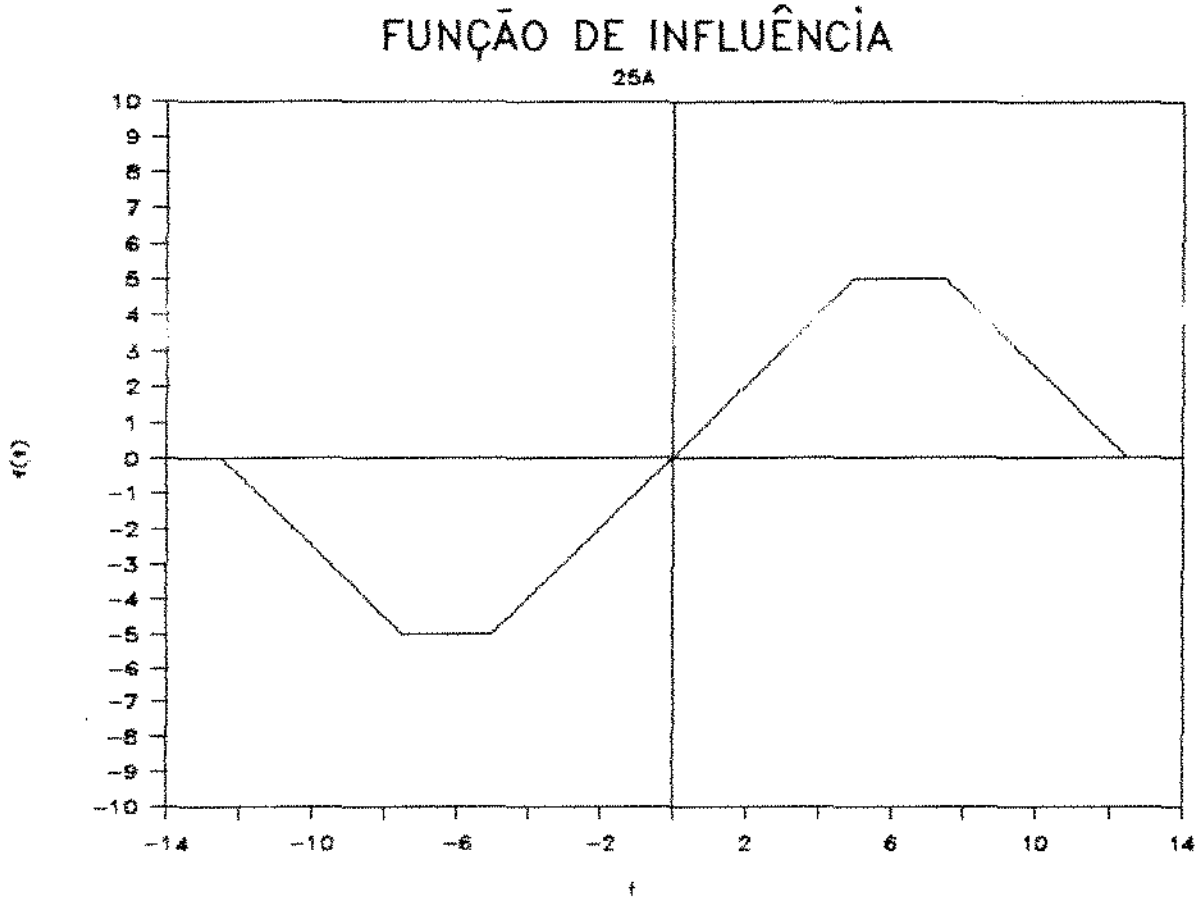
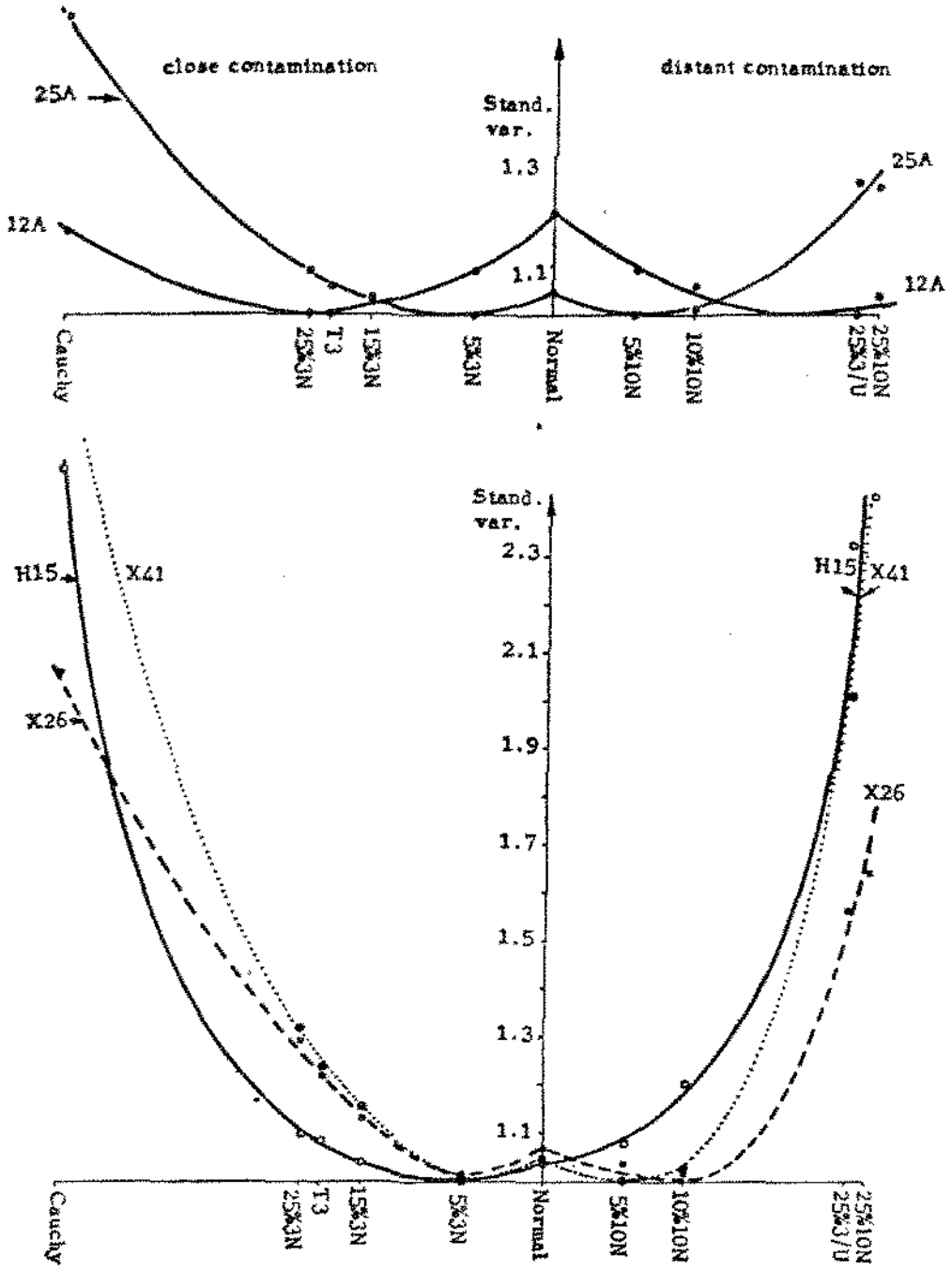

FUNÇÃO DE INFLUÊNCIA
25A
10
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10
-14
-10
-6
-2
2
6
10
14
t
t(t)

GRÁFICO 3



Observação: 12A é obtido fazendo-se d=1.2, b=3.5 e c=8.0 na $\psi(t;a,b,c)$ empregada no 25A.

GRÁFICO 4

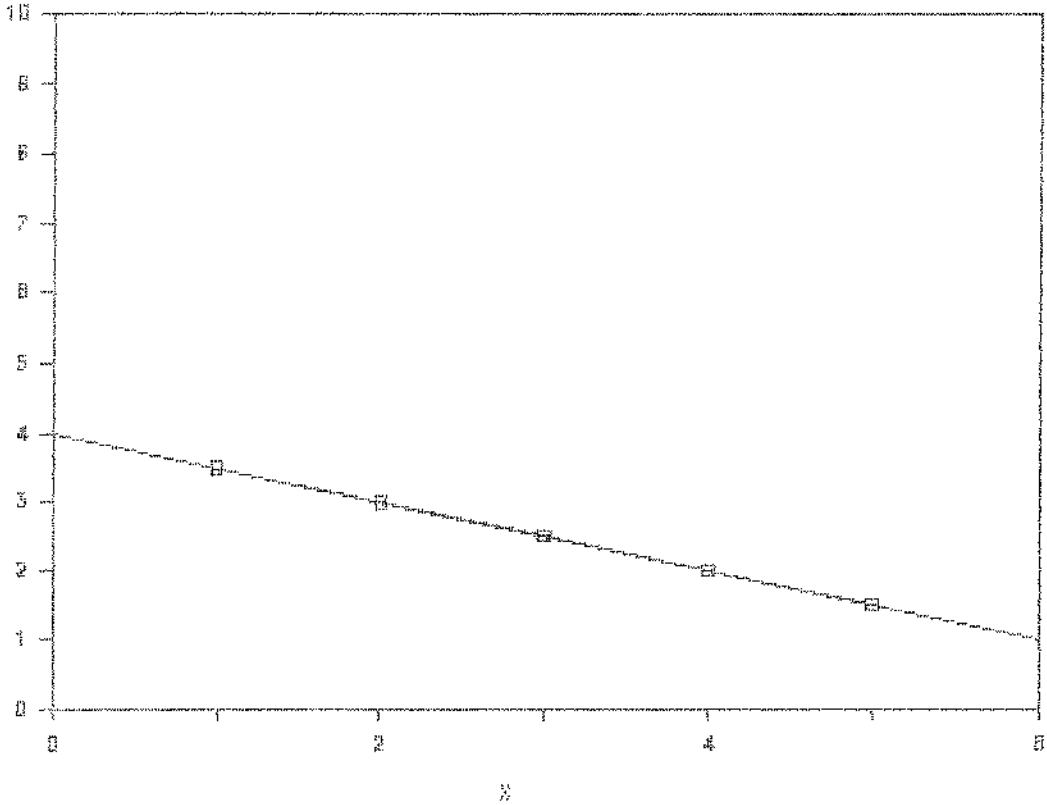


GRÁFICO 5

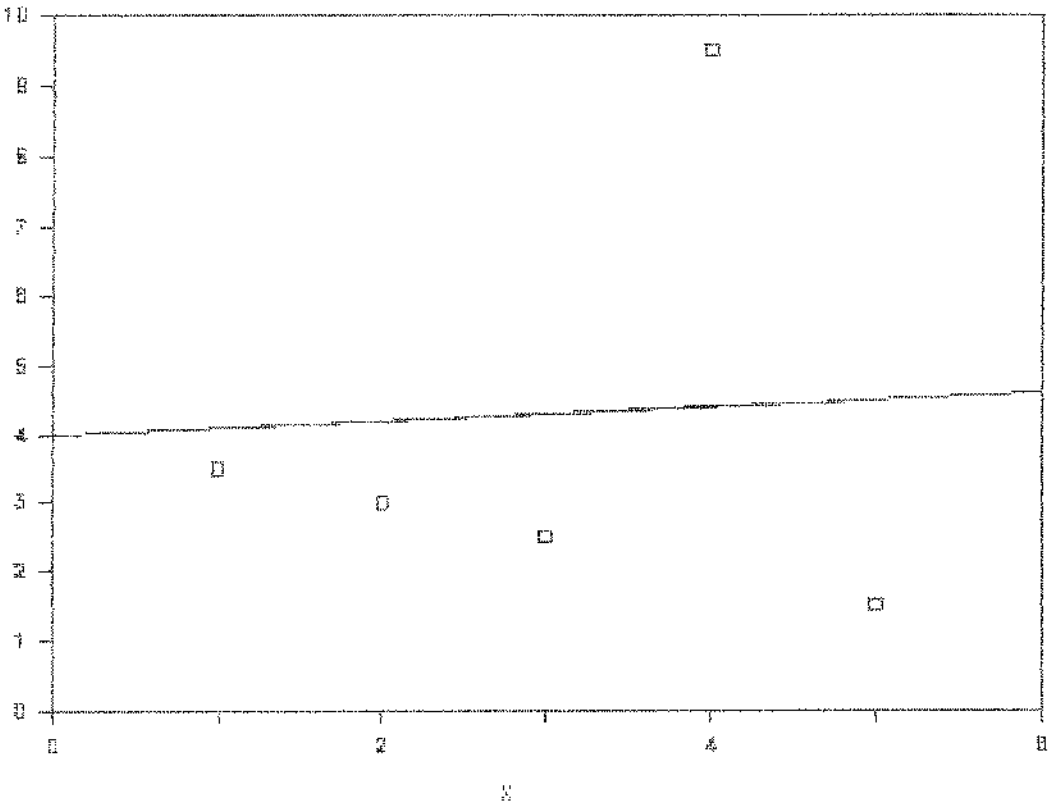


GRÁFICO 6

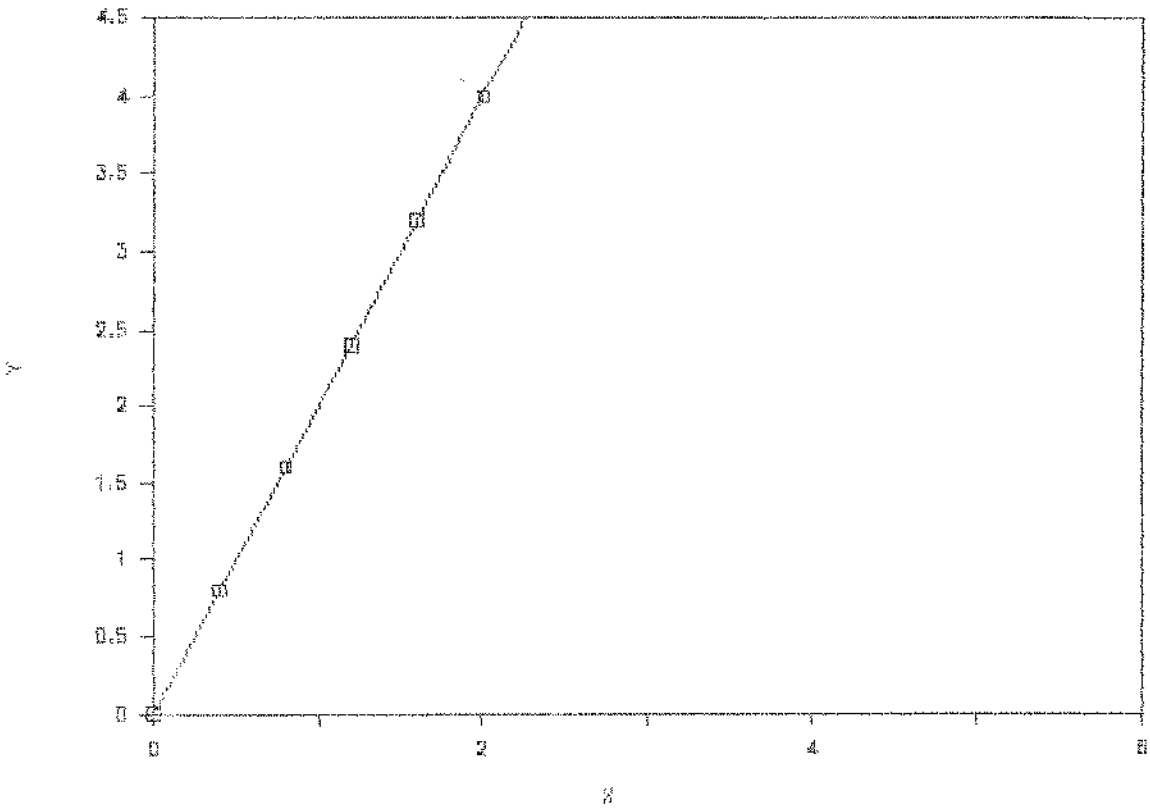


GRÁFICO 7

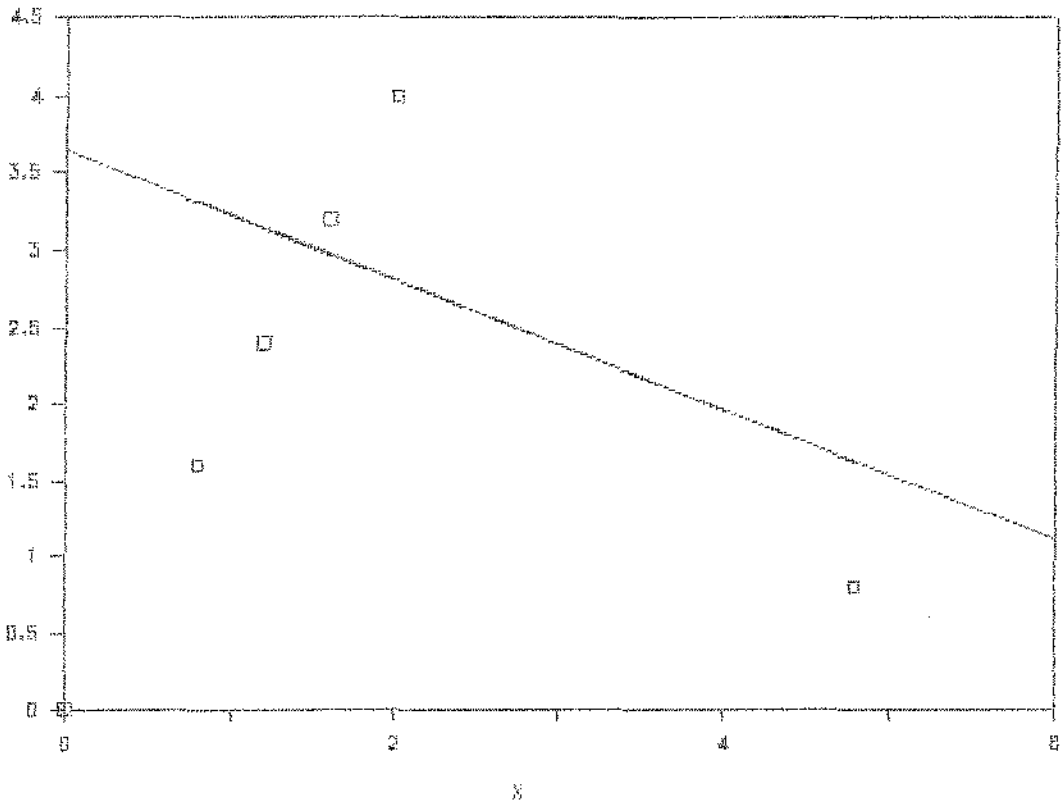


GRÁFICO 8

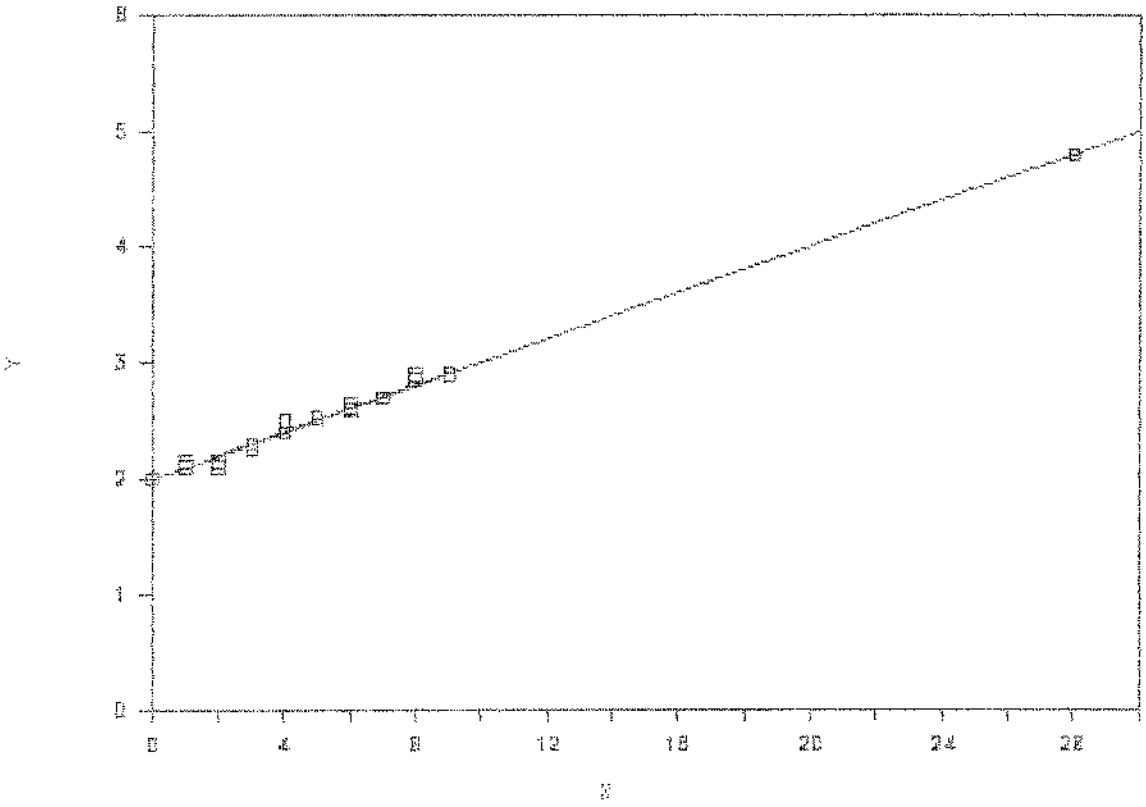


GRÁFICO 9

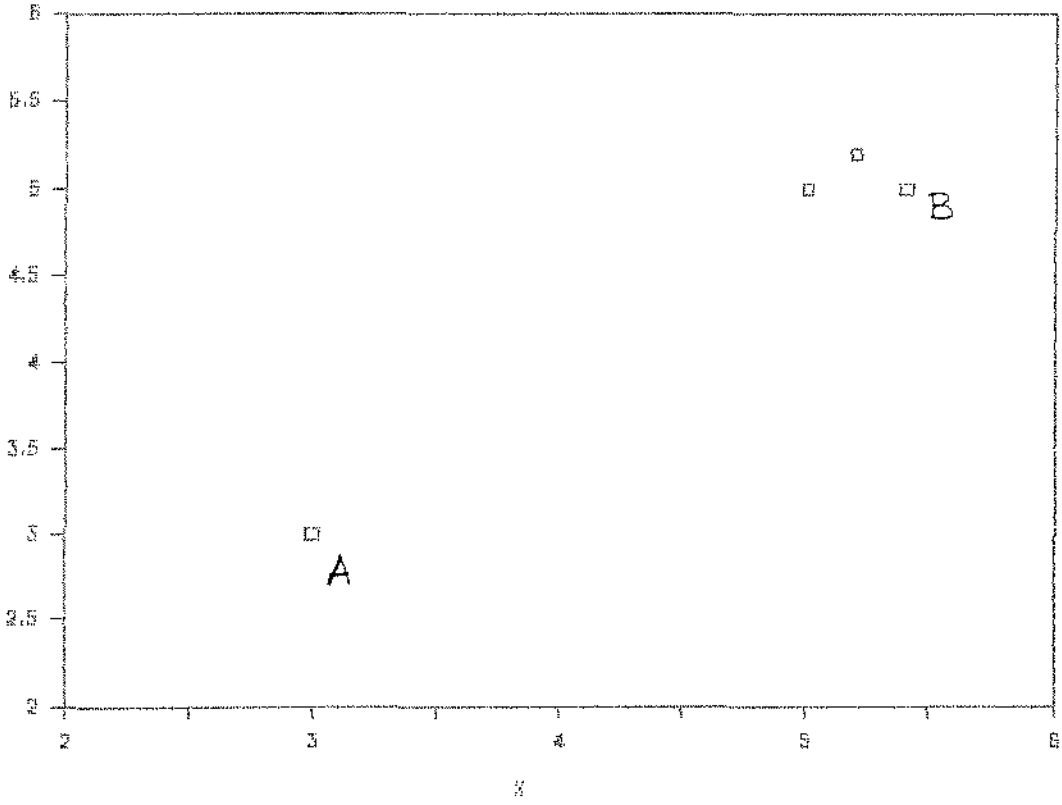


GRÁFICO 10

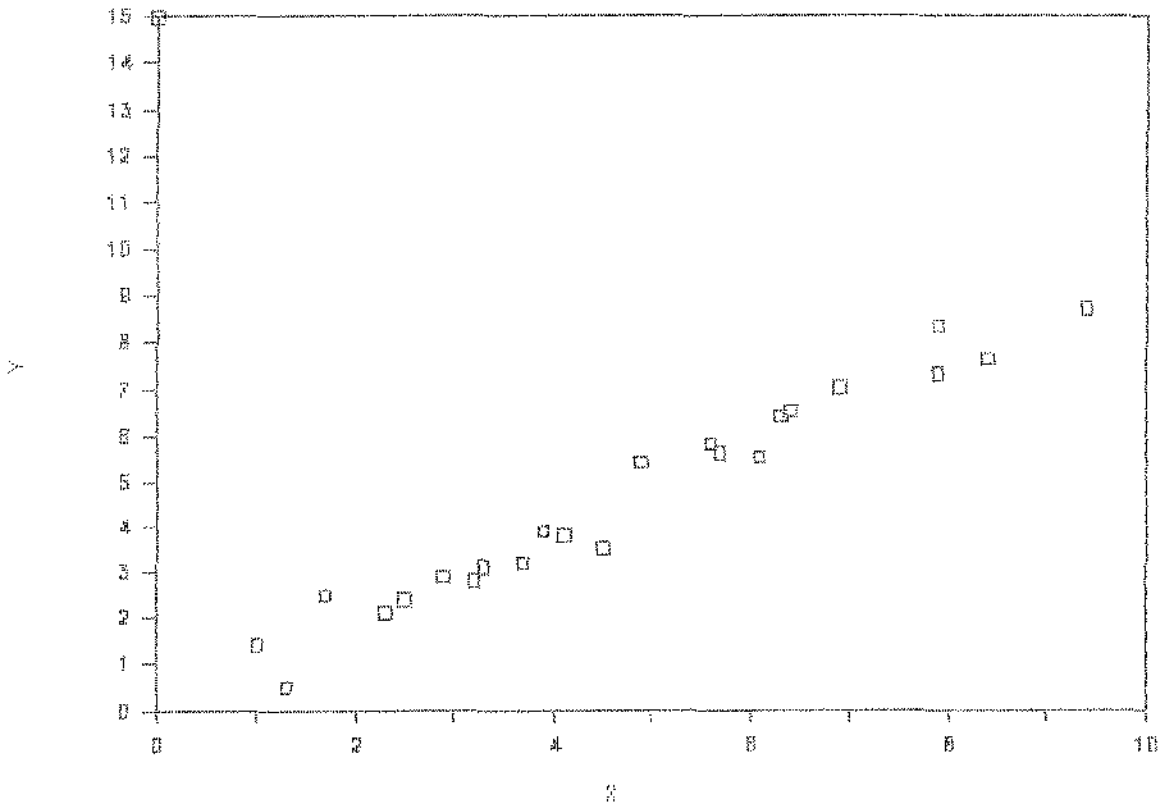


GRÁFICO 11

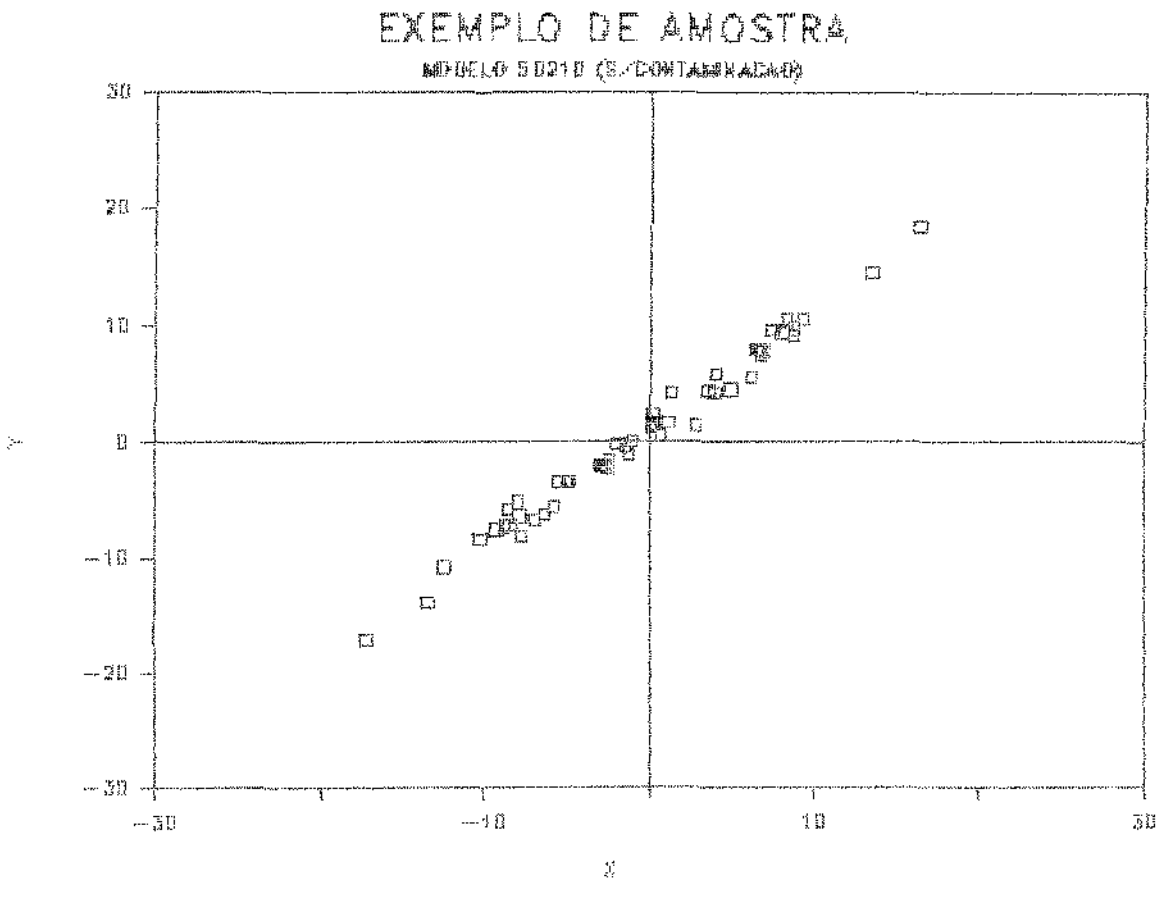


GRÁFICO 12

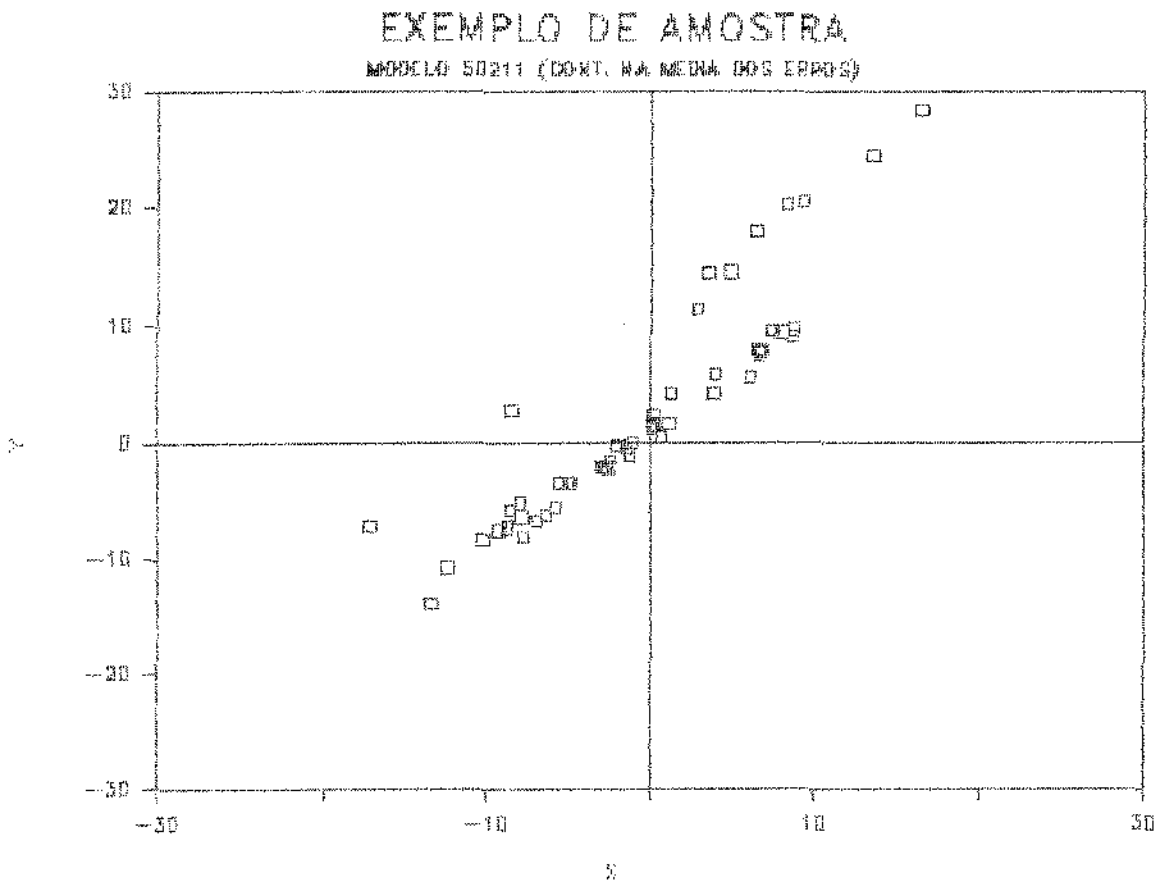

EXEMPLO DE AMOSTRA
MODELO 50211 (CONST. NA MEDIA DOS ERROS)
30
20
10
0
-10
-30
-50
-30
-10
10
30
Y
X

GRÁFICO 13

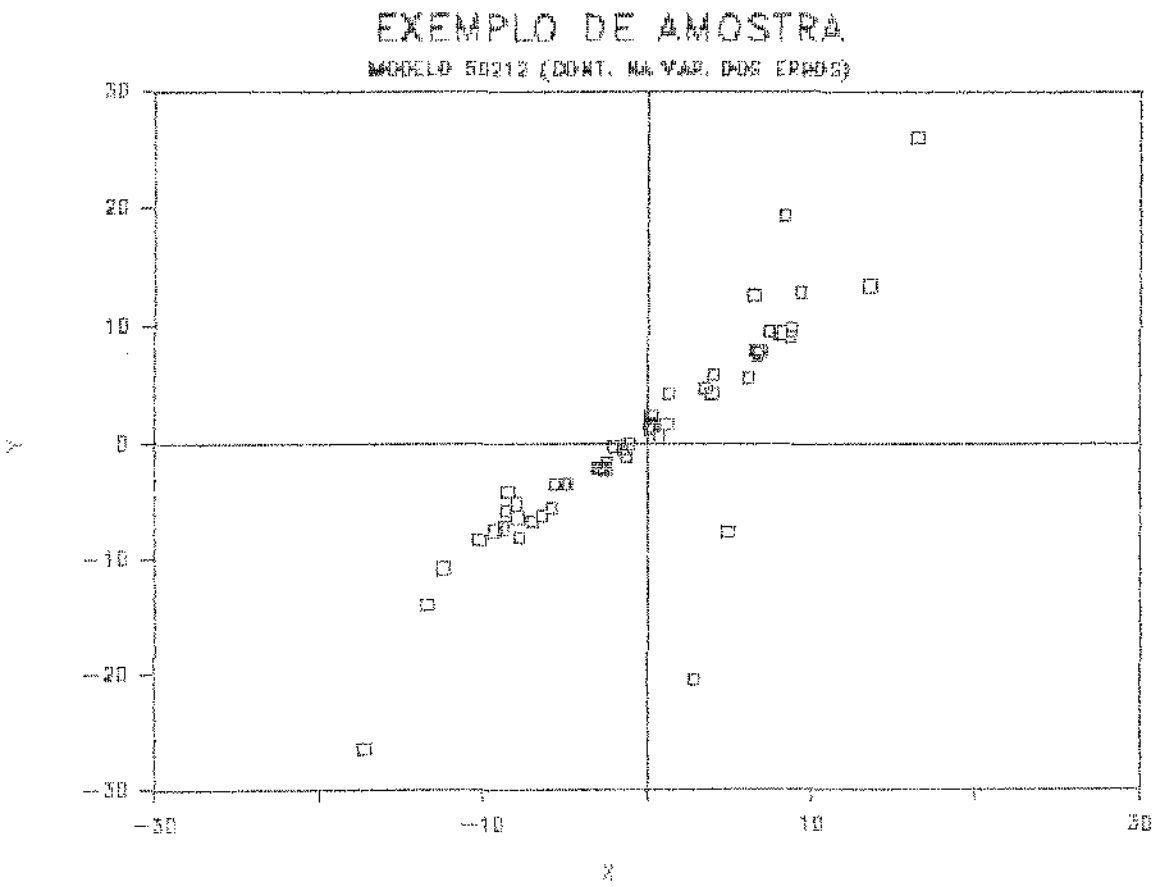

EXEMPLO DE AMOSTRA
MODELO 50213 (CONT. NA VAR. DOS ERROS)
30
20
10
0
-10
-20
-30
Y
-30
-10
10
30
X

GRÁFICO 14

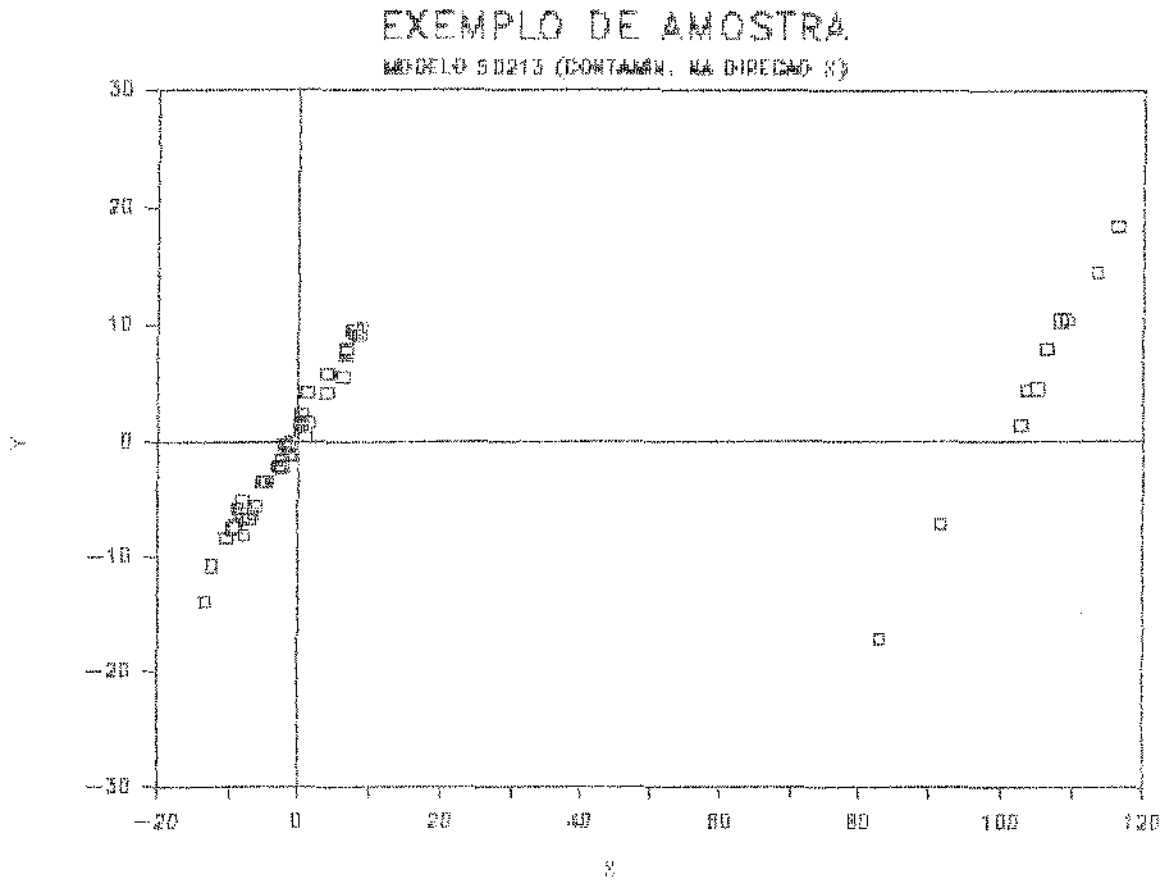

EXEMPLO DE AMOSTRA
MODELO 50215 (CONTAGEM NA DIREÇÃO X)
30
20
10
0
-10
-20
-30
-20
0
20
40
60
80
100
120
Y
X

GRÁFICO 15

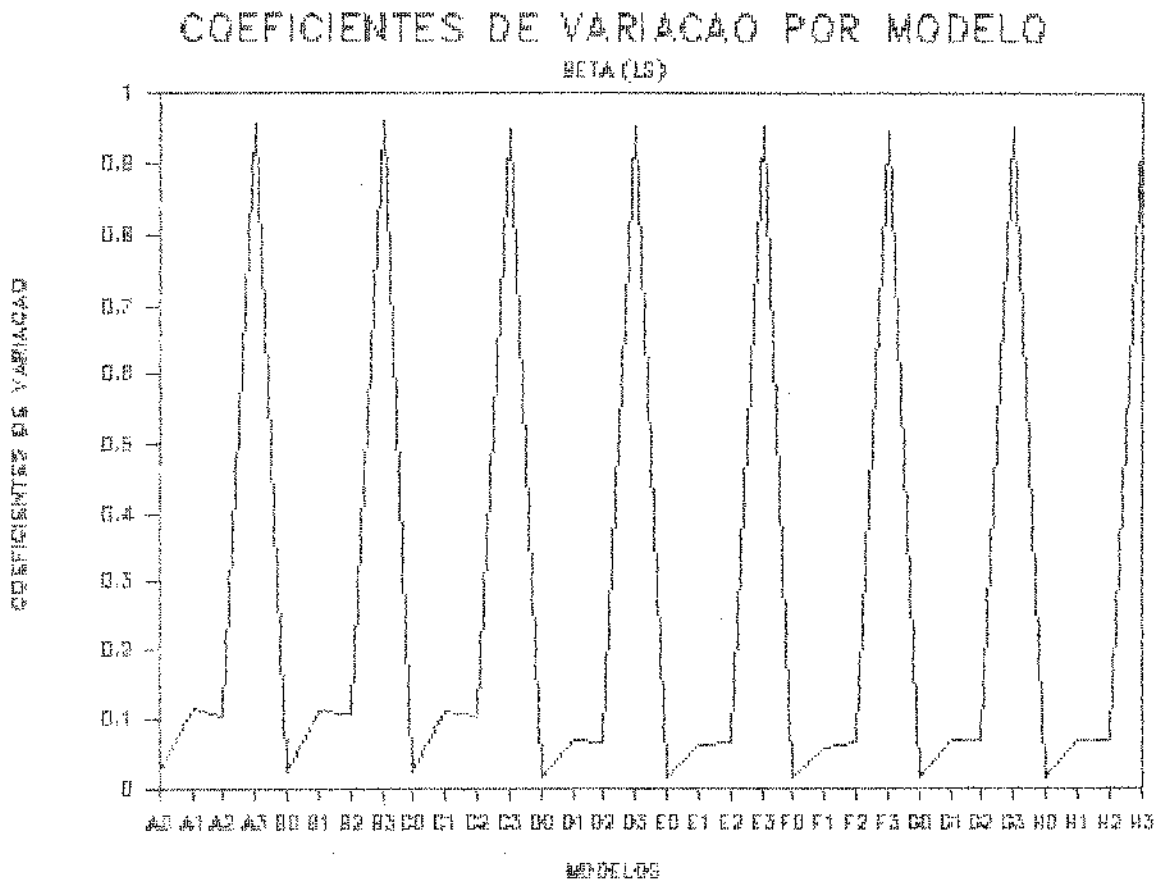

COEFICIENTES DE VARIACAO POR MODELO
BETA(LS)
COEFICIENTES DE VARIACAO
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 15A

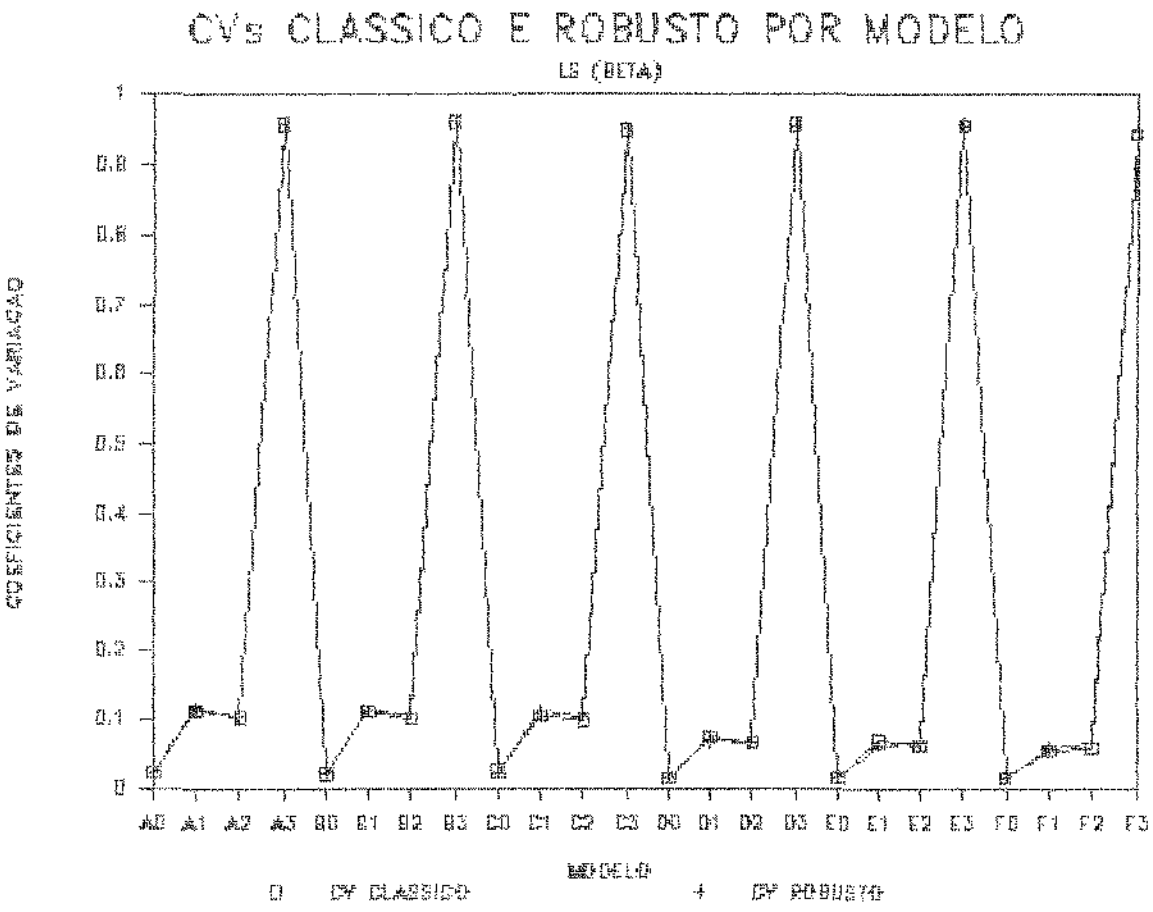

CVs CLASSICO E ROBUSTO POR MODELO
LS (BETA)
COEFICIENTES DE VARIACAO
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3
MODELO
CV CLASSICO
CV ROBUSTO

GRÁFICO 16

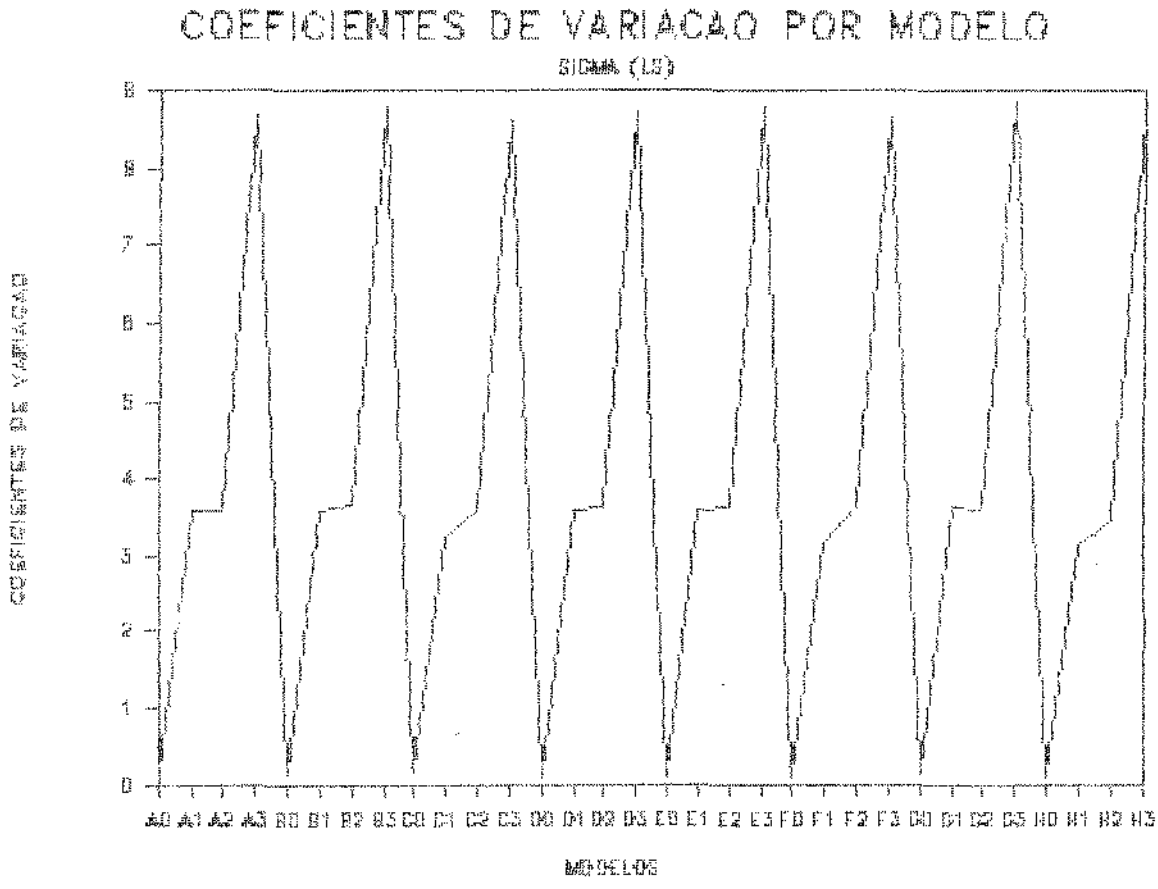

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (LS)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
9
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 17

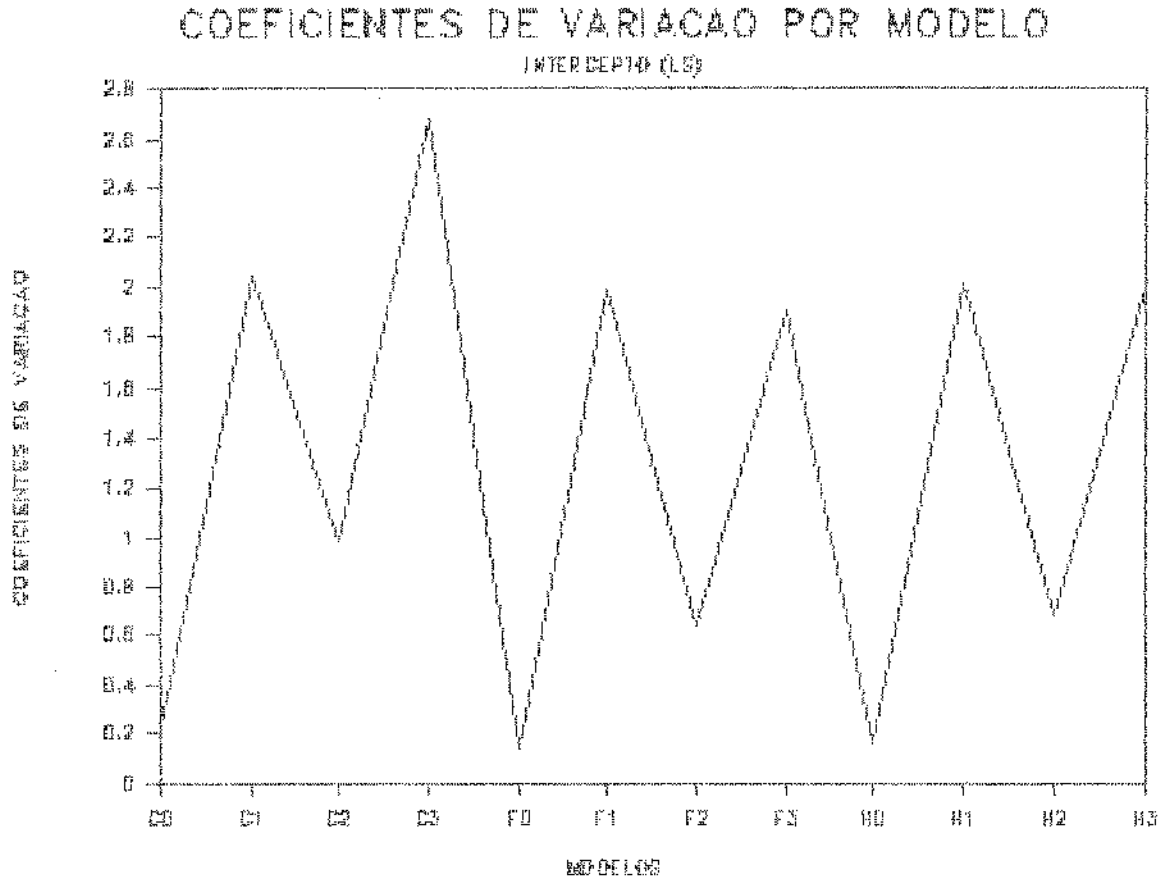


GRÁFICO 18

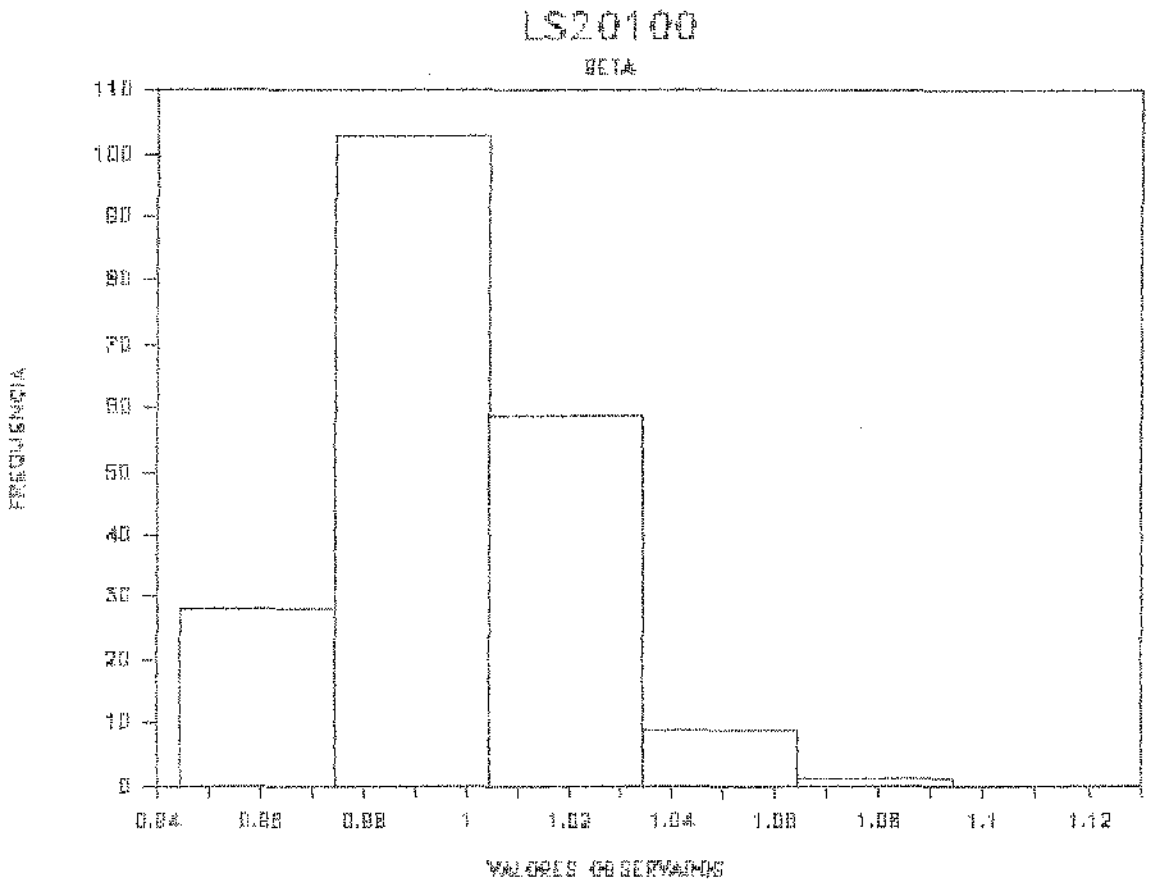

LS20100
BETA
110
100
90
80
70
60
50
40
30
20
10
0
0.94
0.96
0.98
1
1.02
1.04
1.06
1.08
1.1
1.12

GRÁFICO 19

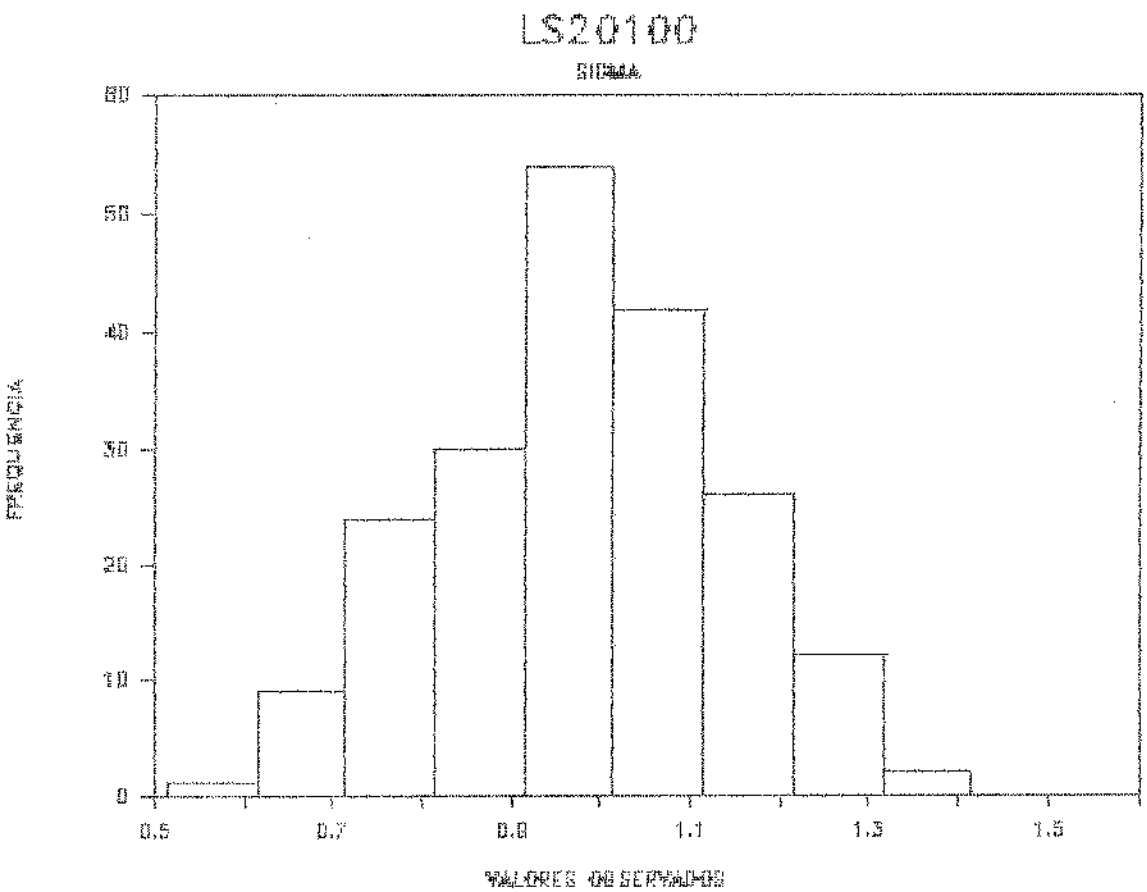

LS20100
SIGMA
FREQUENCIA
60
50
40
30
20
10
0
0.5
0.7
0.9
1.1
1.3
1.5
VALORES OBSERVADOS

GRÁFICO 21

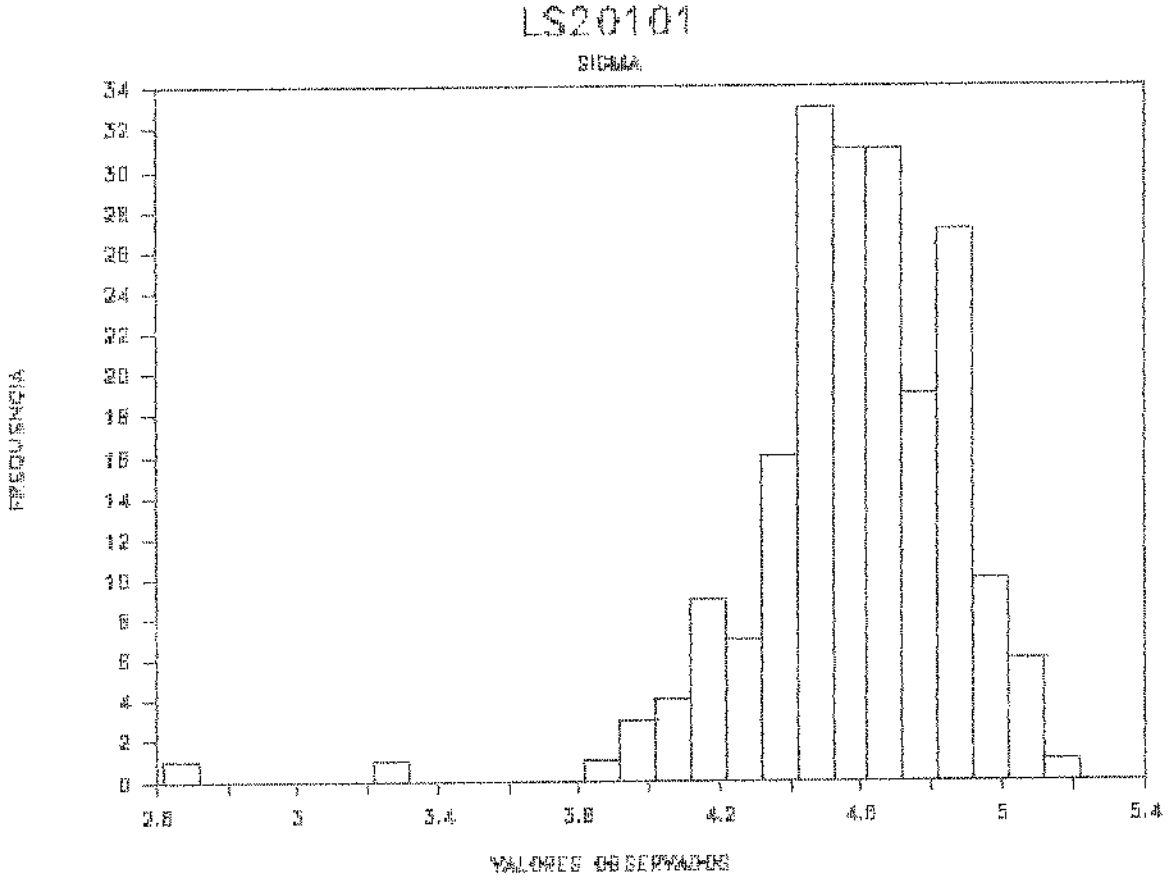

LS20101
FREQUENCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
2.8
3
3.4
3.8
4.2
4.6
5
5.4

GRÁFICO 22

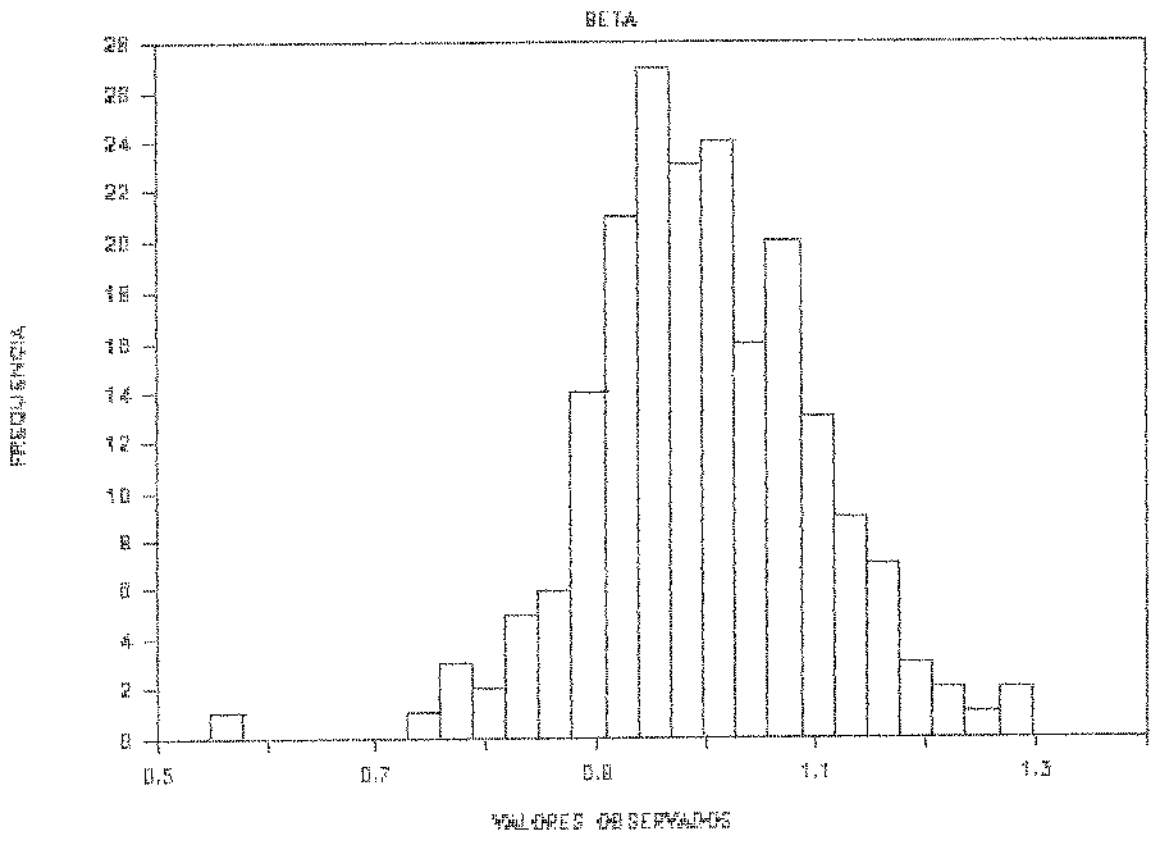

LS20102
BETA
FREQUÊNCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
0.5
0.7
0.9
1.1
1.3
VALORES OBSERVADOS

GRÁFICO 23

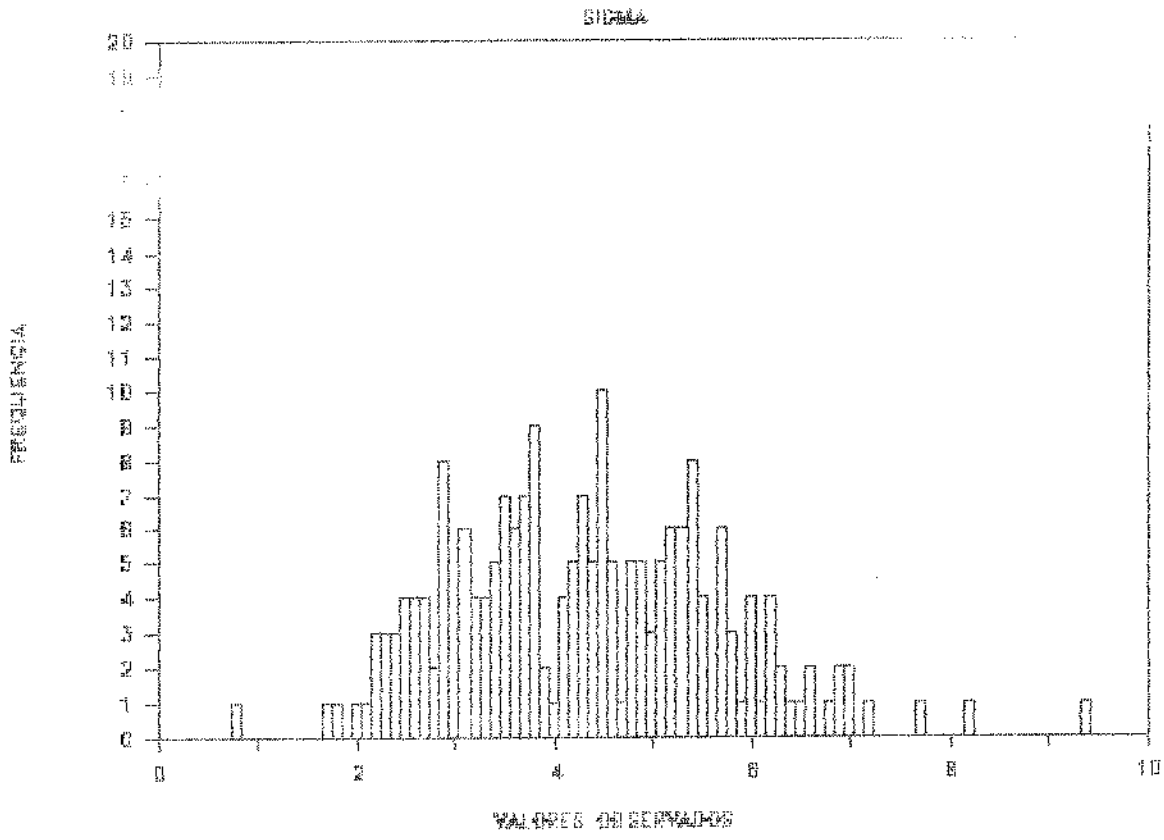


GRÁFICO 24

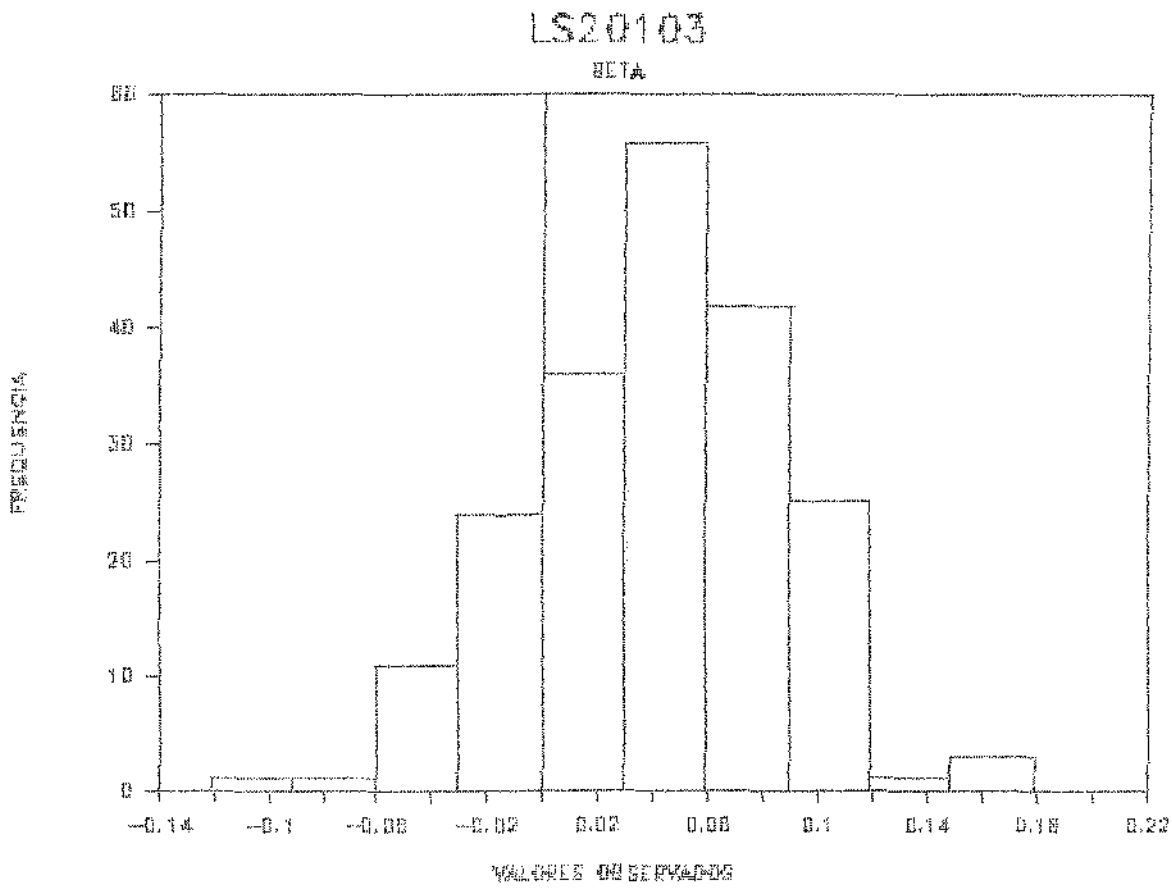

LS20103
BETA
60
50
40
30
20
10
0
-0.14
-0.1
-0.06
-0.02
0.02
0.06
0.1
0.14
0.18
0.22

GRÁFICO 25

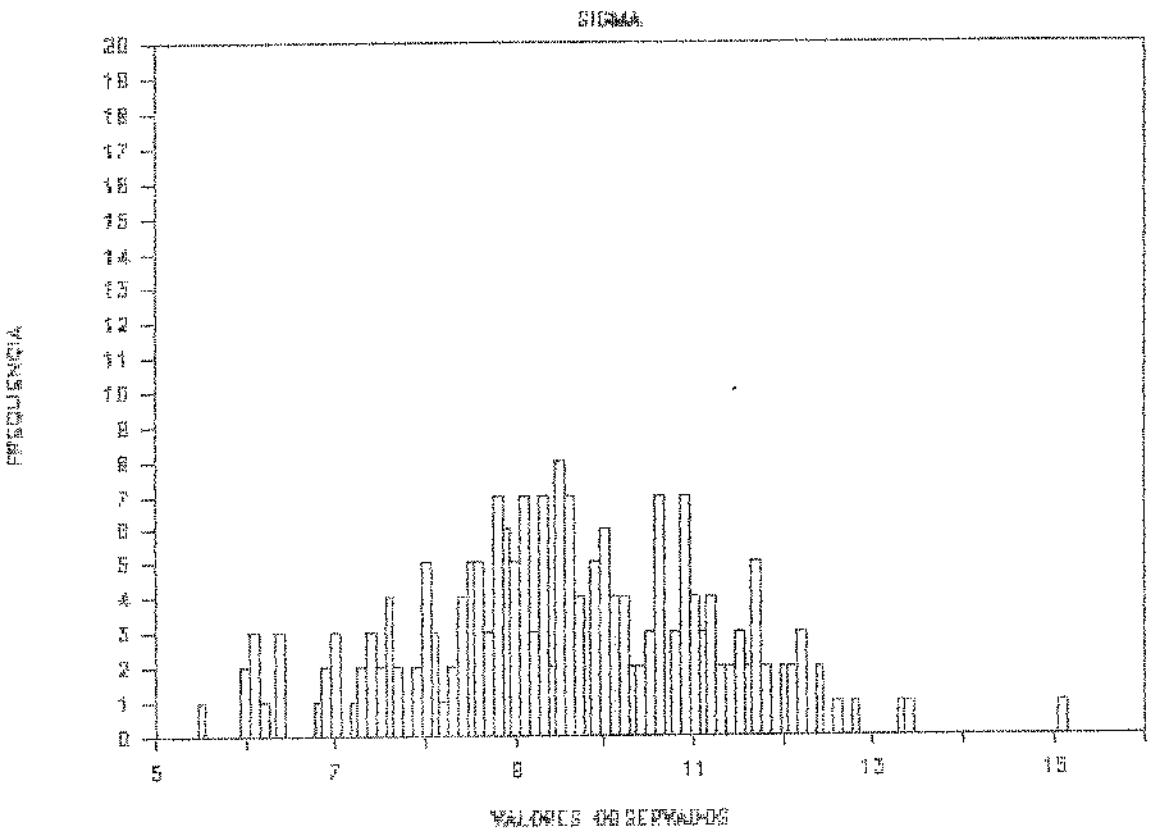

LS20103
SIGMA
FRECUENCIA
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
5
7
8
11
13
15
VALORES OBSERVADOS

GRÁFICO 26

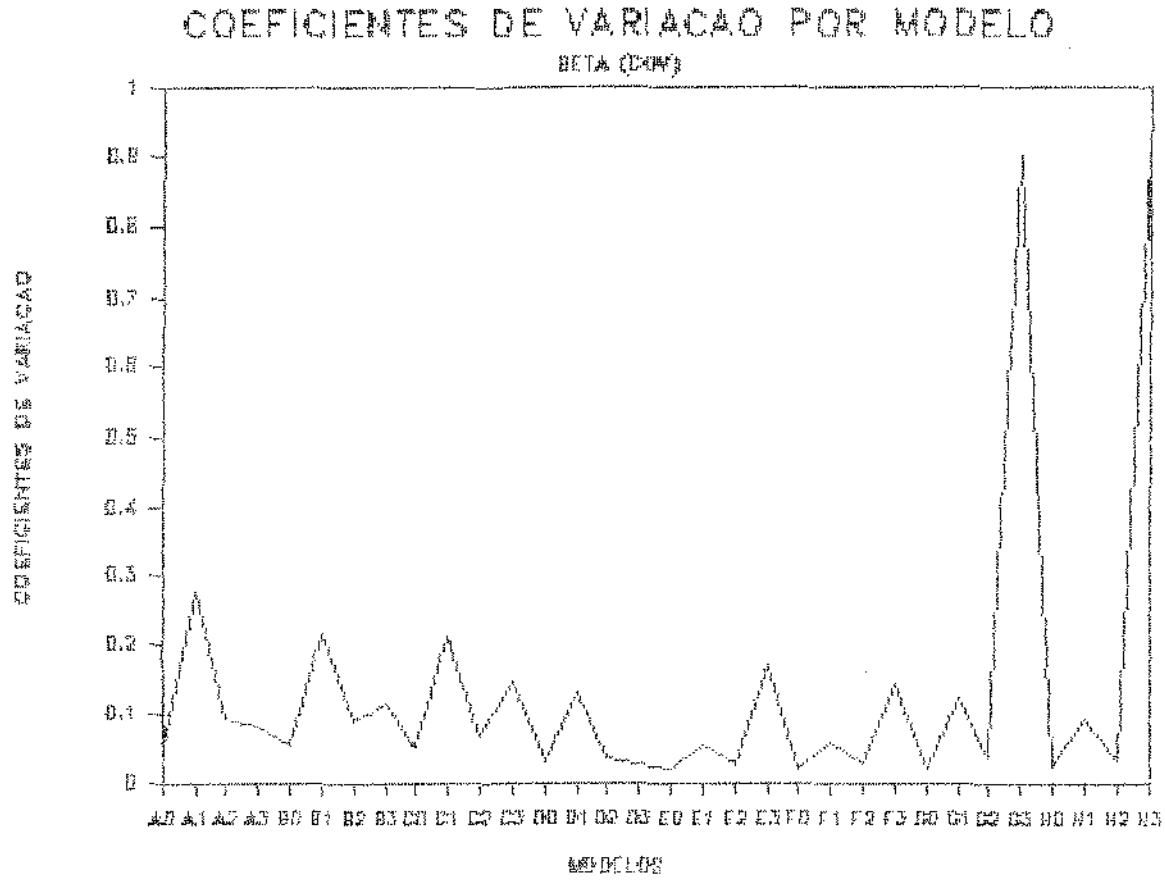

COEFICIENTES DE VARIACAO POR MODELO
BETA (D=W)
COEFICIENTES DE VARIACAO
MODELOS

GRÁFICO 26A

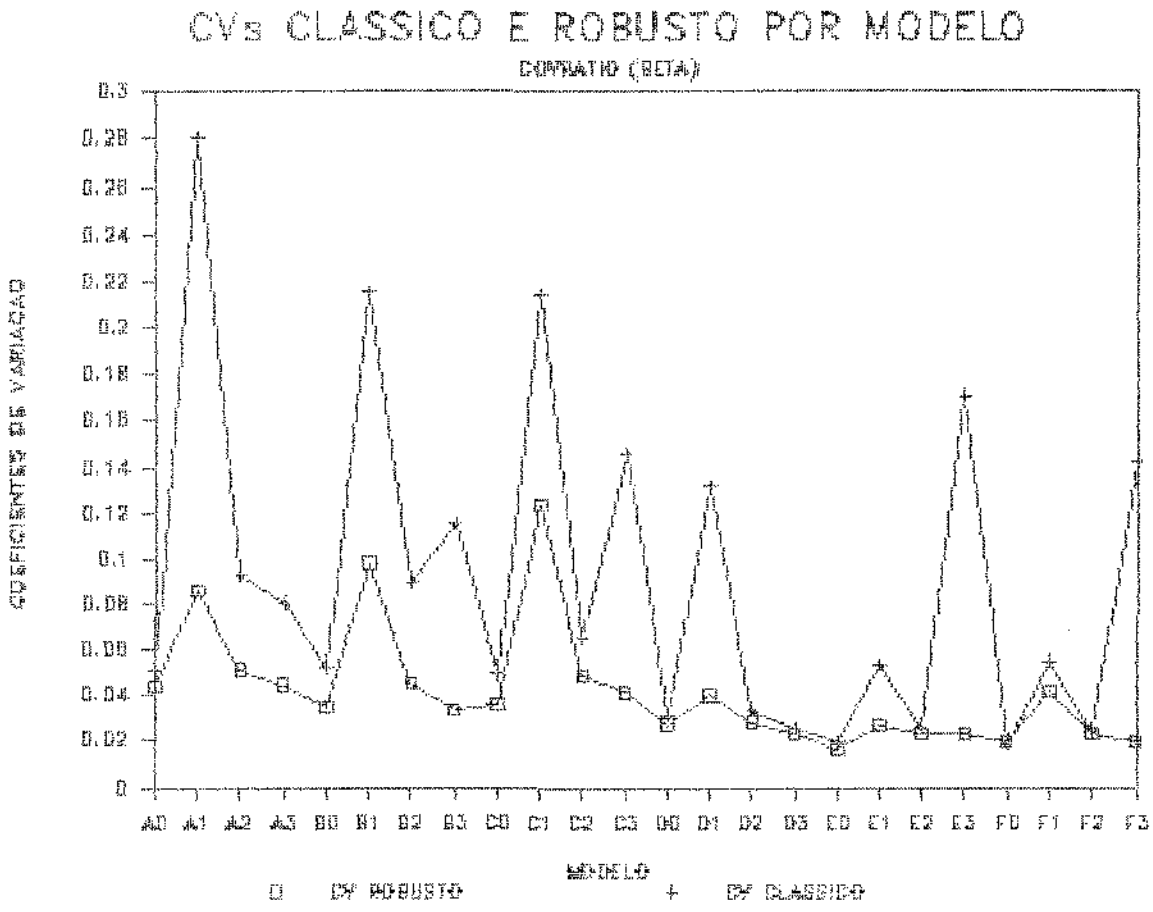

CVs CLASSICO E ROBUSTO POR MODELO
CONTRATO (BETA)
COEFICIENTES DE VARIACAO
0.3
0.28
0.26
0.24
0.22
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3
MODELO
CV ROBUSTO
CV CLASSICO

GRÁFICO 27

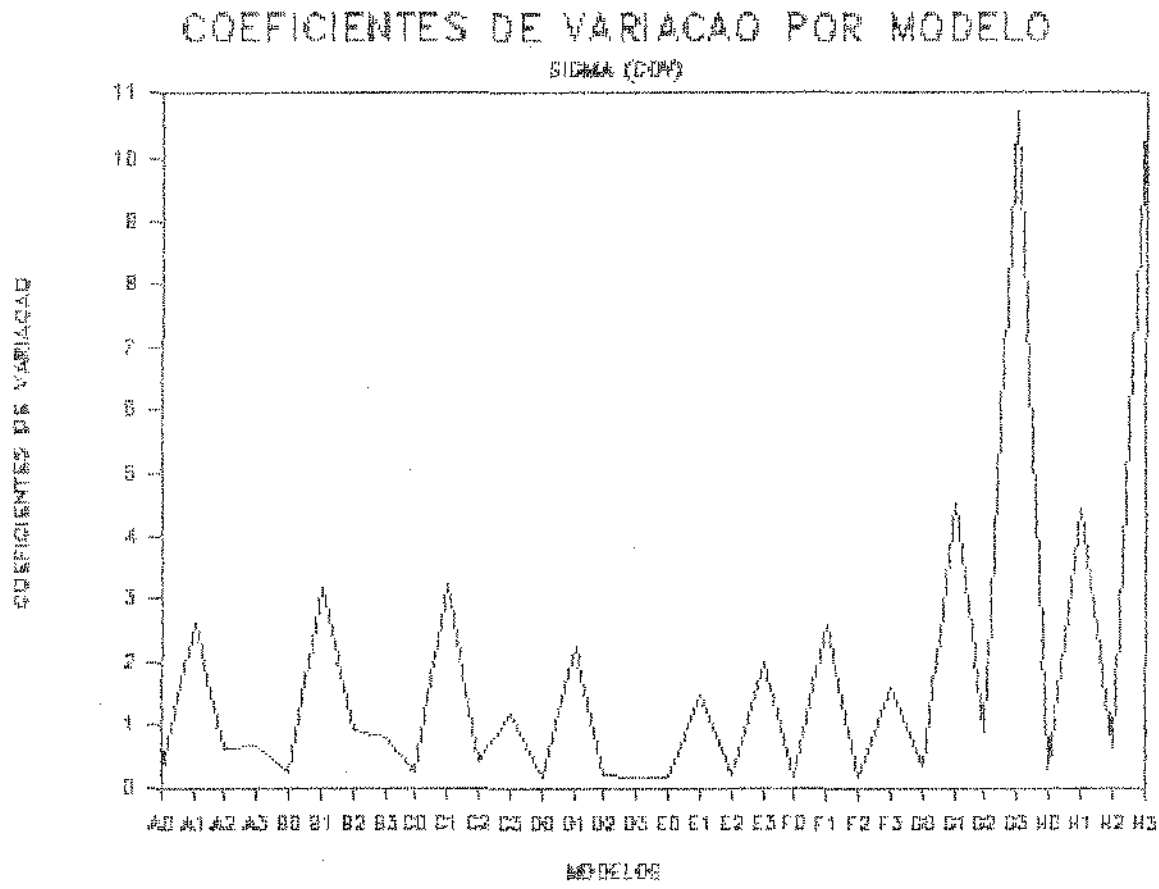

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (CDP)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
9
10
11
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 27A

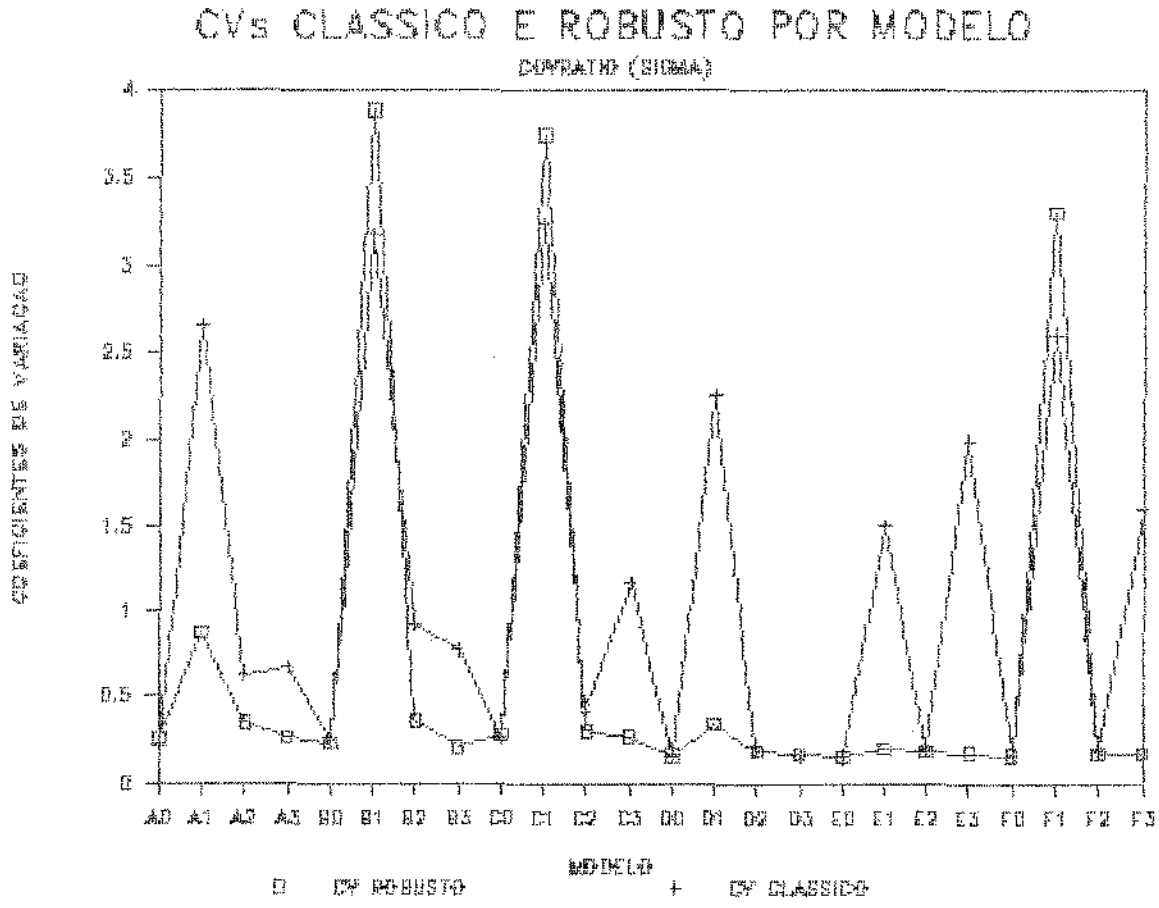

CVs CLASSICO E ROBUSTO POR MODELO
CONTRATO (SIGMA)
COEFICIENTES DE VARIACAO
0
0.5
1
1.5
2
2.5
3
3.5
4
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3
MODELO
CV ROBUSTO
CV CLASSICO

GRÁFICO 28

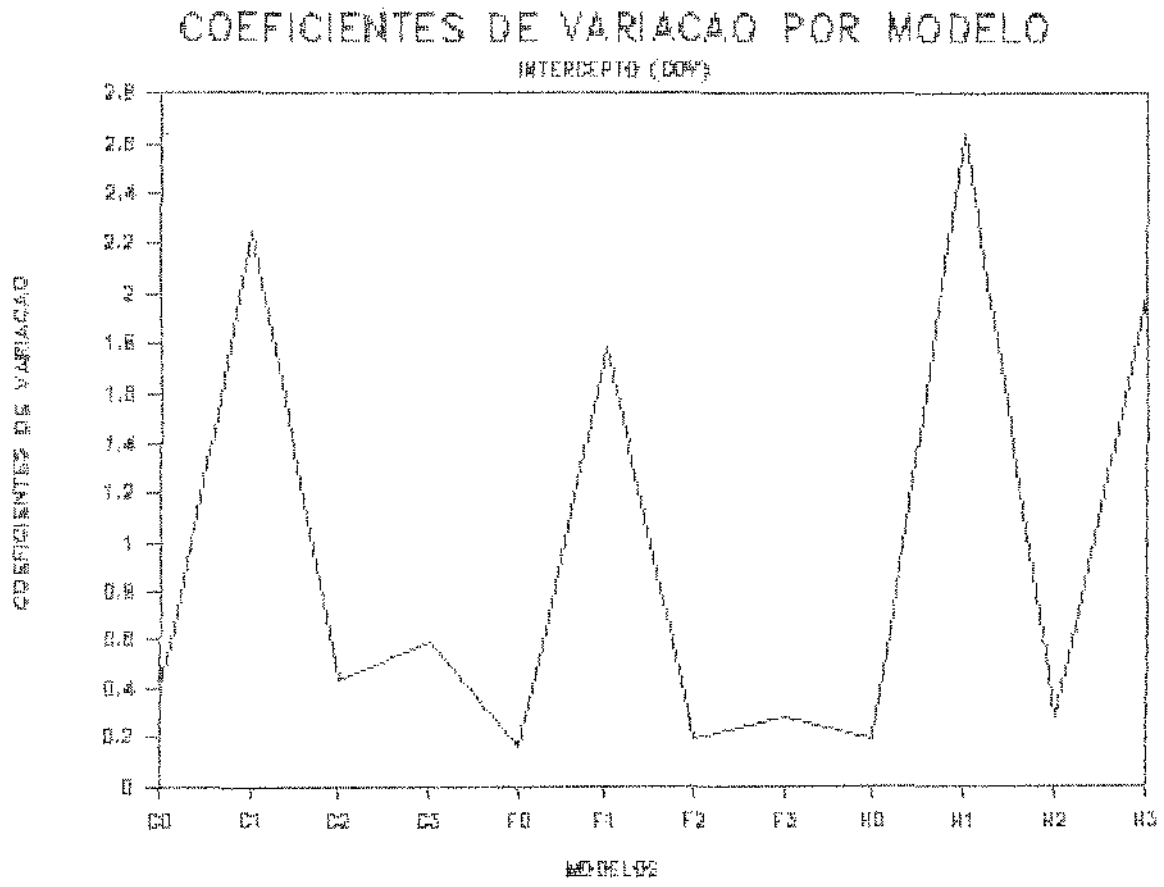

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (DAP)
COEFICIENTE DE VARIACAO
0
0,2
0,4
0,6
0,8
1
1,2
1,4
1,6
1,8
2
2,2
2,4
2,6
2,8
3
C0
C1
C2
C3
F0
F1
F2
F3
H0
H1
H2
H3
MODELOS

GRÁFICO 29

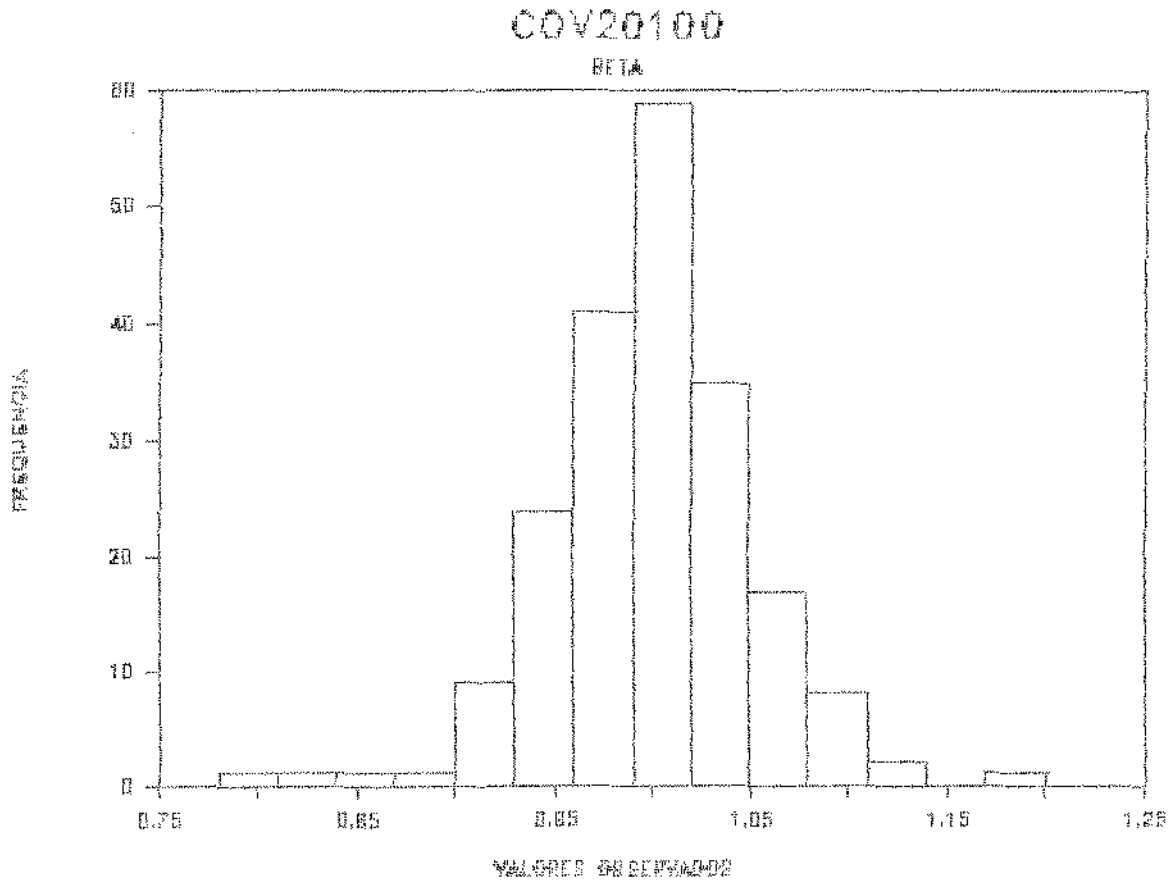


GRÁFICO 30

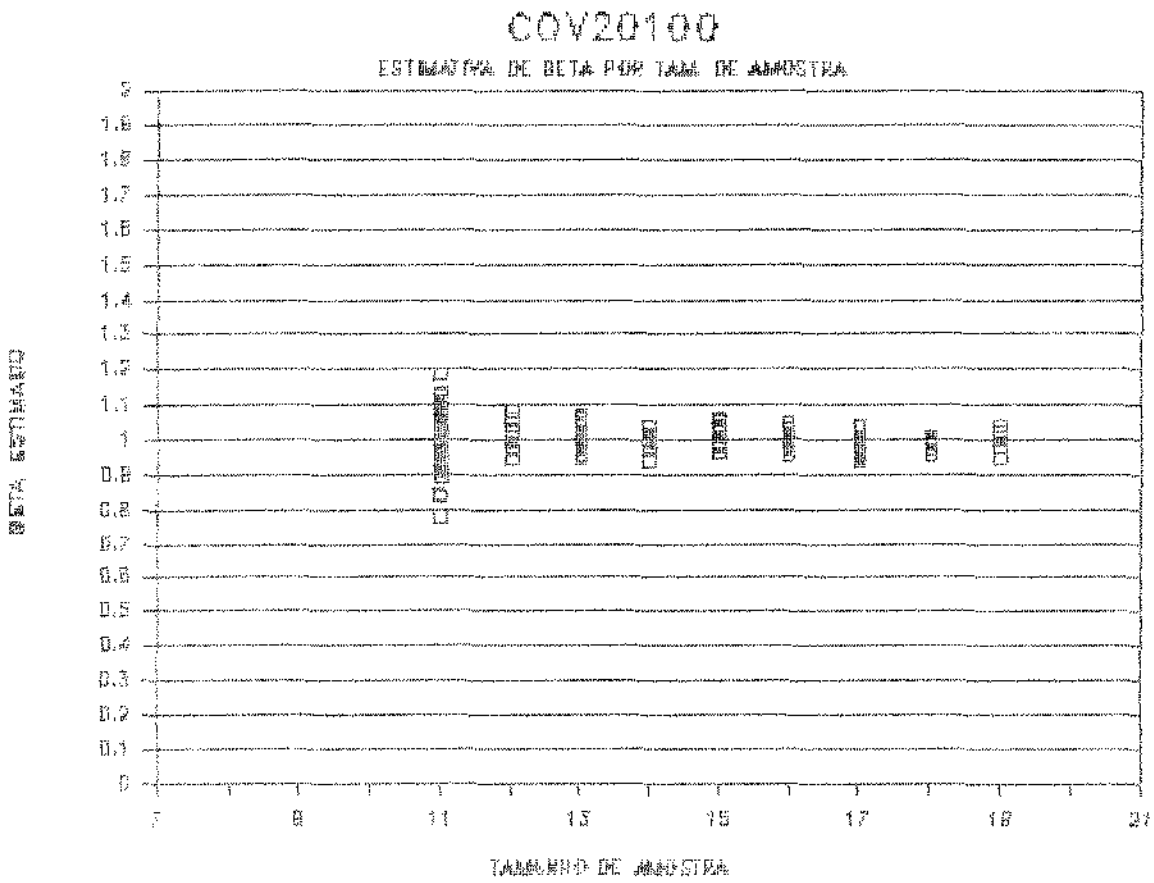

COV20100
ESTIMATIVA DE BETA POR TAM. DE AMOSTRA
ESTIMATIVA BETA
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 31

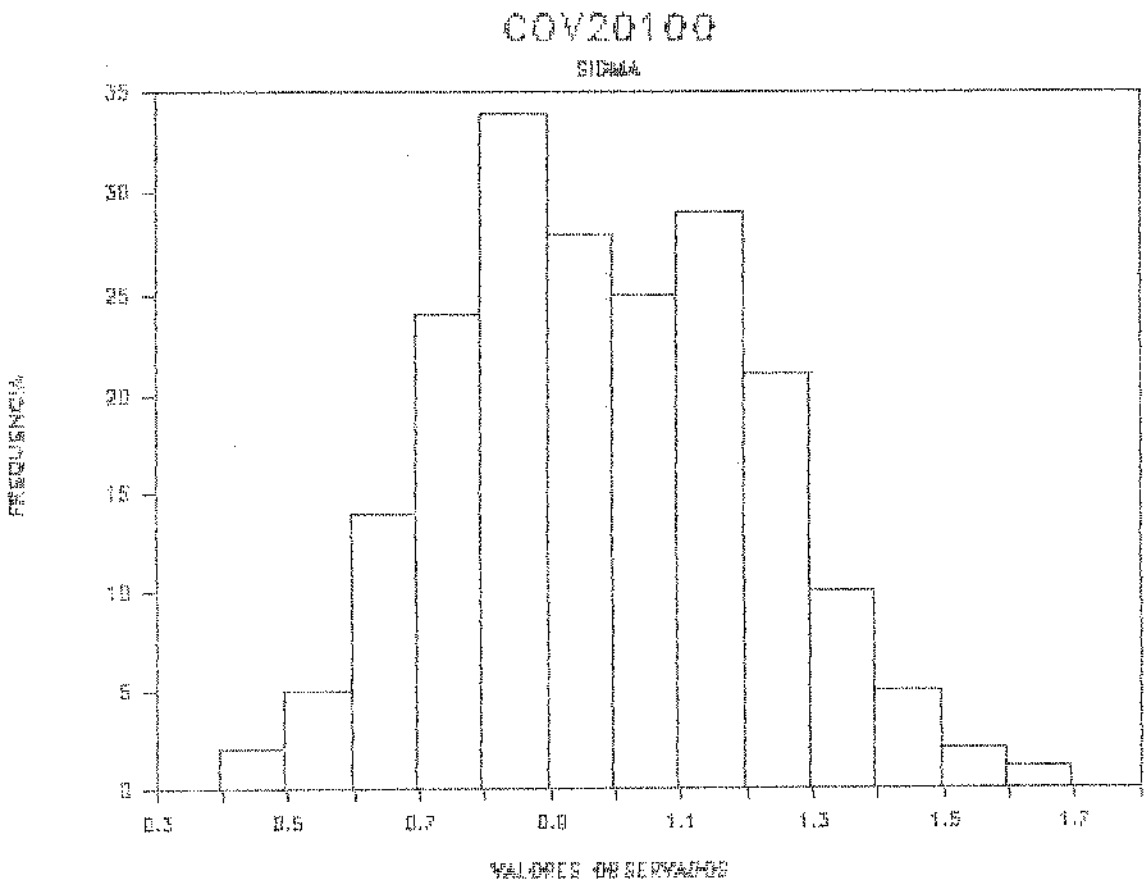

COV20100
SIGMA
FRECUENCIA
0
5
10
15
20
25
30
35
0.3
0.5
0.7
0.9
1.1
1.3
1.5
1.7
VALORES OBSERVADOS

GRÁFICO 32

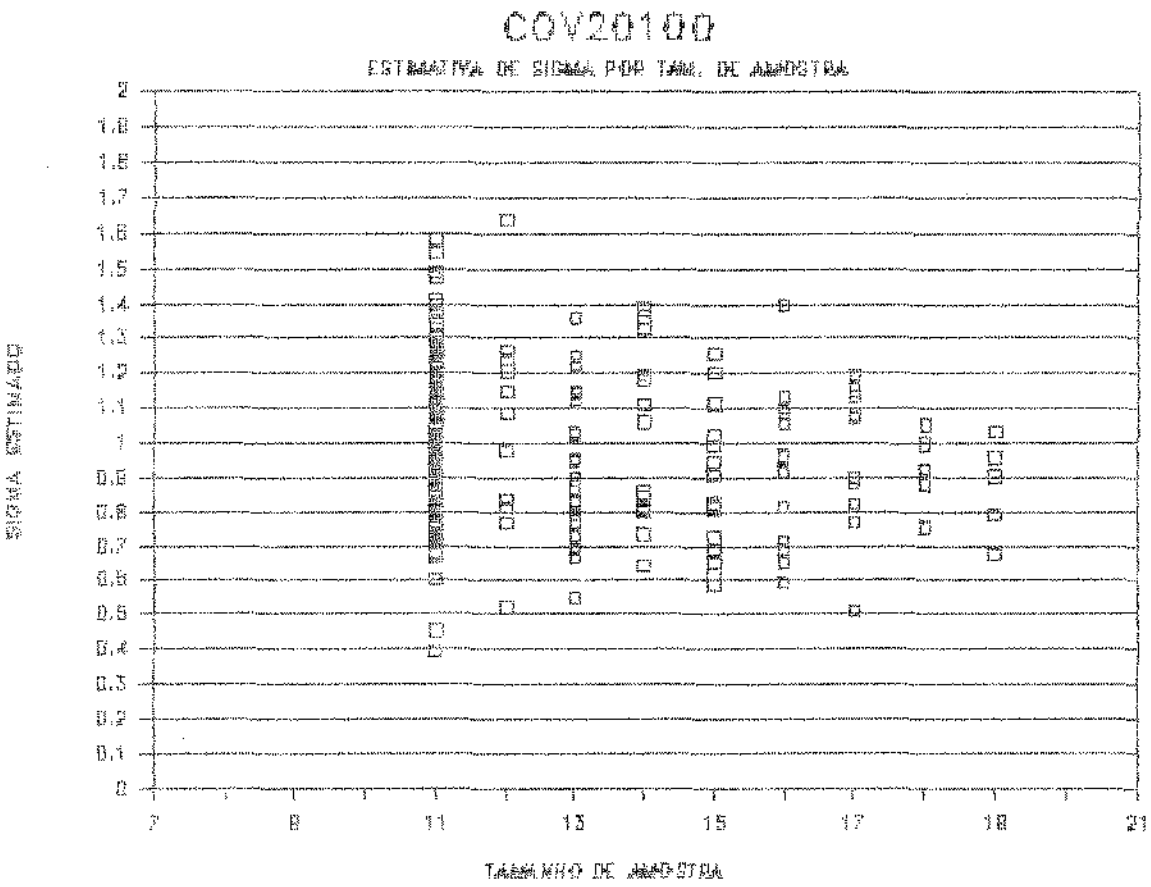

COV20100
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DE AMOSTRA
2
1,8
1,6
1,4
1,2
1
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21

GRÁFICO 33

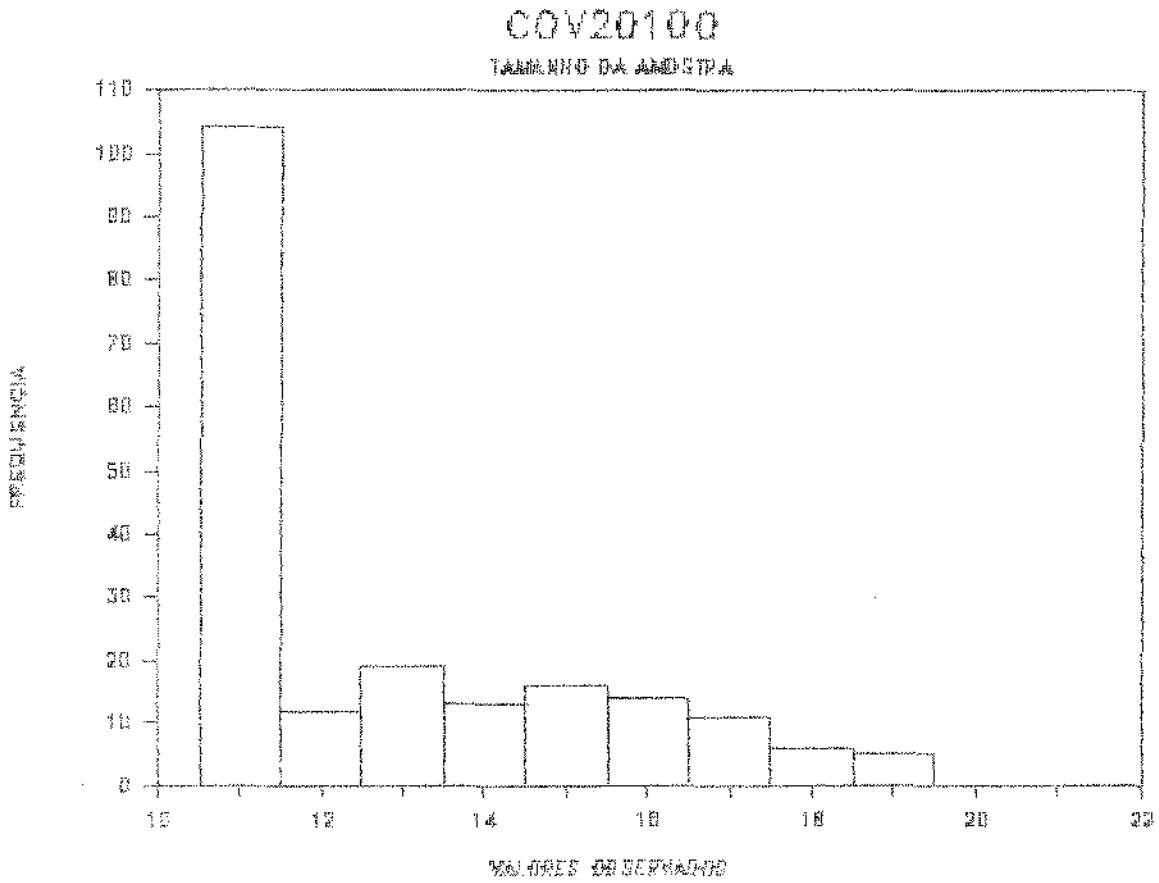

COV20100
TAMANHO DA AMOSTRA
110
100
90
80
70
60
50
40
30
20
10
0
12
13
14
15
16
17
18

GRÁFICO 34

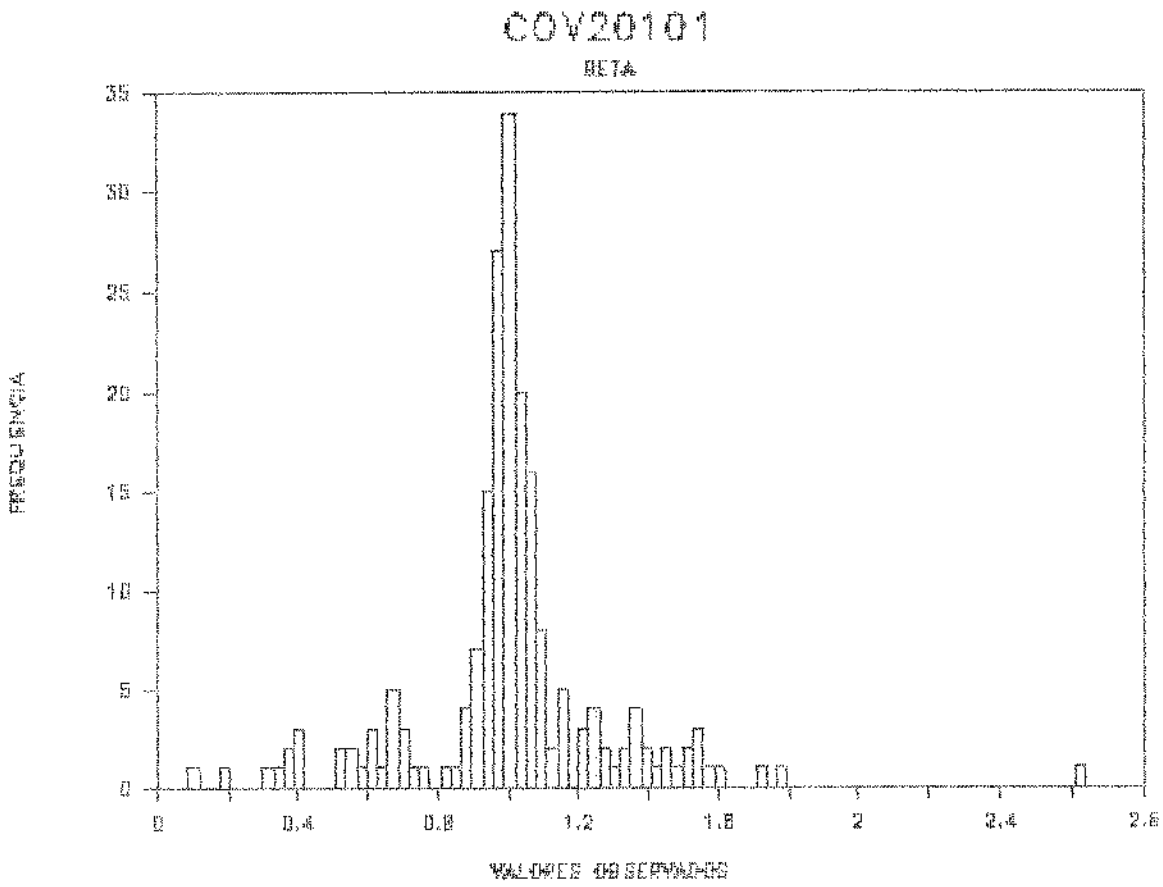

COV20101
BETA
FREQUÊNCIA
0
5
10
15
20
25
30
35
0
0,4
0,8
1,2
1,6
2
2,4
2,8
VALORES OBSERVADOS

GRÁFICO 35

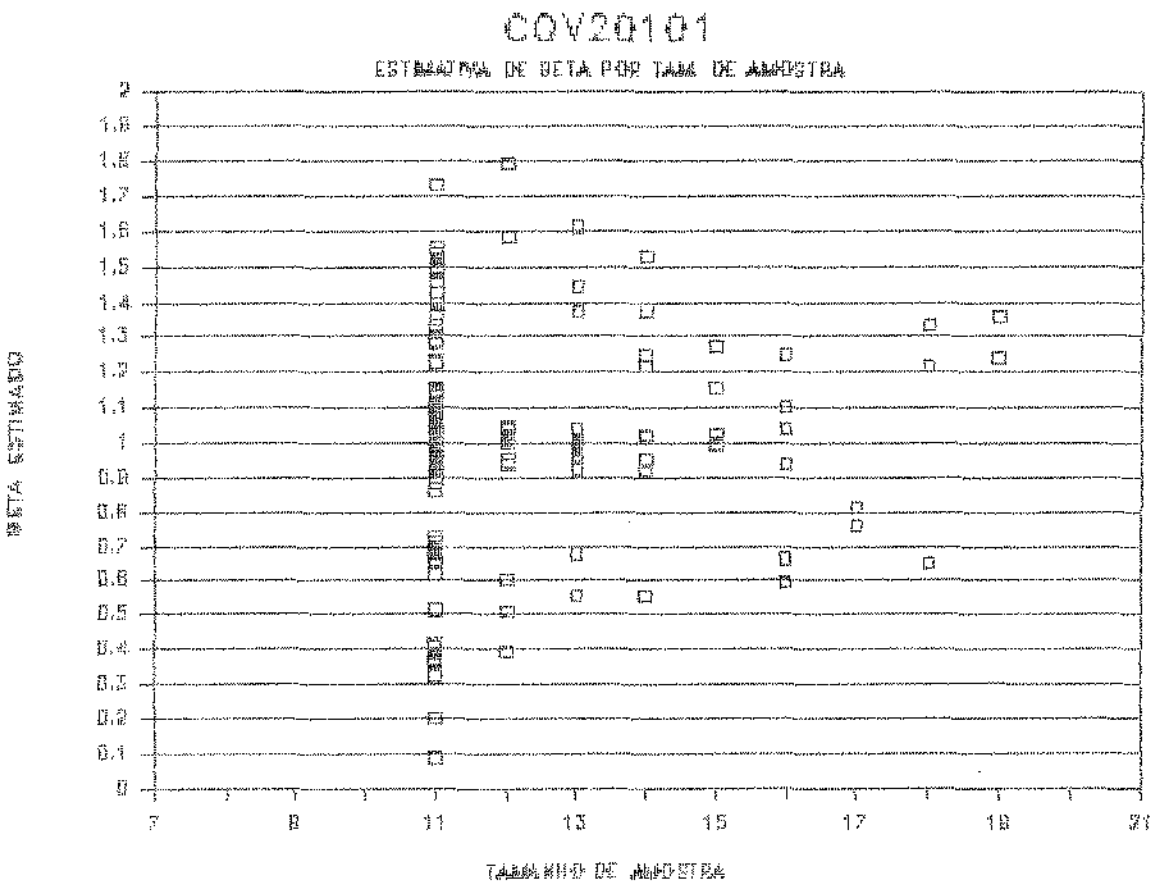

COV20101
ESTIMATIVA DE BETA POR TAM DE AMOSTRA
BETA ESTIMADO
TAMANHO DE AMOSTRA
2
1,8
1,6
1,5
1,4
1,2
1
0,8
0,6
0,5
0,4
0,2
0
7
9
11
13
15
17
19
21

GRÁFICO 36

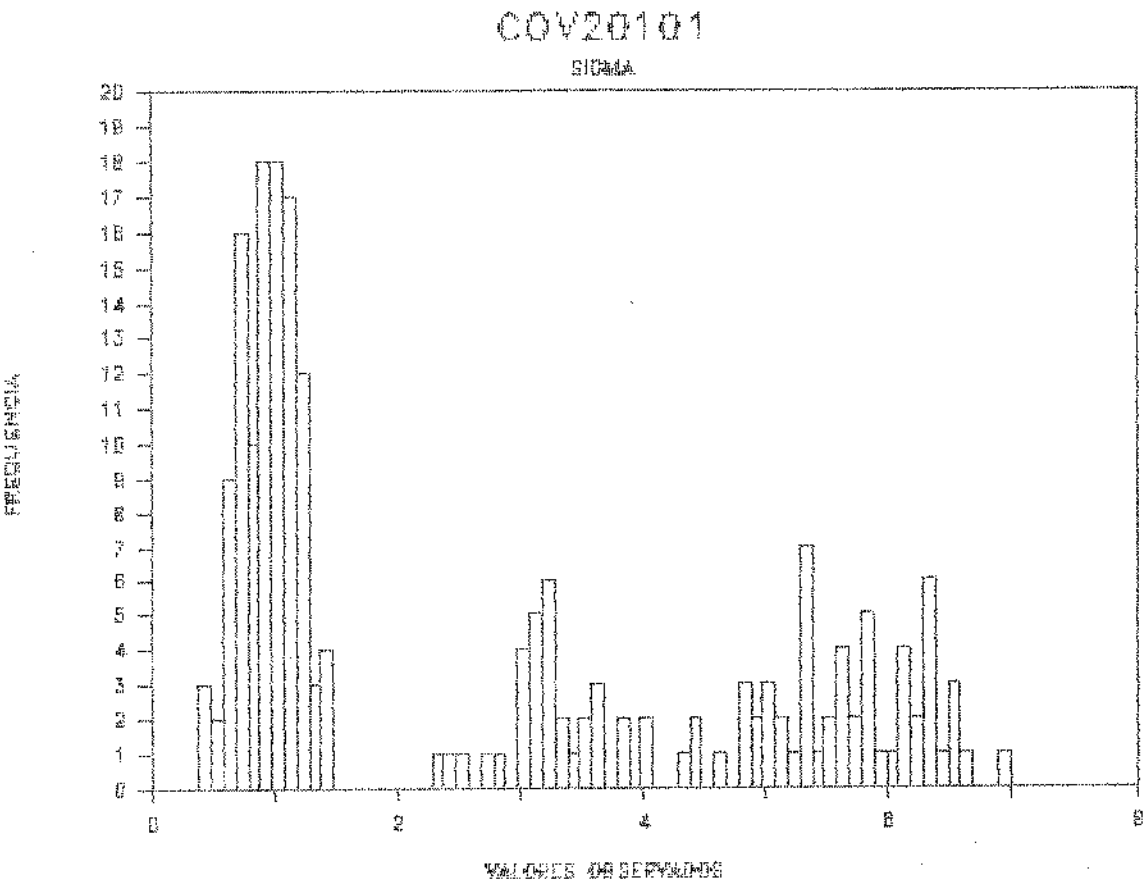

COV20101
SIGMA
FRECUENCIA
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0
2
4
6
8
VALORES OBSERVADOS

GRÁFICO 37

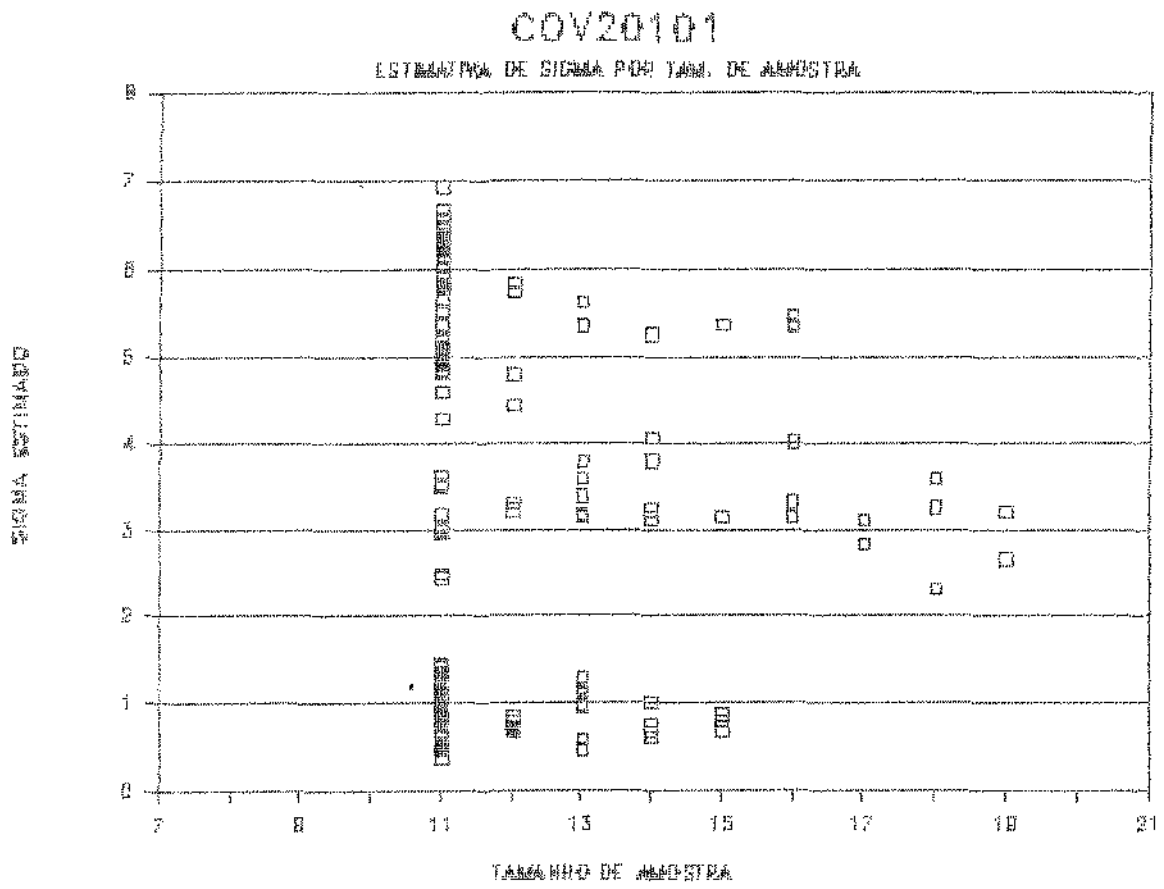

COV20101
ESTIMACION DE SIGMA POR TAM. DE MUESTRA
SIGMA ESTIMADO
0
1
2
3
4
5
6
7
8
7
8
11
13
15
17
19
21
TAMAÑO DE MUESTRA

GRÁFICO 38

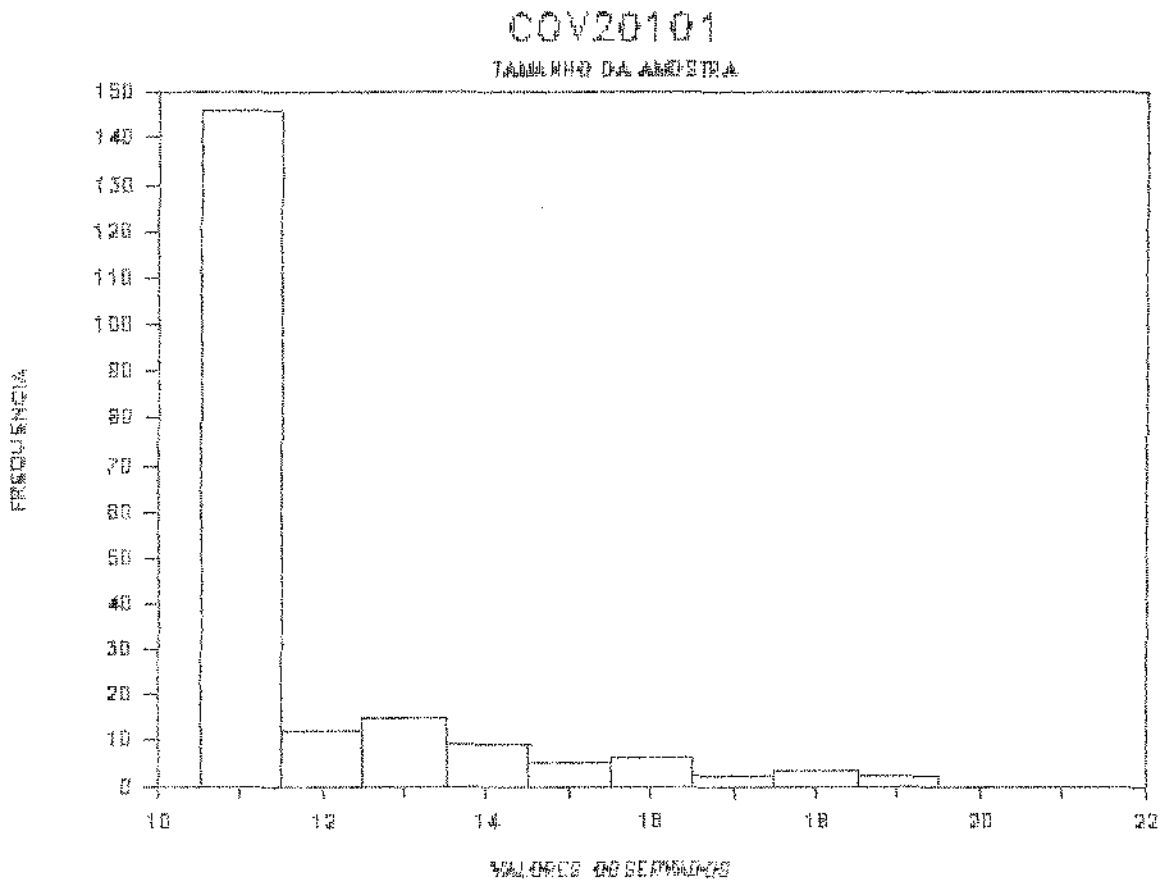

COV20101
TAMANHO DA AMOSTRA
FREQUÊNCIA
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES DO SEGMENTO

GRÁFICO 39

COV20102

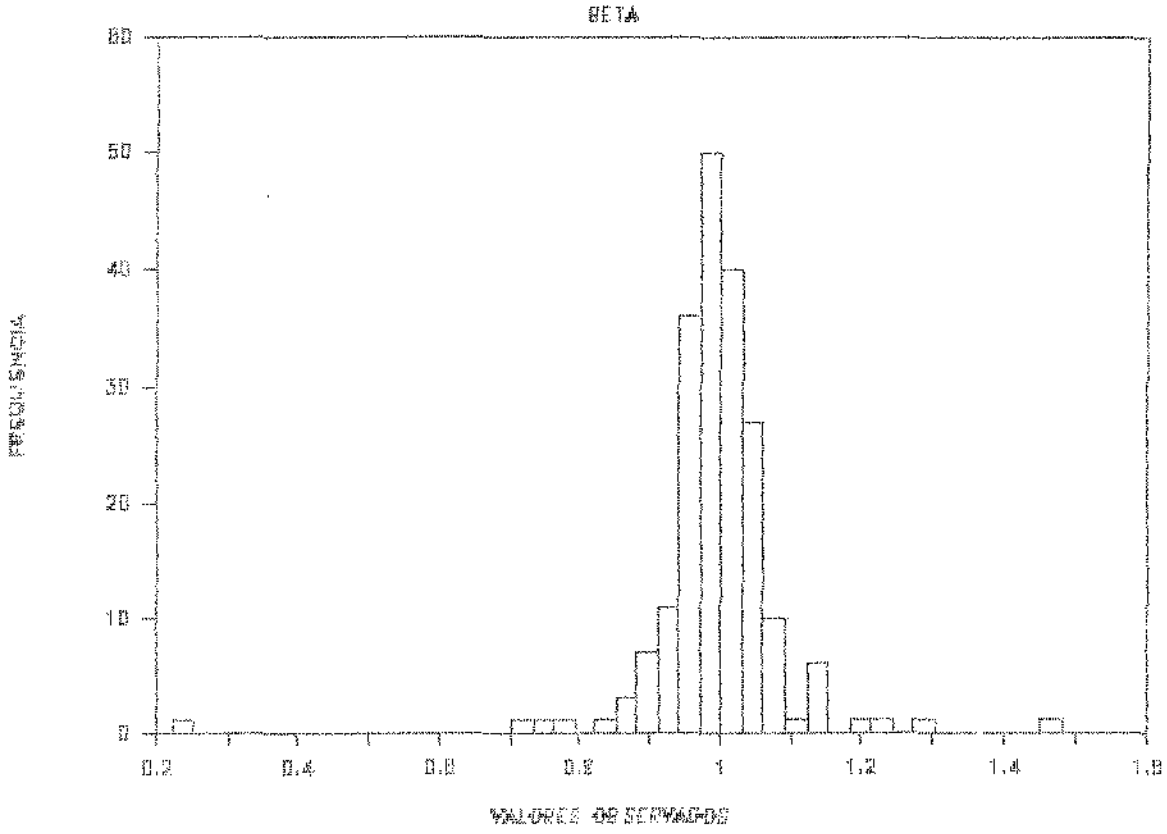


GRÁFICO 40

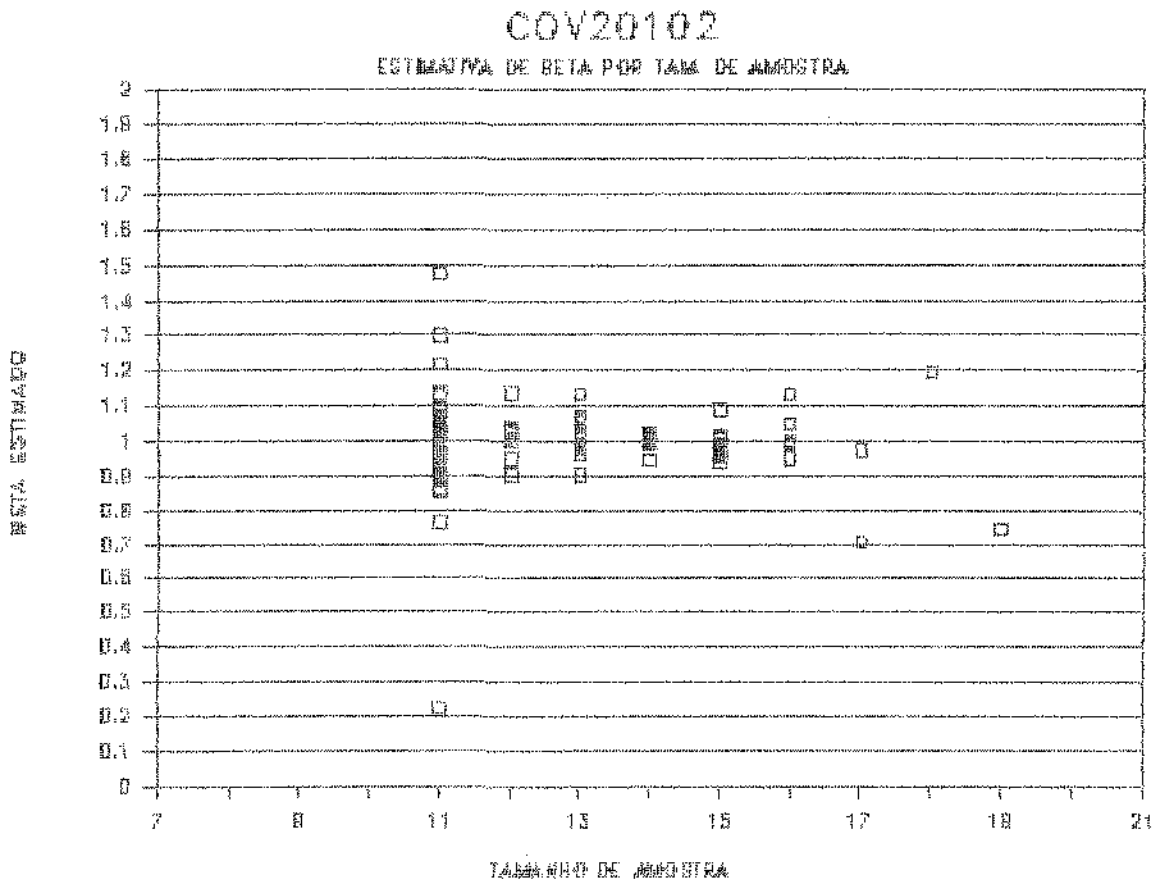

COV20102
ESTIMATIVA DE BETA POR TAM DE AMOSTRA
BETA ESTIMADO
TAMANHO DE AMOSTRA
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21

GRÁFICO 41

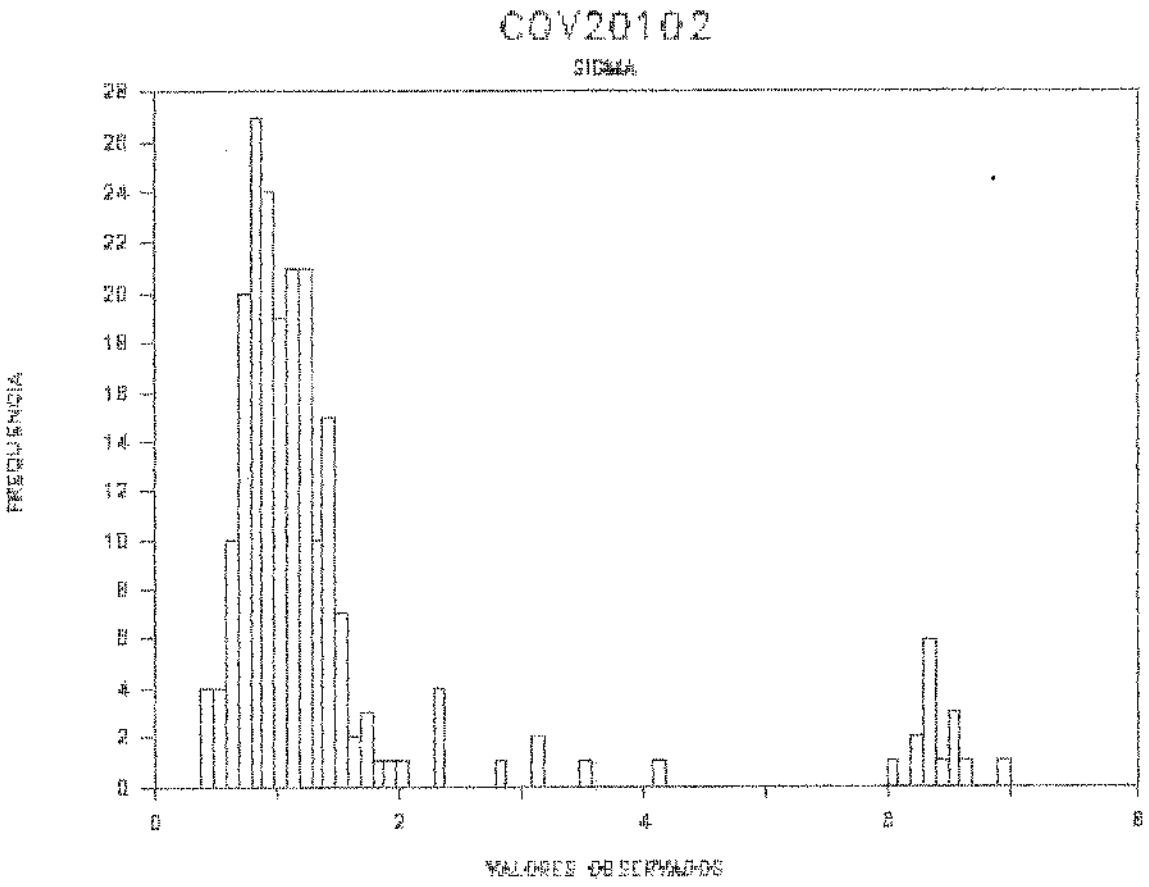


GRÁFICO 42

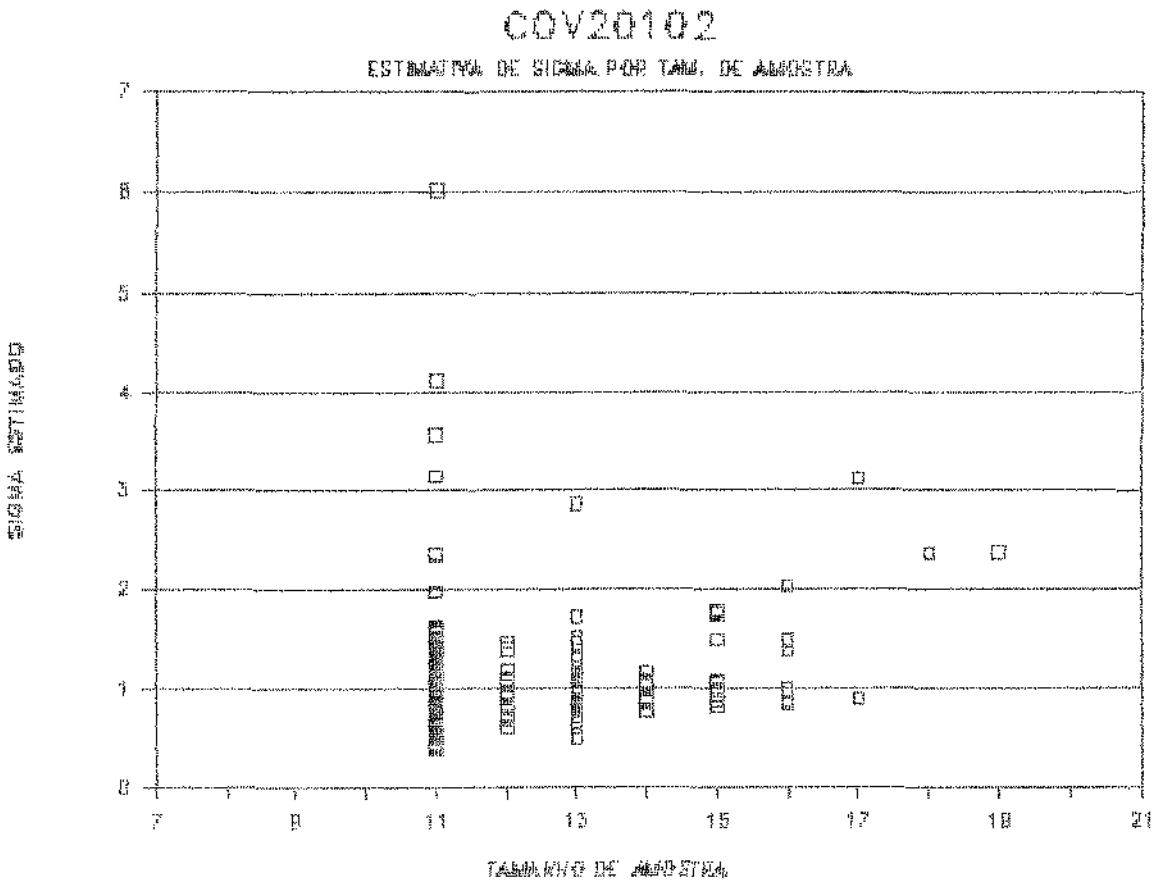

COV20102
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DE AMOSTRA
7
6
5
4
3
2
1
0
7
9
11
13
15
17
19
21

GRÁFICO 43

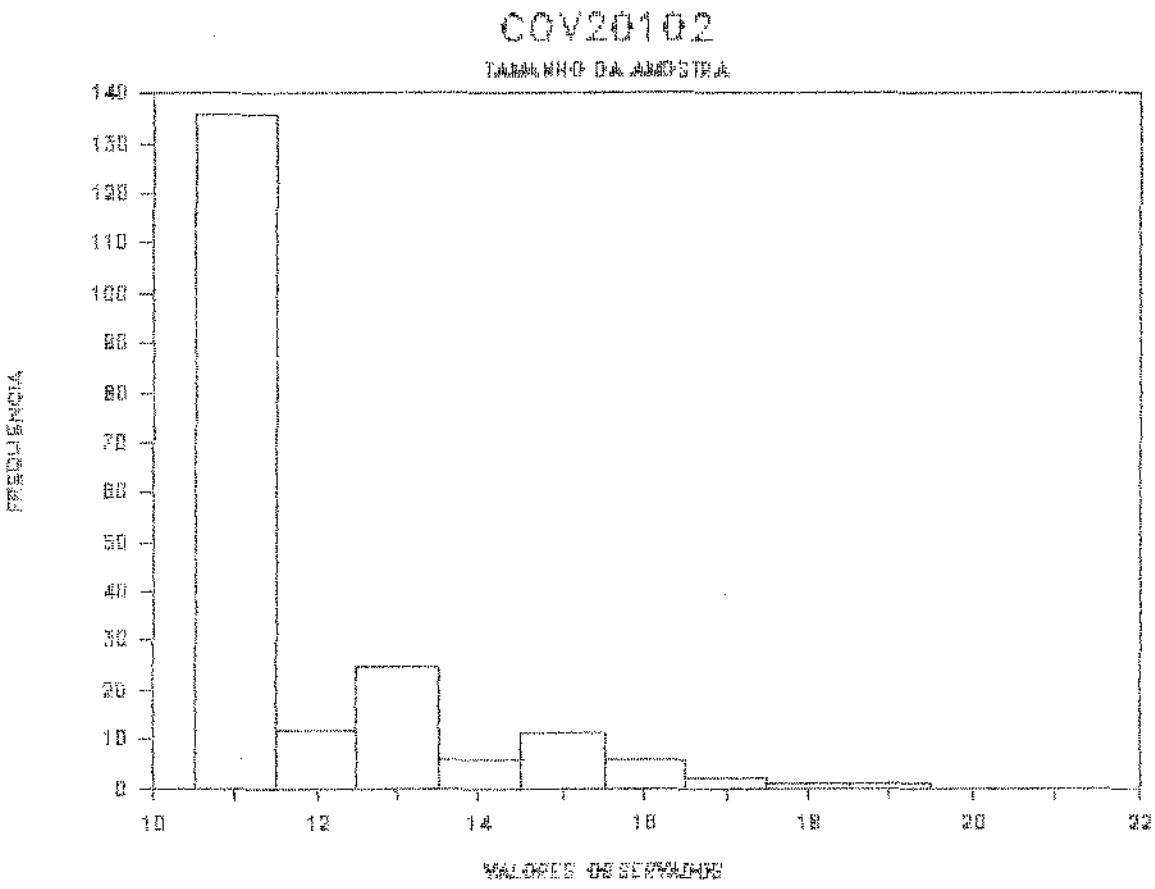

COV20102
TAMANHO DA AMOSTRA
FREQUÊNCIA
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
10
12
14
16
18
20
22
VALORES DE SEPARAÇÃO

GRÁFICO 44

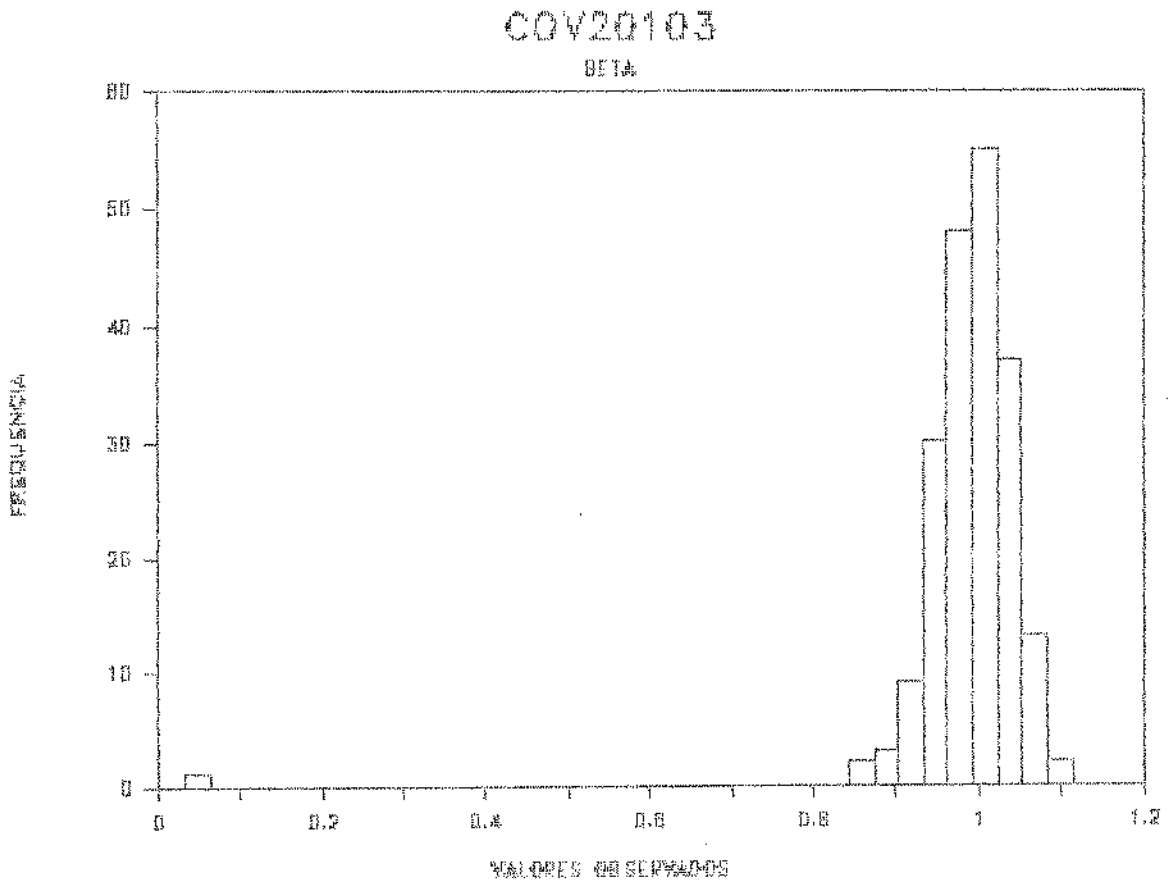


GRÁFICO 45

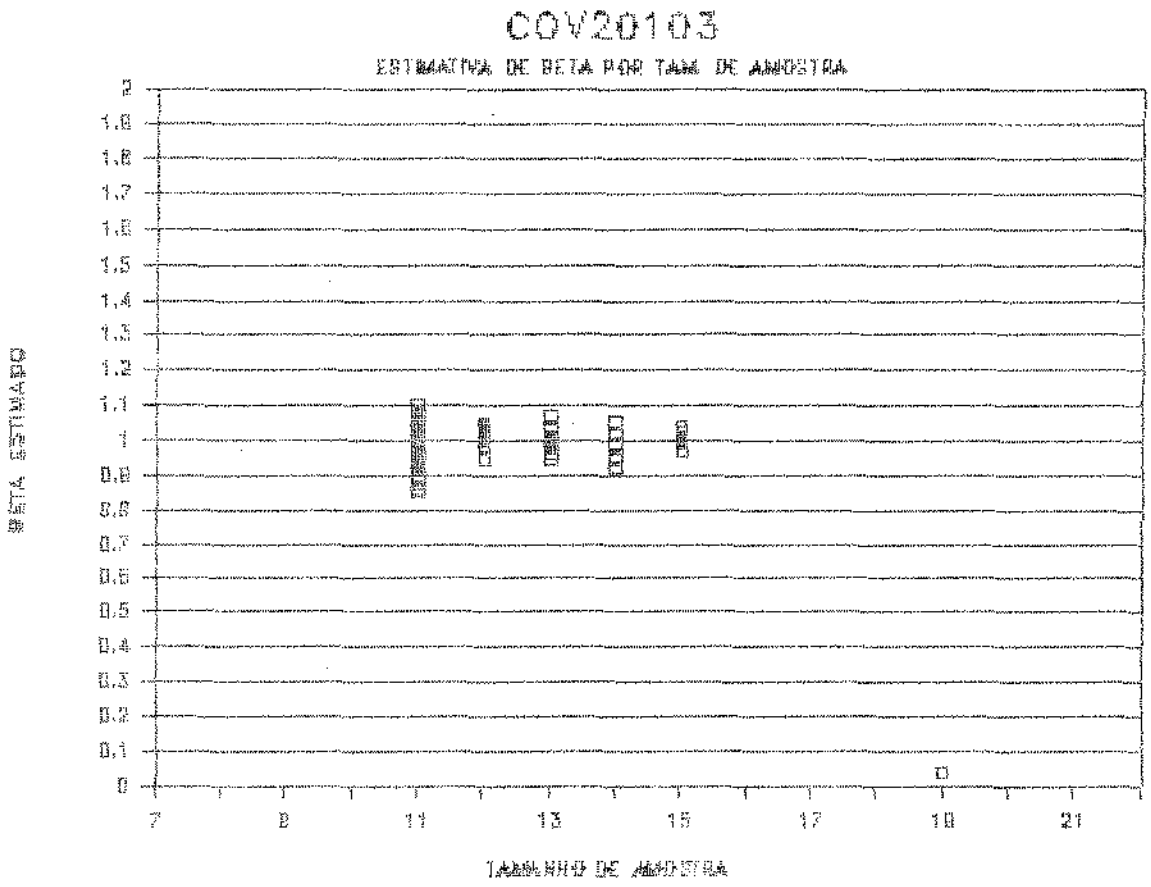

COV20103
ESTIMATIVA DE BETA POR TAM. DE AMOSTRA
BETA ESTIMADO
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 46

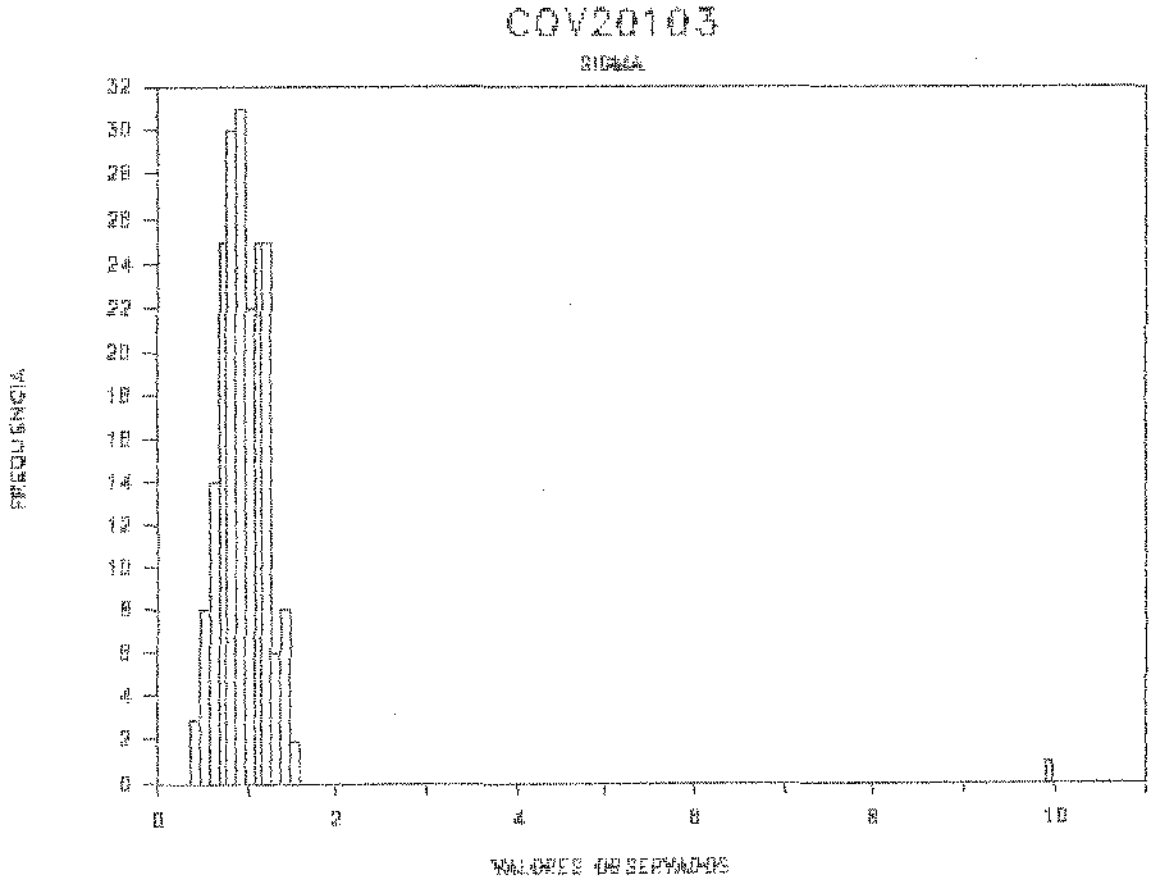

COV20103
SIGMA
FRECUENCIA
0
2
4
6
8
10
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0

GRÁFICO 47

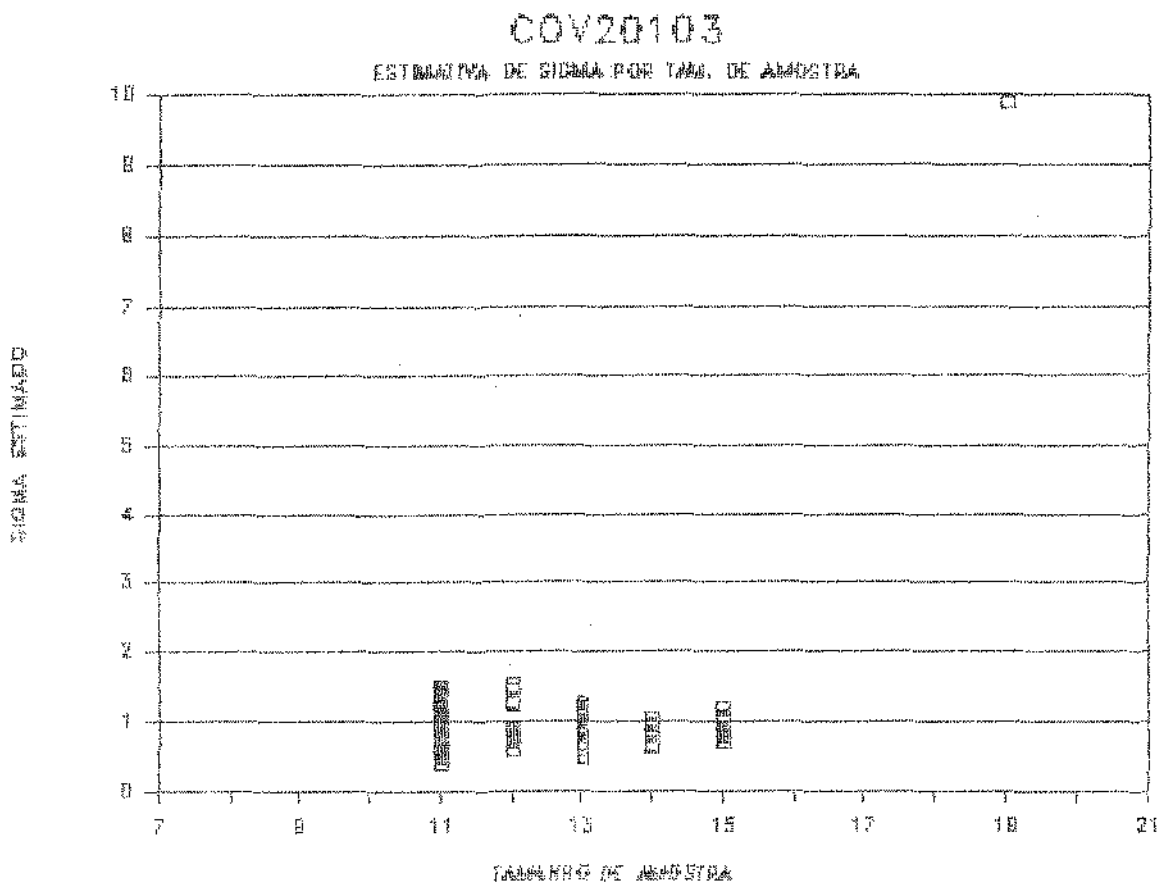

COV20103
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DE AMOSTRA
0
1
2
3
4
5
6
7
8
9
10
7
9
11
13
15
17
19
21

GRÁFICO 48

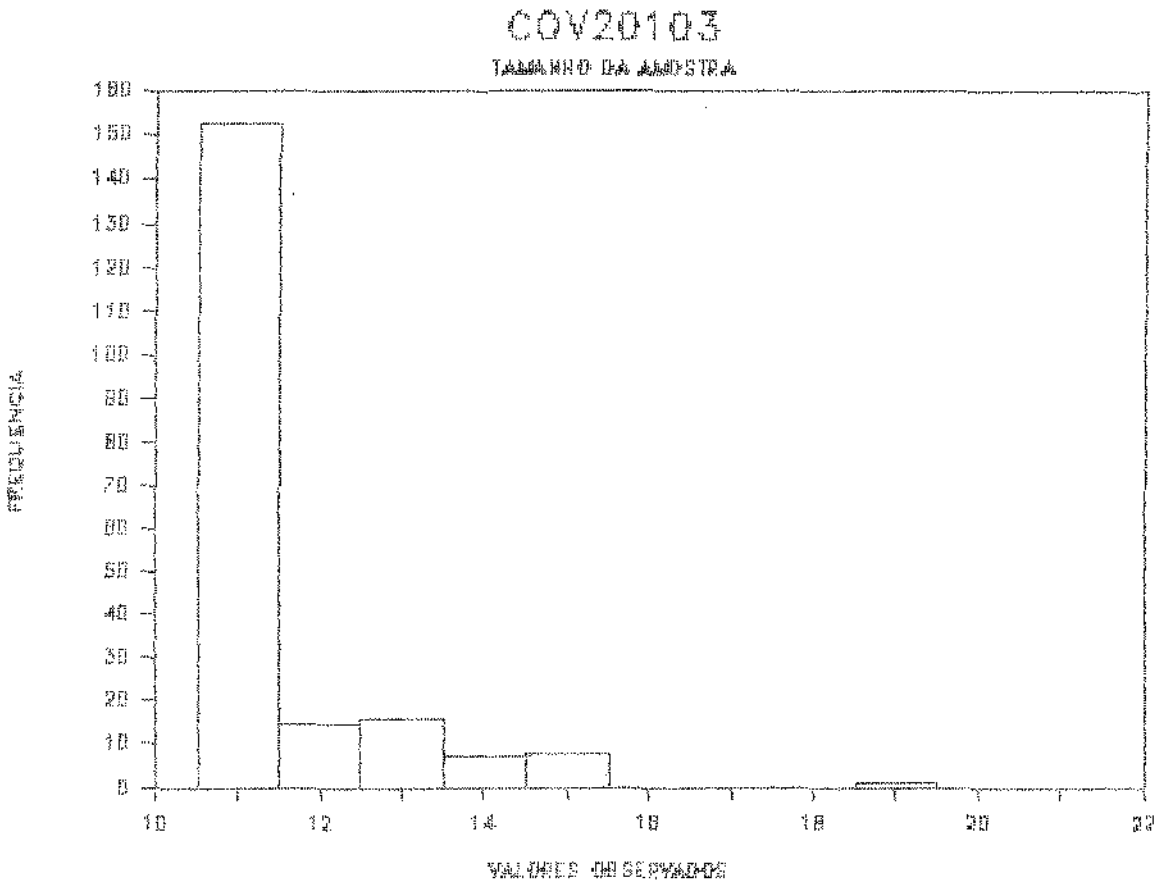

COV2103
TAMANHO DA AMOSTRA
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 49

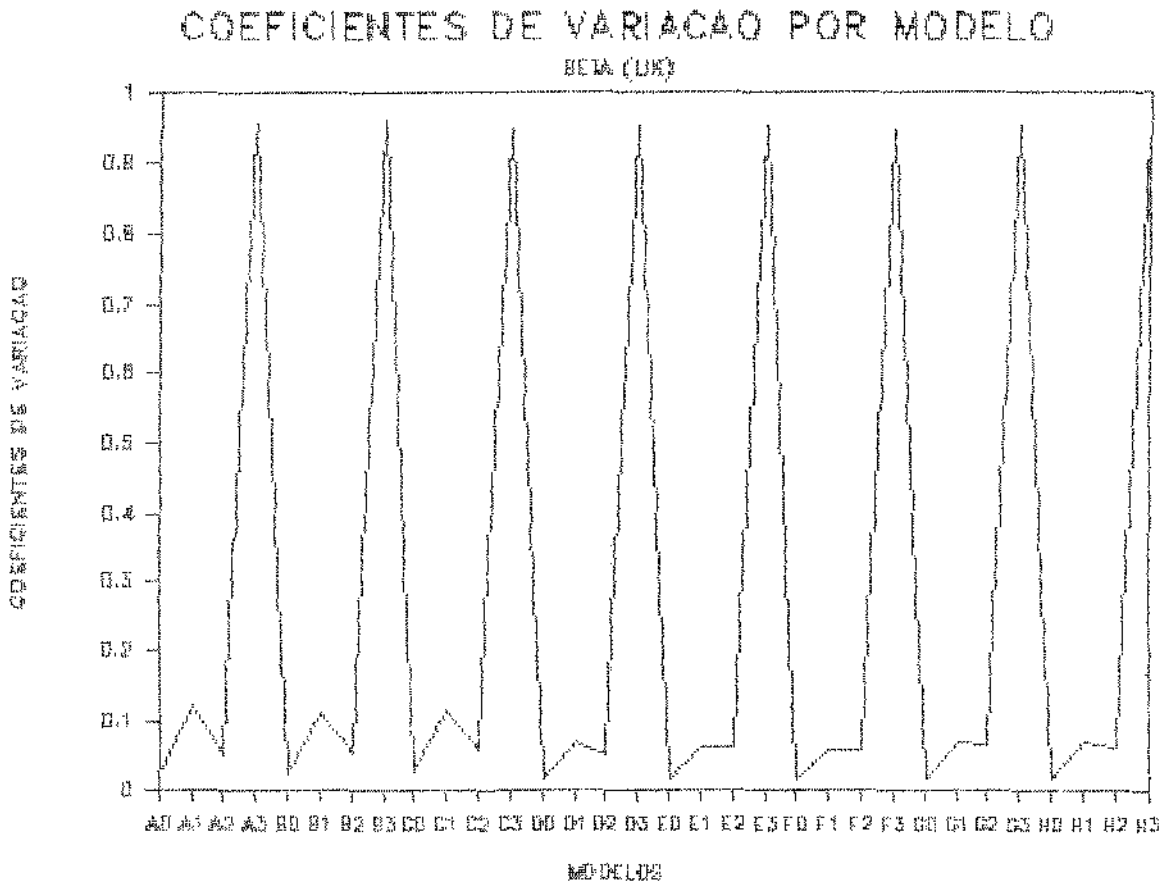

COEFICIENTES DE VARIACAO POR MODELO
BETA (UM)
COEFICIENTES DE VARIACAO
1
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 50

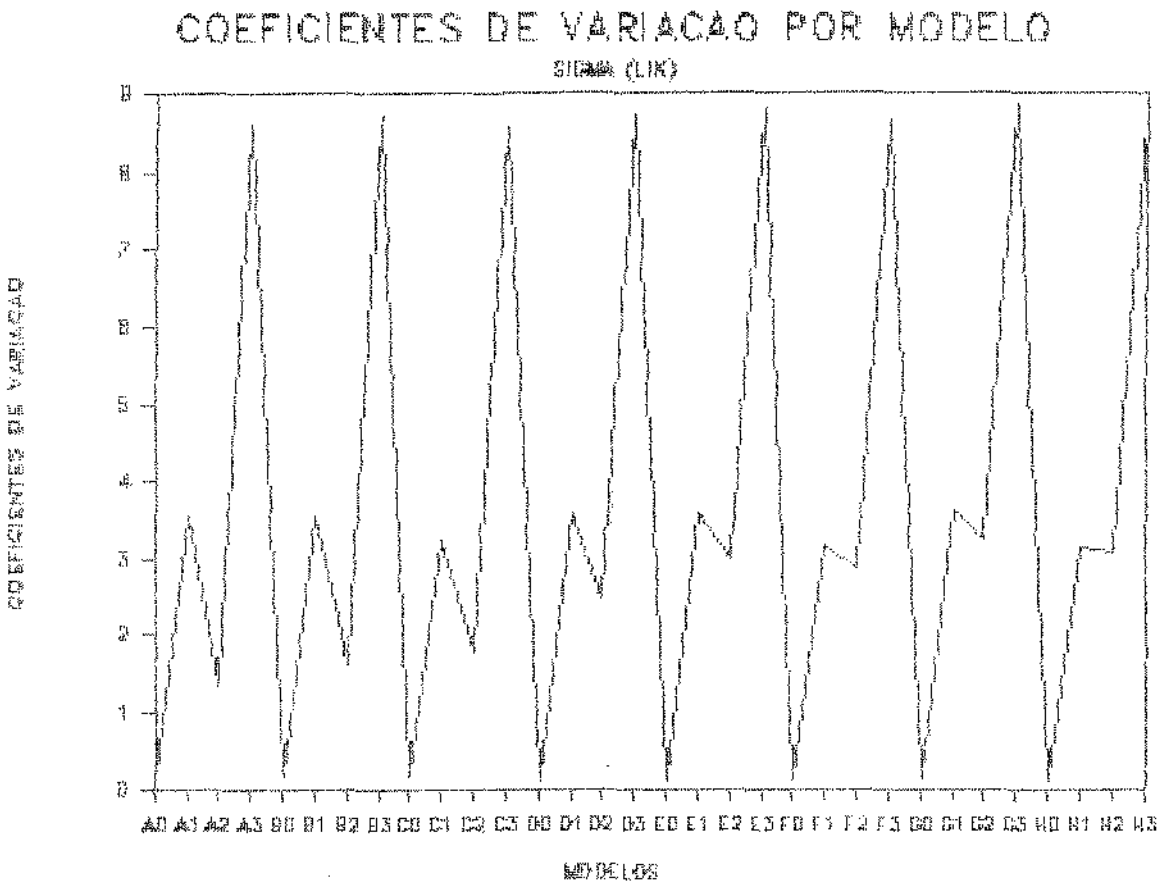

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (LIN)
COEFICIENTES DE VARIACAO
MODELOS

GRÁFICO 51

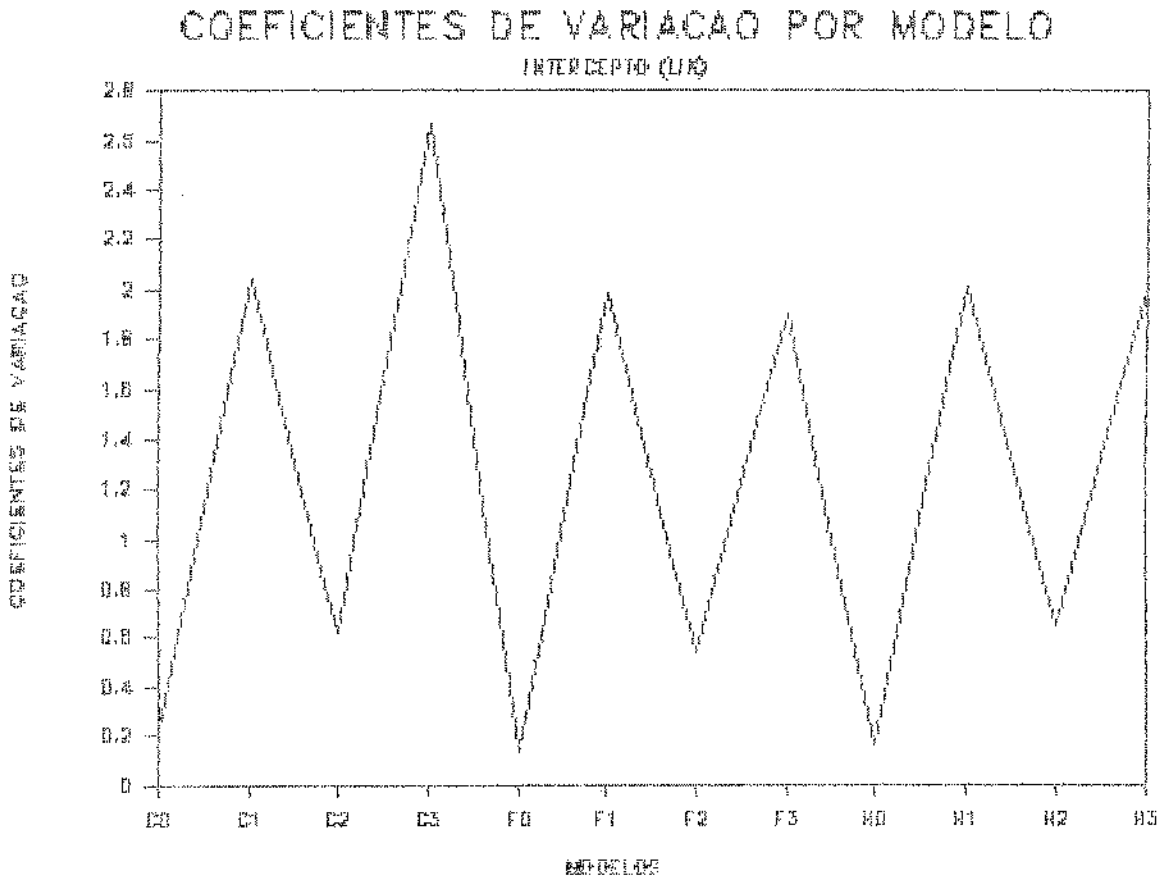

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (LIX)
COEFICIENTES DE VARIACAO
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
C0
C1
C2
C3
F0
F1
F2
F3
H0
H1
H2
H3
MODELOS

GRÁFICO 52

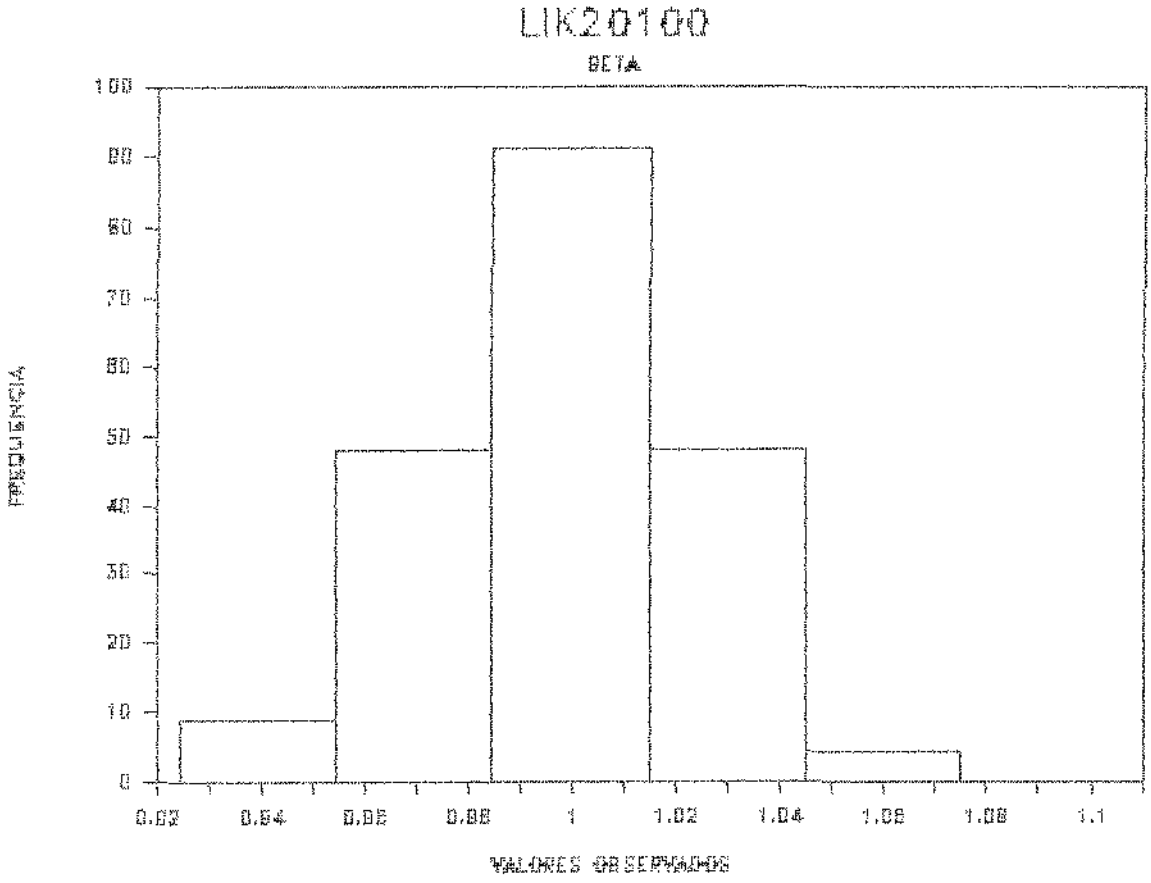

LIK20100
BETA
100
90
80
70
60
50
40
30
20
10
0
FREQUENCIA
0.92
0.94
0.96
0.98
1
1.02
1.04
1.06
1.08
1.1
VALORES OBSERVADOS

GRÁFICO 53

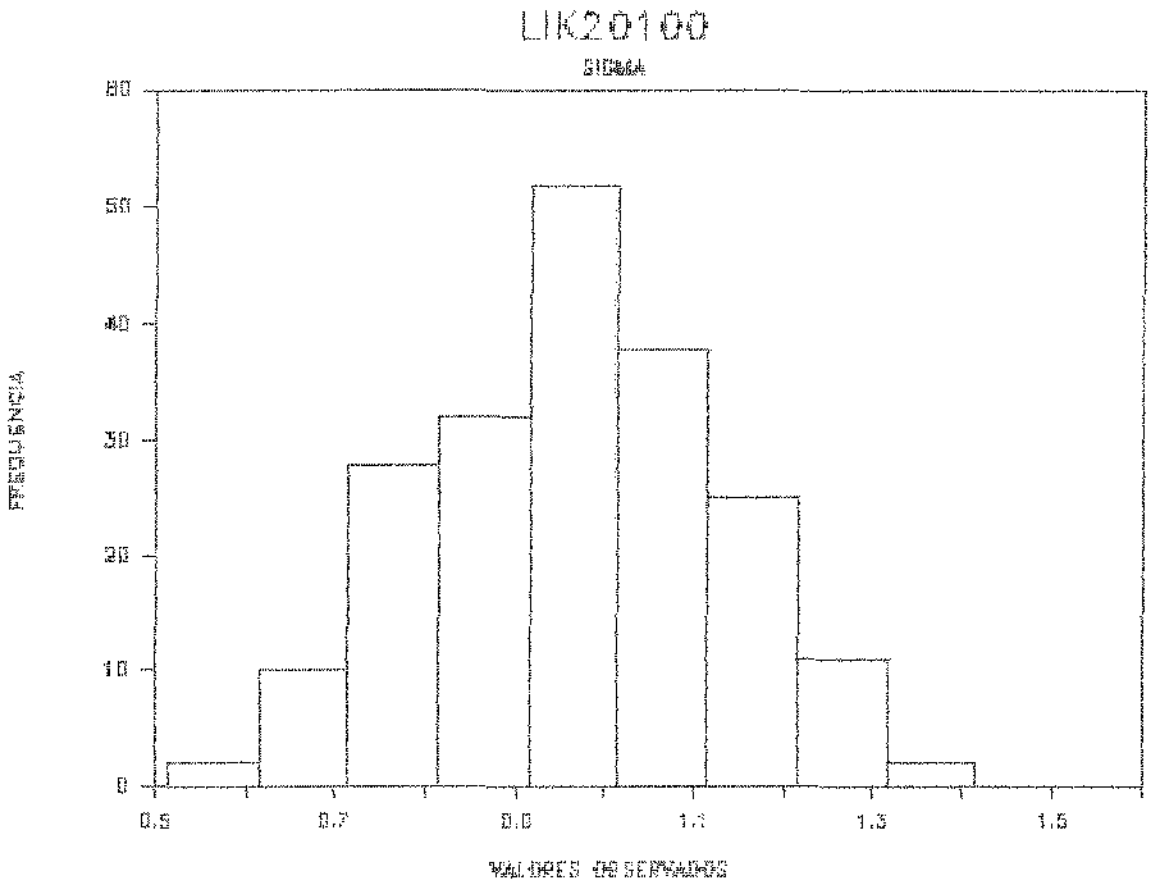

LIK20100
FREQUÊNCIA
80
60
40
30
20
10
0
0.5
0.7
0.9
1.1
1.3
1.5
VALORES OBSERVADOS

GRÁFICO 54

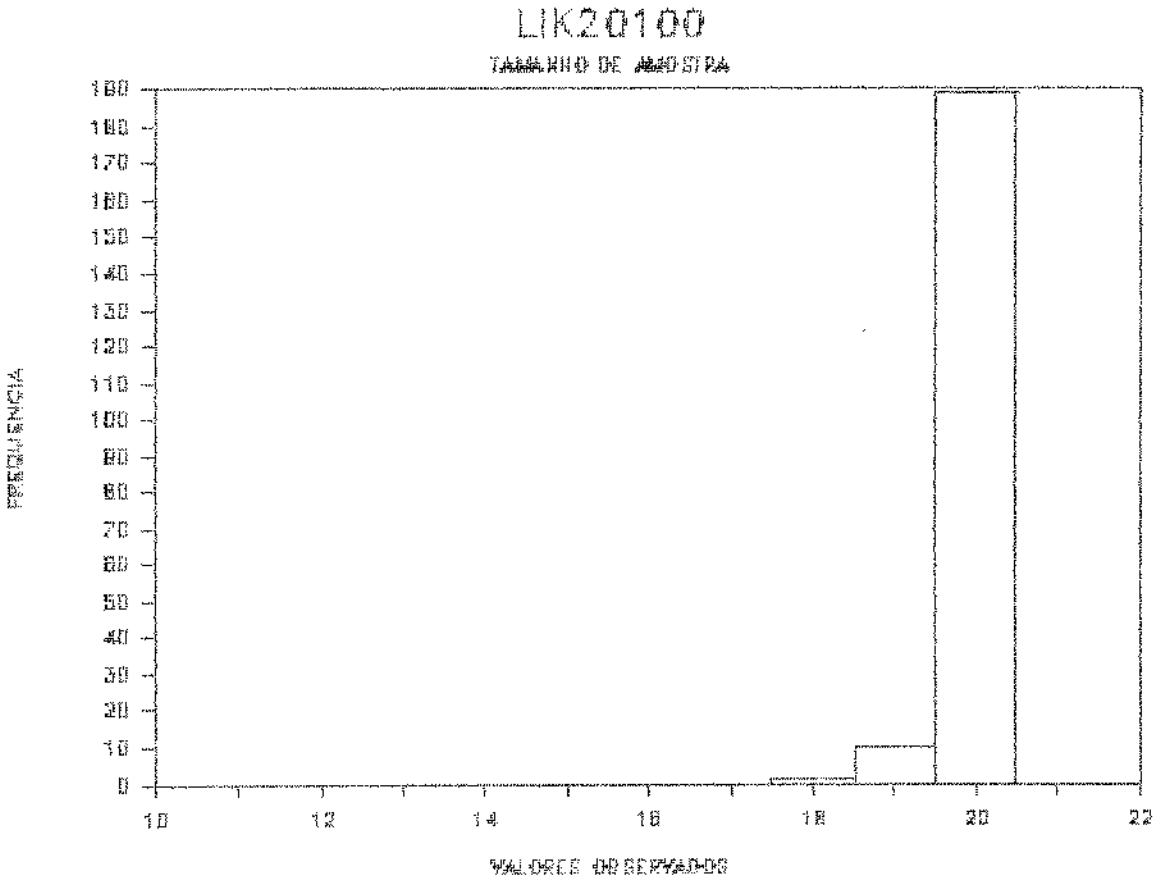

LIK20100
TAMANHO DE AMOSTRA
FREQUÊNCIA
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 55

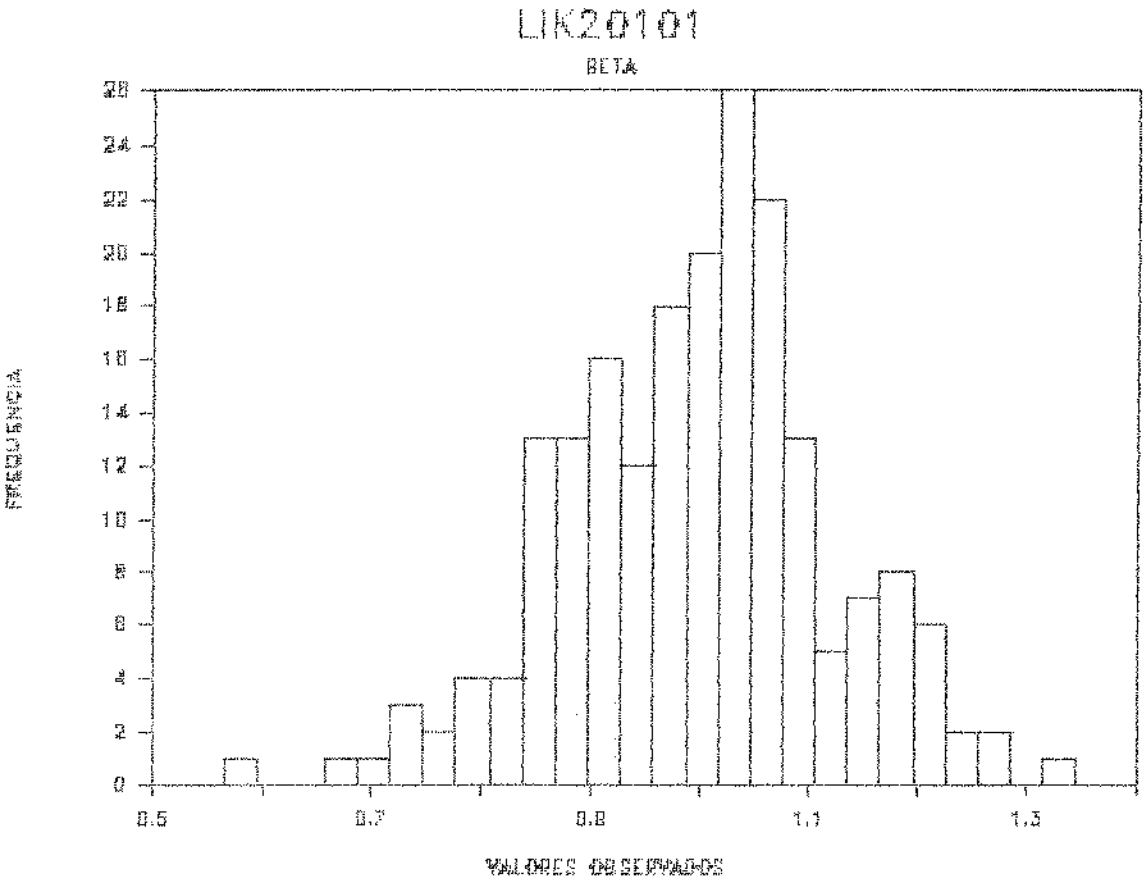

LIK20101
BETA
FREQUENCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
26
0.5
0.7
0.8
1.1
1.3
VALORES OBSERVADOS

GRÁFICO 56

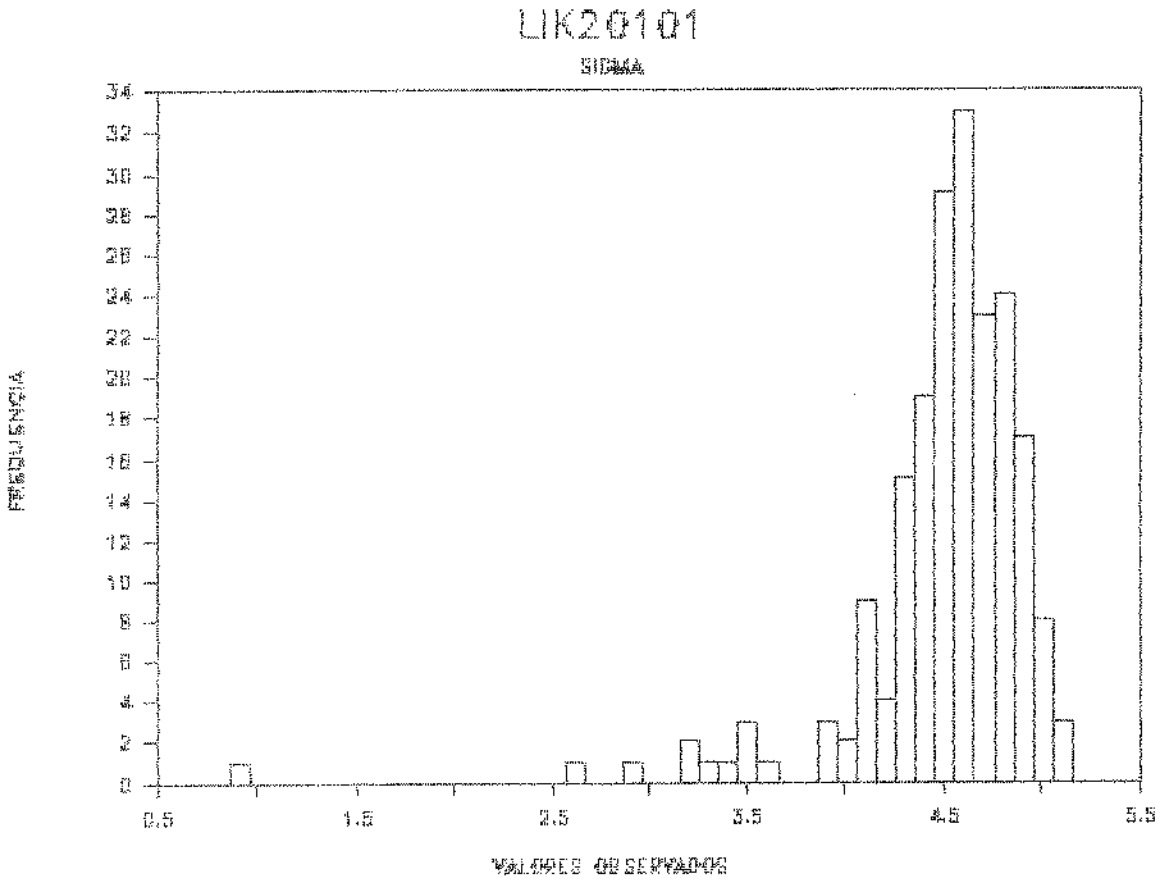


GRÁFICO 57

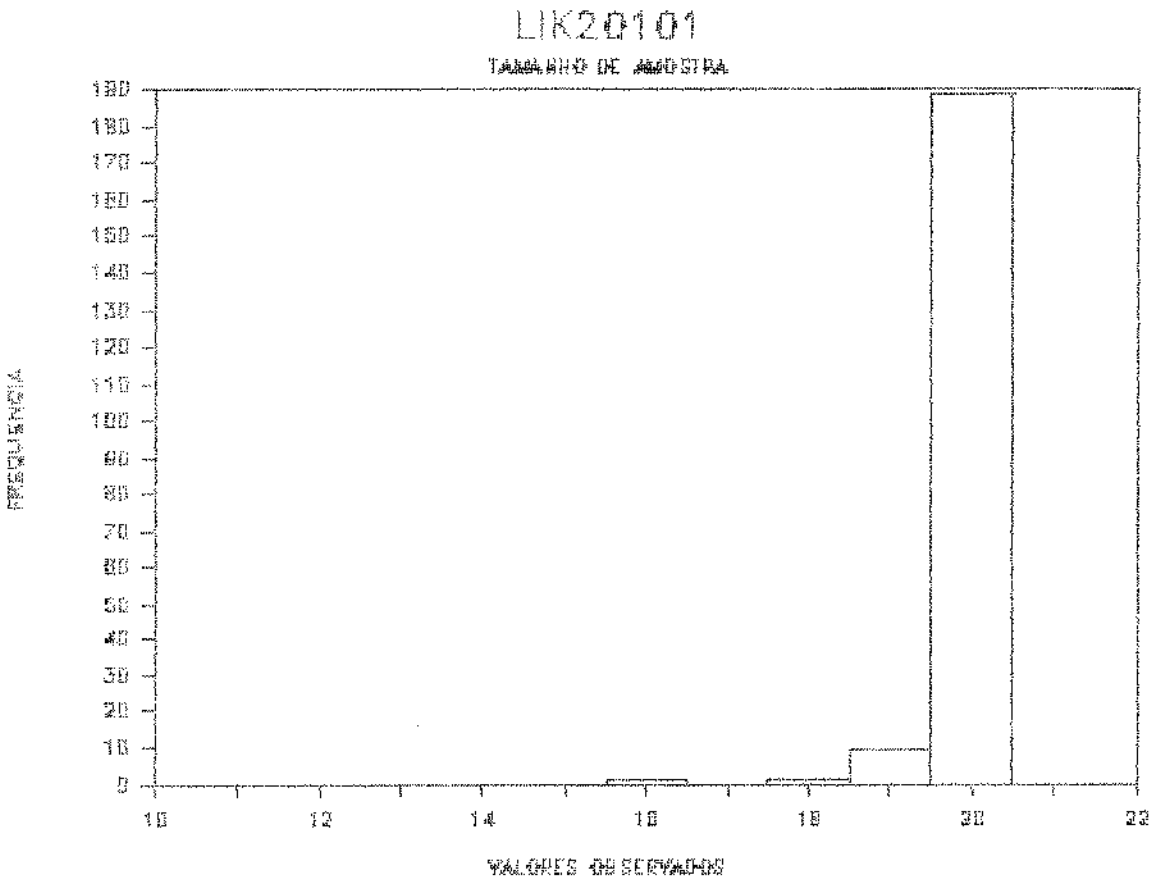

LIK20101
TAMAÑO DE MUESTRA
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 58

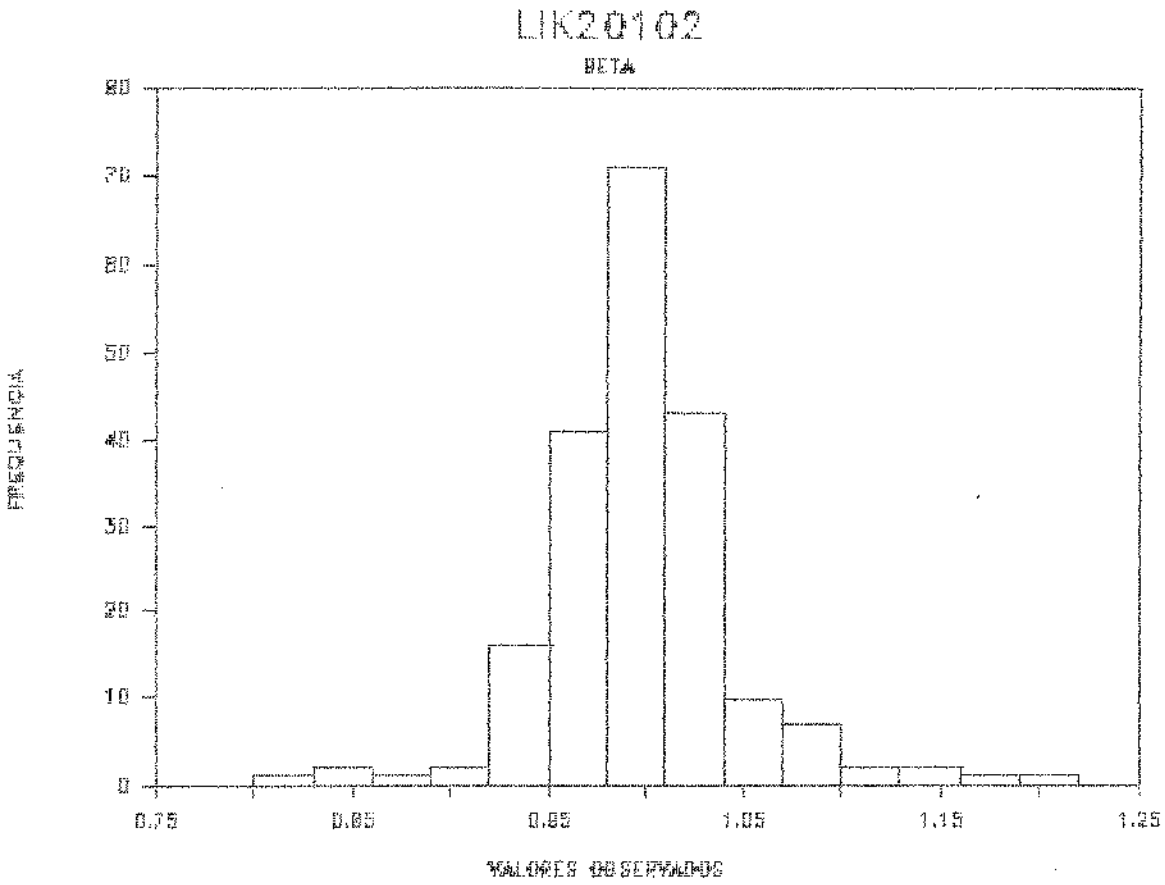


GRÁFICO 59

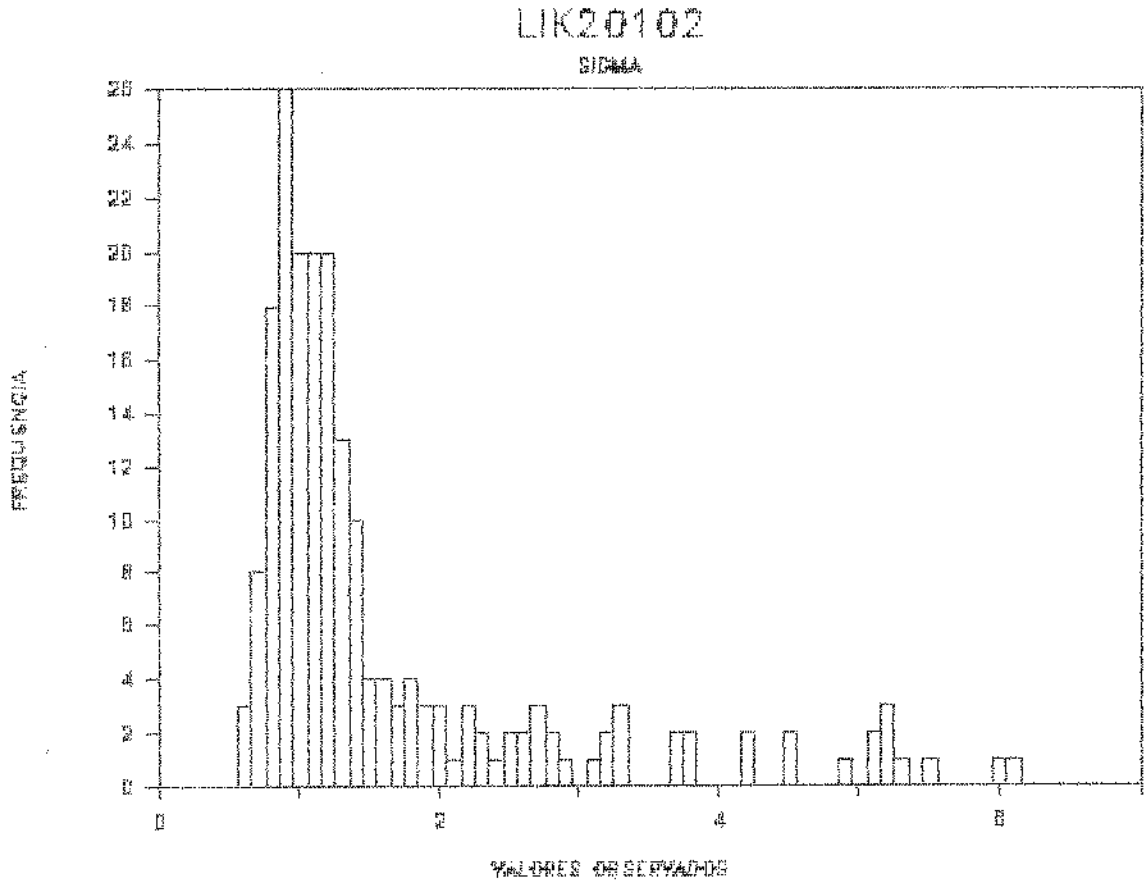


GRÁFICO 60

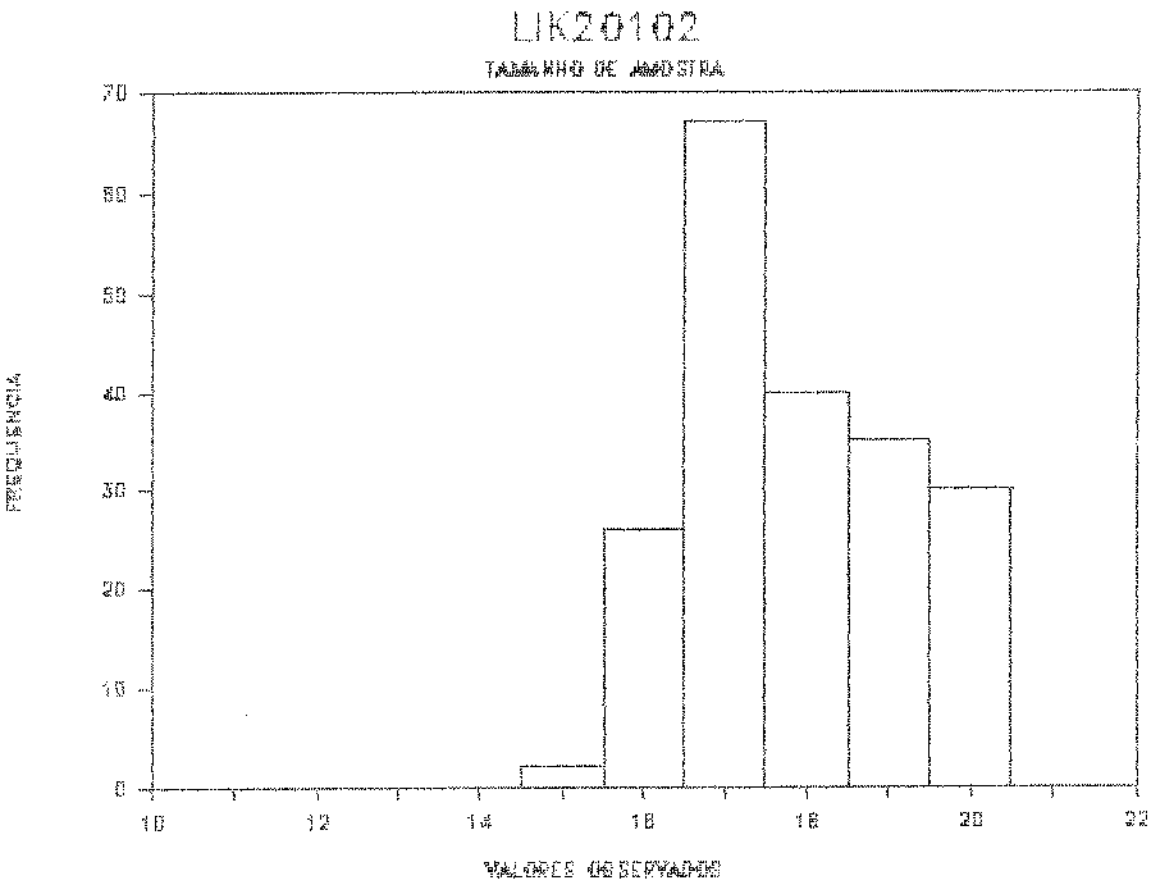

LIK20102
TAMANHO DE AMOSTRA
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 61

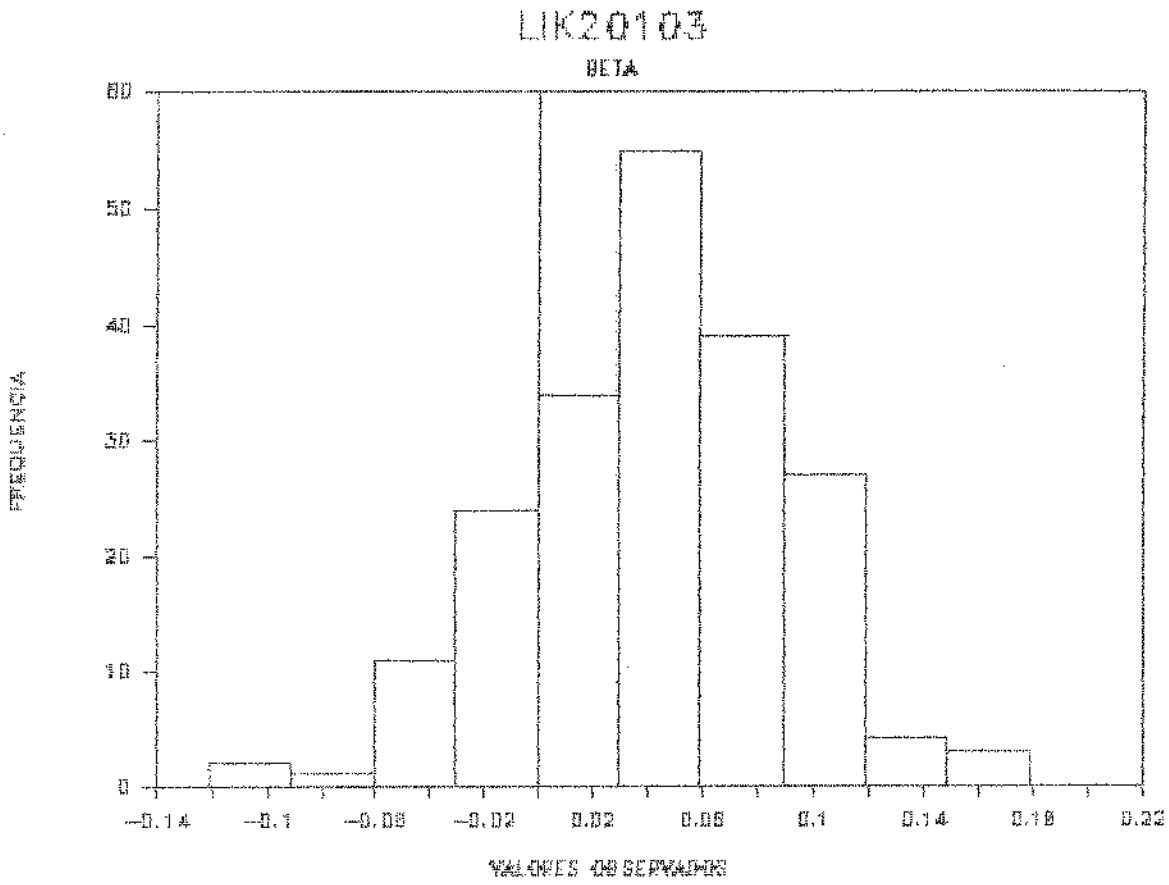

LIK20103
BETA
60
50
40
30
20
10
0
-0.14
-0.1
-0.06
-0.02
0.02
0.06
0.1
0.14
0.18
0.22

GRÁFICO 62

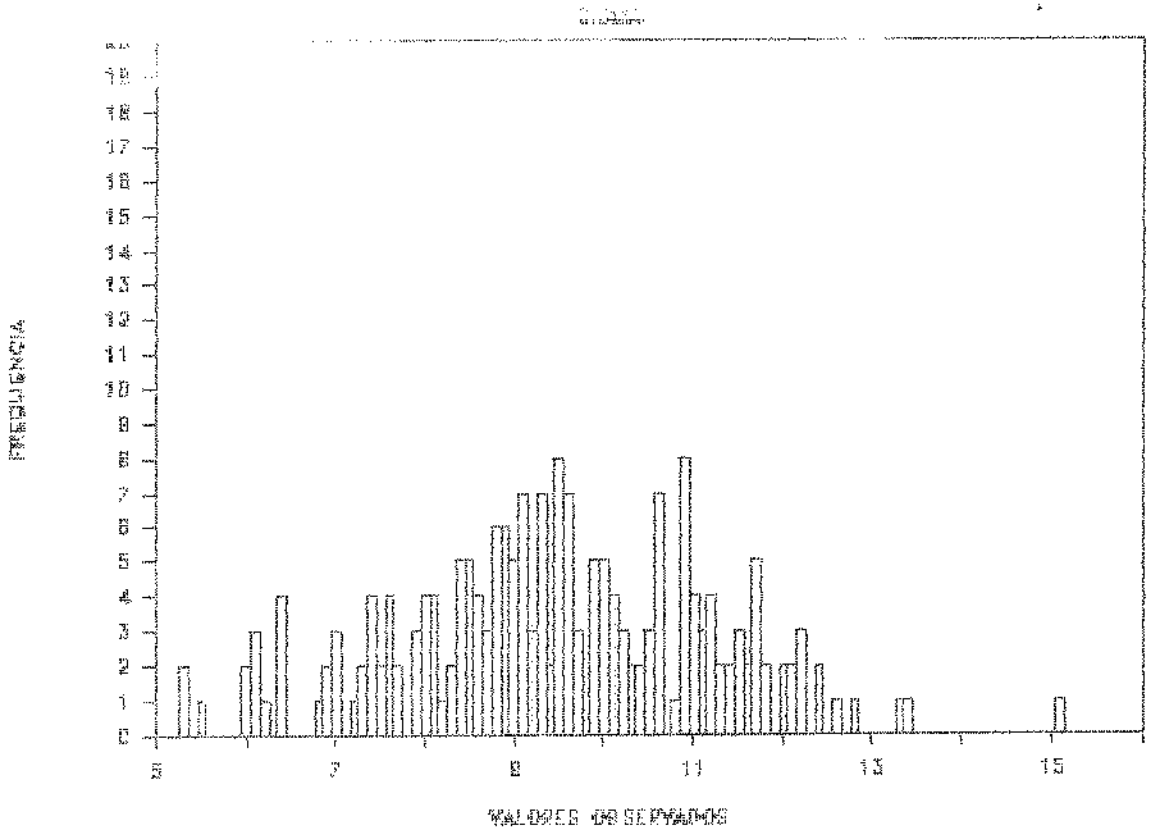


GRÁFICO 63

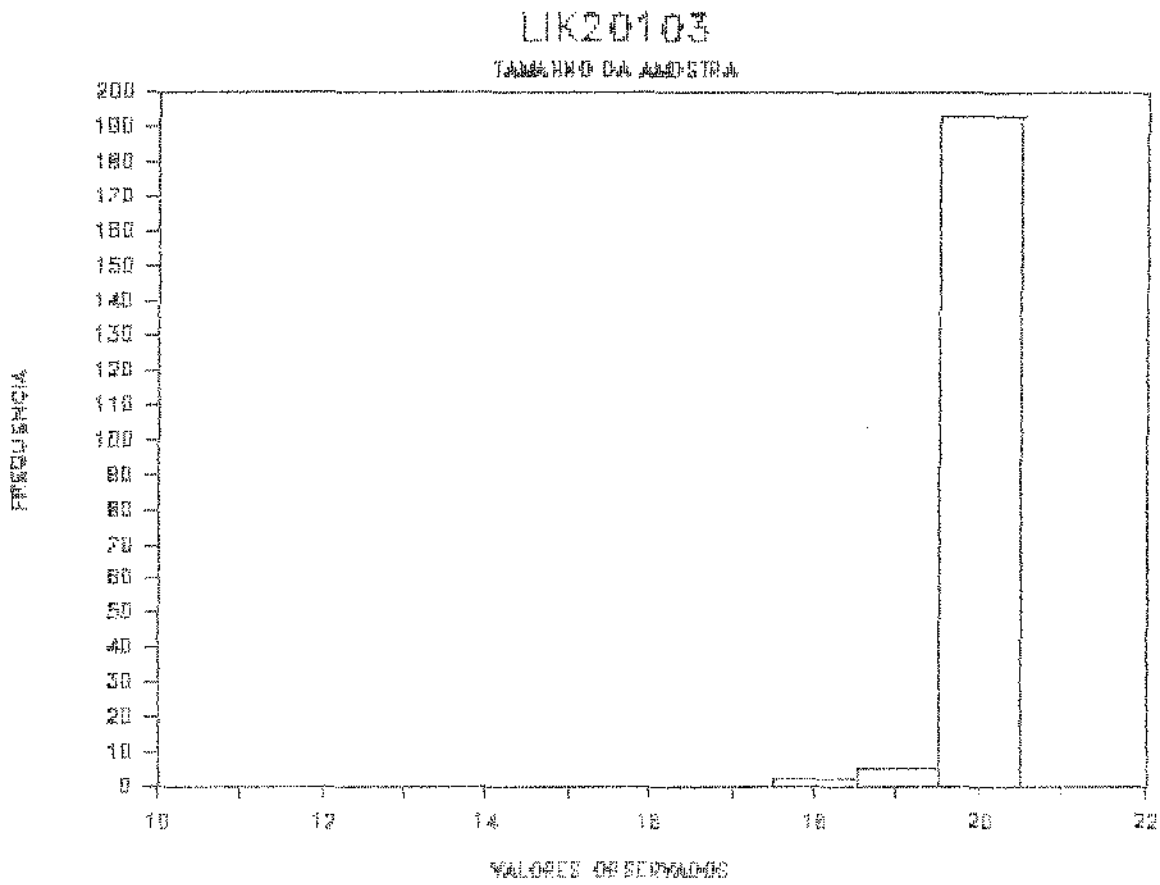

LIK20103
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22

GRÁFICO 64

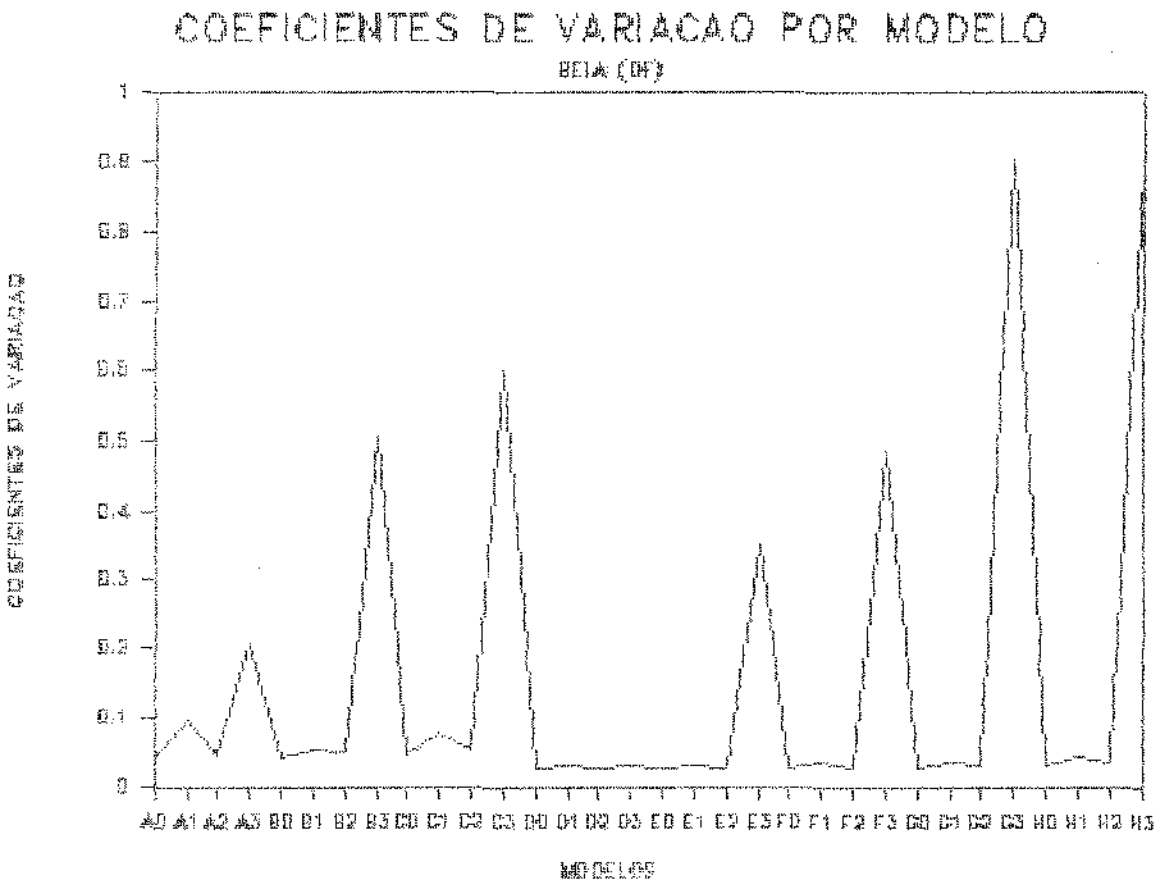

COEFICIENTES DE VARIACAO POR MODELO
BETA (DW)
COEFICIENTES DE VARIACAO
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 64A

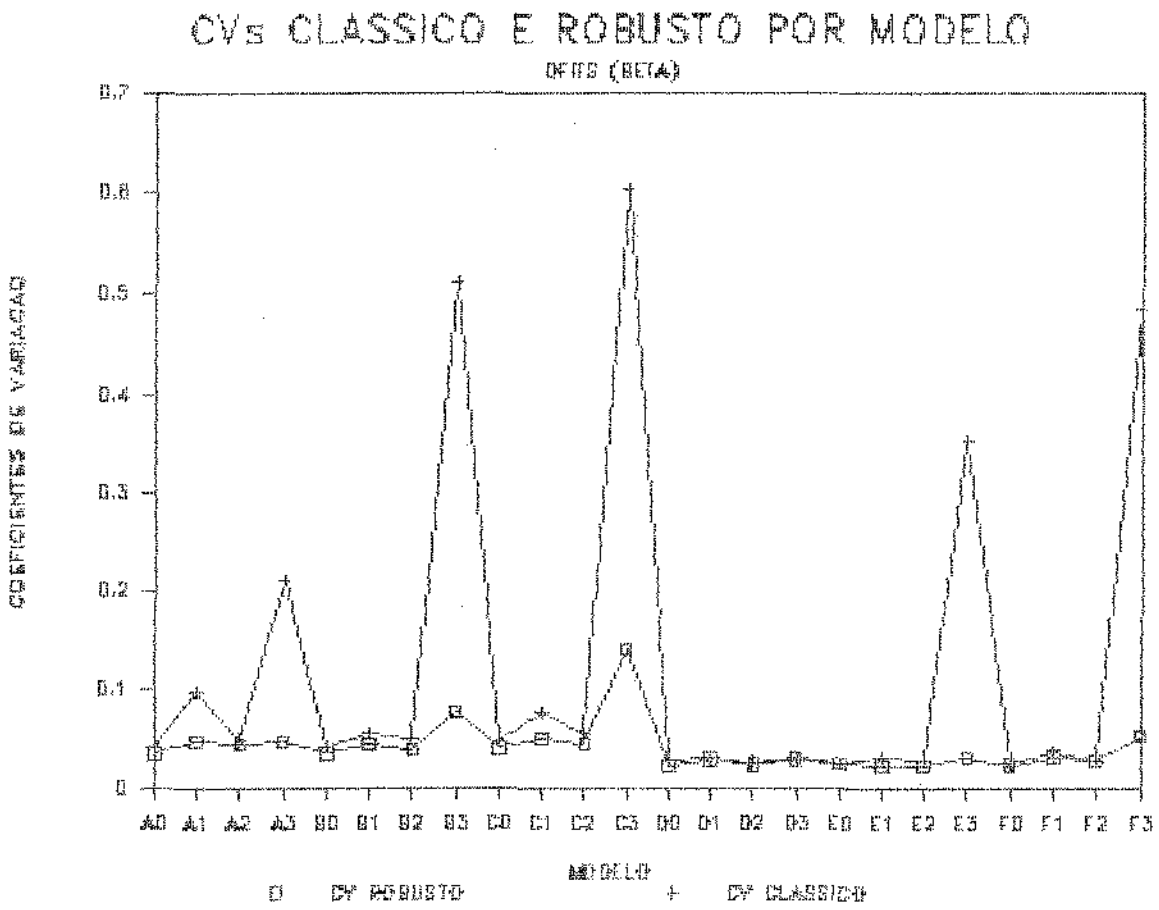

CVs CLASSICO E ROBUSTO POR MODELO
DADOS (BETA)
COEFICIENTES DE VARIACAO
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3
MODELO
□ CV ROBUSTO
+ CV CLASSICO

GRÁFICO 65

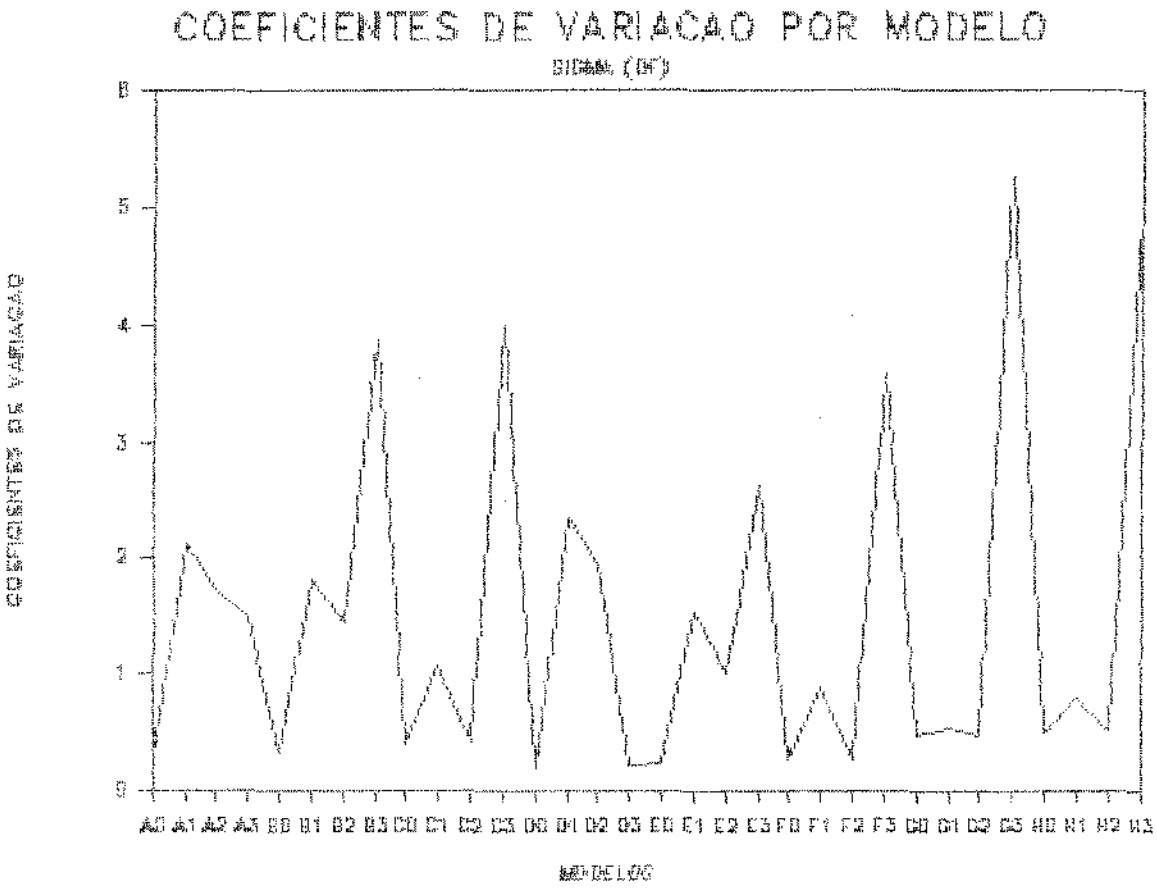

COEFICIENTES DE VARIACAO POR MODELO
SIGLA (DF)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 65A

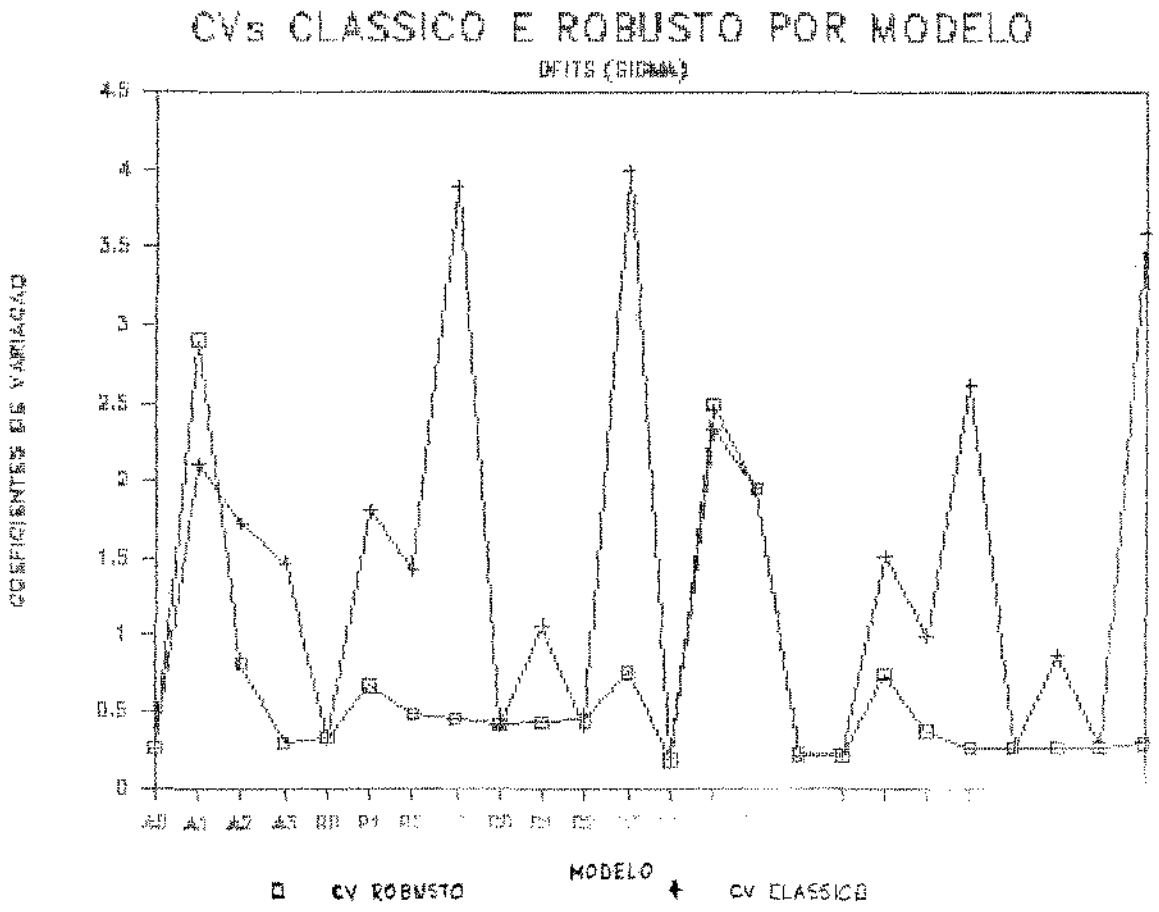

CVs CLASSICO E ROBUSTO POR MODELO
COEFICIENTES DE VARIAÇÃO
MODELO
CV ROBUSTO
CV CLASSICO

GRÁFICO 66

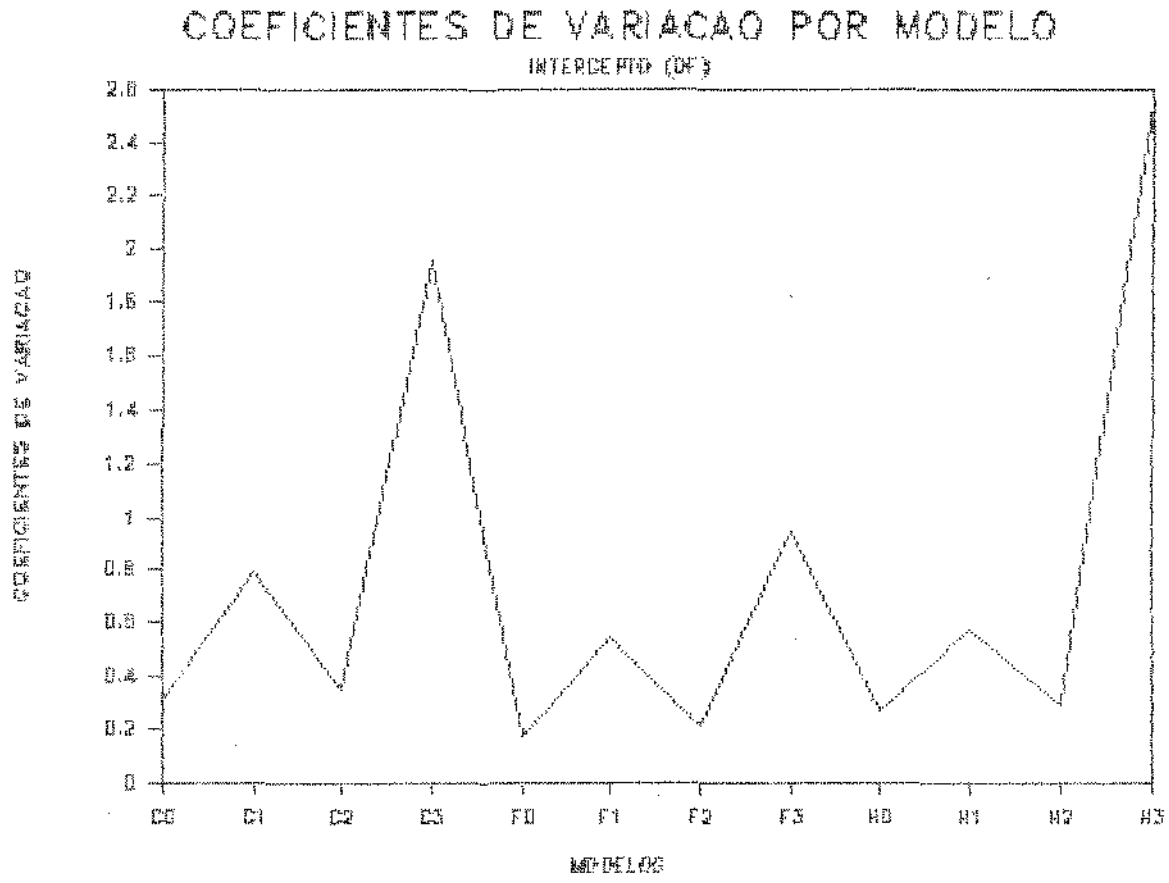

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (DF)
COEFICIENTE DE VARIACAO
2.8
2.4
2.2
2
1.6
1.4
1
0.8
0.6
0.4
0.2
0
C0
C1
C2
C3
F0
F1
F2
F3
M0
M1
M2
M3
MODELOS

GRÁFICO 67

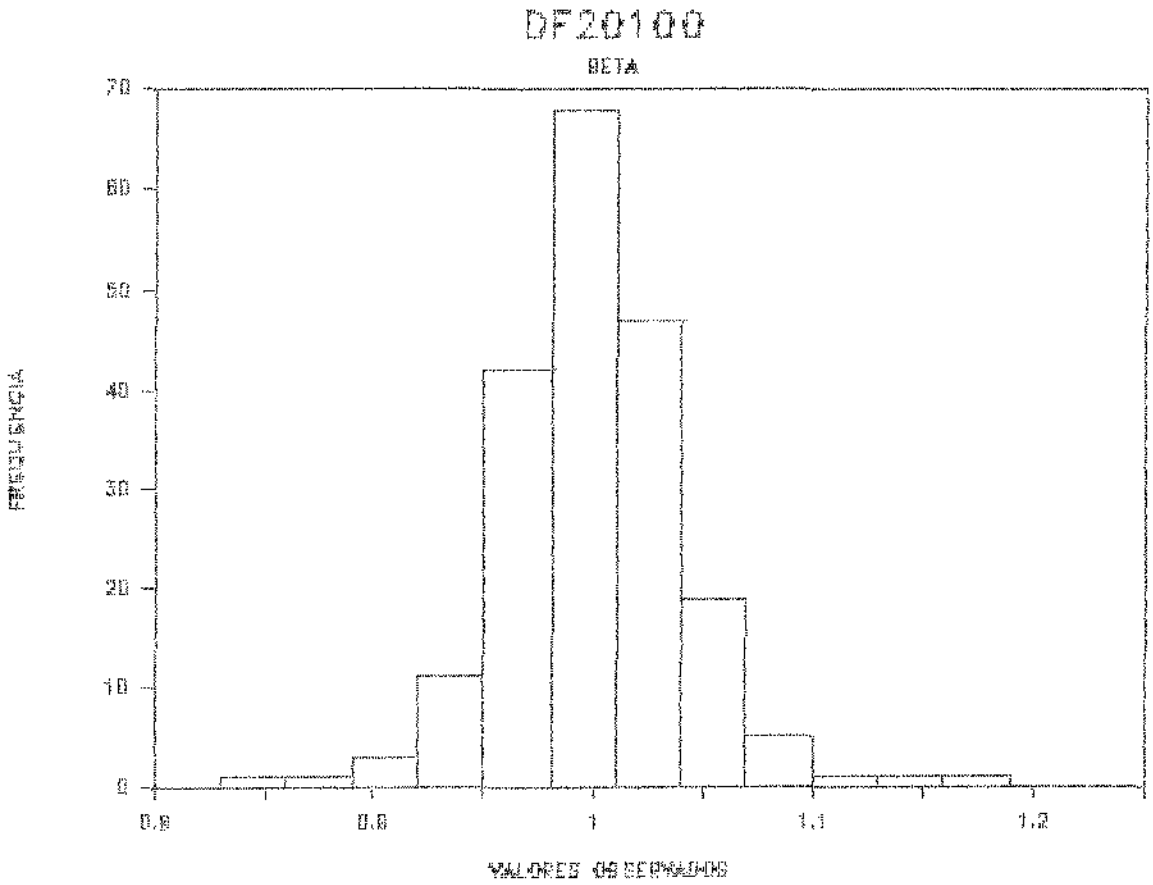


GRÁFICO 68

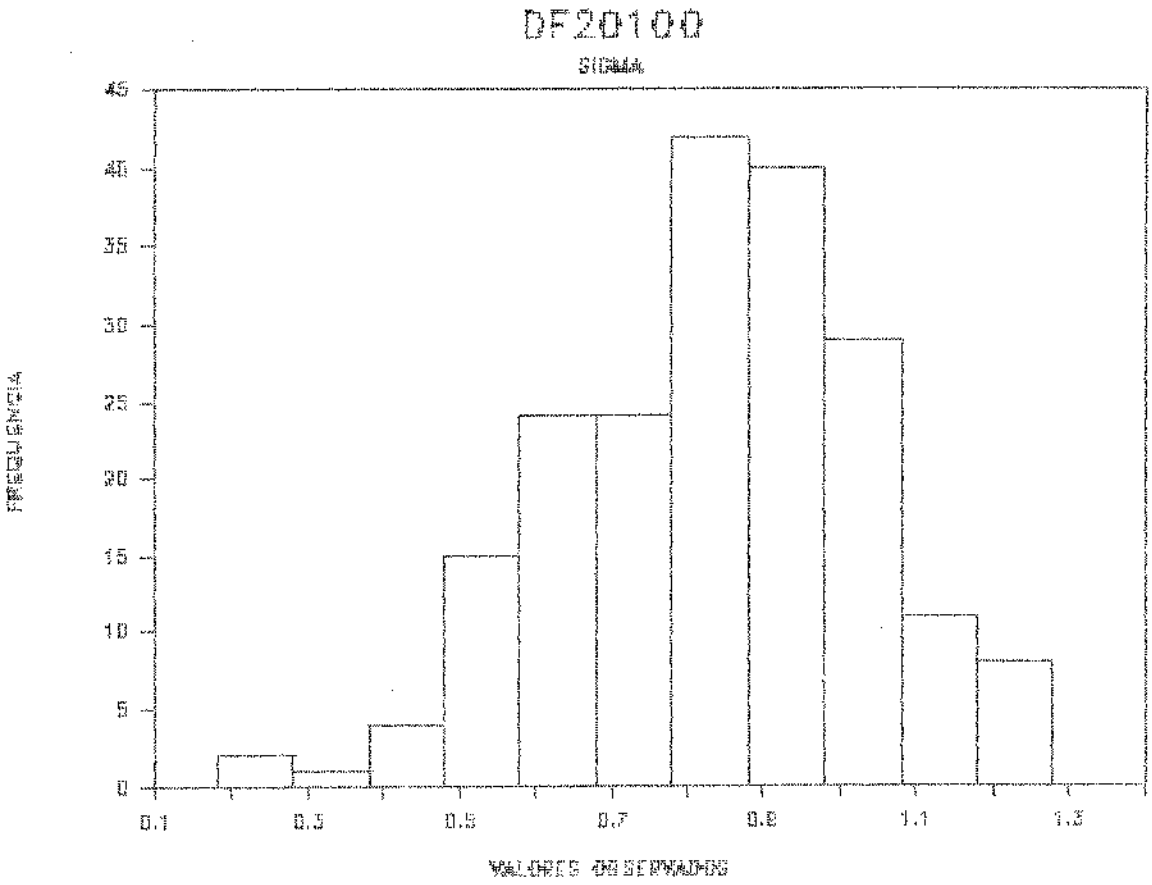

DF20100
SIGMA
FREQUÊNCIA
45
40
35
30
25
20
15
10
5
0
0.1
0.3
0.5
0.7
0.9
1.1
1.3
VALORES OBSERVADOS

GRÁFICO 69

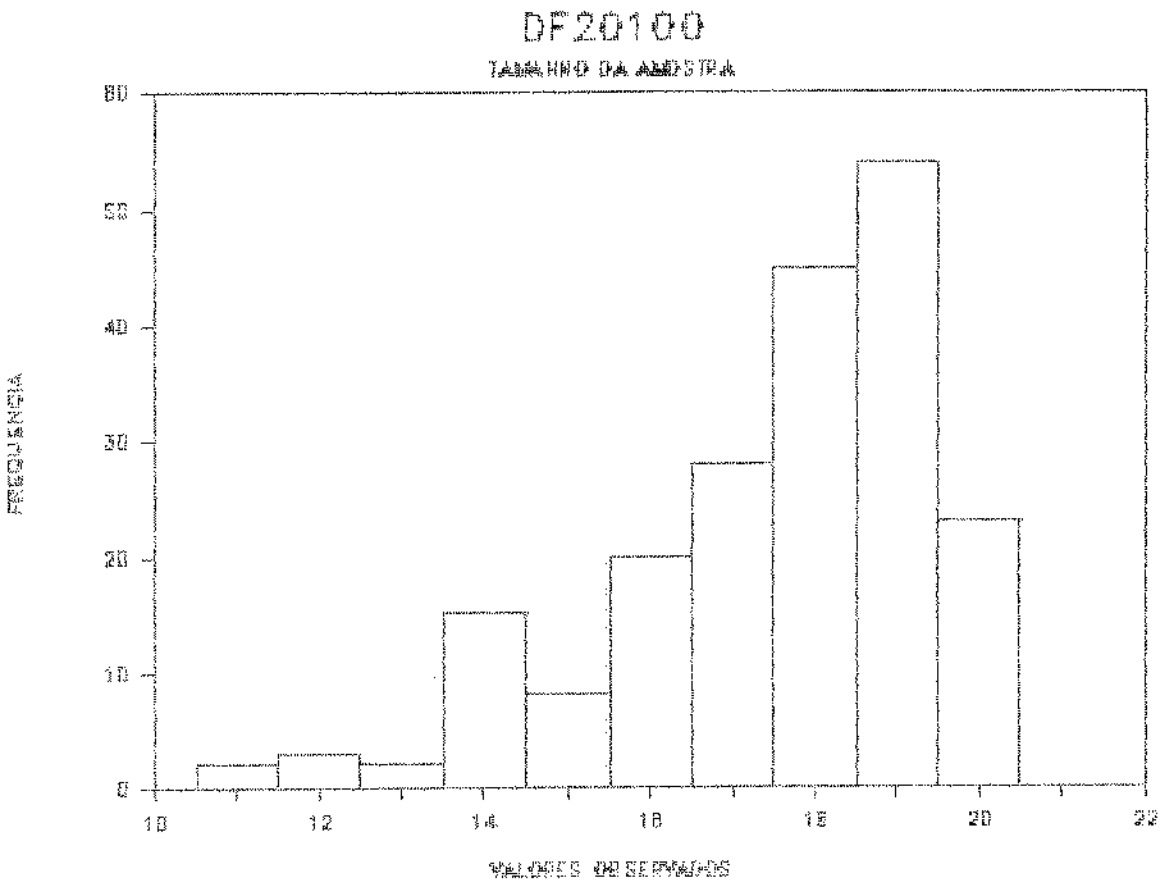

DF20100
TAMANHO DA AMOSTRA
FREQUÊNCIA
0
10
20
30
40
50
60
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 70

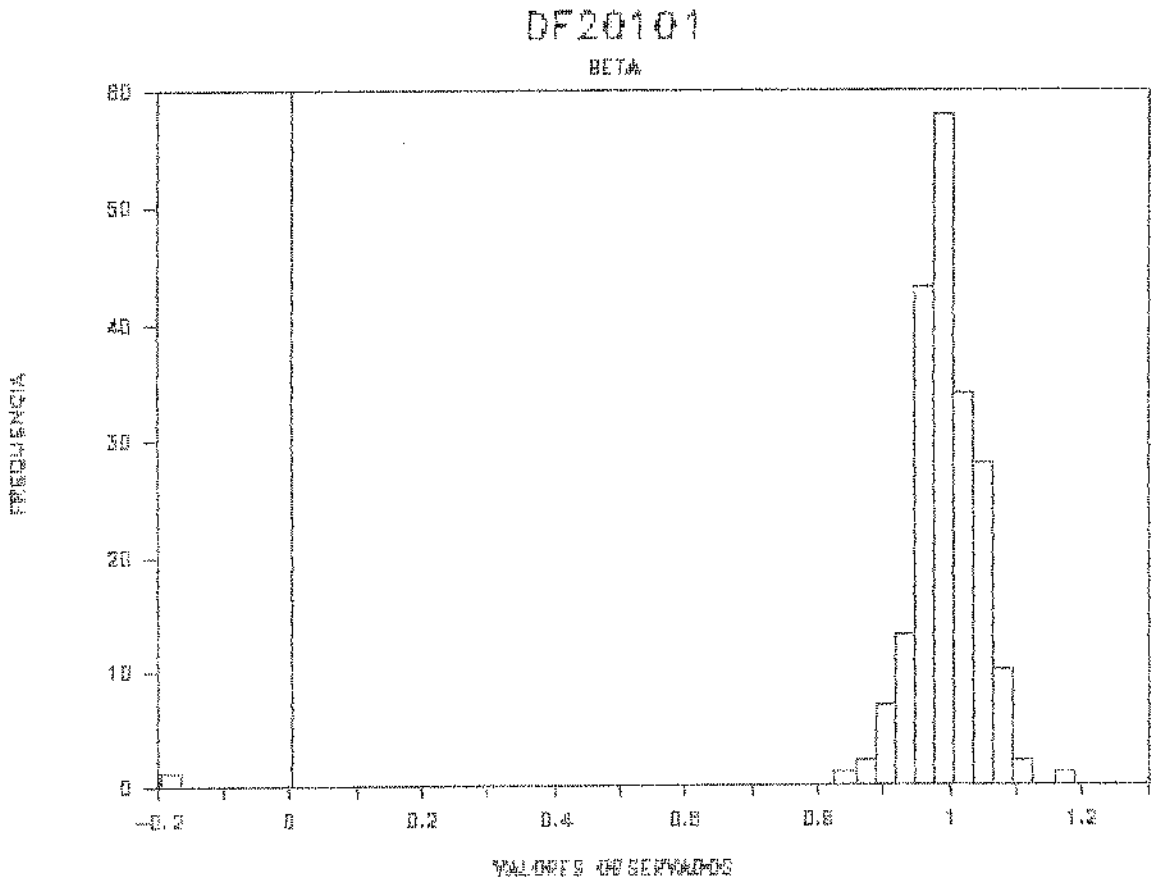

DF20101
BETA
FREQUENCIA
VALORES OBSERVADOS
0
10
20
30
40
50
60
-0.2
0
0.2
0.4
0.6
0.8
1
1.2

GRÁFICO 71

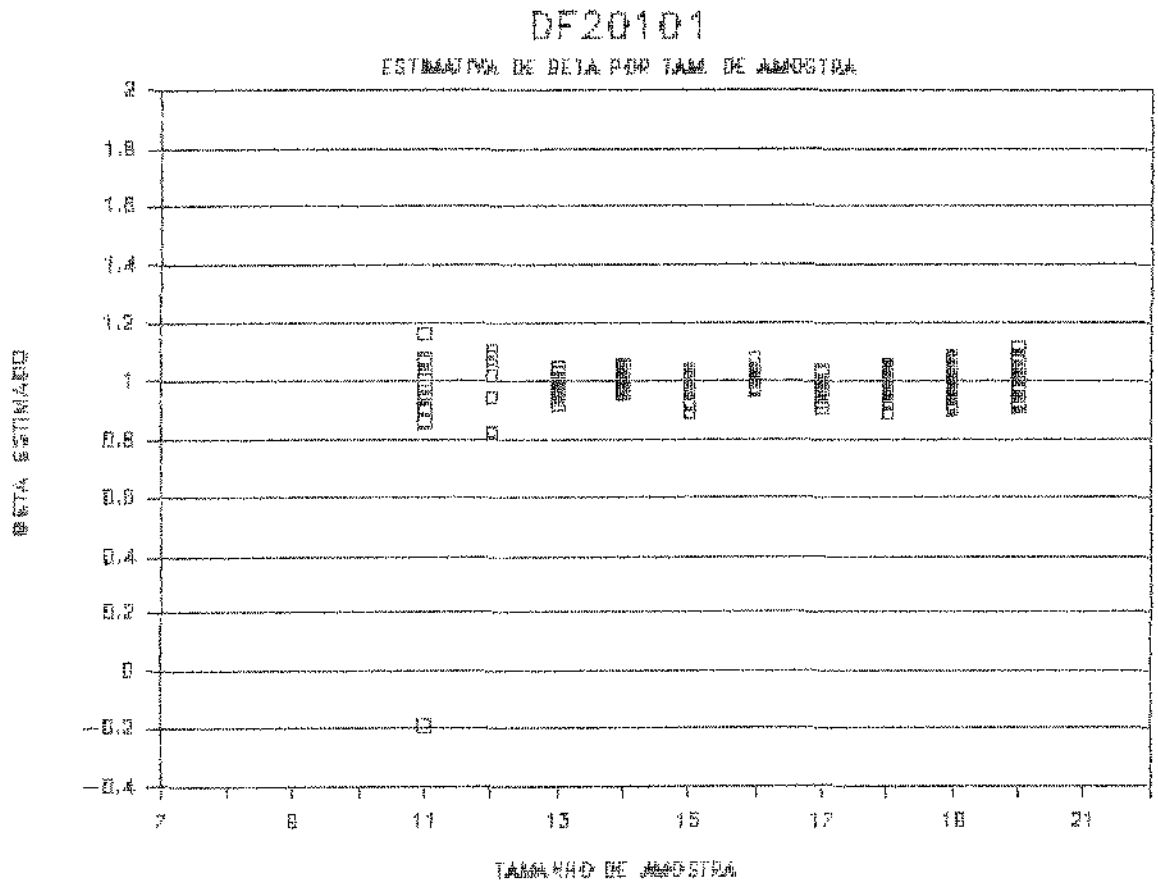

DF20101
ESTIMATIVA DE BETA POR TAM. DE AMOSTRA
BETA ESTIMADO
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 72

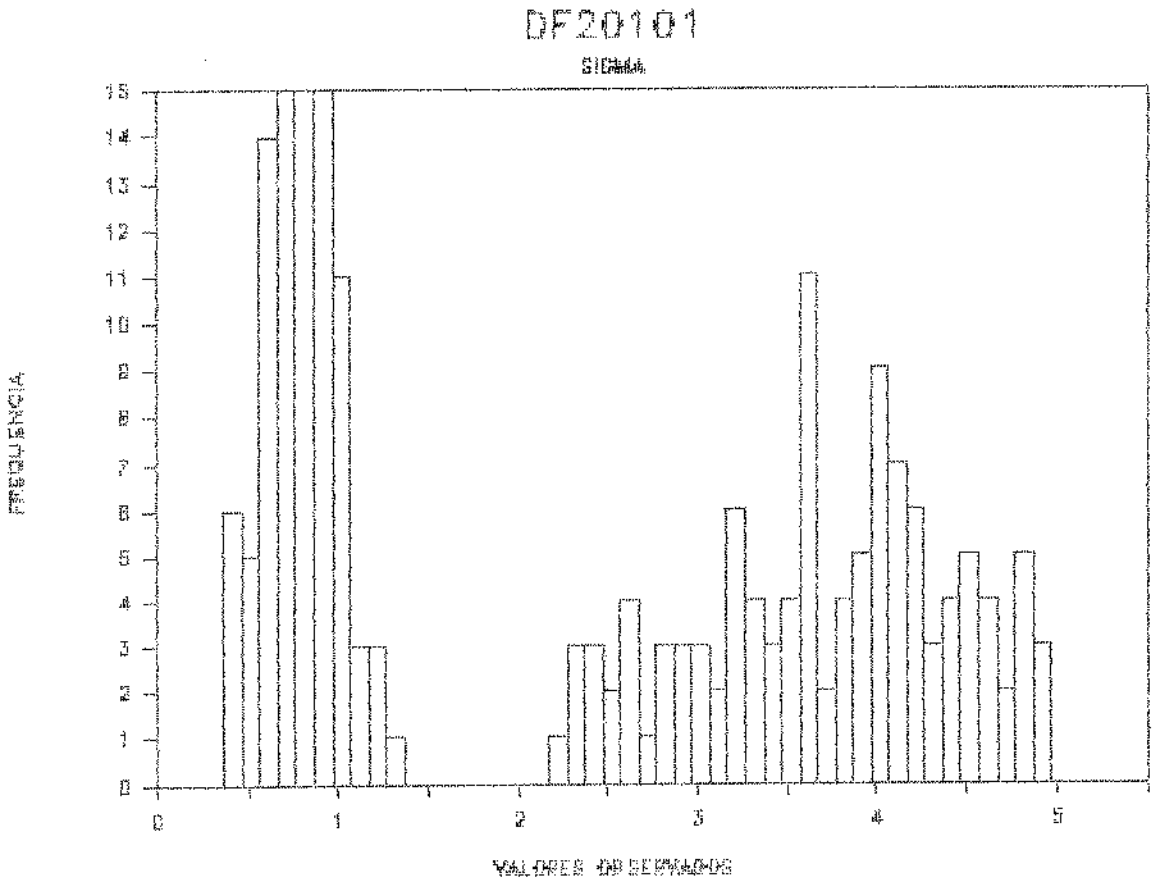


GRÁFICO 73

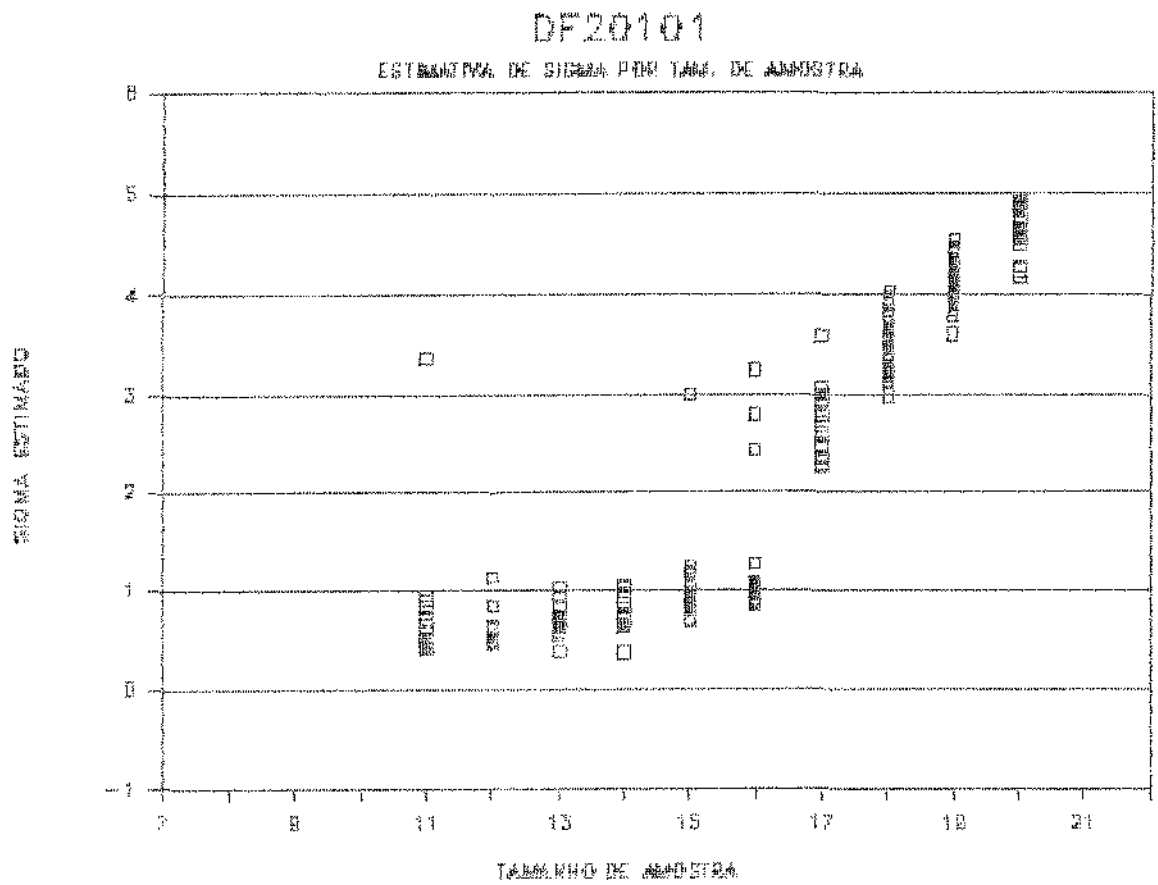

DF20101
ESTIMATIVA DE SIGMA PELO TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DE AMOSTRA
7
8
11
13
15
17
19
21

GRÁFICO 74

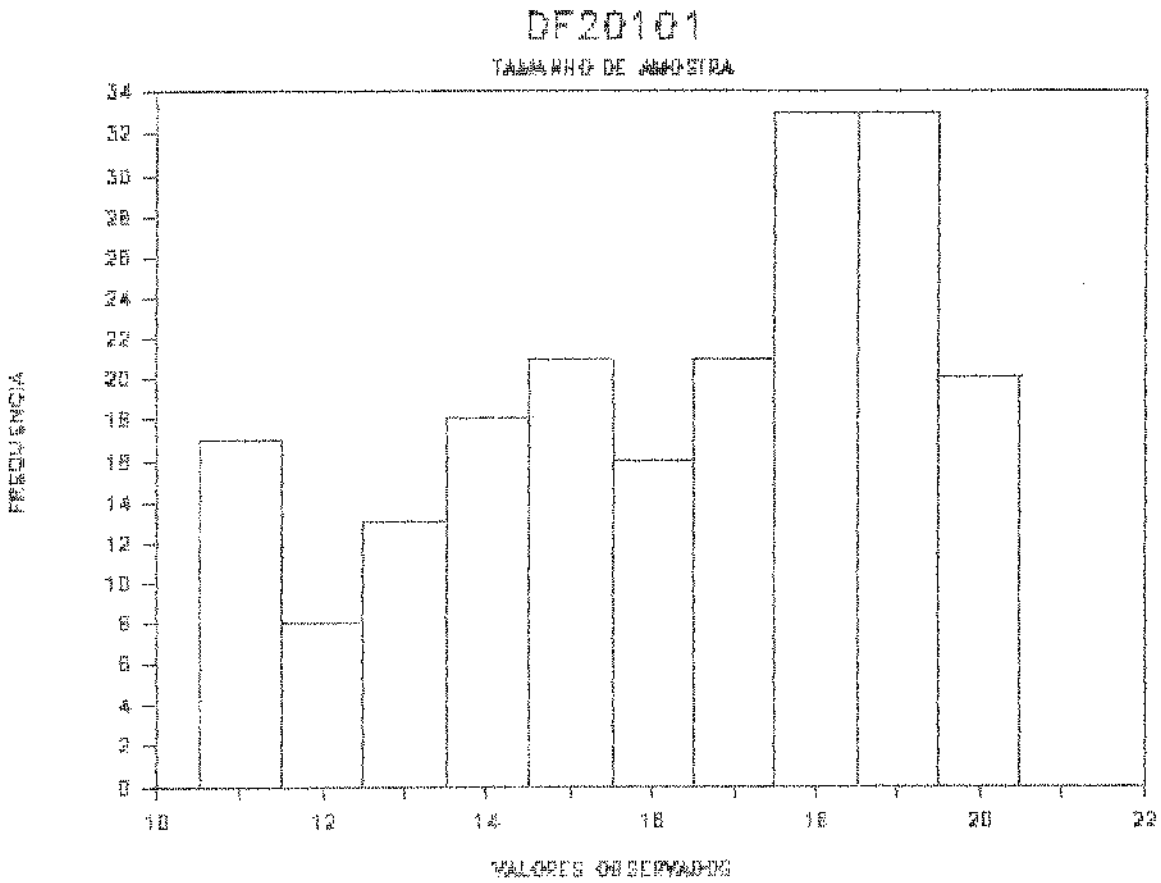

DF20101
TAMANHO DE AMOSTRA
FREQUÊNCIA
0
4
8
12
16
20
24
28
32
36
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 75

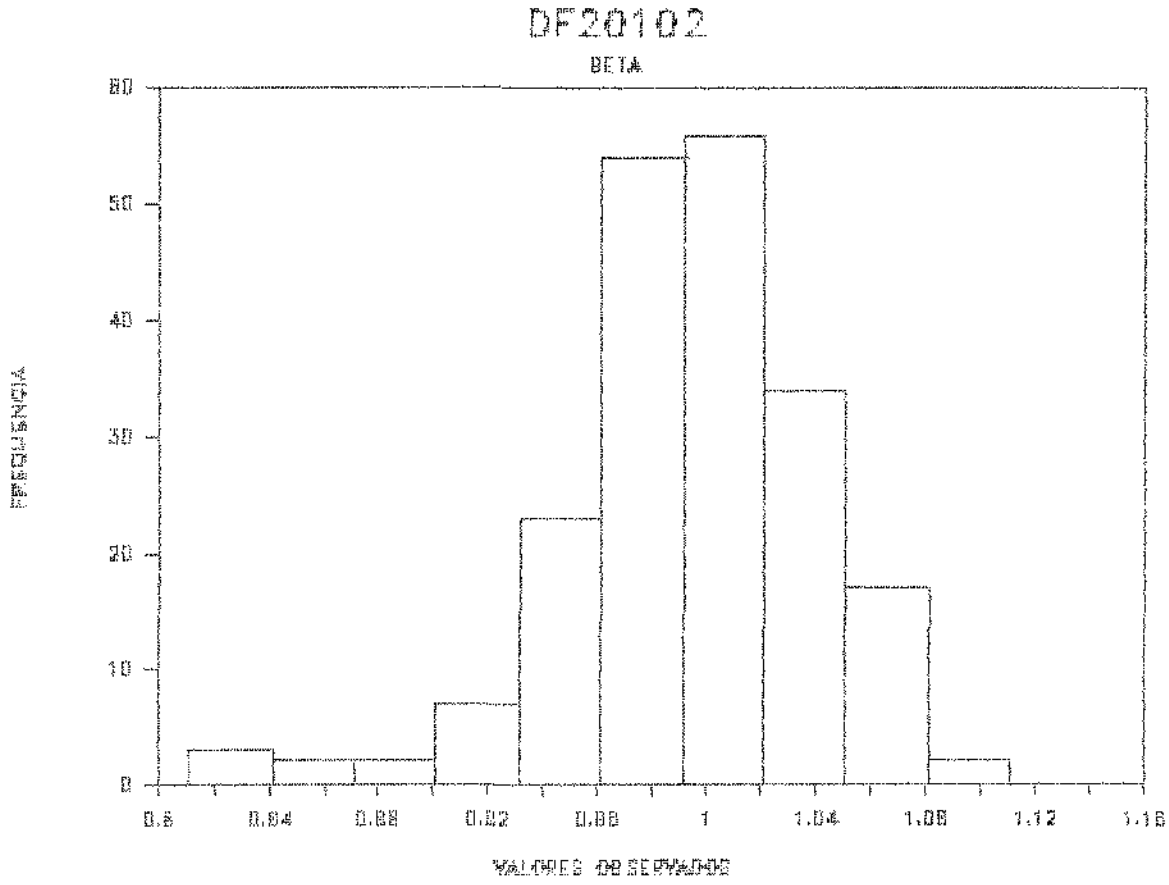

DF20102
BETA
80
60
40
30
20
10
0
FREQUENCIA
0.8
0.84
0.88
0.92
0.96
1
1.04
1.08
1.12
1.16
VALORES OBSERVADOS

GRÁFICO 76

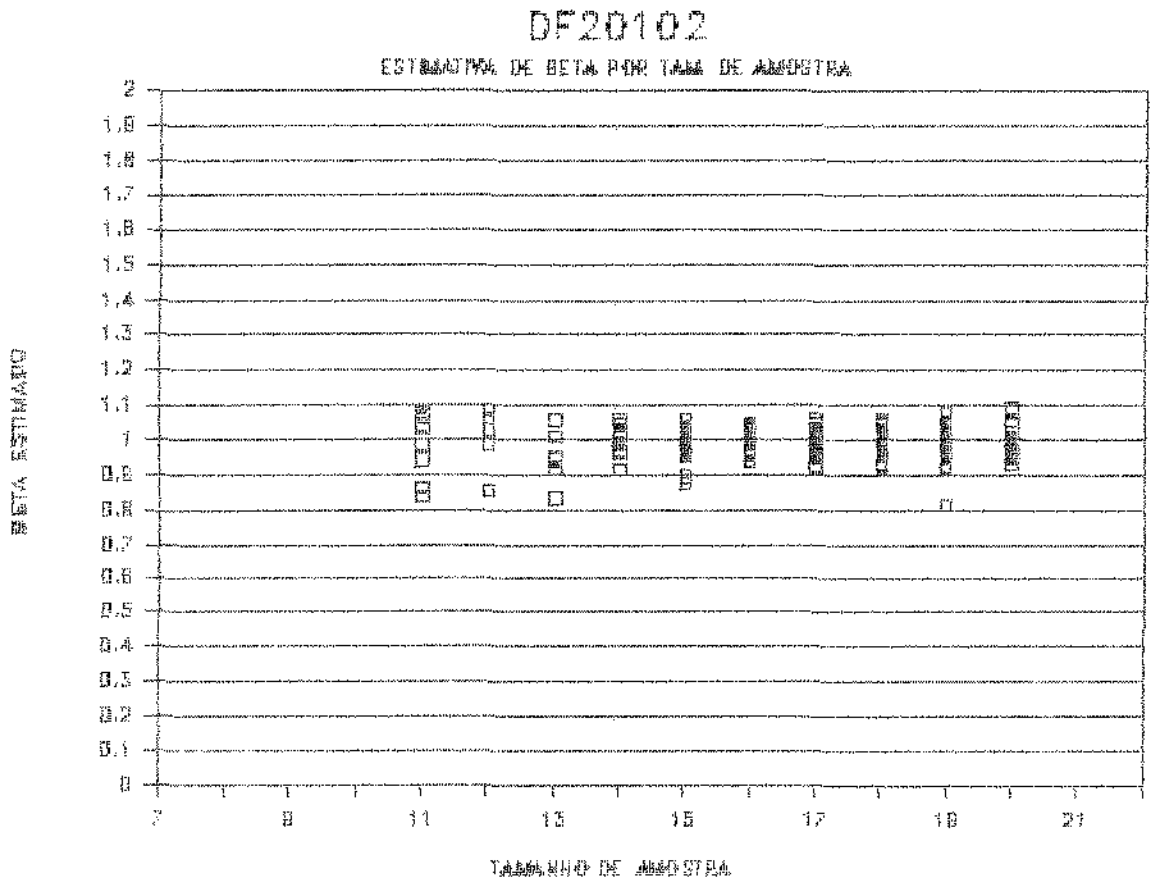

DF20102
ESTIMATIVA DE BETA POR TAM DE AMOSTRA
BETA ESTIMADO
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 77

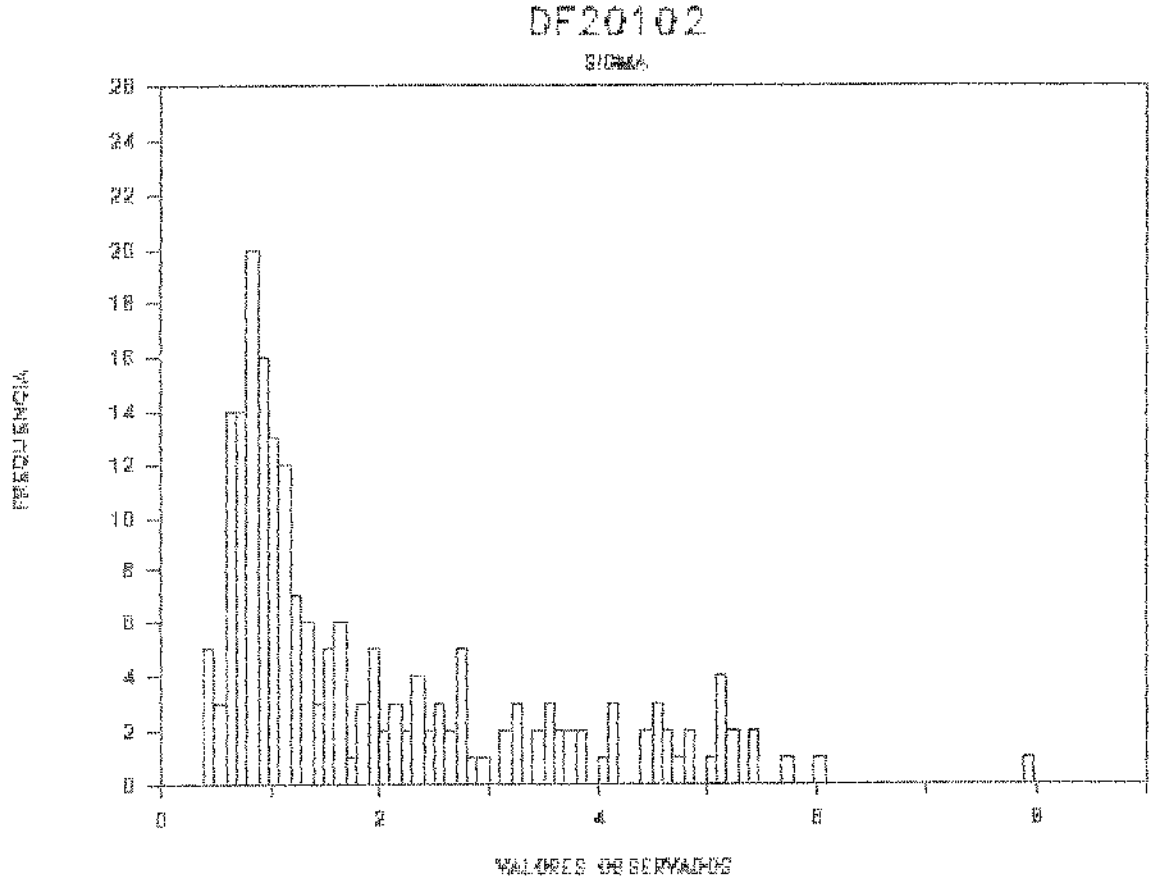


GRÁFICO 78

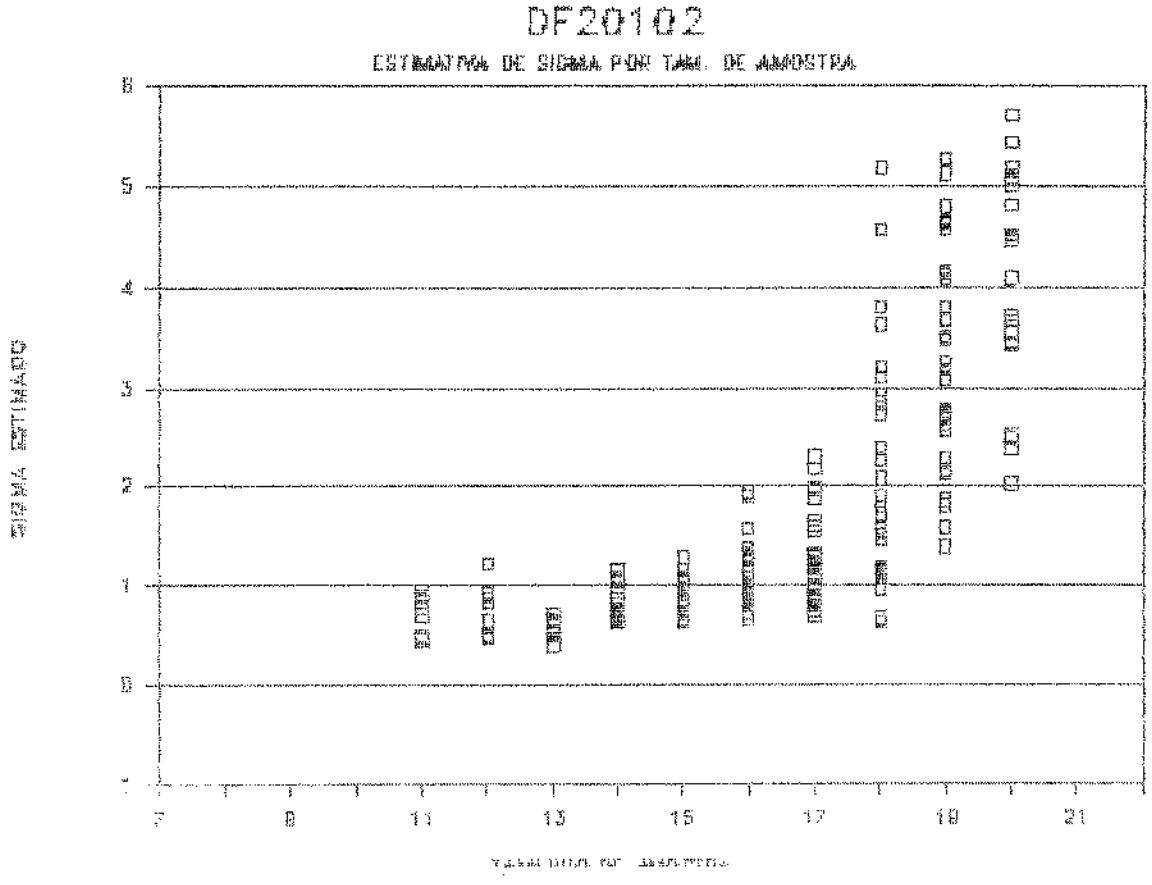


GRÁFICO 79

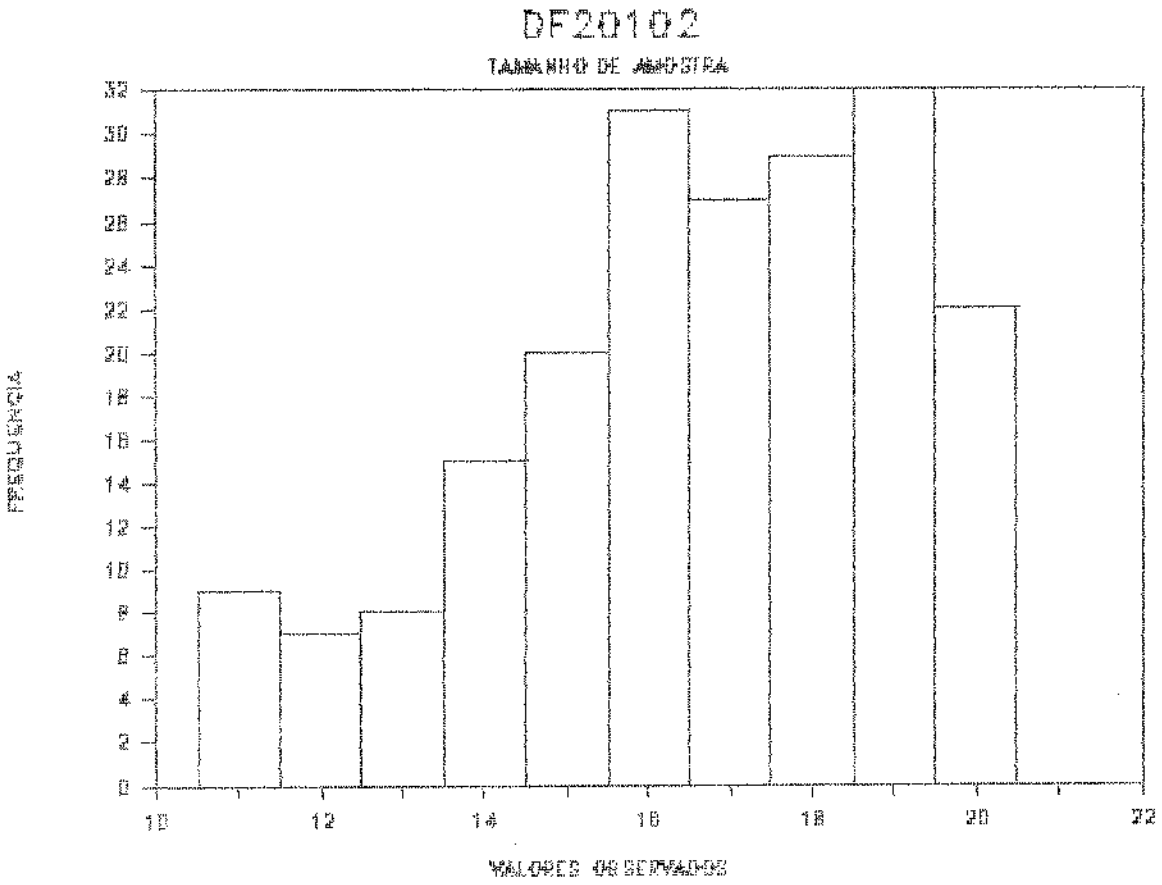

DF20102
TAMANHO DE AMOSTRA
FREQUÊNCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 80

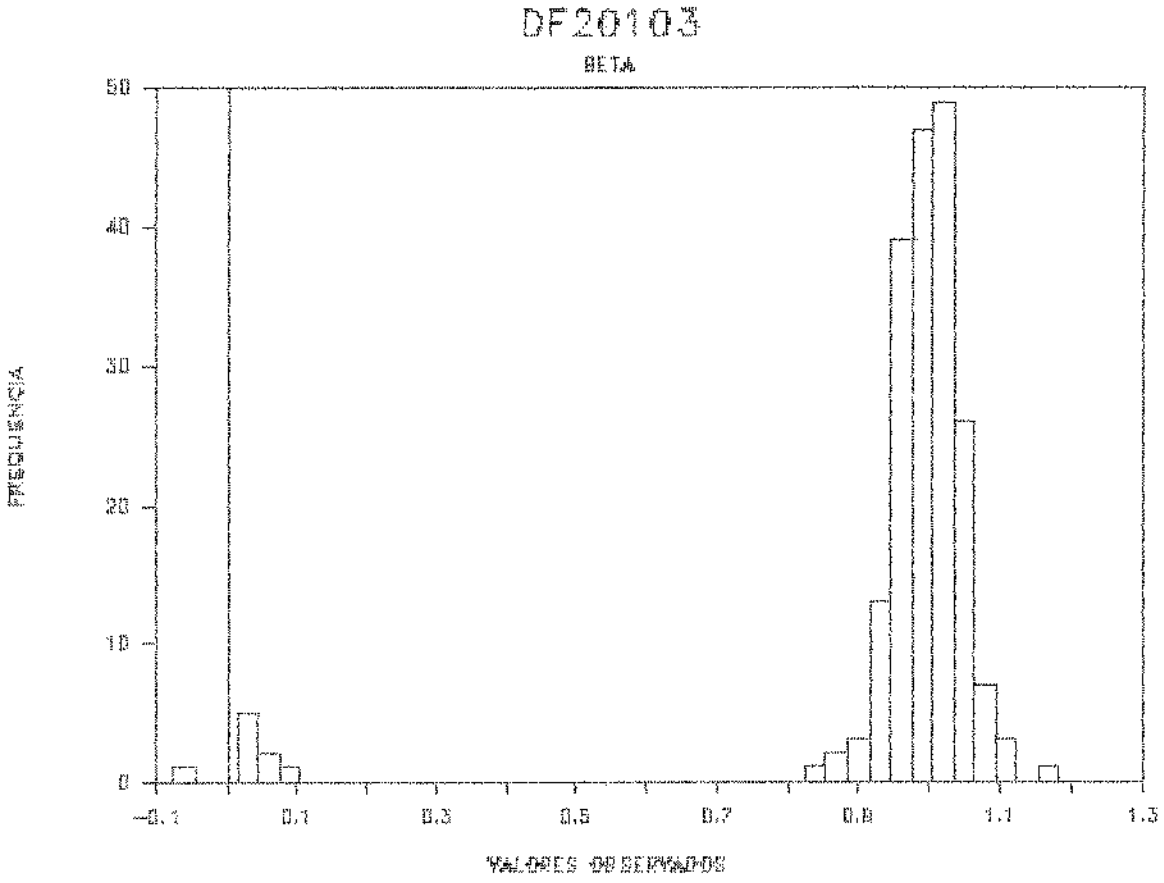

DF20103
BETA
FREQUENCIA
50
40
30
20
10
0
-0.1
0.1
0.3
0.5
0.7
0.9
1.1
1.3
VALORES OBSERVADOS

GRÁFICO 81

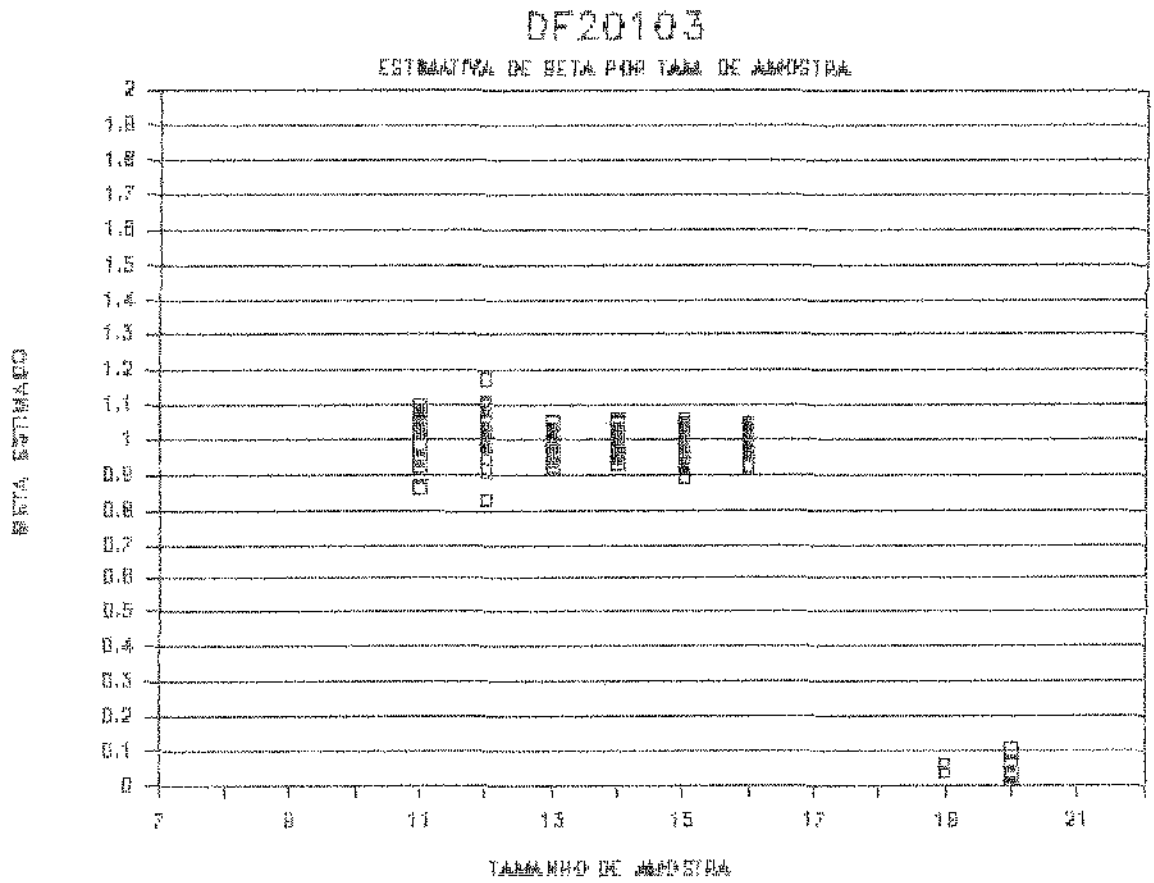

DF20103
ESTIMATIVA DE BETA POR TAMANHO DE AMOSTRA
BETA ESTIMADO
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 82

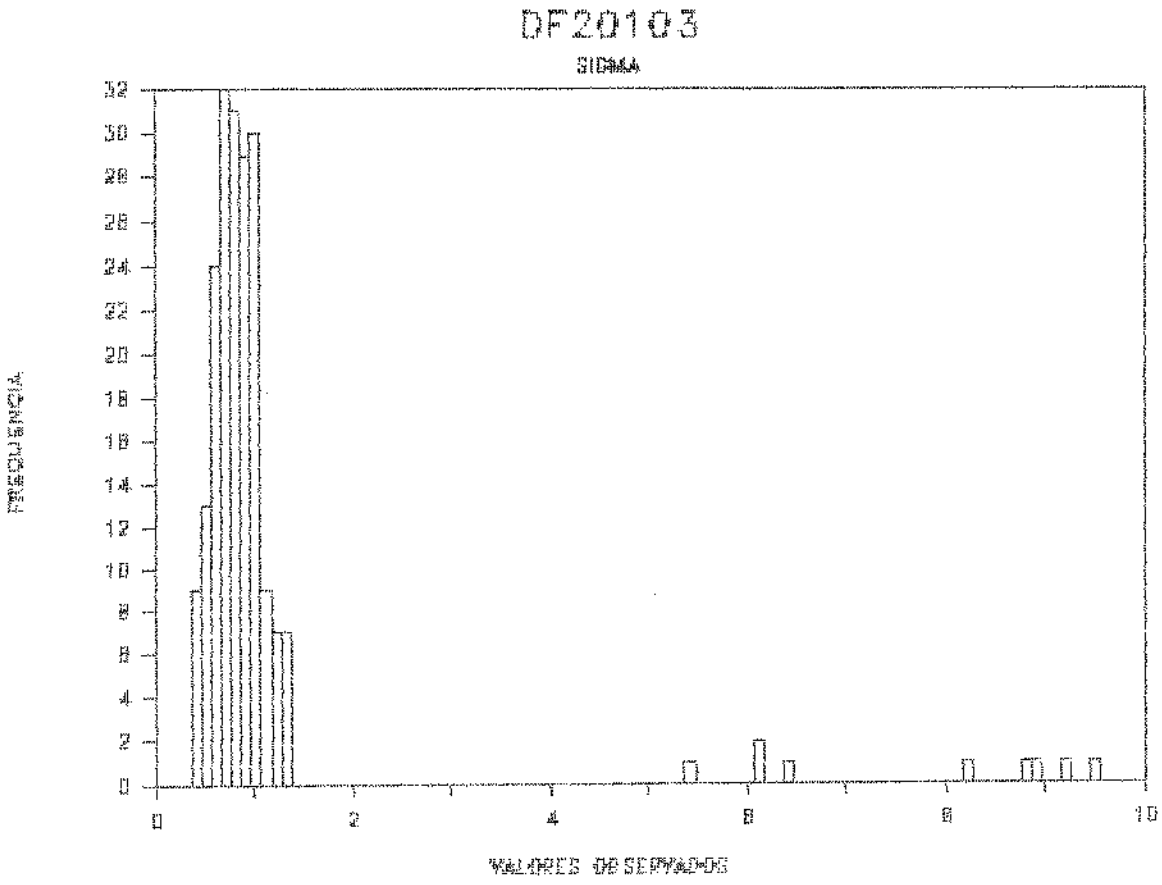


GRÁFICO 83

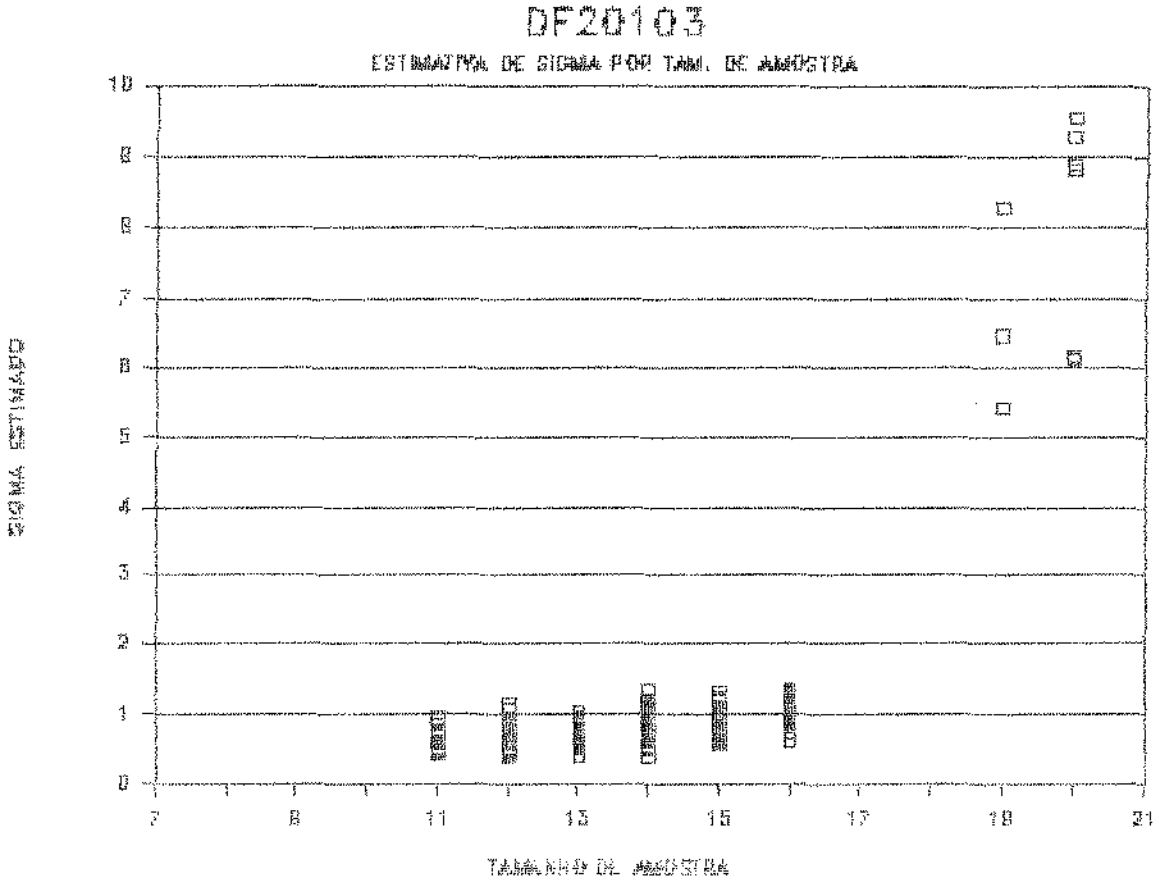

DF2103
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DA AMOSTRA
0
1
2
3
4
5
6
7
8
9
10
7
8
11
13
15
17
19
21

GRÁFICO 84

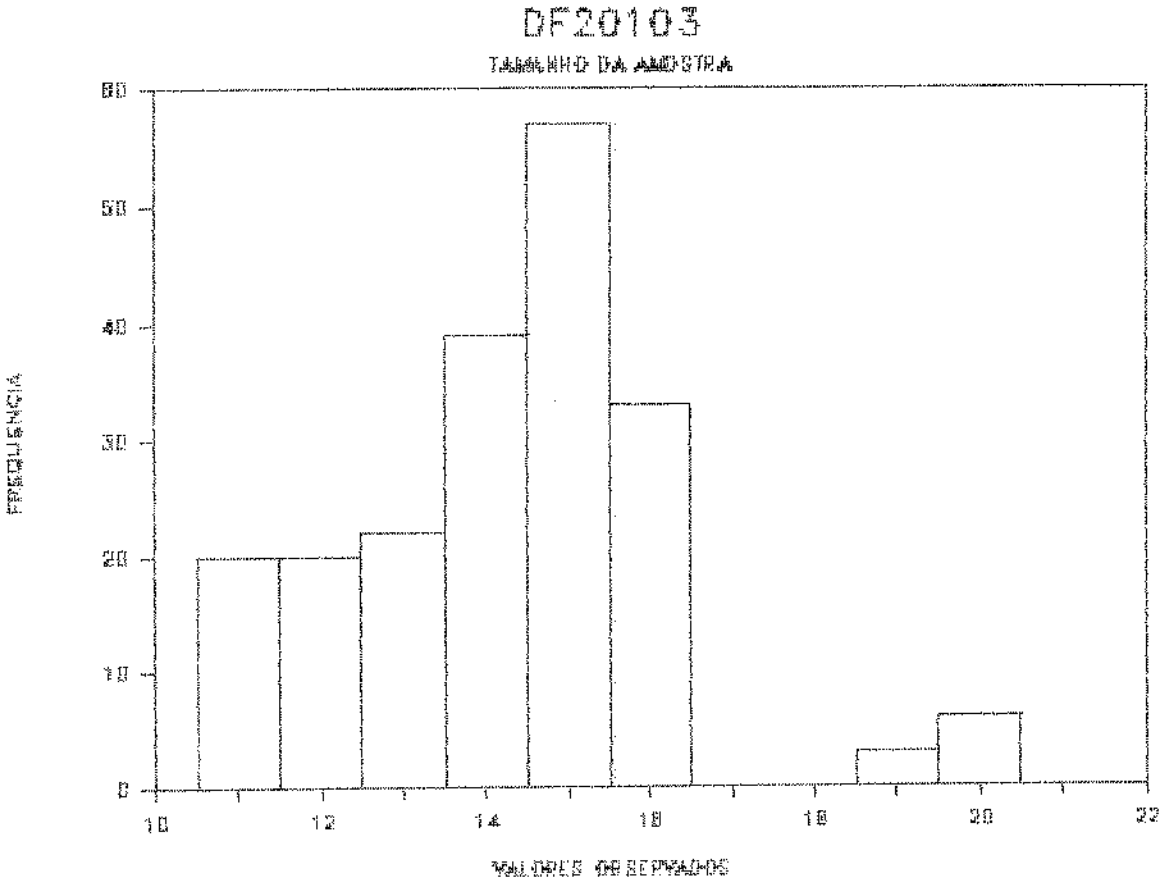

DF20103
TAMANHO DA AMOSTRA
60
50
40
30
20
10
0
10
12
14
16
18
20
22

GRÁFICO 85

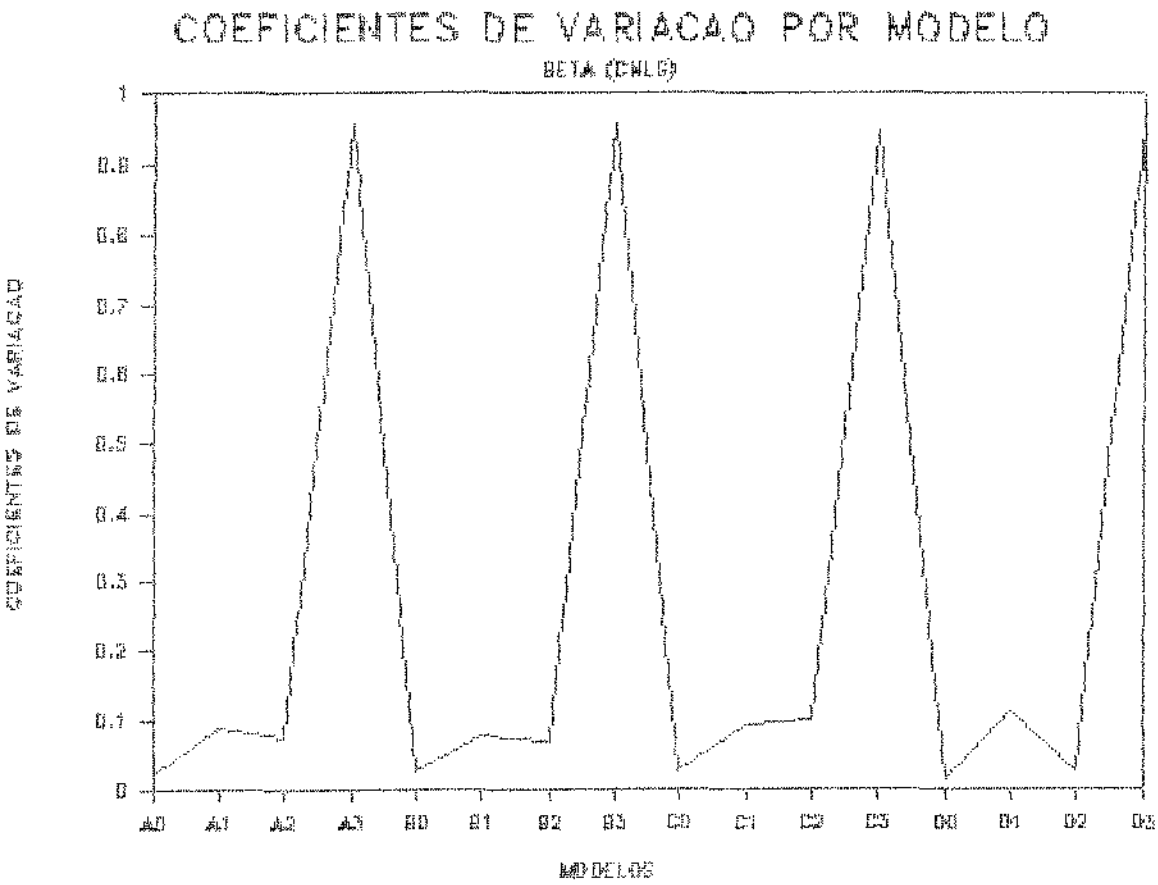

COEFICIENTES DE VARIACAO POR MODELO
BETA (CMLG)
COEFICIENTE DE VARIACAO
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
A0
A1
A2
A3
B0
B1
B2
B3
C0
C1
C2
C3
D0
D1
D2
D3
MODELOS

GRÁFICO 85A

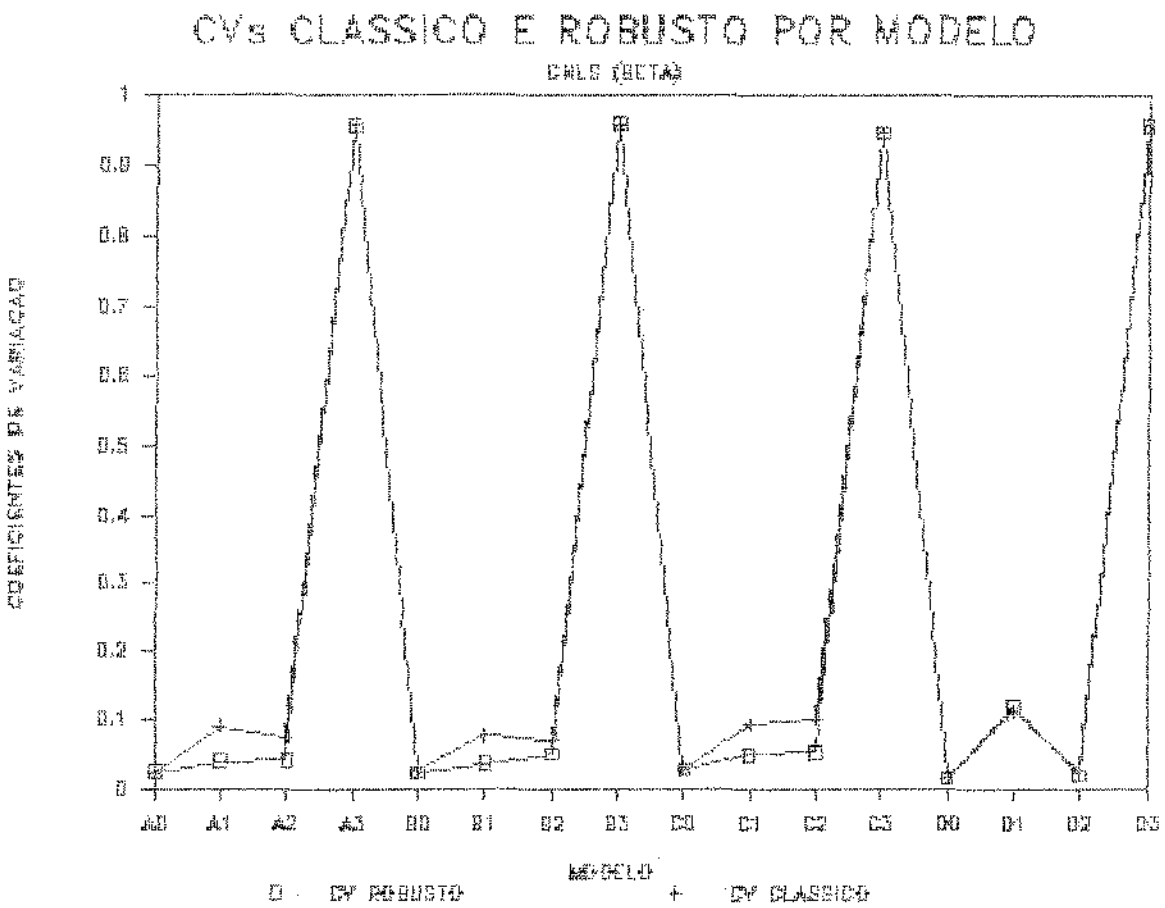

CVs CLASSICO E ROBUSTO POR MODELO
CVLS (BETA)
COEFICIENTES DE VARIACAO
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3
MODELO
CV ROBUSTO
CV CLASSICO

GRÁFICO 86

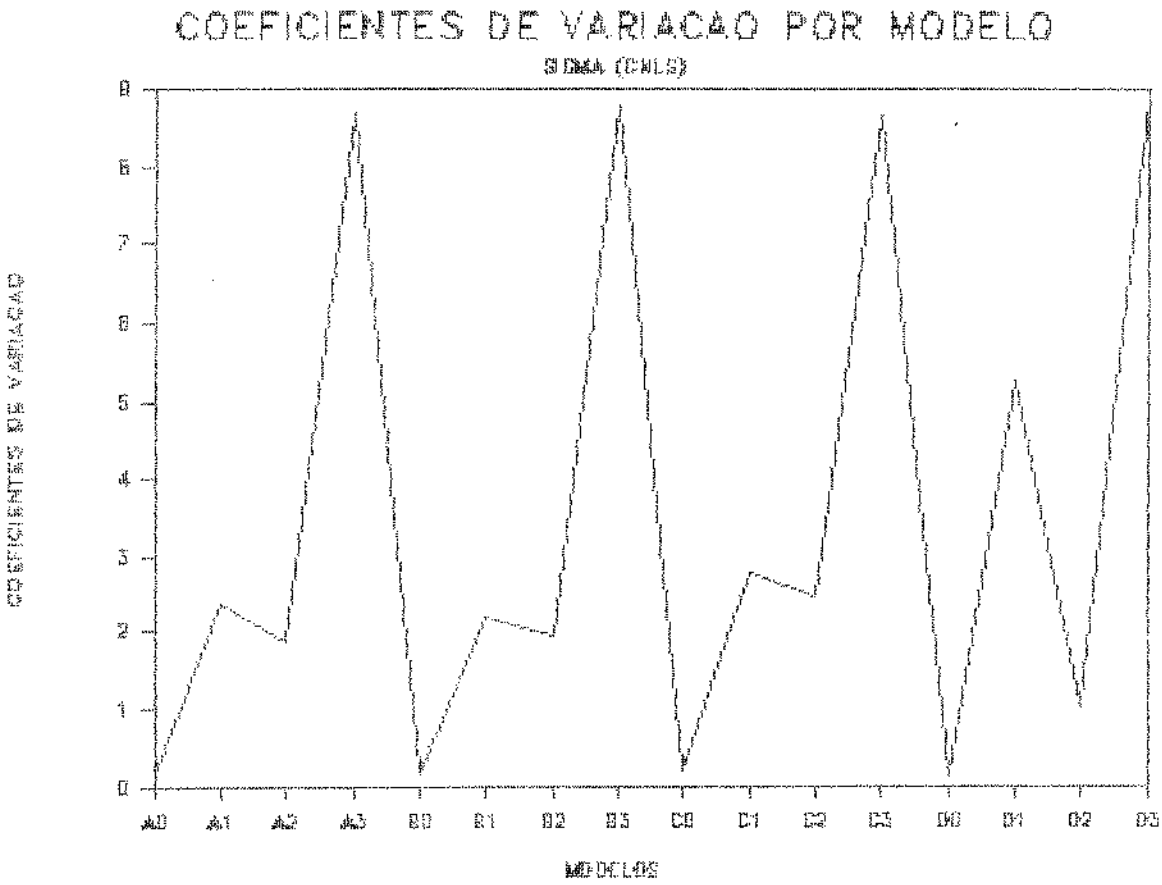

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (CNLS)
COEFICIENTES DE VARIACAO
MODELOS

GRÁFICO 87

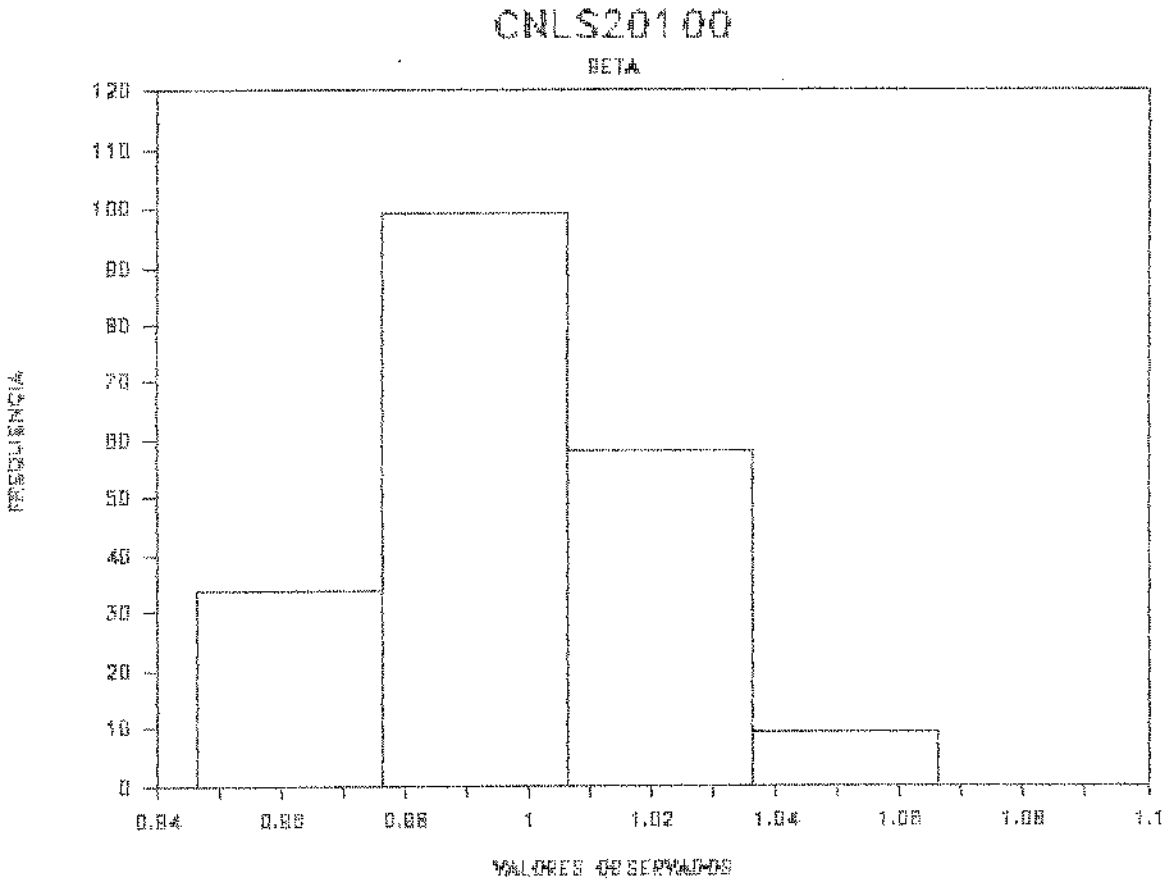

CNLS20100
BETA
120
110
100
90
80
70
60
50
40
30
20
10
0
0.94
0.96
0.98
1
1.02
1.04
1.06
1.08
1.1

GRÁFICO 88

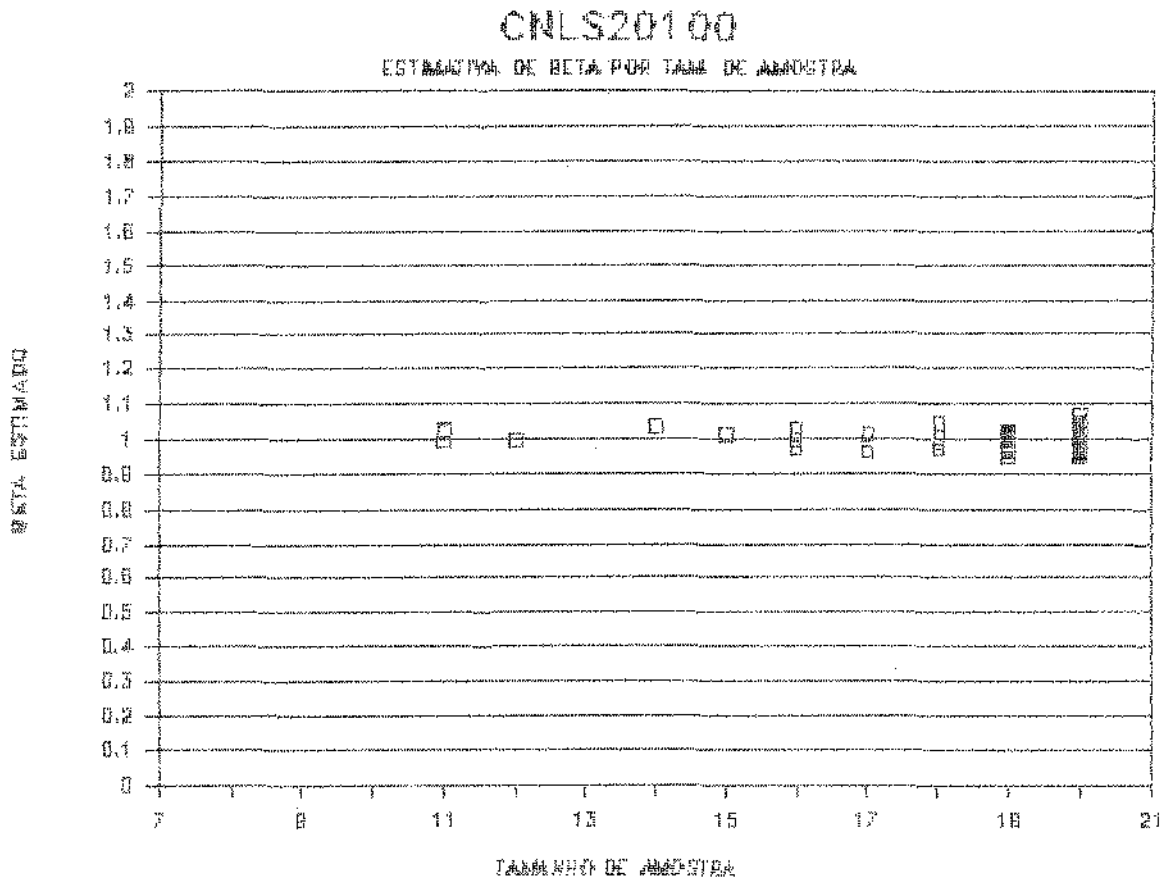

CNLS20100
ESTIMACIÓN DE BETA POR TAMAÑO DE MUESTRA
BETA ESTIMADO
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21
TAMAÑO DE MUESTRA

GRÁFICO 89

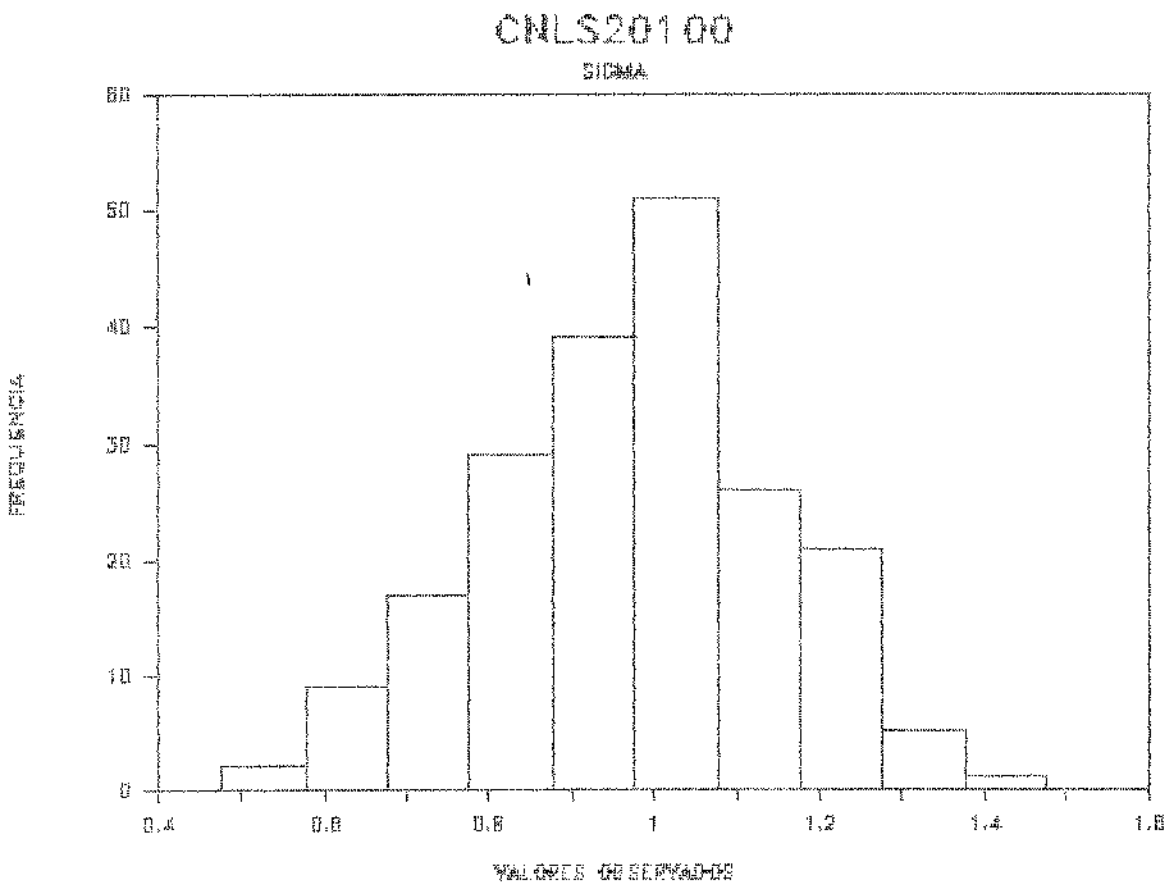


GRÁFICO 90

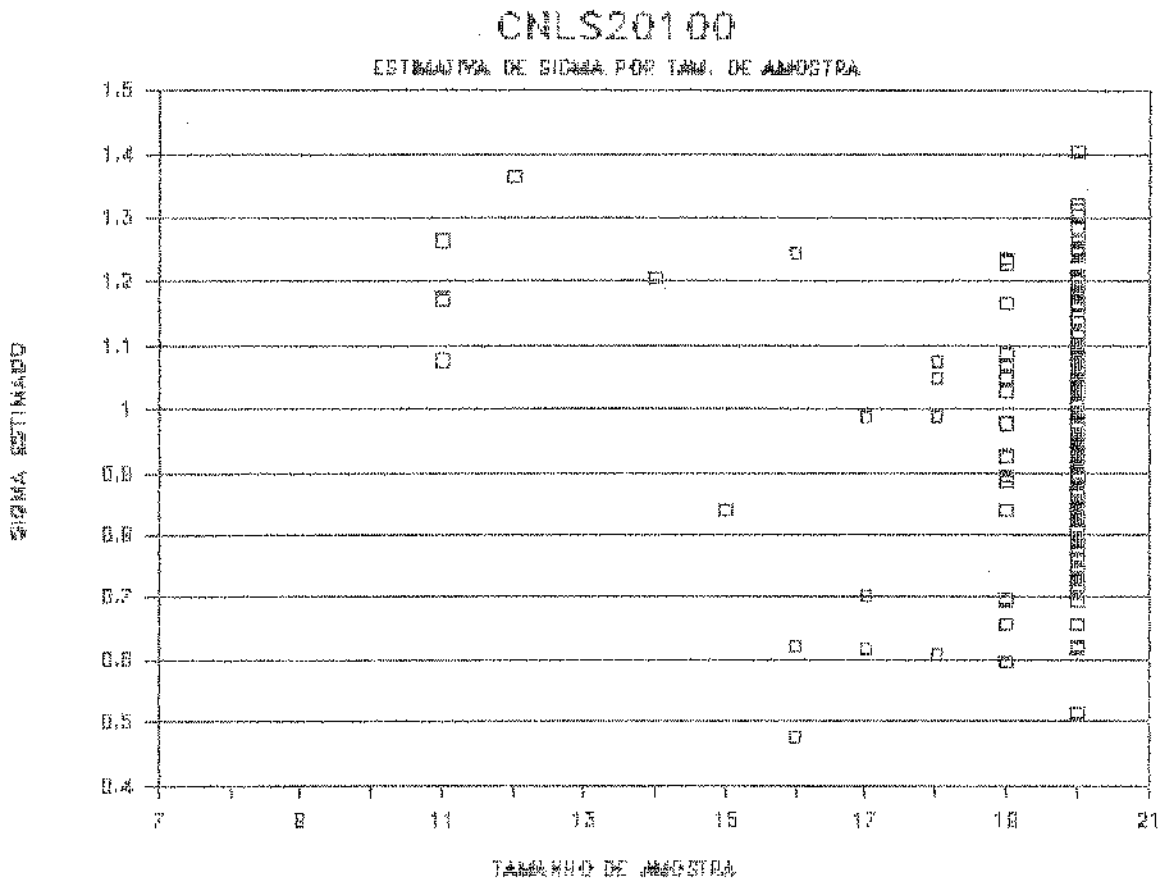

CNLS20100
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 91

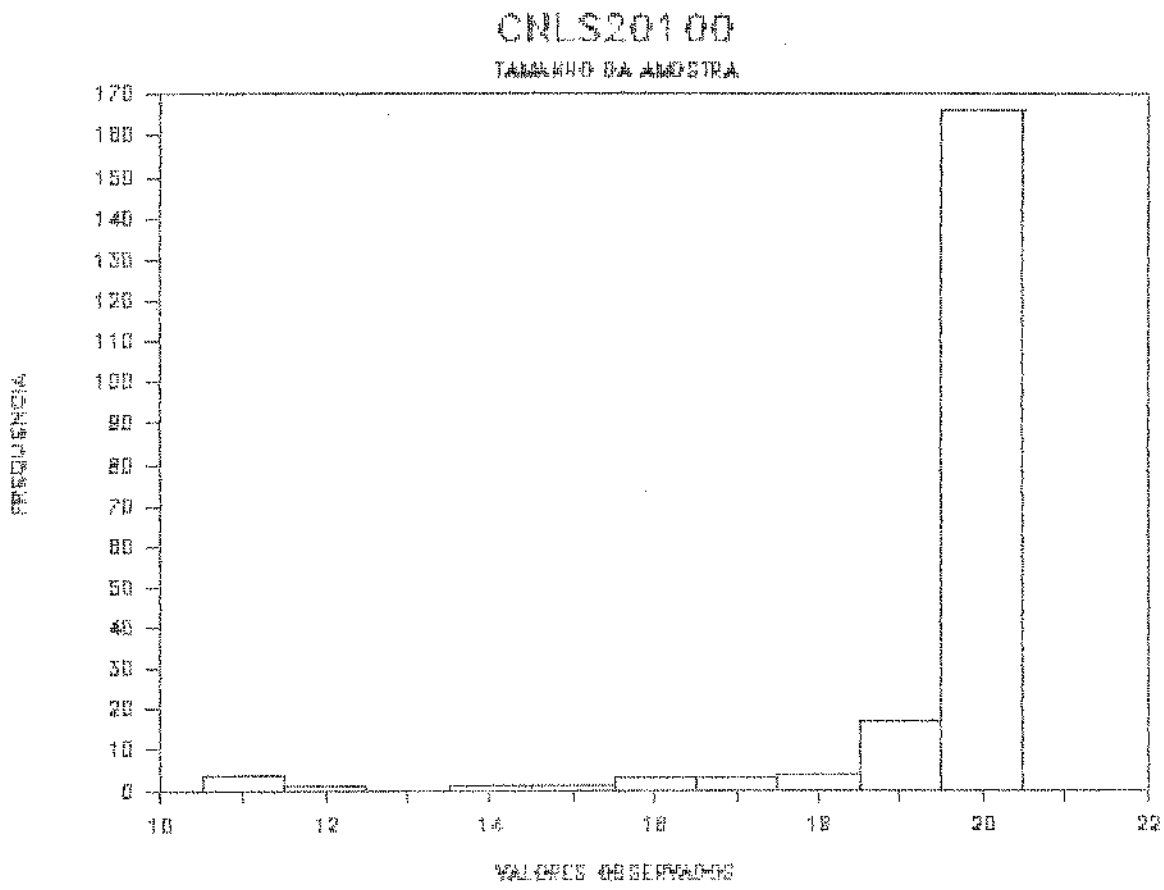

CRLS20100
TAMANHO DA AMOSTRA
FREQUÊNCIA
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 92

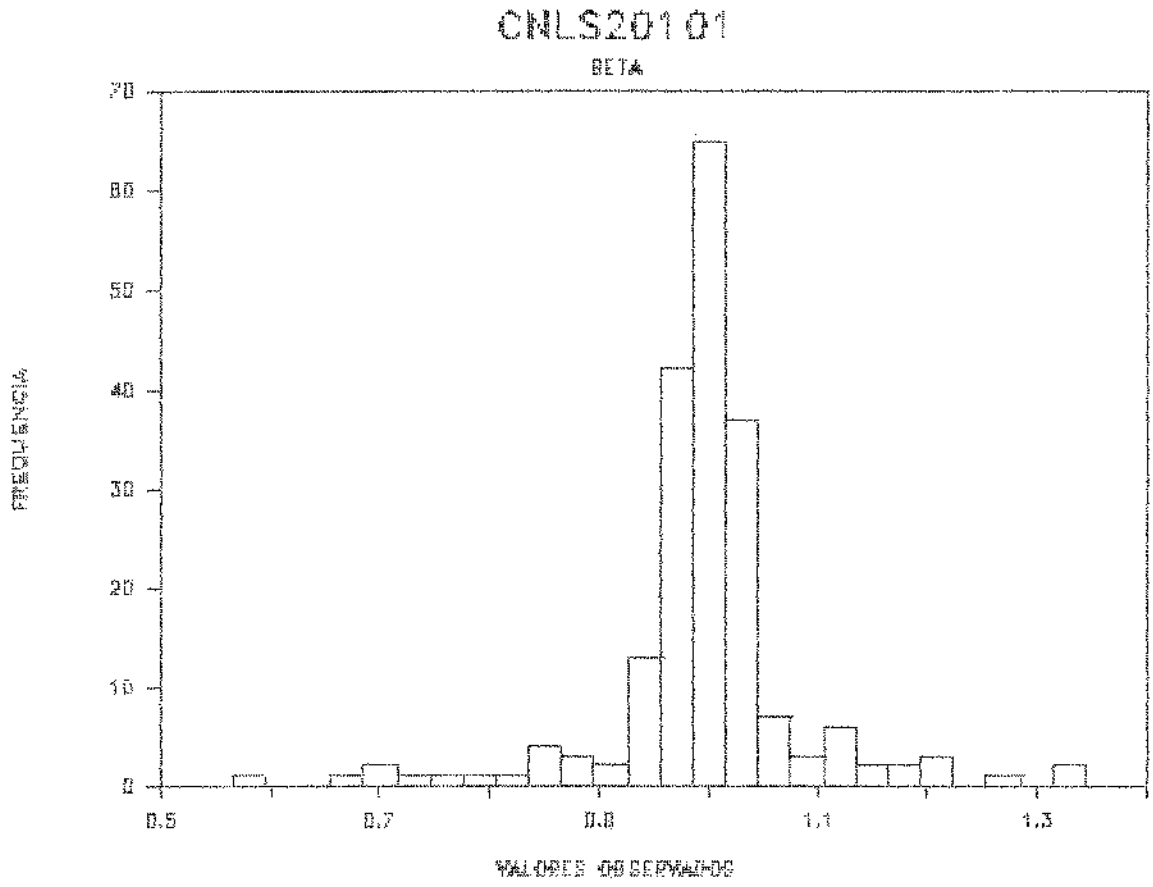


GRÁFICO 93

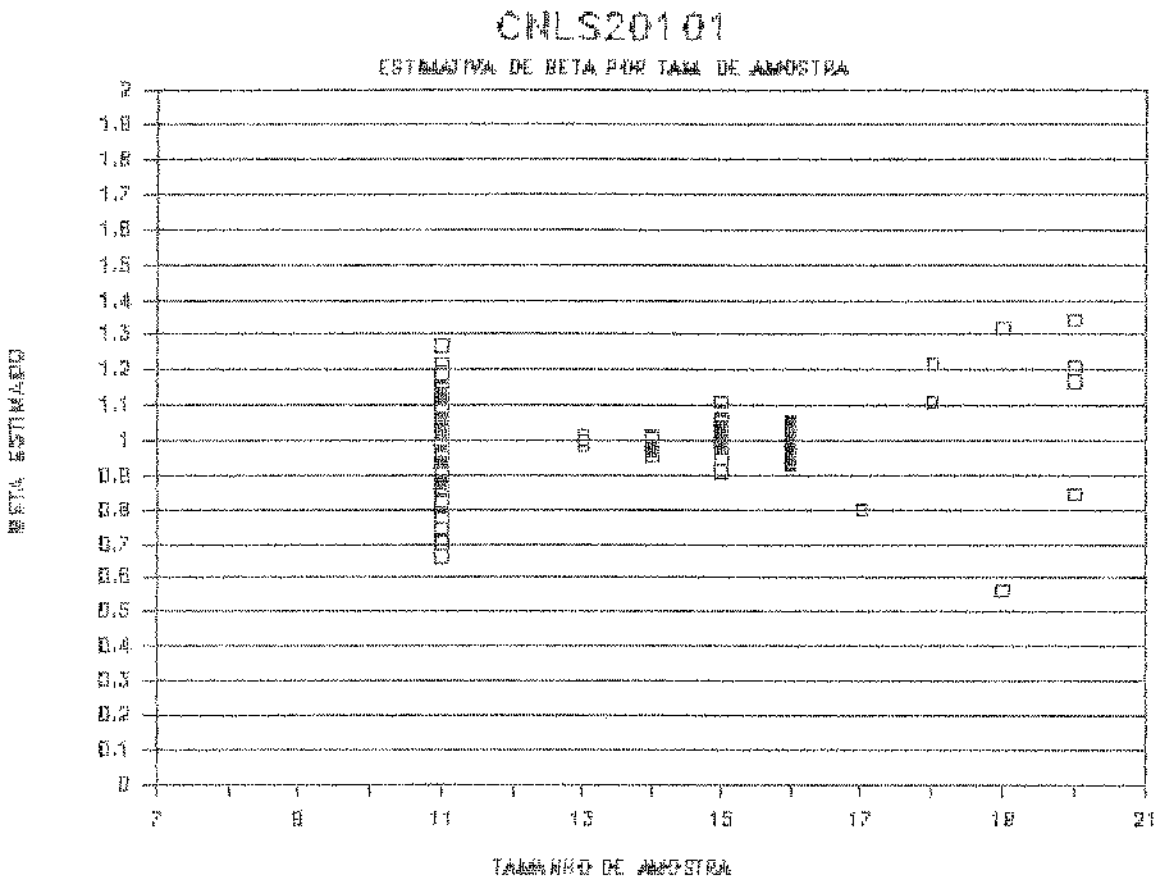

CNLS20101
ESTIMATIVA DE BETA POR TAMANHO DE AMOSTRA
BETA ESTIMADO
TAMANHO DE AMOSTRA
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21

GRÁFICO 94

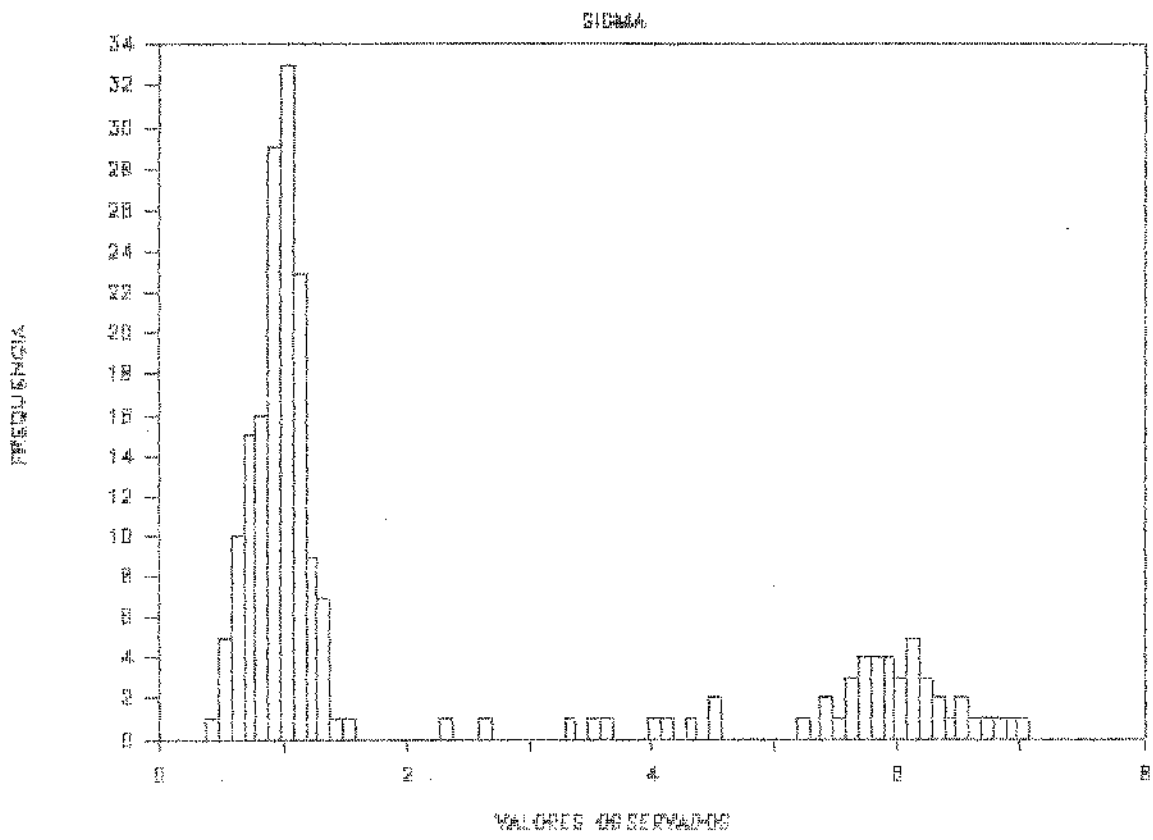


GRÁFICO 95

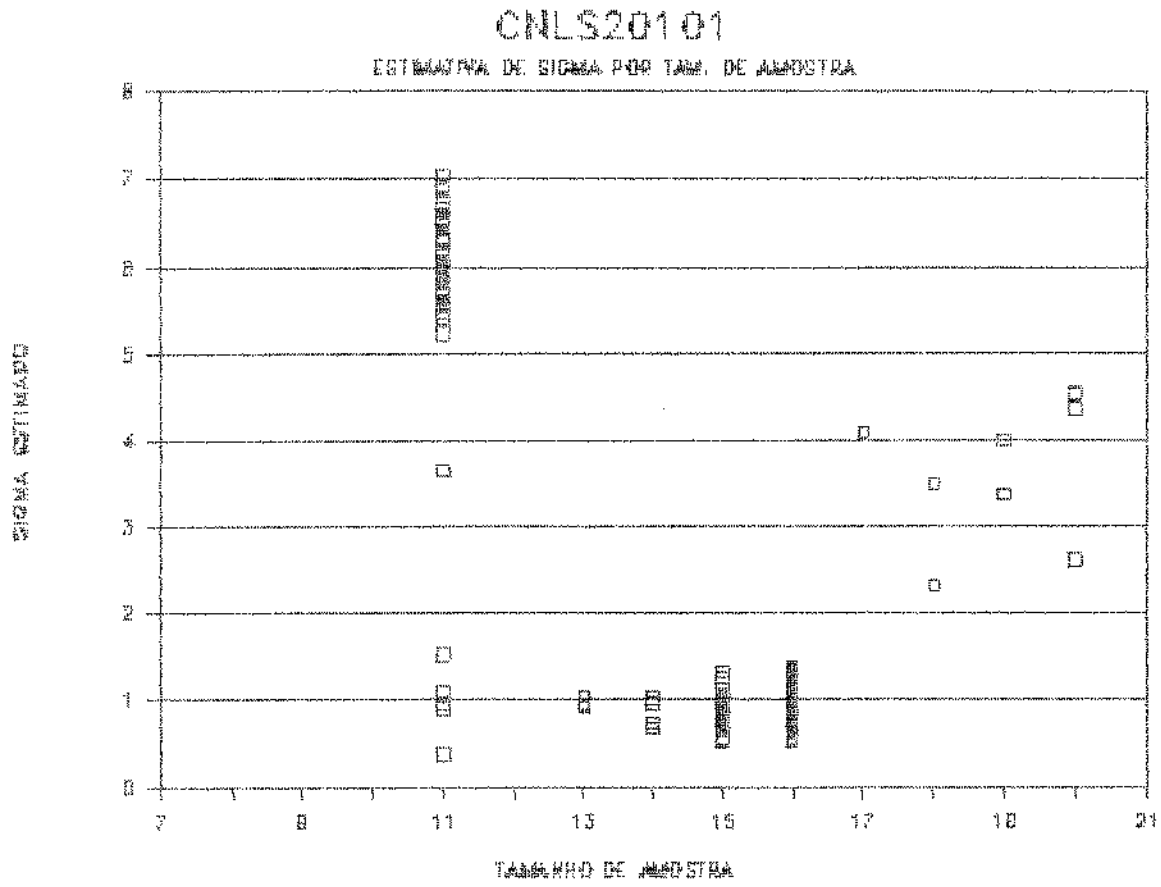

CNLS20101
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DE AMOSTRA
7
8
11
13
15
17
18
21

GRÁFICO 96

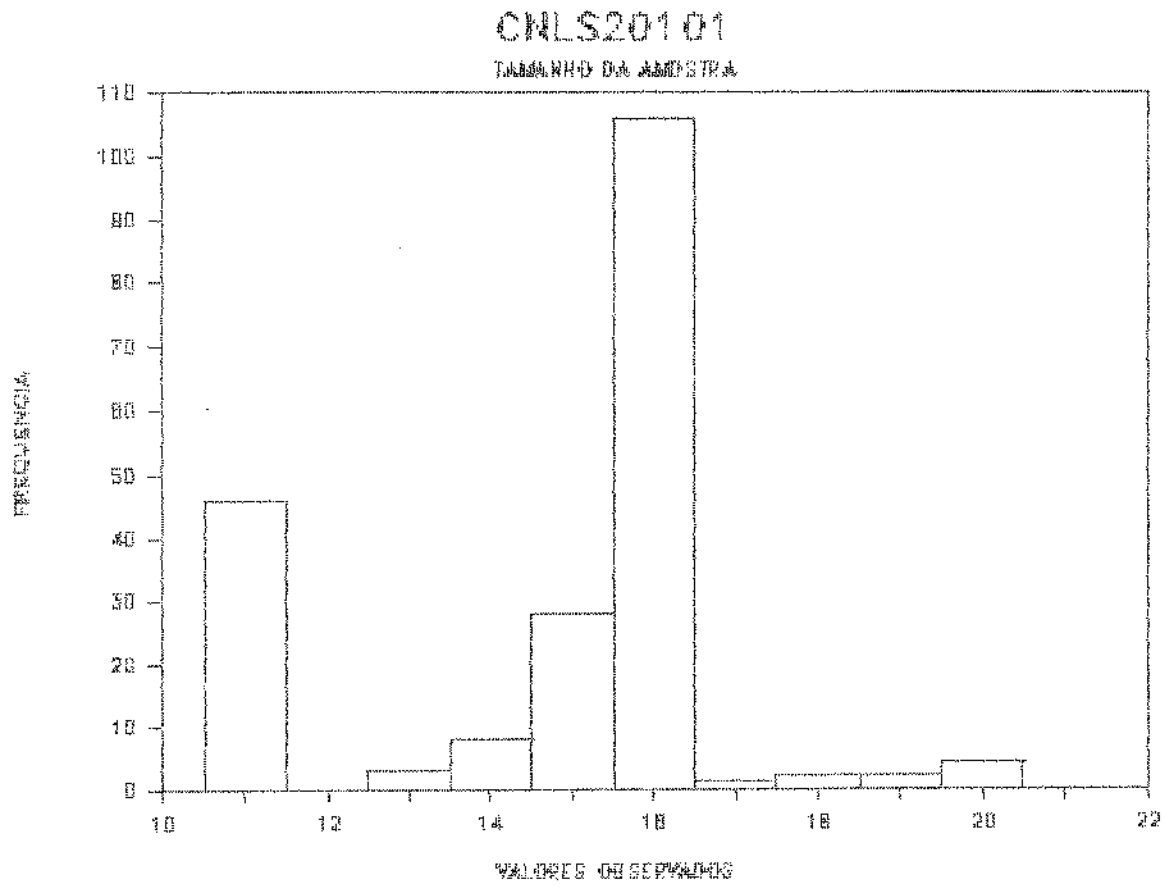

CNLS20101
TAMANHO DA AMOSTRA
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 97

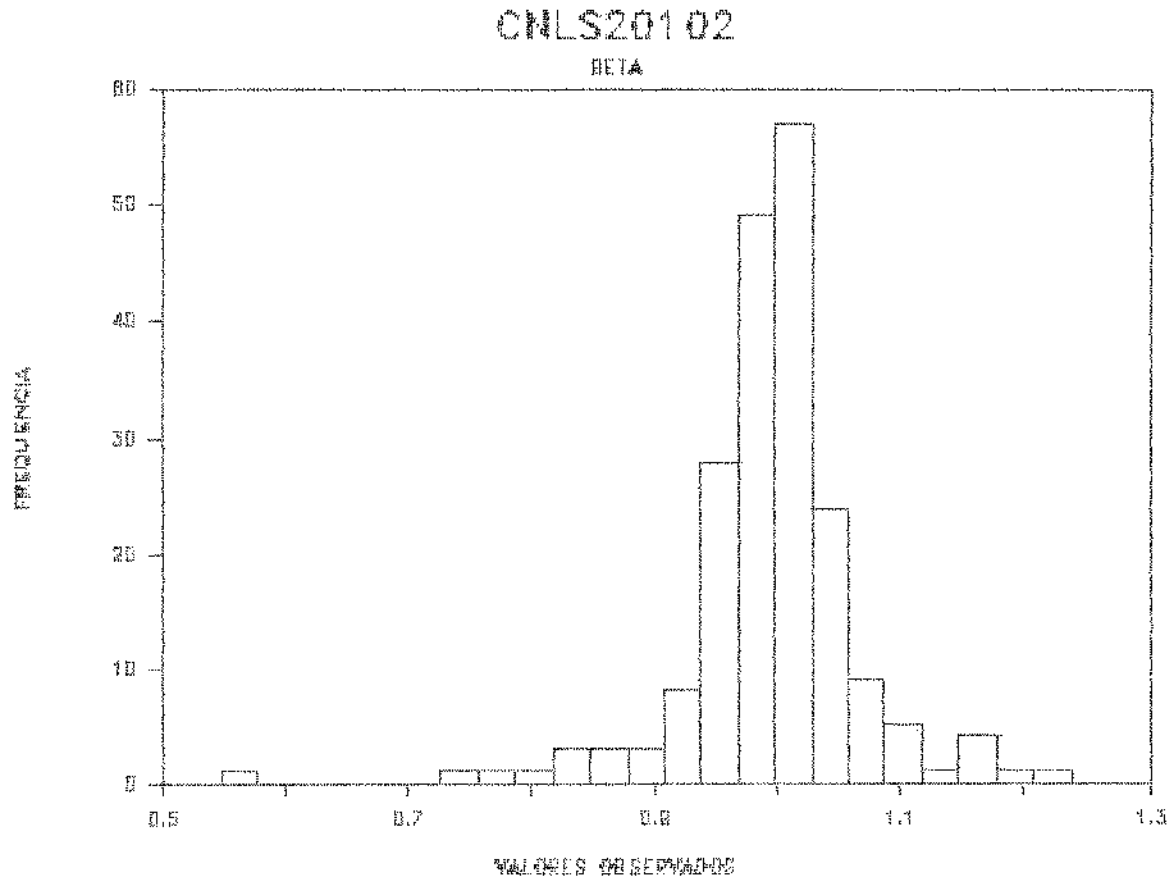


GRÁFICO 98

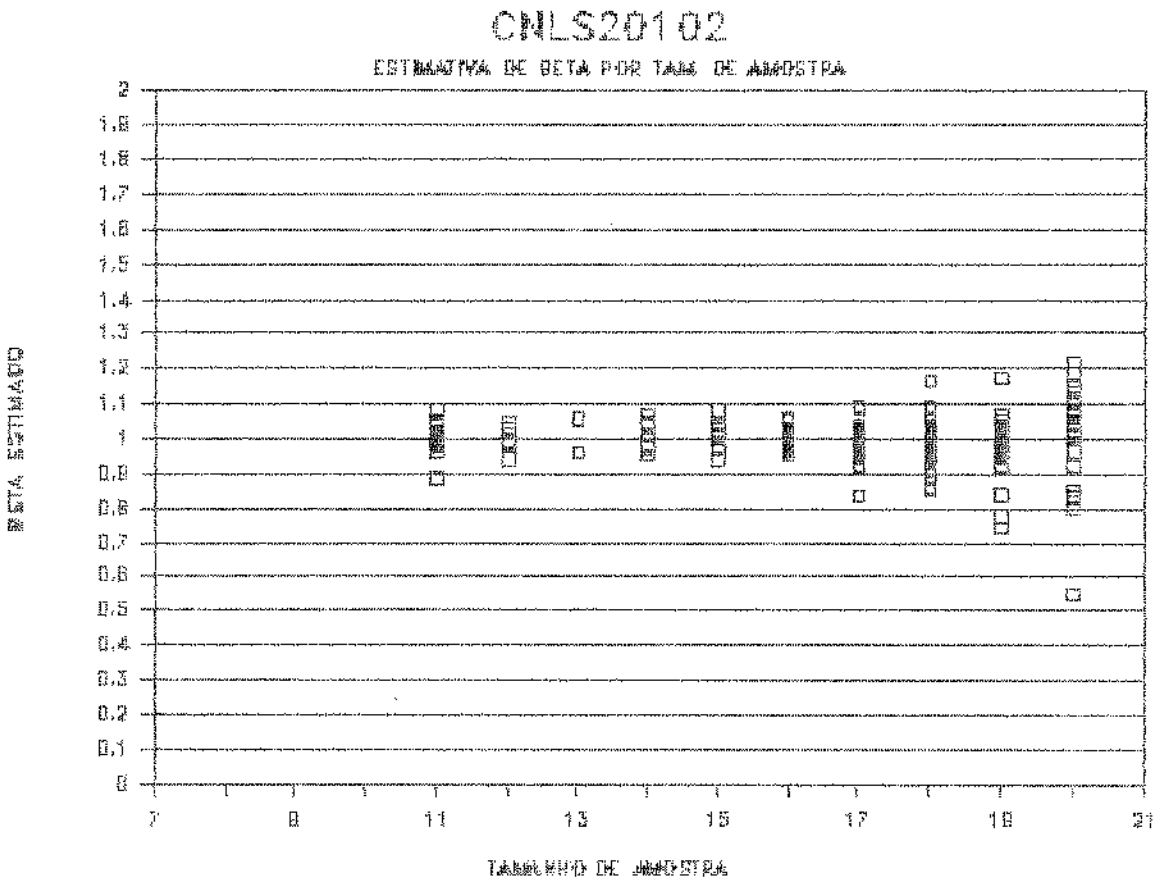

CNLS20102
ESTIMATIVA DE BETA POR TAMANHO DE AMOSTRA
BETA ESTIMADO
2
1,9
1,8
1,7
1,6
1,5
1,4
1,3
1,2
1,1
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 99

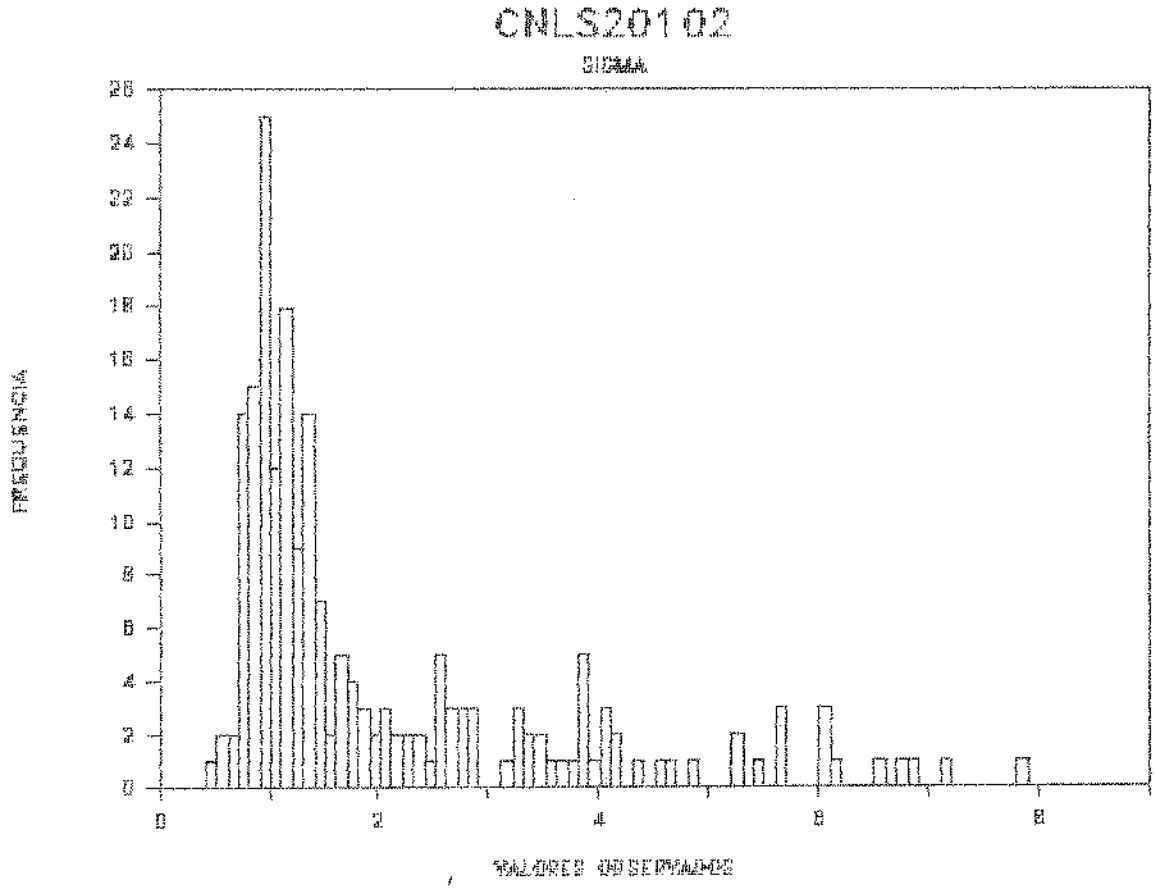


GRÁFICO 100

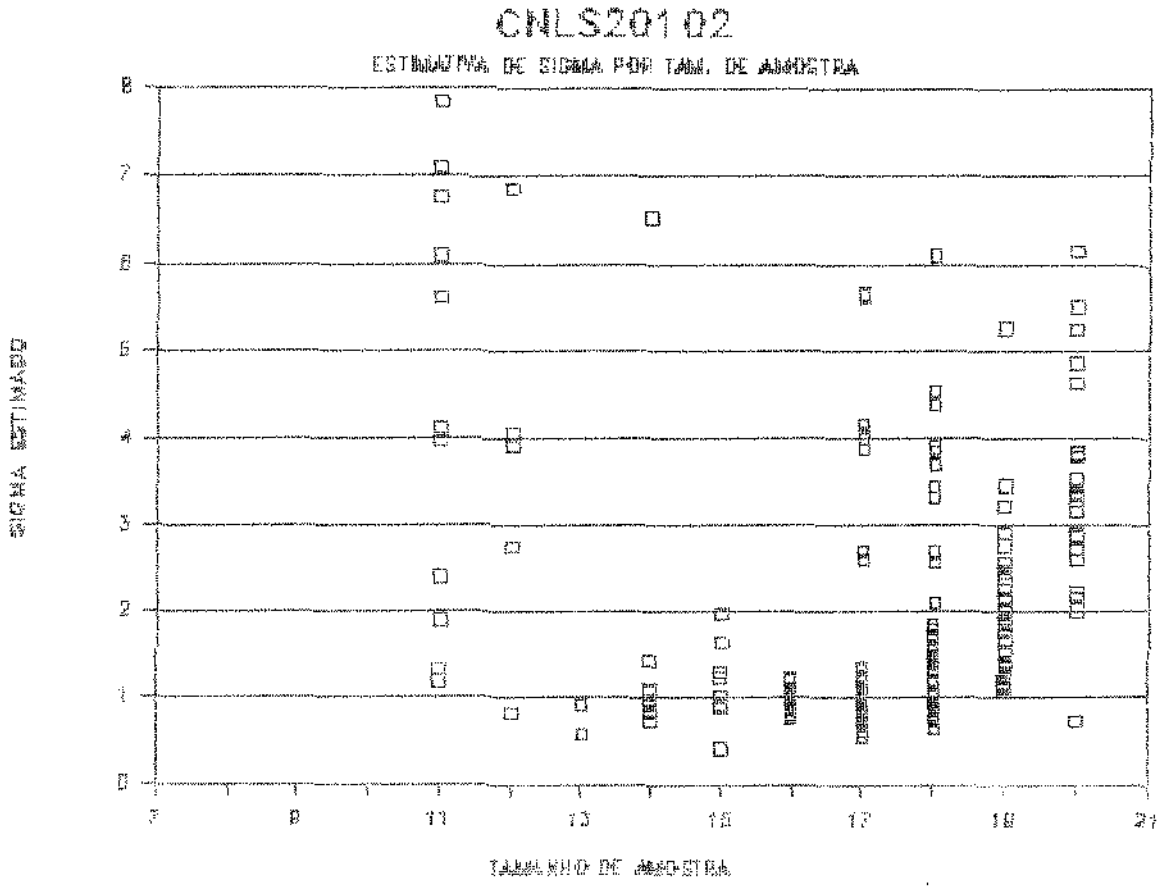

CNLS20102
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
SIGMA ESTIMADO
TAMANHO DE AMOSTRA
0
2
4
6
8
7
9
11
13
15
17
19
21

GRÁFICO 101

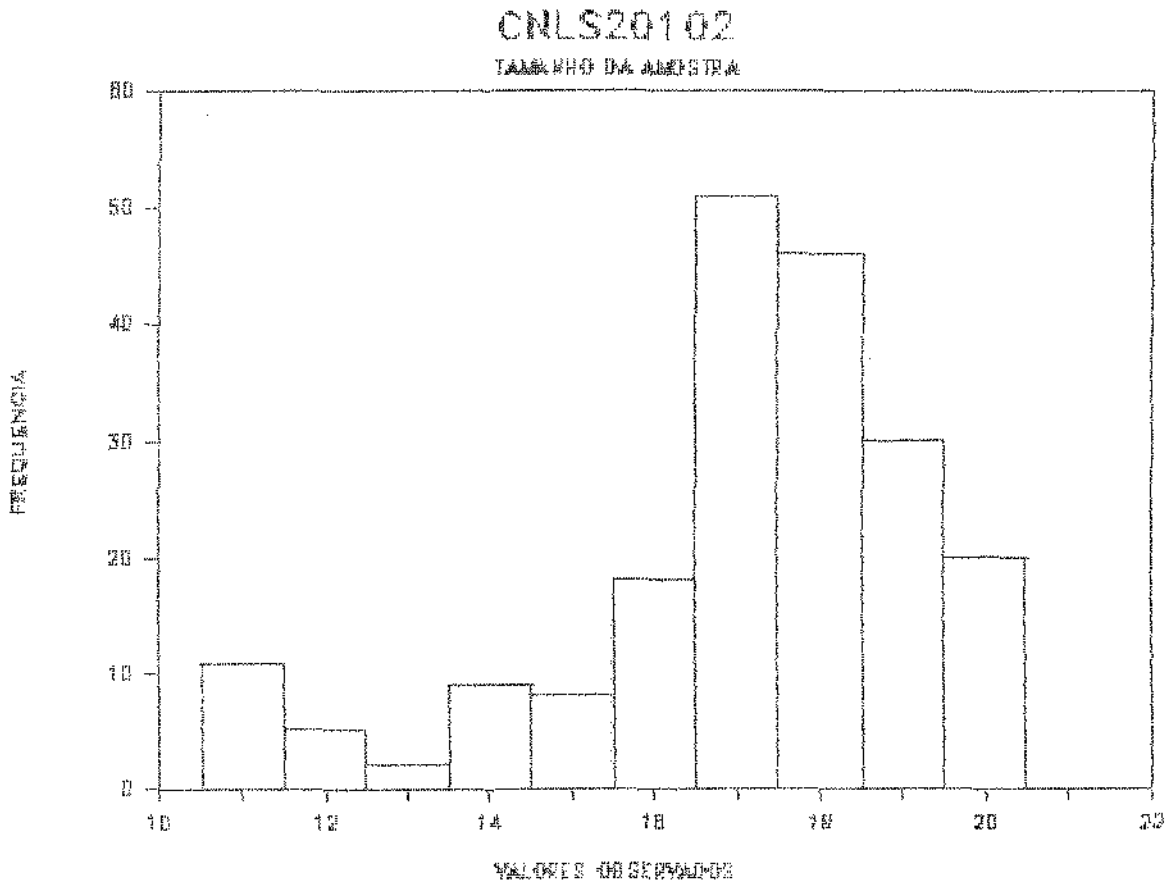

CNLS20102
TAMANHO DA AMOSTRA
FREQUENCIA
0
10
20
30
40
50
60
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 102

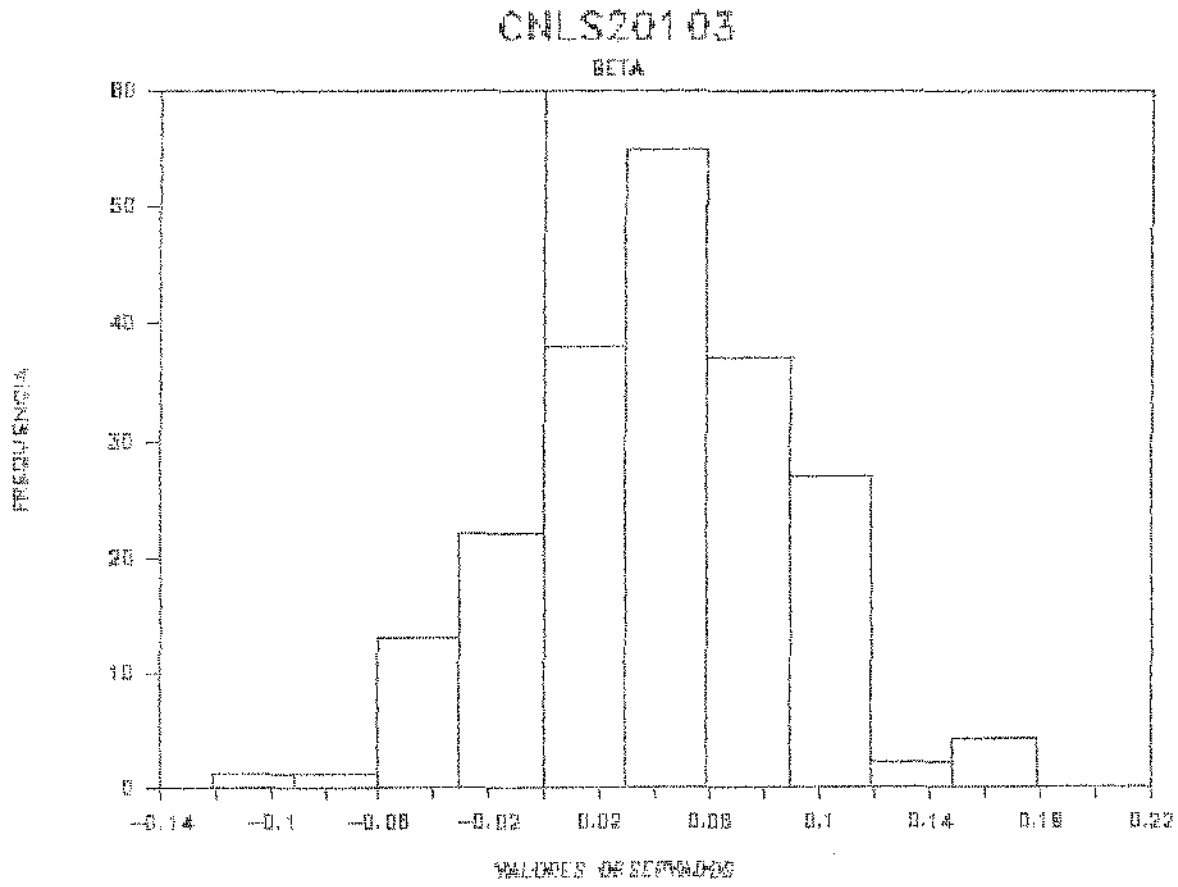

CNLS20103
BETA
60
50
40
30
20
10
0
-0.14
-0.1
-0.06
-0.02
0.02
0.06
0.1
0.14
0.18
0.22

GRÁFICO 103

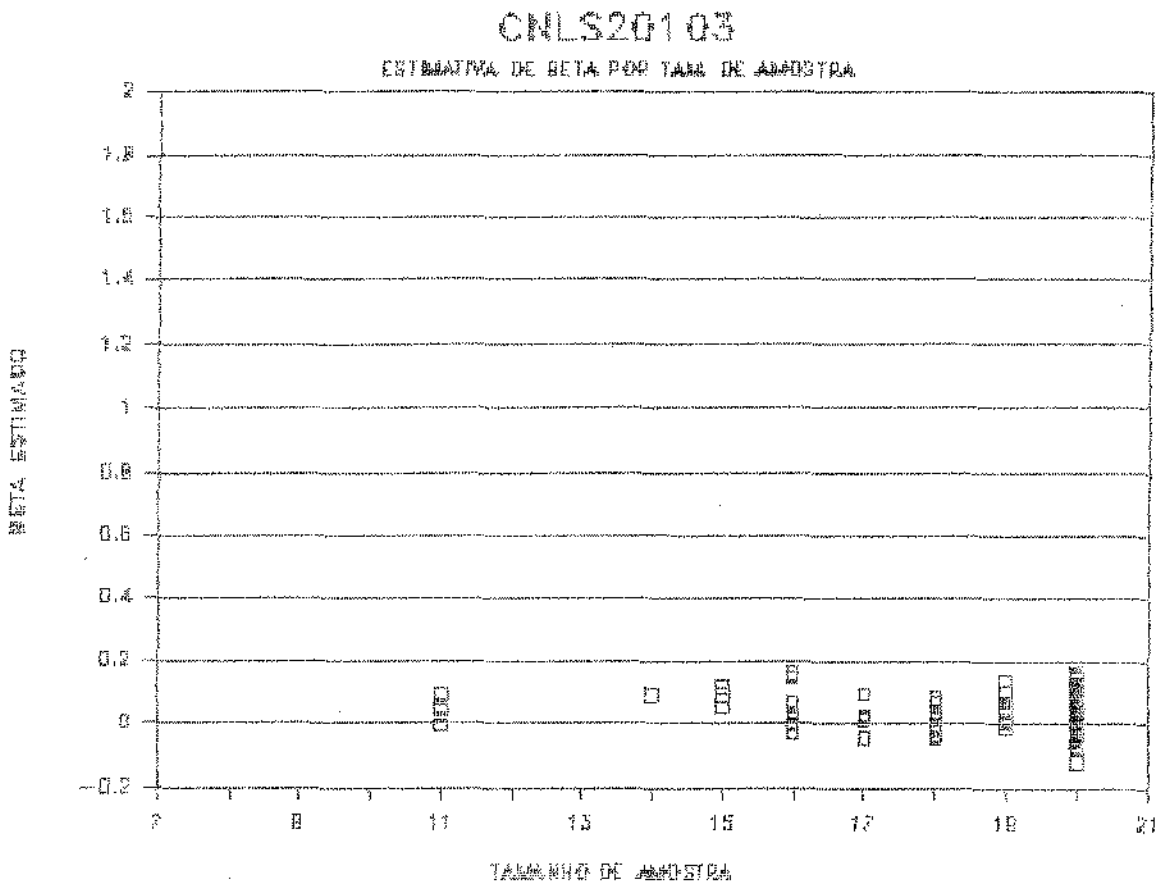

CNLS20103
ESTIMATIVA DE BETA POR TAM DE AMOSTRA
BETA ESTIMADO
TAMANHO DE AMOSTRA
1.2
1.0
0.8
0.6
0.4
0.2
0
-0.2
7
9
11
13
15
17
19
21

GRÁFICO 104

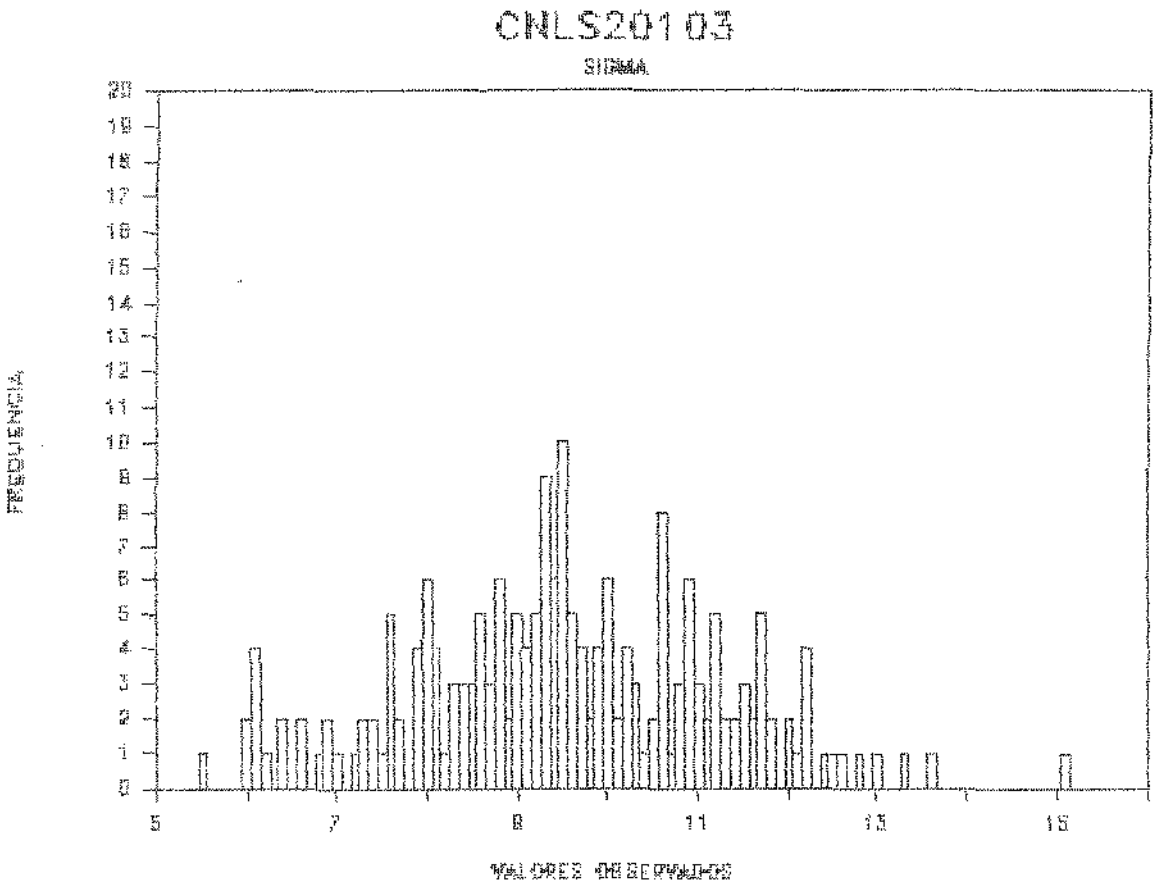


GRÁFICO 105

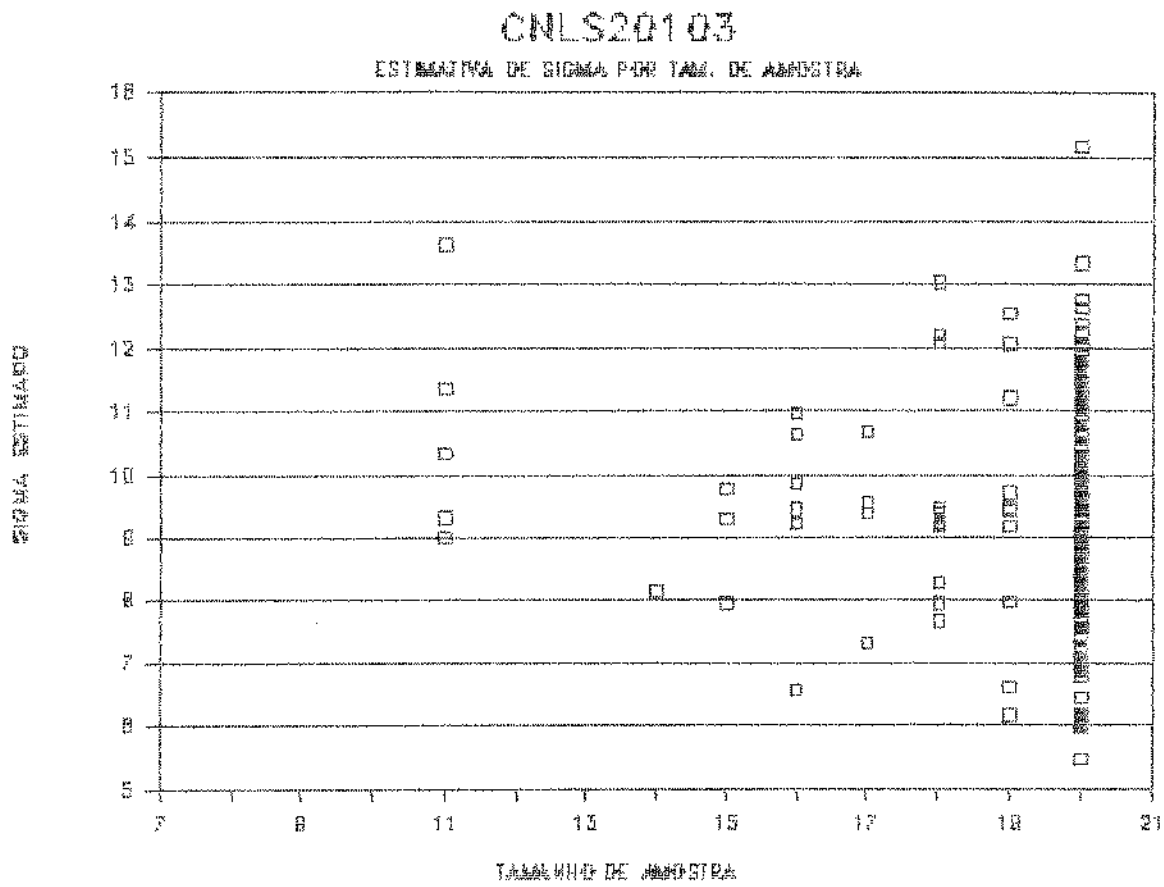

ONLS2103
ESTIMATIVA DE SIGMA POR TAM. DE AMOSTRA
16
15
14
13
12
11
10
9
8
7
6
5
7
9
11
13
15
17
19
21
TAMANHO DE AMOSTRA

GRÁFICO 106

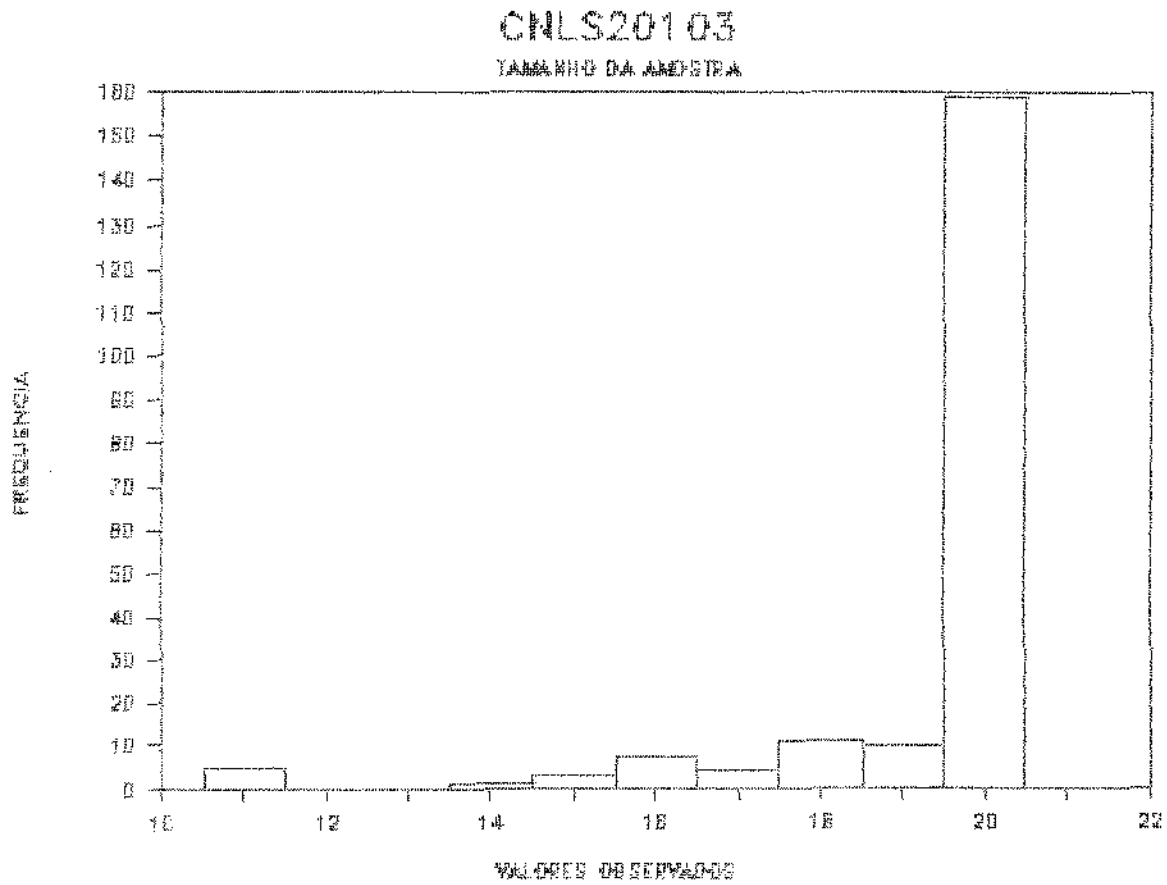

CNLS20103
TAMANHO DA AMOSTRA
FREQUÊNCIA
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 107

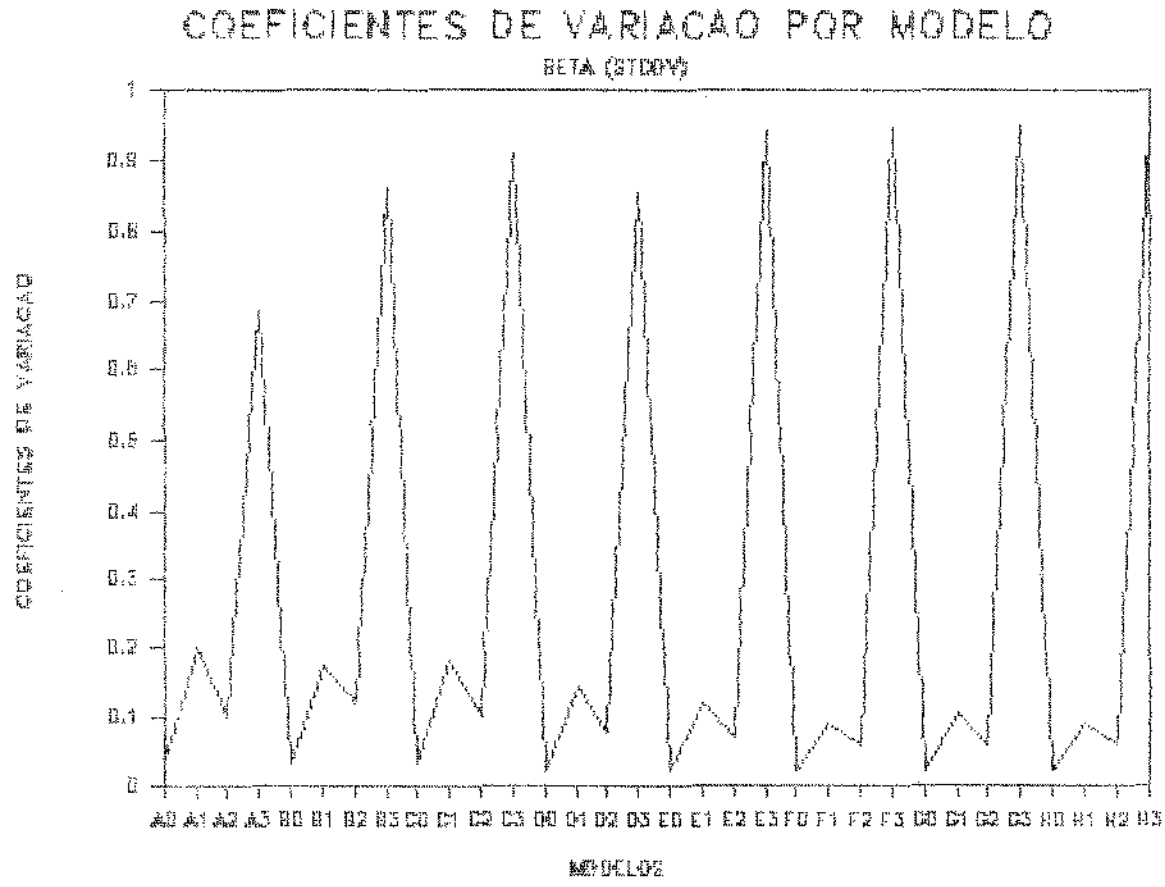

COEFICIENTES DE VARIACAO POR MODELO
BETA (STDDEV)
COEFICIENTES DE VARIACAO
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 108

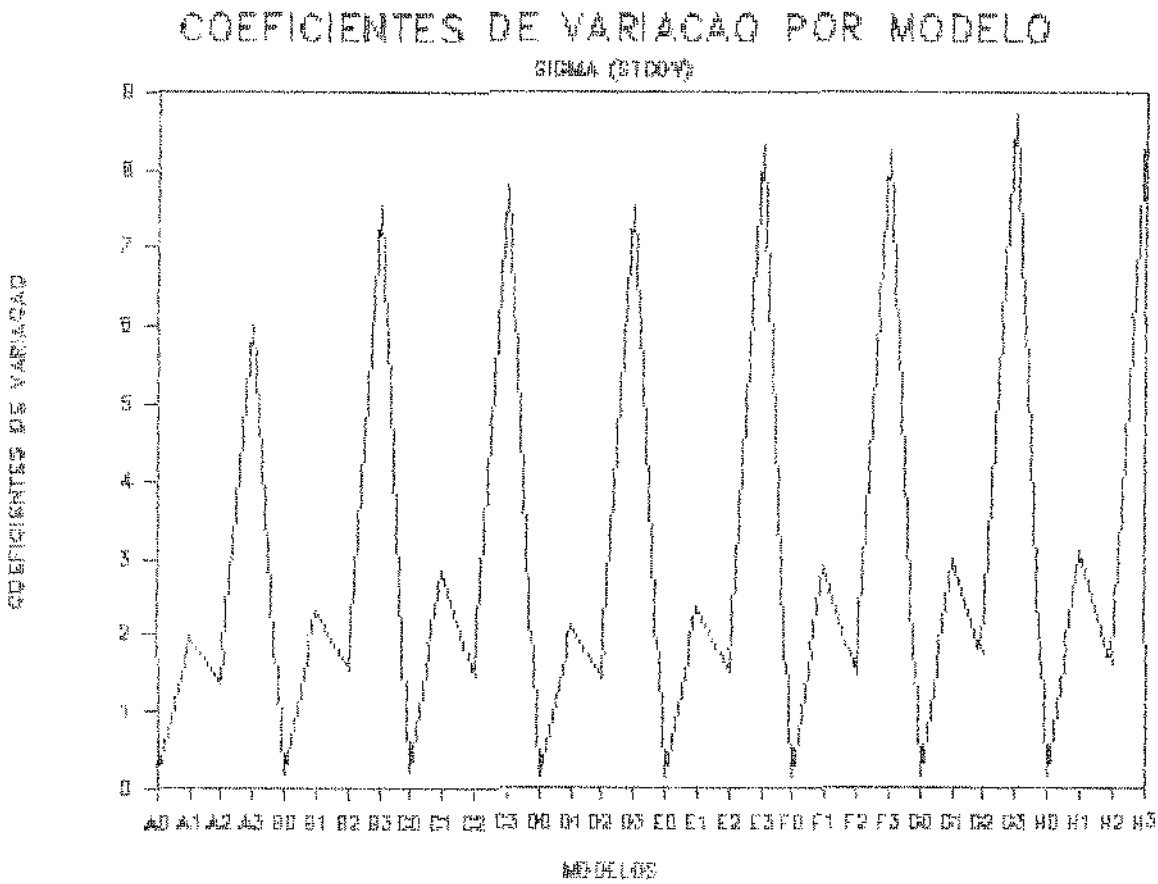

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (STDDEV)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 109

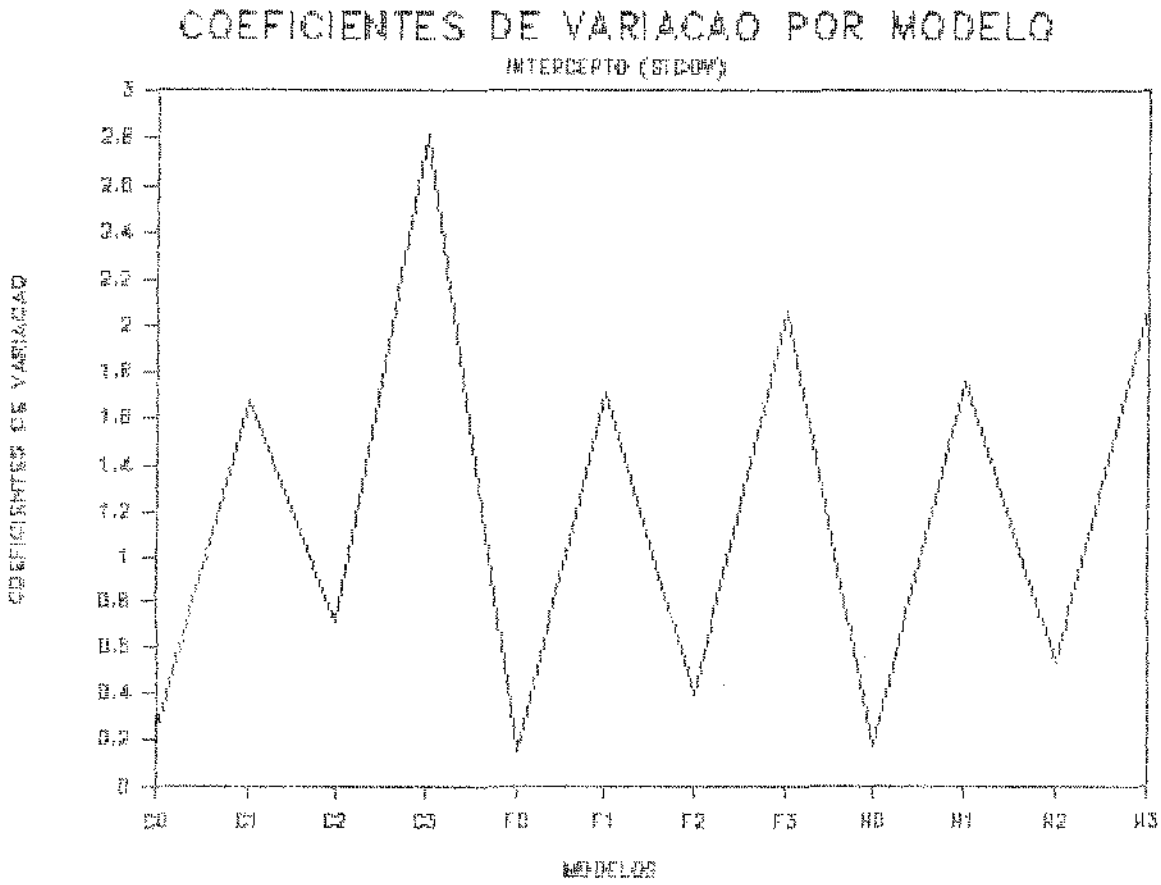

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (STDDW)
COEFICIENTES DE VARIACAO
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
C0
C1
C2
C3
F0
F1
F2
F3
H0
H1
H2
H3
MODELOS

GRÁFICO 110

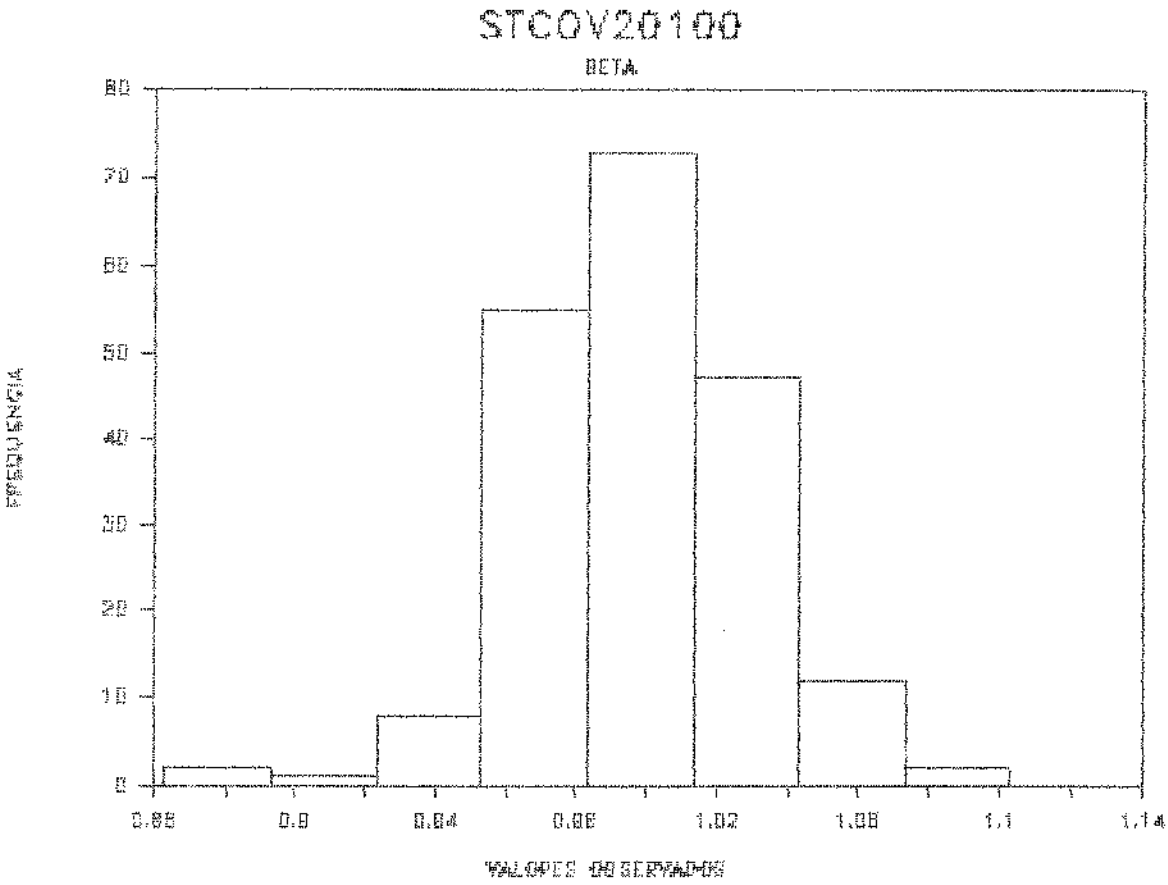

STCOV20100
BETA
80
70
60
50
40
30
20
10
0
0.88
0.9
0.94
0.98
1.02
1.06
1.1
1.14

GRÁFICO 111

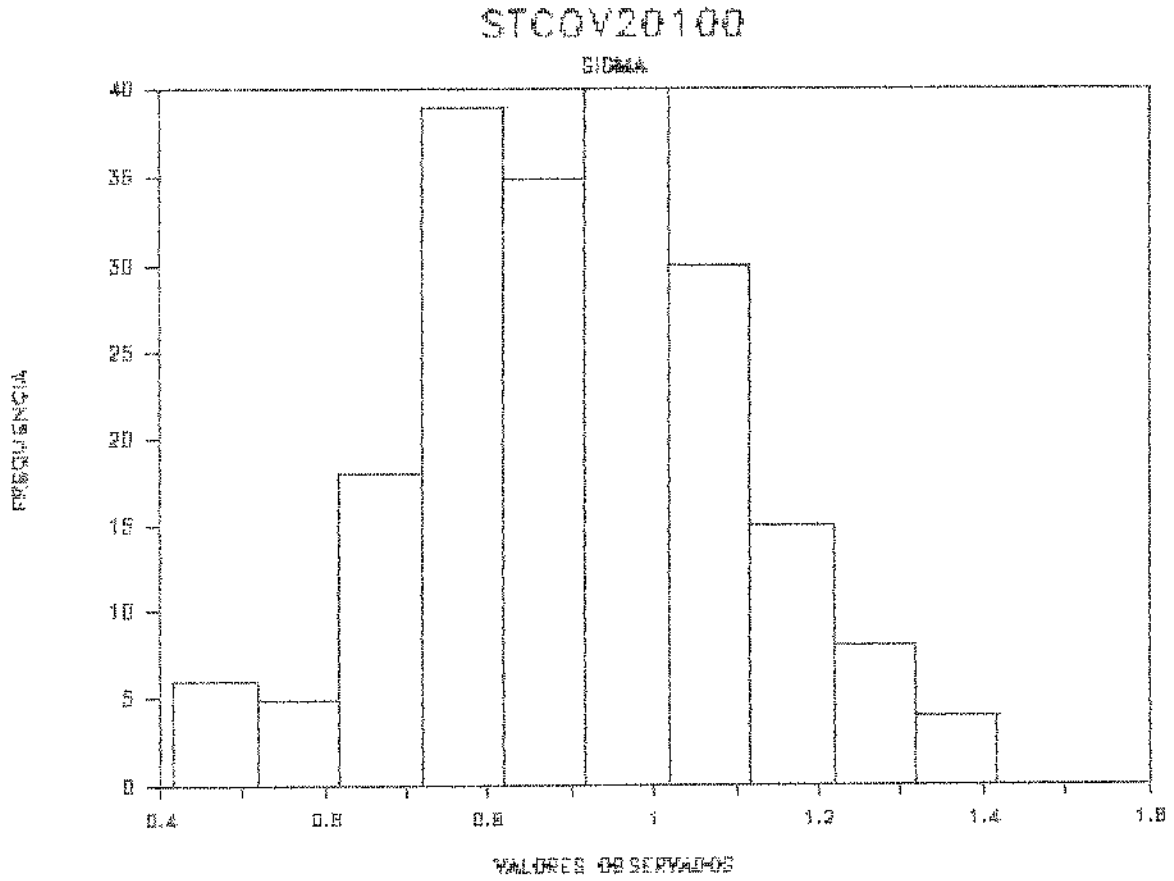


GRÁFICO 112

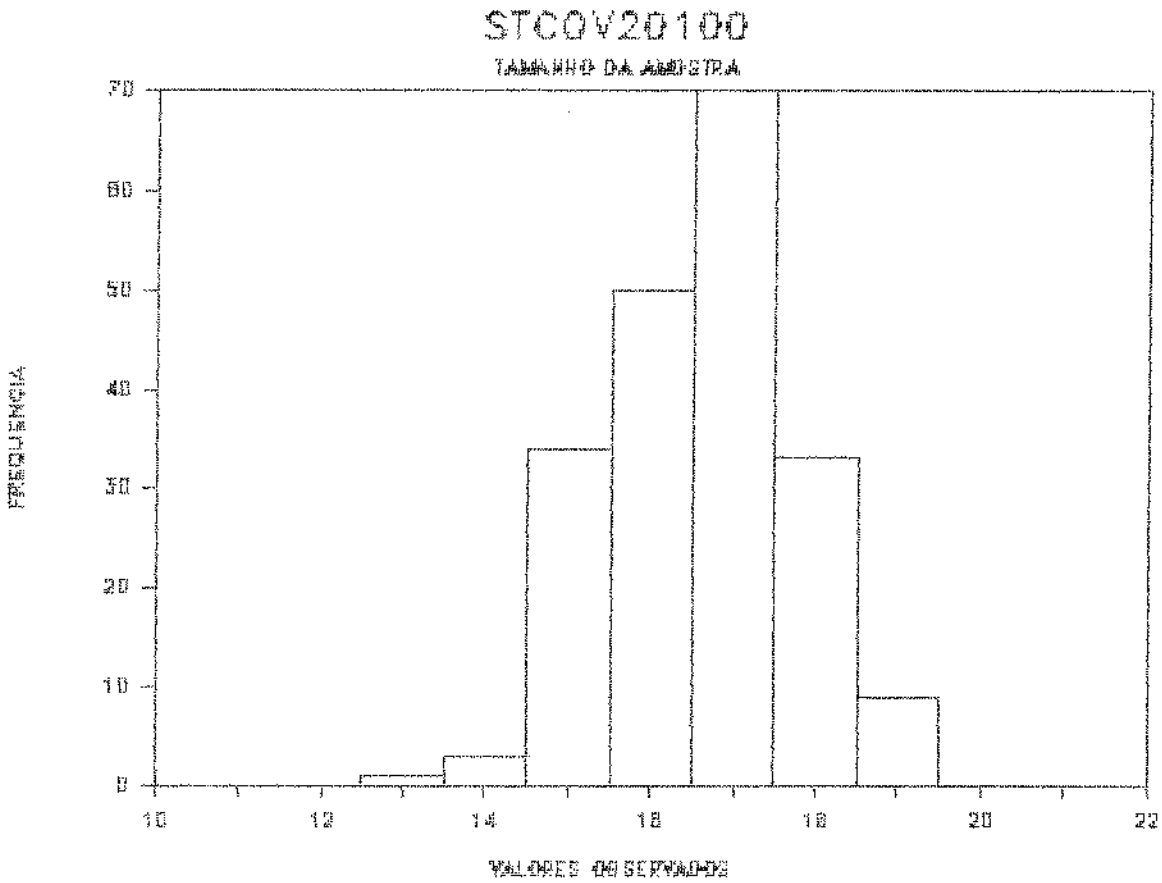

STCOV2010
TAMANHO DA AMOSTRA
FREQUÊNCIA
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 113

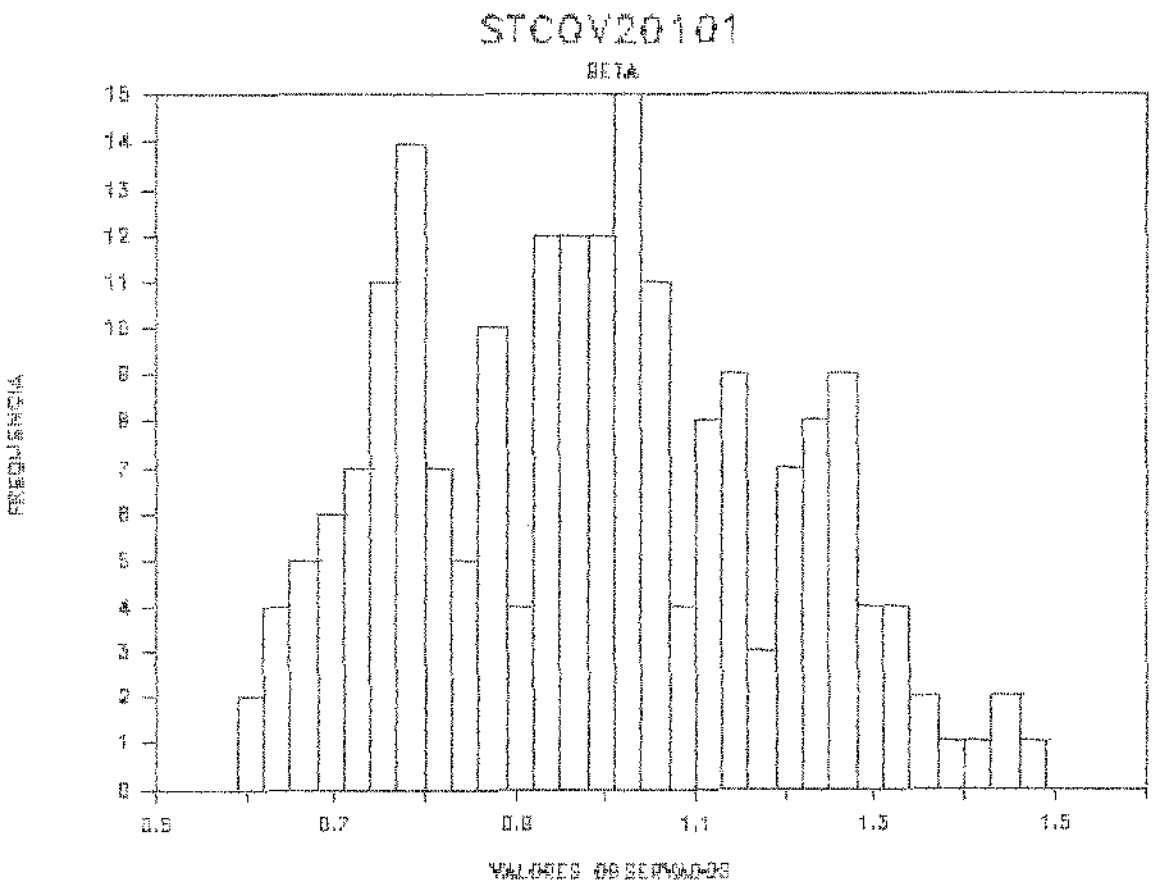


GRÁFICO 114

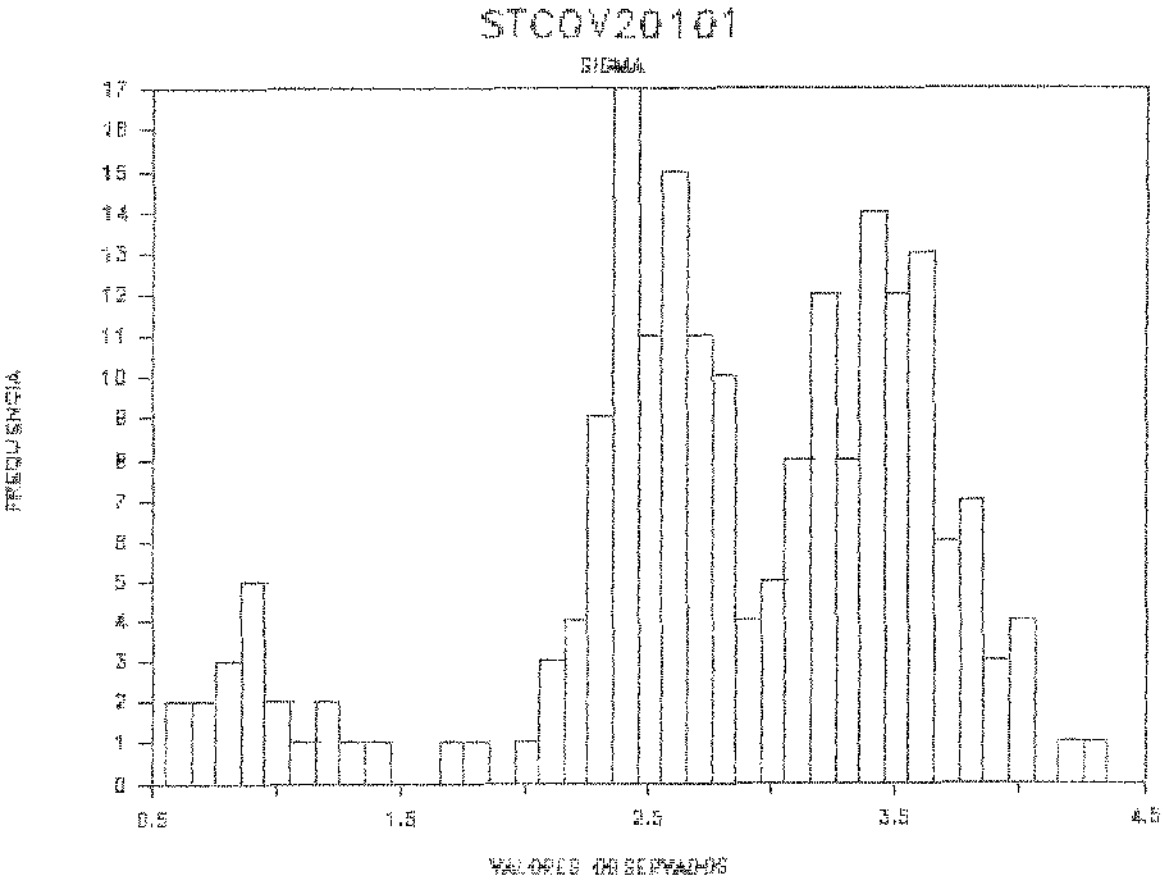


GRÁFICO 115

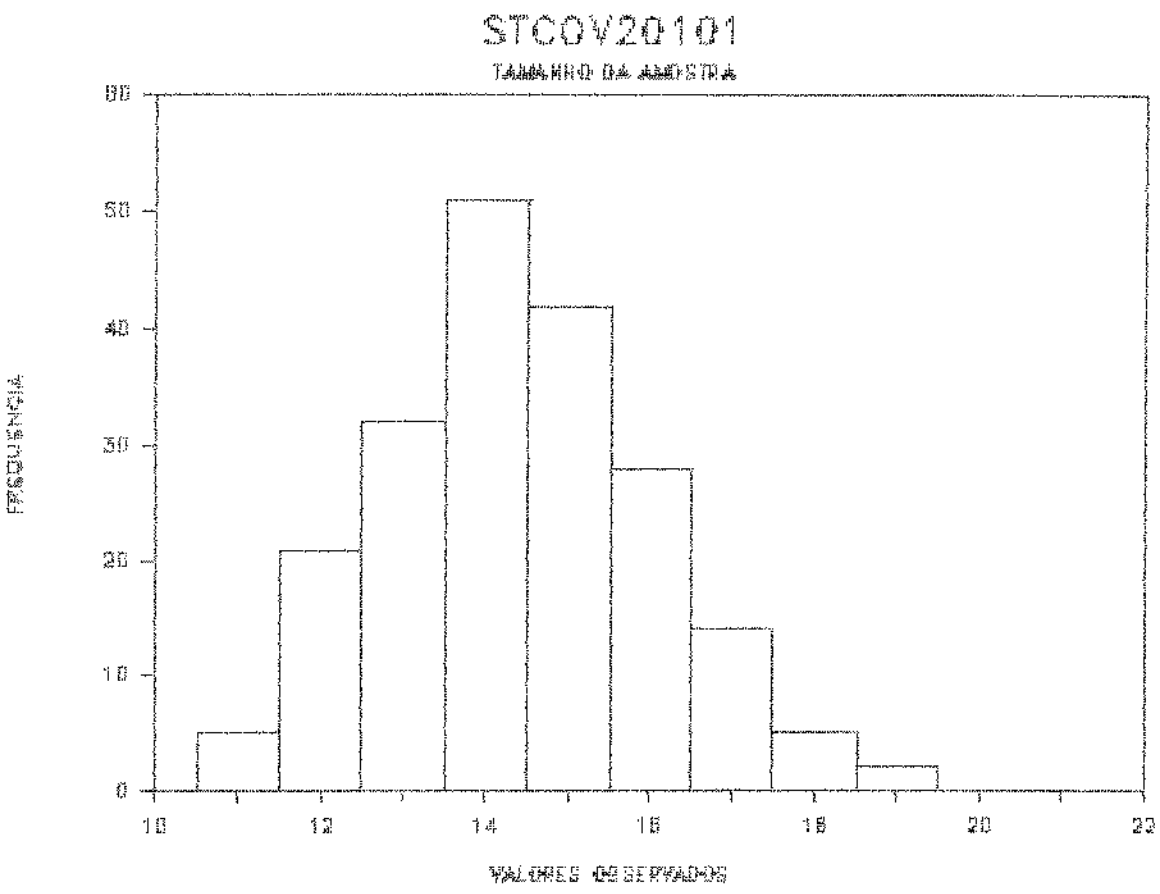

STCOV20101
TAMANHO DA AMOSTRA
FREQUÊNCIA
0
10
20
30
40
50
60
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 116

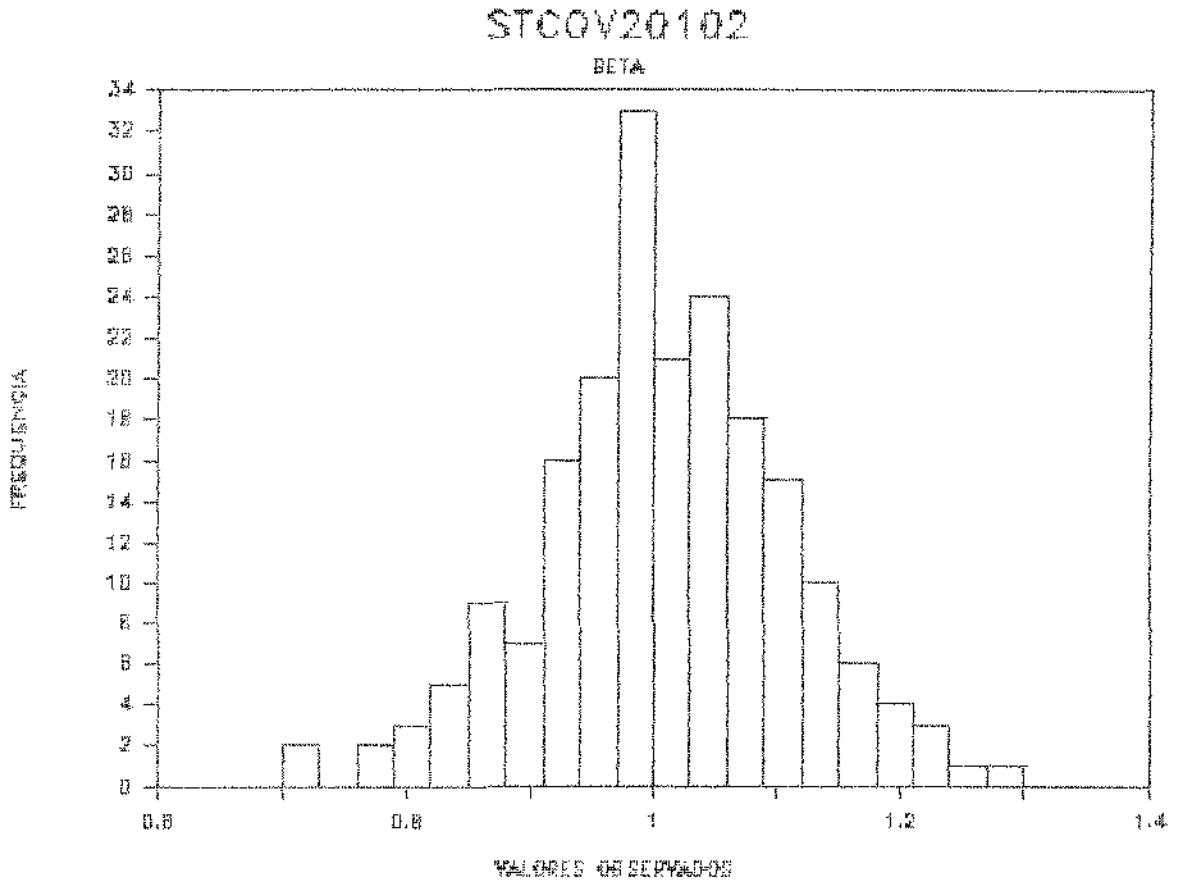

STCOV20102
BETA
FRECUENCIA
0.6
0.8
1
1.2
1.4
VALORES OBSERVADOS

GRÁFICO 117

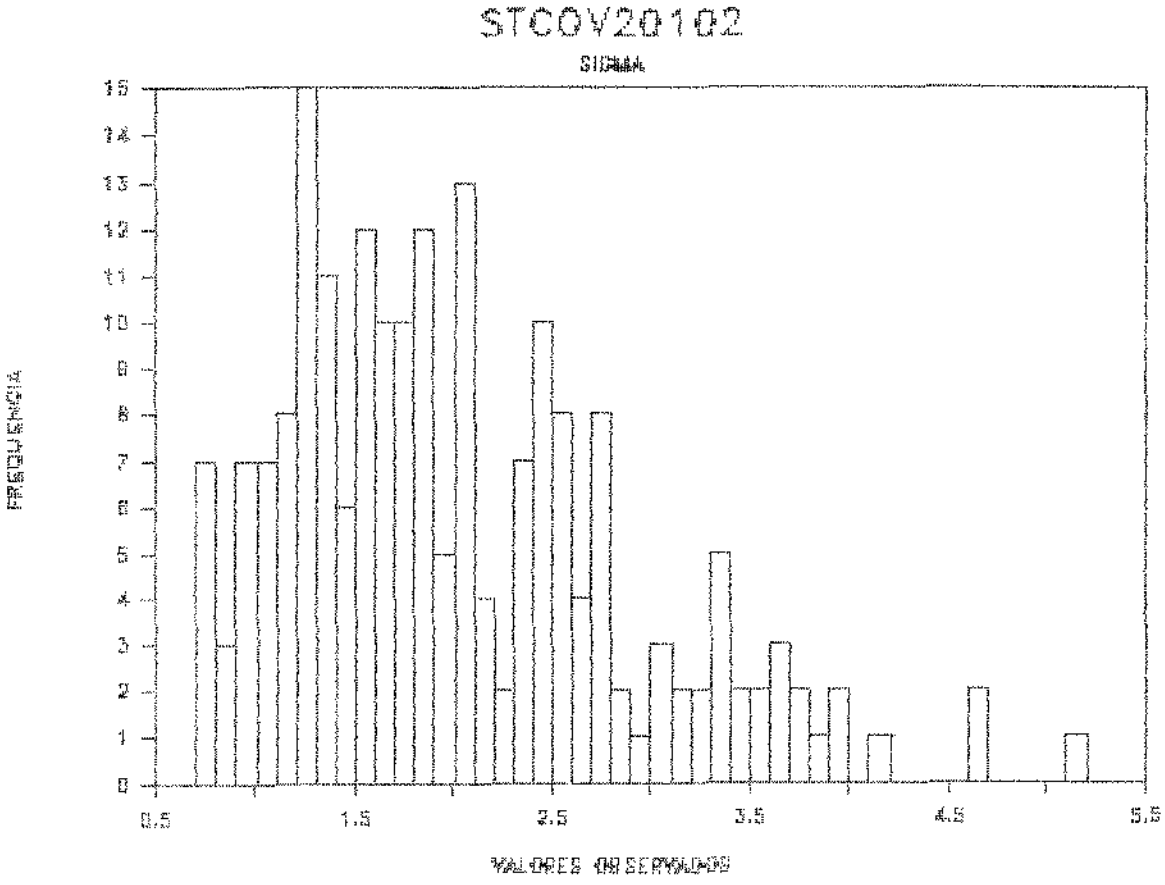

STCOV20102
SIGMA
FRECUENCIA
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
0.5
1.5
2.5
3.5
4.5
5.5
VALORES OBSERVADOS

GRÁFICO 118

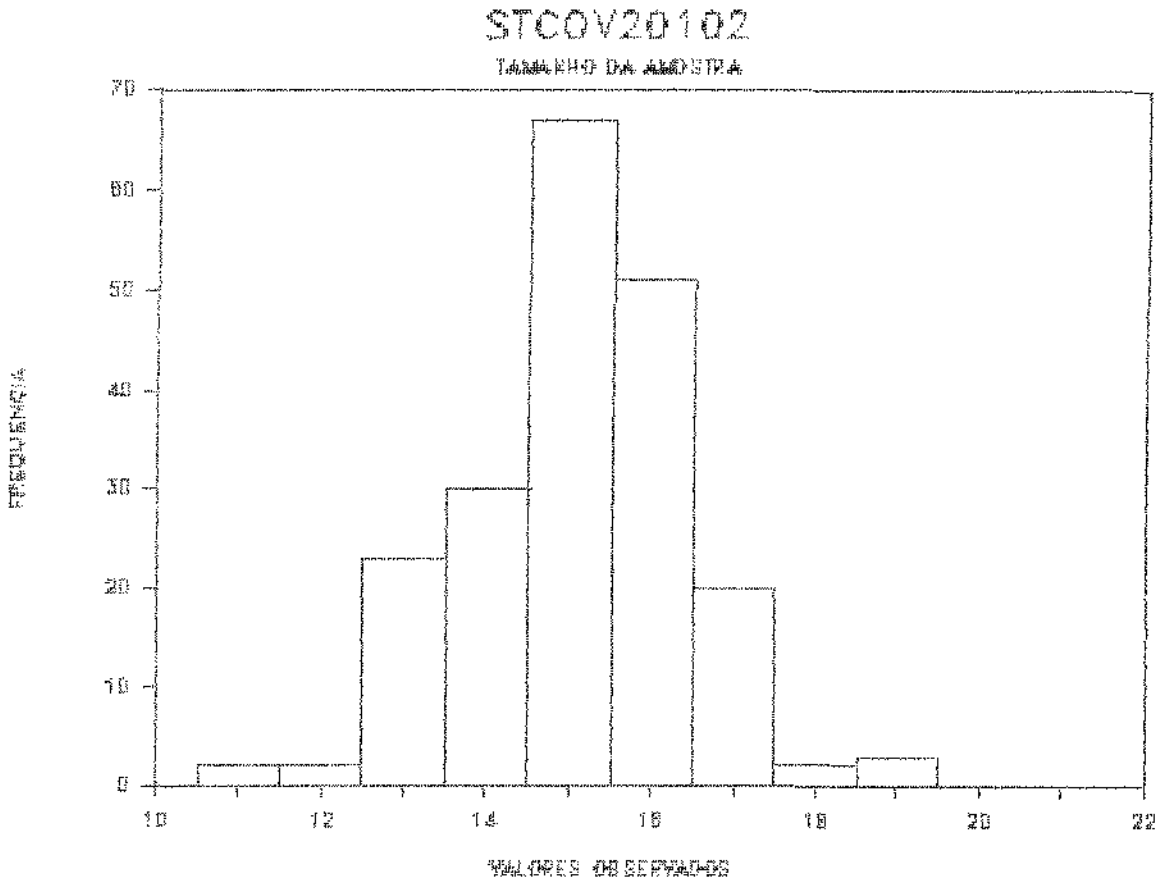

STCOV20102
TAMANHO DA AMOSTRA
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22

GRÁFICO 119

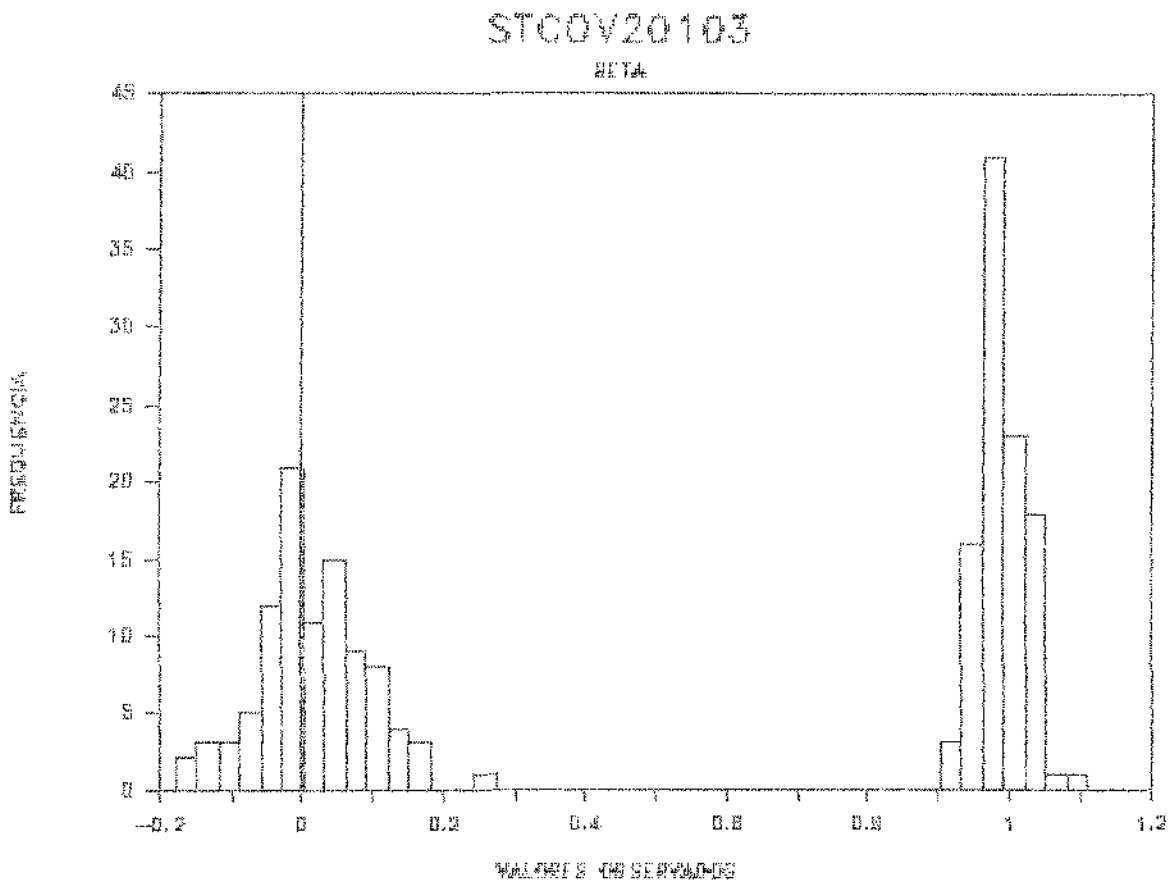


GRÁFICO 120

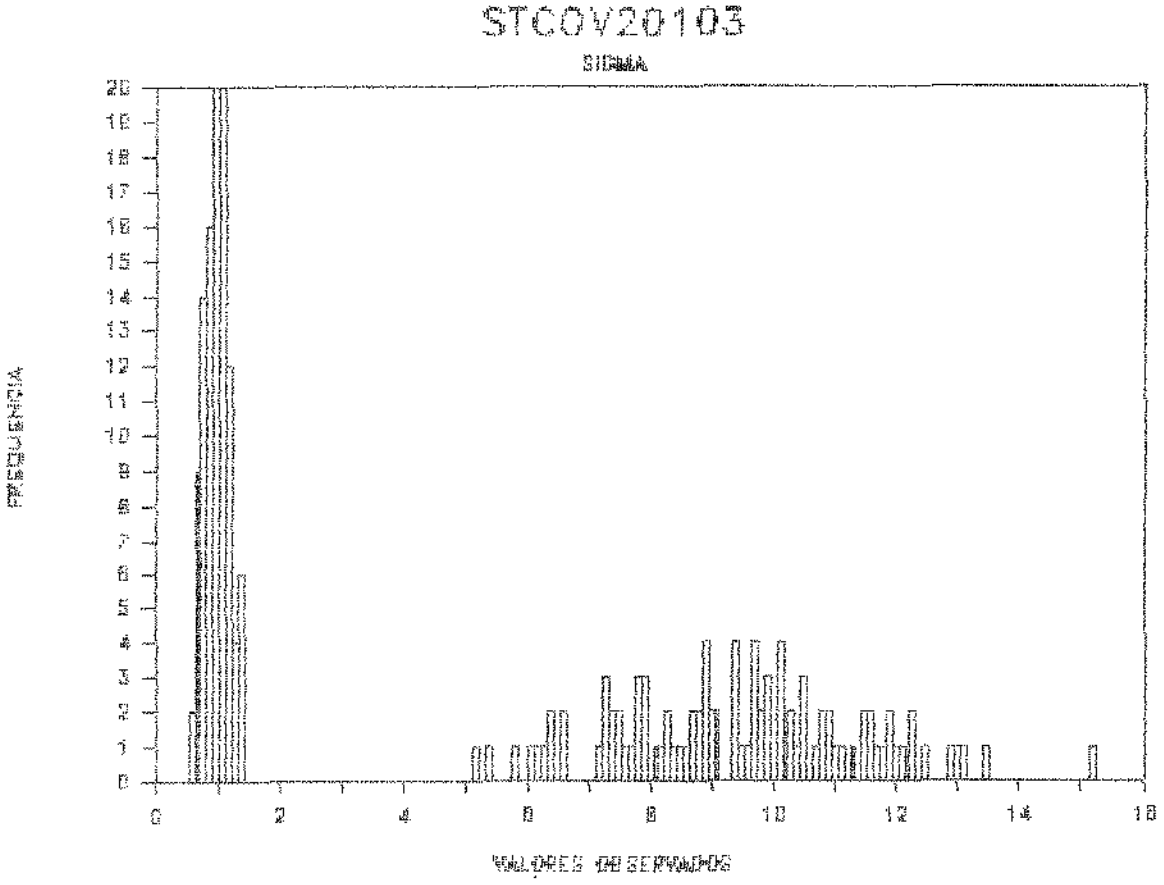


GRÁFICO 121

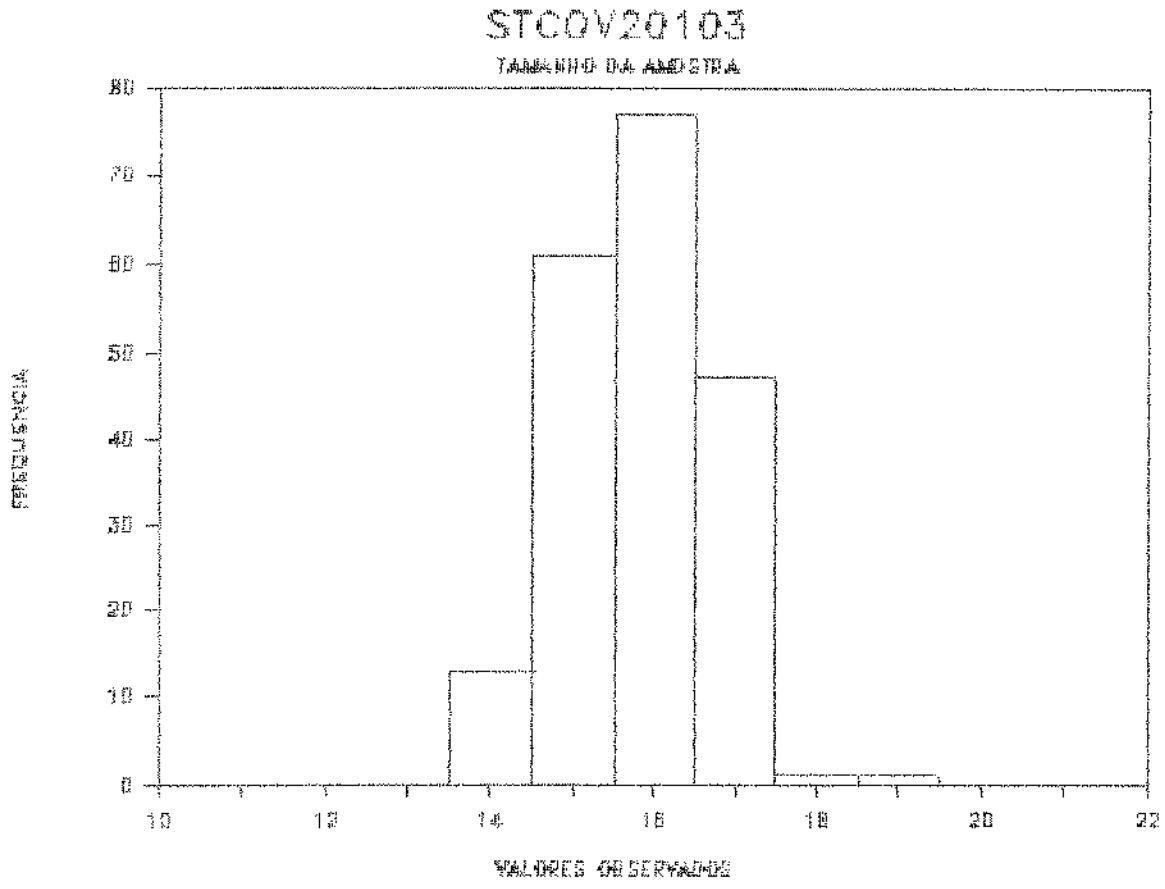

STCOV2103
TAMANHO DA AMOSTRA
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 122

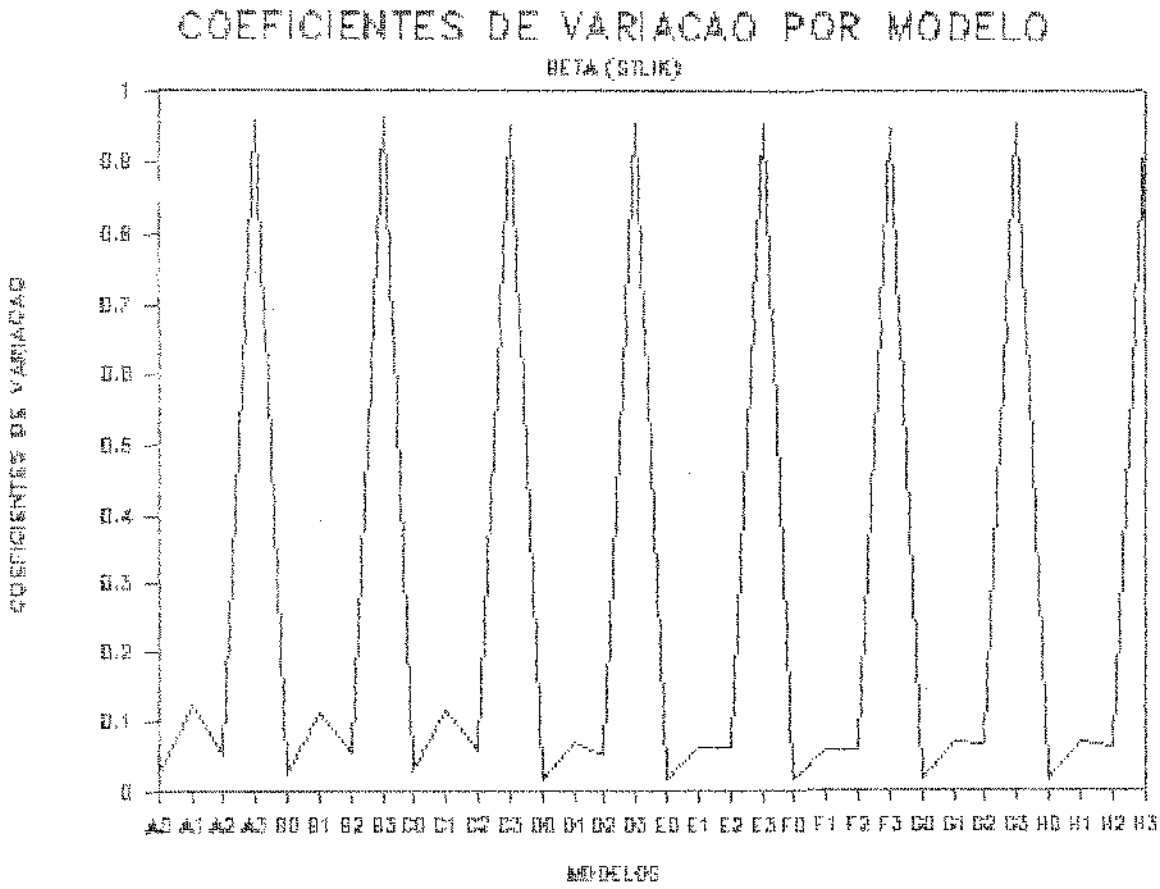

COEFICIENTES DE VARIACAO POR MODELO
BETA (STRIKE)
COEFICIENTES DE VARIACAO
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
1
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 123

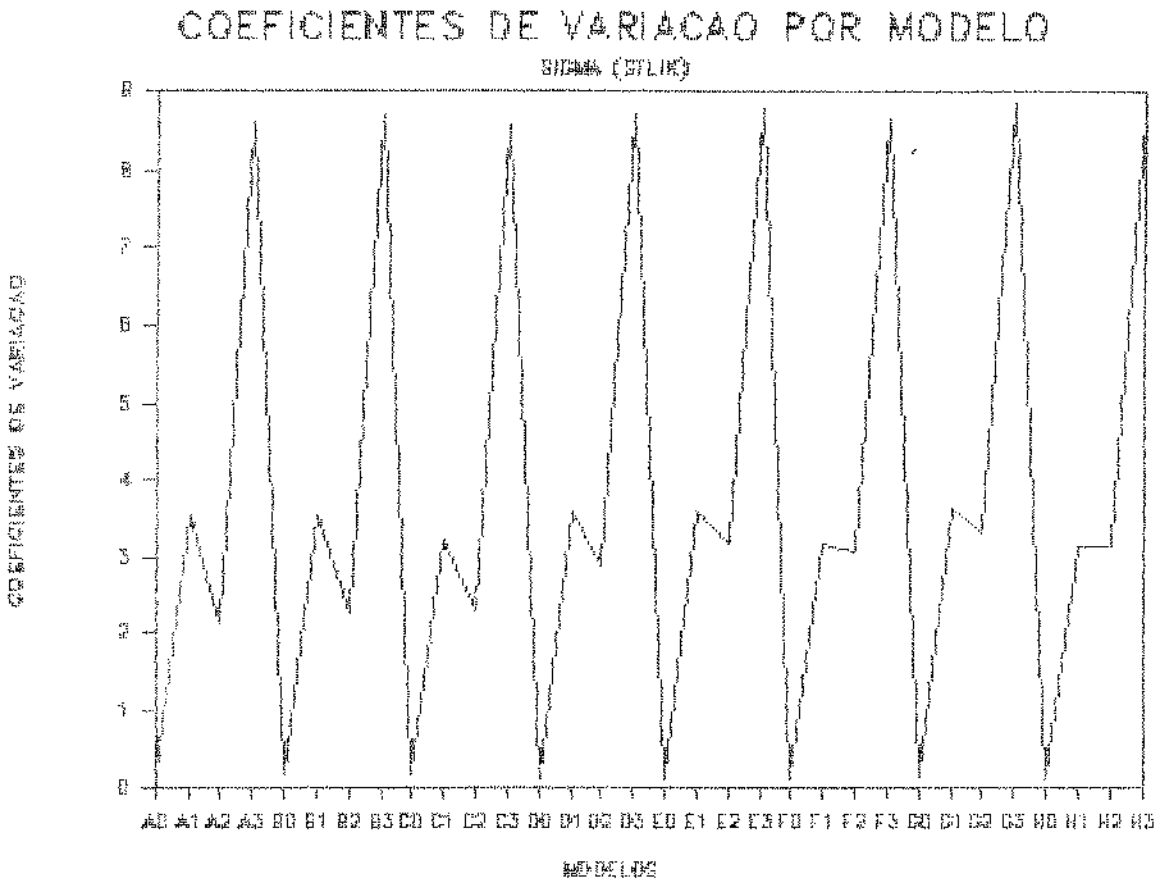

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (STLIN)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 124

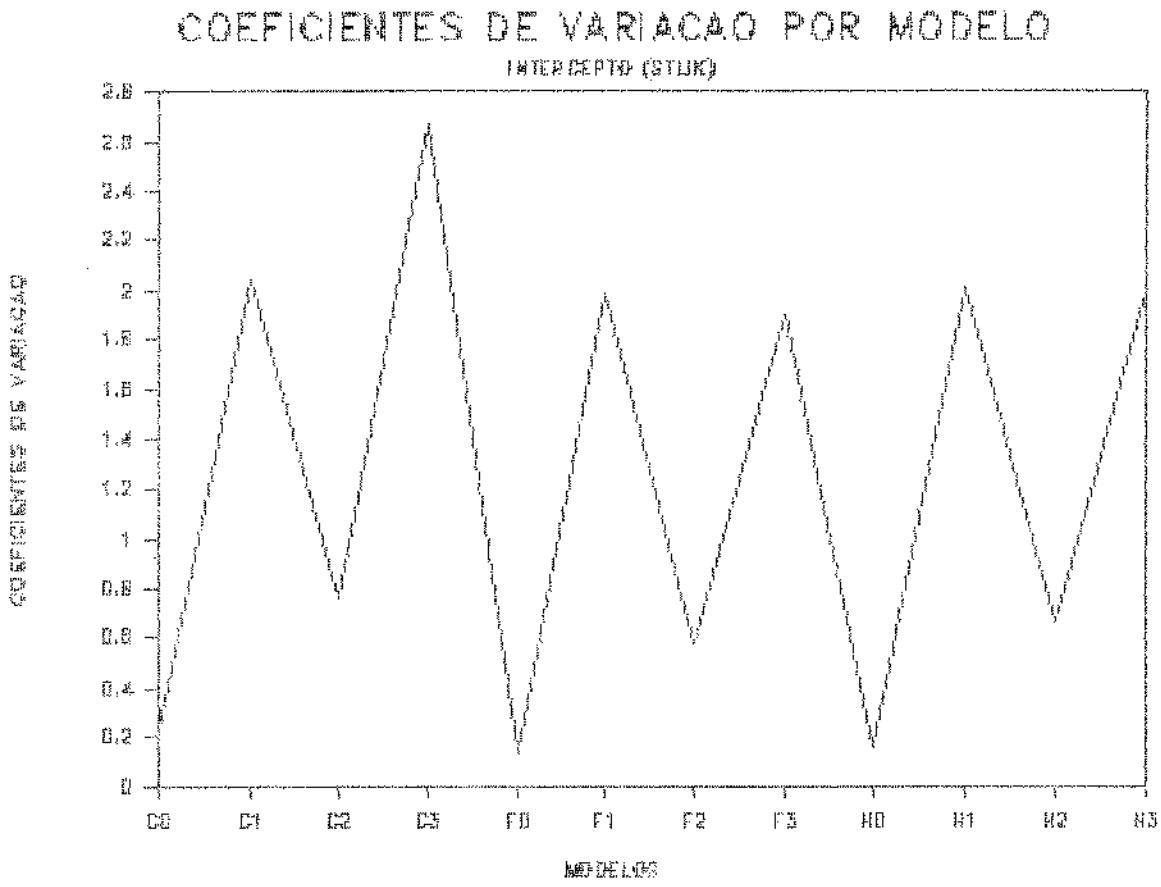

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (STUK)
COEFICIENTES DE VARIACAO
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
C0
C1
C2
C3
F0
F1
F2
F3
H0
H1
H2
H3

GRÁFICO 125

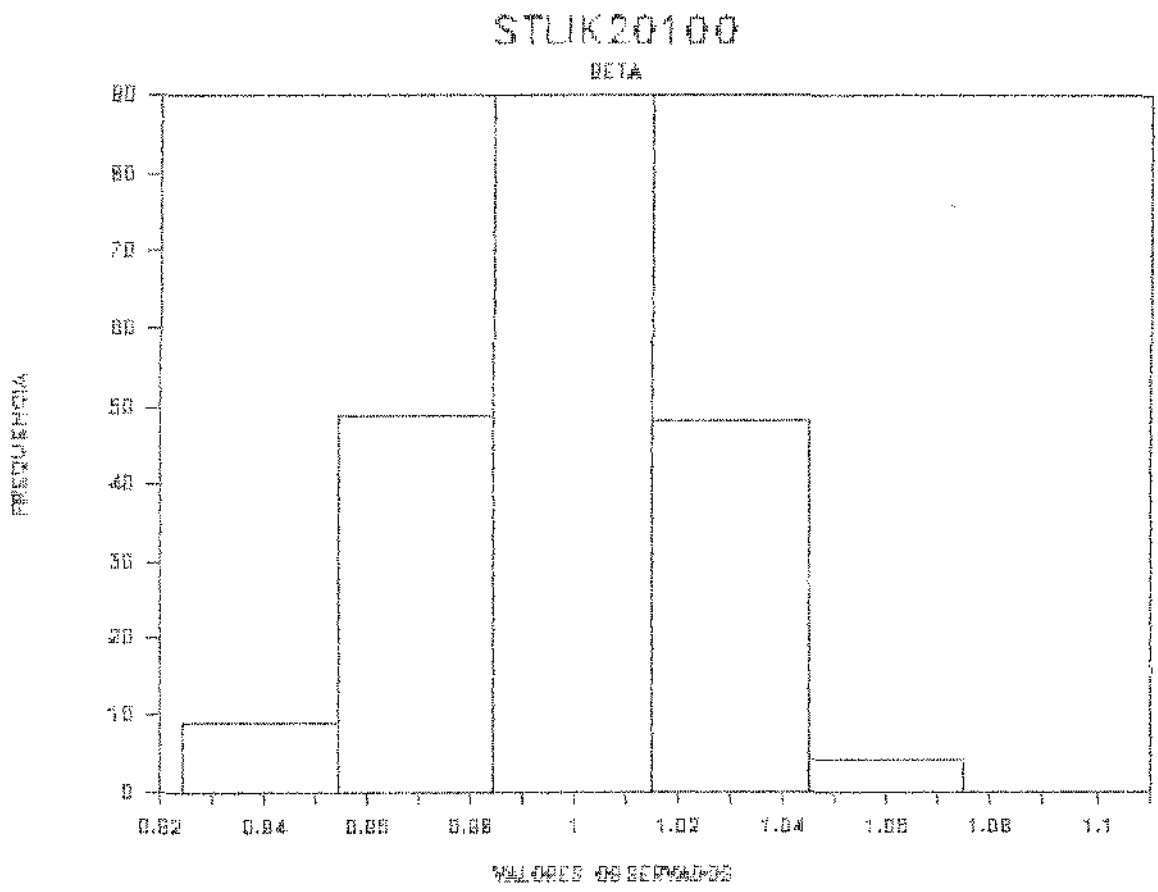

STUIK20100
BETA
80
70
60
50
40
30
20
10
0
0.92
0.94
0.96
0.98
1
1.02
1.04
1.06
1.08
1.1

GRÁFICO 126

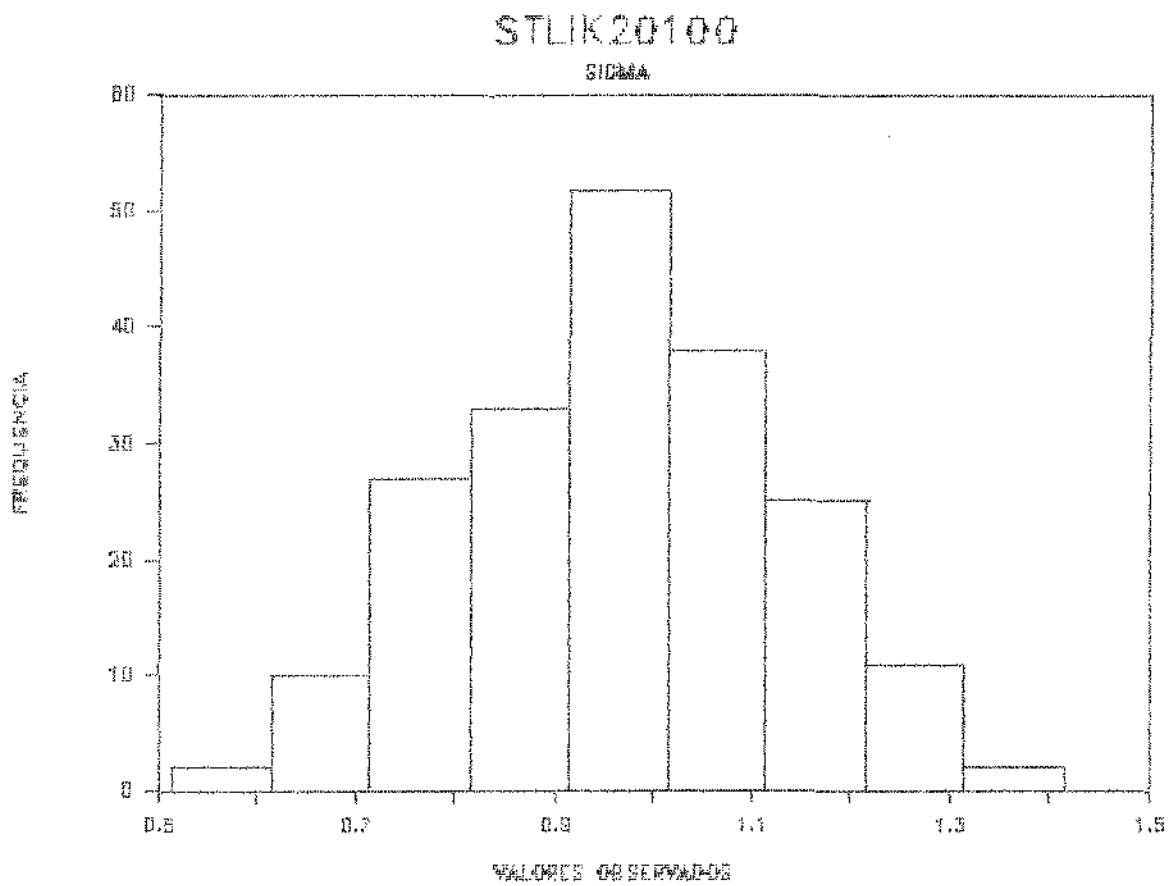

STLIK20100
SIGMA
FREQUÊNCIA
60
50
40
30
20
10
0
0.6
0.7
0.8
1.1
1.3
1.5
VALORES OBSERVADOS

GRÁFICO 127

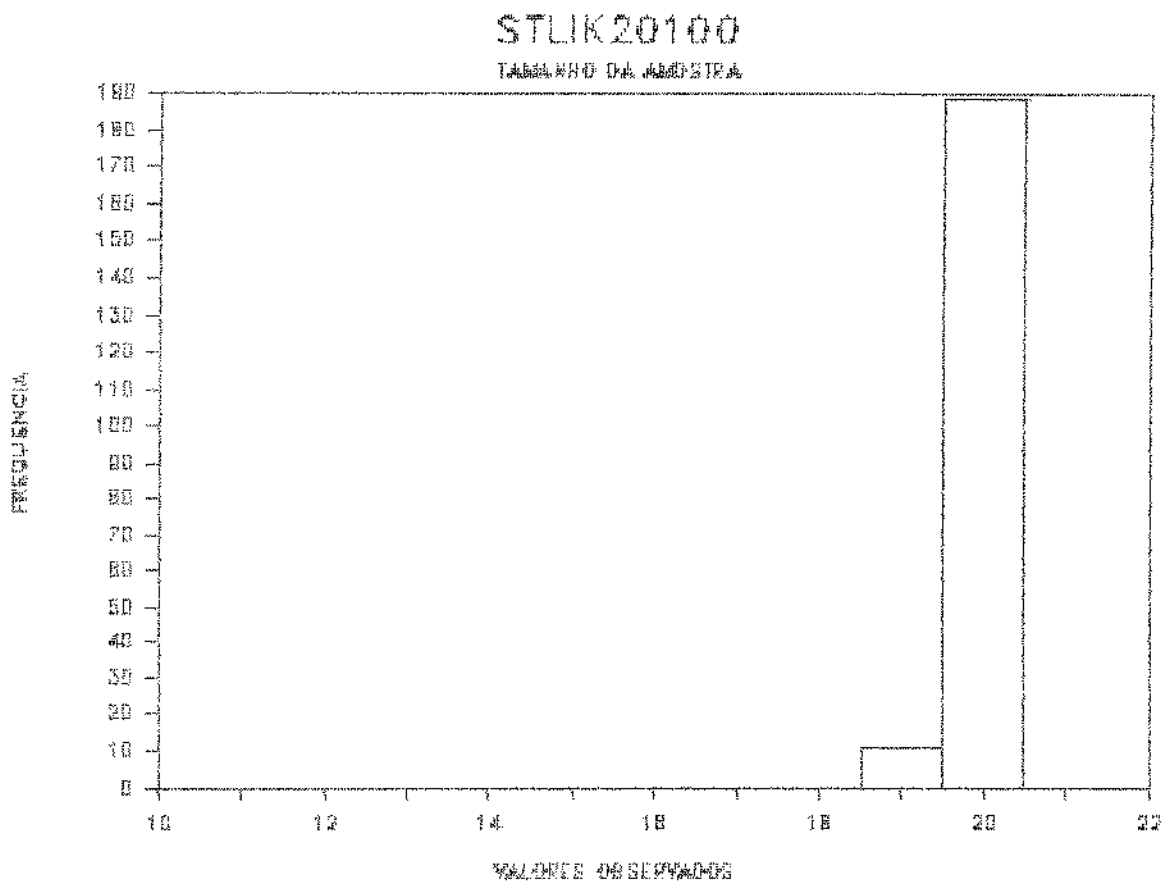

STLIK20100
TAMANHO DA AMOSTRA
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 128

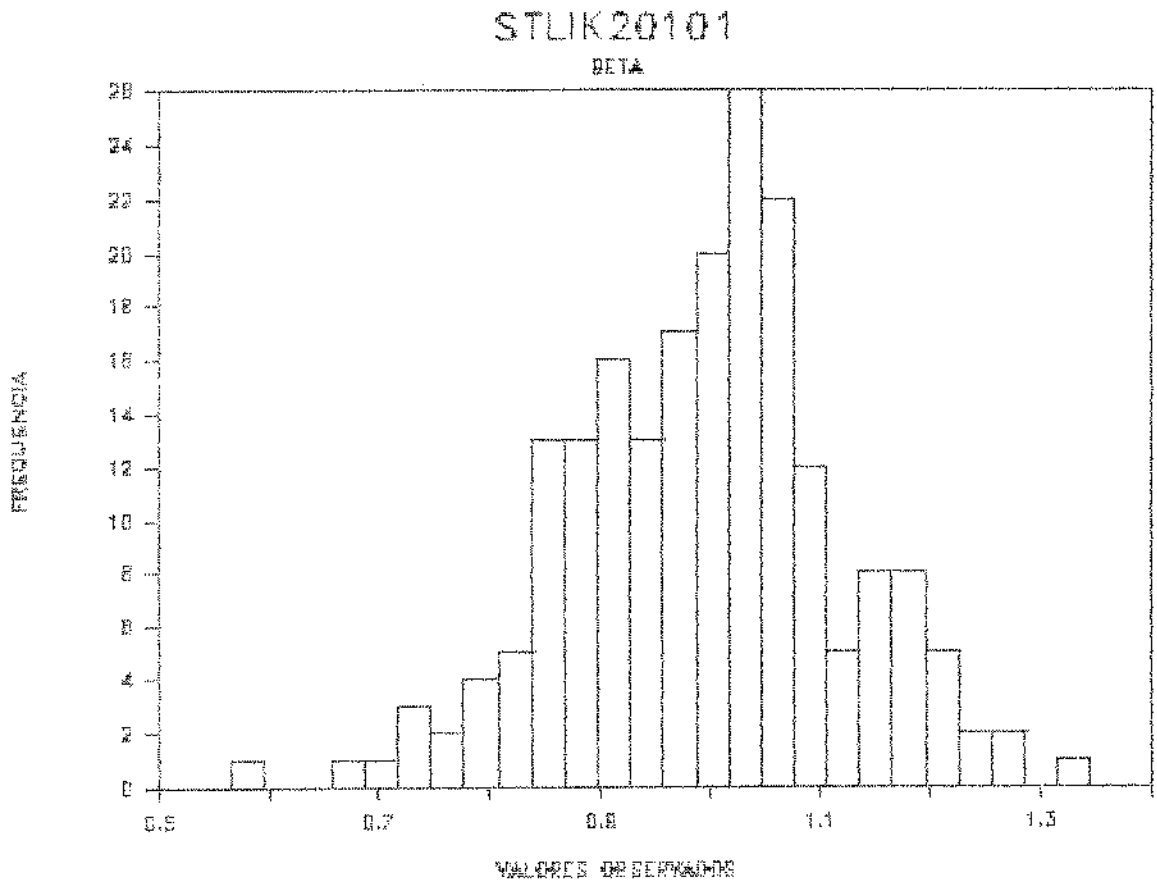

STLIK20101
BETA
FREQUÊNCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
0.6
0.7
0.8
1.1
1.3
VALORES OBSERVADOS

GRÁFICO 129

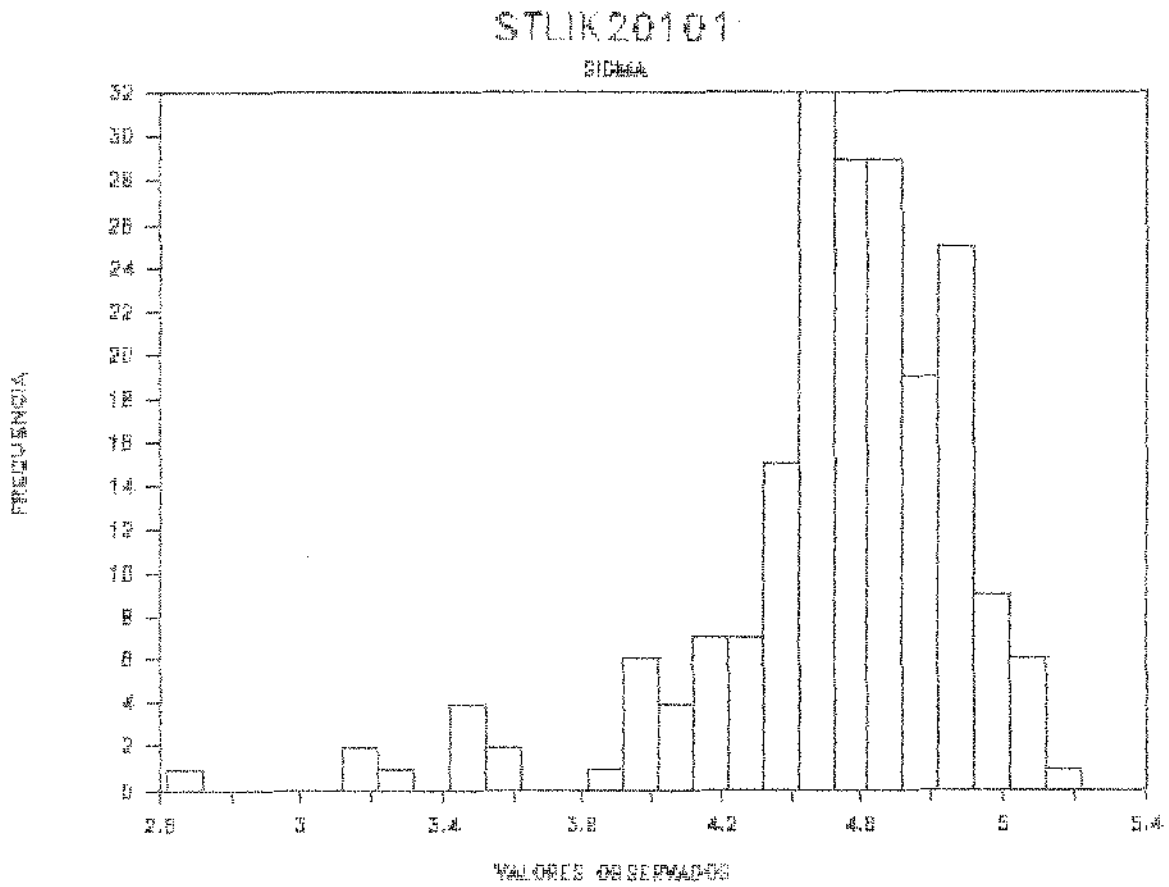

STLIK20101
FRECUENCIA
2.8
3
3.4
3.8
4.2
4.6
5
5.4
VALORES OBSERVADOS

GRÁFICO 130

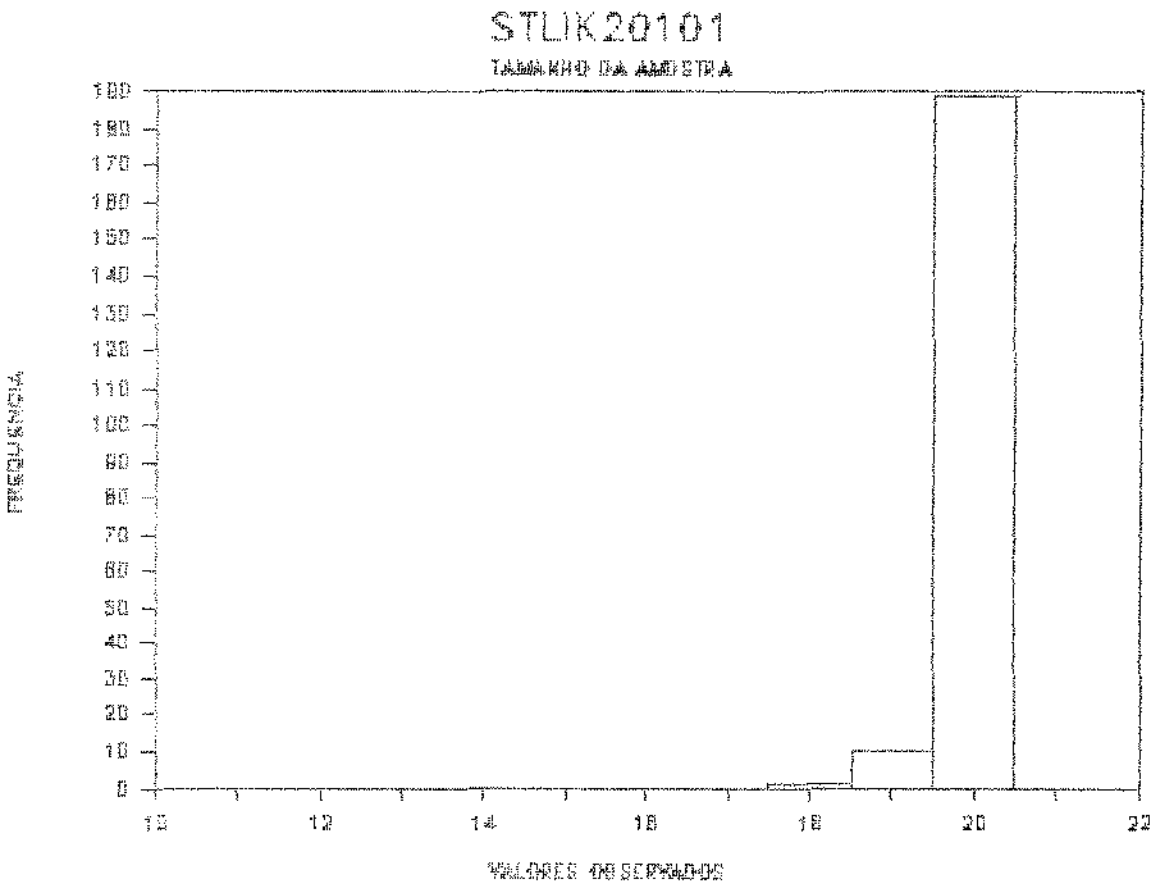

STUK20101
TAMANHO DA AMOSTRA
FREQUÊNCIA
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 131

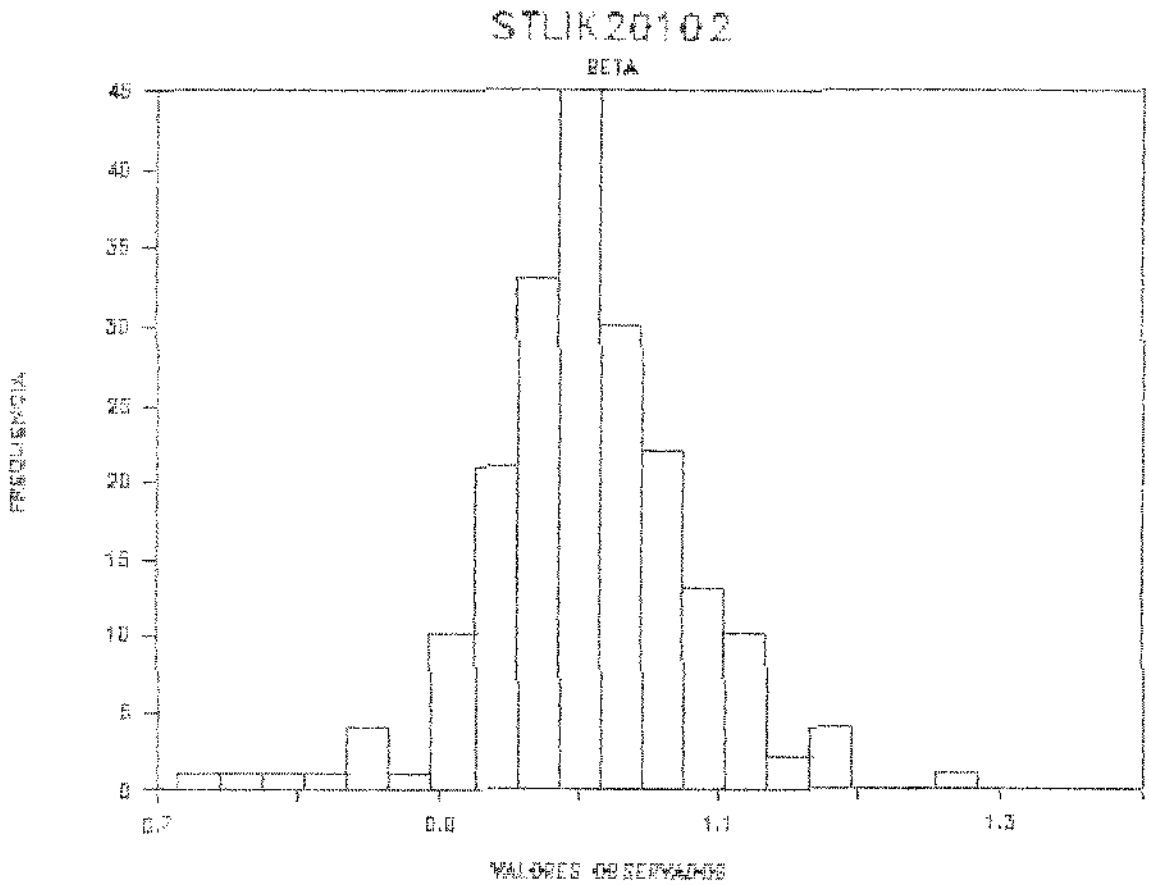

STLIK2012
BETA
FREQUÊNCIA
0
5
10
15
20
25
30
35
40
45
-0.5
0.0
0.5
1.0
1.5
VALORES OBSERVADOS

GRÁFICO 132

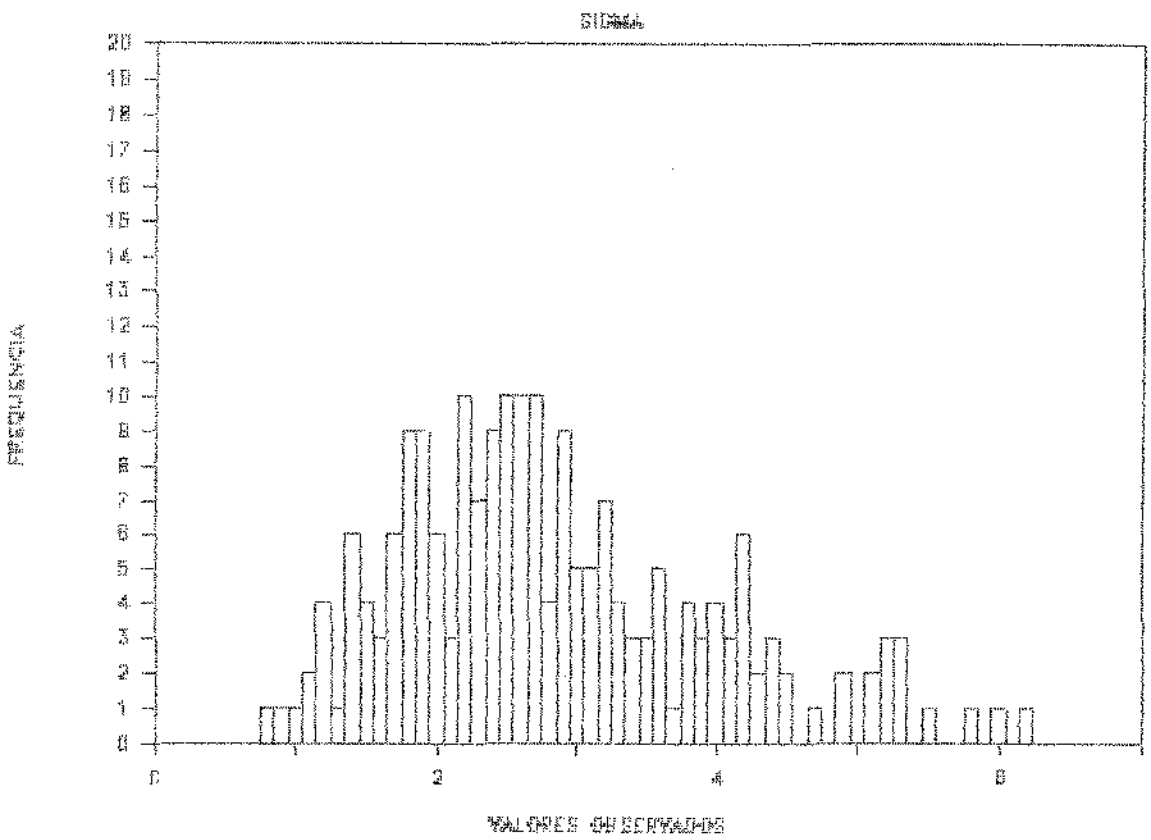


GRÁFICO 133

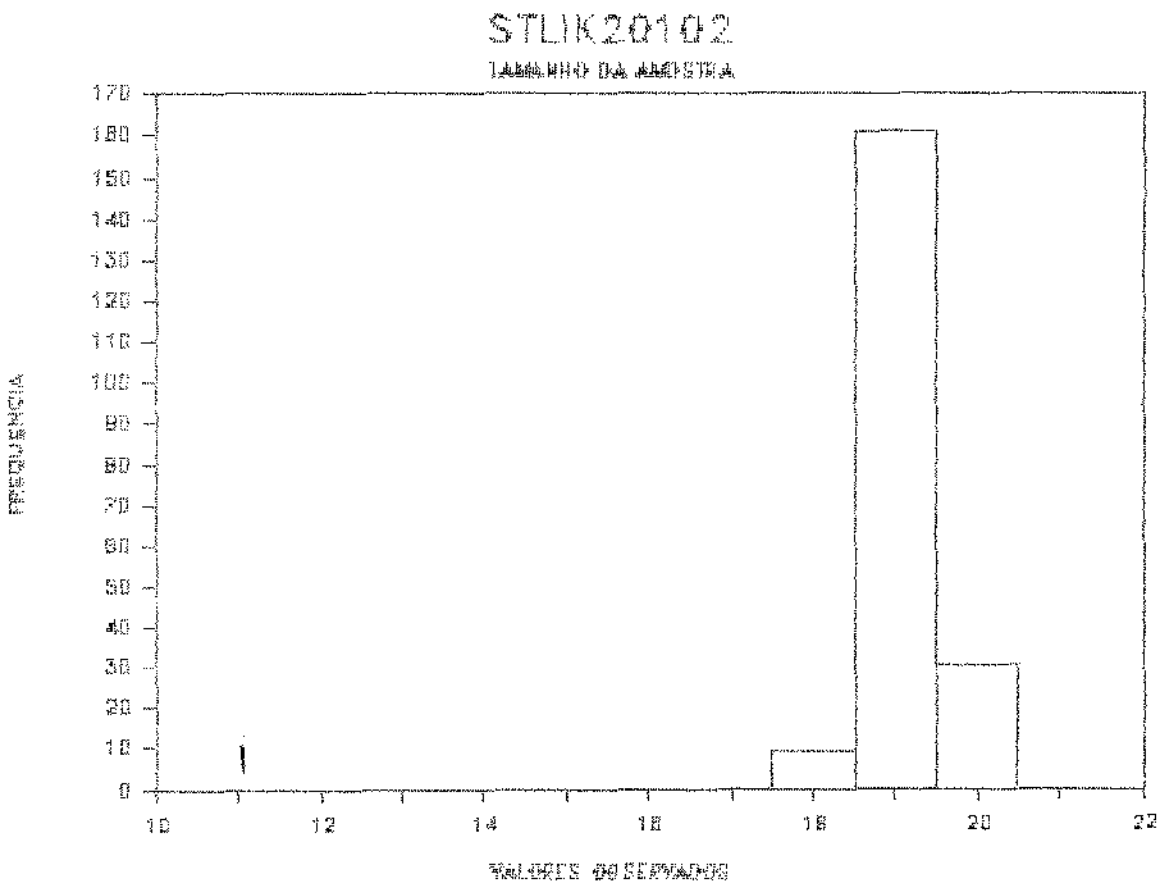

STLIK2102
FREQUÊNCIA
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22

GRÁFICO 134

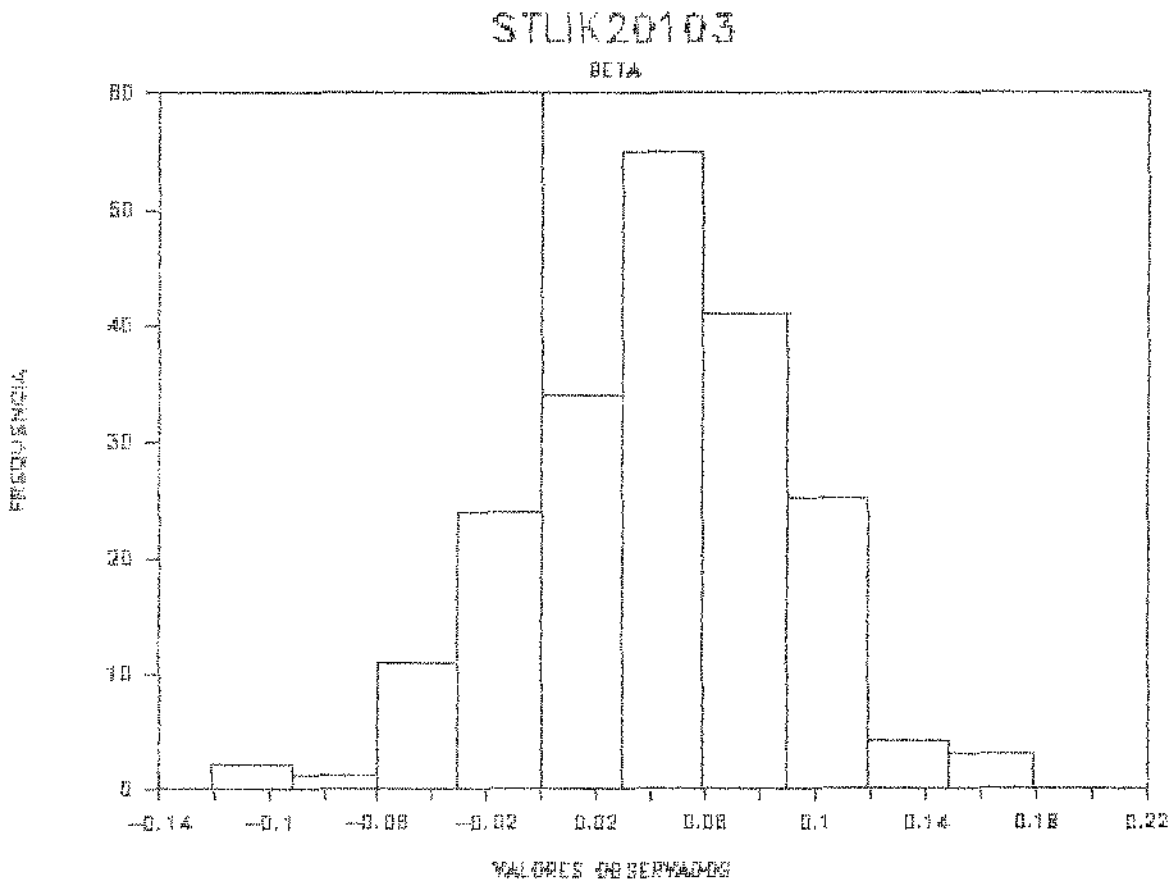

STUK20103
BETA
80
60
40
30
20
10
0
-0.14
-0.1
-0.06
-0.02
0.02
0.06
0.1
0.14
0.18
0.22
VALORES OBSERVADOS

GRÁFICO 135

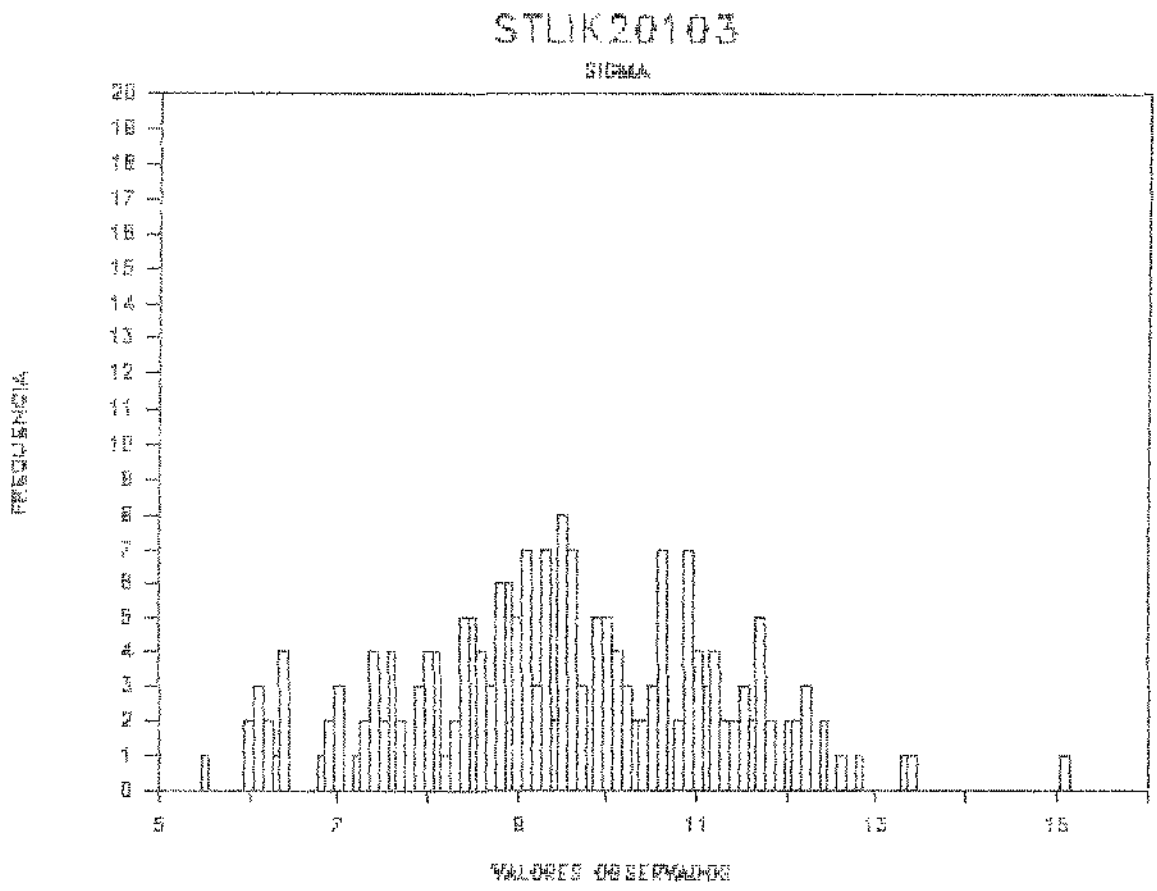

STUK20103
SIGMA
FRECUENCIA
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
5
7
9
11
13
15
VALORES OBSERVADOS

GRÁFICO 136

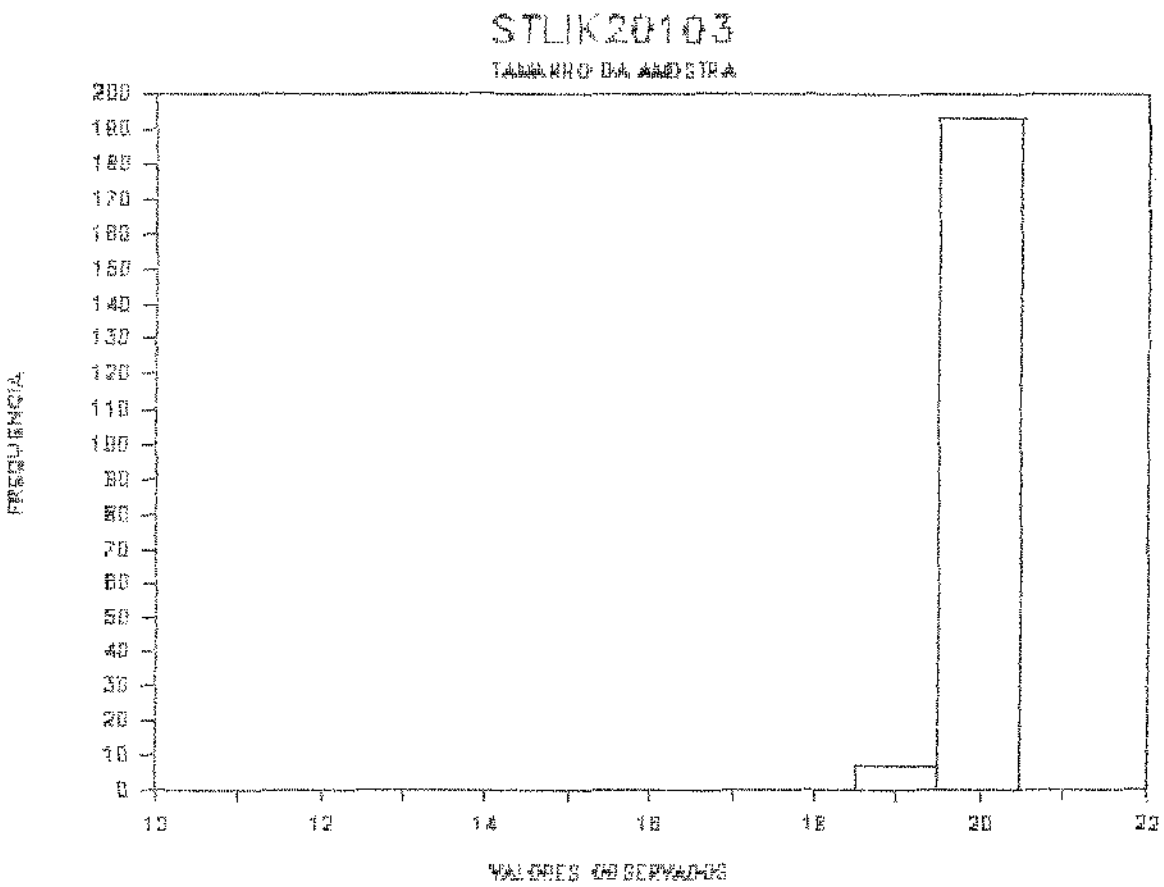

STLIK20103
TAMANHO DA AMOSTRA
200
190
180
170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22

GRÁFICO 137

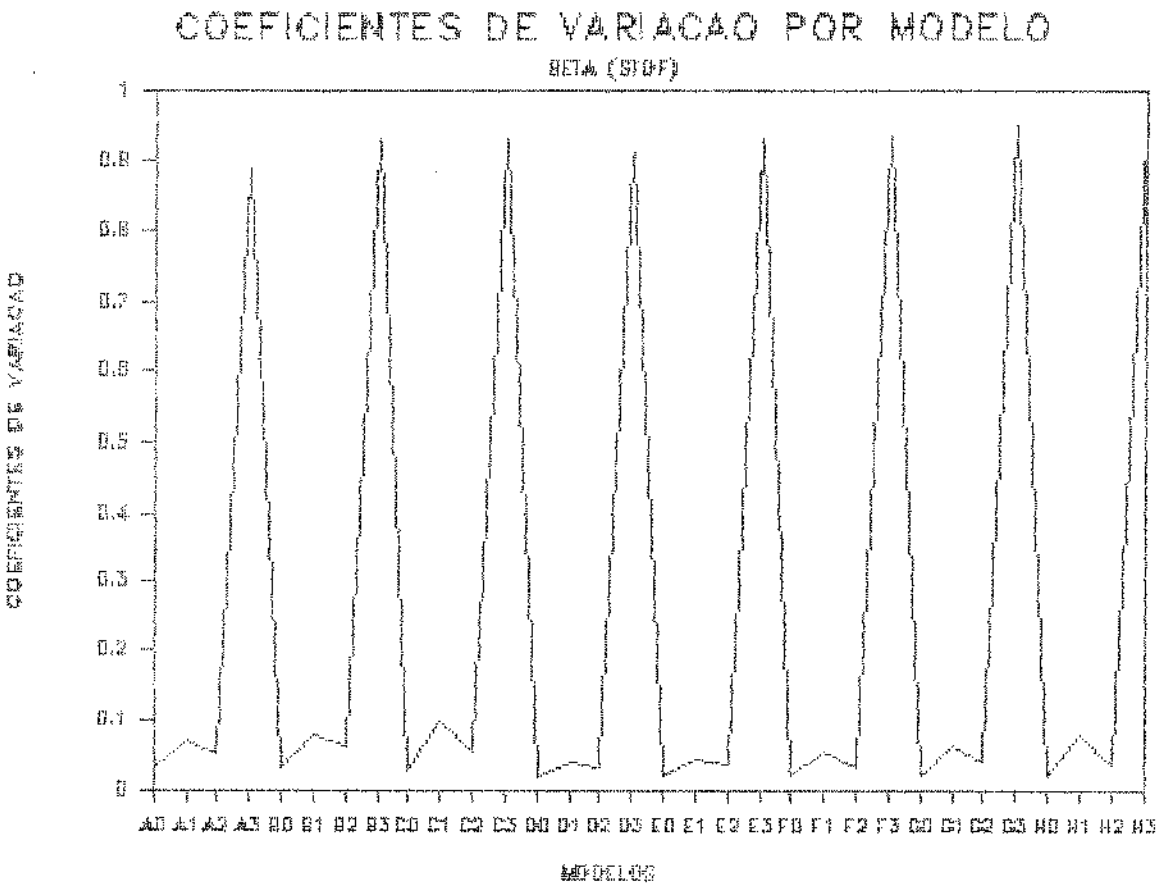

COEFICIENTES DE VARIACAO POR MODELO
BETA (STOP)
COEFICIENTES DE VARIACAO
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 138

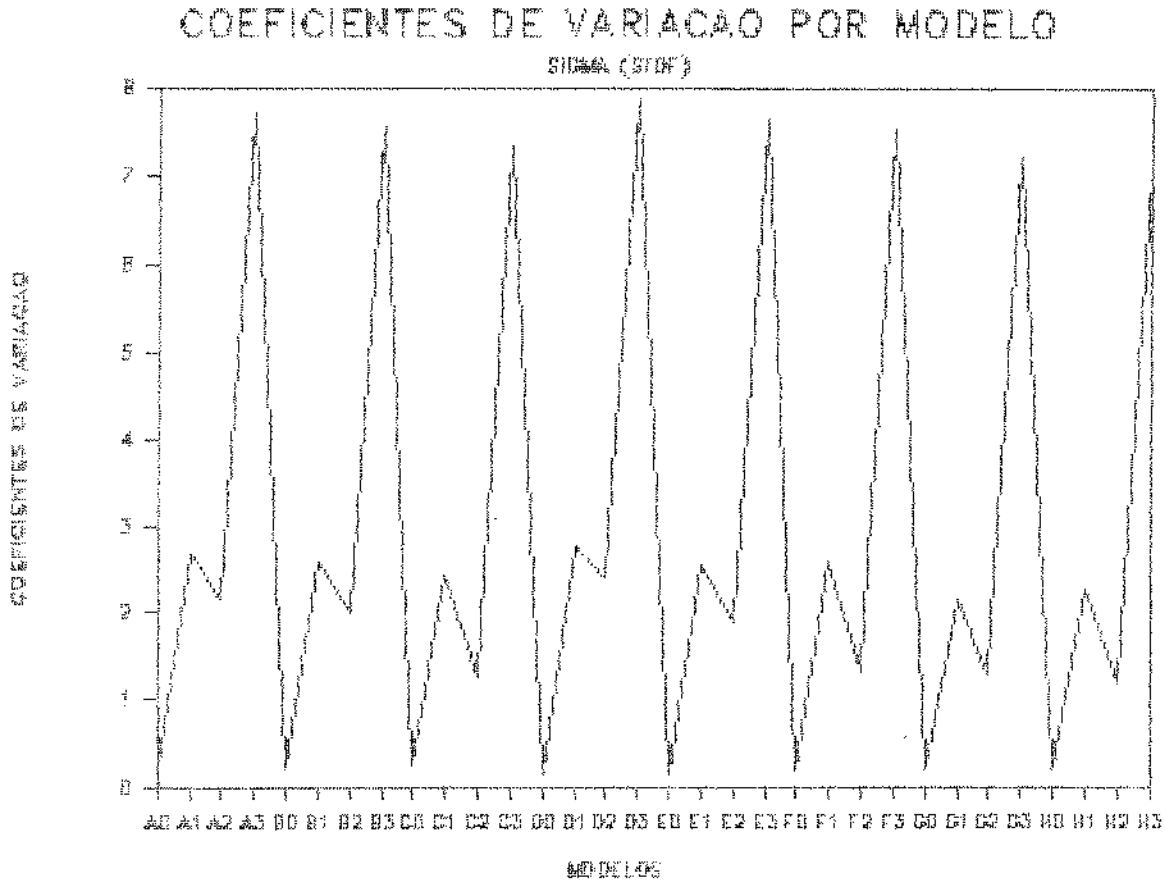

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (STEP)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 139

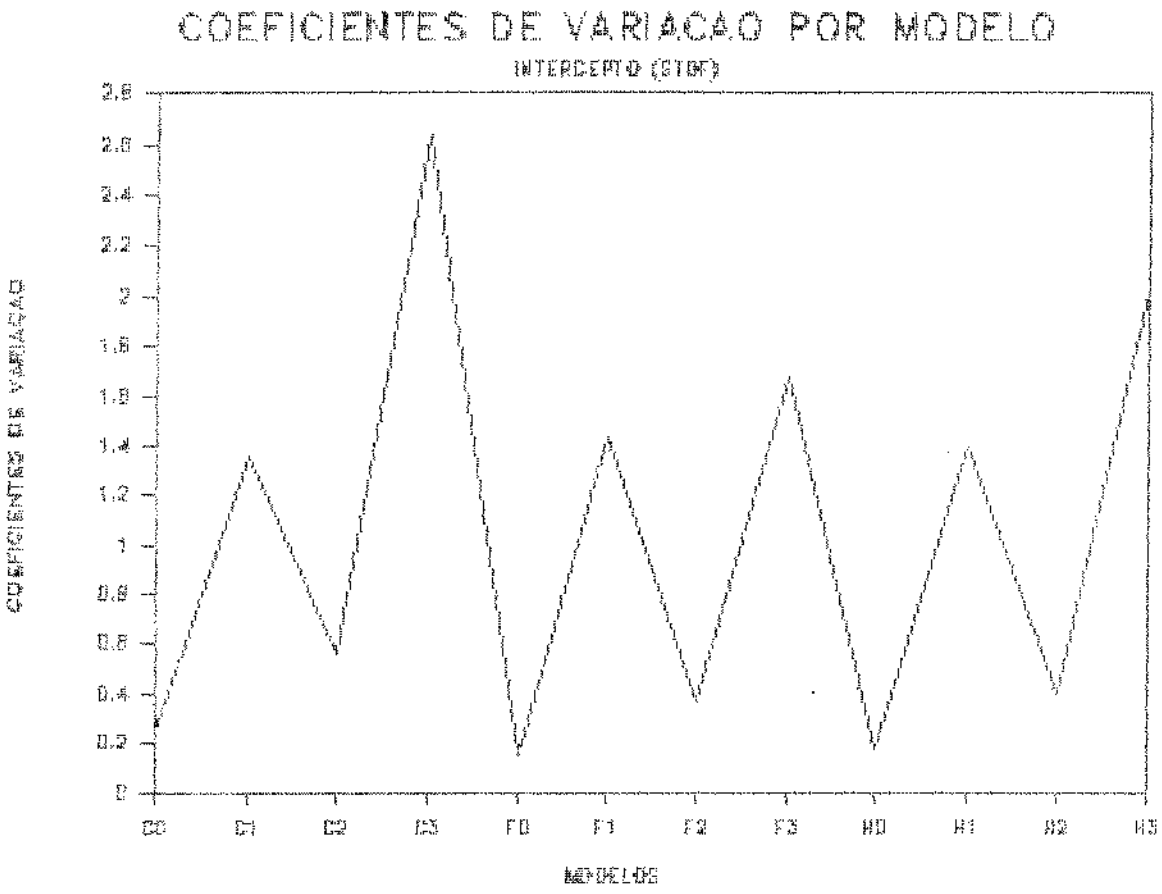

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (STDF)
COEFICIENTES DE VARIACAO
MODELOS
C0
C1
C2
C3
F0
F1
F2
F3
H0
H1
H2
H3

GRÁFICO 140

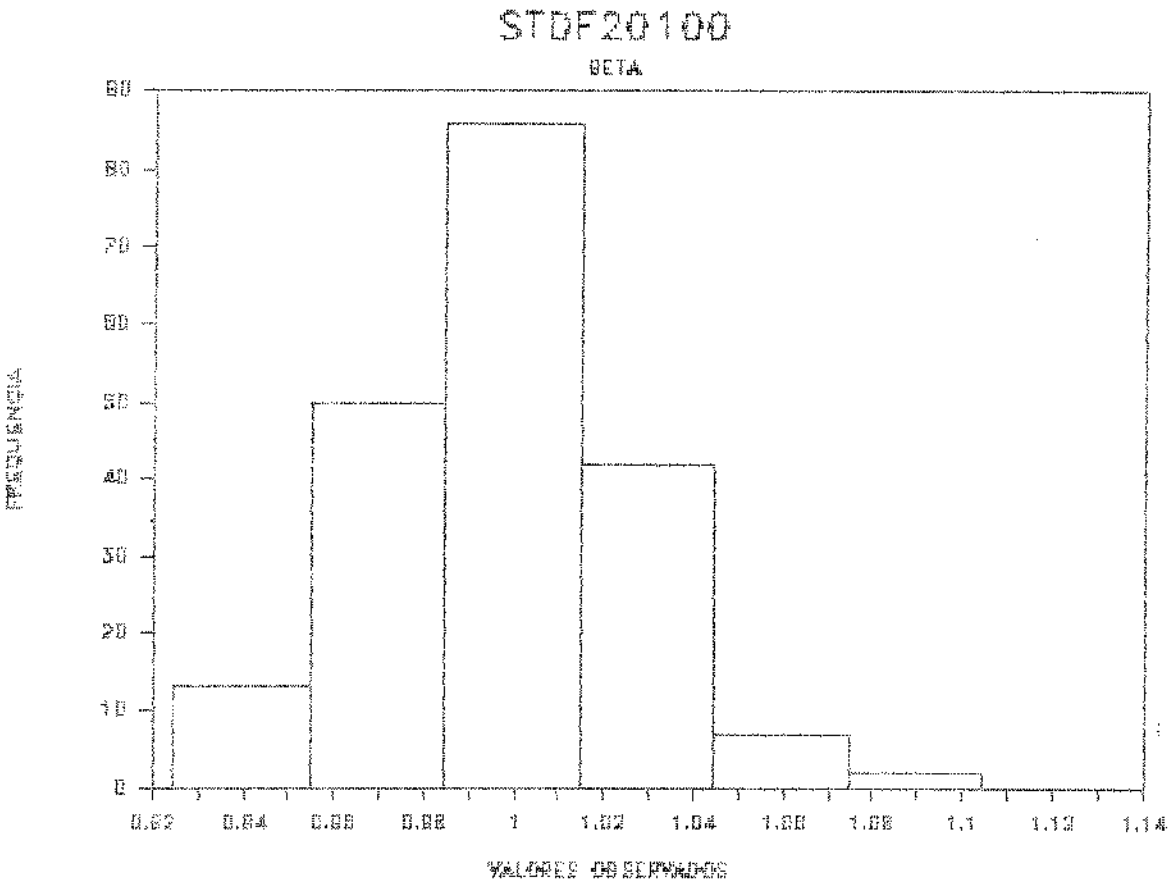

STDF2010D
BETA
FREQUENCIA
80
70
60
50
40
30
20
10
0
0.82
0.84
0.86
0.88
1
1.02
1.04
1.06
1.08
1.1
1.12
1.14

GRÁFICO 141

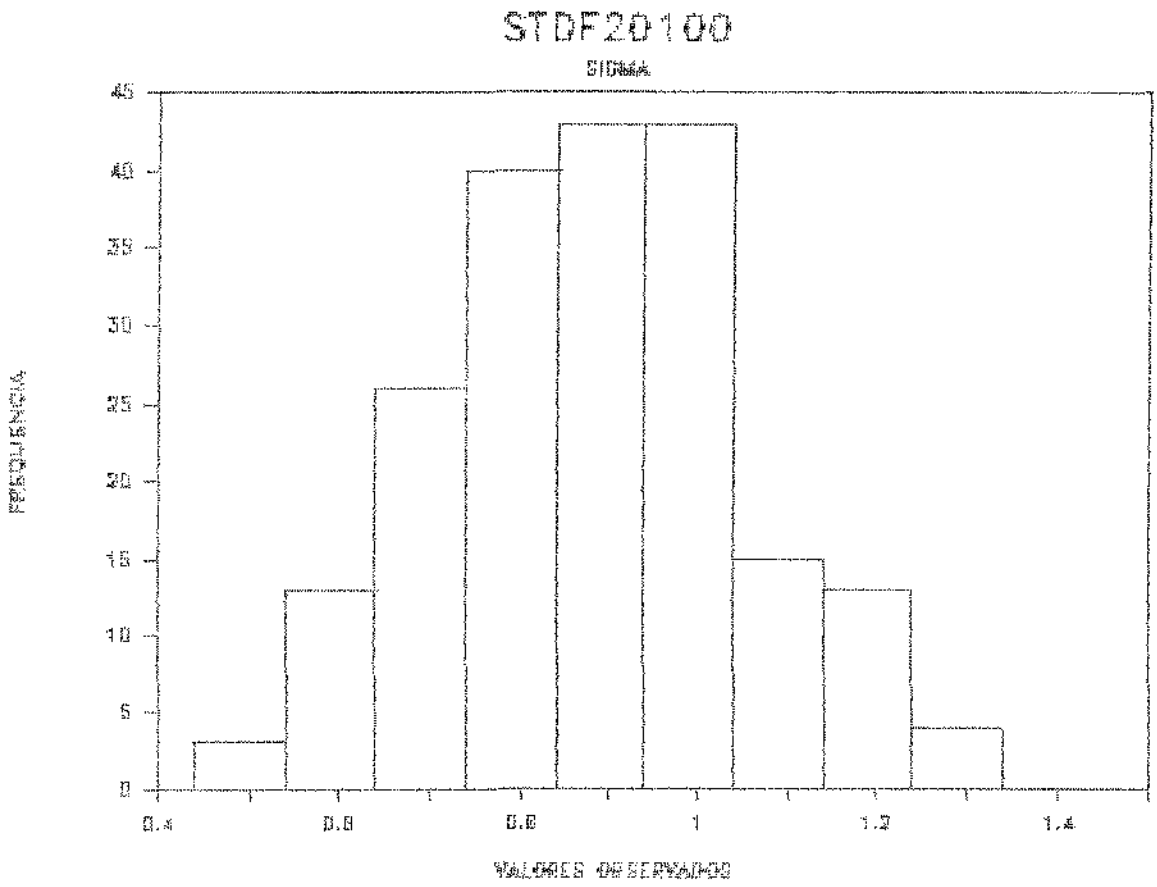


GRÁFICO 142

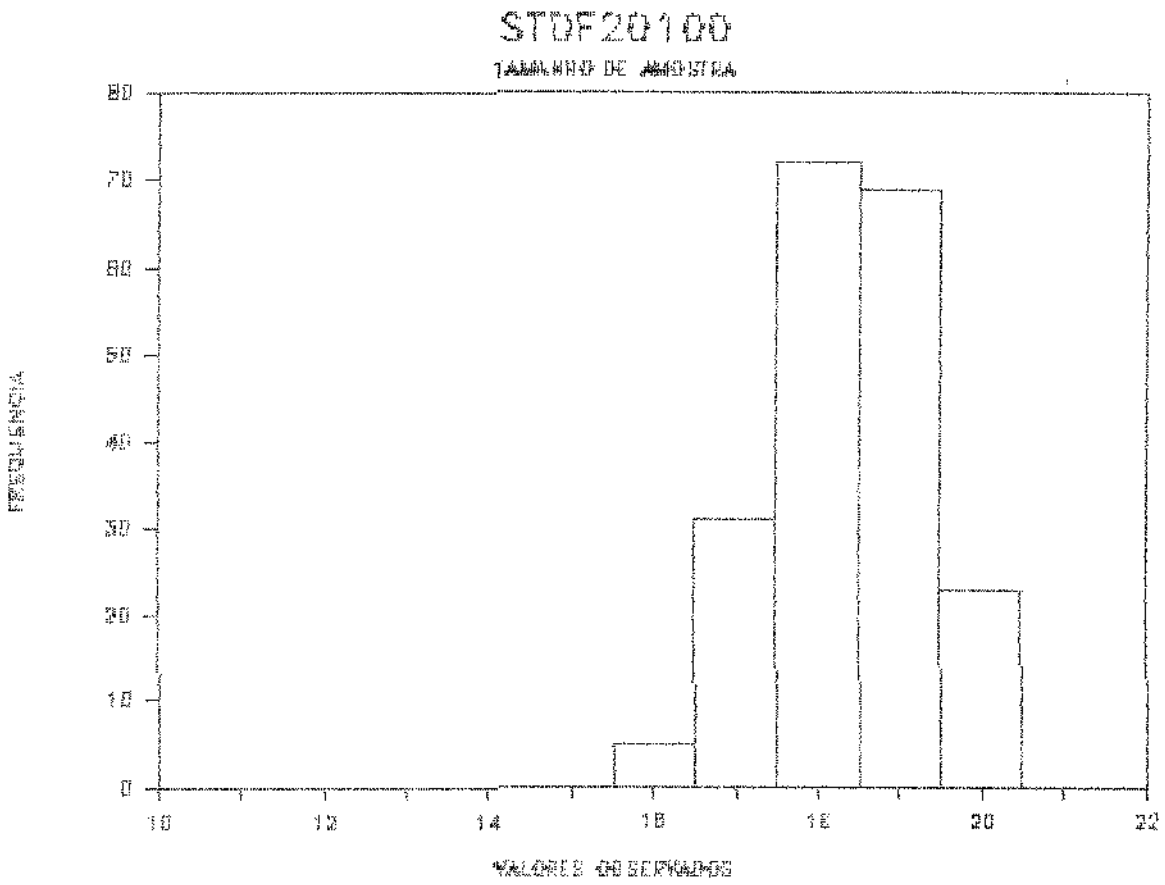


GRÁFICO 143

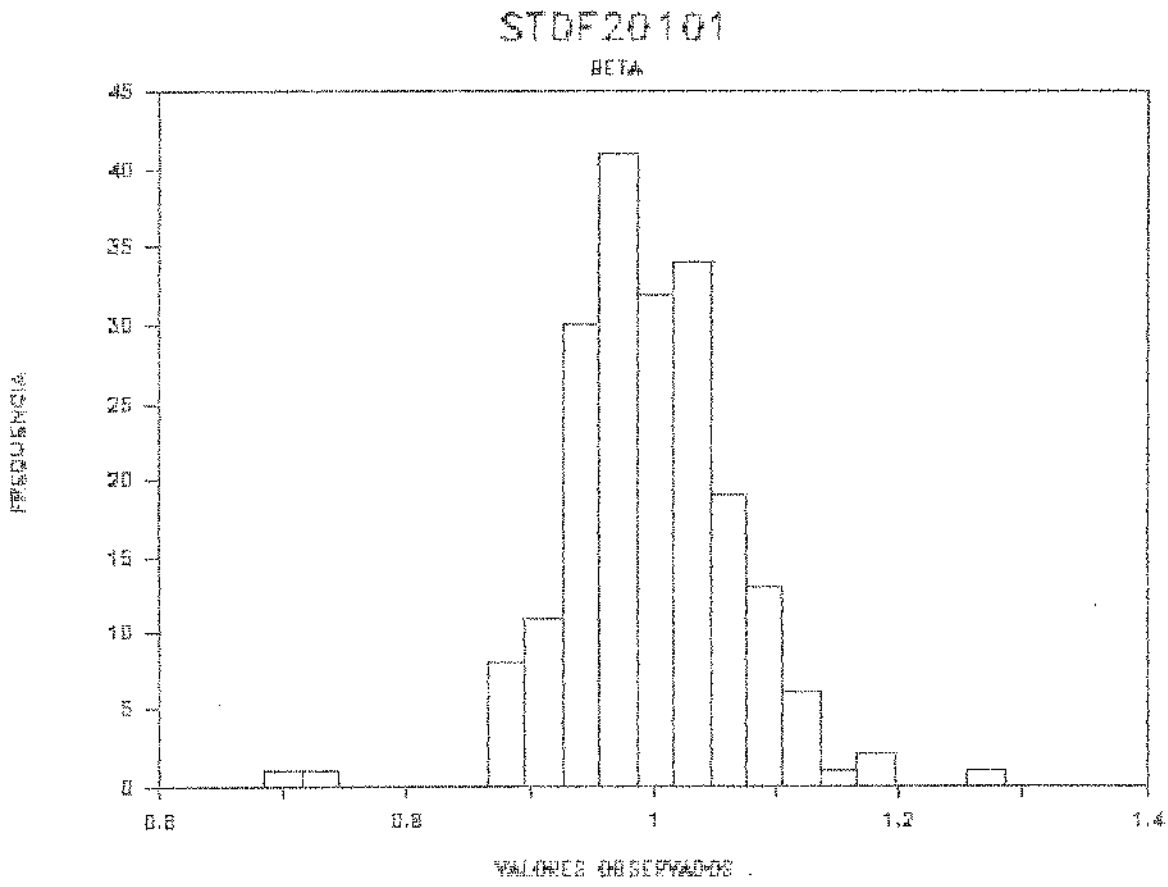

STDF20101
BETA
FREQUENCIA
0
5
10
15
20
25
30
35
40
45
0.6
0.8
1
1.2
1.4
VALORES OBSERVADOS

GRÁFICO 144

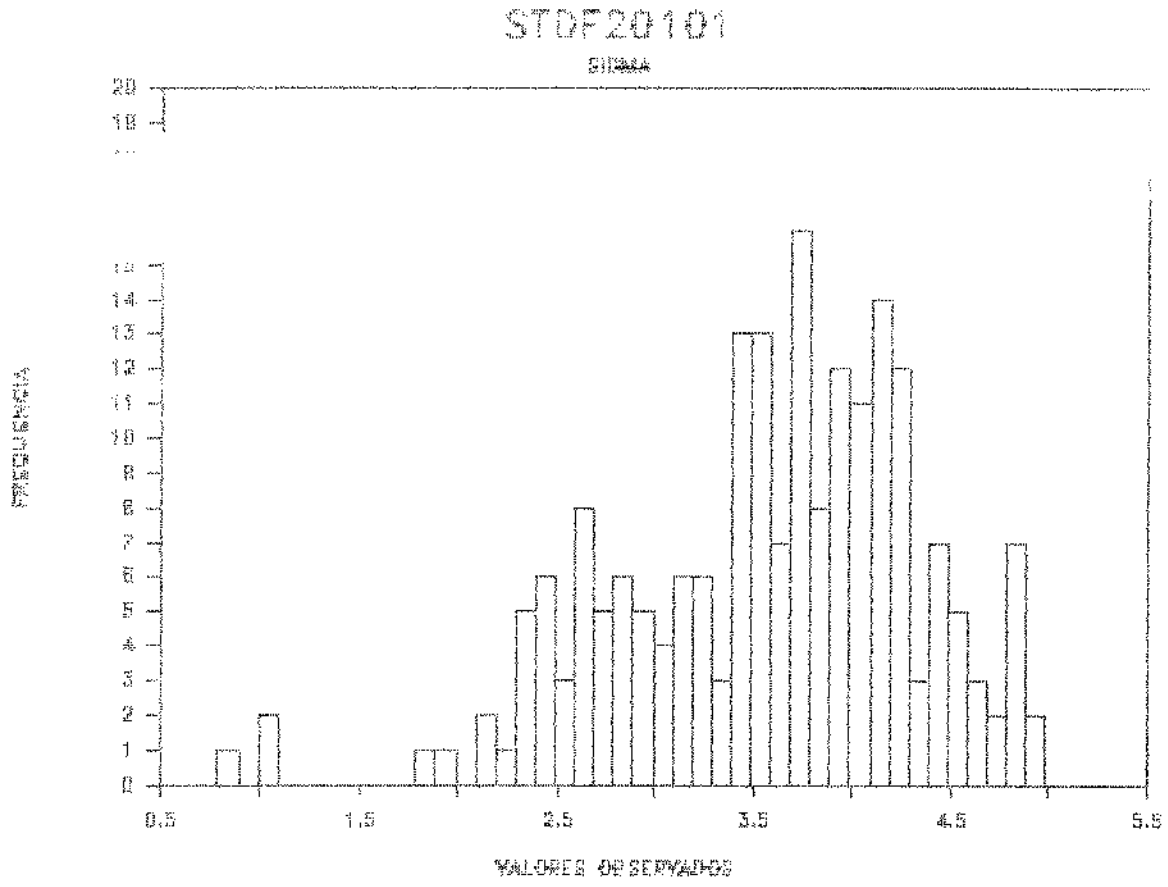


GRÁFICO 145

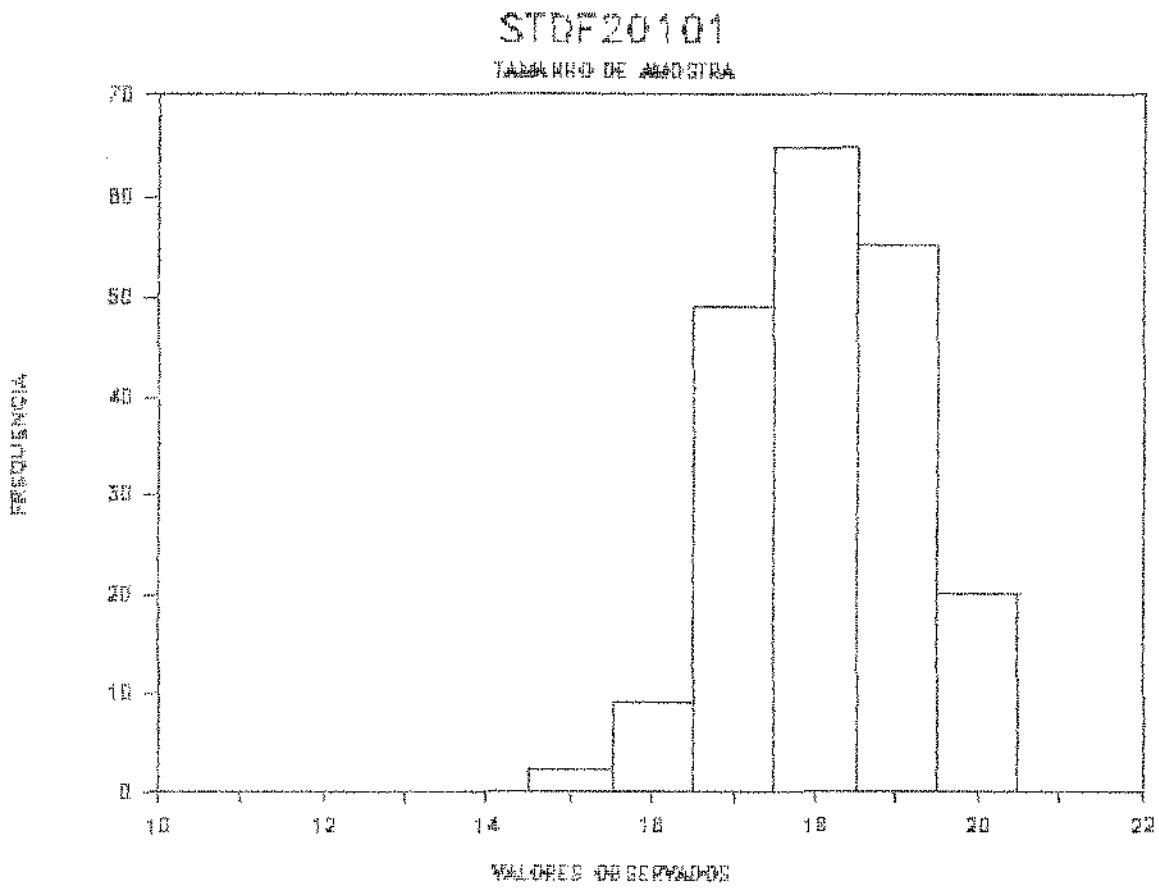

STDF20101
TAMANHO DE AMOSTRA
FREQUÊNCIA
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 146

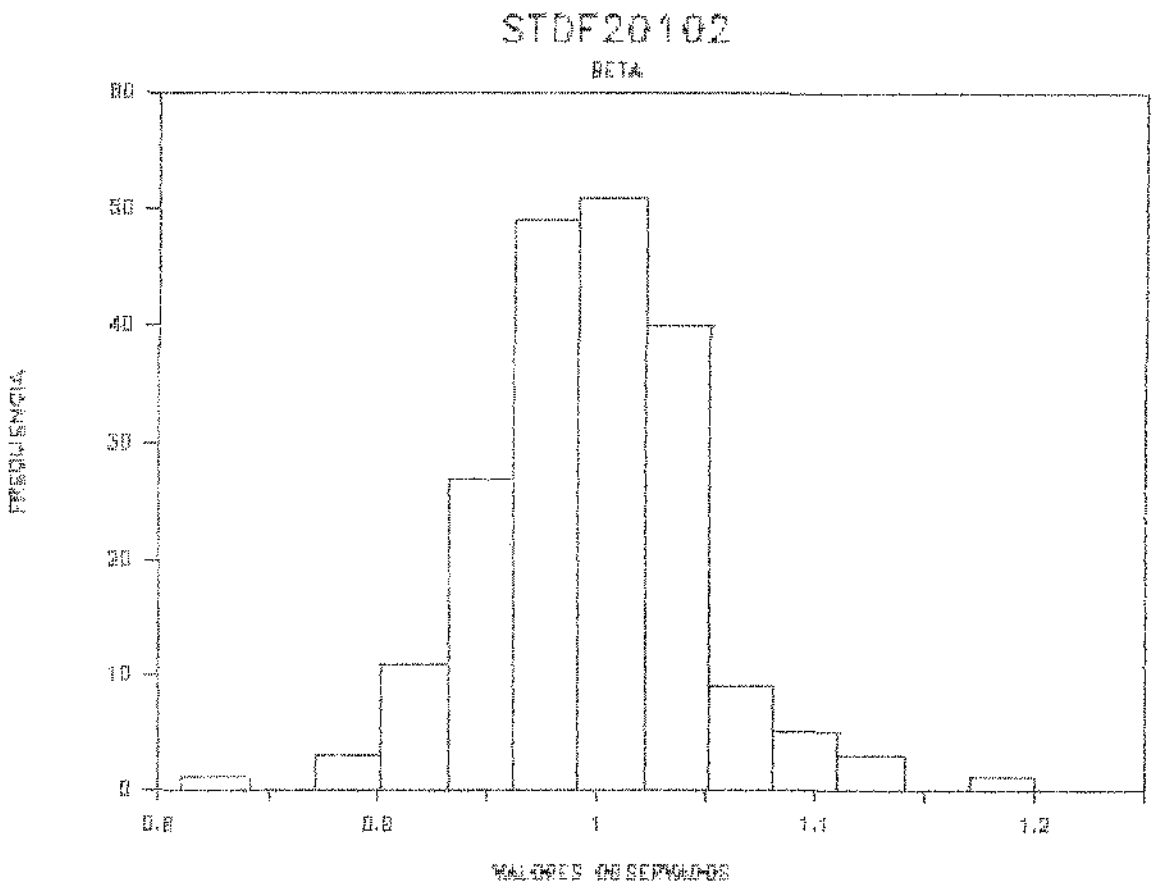

STDF20102
BETA
60
50
40
30
20
10
0
0.8
0.9
1
1.1
1.2
FRECUENCIA
VALORES OBSERVADOS

GRÁFICO 147

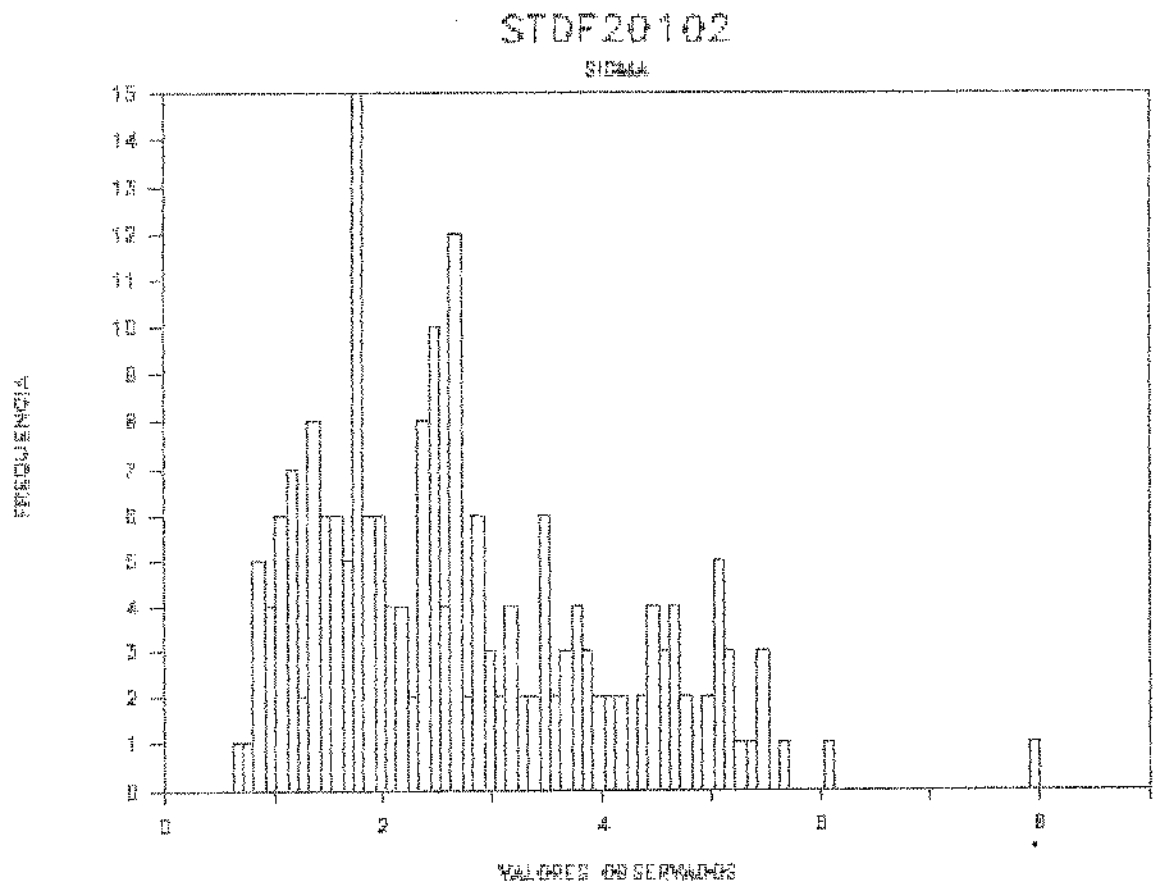


GRÁFICO 148

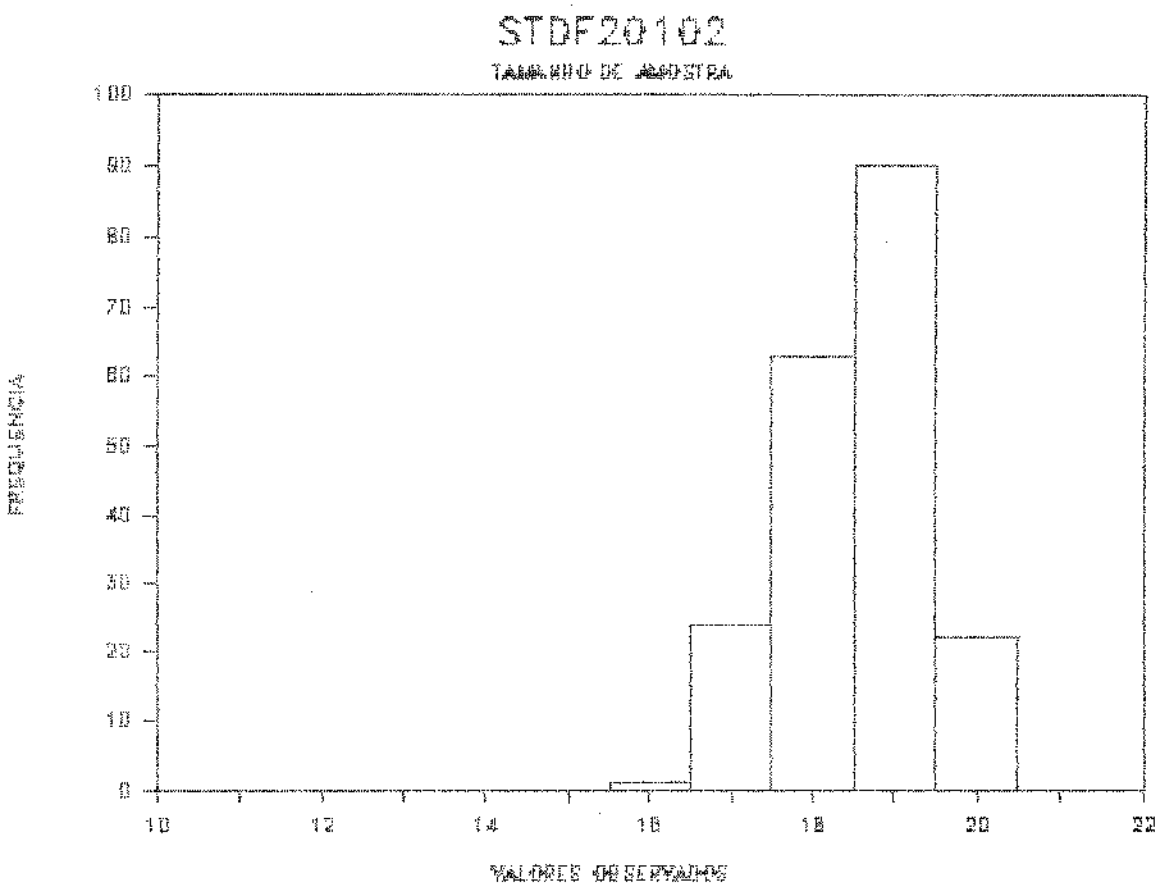

STDF20102
TAMAÑO DE MUESTRA
FRECUENCIA
100
90
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 149

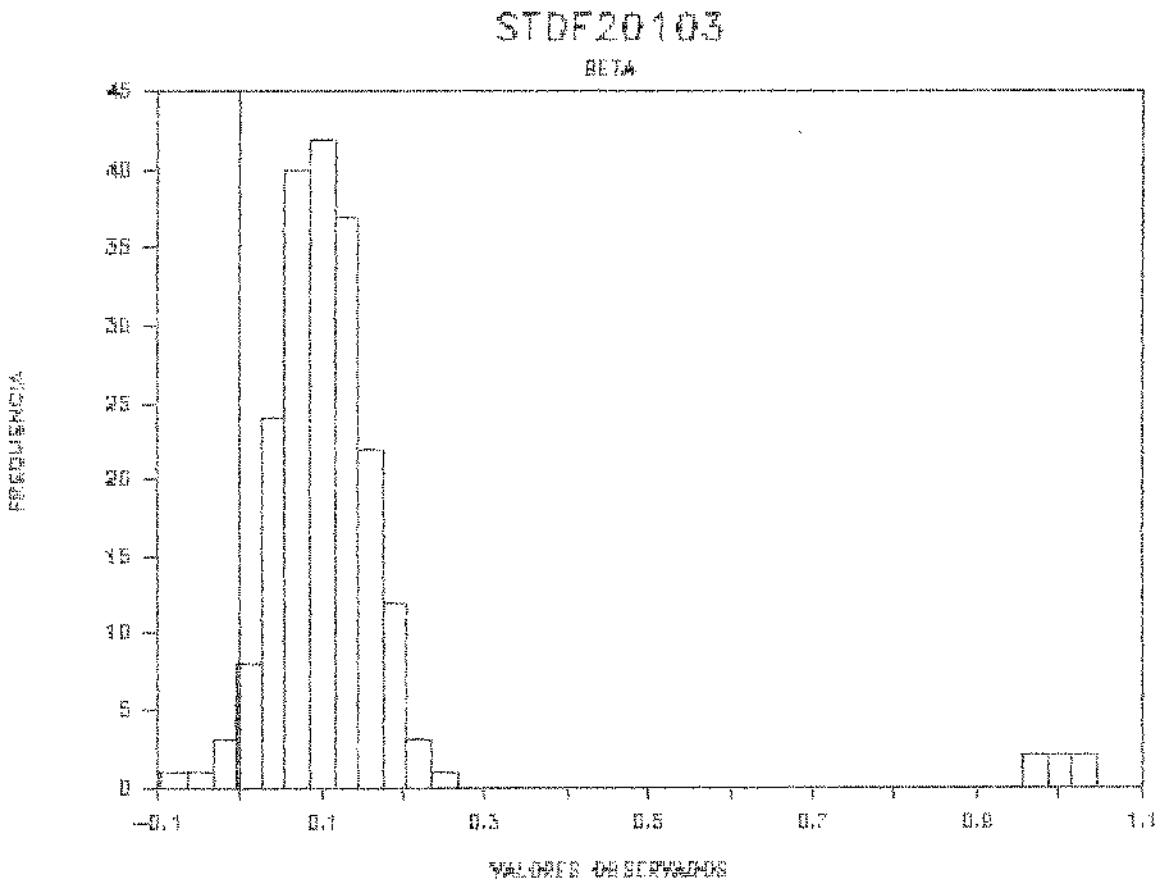

STDF20103
BETA
FREQUENCIA
0
5
10
15
20
25
30
35
40
45
-0.1
0.1
0.3
0.5
0.7
0.9
1.1
VALORES OBSERVADOS

GRÁFICO 150

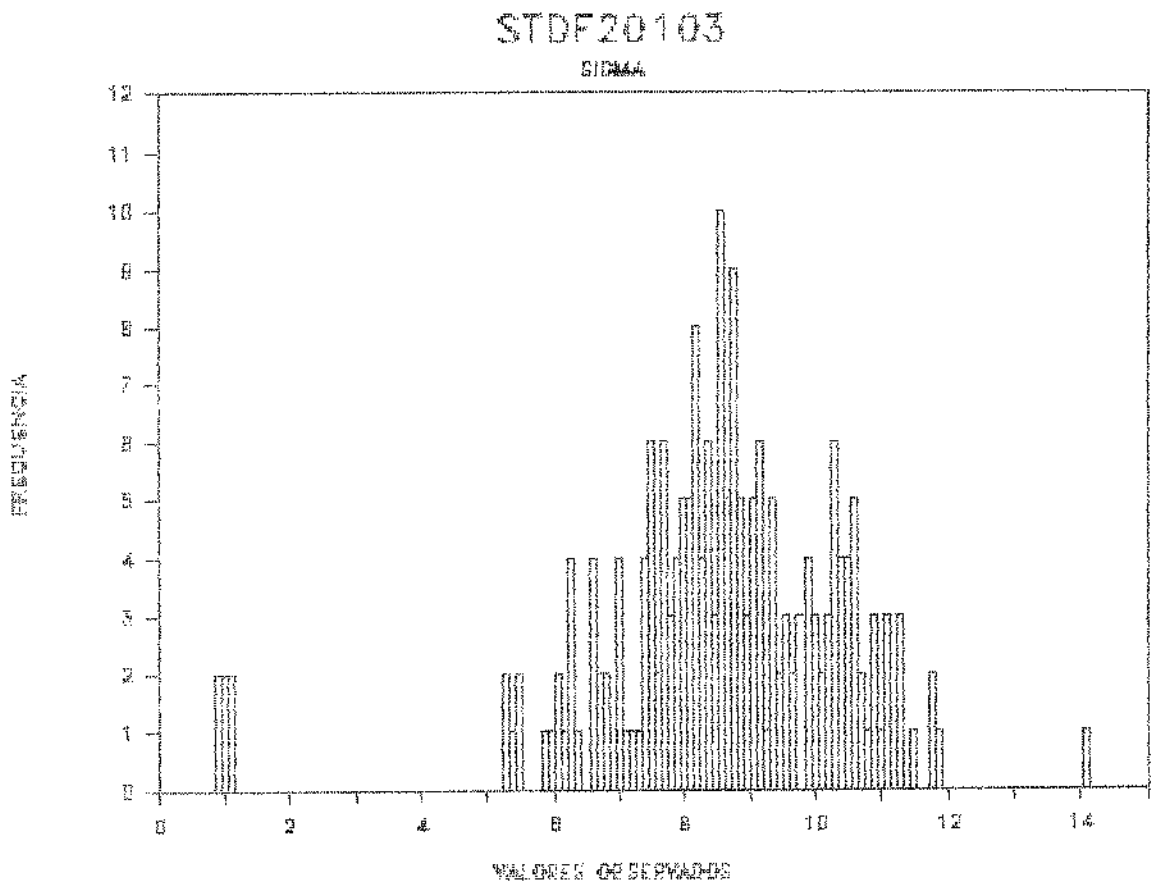


GRÁFICO 151

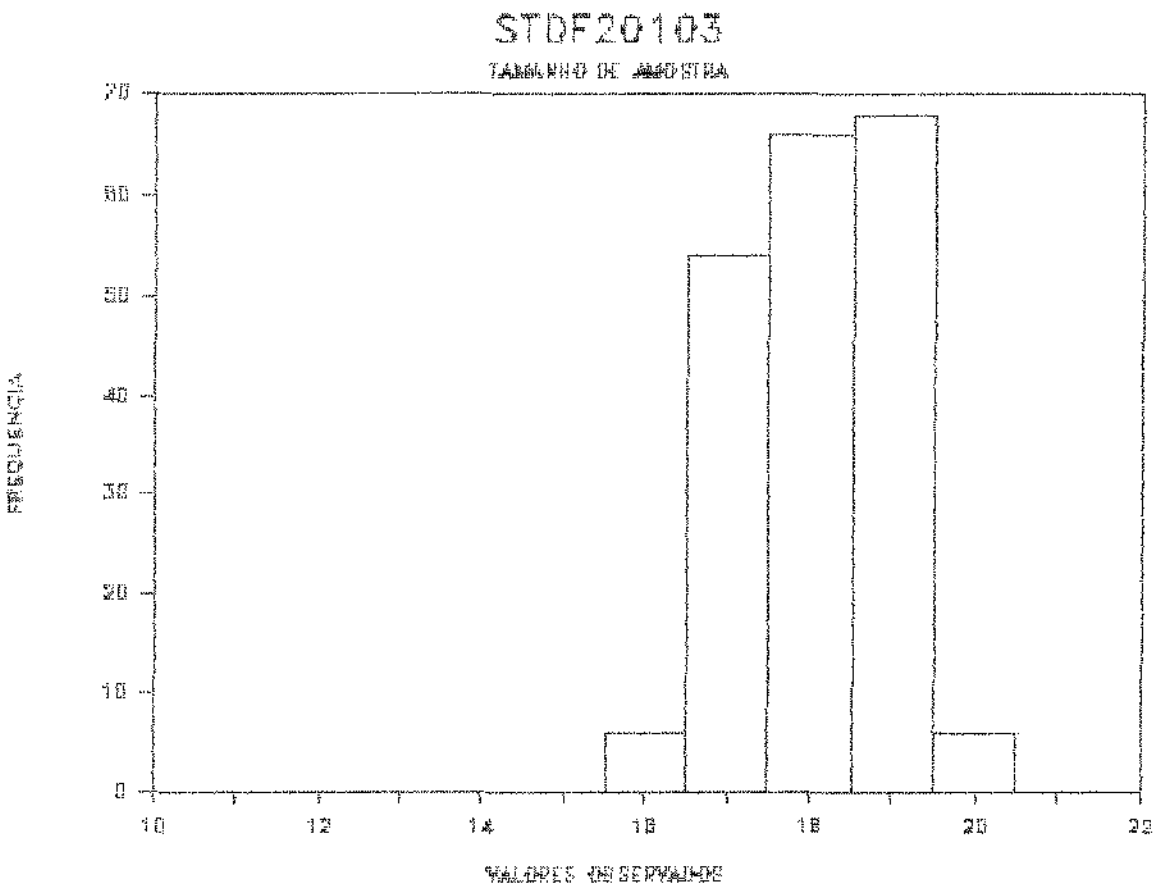


GRÁFICO 152

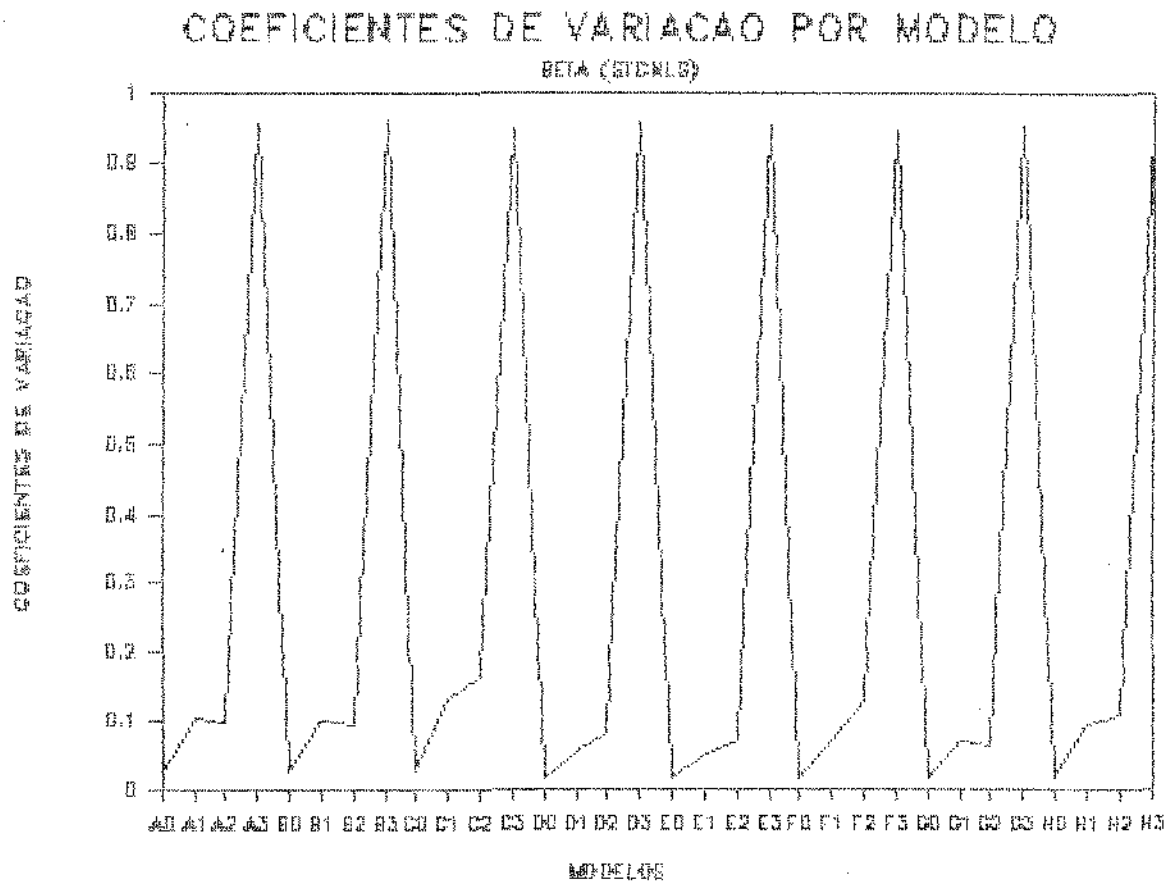

COEFICIENTES DE VARIACAO POR MODELO
BETA (STENLS)
COEFICIENTES DE VARIACAO
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 153

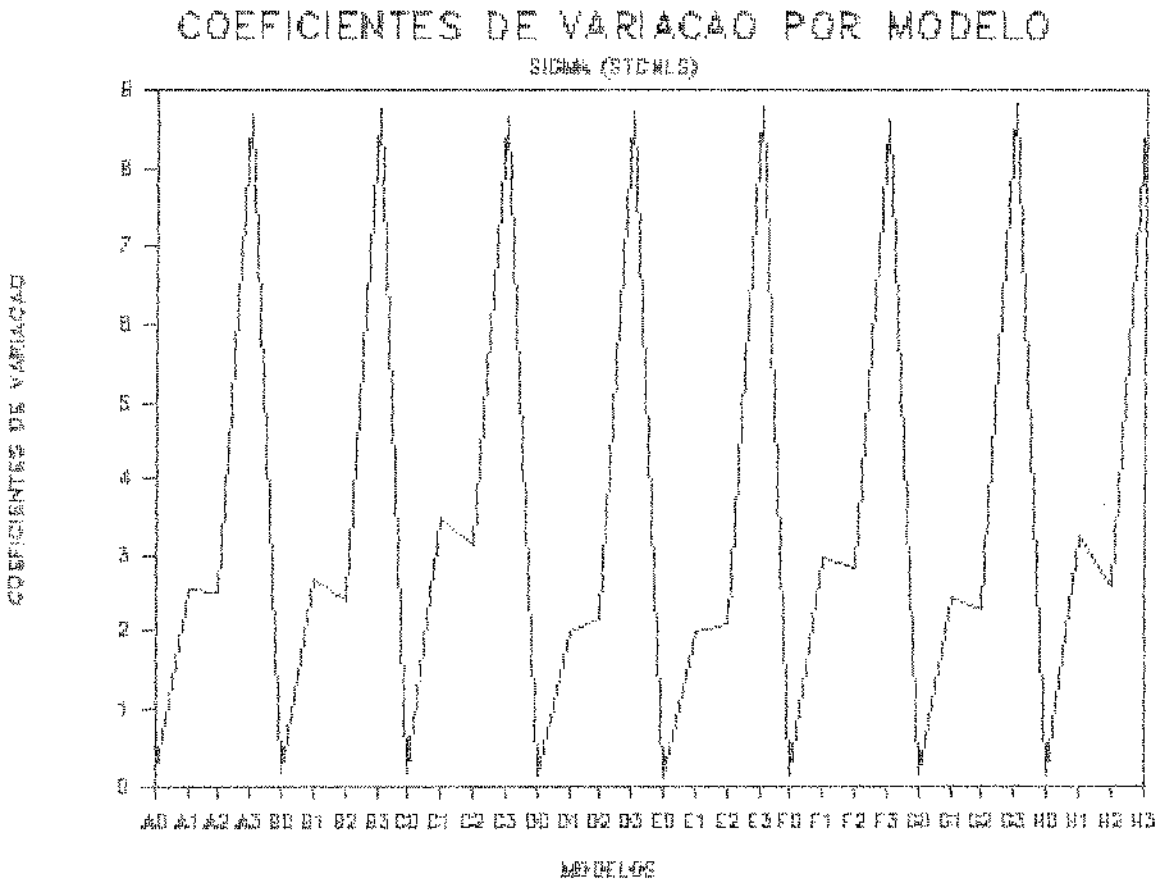

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (STDMLS)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
A0 A1 A2 A3 B0 B1 B2 B3 C0 C1 C2 C3 D0 D1 D2 D3 E0 E1 E2 E3 F0 F1 F2 F3 G0 G1 G2 G3 H0 H1 H2 H3
MODELOS

GRÁFICO 154

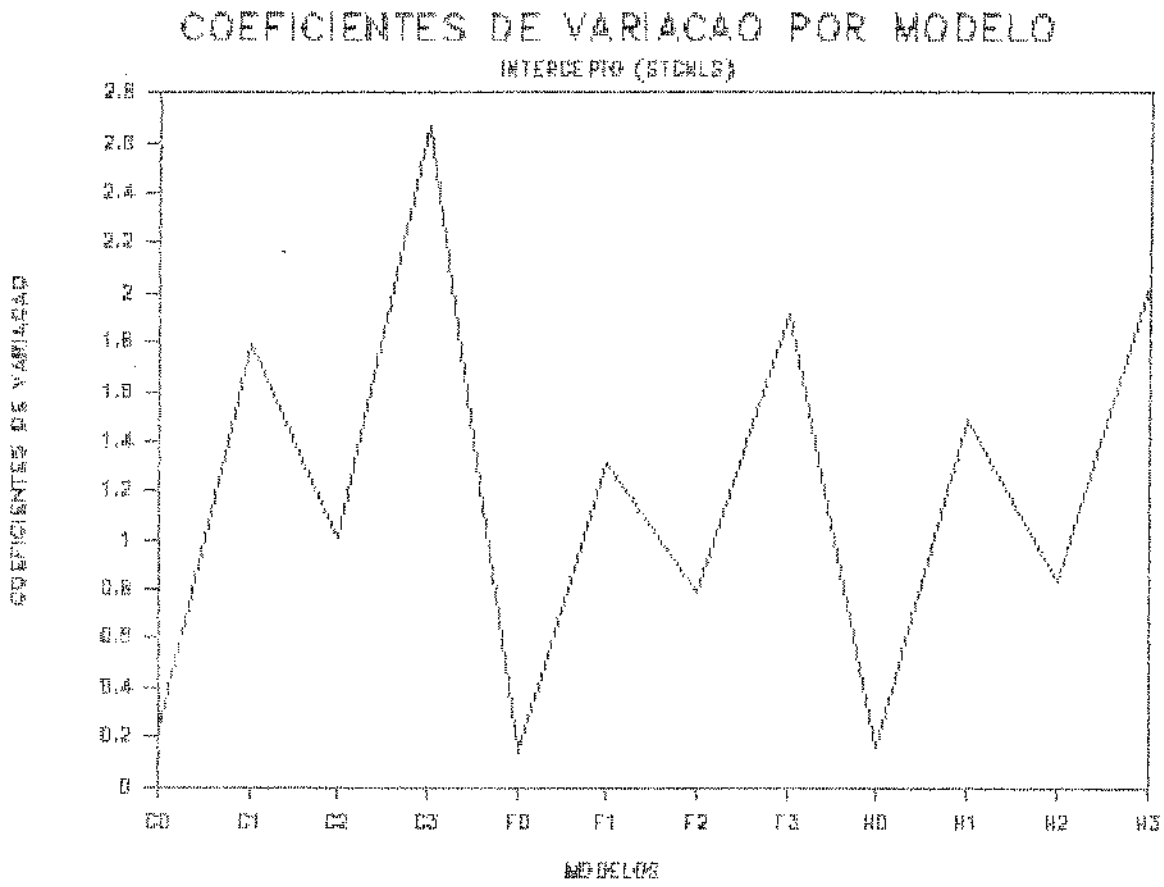

COEFICIENTES DE VARIACAO POR MODELO
INTERCEPTO (STDWLS)
COEFICIENTES DE VARIACAO
2,8
2,6
2,4
2,2
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
C0
C1
C2
C3
F0
F1
F2
F3
H0
H1
H2
H3
MODELOS

GRÁFICO 155

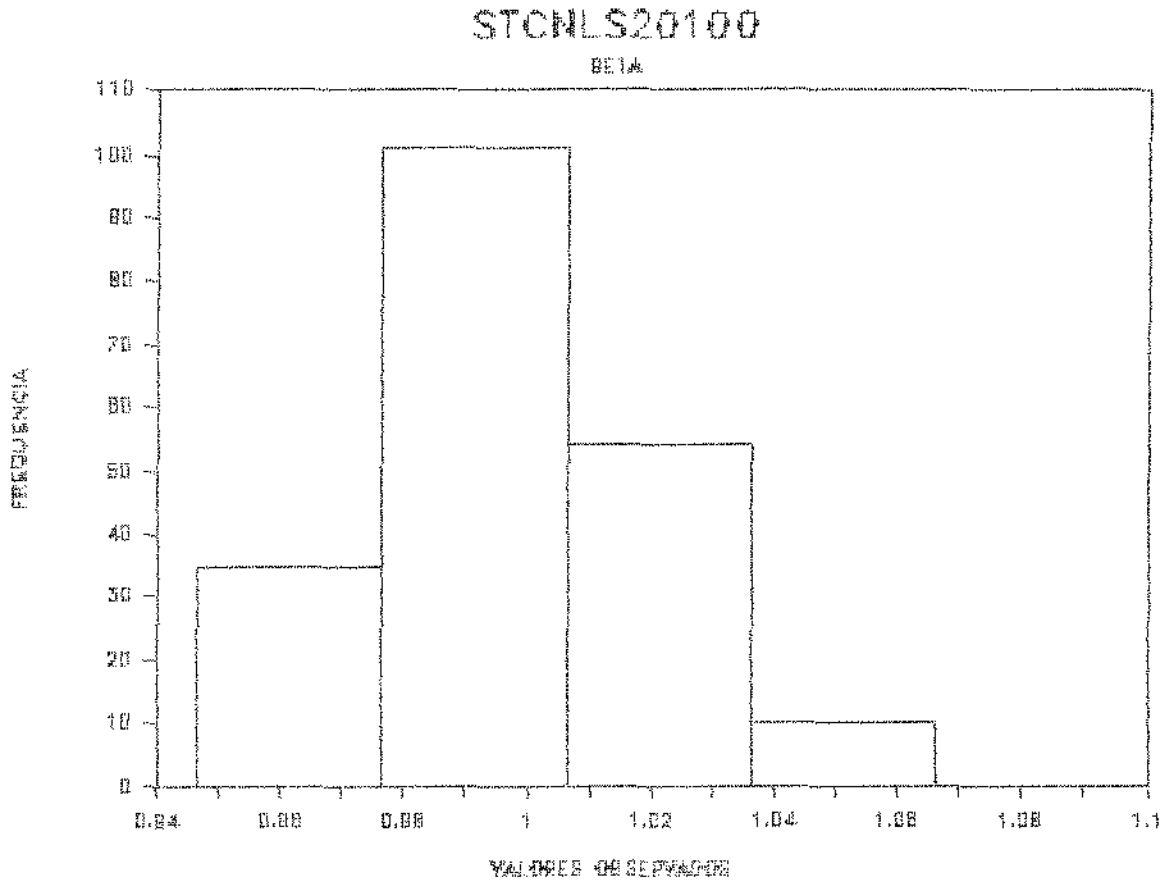

STCNLS20100
BETA
FRECUENCIA
110
100
90
80
70
60
50
40
30
20
10
0
0.94
0.96
0.98
1
1.02
1.04
1.06
1.08
1.1

GRÁFICO 156

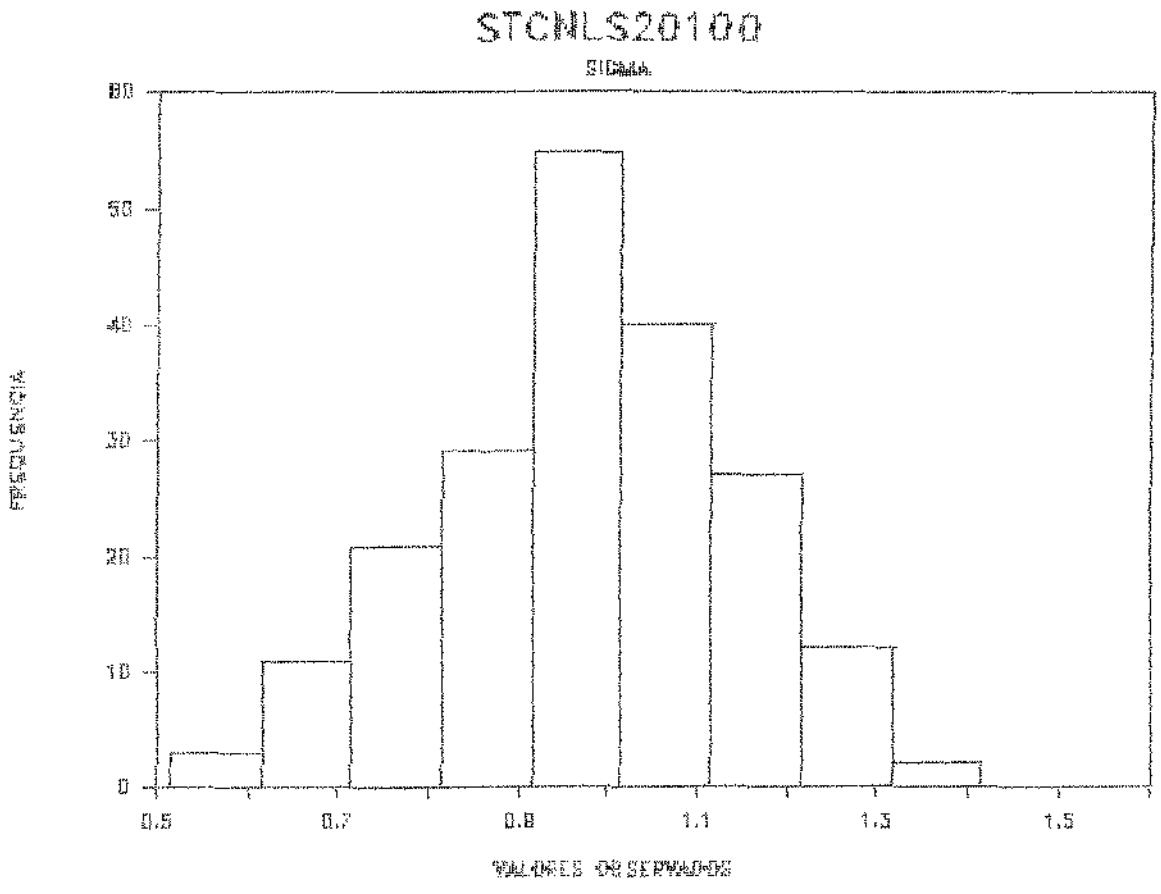


GRÁFICO 157

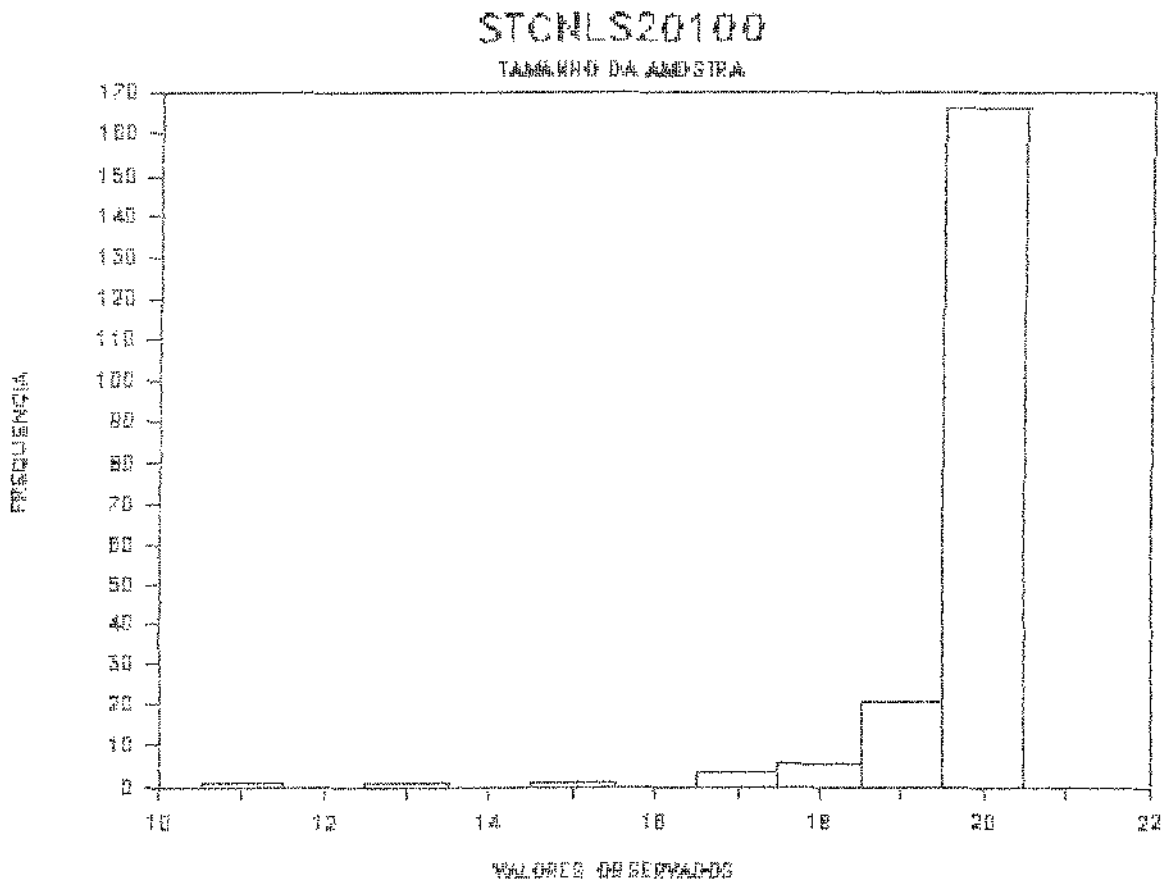

STONLS20100
TAMANHO DA AMOSTRA
FREQUÊNCIA
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 158

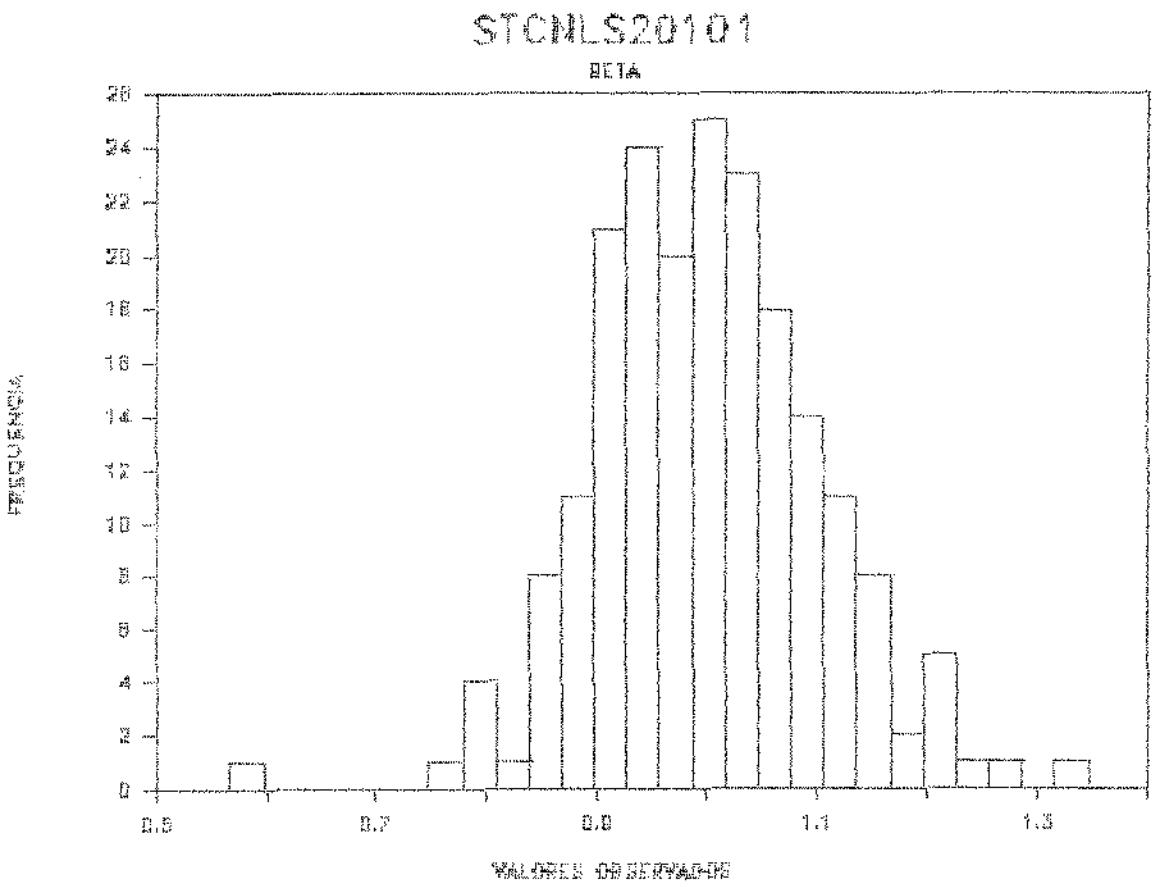

STCNLS20101
BETA
FREQUENCIA
0.6
0.7
0.8
1.1
1.3
VALORES OBSERVADOS

GRÁFICO 159

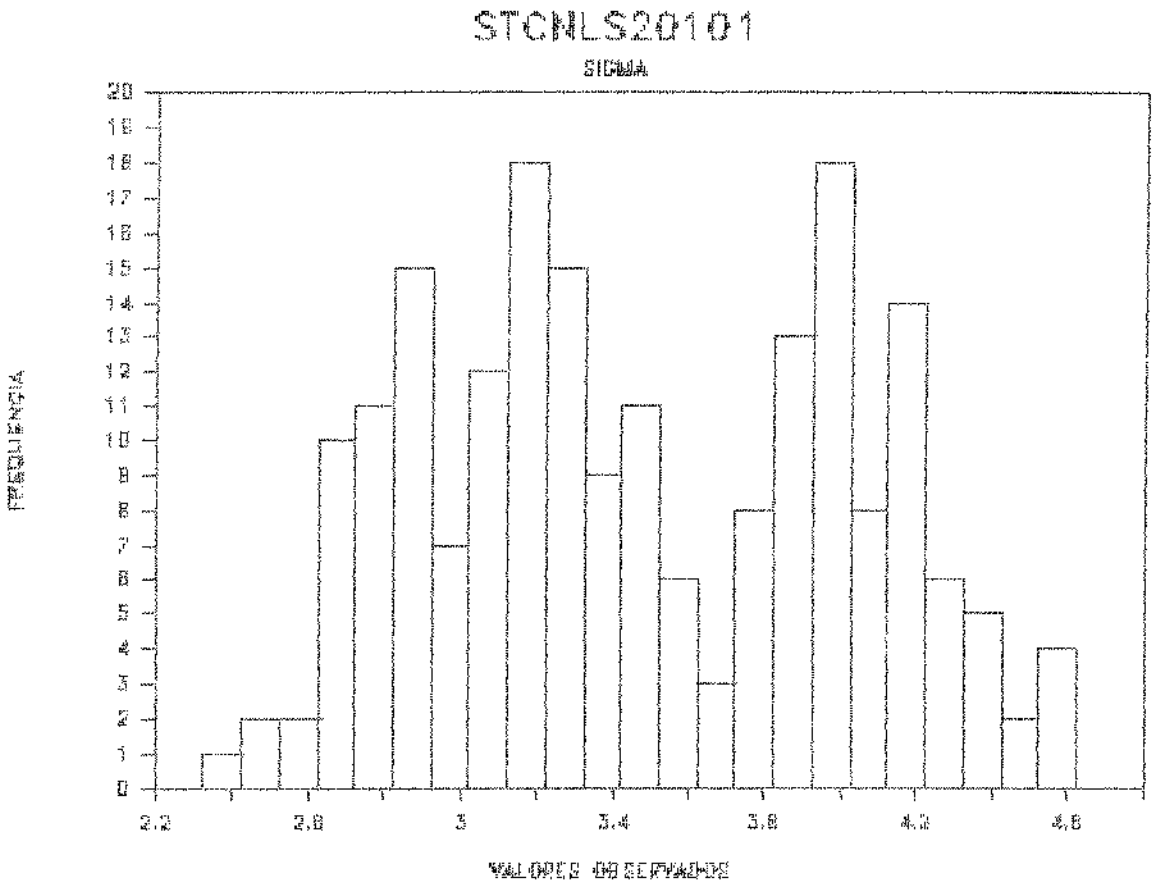


GRÁFICO 160

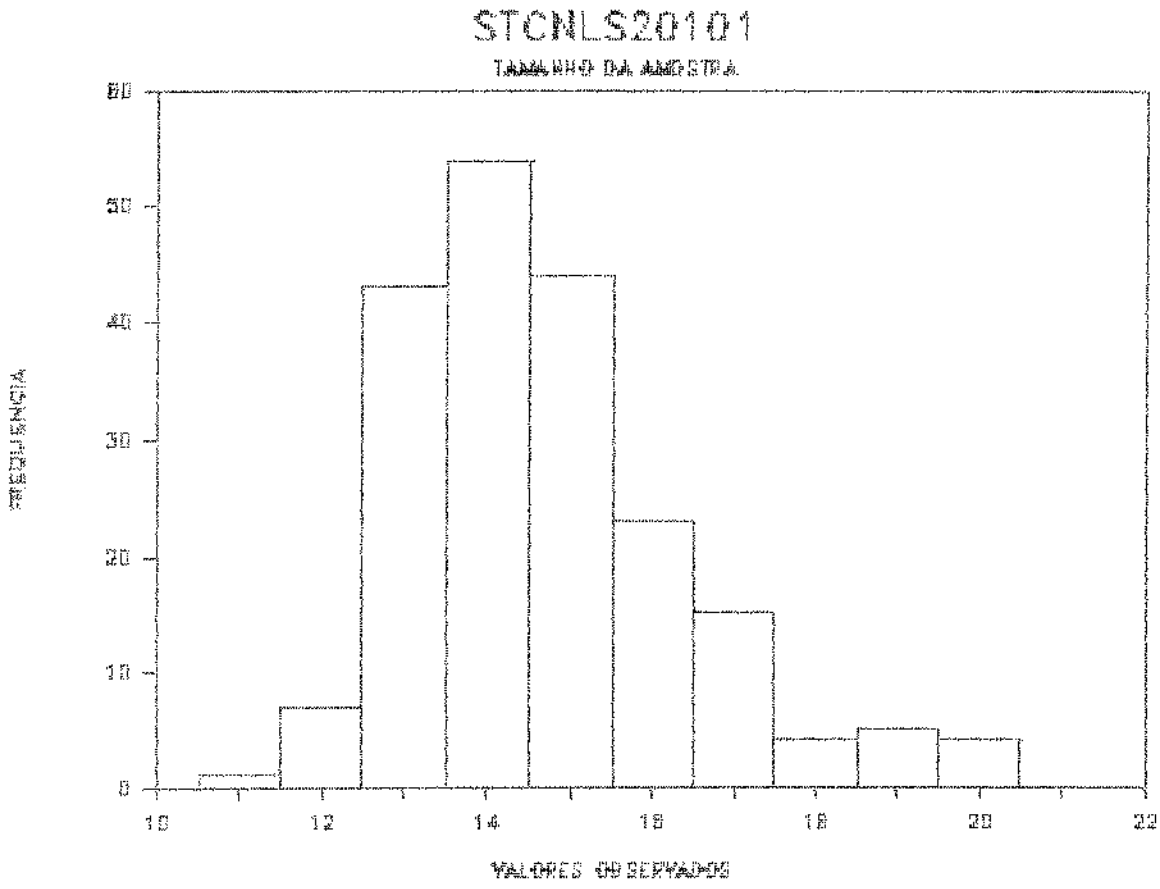

STCNLS20101
TAMANHO DA AMOSTRA
FREQUÊNCIA
60
50
40
30
20
10
0
10
12
14
16
18
20
22
VALORES OBSERVADOS

GRÁFICO 161

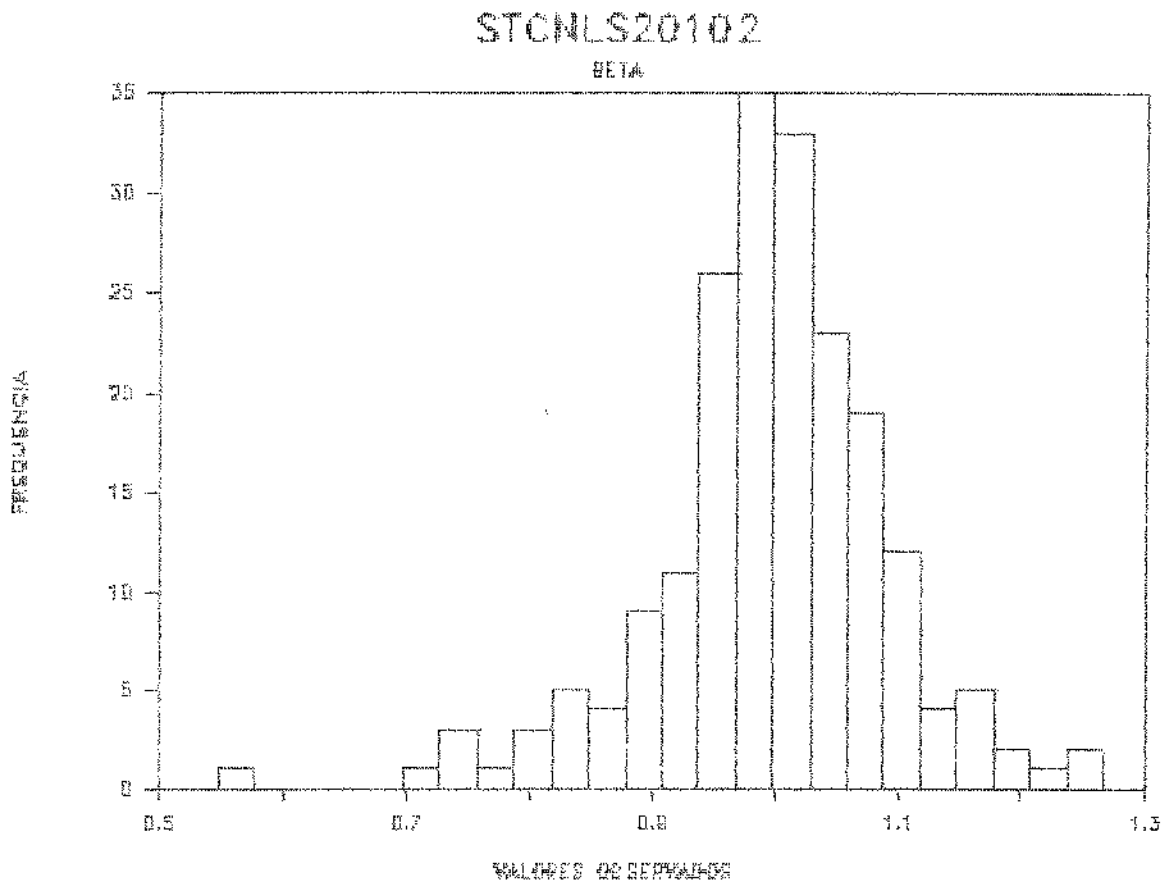


GRÁFICO 162

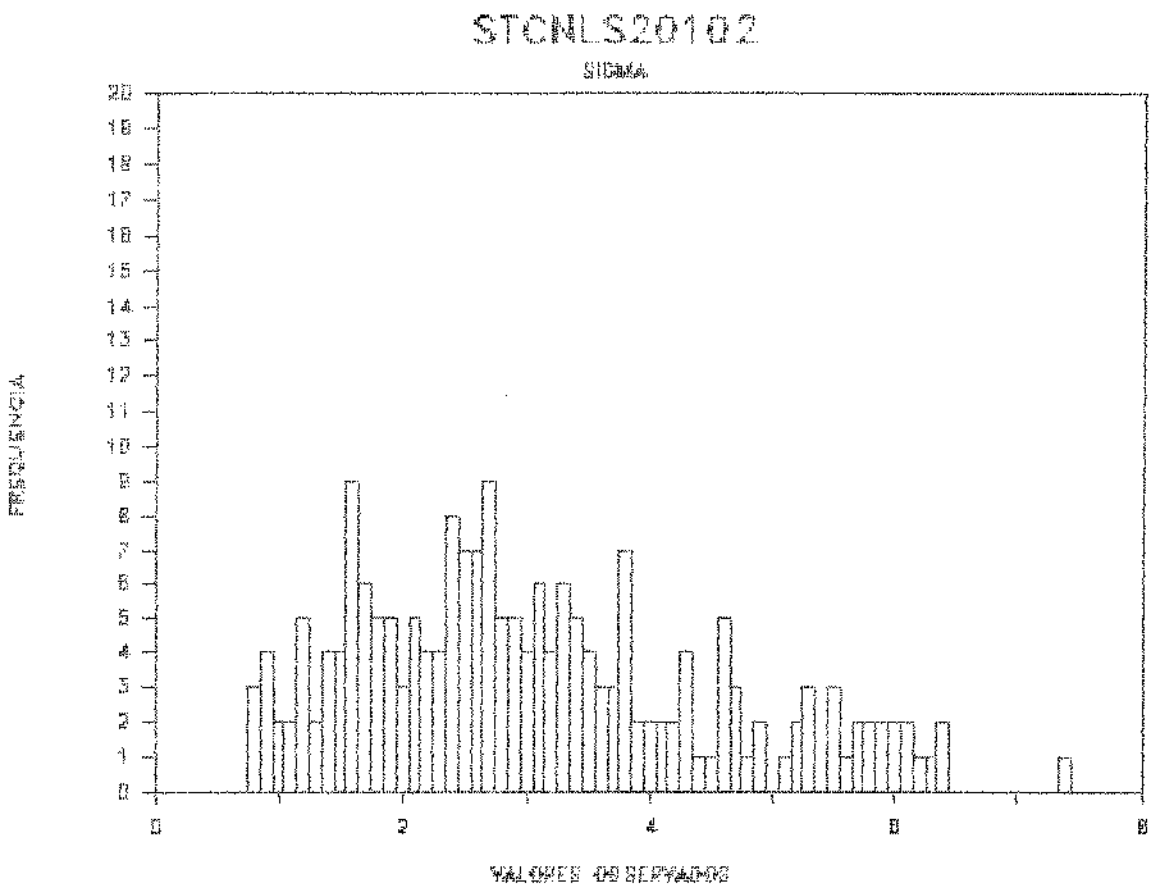


GRÁFICO 163

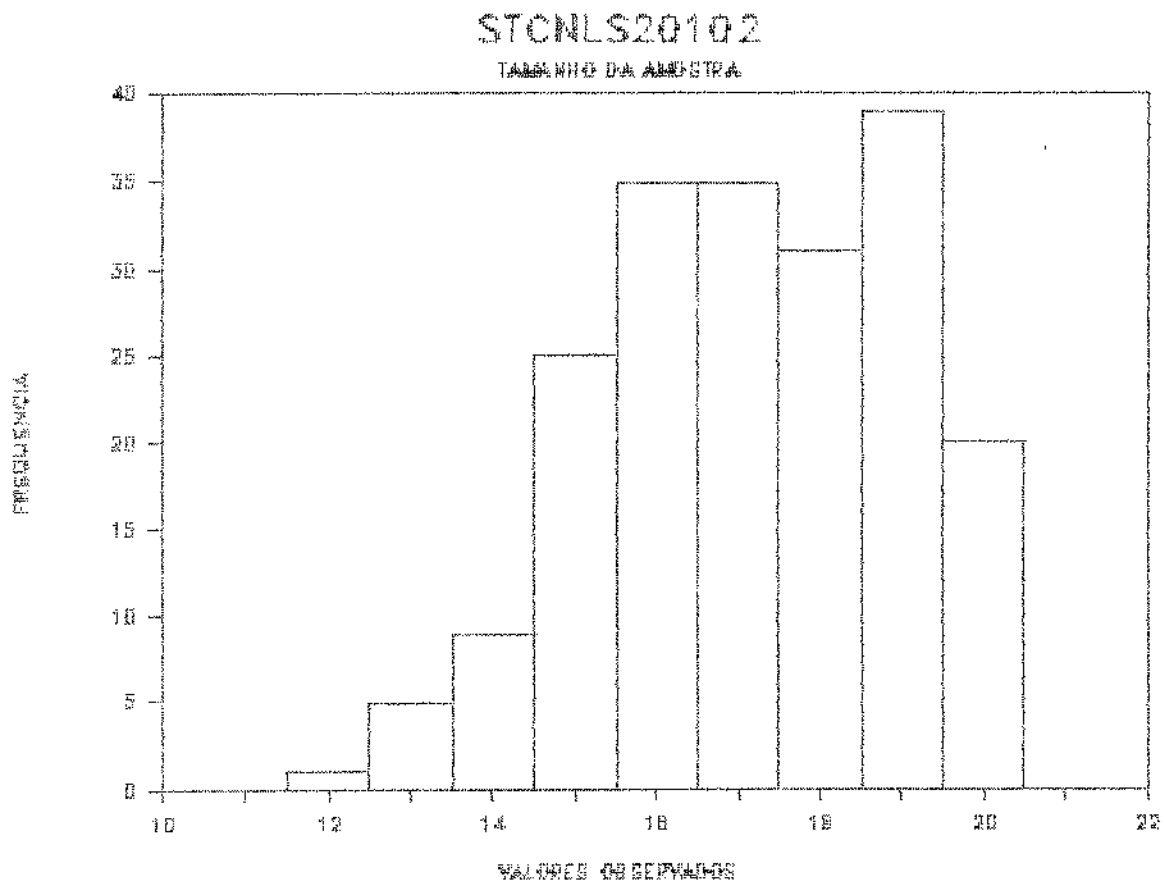

STCNLS20102
TAMANHO DA AMOSTRA
40
35
30
25
20
15
10
5
0
10
12
14
16
18
20
22

GRÁFICO 164

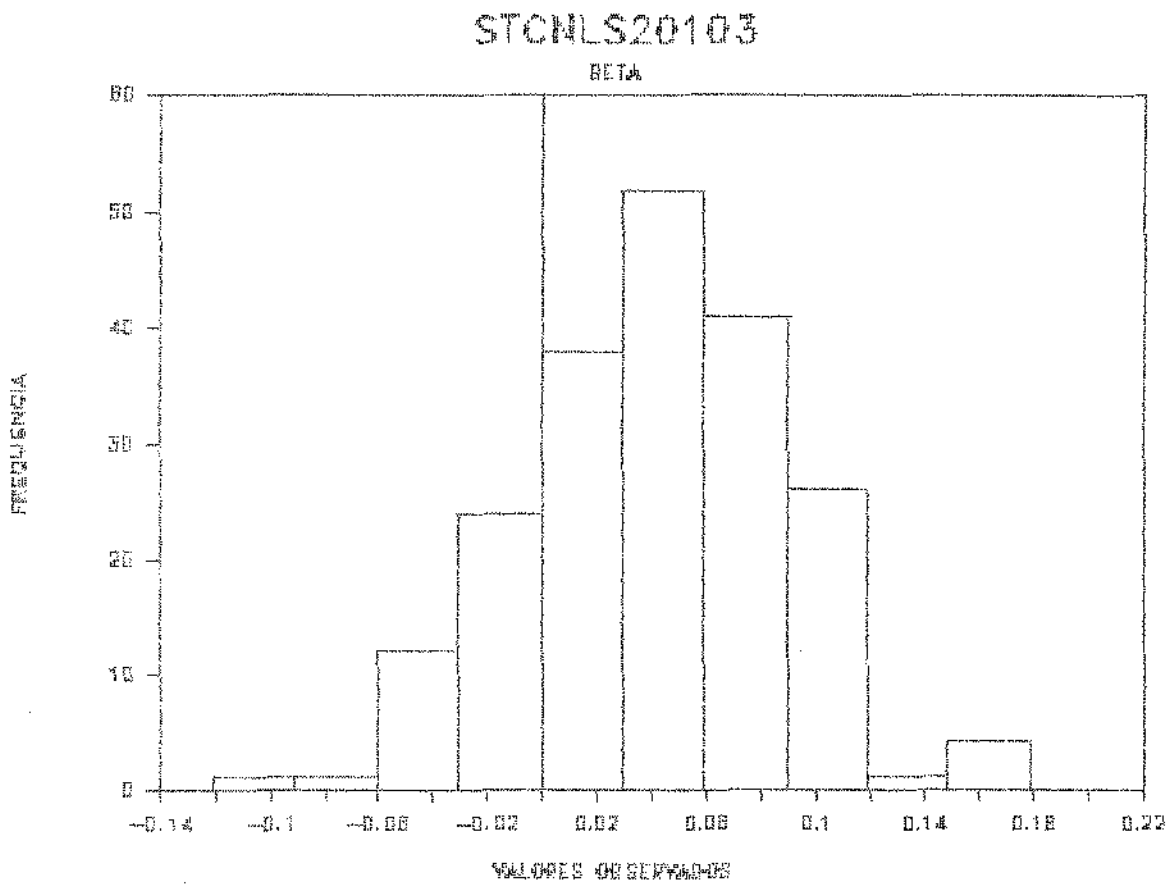

STCNLS20103
BETA
FREQUENCIA
0
10
20
30
40
50
60
-0.14
-0.1
-0.06
-0.02
0.02
0.06
0.1
0.14
0.18
0.22
VALORES OBSERVADOS

GRÁFICO 165

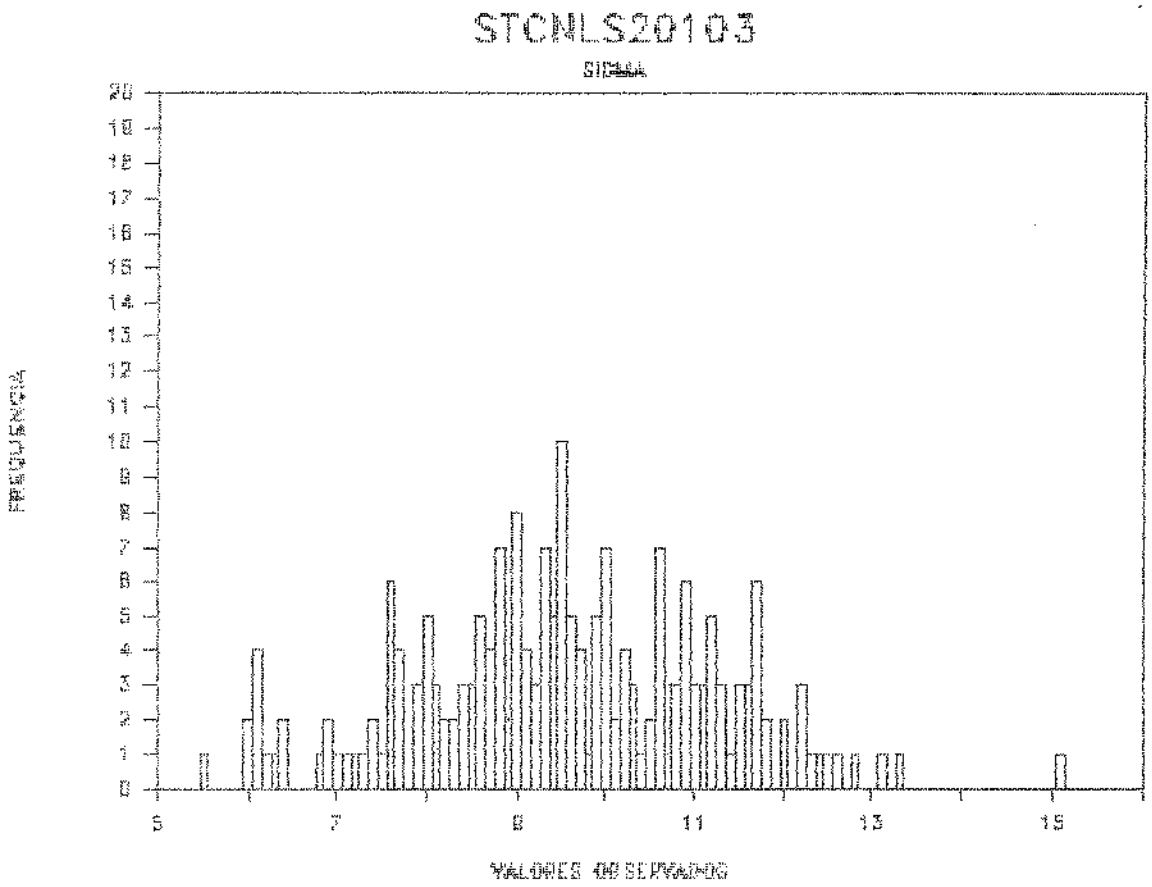


GRÁFICO 166

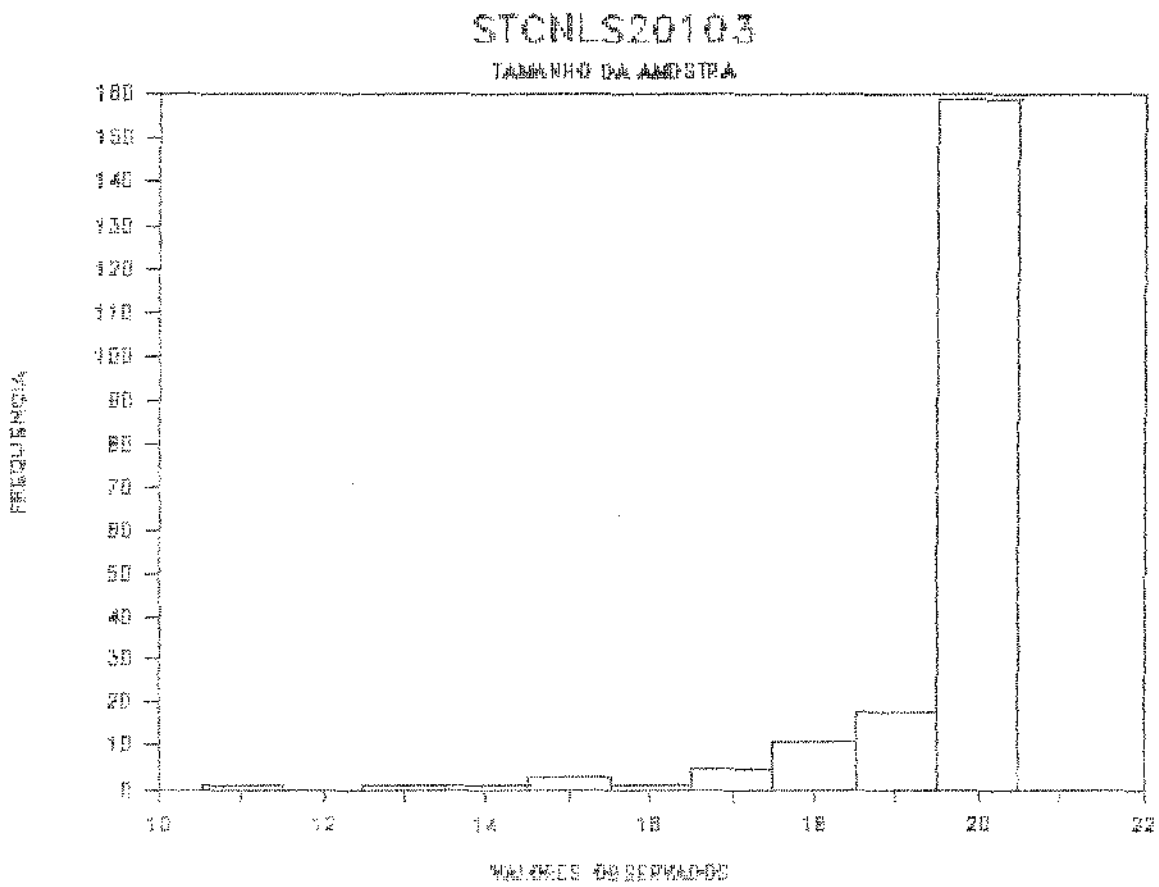

STCNLS20103
TAMANHO DA AMOSTRA
180
160
140
120
100
80
70
60
50
40
30
20
10
0
10
12
14
16
18
20
22

GRÁFICO 167

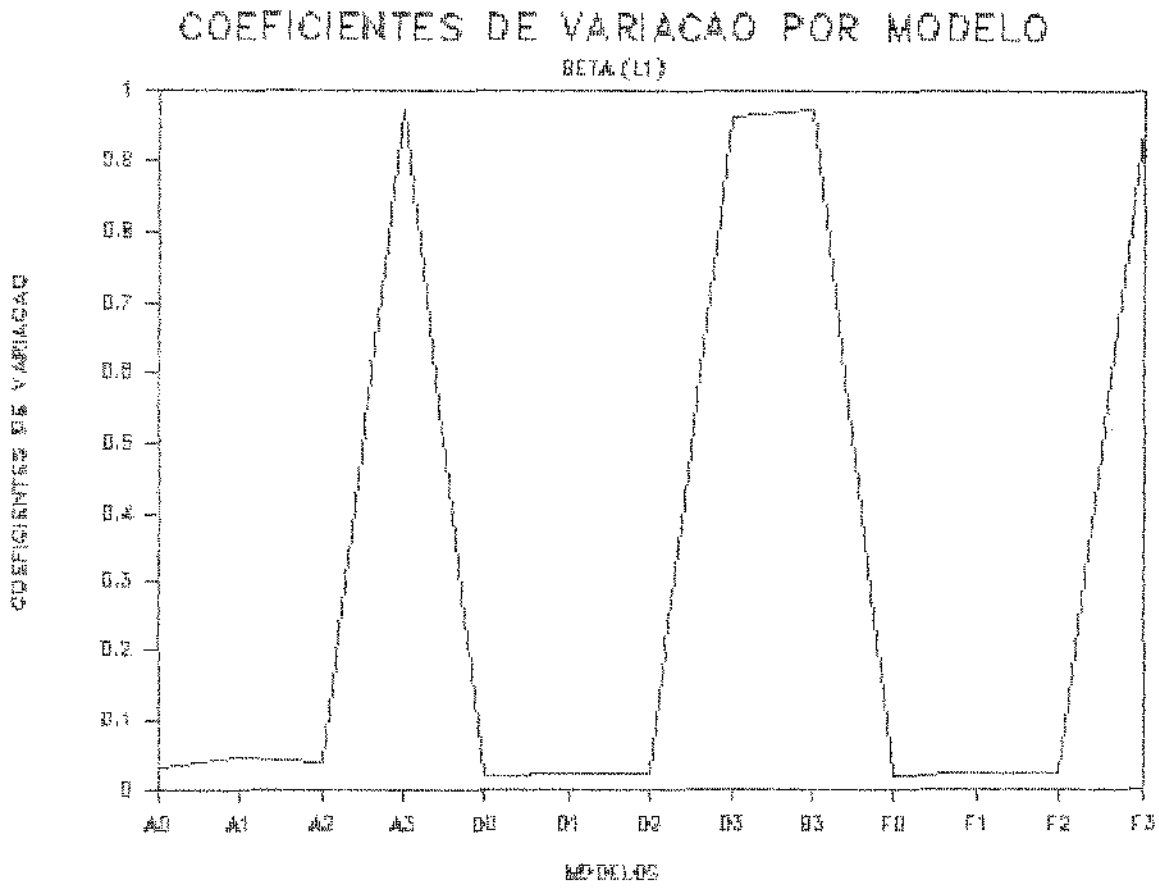

COEFICIENTES DE VARIACAO POR MODELO
BETA(L1)
COEFICIENTES DE VARIACAO
1
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A0
A1
A2
A3
B0
B1
B2
B3
F0
F1
F2
F3
MODELOS

GRÁFICO 168

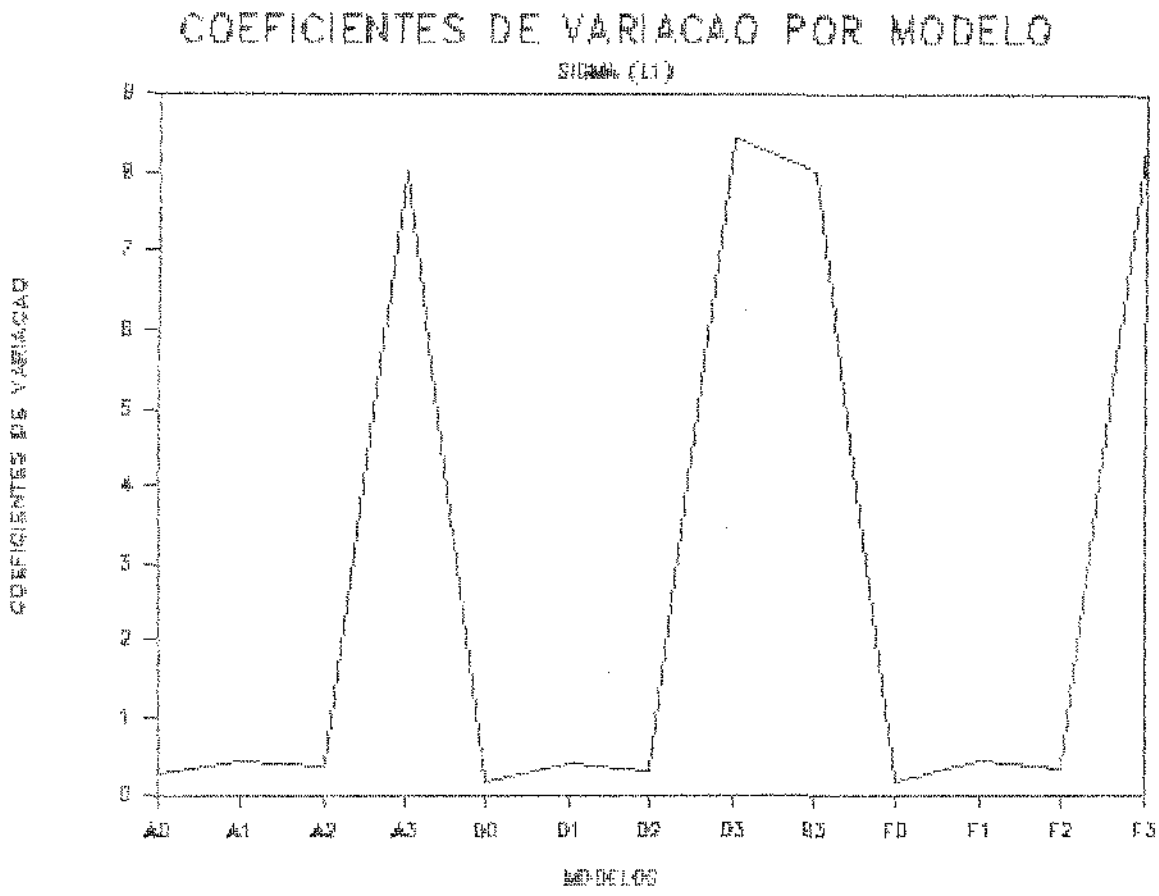

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (L1)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
A0
A1
A2
A3
B0
B1
B2
B3
B5
F0
F1
F2
F3
MODELOS

GRÁFICO 169

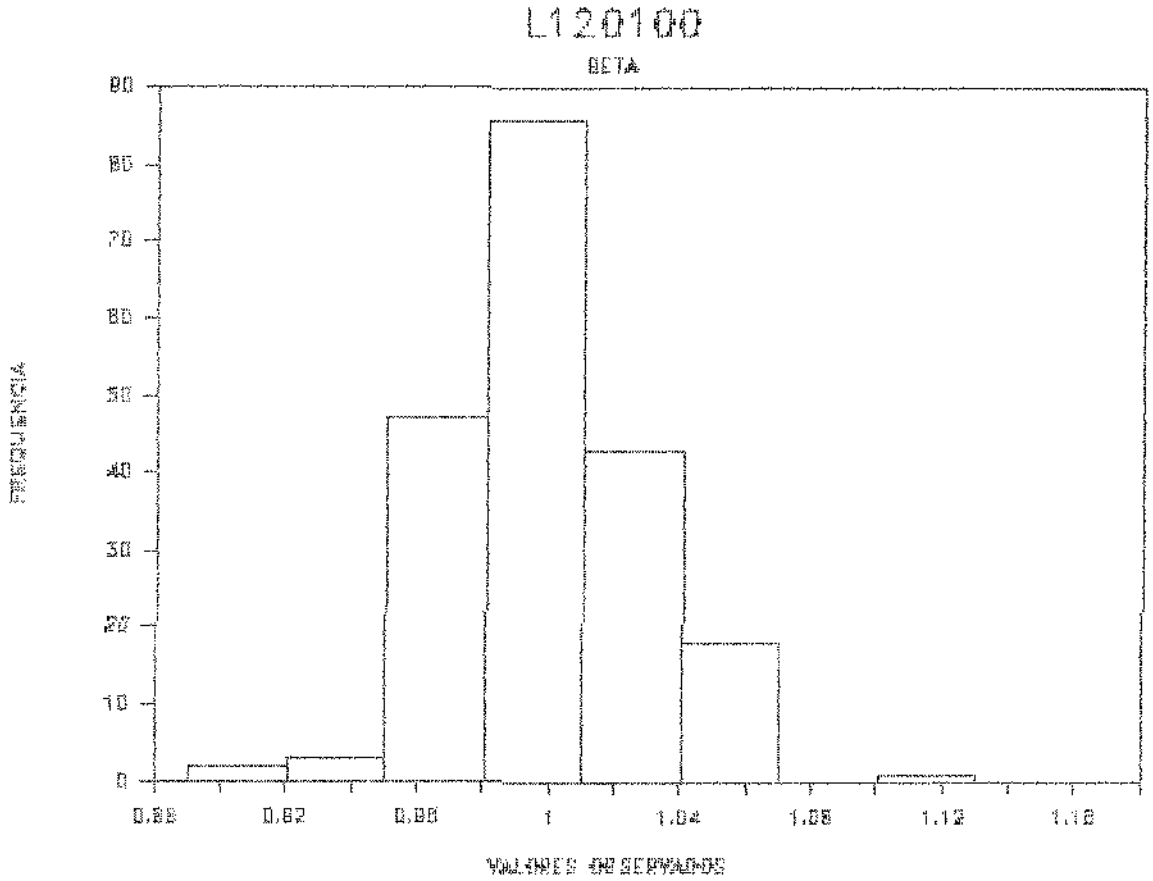


GRÁFICO 170

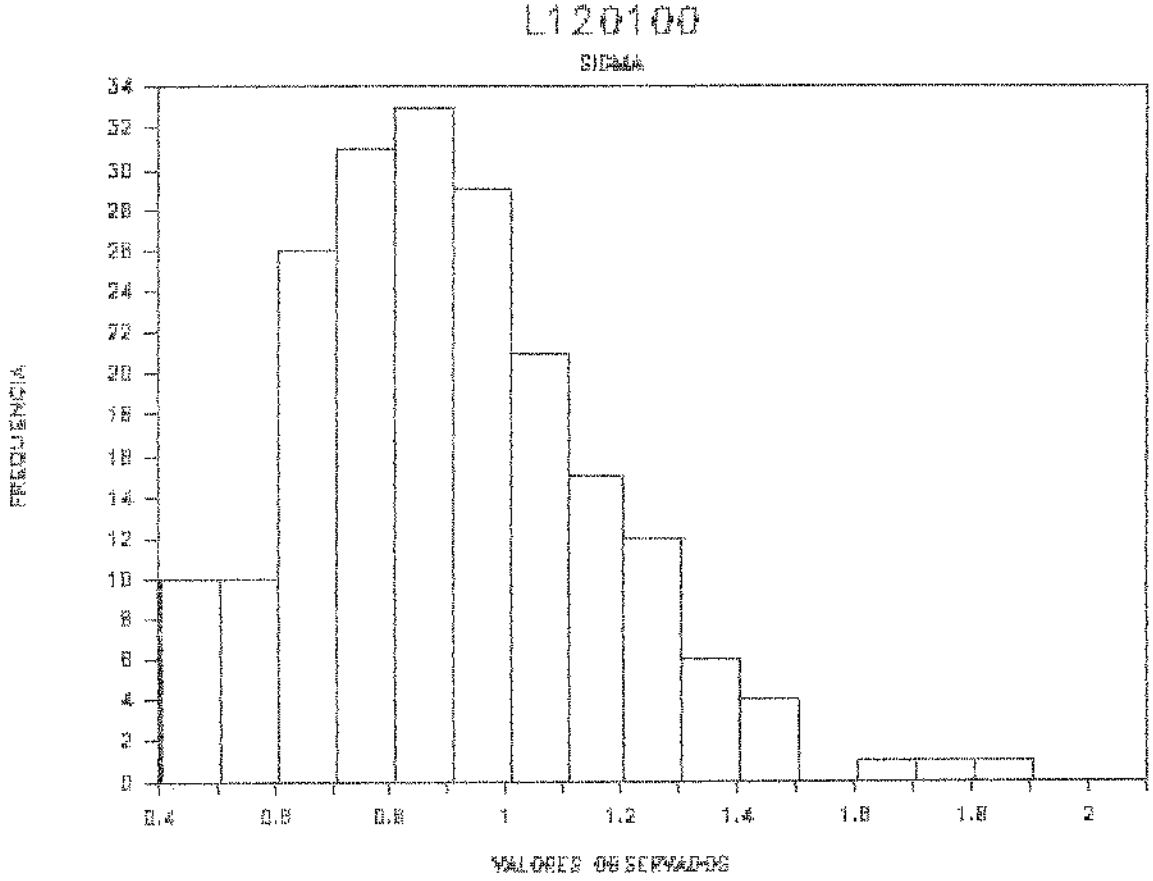


GRÁFICO 171

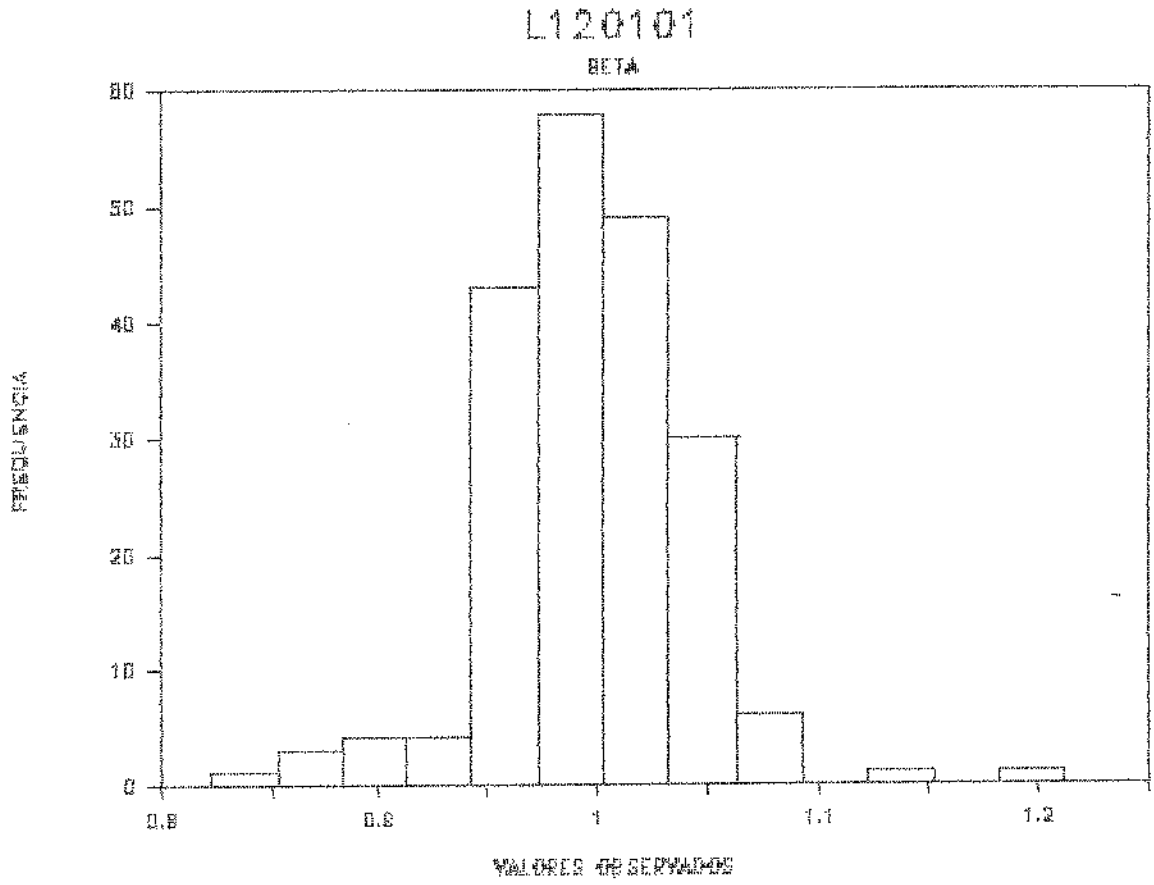


GRÁFICO 172

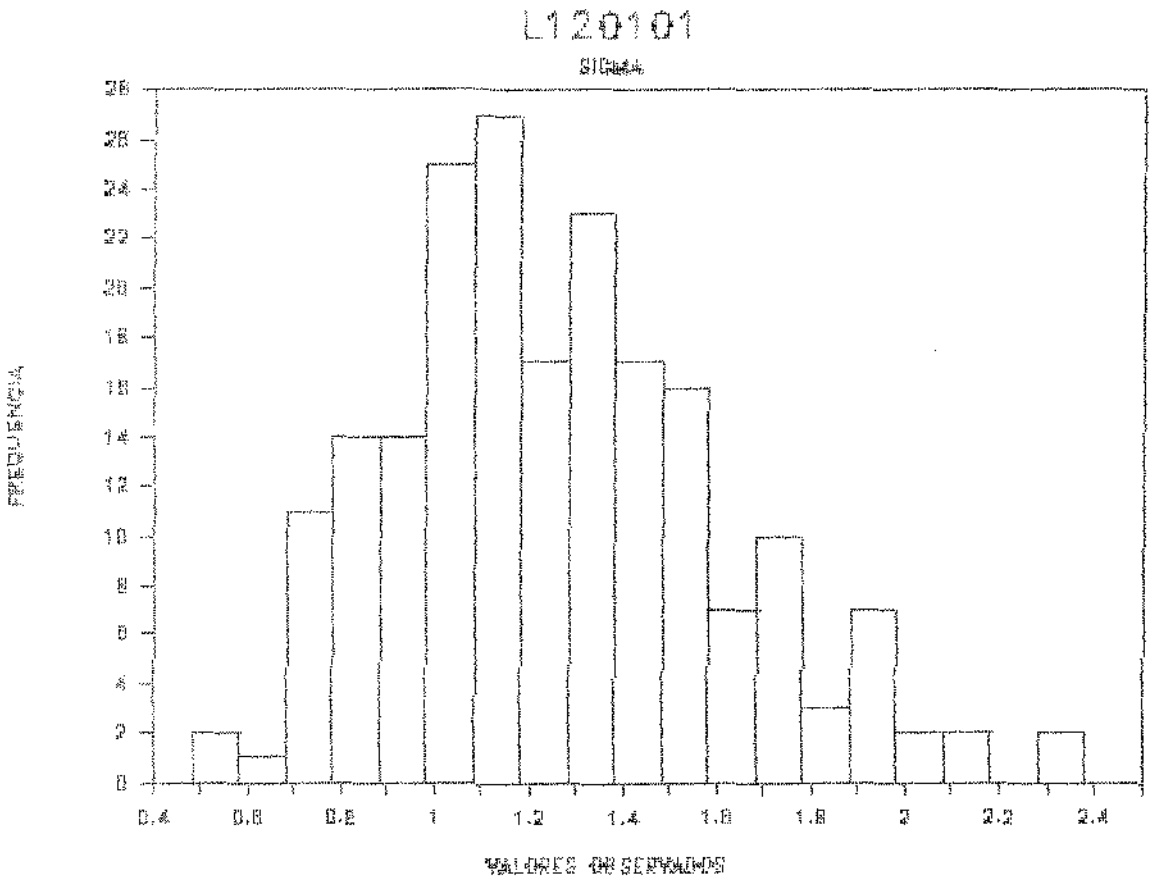

L120101
SIGMA
FRECUENCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
VALORES OBSERVADOS

GRÁFICO 173

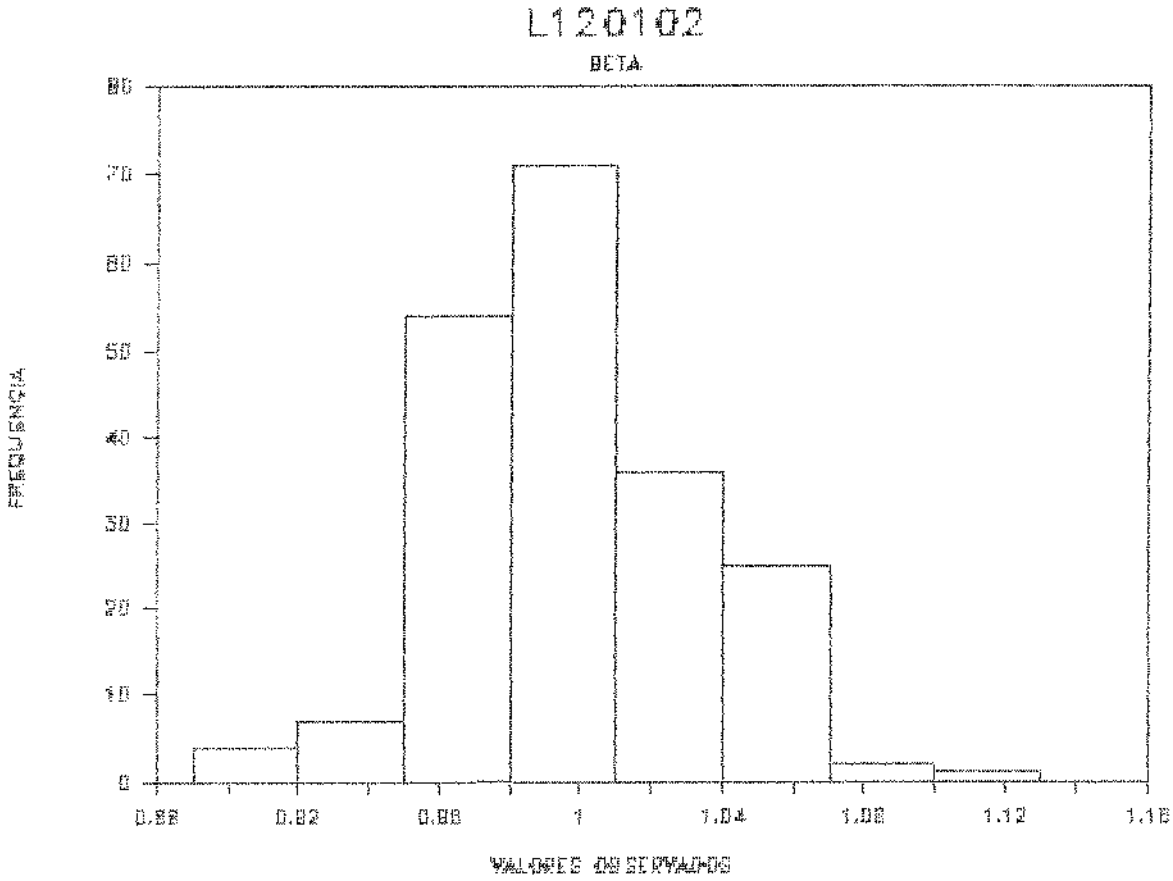

L120102
BETA
80
70
60
50
40
30
20
10
0
FRECUENCIA
0.98
0.82
0.88
1
1.04
1.08
1.12
1.16
VALORES OBSERVADOS

GRÁFICO 174

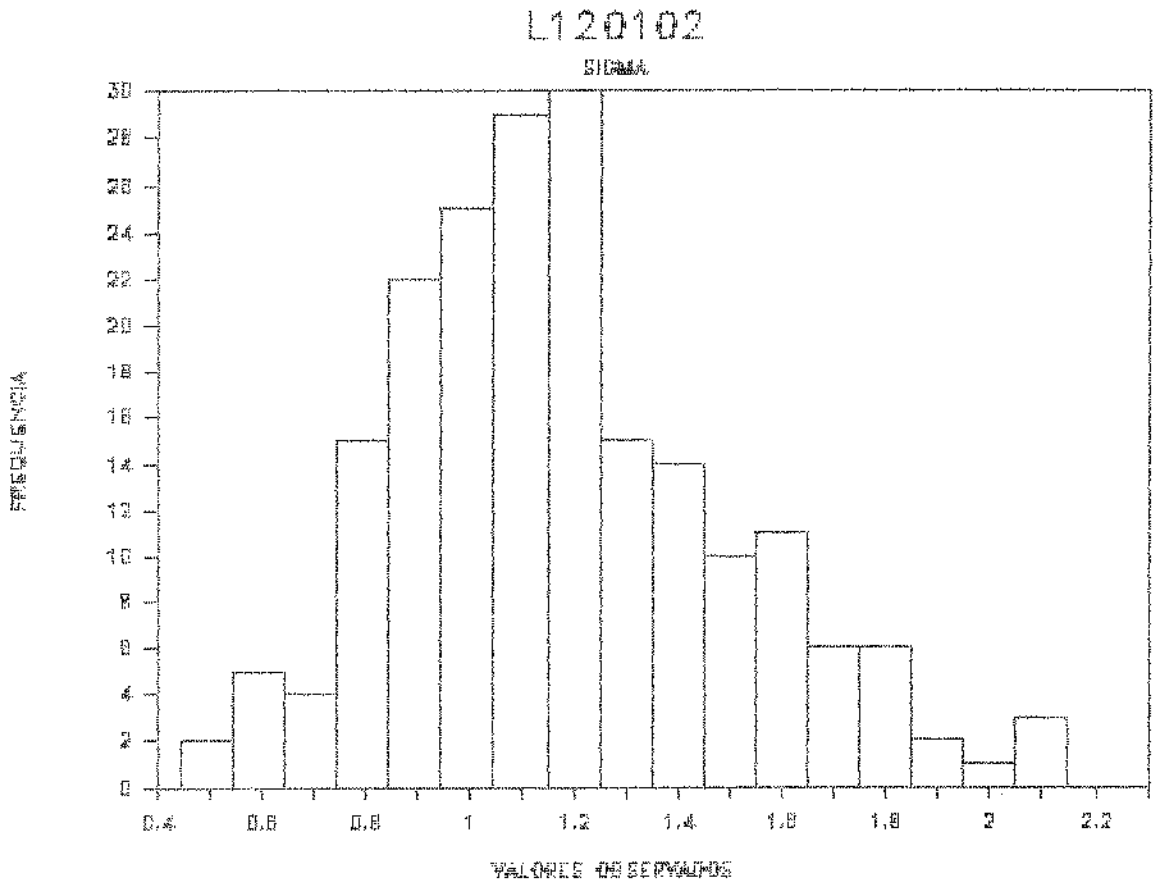

L120102
SIGMA
FREQUENCIA
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
0,6
0,8
1
1,2
1,4
1,6
1,8
2
2,2
VALORES OBSERVADOS

GRÁFICO 175

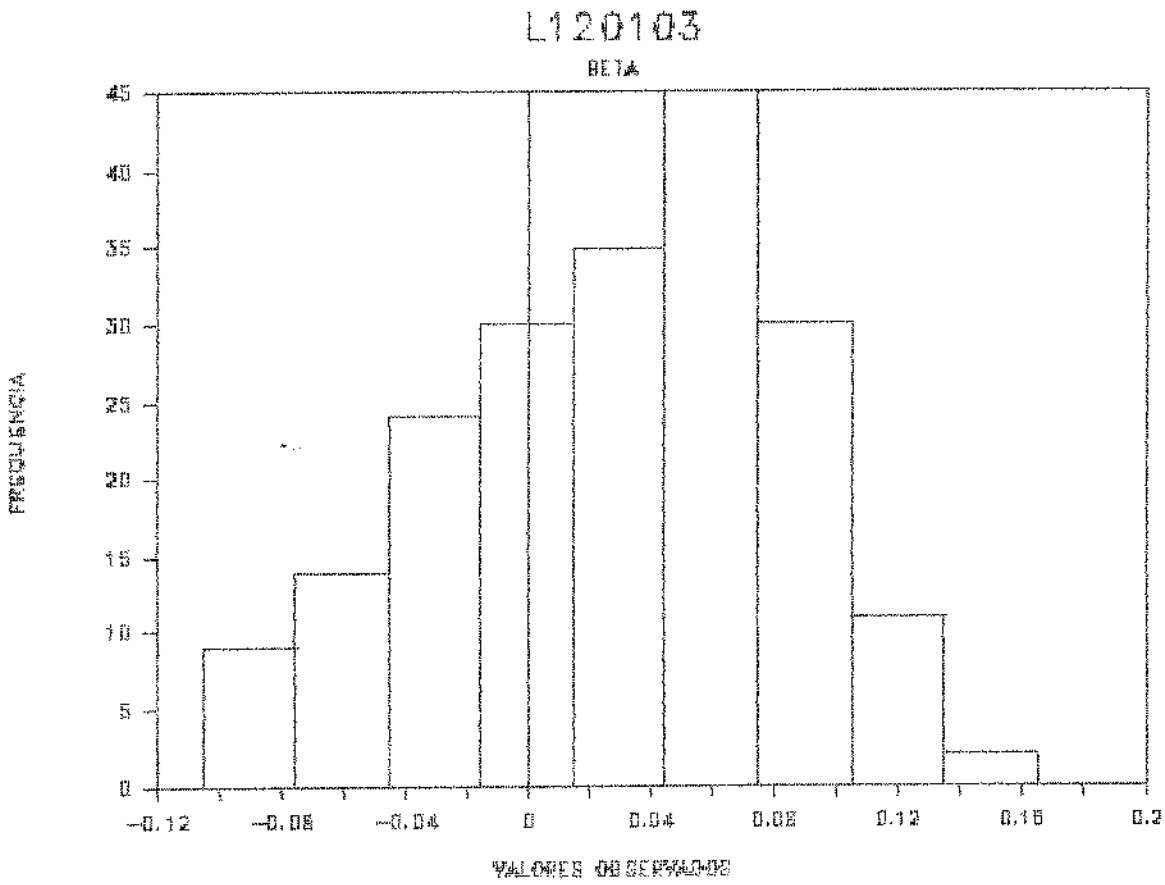

L120103
BETA
45
40
35
30
25
20
15
10
5
0
-0.12
-0.08
-0.04
0
0.04
0.08
0.12
0.16
0.2

GRÁFICO 176

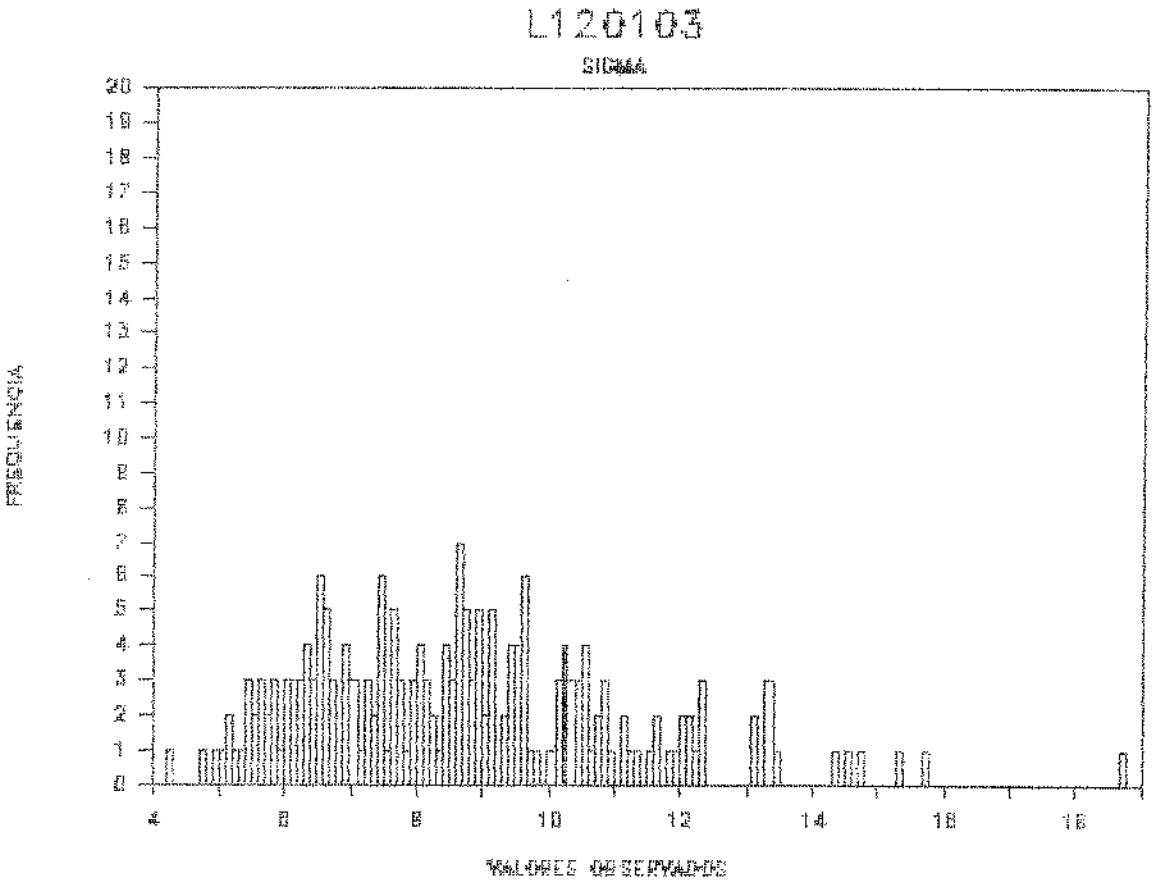

L120103
SIGMA
FRECUENCIA
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
4
6
8
10
12
14
16
18
VALORES OBSERVADOS

GRÁFICO 177A

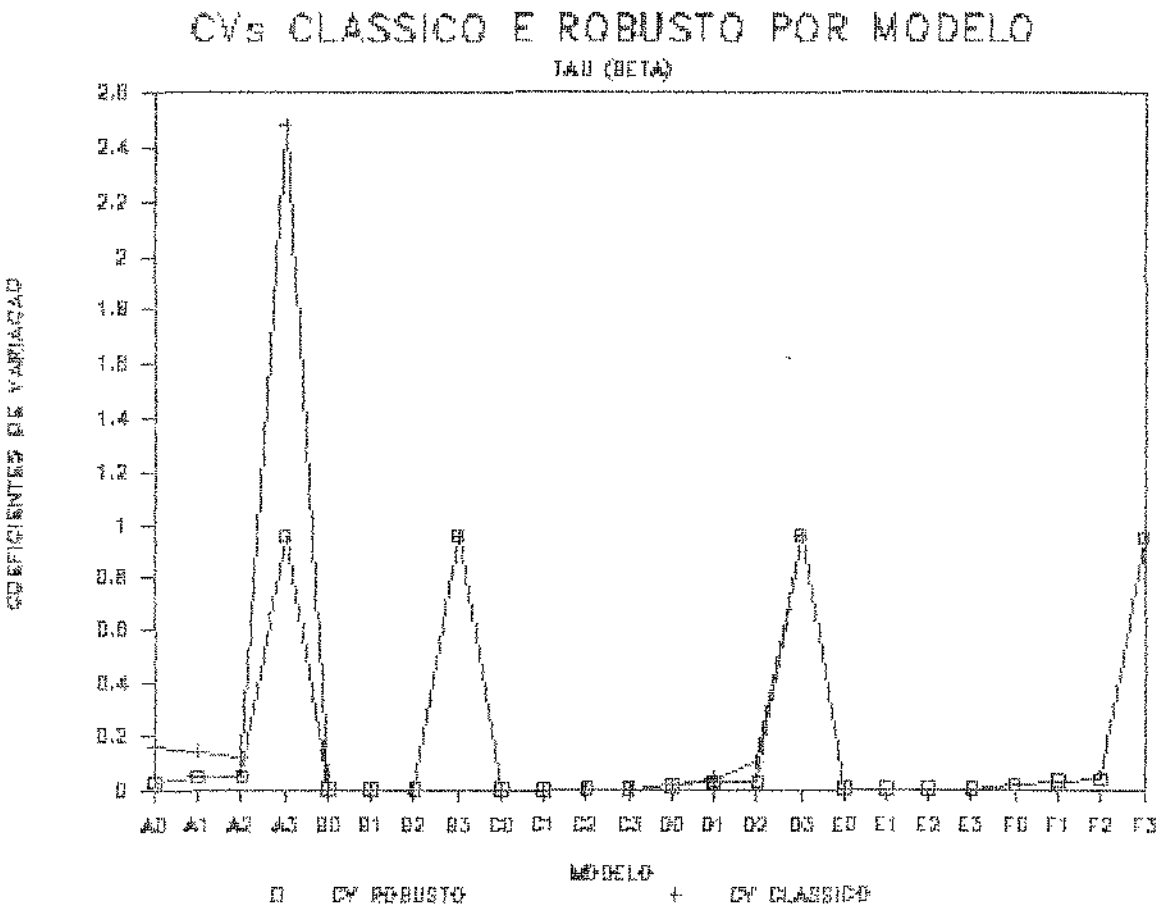


GRÁFICO 177

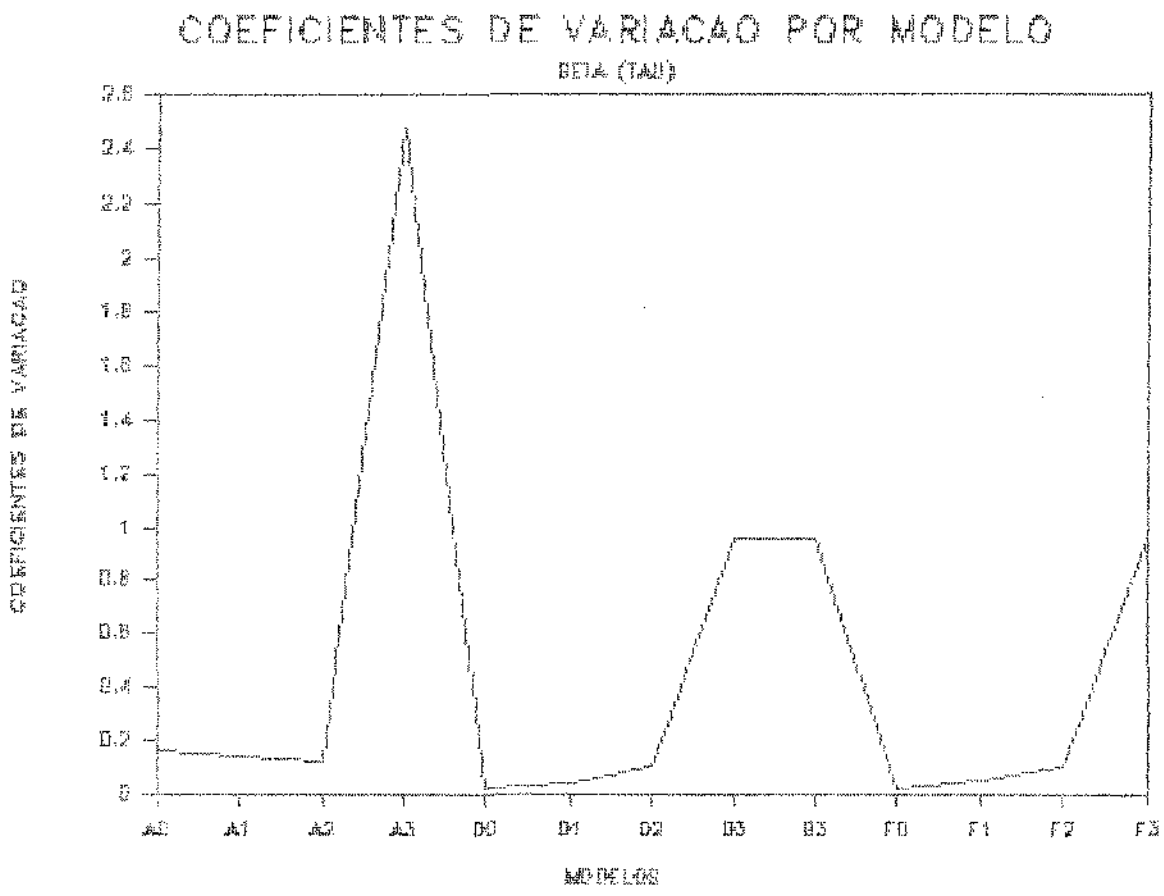

COEFICIENTES DE VARIACAO POR MODELO
BETA (TAU)
COEFICIENTES DE VARIACAO
2.6
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
A0
A1
A2
A3
B0
B1
B2
B3
B5
F0
F1
F2
F3
MODELOS

GRÁFICO 178

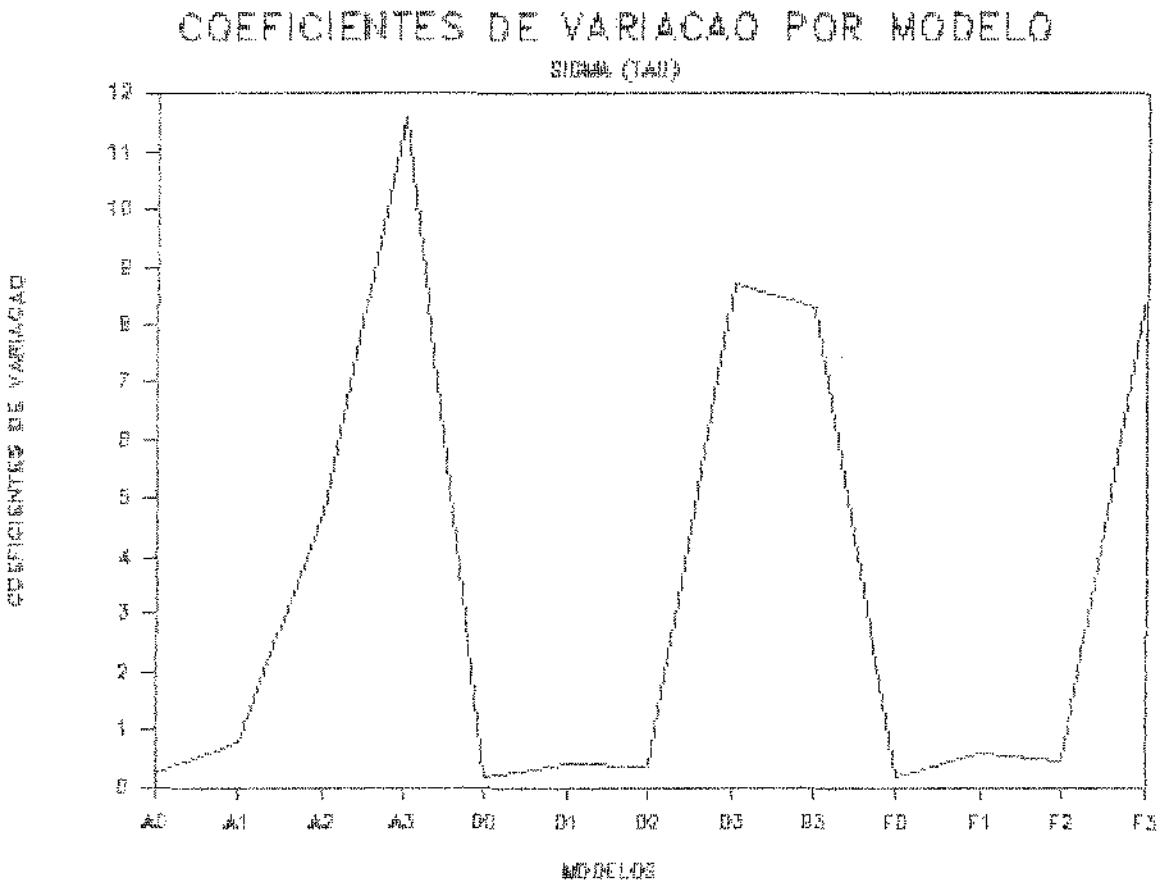

COEFICIENTES DE VARIACAO POR MODELO
SIGMA (TAU)
COEFICIENTES DE VARIACAO
0
1
2
3
4
5
6
7
8
9
10
11
12
A0
A1
A2
A3
B0
B1
B2
B3
B3
F0
F1
F2
F3
MODELOS

GRÁFICO 179

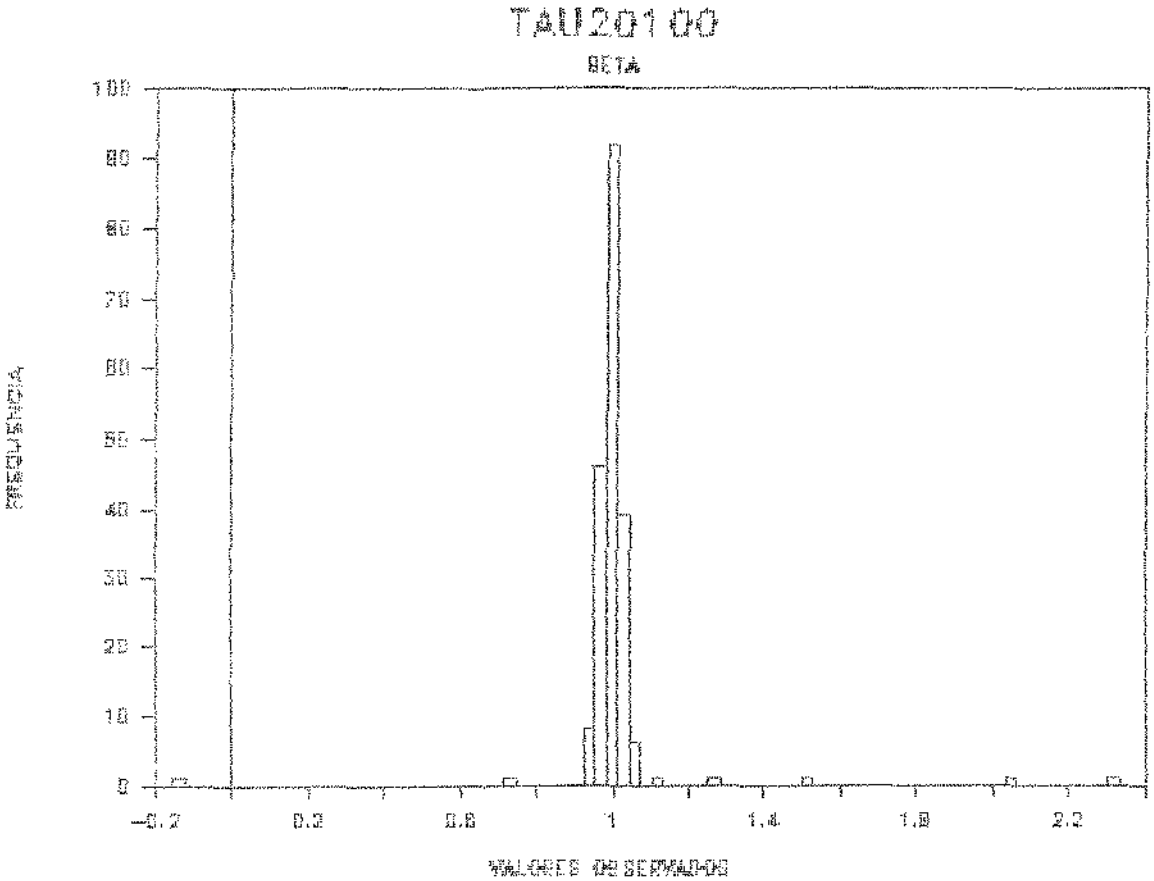


GRÁFICO 180

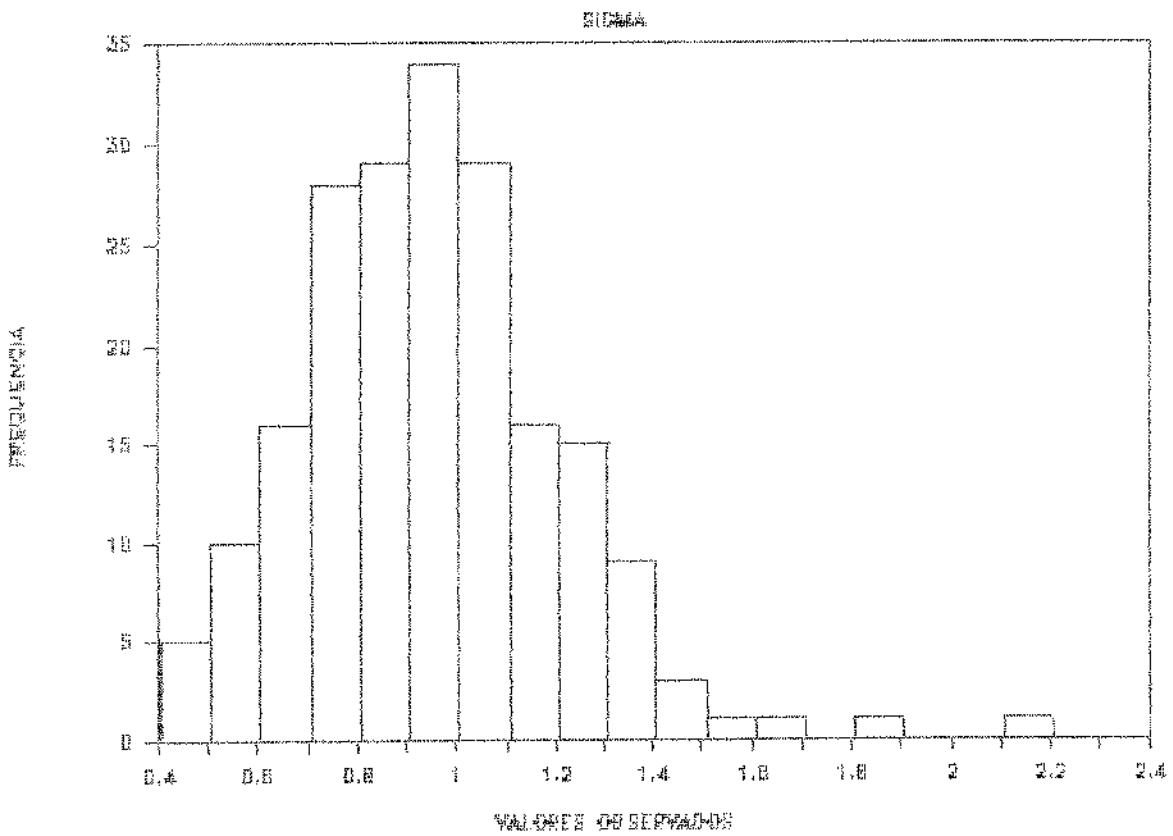

TAU20100
SIGMA
FREQUENCIA
0
5
10
15
20
25
30
35
0,4
0,6
0,8
1
1,2
1,4
1,6
1,8
2
2,2
2,4
VALORES OBSERVADOS

GRÁFICO 181

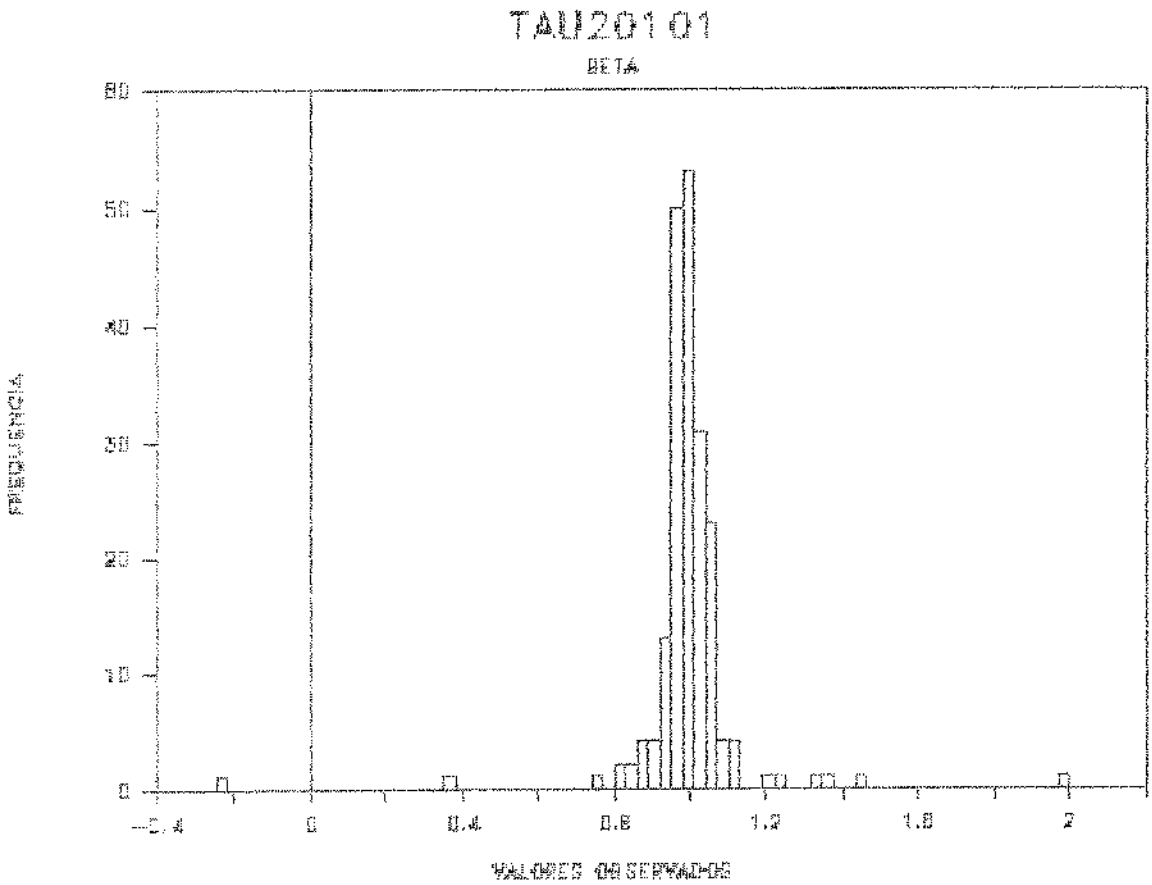

TAU2101
BETA
FRECUENCIA
60
50
40
30
20
10
0
-0.4
0
0.4
0.8
1.2
1.6
2
VALORES OBSERVADOS

GRÁFICO 182

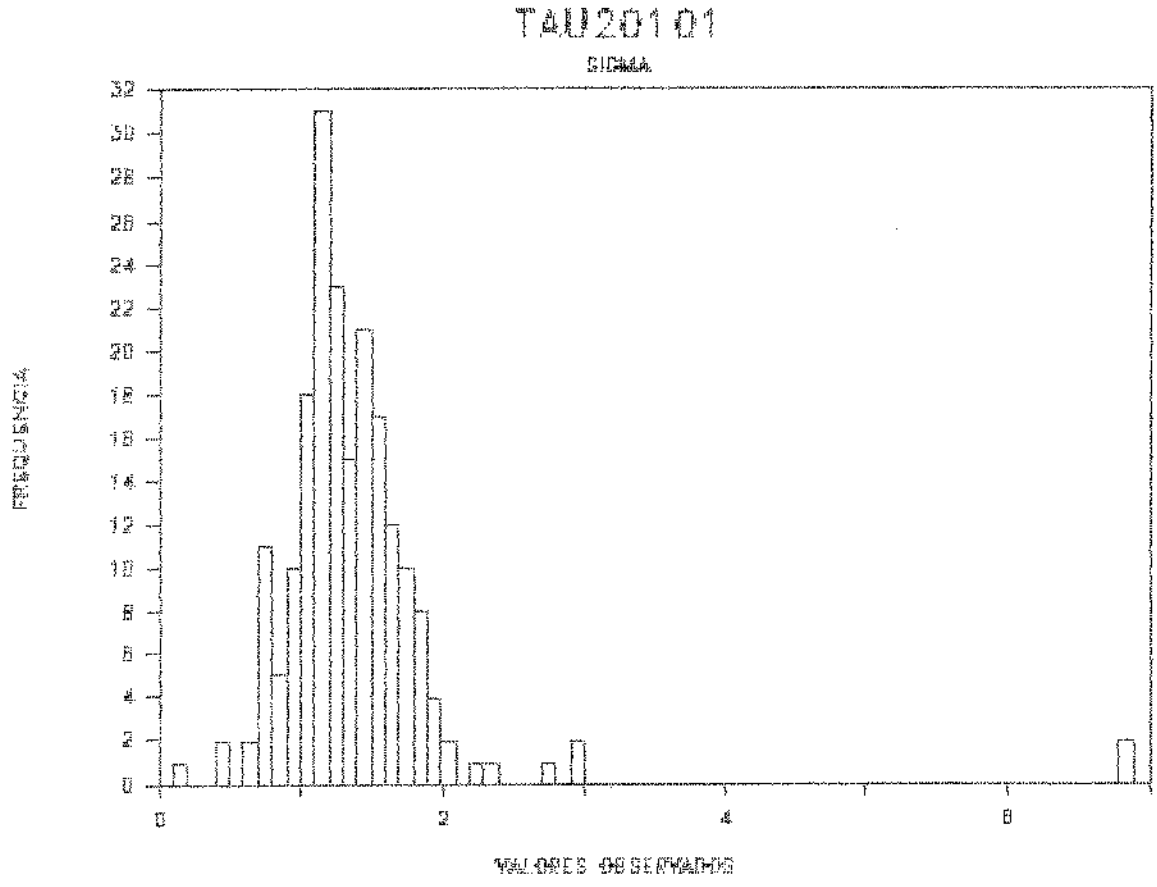


GRÁFICO 183

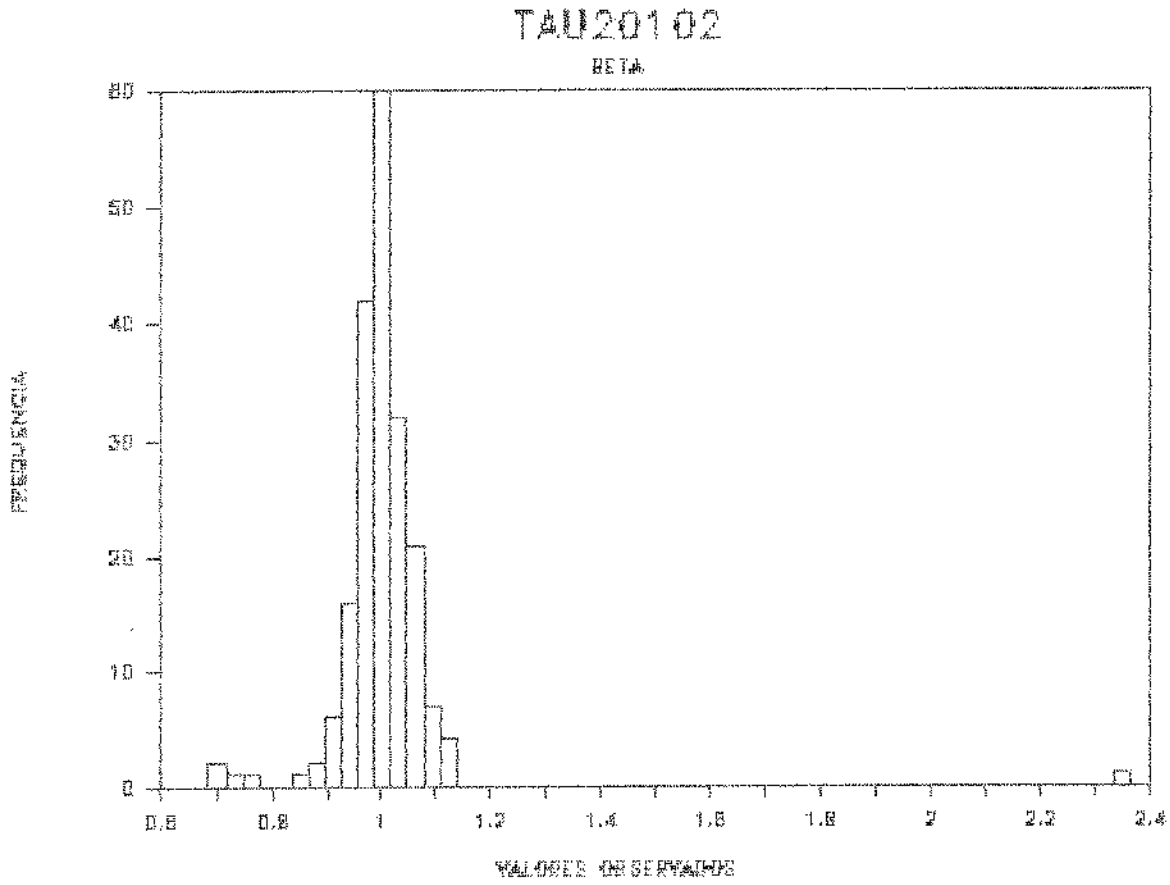


GRÁFICO 184

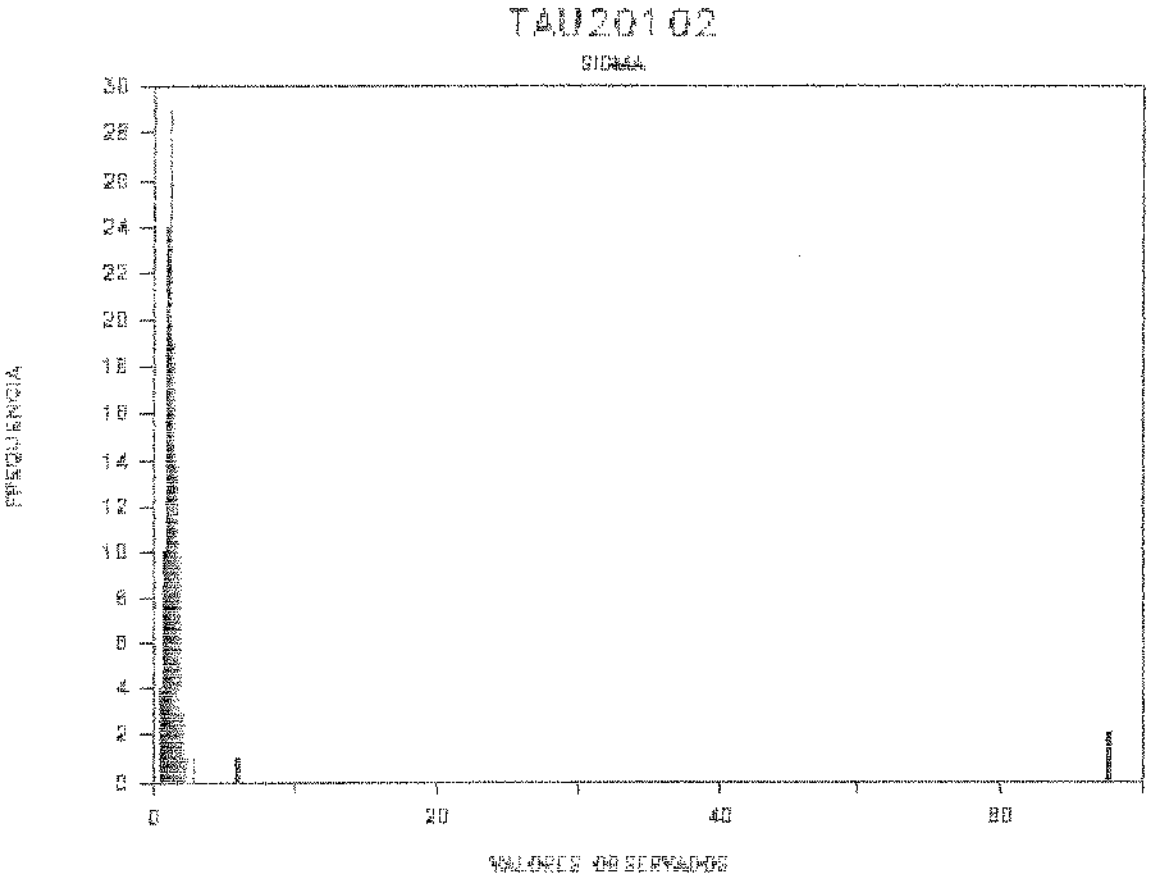


GRÁFICO 185

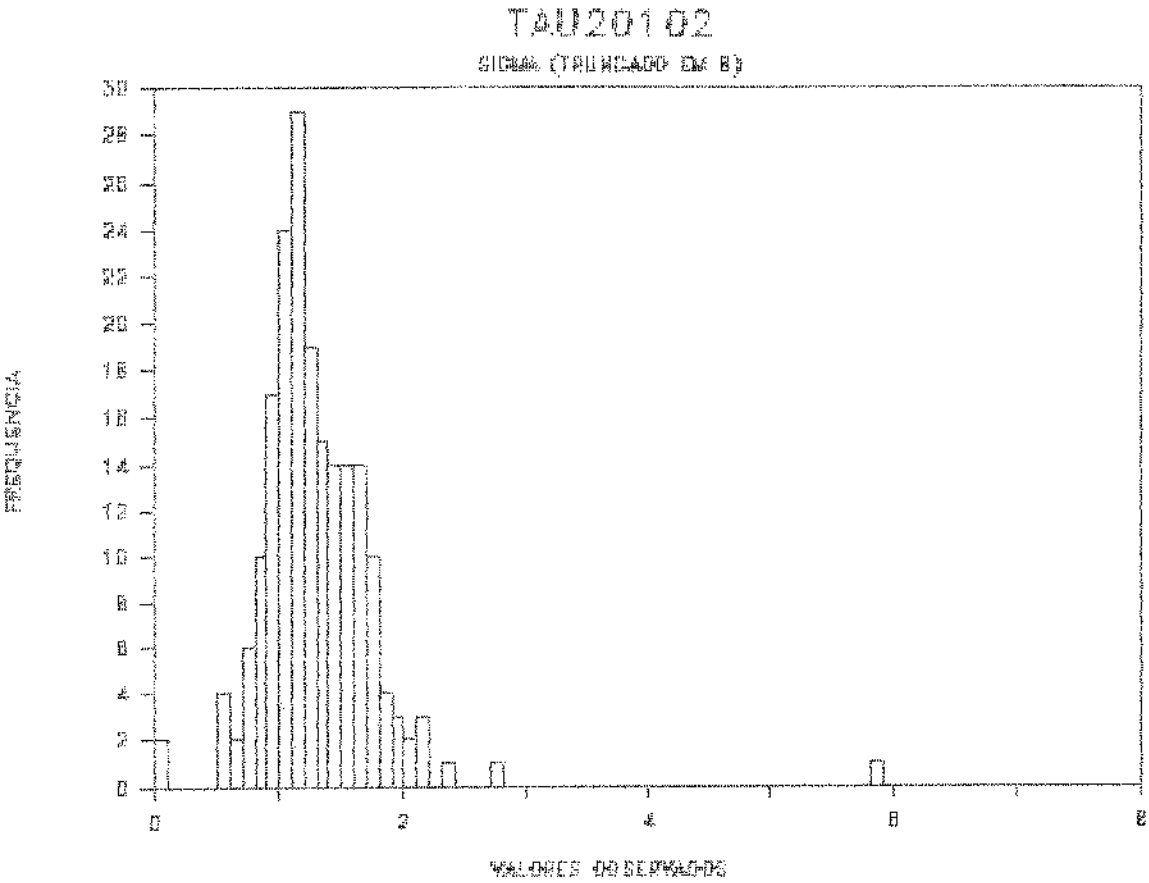

TAU20102
SIGMA (TRUNCADO EM 8)
FREQUÊNCIA
VALORES DO DESVIOS

GRÁFICO 186

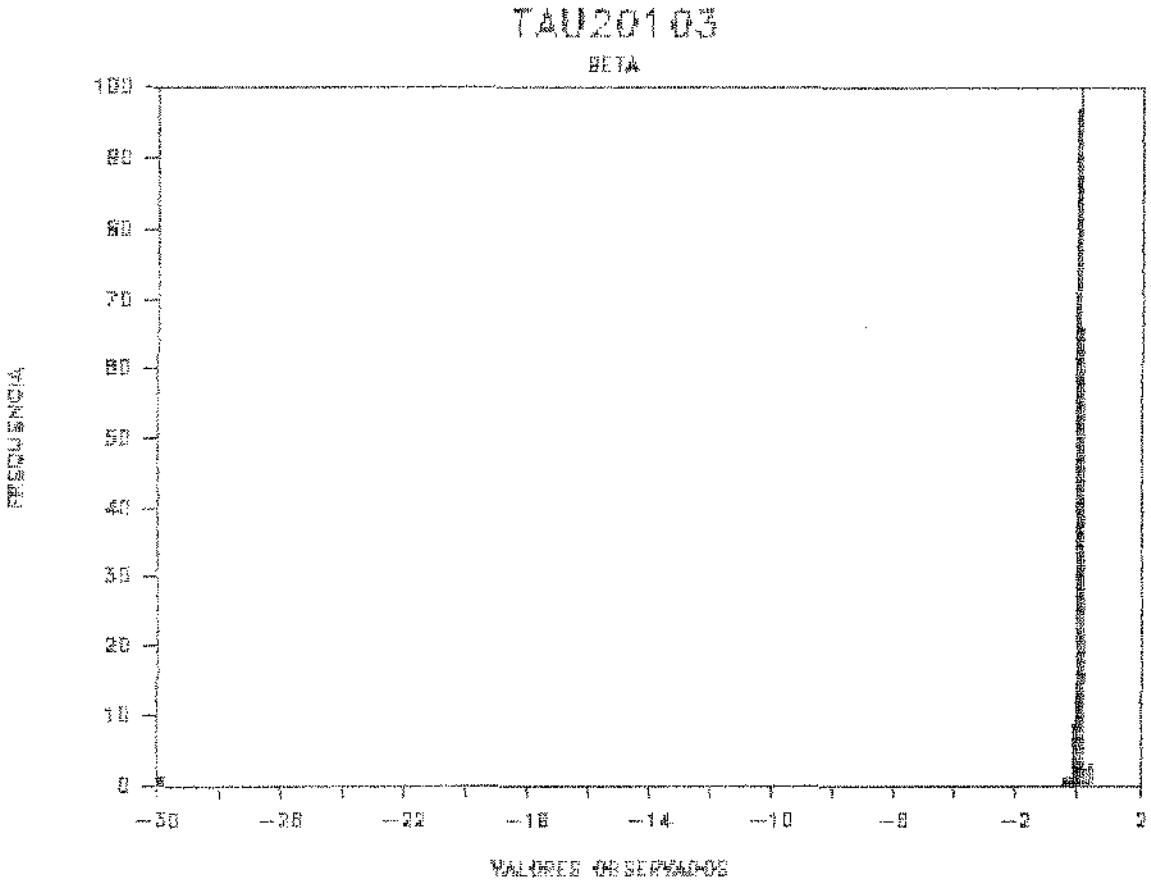

TAU20103
BETA
FREQUENCIA
100
90
80
70
60
50
40
30
20
10
0
-30
-26
-22
-18
-14
-10
-6
-2
2
VALORES OBSERVADOS

GRÁFICO 187

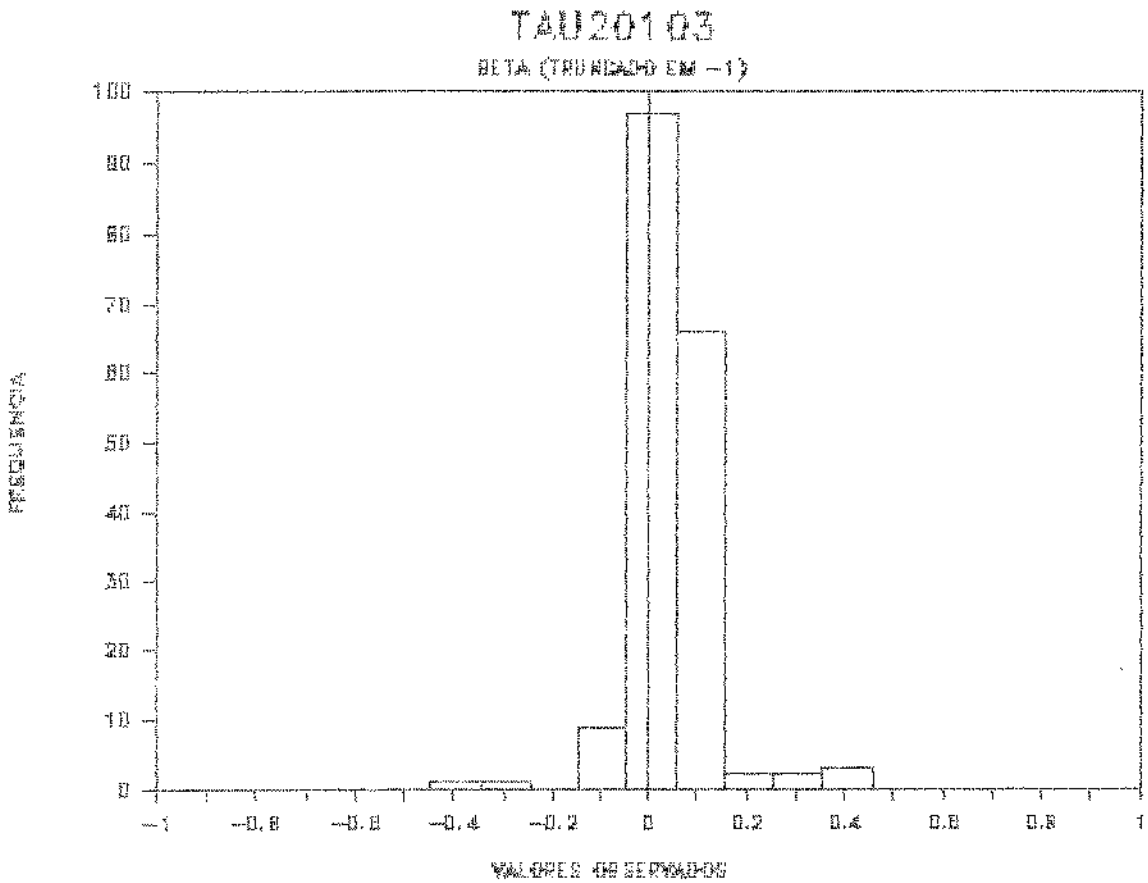

TAU20103
BETA (TRUNCADO EM -1)
FREQUÊNCIA
100
90
80
70
60
50
40
30
20
10
0
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
VALORES OBSERVADOS

GRÁFICO 188

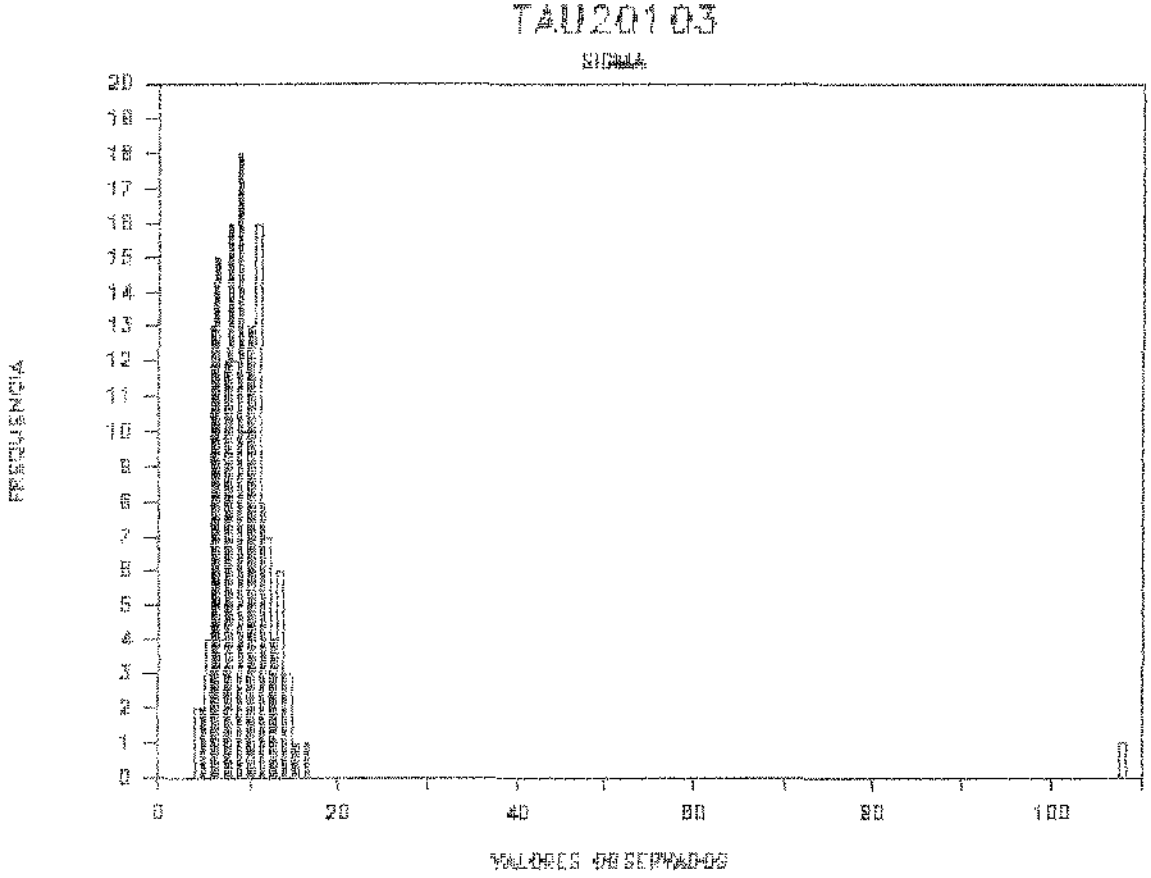

TAU20103
SIGMA
FRECUENCIA
0
20
40
60
80
100
VALORES OBSERVADOS

GRÁFICO 189

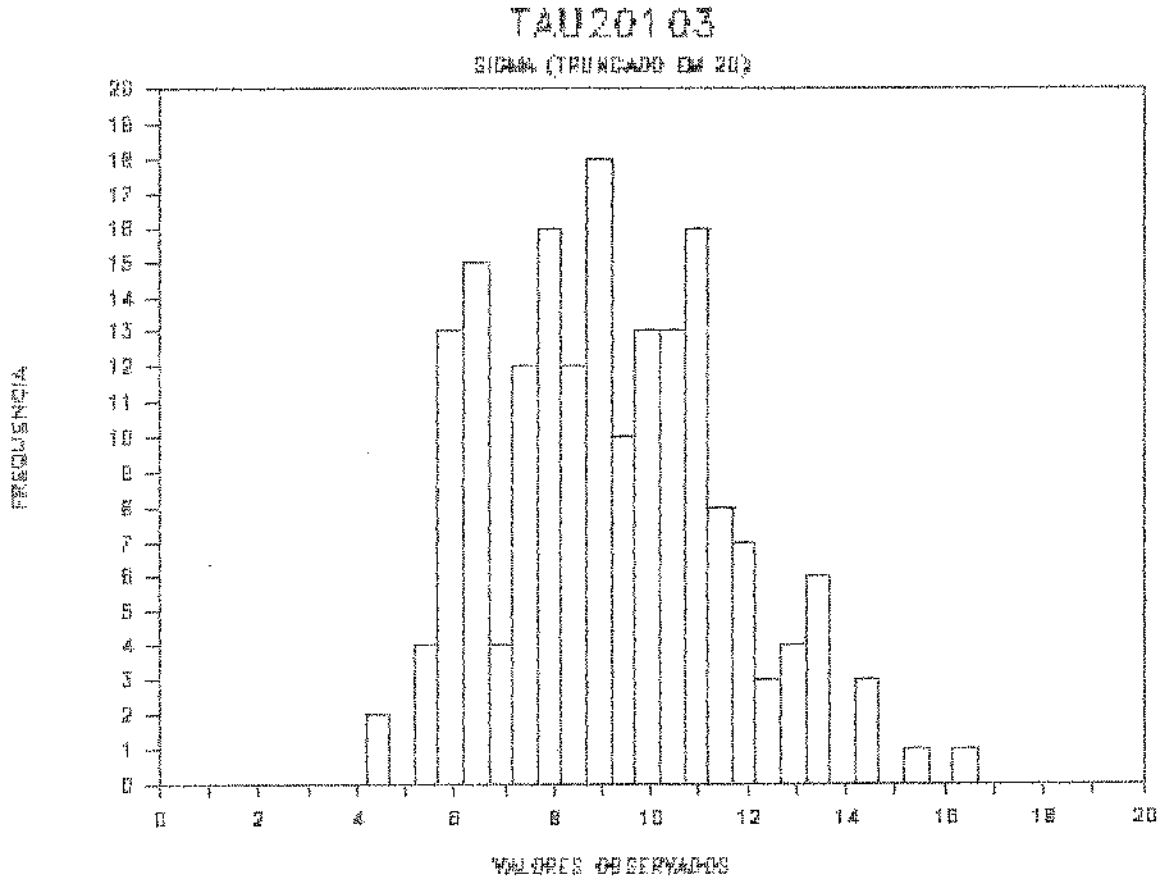

TAU201 03
SIGMA (TRUNCADO EM 20)
FREQUÊNCIA
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0
2
4
6
8
10
12
14
16
18
20
VALORES OBSERVADOS

GRÁFICO 190

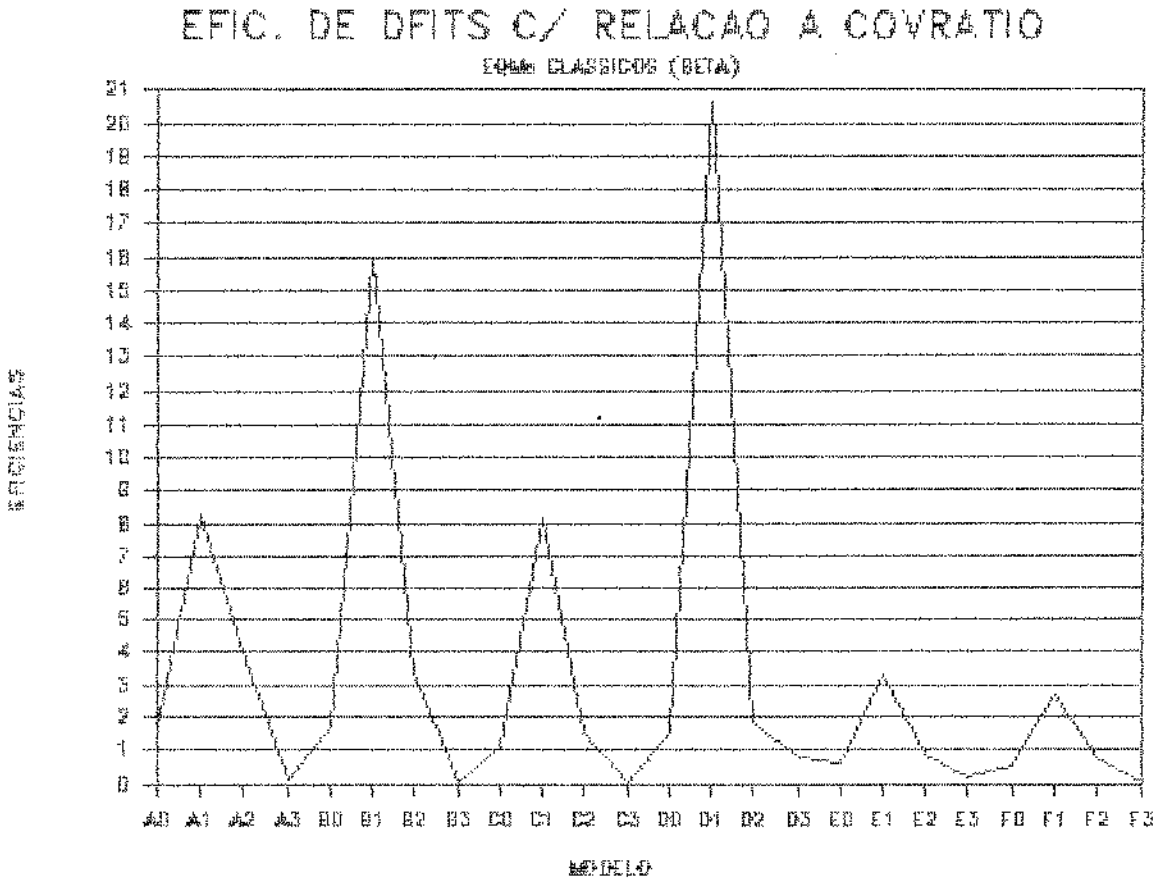

EFIC. DE DFITS C/ RELACAO A COVRATIO
EQUA CLASSICOS (BETA)
MODELO

GRÁFICO 191