USO DE TRANSFORMAÇÃO EM ANÁLISE DE VARIÂNCIA E ANÁLISE DE REGRESSÃO

ARMINDA LUCIA SIQUEIRA

DISSERTAÇÃO APRESENTADA

AO

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

DA

UNIVERSIDADE DE SÃO PAULO

PARA OBTENÇÃO DO GRAU DE MESTRE

EM

ESTATÍSTICA

ORIENTADOR: PROF. DR. CLOVIS DE ARAUJO PERES

- SÃO PAULO, SETEMBRO DE 1983 -

ERRATA

PÁGINA	LINHA	ONDE SE LÊ	LEIA-SE
SUMÁRIO	3.1	tranformação	transformação
19	↓14	$Y = \Theta + \varepsilon$	$Y = X\Theta + \varepsilon$
19	↑ 3	$\tilde{\Theta} = (X'V^{-1}X)^{-1}X'Y$	$\tilde{\Theta} = (X'V^{-1}X)^{-1}X'V^{-1}Y$
21	↓ 7	γ_2	$\gamma_2 - 3$
21	↑12	não-normalidade	normalidade
22	tabela	$(n_1 = n_2)$	(n_1 , n_2)
26	↓10	*da*	*de*
34	↑ 1	são iguais	não são iguais
43	↓ 6	logarítimica	logarítmica
45	↓ 3	Tomando-se o logarítmo natural	Assim
54	↓ 1	logarítimica	logarítmica
56	↑3,↑14	logarítimica	logarítmica
72	↓ 8	não-aditividade	interação
78	↑10	é muito utilizada	são muito utilizadas
79	↑ 2	tipo	tipos
88	↓ 4	$\sigma^2_\lambda(y)$	$\hat{\sigma}^2_\lambda\ (y)$
95	↑ 2	$\lambda = 1,0$	$\lambda = -1,0$
99	↓ 4	$\beta_3 X_2^2$	$\beta_8\ X_2^2$
107	↓ 5,6	encontra-se	encontram-se
108	↓ 4	mesófilas	coliformes
111	↓ 8	variávies	variáveis
113	↓ 9	da variância	da análise de variância
119	↓ 8	SQ_{RES} / σ^2	$SQ_{\tilde{N}\ ADIT} / \sigma^2$
120	numerador de $SQ_{\tilde{N}\ ADIT}$	$y_{i..}$	$\bar{y}_{i..}$
142	denominador de z	(0,60)	(-0,60)

A meus Pais

AGRADECIMENTOS

Ao Professor Clóvis de Araújo Peres, pela sugestão do tema, paciência com que me transmitiu seus conhecimentos, seriedade com que me orientou e pela forma atenciosa e agradável com que sempre me atendeu.

Ao Professor José Francisco Soares (Chico), colega do Departamento de Estatística da UFMG, pela disposição para muitas discussões, valiosas sugestões e pelo incentivo neste trabalho.

Ao Professor Wilton de Oliveira Bussab, pela orientação na Iniciação Científica, orientação nos trabalhos realizados no Setor de Estatística Aplicada do IME / USP e por ter despertado em mim o gosto pela Estatística Aplicada.

Aos demais Professores do Departamento de Estatística do IME / USP, pela contribuição efetiva na minha formação.

S U M Á R I O

Página

CAPÍTULO 1

INTRODUÇÃO

1.1-PRELIMINARES

A Análise de Variância e a Análise de Regressão são técnicas estatísticas extensivamente utilizadas nas mais variadas áreas de estudo.

Para o desenvolvimento teórico dessas técnicas, faz-se determinadas suposições sobre os erros do modelo. Entretanto, na prática, nem sempre tais pressuposições são verdadeiras.

A rigor, a Análise de Variância e Regressão, só poderiam ser aplicadas se as suposições estivessem completamente satisfeitas. Como o teste F da análise de variância não é sensível a alguns desvios das suposições, em muitas situações práticas em que as suposições não estão perfeitamente verificadas, tais técnicas podem ser utilizadas. Entretanto, se o desvio for acentuado, os resultados da análise podem ser bastante alterados, comprometendo assim o emprego dessas técnicas.

Ao se analisar dados que não estão de acordo com as suposições da técnica a ser utilizada, pode-se ter, basicamente, dois procedimentos : (i) buscar novos métodos de análise que se ajustam melhor aos dados; (ii) adequar os dados às suposições através de uma transformação.

Na primeira solução, modifica-se o método ou parte do

método a ser utilizado e na segunda, fixa-se as suposições exigidas pelo modelo adotado e modifica-se a variável analisada.

Transformação significa uma troca da métrica da variável original por uma outra escala. Ao invés de se trabalhar com a variável original Y, utiliza-se log Y, $\sqrt{Y}$, 1/Y, etc...

A idéia central é que, se para a variável original as suposições não são adequadas, pode existir uma transformação conveniente tal que, na nova métrica, elas sejam razoavelmente satisfeitas.

Além de corrigir os desvios das suposições, a transformação também pode ter outras aplicações, tal como linearizar ou simplificar um modelo de Regressão ou ainda eliminar a interação de um modelo de Análise de Variância.

1.2-RESUMO HISTÓRICO

Segundo um levantamento bibliográfico bastante completo, as primeiras publicações sobre o uso de transformação datam do início da década de 30. Originalmente eram investigados métodos para transformar distribuições não-normais na distribuição normal.

No período de 1936 a 1940, foram publicados artigos sobre a transformação raiz quadrada, logarítmica e sobre a transformação angular, no contexto de Análise de Variância. Surgem então muitos trabalhos aplicados em várias á-

reas de estudo, utilizando transformações. Citamos o trabalho de BEALL (1942) sobre experimentos entomológicos.

A literatura disponível sobre o assunto era quase que descritiva, até que CURTISS (1943) publica um artigo com maior formalização matemática.

O ano de 1947 foi um marco no estudo de transformações. Neste ano foram publicados três artigos na revista "Biometrics" que tiveram grande significado dentro do assunto. No primeiro, EISENHART comenta sobre as suposições usuais da Análise de Variância. No segundo, COCHRAN discute sobre algumas consequências do desvio das suposições e, no terceiro, BARTLETT apresenta o uso de transformação como uma solução para corrigir um desvio específico.

São publicados então muitos trabalhos sobre o assunto, sempre considerando cada transformação em particular, até que, em 1954, MOORE & TUKEY e ANSCOMBE & TUKEY sugerem o uso de uma família de transformações.

Em 1964, BOX & COX publicam um artigo que é considerado um marco dentro dessa nova linha que trabalha com uma família de transformações. Apresentam um critério de escolha da transformação bastante simples mas original.

Esse artigo proporcionou um grande avanço ao estudo de transformações e impulsionou o aparecimento de novos trabalhos sobre o assunto. Em muitas publicações são apresentadas modificações do método proposto por BOX & COX além de estudos adicionais. Citamos algumas dessas publicações no capítulo 5 deste trabalho.

O assunto não se esgotou e frequentemente são publicados novos trabalhos sobre o uso de transformações, sendo que, ainda existem muitos pontos a serem esclarecidos.

1.3-APRESENTAÇÃO DOS CAPÍTULOS

No capítulo 2 comentamos sobre as suposições teóricas do modelo de Análise de Variância e Análise de Regressão. Discutimos algumas causas dos desvios das suposições, como detectar e como evitar alguns desvios. Apresentamos ainda consequências da violação das suposições seguidas das soluções então adotadas.

No capítulo 3 apresentamos algumas considerações gerais sobre transformações, com o objetivo de esclarecer aspectos importantes dentro do assunto estudado. Destacamos as seções 3.3 e 3.8, cujo conteúdo praticamente não é tratado na literatura.

Os capítulos 4 e 5 referem-se ao uso de transformação com o objetivo de corrigir os desvios das suposições dos modelos de Análise de Variância e Regressão.

No capítulo 4, apresentamos o tratamento clássico do uso de transformação, discutindo sobre cada suposição separadamente. Fazemos comentários sobre as transformações usuais : logarítmica, raiz quadrada, recíproca e as transformações usadas no estudo de proporções. Ilustramos com exemplos selecionados da literatura.

No capítulo 5, apresentamos o método de determinação da

transformação,proposto por BOX & COX (1964). Ilustramos com dois exemplos extraídos desse artigo.

Finalmente, no capítulo 6, apresentamos dois exemplos de dados reais, procurando aplicar as informações contidas nos capítulos anteriores.

1.4-OBJETIVOS DESTE TRABALHO

Os principais objetivos deste trabalho são :

(i) alertar contra o uso indevido das técnicas de Análise de Variância e Análise de Regressão em algumas situações em que ocorre a violação das suposições do modelo.

(ii) apresentar o uso de transformação como uma possível solução, quando a Análise de Variância ou a Análise de Regressão são técnicas indicadas mas as suposições não estão satisfeitas.

(iii) apresentar informações importantes sobre o uso de transformação, que se encontram de forma muito dispersa na literatura.

(iv) fornecer algumas orientações de como detectar a necessidade do uso de transformação e como escolher a transformação adequada a um conjunto de dados.

(v) destacar certos cuidados que devem ser tomados na utilização de transformação, principalmente na interpretação dos resultados.

CAPÍTULO 2

DISCUSSÕES SOBRE AS SUPOSIÇÕES DO MODELO LINEAR GERAL.

2.1-AS SUPOSIÇÕES

As técnicas de Análise de Variância e Análise de Regressão são tratadas na literatura através do Modelo Linear Geral

$$Y = X\Theta + \varepsilon$$

onde,

Y : vetor das observações

X : matriz de planejamento

Θ : vetor dos parâmetros

ε : vetor dos erros

No caso de Análise de Variância, a matriz X é constituida por 0's e 1's de forma a associar cada unidade experimental ao seu respectivo grupo. Em Análise de Regressão, X é construída utilizando-se os valores das variáveis explicativas.

Para o desenvolvimento teórico dessas técnicas são feitas suposições sobre os erros do modelo. As suposições usuais sobre cada componente de $\varepsilon(\varepsilon_i)$ são as seguintes :

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

(ii) $VAR(\varepsilon_i) = \sigma^2, \quad i=1,\ldots,n$

(iii) $COV(\varepsilon_i,\varepsilon_j)=0, \quad i,j=1,\ldots,n, \; i\neq j$

Para se realizar inferências estatísticas (estimação por intervalos e teste de hipóteses), torna-se necessário acrescentar uma suposição sobre a distribuição dos erros. É usual supor que :

(iv) $\varepsilon_i \sim$ Normal, $\quad i=1,\ldots n$

Como consequência de (iv), a suposição (iii) é equivalente a :

(iii)' ε_i e ε_j são independentes, $i,j=1,\ldots,n, \; i \neq j$.

Essas quatro suposições são conhecidas como *suposições básicas* (ou *fundamentais*) do Modelo Linear Geral e podem ser resumidas por :

$$\varepsilon \sim N(0,\sigma^2 I)$$

Como consequência,

$$Y \sim N(X\theta, \sigma^2 I)$$

Uma outra suposição, relacionada à estrutura do modelo e não aos erros do modelo, refere-se à *aditividade*. Na literatura, o termo aditividade aparece com dois significa -

dos distintos. O primeiro, menos usual, caracteriza modelos de Regressão em que a lei funcional entre as variáveis envolvidadas é linear e não de outra forma, tal como multiplicativa(1). O segundo sentido da palavra, muito utilizado no contexto de transformação e adotado neste trabalho, refere-se à ausência de interação em modelos de Análise de Variância.

2.2-CAUSAS DOS DESVIOS DAS SUPOSIÇÕES

Na seção anterior, indicamos as suposições teóricas do modelo. Como já dissemos, na prática, nem sempre essas condições serão completamente preenchidas. Nesta seção comentamos sobre algumas causas dos desvios da suposição de não-correlação, de homocedasticidade e normalidade dos erros e sobre a não-aditividade do modelo.

As suposições de homogeneidade de variância, normalidade e aditividade, nem sempre podem ser controladas pelo pesquisador. Já a suposição de não-correlação dos erros, em geral, pode ser assegurada através de um esquema de aleatorização apropriado, no estágio do planejamento.

A violação da suposição de *não-correlação* ocorre mais frequentemente quando as observações são tomadas sequencialmente no tempo. Dados correlacionados exigem métodos de análise apropriados. Mesmo em situações em que é possível controlar essa suposição, se não forem tomados os devidos

---

(1) Um exemplo de um modelo multiplicativo é $E(Y) = X_1^{\beta_1} \cdot X_2^{\beta_2}$.

cuidados com a maneira pela qual o experimento é conduzido, pode aparecer correlação entre os erros para diferentes réplicas de um mesmo tratamento. Por exemplo, em experimentos agrícolas, os canteiros vizinhos tendem a ser correlacionados positivamente. Em experimentos de laboratório, as observações feitas pela mesma pessoa e quase simultaneamente, tendem a exibir esse mesmo tipo de correlação.

A *heterocedasticidade* dos erros pode aparecer de várias formas. Por exemplo, pode ser produzida por danos em alguma parte do experimento, causada por contratempos. Pode aparecer devido ao uso de material menos homogêneo em algumas réplicas, ser causado por condições de controle do experimento menos cuidadosas ou ainda pela escolha da unidade experimental.

A natureza dos tratamentos pode fornecer algumas respostas mais variáveis que outras, causando uma heterogeneidade da variância dos erros. Este é o caso de experimentos agrícolas, que consistem em se aplicar diferentes dosagens de adubo em solo ácido. A prática mostra que, em geral, as menores dosagens fornecem uniformemente baixa produção com uma pequena variância. As maiores dosagens, sendo suficientes para cobrir a acidez, fornecem boas produções com uma variância moderada enquanto que, dosagens intermediárias podem fornecer boas produções em alguns lotes e produções baixas em outros, o que causa maior variância.

Em muitos experimentos, os grupos de controle tendem a

apresentar menor ou maior variabilidade que os outros.

Em experimentos biológicos, que tem como objetivo comparar o efeito de tratamentos, a própria aplicação do tratamento pode causar uma variabilidade extra. A extensão disto pode variar com o tipo de tratamento e com o modo como ele é aplicado.

Há um outro tipo de heterocedasticidade que está associado com a não-normalidade dos dados. Aparece nos casos em que a variância é função da média. Um exemplo típico, refere-se a contagens que seguem a distribuição de Poisson, cuja variância é igual à média. Isto pode ser confirmado em um estudo sobre o número de insetos em uma certa área durante um período de vários meses : as variâncias devem ser maiores nos meses de grande infestação.

A heterocedasticidade pode ainda ser causada pela presença de observações atípicas ou espúrias ("outliers").Podemos distinguir dois tipos de "outliers" : os não-genuínos e os genuínos. O primeiro tipo refere-se a dados que foram lidos, anotados ou transcritos erradamente. Tais observações são conhecidas como erros grosseiros. Os "outliers" genuínos podem ser causados por mudanças nas condições experimentais não controláveis, devido a uma forma não correta com que o tratamento foi aplicado ou ainda por uma variabilidade inerente ao experimento.

A *normalidade* perfeita raramente ocorre em dados reais. Em muitas situações a distribuição da variável analisada é

assimétrica e mesmo se a distribuição for simétrica pode ser bem distinta da normal.

Observações referentes a variáveis discretas, que a rigor não tem distribuição normal, só poderão ser tratadas como variáveis normais sob certas condições.

Quando a análise envolve números inteiros e pequenos (cuja distribuição, em geral, é mais próxima da Poisson do que da Normal), deve-se estar de alerta para a não-normalidade.

Além do caso em que a *não-aditividade* é inerente ao modelo construído (devido a interação dos fatores principais), ela pode ser causada pela presença de "outliers" (TUKEY , 1949).

2.3-COMO DETECTAR DESVIOS DAS SUPOSIÇÕES

Existem situações em que, pela natureza do experimento, sabe-se que as suposições não serão satisfeitas. Nestes casos, os dados devem ser cuidadosamente examinados para decidir se os desvios apresentados são suficientes para justificar um tratamento especial dos dados. Mesmo se esse não for o caso, recomenda-se, por precaução, que a validade das suposições seja questionada. Como regra geral, não se deve aplicar uma técnica estatística sem antes verificar se as suposições do modelo estão razoavelmente satisfeitas. É aconselhável que uma análise estatística começe com técnicas exploratórias dos dados. Com isso, ganha-se sensibilidade e informações adicionais sobre a variável estudada.

Comentamos a seguir sobre três estratégias, apresentadas na literatura, para se detectar desvios das suposições. Não são exclusivas, uma pode complementar a outra. As duas primeiras são técnicas descritivas e por isso mesmo exigem uma certa habilidade para produzir bons resultados. A primeira, Análise de Resíduos, só pode ser aplicada após o ajuste do modelo enquanto que, através da segunda, a Análise Descritiva, pode-se detectar desvios das suposições, antes da aplicação da técnica de interesse (Análise de Variância, Análise de Regressão, etc). Finalmente, a terceira, mais formal, utiliza testes de significância.

A utilização dessas técnicas é ilustrada nos exemplos da seção 4.6, 5.3 e nas aplicações do capítulo 6.

2.3.1- Análise de Resíduos

A Análise de Resíduos é uma técnica bastante eficiente para detectar desvios das suposições. Consiste em métodos de análise gráfica, métodos numéricos, e ainda outros mistos. Além de permitir que as suposições sejam examinadas, a Análise de Resíduos pode fornecer informações adicionais sobre os dados. É recomendada como um procedimento de rotina.

Embora a Análise de Resíduos seja muito mais utilizada no contexto de Análise de Regressão, apresenta igual importância em Análise de Variância.

O resíduo é definido por $r = y - \hat{y}$, onde y é o valor observado e $\hat{y}$ é o valor previsto pelo modelo. Esta diferença

mede de certa forma a quantidade da variável estudada que não foi explicada pelo modelo. Também pode ser pensada como sendo a estimativa do erro, se o modelo é correto. Assim, se o modelo ajustado é adequado, os resíduos devem apresentar características favoráveis às suposições, ou pelo menos, não indicar evidências contra elas.

Existem vários procedimentos gráficos, envolvendo os resíduos, que potencialmente revelam *não-normalidade* e *heterocedasticidade* dos erros. Encontram-se descritos, comentados e exemplificados em ANSCOMBE & TUKEY(1963), WOODING (1969), NETER & WASSERMAN (1974, Capítulo 4), DRAPER & SMITH (1981, Capítulo 3).

A suposição de *não-correlação* dos erros, às vezes, pode ser verificada através dos resíduos mas, nem sempre isso é fácil. Uma possível indicação de correlação dos erros é a ocorrência de sequências grandes de resíduos com o mesmo sinal.

Quando as medidas são tomadas sequencialmente no tempo, pode-se construir um gráfico dos resíduos em função do tempo, para que seja examinado se existe indicação de auto-correlação dos erros.

A Análise de resíduos também pode ser utilizada para detectar "outliers". Em muitos casos, as observações correspondentes a resíduos "grandes" são "outliers".

Citamos ainda uma outra aplicação da Análise de Resíduos, que é verificar a adequação de modelos aditivos. BLISS (1967, capítulo 11) descreve um desses procedimentos com o

qual é possível verificar a presença ou não da interação e ainda detectar se a não-aditividade foi causada por algum "outlier".

Finalmente, destacamos a facilidade do uso da técnica de Análise de Resíduos através dos "pacotes" usuais (BMDP, MINITAB, SAS, SPSS, etc), especialmente nos procedimentos referentes à Análise de Regressão. No caso de Análise de Variância, os resíduos não são calculados diretamente mas, a média geral e as médias dos níveis dos fatores e das celas, necessárias para o cálculo dos resíduos, podem ser obtidas através dos "pacotes" usuais.

2.3.2- Análise Descritiva

Enquanto que a técnica de Análise de Resíduos é mais utilizada em Análise de Regressão, a Análise Descritiva dos dados é a forma mais comum de se detectar desvios das suposições, em Análise de Variância. Naturalmente também pode ser utilizada em Análise de Regressão.

Apresentamos a seguir, algumas sugestões que podem auxiliar a detectar desvios das suposições :

a) *Não-normalidade dos erros*

(i) construção do histograma : pode indicar não-normalidade se apresentar acentuada assimetria.

(ii) uso de papel de probabilidade normal : pode indicar falta de normalidade se o gráfico construído desviar acentuadamente de uma reta.

b) *Heterocedasticidade dos erros*

(i) inspeção das variâncias amostrais das caselas ou grupos analisados : pode indicar não homogeneidade de variância, especialmente se esta for acentuada. Uma forma empírica para isto é através do quociente da maior variância(desvio padrão) pela menor variância (desvio padrão); se esse valor for muito "grande", existe indicação de heterocedasticidade dos erros. Em casos de dúvida deve-se fazer um dos testes indicados na seção 2.3.3.

(ii) cálculo da amplitude de variação : pode indicar heterocedasticidade, se as amplitudes,para os diferentes grupos analisados,forem bastante distintas.

(iii) cálculo do coeficiente de variação : pode indicar heterocedasticidade, por exemplo, se todos os grupos apresentarem coeficientes de variação muito próximos, indicando que a variância cresce com a média. Por outro lado, se as médias forem próximas, coeficientes de variação muito diferentes podem indicar que as variâncias não são constantes.

c) *Não-aditividade do modelo*

(i) construção de gráficos de perfil : podem indicar a existência ou não de interação. Para verificar se a interação é causada pela presença de "outlier", tal observação deve ser retirada e refazer-se o gráfico. Se este mantiver a forma do gráfico anterior, existe indicação de que a interação é inerente ao fenômeno e que não foi causada pela presença do "outlier".

2.3.3-Testes de Significância

Uma outra forma de se verificar a ocorrência de al - gum desvio das suposições é através de testes de hipóteses, específicos a cada suposição. Existem situações em que o desvio da suposição é tão evidente que os testes de hipóteses podem ser dispensados. Em caso de dúvida, deve-se testar se a hipótese é adequada mas deve-se tomar certo cuidado com a alteração no nível de significância(NETER & WASSERMAN, 1974, capítulo 17).

A seguir, citamos testes apropriados para cada suposição do Modelo Linear Geral.

a) *Testes para detectar não-normalidade.*

Destacamos os testes clássicos que podem detectar desvio de normalidade : teste de χ^2 de Pearson (SNEDECOR & COCHRAN, 1980), teste de Kolmogorov e teste de Lilliefors (CONOVER, 1980).

Citamos ainda duas medidas importantes no estudo de normalidade : coeficiente de assimetria (γ_1) e curtose, ou coeficiente de achatamento, (γ_2) , definidas respectivamente por : $\gamma_1 = \frac{E(Y - \mu)^3}{\sigma^3}$ e $\gamma_2 = \frac{E(Y - \mu)^4}{\sigma^4}$. O valor de γ_1 para distribuições simétricas é zero. Se $\gamma_1 > 0$ a distribuição é assimétrica à direita e quando a distribuição é assimétrica à esquerda, $\gamma_1 < 0$. Se $\gamma_2 = 3$, que é o caso da normal, a distribuição é denominada mesocúrtica;

se $\gamma_2 > 3$, leptocúrtica e se $\gamma_2 < 3$, platicúrtica.

Uma outra opção para verificar desvios de normalidade é testar as hipóteses $H_o: \gamma_1 = 0$ e $H_o : \gamma_2 = 3$ (SNEDECOR & COCHRAN, 1980).

b) *Testes para detectar heterocedasticidade*

Existem vários testes de homogeneidade de variância. Para o caso de populações normais, citamos o teste de Cochran, o teste de Hartley e o teste de Bartlett, sendo que os dois primeiros são apropriados para dados balanceados. Encontram-se descritos, por exemplo, em WINER (1970) e CUNHA (1978).

Quando as populações não são normais, BOX & ANDERSEN (1955) propõem um teste aproximado para o caso em que as médias populacionais são conhecidas e para o caso mais real em que elas são desconhecidas. Esses testes são apresentados de forma didática em CUNHA (1978).

c) *Testes para detectar correlação dos erros*

Para testar correlação dos erros, WOODING(1969) sugere o teste baseado em postos proposto por Spearman, descrito em CONOVER (1980). Podemos indicar ainda o teste do sinal (CONOVER, 1980), o teste de aleatoriedade e o teste de correlação serial de Durbin-Watson, descritos em DRAPER & SMITH - (1981).

d) *Testes para detectar não-aditividade*

Nos casos em que há graus de liberdade suficiente para as partes não-aditivas do modelo, as interações podem ser testadas através da estatística F da análise de variância.

Caso contrário, deve-se usar testes apropriados. O mais usado é conhecido como *teste de Tukey para não-aditividade* e foi proposto em 1949. Tukey considerou o caso de experimentos cruzados com dois fatores fixos, sem réplicas. O teste consiste em utilizar 1 grau de liberdade do resíduo para testar a não-aditividade. Se o teste for significante, o uso de transformação pode ser recomendável. A construção do teste encontra-se no Apêndice 1.

WILK & KEMPTHORNE (1957) desenvolvem o teste de não-aditividade para planejamento em Quadrado-Latino.

2.4-CONSEQUÊNCIAS DOS DESVIOS DAS SUPOSIÇÕES

Após a escolha da técnica estatística a ser utilizada, duas questões relevantes podem ser levantadas : (i) o conjunto de dados analisados satisfaz as suposições da técnica?; (ii) Se as suposições não estiverem completamente satisfeitas, quais são as consequências? as consequências são graves?

A primeira questão foi abordada na seção anterior e agora tratamos da segunda. Essa seção não pretende ser técnica mas, tem por objetivo apresentar de forma resumida alguns resultados de estudos sobre a questão. Com isso queremos alertar contra o uso indevido das técnicas de Análise de Variância e Regressão, em algumas situações.

Um procedimento estatístico é denominado *Robusto* se não é muito sensível a desvios das suposições. Mesmo para as técnicas robustas, como é o caso do teste F da análise de variância, a precisão das inferências depende marcadamente do

grau com que as observações se ajustam às suposições. Daí a grande importância de se conhecer as consequências da violação das suposições, ou seja, saber qual a extensão do efeito dos desvios das suposições.

Em geral, a falha de uma suposição *altera o nível de significância*. Por exemplo, quando o pesquisador pensa que está testando a um nível de significância de 5%, ele pode realmente estar testando, digamos, a um nível de 8%.

O desvio de uma suposição pode produzir uma *perda de "sensibilidade"*, no sentido de que um teste mais poderoso poderia ser construído. Pode ainda causar uma *perda de precisão* dos estimadores envolvidos. Para ilustrar essa última consequência citada, suponhamos que no Modelo Linear Geral, $Y = X\theta + \varepsilon$, a forma da matriz de variância - covariância seja dada por $\Sigma = \sigma^2 V$ (onde V é uma matriz positiva definida simétrica) e não $\sigma^2 I$. Essa forma geral de Σ inclui casos da violação da homocedasticidade e da suposição de que os erros são não-correlacionados. Neste caso, para a estimação de θ, deve ser utilizado o método de mínimos quadrados generalizados (DRAPER & SMITH, 1981). O estimador obtido é dado por $\tilde{\theta} = (X' V^{-1} X)^{-1} X' V^{-1} Y$ e $Var(\tilde{\theta}) = (X' V^{-1} X)^{-1} \sigma^2$.

Se $\Sigma = V\sigma^2$, e o método de mínimos quadrados simples for utilizado, os estimadores obtidos são ainda não viesados porém, não terão variância mínima. Ou seja, se for utilizado o estimador padrão $\hat{\theta} = (X'X)^{-1} X' Y$ e não $\tilde{\theta} = (X' V^{-1} X)^{-1} X' Y$, tem-se que :

$$E(\hat{\theta}) = (X'X)^{-1} X'X \theta = \theta$$

mas

$$\text{Var}\ (\hat{\theta}) = (X'X)^{-1}\ X'\ V\ X\ (X'X)^{-1}\ \sigma^2.$$

Em geral, essa matriz fornece variâncias maiores do que as variâncias da matriz Var $(\tilde{\theta}) = (X'V^{-1}X)^{-1}\ \sigma^2$.

A seguir, apresentamos consequências específicas ao desvio de cada suposição, em alguns casos particulares. Destacamos o efeito dos desvios no nível de significância dos testes.

2.4.1-Desvio da suposição de normalidade dos erros

Resumimos alguns resultados, baseados principalmente na publicação de COCHRAN(1947) e no capítulo 10 de SCHEFFÉ (1959).

(i) O teste t bicaudal sobre uma média não é sensível à assimetria dos erros. Já o teste t monocaudal é mais vulnerável a esse tipo de desvio. Embora o coeficiente de curtose tenha algum efeito na distribuição t, em geral, esse efeito pode ser considerado desprezível.

(ii) Em testes de comparação de duas médias ("teste t"), o efeito do coeficiente de assimetria (γ_1) não nulo no nível de significância do teste é pequeno se os dois grupos são de igual tamanho, apresentam o mesmo valor de γ_1 e suas variâncias são iguais. Entretanto se uma dessas condições não for verdadeira, o efeito de γ_1 tenderá a ser aumentado e o nível de significância será alterado.

(iii) Em testes de comparação de mais de duas médias ("teste F") a não normalidade dos erros produz pequenas alterações no nível de significância do teste.

(IV) Em inferências sobre uma variância (σ^2), no caso em que a variável estudada não tem distribuição normal, o coeficiente de confiança, o nível de significância e o poder do teste dependem do coeficiente de curtose (γ_2). Se $\gamma_2 - 3$ for muito diferente de zero, o nível de significância é bastante alterado, como pode ser visto na tabela abaixo(α fixado em 5% e n grande).

$\gamma_2 - 3$	-1,5	-1,0	-0,5	0	0,5	1,0	2,0	4,0	7,0
Nível de significância	9×10^{-5}	0,006	0,024	0,050	0,080	0,11	0,17	0,26	0,36

Um resultado mais geral é que, em inferências que envolvem apenas fatores fixos, *inferências sobre médias*, o efeito do desvio da normalidade dos erros é praticamente desprezível(especialmente para amostras grandes), a não ser que o desvio seja acentuado. Segundo alguns estudos, se o nível de significância especificado for de 5% (1%) os níveis reais podem variar de aproximadamente 4 a 7% (de 0,5 a 2%), quando a distribuição não é normal.

Em inferências que envolvem fatores aleatórios, *inferências sobre variâncias*, o efeito da não-normalidade pode ter implicações sérias. Nestes casos, os estimadores das componentes de variância ainda são não-viesadas mas o coeficiente de confiança dos intervalos pode ser bastante diferente do especificado.

2.4.2-Desvio da suposição de igualdade de variância dos erros

Consideremos inicialmente o efeito da violação da suposição de igualdade de variância no caso de comparação de duas médias. Se os tamanhos das amostras são iguais ($n_1 = n_2$), *experimentos balanceados*, e as amostras são grandes, o nível de significância, calculado segundo a suposição de normalidade e igualdade de variância, é válido, mesmo se essas suposições são violadas. Entretanto se o experimento for *não-balanceado*, o nível de significância do teste de comparação de duas médias de populações normais é alterado, como pode ser visto na tabela abaixo (α fixado em 5%, σ_1^2 e σ_2^2 : variâncias populacionais).

(n_1, n_2) \ σ_1^2/σ_2^2	1	2	5	10	∞
(15,5)	0,050	0,025	0,008	0,005	0,002
(5,3)	0,050	0,038	0,031	0,030	0,031
(7,7)	0,050	0,051	0,058	0,063	0,072

No caso em que as duas amostras tem igual tamanho (7,7) e que a variância σ_1^2 é 10 vezes maior que σ_2^2 (um desvio de homogeneidade de variância relativamente grande), o nível de significância muda de 5% para 6,3%, uma alteração que na prática pode ser considerada desprezível. Entretanto, no caso de (15,5), em que n_1 é 3 vezes maior que n_2, o nível de significância é alterado de 5% para 0,5%, quando $\sigma_1^2/\sigma_2^2 = 10$.

BOX (1954a) estuda o efeito da desigualdade de variância em modelos de Análise de Variância a um fator fixo com três ou mais níveis. Apresenta um resultado numérico, que se encon

tra na tabela a seguir,para ilustrar o efeito da heteroce-dasticidade no nível de significância do teste F da análi se de variância (α fixado em 5%).

Número de níveis	Razão das variân cias dos níveis	Número de Observa ções dos níveis	Número total de observações	Nível de significância
3	1:1:1	(qualquer)	15	0,050
3	1:2:3	5 5 5	15	0,056
3	1:2:3	3 9 3	15	0,056
3	1:2:3	7 5 3	15	0,092
3	1:2:3	3 5 7	15	0,040
3	1:1:3	5 5 5	15	0,059
3	1:1:3	7 5 3	15	0,11
3	1:1:3	9 5 1	15	0,17
3	1:1:3	1 5 9	15	0,013
5	1:1:1:1:1	(qualquer)	25	0,05
5	1:1:1:1:3	5 5 5 5 5	25	0,074
5	1:1:1:1:3	9 5 5 5 1	25	0,14
5	1:1:1:1:3	1 5 5 5 9	25	0,025
7	1:1:1:1:1:1:7	3 3 3 3 3 3 3	21	0,12

Esses resultados evidenciam que :

(i) Moderados desvios da suposição de homogeneidade de va riância não afetam seriamente o nível de significância do teste, se o experimento é balanceado.

(ii) Se o experimento é não-balanceado a alteração do nível

de significância pode ser acentuada.

(iii) A discrepância entre o valor real do nível de significância e o valor fixado acentua-se quando apenas uma variância é diferente.

(iv) A alteração do nível de significância depende do número de níveis considerados.

BOX(1954b) estuda o efeito da não homogeneidade de variância que ocorre apenas no fator coluna em experimentos cruzados com dois fatores fixos, sem réplicas. Mostra que o teste de comparação de linhas é exato e propõe um teste aproximado para a comparação das colunas. Apresenta o seguinte resultado numérico (α fixado em 5%):

Número de linhas	Número de colunas	Razão das variâncias do fator coluna	Probabilidade do erro do tipo I	
			Linha	Coluna
11	3	1:2:3	0,042	0,055
5	3	1:2:3	0,043	0,056
11	3	1:1:3	0,038	0,059
5	3	1:1:3	0,039	0,060
3	5	1:1:1:1:3	0,045	0,068
3	11	1:1:.....1:3	0,049	0,071

A partir desses resultados podemos estabelecer as seguintes comparações entre o nível de significância real e o fixado :

(i) As discrepâncias nos testes de comparação de linhas e colunas não são acentuadas.

(ii) A heterogeneidade das variâncias dos níveis do fator

coluna torna o nível de significância real do teste de comparação de colunas superior ao nível fixado. Esse efeito acentua-se somente se a diferença entre as variâncias for grande.

(iii) Nos testes de comparação de linhas, o efeito aparece em direção oposta tornando o nível de significância real menor que o nível fixado.

(iv) A comparação da primeira linha da tabela anterior com a terceira e da segunda linha com a quarta indicam que o efeito da não-homogeneidade de variância acentua-se quando apenas uma variância é diferente das outras.

Como resultados mais gerais podemos dizer que, quando não há homogeneidade de variância em modelos fixos de Análise de Variância, o nível de significância do teste de comparação de médias é pouco afetado se o experimento é balanceado. Nessas condições, o procedimento de comparações múltiplas de Scheffé também é pouco alterado. Assim o uso de igual tamanho de amostras para todos os níveis do fator não só simplifica os cálculos mas também minimiza o efeito da não homogeneidade de variância no teste F.

Se os fatores forem aleatórios, a não-homogeneidade de variância dos erros pode ser bastante séria em inferências sobre as componentes de variância, mesmo se o experimento for balanceado.

2.4.3-Desvio da suposição de não-correlação dos erros.

A violação da suposição de não-correlação dos erros, em geral, produz consequências bastante sérias, para modelos fixos ou aleatórios. A correlação pode produzir um erro sistemático.

Comentaremos sobre um tipo de correlação de particular interesse prático, denominado *correlação serial*. Esse tipo de correlação pode aparecer se as observações são tomadas em intervalos de tempo igualmente espaçados.

Definimos abaixo a *correlação serial da primeira ordem*.

DEFINIÇÃO :

Seja $Y_1,\ldots,Y_n$ uma amostra aleatória e denotemos a correlação entre Y_i e Y_j por $COR(Y_i,Y_j)$. Se $COR(Y_iY_{i+1})=\rho$, $i=1,\ldots,n-1$, $COR(Y_i,Y_i)=1$, $i=1,\ldots,n$ e nos outros casos a correlação é nula, dizemos que as observações são correlacionadas serialmente e que ρ é o coeficiente de correlação serial de primeira ordem.

Consideremos a situação em que ocorre a homogeneidade de variância dos erros mas que exista uma correlação serial de primeira ordem. Então a matriz de variância-covariância (Σ) é dada por :

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & 0 & 0 & \ldots & 0 & 0 \\ \rho & 1 & \rho & 0 & \ldots & 0 & 0 \\ 0 & \rho & 1 & \rho & \ldots & 0 & 0 \\ \vdots & & & & & & \\ 0 & 0 & 0 & 0 & \ldots & \rho & 1 \end{bmatrix}$$

Para se assegurar que Σ seja positiva definida, ρ deve satisfazer a desigualdade :

$$-(2\cos(\frac{\pi}{n+1}))^{-1} < \rho < (2\cos(\frac{\pi}{n+1}))^{-1}$$

Pode-se mostrar que o efeito da presença de correlação serial de primeira ordem pode ser sério. A tabela a - baixo ilustra esse fato, em inferência sobre uma média(α fixado em 5% e n "grande").

ρ	-0,4	-0,3	-0,2	-0,1	0	0,1	0,2	0,3	0,4
nível de significância	10^{-5}	0,002	0,011	0,028	0,05	0,074	0,098	0,12	0,14

Esses resultados mostram que a correlação serial positiva tende a aumentar o nível de significância do teste, ocorrendo o inverso se ρ é negativo.

BOX (1954b) estuda esse mesmo tipo de correlação em modelos cruzados com dois fatores (linha e coluna). Supõe que os erros sejam correlacionados serialmente apenas dentro de um fa

tor (linha).

Um exemplo típico dessa situação é o estudo de frequências de chuvas em um certo local durante um determinado período de tempo. Se os níveis do fator linha são as 24 horas do dia e os níveis do fator coluna são os 12 meses do ano, podemos esperar que as sucessivas observações das frequências nas 24 horas, para um determinado mês, devem ser correlacionadas positivamente, já que a probabilidade de chover durante algum período de uma hora é maior ou menor, dependendo se choveu ou não na hora anterior. Entretanto, o efeito de correlação entre os meses deve ser bem menor e possivelmente desprezível.

Consideremos um experimento cruzado a dois fatores (linha e coluna), cada um com 5 níveis, e que apenas exista correlação serial de primeira ordem entre os níveis do fator linha. Na tabela abaixo podemos observar como a presença da correlação (ρ) afeta o nível de significância real em testes de comparação dos níveis do fator linha e do fator coluna (α fixado em 5%).

ρ	Nível de Significância	
	linha	coluna
-0,4	0,0003	0,059
-0,2	0,010	0,053
0,0	0,05	0,05
0,2	0,13	0,054
0,4	0,25	0,064

Esses resultados evidenciam que :

(i) A presença de correlação produz pequena mudança no nível de significância do teste de comparação das colunas mas, a mudança é notável no teste de comparação das linhas.
(ii) A correlação positiva, que é o tipo mais frequente , aumenta o nível de significância do teste da comparação das linhas, e ocorre o inverso se a correlação for negativa.

2.4.4-Desvio da suposição de aditividade do modelo

Em análise de Quadrados-Latinos e experimentos cruzados a dois fatores com uma observação por casela, supomos que o modelo é aditivo. Se isso for verdade, não há problema nenhum quanto à validade da análise. Entretanto, se houver interação entre os fatores, os estimadores dos efeitos dos fatores podem tornar-se viesados. Por exemplo, em modelos cruzados com dois fatores (A e B), sem réplicas, o quadrado médio residual (s^2) é estimador viesado de σ^2 se σ^2_{AB} for não-nulo, isto é, $E(s^2) = \sigma^2 + \sigma^2_{AB}$.

Para análise de Quadrados-Latinos com fatores fixos (linha, coluna e tratamento) a esperança do quadrado médio do resíduo contém um termo envolvendo a interação que é não negativo e pode ou não ser apreciável. Se os fatores são aleatórios e o modelo realmente não é aditivo, o quadrado médio do fator linha, do fator coluna, do tratamento e do resíduo são todos viesados (SCHEFFÉ, 1959, capítulo 5).

2.5-SOLUÇÕES ADOTADAS QUANDO OCORREM DESVIOS DAS SUPOSIÇÕES

Se for detectado o desvio de alguma suposição que não traz consequências sérias (como a não-normalidade ou a não-homogeneidade das variâncias em experimentos balanceados com fatores fixos) a análise de variância pode ser utilizada pois, as conclusões não serão praticamente alteradas. No entanto, se o efeito causado pelo desvio de uma suposição for grande, torna-se necessário o uso de medidas corretivas.

Existem medidas corretivas específicas ao desvio de cada suposição e, muitas vezes, para situações particulares.

Inicialmente vamos destacar algumas medidas utilizadas no caso em que a suposição de não-correlação dos erros não é verificada.

Já comentamos que a aleatorização, em geral, é uma forma eficiente para se garantir a não-correlação dos erros. Entretanto, existem situações em que a aleatorização não é possível, como por exemplo, estudos envolvendo o tempo ou a posição. Em muitos casos, a introdução de uma nova variável no modelo (tal como um termo linear no tempo) pode remover a correlação dos erros.

Um outro problema interessante refere-se ao planejamento de experimentos em blocos completamente casualizado (EBCC). Neste caso, as unidades experimentais não são todas homogêneas: as unidades dentro de um mesmo bloco são mais parecidas entre si do que com as unidades de blocos diferentes. A aleatorização das unidades experimentais não é completa mas esta ocorre dentro de cada bloco. Esses dois fatos, que ca

racterizam o planejamento em blocos, podem causar uma forte correlação entre as respostas das unidades experimentais dentro de um mesmo bloco.

Em geral, o modelo associado ao EBCC é exatamente o mesmo de um planejamento cruzado a dois fatores completamente casualizado (ECC) mas, as suposições usuais do modelo ECC não são realistas, quando o planejamento é em blocos.

Se os blocos são considerados aleatórios, a matriz de variância-covariância das respostas e dos erros não é a mesma já que as respostas em um mesmo bloco são correlacionadas enquanto que os erros não são.

Se os blocos são fixos, pode-se modificar o modelo ou as suposições sobre os erros do modelo para permitir a correlação entre as respostas de um mesmo bloco. Neste caso, ANDERSON (1970) sugere que as suposições sejam mantidas e apresenta uma modificação do modelo, introduzindo um termo aleatório correspondente à restrição da aleatorização causada pelos blocos. PERES (1981), mantendo o modelo usual, discute como os efeitos de tratamentos e blocos devem ser testados, em diferentes suposições que podem ser feitas sobre os erros : variâncias e covariâncias desiguais, variâncias desiguais e covariâncias iguais e ainda variâncias e covariâncias iguais.

Em muitos casos, a modificação do modelo permite a utilização de técnicas padrão. Se isso não for possível, devem ser utilizadas novas técnicas. Uma delas é uma técnica de Análise Multivariada denominada Curvas de Crescimento.

A desvantagem dessa técnica é que o número de parâmetros a ser estimado aumenta demasiadamente, exigindo assim um maior número de observações para fornecer boa precisão.

A seguir, apresentamos procedimentos mais gerais que podem ser utilizados no caso em que as suposições usuais do Modelo Linear Geral não estão satisfeitas.

a) *Método de aleatorização completa*

O teste de aleatorização completa, ou de permutação, foi proposto por Fisher em 1925.

A hipótese a ser testada é que cada unidade experimental, alocada em qualquer grupo do esquema do planejamento, fornece o mesmo valor da variável resposta. O procedimento para se testar essa hipótese consiste basicamente em se construir todas as possíveis permutações dos valores e então calcular o nível de significância do teste. Tal procedimento naturalmente não depende das suposições usuais do Modelo Linear Geral.

Para o teste é necessário o cálculo do quadrado médio do resíduo de todas as possíveis permutações. Isso, em geral, é bastante trabalhoso ou mesmo inviável. Por exemplo, em um experimento a um fator com 4 níveis, cada um com 6 réplicas, tem-se $24!/(6!6!6!6!) = 164\ 910\ 249\ 500$ maneiras possíveis de se permutar os resultados. Entretanto, uma simples modificação da teoria normal, em geral, fornece uma boa aproximação, conforme descrito em JOHNSON & LEONE (1964, capítulo 13).

SCHEFFÉ (1959, capítulo 9) descreve o teste de permutação para planejamentos em blocos e para o Quadrado-Latino

b) *Métodos não-paramétricos*

Para os testes não-paramétricos são exigidas suposições bem menos específicas do que os testes paramétricos. Por exemplo, a suposição clássica de normalidade exigida nos testes paramétricos é substituída por suposições mais gerais, tal como continuidade ou simetria da distribuição. O fato das suposições serem mais gerais faz com que os testes não-paramétricos sejam inerentemente robustos.

As suposições básicas dos testes não-paramétricos são sobre a independência das observações e sobre o tipo de escala de medida.

Sabe-se que, se as suposições estiverem satisfeitas, muitos dos testes paramétricos são mais poderosos do que os correspondentes testes não-paramétricos. Entretanto, se as suposições não estiverem satisfeitas os testes não-paramétricos podem ser mais indicados (especialmente para amostras pequenas), por serem, em geral, robustos.

Em alguns casos pode-se construir a distribuição exata da estatística do teste mas, em grande parte dos métodos não-paramétricos, a estatística do teste apresenta uma distribuição complicada, obrigando então a utilização de distribuição assintótica.

Uma crítica aos testes não-paramétricos baseados em postos é que não são utilizadas todas as informações contidas na amostra. Por outro lado, convém destacar a simplicidade de cálculos dos testes não-paramétricos.

c) *Processos Aproximados*

São testes aproximados, construídos no caso em que ocorrem desvios das suposições usuais do modelo. Consistem, em geral, na modificação das estatísticas usuais e/ou determinação de um fator de correção dos graus de liberdade dos testes usuais.

CUNHA (1978) apresenta uma monografia sobre testes de hipóteses sobre variâncias e médias nos casos de desvio da normalidade, homogeneidade de variância e independência dos erros.

PERES (1981) estuda as consequências da forte correlação existente entre as respostas das unidades experimentais de um mesmo bloco em experimentos em blocos completamente casualizados. Verifica que o teste F usual para avaliar o efeito dos tratamentos pode ser usado mesmo se houver uma covariância constante entre as respostas das unidades experimentais de um mesmo bloco. Propõe uma solução aproximada para testar o efeito dos tratamentos, quando as variâncias dos tratamentos não são iguais.

d) *Transformação*

Uma outra solução corretiva que pode ser adotada é adequar os dados às suposições através da transformação dos mesmos.

O uso de transformação é um procedimento bem geral que pode ser adotado para qualquer modelo de Análise de Variância e Regressão, em experimentos balanceados ou não e para amostras grandes e pequenas. As estatísticas utilizadas são exatamente as usuais e os graus de liberdade são mantidos. Portanto, não há perda de precisão na análise e nem perda de " sensibilidade" do teste.

No próximo capítulo apresentamos detalhes interessantes sobre transformação e nos capítulos 4 e 5 discutimos sobre o uso de transformação da variável resposta, com o objetivo de corrigir a não-normalidade, estabilizar a variância e obter aditividade do modelo.

CAPÍTULO 3

CONSIDERAÇÕES GERAIS SOBRE TRANSFORMAÇÃO

3.1-DIFERENTES OBJETIVOS DO USO DE TRANSFORMAÇÃO

Em Análise de Variância e Análise de Regressão, a transformação pode ser aplicada com um ou mais dos seguintes objetivos :

(i) linearizar o modelo

(ii) corrigir desvios das suposições do modelo

(iii) simplificar o modelo

O primeiro objetivo listado restringe-se a modelos de Regressão. A utilidade dessa aplicação do uso de transformação justifica-se pelo fato de que os procedimentos estatísticos, em geral, são mais complicados para relações não-lineares do que para as lineares. Podemos distinguir dois tipos de *modelos não-lineares nos parâmetros* : os intrinsecamente não-lineares e os linearizáveis.

Um *modelo intrinsecamente não-linear nos parâmetros* não pode ser expresso na forma $g(Y) = \beta_o X_o^* + \beta_1 X_1^* + \ldots + \beta_p X_p^* + \varepsilon$, onde X_i^* é uma variável explicativa ou uma função dessa variável que não depende de nenhum parâmetro. São exemplos :

$Y = \beta_o + \beta_1 e^{-\beta_2 X} + \varepsilon$, $Y = \beta_o + \beta_1 X + \beta_2 (\beta_3)^X + \varepsilon$.

Tais modelos são tratados no capítulo de Análise de Regressão como Modelos Não-Lineares (DRAPER & SMITH, 1981, capítulo 10).

Modelos linearizáveis, como o próprio nome indica, são os modelos não-lineares nos parâmetros mas que, após uma transformação, tornam-se lineares. São exemplos de modelos linearizáveis :

(i) $Y = \beta_o e^{\beta_1 X} \cdot \varepsilon$ (modelo exponencial)

(ii) $Y = \beta_o X^{\beta_1} \cdot \varepsilon$ (modelo de potência)

(iii) $Y = \beta_o X_1^{\beta_1} X_2^{\beta_2} \cdot \varepsilon$ (modelo multiplicativo)

(iv) $Y = \dfrac{1}{\beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$ (modelo recíproco)

(v) $Y = \dfrac{1}{1+\exp\{-(\beta_o + \beta_1 X + \varepsilon)\}}$ (modelo logístico)

Após a transformação, os correspondentes modelos são :

(i) $\log Y = \log\beta_o + \beta_1 X + \log\varepsilon = \beta_o^* + \beta_1 X + \varepsilon^*$

(ii) $\log Y = \log\beta_o + \beta_1 \log X + \log\varepsilon = \beta_o^* + \beta_1 X^* + \varepsilon^*$

(iii) $\log Y = \log\beta_o + \beta_1 \log X_1 + \beta_2 \log X_2 + \log\varepsilon = \beta_o^* + \beta_1 X_1^* + \beta_2 X_2^* + \varepsilon^*$

(iv) $\dfrac{1}{Y} = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

(v) $\log \dfrac{Y}{1 - Y} = \beta_o + \beta_1 X + \varepsilon$

O último modelo é usado em curvas de dose-resposta e a transformação correspondente é conhecida como "logit".

Nos exemplos (i), (iv) e (v) somente a variável resposta é transformada enquanto que, nos exemplos (ii) e (iii) tanto a variável resposta como as variáveis explicativas são transformadas.

A transformação das variáveis explicativas não afeta as suposições sobre os erros, quando é suposto que tais variáveis não são aleatórias. Quando a variável resposta for transformada deve ser verificado se a suposição acerca dos erros não foi violada pela transformação.

Uma atenção deve ser dada à estrutura dos erros: para se aplicar o método de mínimos quadrados, o erro deve ser aditivo na variável transformada, o que implica que em muitos casos, o erro no modelo original seja multiplicativo. Este é o caso dos exemplos (i) a (iii) e, após a transformação, a suposição a ser verificada é que $\log\varepsilon_i \sim N(o,\sigma^2)$. Estimados os parâmetros do modelo transformado através do método de mínimos quadrados, pode-se voltar ao modelo original mas os estimadores assim obtidos podem ser viesados.

Um outro objetivo do uso da transformação é corrigir a não-normalidade e estabilizar a variância. Esses tópicos são discutidos no capítulo 4 e 5 deste trabalho.

O terceiro objetivo citado do uso de transformação é a simplificação do modelo, ou seja, queremos achar, se possível, uma métrica na qual o modelo adotado seja expresso de forma mais simples. Podemos considerar dois casos : transformação

da variável resposta e transformação das variáveis explicativas.

a) *Transformação da variável resposta*

Em Análise de Variância, uma transformação conveniente da variável resposta pode tornar um modelo não-aditivo em aditivo, ou seja, pode ser que na métrica original o modelo aditivo não seja apropriado e, após a transformação, torne-se adequado. Neste caso, dizemos que a transformação eliminou a interação do modelo. Discutimos esse tópico na seção 4.4.

Em Regressão Polinomial, o objetivo da transformação da variável resposta pode ser a obtenção de um modelo de ordem mais baixa. Em muitos casos, se for adotado um modelo de Regressão de 2º grau para os dados originais, pode ser que, depois da transformação da variável resposta, um modelo de 1º grau seja perfeitamente adequado. Essa aplicação da transformação aparece muito em estudos de Engenharia onde as relações de 1ª ordem são, em geral, preferidas. Quase sempre não há equivalência matemática entre os parâmetros dos dois modelos, exceto uma equivalência aproximada que pode ser obtida pela expansão de Taylor. Assim, se ao invés de ajustarmos o modelo $Y = \beta_o + \beta_1 X + \beta_2 X^2 + \varepsilon$, ajustamos $g(Y) = \alpha_o + \alpha_1 X + \varepsilon$, a relação entre β_o, β_1 e β_2 com α_o e α_1 não é clara. Entretanto, essa simplificação pode proporcionar maior facilidade de interpretação do fenômeno estudado.

O problema consiste na escolha da transformação a ser a-

dotada. Um método que pode ser utilizado é discutido no capítulo 5 deste trabalho.

b) *Transformação das variáveis explicativas*

Através da transformação das variáveis explicativas pode ser possível tornar uma relação complicada das variáveis originais, em um modelo mais simples. Neste contexto destacamos a publicação de BOX & TIDWELL (1962) onde é apresentado um processo para decidir qual a transformação a ser utilizada. Apesar da validade original do trabalho, seu valor ficou reduzido com o aparecimento de programas (como o P3R do BMDP, versão 1981) que estimam parâmetros em modelos não-lineares relativamente complicados.

Neste trabalho, limitamo-nos ao estudo da transformação da variável resposta em Análise de Variância e Análise de Regressão.

3.2-CARACTERIZAÇÃO DAS TRANSFORMAÇÕES USUAIS

Na literatura, as transformações da variável resposta restringem-se às funções *não-lineares, monotônicas estritamente crescentes e contínuas*.

Transformações lineares, que envolvem apenas uma mudança de origem e/ou de escala, podem ser úteis para simplificação de cálculos, facilidade de interpretação ou por algum inte-

resse particular. Um exemplo clássico de transformação linear refere-se à padronização de uma variável com distribuição normal. Em Análise de Regressão, quando os valores da variável explicativa são igualmente espaçados, é muito comum uma recodificação para centralizar os dados. Essa transformação linear tem por objetivo a simplificação de alguns cálculos.

Transformações lineares, em geral, não afetam as características essenciais de uma análise estatística. Citamos como exemplo a conhecida propriedade de invariância do coefiente de correlação de Pearson com relação a mudanças lineares. O mesmo acontece com a estatística F da análise de variância: para a transformação linear $g(Y) = a + bY$ ($a \in \mathbb{R}$, $b \in \mathbb{R} - \{0\}$), as somas de quadrados ficam multiplicadas por b^2 mas o valor da estatística F é exatamente o mesmo, para a variável transformada ou não (pois b^2 é cancelado ao se fazer o quociente para o cálculo do valor dessa estatística).

Através de transformações lineares não é possível corrigir desvios das suposições. Se a variável resposta (Y) não é normal, uma transformação linear de Y também não será normal. Se ocorre heterogeneidade das variâncias na escala original, após uma transformação linear, as variâncias também não serão iguais. Portanto, as transformações de importância são as não-lineares, com as quais um certo incremento na escala original, em geral, não corresponde a um incremento igual na nova escala. Esse fato é responsável pelo efeito que a transformação tem na correção dos desvios das suposições.

Uma transformação g(Y) é denominada monotônica estrita -

mente crescente se para todo y'> y", necessariamente tem-se que g(y') > g(y"). Esse tipo de transformação não troca a relação de ordem (<,>). Essa característica é importante no contexto de transformação pois a ordenação das observações deve ser preservada. É desejável que a ordenação das médias dos grupos também seja mantida. Assim, se forem aplicados dois tratamentos, A e B, tal que,para a variável original, as médias populacionais são tais que $\mu_A^* > \mu_B^*$, gostaríamos que, após a transformação, as médias populacionais na nova escala, μ_A e μ_B, mantivessem a ordem, isto é, $\mu_A > \mu_B$. Entretanto, nem sempre isso é possível. De fato, suponhamos que para a variável original as variâncias são diferentes e, com objetivo de obter homocedasticidade foi utilizada a transformação logarítmica.Lembramos que, se $Z = \log Y \sim N(\mu,\sigma^2)$, dizemos que Y tem distribuição log-normal, $E(Y) = e^{\mu + \sigma^2/2}$ e $VAR(Y)=e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2}$. Portanto, na escala original, as médias da população A e B são respectivamente $\mu_A^* = e^{\mu_A+\sigma_A^2/2}$ e $\mu_B^* = e^{\mu_B+\sigma_B^2/2}$. O fato de que $e^{\mu_A+\sigma_A^2/2} > e^{\mu_B+\sigma_B^2/2}$ não implica necessariamente que $\mu_A > \mu_B$. Se entretanto, a transformação efetivamente tem o efeito de estabilizar a variância, isto é, $\sigma_A^2 = \sigma_B^2 = \sigma^2$, a ordenação das médias é mantida pois, se $e^{\mu_A+\sigma^2/2} > e^{\mu_B+\sigma^2/2}$, necessariamente tem-se que $\mu_A > \mu_B$.

Em muitos casos, a ordenação das médias amostrais não é mantida quando se utiliza a transformação raiz quadrada e a logarítmica.

Dentre a classe de funções não-lineares, monotônicas estritamente crescentes e contínuas, destaca-se a transformação do tipo potência ($Y^{\lambda}, \lambda \in \mathbb{R}$) e a logarítmica.

3.3-RELAÇÃO ENTRE AS HIPÓTESES DE IGUALDADE DE MÉDIAS DO MODELO ORIGINAL E TRANSFORMADO

Consideremos um experimento cujo objetivo é a comparação de K tratamentos e seja Y a variável resposta.

Se Y tem distribuição normal e as variâncias dos tratamentos são iguais, a hipótese de igualdade das médias dos tratamentos é equivalente à hipótese de que os K tratamentos produzem o mesmo efeito. Neste caso, a hipótese de igualdade de médias é uma hipótese auto-suficiente. Caso contrário, uma hipótese mais apropriada a ser testada é que as distribuições são as mesmas.

Suponhamos que a hipótese de normalidade e/ou homogeneidade de variância dos K tratamentos não esteja satisfeita. Suponhamos ainda que a hipótese de igualdade de médias seja de interesse, de forma que a comparação dos tratamentos é formalizada pela hipótese :

$$H_o^* = \mu_1^* = \mu_2^* = \ldots = \mu_k^*$$

onde $\mu_i^* = E[Y_{ij}]$, $i=1,\ldots,K$; $j=1,\ldots,n_i$.

Consideremos a transformação $Z = g(Y)$ tal que, após a transformação dos dados, as suposições do Modelo Linear Geral são verdadeiras, isto é $Z_{ij} \sim N(\mu_i, \sigma^2)$, $i=1,\ldots,K$, $j=1,\ldots,n_i$. Para a nova métrica testamos a hipótese :

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_k$$

Uma questão que pode ser levantada é se testar H_o é equivalente a testar H_o^*, no seguinte sentido : se para os dados transformados as médias forem estatisticamente iguais, isto significa que para os dados originais acontece o mesmo?

Vamos estudar a questão, considerando alguns casos :

1º Caso : Transformação Linear

$$Z = a + bY, \quad a \in \mathbb{R}, \quad b \in \mathbb{R} - \{0\}$$

Se $E(Y_{ij}) = \mu_i^*$ então $\mu_i = E(Z_{ij}) = a + b\mu_i^*$, e a equivalência de H_o e H_o^* segue imediatamente.

2º Caso : Transformação Logarítmica

$$Z = \log Y, \quad Y \in \mathbb{R}^+$$

Se $Z_{ij} \sim N(\mu_i, \sigma^2)$ então Y_{ij} tem distribuição log-normal e $E(Y_{ij}) = e^{\mu_i + \sigma^2/2}$, $i=1,\ldots,K$, $j=1,\ldots,n_i$.

Se $\mu_1 = \mu_2 = \ldots = \mu_k$ tem-se que

$$e^{\mu_1+\sigma^2/2} = e^{\mu_2+\sigma^2/2} = \ldots\ldots = e^{\mu_k+\sigma^2/2}.$$

Assim
Tomando-se o logaritmo natural é imediato mostrar que a hipótese H_o é equivalente a H_o^*.

3º Caso : Transformação Potência

$$Z = Y^{\lambda},\ \lambda \in \mathbb{R} - \{0\},\ Y \in \mathbb{R}^+$$

Se $Z_{ij} \sim N(\mu_i, \sigma^2)$ então $\mu_i^* = E(Y_{ij}) = \int_o^{\infty} y^{\frac{1}{\lambda}} \frac{e^{-\frac{(y-\mu_i)^2}{2\sigma^2}}}{\sqrt{2\Pi}\ \sigma}\, dy$,

$i=1,\ldots,k,\quad j=1,\ldots,n_i$.

Se $\mu_1 = \mu_2 = \ldots = \mu_k$, tem-se que

$$\int_o^{\infty} y^{\frac{1}{\lambda}} \frac{e^{\frac{-(y-\mu_1)^2}{2\sigma^2}}}{\sqrt{2\Pi}\,\sigma}\, dy = \int_o^{\infty} y^{\frac{1}{\lambda}} \frac{e^{\frac{-(y-\mu_2)^2}{2\sigma^2}}}{\sqrt{2\Pi}\,\sigma}\, dy = \ldots\ldots =$$

$$= \int_o^{\infty} y^{\frac{1}{\lambda}} \frac{e^{\frac{-(y-\mu_k)^2}{2\sigma^2}}}{\sqrt{2\Pi}\,\sigma}\, dy.$$

Novamente, a hipótese H_o é equivalente à hipótese H_o^*.

A questão básica desta seção é se a hipótese H_o^* é de interesse do ponto de vista prático, ou seja, se com es-

ta hipótese pode-se tirar conclusões relevantes. Se este não for o caso, as considerações acima não tem valor prático e surge então o problema da interpretação na escala original ao se utilizar uma transformação.

3.4-SITUAÇÕES EM QUE A TRANSFORMAÇÃO NÃO É INDICADA

No capítulo 2 comentamos sobre várias situações onde a transformação pode ou não ser necessária, dentro do contexto dos efeitos dos desvios das suposições. Nesta seção destacamos duas situações em que o uso de transformação não é indicado.

Em certas análises apenas um grupo (casela,bloco,etc) ou poucos grupos apresentam um comportamento diferente dos demais. Isto pode causar uma grande variabilidade entre as variâncias dos grupos e uma análise menos cuidadosa pode indicar a necessidade de uma transformação. Um caso particular dessa situação é quando apenas um dos grupos não apresenta variabilidade, isto é, todas as observações são iguais (ou quando a variância amostral é aproximadamente zero). Nestes casos, não é aconselhável o uso de transformação mesmo porque, dificilmente alguma transformação seria efetiva. O procedimento mais recomendável consiste na omissão desse grupo que poderá ser analisado separadamente. Um outro procedimento satisfatório é a subdivisão da variância dos erros (ver por exemplo, COCHRAN & COX , 1957, seção 3.6.3). Em modelos mais complexos os cálcu -

los, necessários a esse procedimento, podem ser muito trabalhosos. YATES(1936) apresenta um método para omitir apenas um fator, linha ou coluna, de um Quadrado Latino. YATES & HALE (1939) estendem o processo para o caso de dois fatores, linha e coluna.

Uma outra situação em que o uso da transformação nem sempre é recomendável é quando a heterocedasticidade dos erros é causada por apenas alguns dados. Outros detalhes são discutidos na seção 3.6.

3.5-COMO ESCOLHER A TRANSFORMAÇÃO

Detectada a necessidade de uma transformação, o problema consiste então na escolha da transformação apropriada ao conjunto de dados analisados.

O método de escolha mais elementar é por tentativas. Em muitos casos, tem-se idéia sobre possíveis transformações que poderão ser adequadas e então, o processo de escolha consiste em se aplicar essas transformações e selecionar aquela na qual as suposições do modelo adotado são mais proximamente satisfeitas. Em alguns casos esse procedimento pode ser bastante trabalhoso e dispendioso. Existe na literatura muitas outras sugestões mais práticas de como escolher a transformação a ser utilizada. Citamos a seguir algumas delas.

O procedimento a ser adotado pode depender do objetivo com que a transformação será utilizada, isto é, se o ob

jetivo é obter normalidade, homogeneidade de variância dos erros ou aditividade do modelo. Tais procedimentos são tratados no capítulo 4 deste trabalho.

No capítulo 5, apresentamos um procedimento mais geral, no sentido de que os três objetivos citados acima podem ser atingidos.

ANSCOMBE & TUKEY (1963) empregam a Análise de Resíduos não só como um procedimento para detectar desvios das suposições padrão mas, também sugerem como construir funções dos resíduos que indicam a transformação adequada.

Em muitos casos, o gráfico das observações revelará claramente a necessidade da transformação de um certo tipo (log y, 1/y, etc).

O fato de existir métodos que auxiliam a escolha da transformação, não significa que eles devem ser sempre utilizados. Às vezes, é possível que a escolha da transformação seja baseada em informações teóricas sobre o experimento, ou seja, o conhecimento sobre o fenômeno estudado pode sugerir uma transformação. SCHEFFÉ (1959, seção 10.7) apresenta um exemplo de uma reação química em que uma particular transformação, que tem interpretação dentro do fenômeno, é adequada.

Escolhida a transformação, um outro fator importante refere-se à interpretação dos dados na nova escala. Pode ser possível que um método indique que a transformação raiz quadrada é a melhor escala para se obter normalidade e homogeneidade de variâncias dos erros. Entretanto, se em ou

tra escala, digamos a logarítmica, as suposições estiverem razoavelmente satisfeitas e houver maior facilidade de interpretação, pode ser mais interessante trabalhar com a transformação logarítmica.

Os métodos desenvolvidos para se escolher uma transformação são úteis como guias mas naturalmente deve-se considerar cada caso em particular.

3.6-TRANSFORMAÇÃO E "OUTLIERS"

Os efeitos de "outliers", se não detectados e devidamente tratados são óbvios : distorcem a média do grupo a que pertencem e como aumentam o quadrado médio do resíduo, também afetam as conclusões sobre os outros grupos da análise.

Os "outliers" podem ser detectados através de gráficos (de resíduos, de probabilidade normal,etc) ou através de testes apropriados. Dentre as diversas publicações sobre o assunto destacamos o trabalho de ANSCOMBE (1960).

O procedimento mais aconselhável é examinar a causa da presença do "outlier " para julgar se ele deve ou não ser eliminado.

Se estivermos seguros que um "outlier" foi causado , por exemplo, por erro de execução do experimento, e se não houver possibilidade de retificação, podemos estar inteiramente justificados em descartar essa observação, já que o dado é obviamente incorreto. Neste caso, COCHRAN (1947) su

gere que tais pontos sejam tratados como observações perdidas.

Por outro lado, se nenhuma explicação puder ser dada a uma observação atípica, a eliminação desse dado torna - se mais questionável. Se pudermos estar seguros de que um "outlier " não foi causado por um erro mas por alguma peculiaridade da população estudada (tal como não-normalidade ou uma variabilidade inerente) então essas observações devem receber um tratamento apropriado. Vários procedimentos são sugeridos na literatura : redução ponderada da influência dessa observação na análise, análise desses valores separadamente, modificação do método de mínimos quadrados com pesos dependendo dos resíduos, utilização do método que minimiza a soma dos erros tomados em valor absoluto (NARULA & WELLINGTON, 1982), etc. Destacamos ainda uma outra solução que pode ser adotada, que consiste na transformação dos dados e então aplicação da análise de variância usual aos dados transformados. Em muitos casos, esta solução é bastante razoável pois os "outliers" podem "desaparecer" após a transformação.

Por outro lado, em alguns casos, a presença de "outliers" pode forçar a indicação da necessidade da transformação dos dados. Assim, uma análise menos cuidadosa pode levar à transformação dos dados que seria desnecessária se fossem eliminados os "outliers" da variável original. A presença de "outliers" também pode influenciar fortemente a escolha da transformação a ser utilizada. Essa questão é discutida

por ATKINSON (1982) que estuda não só o efeito de "outliers" mas também destaca o problema de pontos influentes na escolha da transformação.

3.7-VERIFICAÇÃO DA EFETIVIDADE DA TRANSFORMAÇÃO

O fato de que uma determinada transformação foi selecionada como sendo a melhor para um certo objetivo, não significa que seja necessariamente satisfatória. Após a escolha da transformação, segundo qualquer critério, é aconselhável a verificação da efetividade da transformação escolhida a fim de que seja confirmado se realmente o objetivo do uso da transformação foi atingido.

Para se avaliar a efetividade da transformação é recomendável que seja feita uma análise paralela dos dados transformados com os dados originais.

Na seção 2.3 apresentamos como se pode detectar desvios das suposições. Agora sugerimos que os mesmos procedimentos sejam utilizados para a verificação da efetividade da transformação adotada, ou seja, destacamos a importância de se analisar os resíduos do modelo transformado e/ou uma análise descritiva dos dados na nova escala.

3.8-APRESENTAÇÃO E INTERPRETAÇÃO DOS RESULTADOS

Se os dados são analisados utilizando-se uma transformação, é natural considerar-se a seguinte questão: os re-

sultados devem ser apresentados e interpretados em termos da variável transformada ou da variável original?

Na literatura existem poucos comentários sobre esta questão e não há acordo geral sobre qual deve ser o procedimento. Este é portanto um assunto que merece estudos adicionais.

Se a variável transformada tem interpretação prática, é bastante razoável que as conclusões sejam expressas na nova escala, para a qual as suposições usuais são adequadas. Quando isto não ocorre, é desejável que as conclusões sejam apresentadas em termos da variável original. Por exemplo, em um estudo sobre efeito de fertilizantes na produção de um determinado cereal, apresentar as conclusões em termos do logarítmo da produção não é muito satisfatório, já que esta variável não tem significado prático.

Discutimos abaixo alguns detalhes deste problema, considerando separadamente o caso de estimação e o de teste de hipóteses.

a) *Estimação*

(i) Estimação Pontual

Se para a variável analisada, a variância depende da média, a estimativa padrão $\bar{y}$ (a média das observações originais), não é usualmente a melhor estimativa da média populacional. Uma estimativa mais eficiente é obtida utilizando -

se os dados transformados. Este é o caso da distribuição log-normal, para a qual FINNEY (1941) mostra que os estimadores da média e da variância obtidos diretamente dos dados originais são menos eficientes do que aqueles que utilizam os dados transformados.

Um procedimento recomendado por muitos autores clássicos (KEMPTHORNE, 1952; SNEDECOR, 1956; STEEL & TORRIE, 1960, etc) consiste em se achar a estimativa $\hat{\theta}_Z$ de interesse (média, limites de confiança, etc) em termos da variável transformada $Z = g(Y)$ e então aplicar a função inversa para obter a estimativa na escala original : $\hat{\theta}_Y = g^{-1}(\hat{\theta}_Z)$.

Consideremos o problema de estimação da média populacional e seja $\bar{z}$ a média amostral na variável transformada. Na coluna 2 do quadro abaixo apresentamos as estimativas obtidas, segundo o procedimento citado, para várias transformações.

Transformação	Estimativa da média original	Estimativa da média original assintoticamente não-viesada
raiz quadrada	$\bar{z}^2$	$\bar{z}^2 + (n-1)s_z^2/n$
logarítmica	$e^{\bar{z}}$ ou $10^{\bar{z}}$	$e^{(\bar{z} + \frac{(n-1)}{2n} s_z^2)}$ ou $10^{(\bar{z} + 2{,}30 . \frac{(n-1)}{2n} s_z^2)}$
recíproca	$\frac{1}{\bar{z}}$	Não há forma explícita para o ajuste
angular (arc sen $\sqrt{Y}$)	$(\text{sen } \bar{z})^2$	$(\text{sen}\bar{z})^2 + \frac{1}{2}(1-e^{-2s_z^2}) \cos 2\bar{z}$
$Z = g(Y)$, g monotônica	$g^{-1}(\bar{z})$	$g^{-1}(\bar{z})$ + viés $(g^{-1}(\bar{z}))$

Nos casos da transformação logarítmica, a estimativa da média original, obtida segundo esse procedimento, é denominada média geométrica e no caso da transformação recíproca, média harmônica.

Os estimadores da coluna 2 do quadro anterior são viesados e nem sempre o viés é desprezível. Existem vários estudos sobre o problema do viés para estimadores do tipo $\hat{\theta}_Y = g^{-1}(\hat{\theta}_Z)$. FINNEY (1941) determina o viés para estimadores da média e da variância da distribuição log-normal. ANSCOMBE (1948) calcula o viés para o caso em que a variável tem distribuição Poison, Binomial e Binomial Negativa. NEYMAN & SCOTT (1960) derivam expressões para o viés sob a suposição geral de que a variável transformada tem distribuição normal.

Na coluna 3 do quadro anterior apresentamos os ajustes mais frequentes para remover o viés do estimador da média. Nessas expressões s_z^2 é o quadrado médio residual, obtido com os dados transformados, e n é o número de elementos utilizados no cálculo de $\bar{z}$.

Vamos ilustrar o procedimento citado acima com um exemplo.

EXEMPLO :

Um experimento de controle de insetos Pyrausta nubilalis, tem por objetivo comparar sete tratamentos. Foram obtidos os seguintes resultados (BEALL, 1942) :

TRATAMENTO	BLOCO									
	1	2	3	4	5	6	7	8	9	10
1	32	38	27	7	13	14	26	25	22	30
	18	40	39	12	19	26	30	19	18	28
2	6	23	8	4	3	18	26	27	17	19
	9	14	20	13	15	14	15	19	19	10
3	10	21	25	10	13	20	33	48	28	27
	4	21	26	4	9	14	30	18	27	18
4	2	17	11	3	10	10	26	13	22	17
	24	13	13	10	6	14	28	11	34	7
5	13	2	5	0	18	10	33	23	20	34
	17	22	23	8	14	16	26	22	15	34
6	13	10	21	4	10	8	17	15	13	16
	17	9	29	5	18	5	19	16	27	23
7	37	58	28	11	24	44	30	44	56	45
	44	71	55	20	26	27	43	52	39	58

Com o objetivo de estabilizar a variância foi aplicada a transformação logarítmica (base 10).

As médias dos tratamentos para os dados não transformados ($\bar{y}$), para os dados transformados ($\bar{z}$), a média convertida à escala original através da função inversa, sem correção do viés ($\bar{y}^*$) e com correção do viés ($\bar{y}^*_c$) encontram-se na tabela a seguir.

MÉDIAS	TRATAMENTOS						
	1	2	3	4	5	6	7
$\bar{y}$	24,2	15,0	20,3	14,6	17,8	15,2	40,6
$\bar{z}$	1,367	1,152	1,264	1,122	1,170	1,171	1,584
$\bar{y}^*=10^{\bar{z}}$	22,3	13,2	17,4	12,2	13,8	13,8	37,4
$\bar{y}_c^* = 10^{\bar{z}+0,070}$	26,4	15,7	20,6	14,6	16,4	16,4	44,1

Como era esperado, para todos os tratamentos, $\bar{y}^*$ (a média geométrica) é menor que $\bar{y}$. Na escala logarítmica, o quadrado médio residual é 0,064 e o fator de correção do viés é $2,30\frac{(n-1)}{2n} s_z^2 = 2,30.\frac{19}{40} . 0,064 = 0,070$. As médias corrigidas são muito mais próximas das médias das observações originais ($\bar{y}$), embora elas tendem a ter valores maiores que $\bar{y}$.

(ii) Estimação por intervalo :

Suponhamos que após a transformação $Z=g(Y)$ as suposições estejam satisfeitas, isto é, $Z \sim N(\mu,\sigma^2)$, e seja $\mu^* = E(Y)$. Desejamos construir o intervalo de confiança para μ^*.

Consideremos inicialmente a transformação logarítmica ($Z = \log Y$), caso em que a distribuição de Y é log-normal. Como é usual, consideremos os seguintes casos :

1º Caso : σ^2 conhecido

Seja $z_{\frac{\alpha}{2}}$ o quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição N(0,1). Temos que :

$$P\left[\bar{Z} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{Z} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right] =$$

$$= P\left[e^{\bar{Z} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \frac{\sigma^2}{2}} < e^{\mu + \frac{\sigma^2}{2}} < e^{\bar{Z} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \frac{\sigma^2}{2}}\right] = 1-\alpha$$

Portanto, um intervalo para μ^* com coeficiente de confiança $1-\alpha$, é dado por :

$$\left[e^{\bar{Z} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \frac{\sigma^2}{2}}, e^{\bar{Z} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + \frac{\sigma^2}{2}}\right]$$

2º Caso : σ^2 desconhecido

Seja $t_{n-1,\frac{\alpha}{2}}$ o quantil de ordem $1-\frac{\alpha}{2}$ da distribuição t com n-1 graus de liberdade. Temos que :

$$P\left[\bar{Z} - t_{n-1,\frac{\alpha}{2}} \frac{S_t}{\sqrt{n}} < \mu < \bar{Z} + t_{n-1,\frac{\alpha}{2}} \frac{S_t}{\sqrt{n}}\right] =$$

$$P\left[e^{\bar{Z}-t_{n-1,\frac{\alpha}{2}}\frac{S_t}{\sqrt{n}}+\frac{\sigma^2}{2}} < e^{\mu+\frac{\sigma^2}{2}} < e^{\bar{Z}+t_{n-1,\frac{\alpha}{2}}\frac{S_t}{\sqrt{n}}+\frac{\sigma^2}{2}} \right] = 1-\alpha$$

Como σ é desconhecido, um *intervalo aproximado* para μ^* com coeficiente de confiança $1-\alpha$, é dado por :

$$\left[e^{\bar{Z}-t_{n-1,\frac{\alpha}{2}}\frac{S_t}{\sqrt{n}}+\frac{S_t^2}{2}} , \; e^{\bar{Z}+t_{n-1,\frac{\alpha}{2}}\frac{S_t}{\sqrt{n}}+\frac{S_t^2}{2}} \right]$$

Consideremos agora a transformação genérica $Z = g(Y)$, g monotônica estritamente crescente, cuja expressão de $\mu^* = E(Y)$ não é conhecida.

Para construirmos um intervalo de confiança para μ, desenvolvemos $g(Y)$ em série de Taylor, em torno de μ^*, até a 1ª ordem. Teremos :

$$g(Y) \cong g(\mu^*) + g'(\mu^*)(Y-\mu^*)$$

e portanto

$$\mu = E(g(Y)) \cong g(\mu^*) + g'(\mu^*)\, E(Y-\mu^*) = g(\mu^*) \qquad (3.8.1)$$

Seja $[T_1, T_2]$ um intervalo aleatório de confiança para μ, construído da forma usual. Se o coeficiente de confiança é $1-\alpha$, teremos

$$P[\, T_1 < \mu < T_2 \,] = 1-\alpha \qquad (3.8.2)$$

Usando a aproximação (3.8.1), a relação (3.8.2) fica:

$$P[\, T_1 < g(\mu^*) < T_2 \,] = 1-\alpha$$

e portanto, $[\, g^{-1}(T_1),\ g^{-1}(T_2) \,]$ é um *intervalo de confiança aproximado* para μ^*.

Esse procedimento pode ser utilizado em problemas de previsão, através de modelos de Regressão : constrói-se um intervalo de confiança para a esperança do valor previsto na variável transformada e então aplica-se a transformação inversa para os limites do intervalo de confiança.

b) *Teste de hipóteses*

Consideremos o problema da comparação de K tratamentos. Se uma transformação for utilizada, com o objetivo de estabilizar as variâncias, testamos a hipótese de igualdade de médias para a variável transformada. Se o resultado do teste F da análise de variância for *não-significante*, as hipóteses do modelo original e do modelo transformado são equivalentes para as transformações usuais (ver seção 3.3). Neste caso, a interpretação é que em média os tratamentos são iguais mas alguns são mais instáveis que outros.

O problema maior aparece quando o resultado do teste é *significante*. Do ponto de vista teórico, os métodos de comparações múltiplas devem ser aplicados aos dados transformados pois estes obedecem às suposições básicas do Modelo

Linear Geral. Porém, do ponto de vista prático, esse procedimento pode, em algumas situações, trazer dificuldades de interpretação. Nestes casos, sugerimos que o método de comparações múltiplas seja aplicado à variável transformada e que, localizadas as significâncias e as não-significâncias, a interpretação seja dada na variável original levando-se em conta a desigualdade das variâncias.

Para as *comparações não-significantes*, a interpretação pode ser a mesma dada ao caso em que o teste F é não significante, sempre acompanhada de um comentário sobre as variâncias.

Se o método de comparações múltiplas indica que as médias transformadas, referentes a dois tratamentos, são estatisticamente diferentes, (*comparações significantes*), existe evidências de que as médias desses tratamentos também são diferentes. (ver 3.8.1).

Um outro problema, já discutido na seção 3.2 e que merece destaque dentro deste contexto, é a possível troca da ordenação das médias após a transformação.

Se o teste F da análise de variância for *significante* e não houver trocas da ordenação das médias, vale o mesmo comentário anterior sobre a interpretação dos resultados para as comparações múltiplas.

Entretanto, se a ordenação das médias mudar muito, sugerimos que as conclusões sejam tiradas de forma descritiva para os dados originais. Uma possibilidade, no caso de análise de agrupamento de muitas caselas, é formar grupos

de caselas que tenham variâncias mais próximas e, para cada grupo, aplicar a técnica de agrupamento de médias.

CAPÍTULO 4

A TRANSFORMAÇÃO DA VARIÁVEL RESPOSTA DO MODELO LINEAR GERAL

4.1-INTRODUÇÃO

No capítulo 2, discutimos sobre as suposições teóri - cas do Modelo Linear Geral e algumas consequências dos desvios das suposições. Sugerimos então a transformação da variá- vel resposta (Y) como uma solução para validar o uso das técnicas de Análise de Variância e Regressão, quando as suposições não estão completamente satisfeitas.

Neste capítulo, destacamos alguns tópicos relaciona - dos a esse assunto. Tratamos cada suposição separadamente. No capítulo seguinte apresentamos um procedimento alterna tivo mais interessante, que consiste em se procurar corri gir simultaneamente todos os desvios que eventualmente pos sam ocorrer. Entretanto, em determinadas situações o inte resse consiste em se corrigir apenas uma suposição (por e xemplo, às vezes a heterocedasticidade é o único problema a ser contornado). Além disso, o procedimento apresentado no capítulo seguinte requer um custo computacional rela- tivamente alto. Assim, em muitas situações, o enfoque desse capítulo é satisfatório e as informações aqui apresenta - das podem ser utilizadas com sucesso.

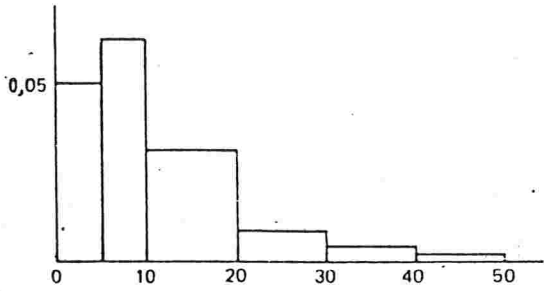
4.2-USO DE TRANSFORMAÇÃO PARA CORRIGIR NÃO-NORMALIDADE

Já comentamos no capítulo 2 que, moderados desvios da suposição de normalidade praticamente não afetam os resultados dos "testes t e F". Porém, se a distribuição apresentar uma assimetria acentuada, se o coefiente de curtose for muito diferente de zero ou ainda se ocorrer algum grande desvio na região das observações extremas, os métodos que supõem a normalidade da distribuição não devem ser utilizados. Nestes casos, em geral, é possível achar uma transformação que torne a distribuição razoavelmente simétrica e possivelmente próxima da normal.

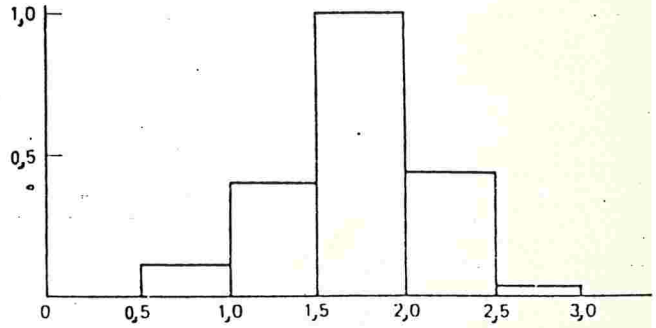
Para ilustrarmos o efeito que uma transformação pode ter em tornar uma distribuição assimétrica em uma distribuição próxima da normal, apresentamos abaixo um conjunto de dados (BHATTACHARYYA & JOHNSON,1977). O histograma desses dados, (a), exibe uma cáuda acentuada à direita. Após a transformação, $Z = \sqrt[4]{Y}$, o histograma (b), torna-se razoavelmente simétrico.

Y					$Z = \sqrt[4]{Y}$				
39,3	14,8	6,3	0,9	6,5	2,50	1,96	1,58	0,97	1,60
3,5	8,3	10,0	1,3	7,1	1,37	1,70	1,78	1,07	1,63
6,0	17,1	16,8	0,7	7,9	1,57	2,03	2,02	0,91	1,68
2,7	26,2	24,3	17,7	3,2	1,29	2,26	2,22	2,05	1,34
7,4	6,6	5,2	8,3	5,9	1,64	1,60	1,51	1,70	1,56
3,5	8,3	44,8	8,3	13,4	1,37	1,70	2,59	1,70	1,91
19,4	19,0	14,1	1,9	12,0	2,07	2,10	1,93	1,17	1,86
19,7	10,3	3,4	16,7	4,3	2,11	1,79	1,36	2,02	1,44
1,0	7,6	28,3	26,2	31,7	1,00	1,66	2,31	2,26	2,37
8,7	18,9	3,4	10,0		1,72	2,09	1,36	1,78	

(a)



(b)



O problema básico consiste na escolha da transforma - ção a ser adotada. Existem trabalhos teóricos que tratam desse assunto mas não existem métodos práticos que indi - cam a transformação adequada. Uma sugestão que pode ser u- tilizada na prática é que, as transformações $\sqrt{Y}$, $\sqrt[4]{Y}$, log Y, 1/Y tem o efeito de diminuir valores grandes enquan to que as transformações Y^3, Y^2 tendem a aumentar os valo- res grandes pequenos. Dependendo da forma do histograma pode ser que uma dessas transformações seja adequada para tornar a dis tribuição mais próxima da normal.

A distribuição conhecida como log-normal é tal que, a- pós a transformação logarítmica, tem-se uma *distribuição normal*. Além disso, CURTISS (1943) mostra o seguinte resul tado, mais geral, relacionado com a transformação logarít- mica:

TEOREMA

Uma condição necessária e suficiente para que Y tenha uma distribuição contínua com função densidade dada por :

$$f(y) = \begin{cases} \dfrac{1}{\sqrt{2\Pi \log (K^2+1)}} \dfrac{1}{Y+\alpha} \exp \dfrac{-\left(\log \dfrac{(y+\alpha)\sqrt{k^2+1)}}{\mu^*+\alpha}\right)^2}{2 \log (K^2+1)}, & y > -\alpha \\ 0, & y \leq -\alpha \end{cases}$$

tal que $VAR(Y) = (K(\mu^* + \alpha))^2$, é que a transformação $T = \log (Y+\alpha)$ tenha uma *distribuição normal*, com média $\mu = \log(\mu^*+\alpha) - \log \sqrt{K^2+1}$ e variância $\sigma^2 = \log (K^2+1)$.

O caso particular desse teorema em que $\alpha = 0, \sigma^2 = \log(K^2+1)$ e $\mu = \log \mu^* - \dfrac{\sigma^2}{2}$, corresponde exatamente à distribuição log-normal.

A importância desse teorema reside no fato de se estabelecer uma condição necessária e suficiente para que o logarítmo de uma variável não-normal seja normal, ou seja, ele mostra que toda distribuição não-normal, que após a transformação logarítmica torna-se normal, possui um padrão.

Do ponto de vista prático, esse resultado mostra que existem situações em que efetivamente a transformação logarítmica torna uma distribuição não-normal em normal.

De fato, a prática tem mostrado que, em muitos casos, a transformação logarítmica tem sucesso na normalização de variáveis.

Um outro resultado importante, discutido por CURTISS

(1943) é que *muitas transformações que tem o efeito de tornar uma distribuição não-normal em normal, também estabilizam a variância.*

Para um estudo mais detalhado sobre o assunto dessa seção, citamos as seguintes referências bibliográficas : WASOW (1956), MOORE (1957), ATKINSON (1973), HINKLEY(1975), CARROLL(1980), HERNANDEZ & JOHNSON (1980), EFRON (1981).

4.3-USO DE TRANSFORMAÇÃO PARA ESTABILIZAR A VARIÂNCIA

Se a suposição de homogeneidade de variância não estiver satisfeita , pode ser que uma transformação da variável dependente estabilize a variância.

Apresentamos, a seguir um desenvolvimento teórico,devido a BARTLETT (1947), que determina a transformação conveniente, no caso em que existe uma relação entre a média (μ) e a variância (σ^2) da variável resposta.

4.3.1-Solução Teórica

Suponhamos que a relação existente entre $\mu = E(Y)$ e $\sigma^2 = VAR\ (Y)$ seja dada por

$$\sigma^2 = f(\mu) \qquad (4.3.1)$$

Procuramos uma transformação de Y, $Z = g(Y)$, tal que VAR(Z) seja constante.

Desenvolvendo $g(Y)$ em série de Taylor, em torno de μ, com aproximação até o 1º grau, obtemos :

$$Z = g(Y) = g(\mu) + (Y - \mu)\, g'(\mu) \qquad (4.3.2)$$

Para esse grau de aproximação, temos que :

$$E(Z) = E\left[g(\mu) + (Y-\mu)\, g'(\mu)\right] = g(\mu) \qquad (4.3.3)$$

e

$$VAR(Z) = E\left[Z - E(Z)\right]^2 = E\left[(Y-\mu) g'(\mu)\right]^2 = \left[g'(\mu)\right]^2 VAR(Y) \qquad (4.3.4)$$

Substituindo (4.3.1) em (4.3.4), temos que

$$VAR(Z) = \left[g'(\mu)\right]^2 . f(\mu) = K \qquad (4.3.5)$$

onde, por hipótese, K é uma constante (positiva).

Dessa relação, segue que :

$$g'(\mu) = \sqrt{\frac{K}{f(\mu)}} \qquad (4.3.6)$$

ou ainda,

$$g(\mu) = \int \sqrt{\frac{K}{f(\mu)}}\, d\mu \qquad (4.3.7)$$

De forma mais geral, temos que

$$g(y) = \int \sqrt{\frac{K}{f(y)}}\, dy \qquad (4.3.8)$$

Determinamos então a forma genérica da transformação que estabiliza a variância.

A seguir, apresentamos alguns exemplos de aplicação desse procedimento.

EXEMPLO 1 : Seja $Y \sim$ Poisson(μ)

$$VAR(Y) = f(\mu) = \mu, \quad g(\mu) = \int \frac{\sqrt{K}}{\sqrt{\mu}}\, d\mu = K_1\sqrt{\mu} + K_2.$$

Então a transformação $Z = \sqrt{Y}$ estabiliza a variância. De fato :

$$VAR(Z) = [g'(\mu)]^2 \cdot VAR(Y) = \left(\frac{1}{2\sqrt{\mu}}\right)^2 \cdot \mu = \frac{1}{4}$$

(constante).

EXEMPLO 2 : Seja Y uma variável aleatória tal que

$$VAR(Y) = \mu^2 \text{ e } E(Y) = \mu$$

$$VAR(Y) = f(\mu) = \mu^2, \quad g(\mu) = \int \frac{\sqrt{K}}{\sqrt{\mu^2}}\, d\mu = K_1 \log \mu + K_2.$$

Então a transformação $Z = \log Y$ estabiliza a variância. De fato :

$$VAR(Z) = [g'(\mu)]^2 . VAR(Y) = \left(\frac{1}{\mu}\right)^2 \mu^2 = 1. \text{ (constante)}.$$

EXEMPLO 3 : Seja Y^* ~ Binomial (n,μ). Consideremos a proporção de sucessos $Y = Y^*/n$.

$$VAR(Y) = f(\mu) = \frac{\mu(1-\mu)}{n}, \quad g(\mu) = \int \frac{\sqrt{nK}}{\sqrt{\mu(1-\mu)}} d\mu = K_1 \text{ arcsen}\sqrt{\mu} + K_2$$

Então $Z = \text{arcsen } \sqrt{Y}$ estabiliza a variância. De fato:

$$VAR(Z) = [g'(\mu)]^2 . VAR(Y) = \left(\frac{1}{2\sqrt{\mu}(1-\mu)}\right) . \frac{\mu(1-\mu)}{n} = \frac{1}{4n}$$

(constante, se n é constante).

A tabela a seguir contém vários tipos de relação da média com a variância, que ocorrem na prática.

Uma característica de muitas distribuições não-normais é que a variância está relacionada com a média. As transformações obtidas por esse procedimento, que eliminam a relação existente entre a variância e a média, tendem a melhorar a aproximação da distribuição dos erros à distribuição normal. Este resultado está de acordo com o resultado obtido por CURTISS (1943), citado na seção anterior.

RELAÇÃO DA VARIÂNCIA EM TERMOS DA MÉDIA μ (K CONSTANTE)	TRANSFORMAÇÃO	VARIÂNCIA APROXIMADA NA NOVA ESCALA	DISTRIBUIÇÃO TÍPICA
μ	$\sqrt{Y}$ (ou $\sqrt{Y+\frac{3}{8}}$ quando os valores de Y são inteiros pequenos)	0,25	Poisson
$K^2 \mu$		$0,25\ K^2$	
$\frac{2\mu^2}{n-1}$	log Y	$\frac{2}{n-1}$	Variâncias Amostrais
$K^2 \mu^2$	log Y, log (Y+1)	K^2	Empírica
	$\log_{10}$ Y, $\log_{10}$ (Y+1)	$0,189\ K^2$	
$\frac{\mu(1-\mu)}{n}$	arc sen $\sqrt{Y}$ (radianos)	$\frac{0,25}{n}$	Binomial
	arc sen $\sqrt{Y}$ (graus)	$\frac{821}{n}$	
	(ou arc sen $\sqrt{\frac{Y+\frac{3}{8}}{n+\frac{3}{4}}}$, quando os valores de Y são inteiros pequenos)		
$K^2 \mu^2 (1-\mu)^2$	log $\left(\frac{Y}{1-Y}\right)$	K^2	Empírica
$\frac{(1-\mu^2)^2}{n-1}$	$\frac{1}{2}$ log $\left(\frac{1+Y}{1-Y}\right)$	$\frac{1}{n-3}$	Correlações Amostrais
$\mu + K^2 \mu^2$	$\frac{1}{K}$ arc senh$(K\sqrt{Y})$	0,25	Binomial Negativa
μ^4	$\frac{1}{Y}$ ou $\frac{1}{Y+K}$	1,00	Empírica

4.3.2-Solução Prática

Na prática, nem sempre a relação entre a variância (σ^2) e a média (μ) é conhecida. Se este for o caso, deve-se tentar perceber se existe algum tipo de relação, utilizando-se as médias e variâncias amostrais ($\bar{y}_i$ e s_i^2).

O gráfico de s_i^2 em função de $\bar{y}_i$, $\log s_i^2$ em função de $\log \bar{y}_i$, ou outros, pode indicar o tipo de relação existente.

4.4-USO DE TRANSFORMAÇÃO PARA OBTER ADITIVIDADE

A escolha entre o modelo aditivo e o não-aditivo, em geral, depende da informação do campo científico da pesquisa. O conhecimento que se tenha sobre o experimento, as experiências adquiridas através de pesquisas passadas, são os melhores guias de decisão entre os dois modelos. Em caso de dúvida, recomenda-se que a interação seja incluída no modelo e então testado seu efeito.

O uso de transformação para obter aditividade (ou eliminar a interação) pode ter como objetivo a simplificação do modelo, o que proporciona maior facilidade de interpretação dos fatores envolvidos no experimento. Essa utilização é razoável quando não existe interesse em se testar a interação.

No caso em que os experimentos são realizados segundo o planejamento em Quadrado-Latino, Quadrado Greco-Lati

no ou ainda experimento cruzado com uma observação por casela, arbitrariamente postula-se o modelo como aditivo (ou seja, supõe-se que as interações sejam nulas) pois não é possível testar as interações. Entretanto, pode ser que o modelo aditivo não seja adequado, e isso pode ser verificado através do Teste de Tukey para não-aditividade (descrito no apêndice 1) ou por outros testes apropriados. A mudança de escala pode eliminar a ~~não-aditividade~~ interação, tornando assim o modelo aplicável.

Apresentamos, a seguir um exemplo artificial que ilustra o efeito que a transformação pode produzir na eliminação da interação. Consideremos um experimento fatorial 2^2. Representemos por A_1 (B_1) o nível baixo ou a ausência do fator A(B) e A_2 (B_2) o nível alto ou a presença do fator A(B).

Nos quadros abaixo, (a) representa o resultado do experimento e (b) a raiz quadrada dos valores de (a) (ARMITAGE, 1977).

(a)

Níveis de A	Níveis de B	
	B_1	B_2
A_1	9	16
A_2	16	25

(b)

Níveis de A	Níveis de B	
	B_1	B_2
A_1	3	4
A_2	4	5

A situação (a) mostra um efeito de interação entre A e B mas (b) evidencia que a transformação raiz quadrada eliminou a interação.

Como nas seções 4.2 e 4.3, o problema que aparece na prática refere-se à escolha da transformação a ser utilizada. Existem alguns estudos sobre o assunto sendo que, os de maior destaque adotam os seguintes critérios para a seleção da transformação :

(a) Minimização do valor da estatística F que testa não-aditividade (TUKEY, 1949): ver apêndice 1.

(b) Minimização da razão do quadrado médio da interação pelo quadrado médio do resíduo (TUKEY, 1950).

(c) Maximização da razão do quadrado médio do tratamento pelo quadrado médio do resíduo (TUKEY, 1950).

ANSCOMBE & TUKEY (1963) apresentam um outro método para a escolha da transformação que remove a aditividade.

4.5-AS TRANSFORMAÇÕES USUAIS

Nesta seção, comentamos sobre algumas transformações muito utilizadas na prática.

Alguns comentários podem ajudar na escolha da transformação a ser adotada mas é importante deixar claro que, não existem normas gerais que garantam o sucesso na escolha da tranformação (depende muito do conjunto de dados analisados). A efetividade da transformação deve ser avaliada a -

través de uma análise de resíduos ou outro tipo de análise dos dados transformados (ver seção 3.7).

Citamos o trabalho de THÖNI (1978) que apresenta um grande levantamento bibliográfico sobre as transformações usuais.

4.5.1-A transformação logarítmica

A base 10 e a natural são as mais utilizadas por conveniência mas, qualquer base fornece conclusões equivalentes. A justificativa para isto é que, os valores de logarítmos em bases distintas diferem apenas por um fator constante e, os resultados da estatística F da análise de variância não se alteram com transformações lineares.

Quando aparecer algum valor zero ou valores negativos nas observações originais, a transformação logarítmica não pode ser utilizada diretamente. Nestes casos, é sugerida na literatura, a transformação log(Y+1) ou log(Y+K), onde K é uma constante conveniente.

O valores abaixo mostram o efeito da transformação logarítmica

y	2	20	200	2000
$\log_{10} y$	0,3	1,3	2,3	3,3

É intuitivo esperar que essa transformação tende a estabilizar a variância de grupos que na variável original tenham variâncias muito distintas. É apropriada quando, na

escala original, o desvio padrão é proporcional à média (ver tabela da seção 4.3).

Em muitas situações, a transformação logarítmica também torna distribuições não-normais mais próximas da normal (ver seção 4.2) e estudos tem mostrado que é particularmente efetiva se a assimetria é positiva (ARMITAGE , 1977).

A literatura é rica em comentários sobre situações em que a transformação logarítmica pode ser utilizada com sucesso. É extensivamente utilizada em estudos biológicos e em geral, produz excelente resultado na estabilização de variâncias para variáveis do tipo contagem, especialmente se a amplitude de variação for grande (BARTLETT, 1947).

Também é utilizada para obter relações lineares, conforme já comentado na seção 3.1 do capítulo 3. É recomendada em estudos do efeito de drogas, onde frequentemente os logarítmos das tolerâncias são normalmente distribuídos (FINNEY, 1964 e 1971).

4.5.2-A transformação raiz quadrada

Se a variável resposta (Y) apresentar valores negativos, a transformação raiz quadrada não pode ser aplicada diretamente e deve ser utilizada a transformação $\sqrt{Y+K}$, onde K é uma constante conveniente.

Se, na variável original, a variância é proporcional à média, a transformação raiz quadrada é efetiva na estabilização da variância (ver tabela da seção 4.3). A dis -

tribuição de Poisson é um caso típico dessa situação.

A literatura apresenta vários resultados empíricos relacionados à utilização da transformação raiz quadrada. Citamos alguns deles, sendo que maiores detalhes poderão ser encontrados nas referências citadas.

Se alguns valores da variável resposta (Y) são pequenos tal que as médias dos grupos comparados esteja entre 2 e 10 e, especialmente se aparecer zeros, a transformação $\sqrt{Y + 1/2}$ é recomendada (BARTLETT, 1936, 1947).

Para o caso em que algumas contagens são pequenas (menor que 10), são sugeridas as transformações $\sqrt{Y+1}$ ou $\sqrt{Y}$ + + $\sqrt{Y+1}$ (SNEDECOR & COCHRAN, 1980) ou ainda $\sqrt{Y+3/8}$ (ANSCOMBE, 1948), que são mais efetivas para estabilizar a variância.

Quando a média e a variância são inversamente proporcionais, a transformação $\sqrt{Y_{máximo}} - \sqrt{Y_{máximo} - Y}$, tem-se mostrado eficiente na estabilização da variância (ANDERSON & MCLEAN, 1974).

A transformação raiz quadrada aparece muito em estudos bacteriológicos (quando a variabilidade não é excessiva), em estudos entomológicos e em estudos sobre pragas de plantações.

4.5.3-A transformação recíproca

A transformação recíproca em geral é utilizada como $Z = 1/Y$ ou, se houver alguma observação com valor zero, $Z = 1/(Y+1)$ ou $Z = 1/(Y+K)$, K constante.

A transformação recíproca estabiliza a variância se a variância de Y for proporcional à potência quarta da média, uma forma incrível de variação (ver tabela da seção 4.3).

Para valores convenientes de K, a transformação $Z = 1/(Y+K)$ pode ter o efeito de aproximar uma distribuição não-normal à normal (THÖNI, 1978).

A transformação recíproca é muito utilizada em análise de tempo de sobrevivência de animais e plantas, em estudos de tempo de cura ou cicatrização, em muitos estudos farmacológicos e em estudos de densidade de plantas por unidade de área, em situações em que a densidade dos grupos comparados é muito variável.

4.5.4-A transformação angular (arcsen $\sqrt{Y}$)

A transformação angular é utilizada para estabilizar a variância da variável "proporção de sucessos", quando a variável "número de sucesso" segue a distribuição Binomial - (ver exemplo 3 da seção 4.3).

É recomendada especialmente quando as porcentagens dos grupos a serem comparados cobrem uma grande amplitude de variação. Se todas as porcentagens variarem entre 30% a 70%, a transformação angular, em geral, não é necessária pois o produto p(1-p)(onde p é a probabilidade de "sucesso") varia pouco nesse intervalo e então a variância se mantem razoavelmente constante.

A variância na nova escala depende de n, o número de observações em que a proporção é baseada (ver tabela da seção 4.3). Assim, somente se o experimento for balanceado ou quase balanceado, é que a condição de homogeneidade de variância será atingida.

Quando n < 50, o efeito de estabilização da variância pode ser melhorado, substituindo-se as proporções iguais a 0 e 1 por 1/4n e 1-1/4n, respectivamente (BARTLETT, 1947).

Em estudo de *porcentagens* em que o denominador não é fixo mas sim uma variável aleatória, a transformação angular não é apropriada pelo desenvolvimento de BARTLETT, segundo o qual esta transformação foi obtida para dados que seguem a distribuição binomial. Outras transformações, como a raiz quadrada e a logarítmica, podem ser eficientes na estabilização da variância.

A principal utilização da transformação angular refere-se à análise de variância aplicada a dados biológicos de resposta quantal. Neste caso também são muito utilizadas as trasformações "Probit" e "Logit" mas a transformação angular estabiliza a variância, enquanto que estas duas podem não estabilizar.

Tabelas da transformação angular podem ser encontradas em FISHER & YATES (1971) em SNEDECOR & COCHRAN (1980). Os valores dessa transformação também podem ser obtidos facilmente através das funções ASIN e SQRT, disponíveis nos "pacotes" BMDP e MINITAB. O conjunto de funções do SPSS não inclui a função arco-seno.

4.5.5-Outras transformações

São citadas na literatura muitos outros tipos de transformações. Destacamos :

(i) Inverso do Seno Hiperbólico : estabiliza a variância de uma variável Binomial Negativa (BEALL, 1942).

(ii) "Legit" introduzida por FISHER (1950) para estudo de frequência de gens.

(iii) "Probit" : a função distribuição da N(0,1) é usada para modelar frequências acumuladas, principalmente para relacionar estas frequências com variáveis explicativas (FINNEY, 1971).

(iv) "Logit" : similar à transformação "Probit," só que é baseada na distribuição logística ao invés da N(0,1). (BERKSON 1944).

4.6-EXEMPLOS

EXEMPLO 1 :

Os dados a seguir referem-se ao número estimado de quatro tipos de plânctons, coletados com rede de arrasto - (WINSOR & CLARKE, 1940).

NÚMERO DA REDE DE ARRASTO	TIPO DE PLÂNCTON			
	I	II	III	IV
1	895	1 520	43 000	11 000
2	540	1 610	32 800	8 600
3	1 020	1 900	28 800	8 260
4	470	1 350	34 600	9 830
5	428	980	27 800	7 600
6	620	1 710	32 800	9 650
7	760	1 930	28 100	8 900
8	537	1 960	18 900	6 060
9	845	1 840	31 400	10 200
10	1 050	2 410	39 500	15 500
11	387	1 520	29 000	9 250
12	497	1 695	22 300	7 900
Média ($\bar{y}$)	670,75	1 701,25	30 775,00	9 395,83
Amplitude	633	1 480	24 400	9 440
Desvio Padrão (s_y)	233,92	356,54	6 688,68	2 326,04
$s_y / \bar{y}$	0,35	0,21	0,22	0,25

A simples inspeção das amplitudes e dos desvios padrão desta tabela evidencia que não é razoável supor homocedasticidade.

Como $s_y/\bar{y}$ é aproximadamente constante para os quatro tipos de plânctons, existe indicação de que a tranformação

logarítmica estabiliza a variância. De fato, as estatísticas apresentadas no quadro abaixo mostram como a transformação logarítmica (base 10) produziu uma sensível homogeneidade das variâncias dos quatro tipos de plânctons.

ESTATÍSTICA	TIPO DE PLÂNCTON			
	I	II	III	IV
Média ($\bar{z}$)	2,80	3,22	4,48	3,96
Amplitude	0,43	0,39	0,36	0,41
Desvio Padrão (s_z)	0,15	0,10	0,10	0,10

EXEMPLO 2 :

Um experimento de controle de pragas em plantação de aveia consistiu na aplicação de cinco tratamentos destinados a combater certas ervas daninhas. Após a aplicação dos tratamentos, contou-se o número de ervas daninhas por unidade de área e o resultado obtido encontra-se no quadro abaixo (BARTLETT, 1936).

BLOCO	TRATAMENTO				
	A	B	C	D	E
1	438	538	77	17	18
2	442	422	61	31	26
3	319	377	157	87	77
4	380	315	52	16	20
Média ($\bar{y}$)	395	413	87	38	35
Amplitude	123	223	105	71	59

A variação das amplitudes dos tratamentos sugere a não homogeneidade de variância.

Por experiência sabe-se que, a variável analisada segue aproximadamente a distribuição de Poisson, cuja transformação utilizada para estabilizar a variância é a raiz quadrada. O quadro abaixo mostra uma considerável redução da variação das amplitudes na nova escala.

ESTATÍSTICA	TRATAMENTO				
	A	B	C	D	E
Média ($\bar{z}$)	19,8	20,2	9,1	5,8	5,6
Amplitude	3,1	5,5	5,3	5,3	4,6

EXEMPLO 3 :

Em um estudo sobre o dano causado por larvas na cultura do milho, foram utilizados, além do tratamento padrão (A), seis métodos de controle (B, C, D, E, F e G).

Os valores da tabela a seguir representam as porcentagens da colheita que ficaram danificadas (COCHRAN, 1940).

TRATAMENTO	BLOCO					
	1	2	3	4	5	6
A	42,4	34,3	24,1	39,5	55,5	49,1
B	23,5	15,1	11,8	9,4	31,7	15,9
C	33,3	33,3	5,0	26,3	30,2	28,6
D	11,4	13,5	2,5	16,6	39,4	11,1
E	14,3	29,0	10,8	21,9	30,8	15,0
F	8,5	21,9	6,2	16,0	13,5	15,4
G	16,6	19,3	16,6	2,1	11,1	11,1

As porcentagens variam de 2,1% a 55,5%, um intervalo razoavelmente grande. O número de unidades experimentais dos blocos submetido a cada tratamento (n) não era constante mas com variação pequena (em torno de 36,5). Com objetivo de estabilizar a variância foi aplicada a transformação angular.

Para os dados transformados, o quadrado médio residual é 32,76, bem diferente da variância teórica que é dada por 821/36,5 = 22,49 (ver tabela da seção 4.3). Existe portanto uma indicação de que a transformação não foi muito adequada. Talvez a causa disto é que a distribuição da variável analisada não seja Binomial e, neste caso, outras transformações devem ser procuradas.

CAPÍTULO 5

A TRANSFORMAÇÃO DE BOX-COX

5.1 - INTRODUÇÃO

No capítulo 4 discutimos o uso de transformações para a adequação de um conjunto de dados às suposições do Modelo Linear Geral, tratando cada suposição separadamente.

Neste capítulo apresentamos o procedimento desenvolvido por BOX & COX (1964) que fornece uma transformação da variável resposta (Y) tal que as condições de estrutura simples do modelo, homocedasticidade e distribuição normal podem ser simultaneamente satisfeitas.

Naturalmente essas suposições são muito menos restritivas do que a hipótese usual de que as suposições são todas satisfeitas para a variável original.

A maioria das publicações citadas no capítulo anterior trata de transformações específicas (raiz quadrada, logarítmica, etc). Em 1954, MOORE & TUKEY e ANSCOMBE & TUKEY lançam uma idéia bastante interessante e mais geral que os procedimentos até então apresentados: consistia em se trabalhar com famílias de transformações.

TUKEY(1957) sugere a classe de funções do tipo potência, definida por:

$$y_T^{(\lambda)} = \begin{cases} y^{\lambda} & , \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases} \qquad (5.1.1)$$

onde λ é um número real.

Essa classe de transformações inclui casos especiais muito utilizados na prática: logarítmica ($\lambda = 0$), recíproca ($\lambda = -1$) e a transformação raiz quadrada ($\lambda = 1/2$).

BOX & COX (1964) alteram essa família para evitar descontinuidade para $\lambda = 0$. Definem então a seguinte família de transformações:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & , \quad \lambda \neq 0 \\ \log y & , \quad \lambda = 0 \end{cases} \qquad (5.1.2)$$

onde log representa o logarítmo natural.

O valor da estatística F da análise de variância é o mesmo para as famílias (5.1.1.) e (5.1.2) e portanto as conclusões não se alteram. Adotaremos a segunda, por ser aquela que ganhou grande destaque na literatura.

Para que essa família esteja bem definida, devemos ter $y > 0$. Com esta restrição, $Y^{(\lambda)}$ tem exatamente a distribuição normal, somente se λ é zero (ver seção 4.2). Se $\lambda > 0$, $y^{(\lambda)} > -1/\lambda$ e se $\lambda < 0$ $y^{(\lambda)} < -1/\lambda$. Entretanto, do ponto de vista prático, podemos obter uma transformação do tipo potência que tenha aproximadamente a distribuição normal, embora limitada superior ou inferiormente.

A idéia básica do procedimento de BOX-COX é considerar λ como um parâmetro adicional e desconhecido do modelo e estimá-lo pelos métodos padrão da inferência estatística.

BOX & COX (1964) apresentam dois enfoques para a estimação de λ: método de máxima verossimilhança e método Baysiano. Consideraremos apenas a primeira abordagem já que, em geral,

os dois métodos produzem estimativas muito próximas e também porque a estimação Baysiana apresenta dificuldades computacionais. Outras razões para a escolha do método de máxima verossimilhança encontram-se nas discussões no final do artigo citado no início do parágrafo.

5.2 - A DETERMINAÇÃO DA TRANSFORMAÇÃO

5.2.1 - Estimação de λ

Seja $Y = (Y_1, Y_2, \ldots, Y_n)$ o vetor das observações independentes da variável resposta. Suponhamos que para algum λ desconhecido, o vetor das observações transformadas $Y^{(\lambda)} = (Y_1^{(\lambda)}, Y_2^{(\lambda)}, \ldots, Y_n^{(\lambda)})$ satisfaz as suposições da teoria normal. Em outras palavras assumimos que, para uma escolha conveniente de λ , o Modelo Linear Geral é tal que:

$$Y^{(\lambda)} = X\,\theta + \varepsilon$$

onde $\varepsilon \sim N(0, \sigma^2 I)$. Além disso, suponhamos que o modelo transformado tenha a estrutura "mais simples possível".

A função de verossimilhança de $Y^{(\lambda)}$ é dada por:

$$\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp \left\{ - \frac{(Y^{(\lambda)} - X\,\theta)'(Y^{(\lambda)} - X\,\theta)}{2\sigma^2} \right\}$$

e para as observações originais:

$$\frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ - \frac{(Y^{(\lambda)} - X\,\Theta)' (Y^{(\lambda)} - X\,\Theta)}{2\sigma^2} \right\} J(\lambda;Y)$$

onde $J(\lambda;Y) = \prod_{i=1}^{n} \left| \frac{d}{dy_i} y_i^{(\lambda)} \right|$ é o jacobiano da transformação e $Y^{(\lambda)}$ é expresso em termos de Y.

Para simplificação, tomamos o logarítmo natural dessa função. Obtemos:

$$L_{\Theta,\sigma^2}(\lambda) = - \frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{(Y^{(\lambda)} - X\,\Theta)' (Y^{(\lambda)} - X\Theta)}{2\sigma^2} + \log J(\lambda;Y)$$

Os estimadores de máxima verossimilhança dos parâmetros envolvidos em $L_{\Theta,\sigma^2}(\lambda)$ são obtidos em duas etapas. Primeiramente estimamos Θ e σ^2, para λ fixo. Se X tem posto completo, os estimadores são dados, respectivamente, por:

$$\hat{\Theta}_\lambda(Y) = (X'X)^{-1} X' Y^{(\lambda)} \quad \text{e}$$

$$\hat{\sigma}^2_\lambda(Y) = ((Y^{(\lambda)} - X\,\hat{\Theta})'(Y^{(\lambda)} - X\,\hat{\Theta}))/n = (Y^{(\lambda)'} A\, Y^{(\lambda)})/n = S(\lambda;Y)/n$$

onde $A = I - X(X'X)^{-1}X'$ e $S(\lambda;Y)$ é a soma de quadrados residual referente à variável transformada.

Se X não tem posto completo, substituimos $(X'X)^{-1}$ pela inversa generalizada de X'X (SEARLE, 1971).

Para λ fixo, o máximo de $L_{\Theta,\sigma^2}(\lambda)$ é, exceto por uma constante,

$$L_{max}(\lambda) = -\frac{n}{2} \log \hat{\sigma}^2_{\lambda}(Y) + \log J(\lambda;Y)$$

Para a família considerada, $J(\lambda;Y) = \prod_{i=1}^{n} y_i^{\lambda-1}$, para todo λ. Teremos então,

$$L_{max}(\lambda) = -\frac{n}{2} \log \hat{\sigma}^2_{\lambda}(Y) + (\lambda-1) \sum_{i=1}^{n} \log y_i$$

$$= -\frac{n}{2} \log \frac{(Y^{(\lambda)'} A\, Y^{(\lambda)})}{n} + (\lambda-1) \sum_{i=1}^{n} \log y_i$$

ou ainda

$$L_{max}(\lambda) = -\frac{n}{2} \log \frac{S(\lambda;Y)}{n} + (\lambda-1) \sum_{i=1}^{n} \log y_i \qquad (5.2.1)$$

A segunda etapa consiste em se determinar o estimador de máxima verossimilhança de λ. Para issso, calculamos:

$$\frac{d}{d\lambda} L_{max}(\lambda) = \frac{d}{d\lambda}\left[-\frac{n}{2} \log \left(\frac{Y^{(\lambda)'} A\, Y^{(\lambda)}}{n} \right) + (\lambda-1) \sum_{i=1}^{n} \log y_i \right]$$

$$= -\frac{n}{2} \frac{n}{Y^{(\lambda)'} A\, Y^{(\lambda)}} 2 \frac{Y^{(\lambda)'} A}{n} \frac{d}{d\lambda} Y^{(\lambda)} + \sum_{i=1}^{n} \log y_i$$

Se $\lambda \neq 0$, temos que

$$\frac{d}{d\lambda} L_{max}(\lambda) = -\frac{n\, Y^{(\lambda)'} A}{Y^{(\lambda)'} A\, Y^{(\lambda)}} (U^{(\lambda)} - \frac{1}{\lambda} Y^{(\lambda)}) + \sum_{i=1}^{n} \log y_i$$

$$= - n \frac{Y^{(\lambda)'} A\, U^{(\lambda)}}{Y^{(\lambda)'} A\, Y^{(\lambda)}} + \frac{n}{\lambda} + \sum_{i=1}^{n} \log y_i$$

onde $U^{(\lambda)}$ é o vetor de componentes de $\{\lambda^{-1} y_i^{\lambda} \log y_i\}$

Se $\lambda = 0$, usamos o fato de que

$$\frac{d}{d\lambda} y_i^{(\lambda)} \Big|_{\lambda=0} = \lim_{\lambda \to 0} \frac{1}{\lambda} \left(\frac{y_i^{\lambda} - 1}{\lambda} - \log y_i \right) = \frac{(\log y_i)^2}{2}$$

Neste caso,

$$\frac{d}{d\lambda} L_{max}(\lambda) = - \frac{n}{2} \frac{V' A W}{V' A V} + \sum_{i=1}^{n} \log y_i$$

onde V e W são respectivamente os vetores de componentes $\{\log y_i\}$ e $\{(\log y_i)^2\}$. Assim, teremos:

$$\frac{d}{d\lambda} L_{max}(\lambda) = \begin{cases} -n \dfrac{Y^{(\lambda)'} A\, U^{(\lambda)}}{Y^{(\lambda)'} A\, Y^{(\lambda)}} + \dfrac{n}{\lambda} + \displaystyle\sum_{i=1}^{n} \log y_i, & \lambda \neq 0 \\ -\dfrac{n}{2} \dfrac{V' A W}{V' A V} + \displaystyle\sum_{i=1}^{n} \log y_i, & \lambda = 0 \end{cases} \qquad (5.2.2)$$

Igualando-se a zero essa derivada não é possível explicitar λ. Métodos aproximados são portanto necessários.

O desenvolvimento acima pode ser expresso de forma mais simples trabalhando-se com a transformação normalizada

$$\underset{\sim}{z}(\lambda) = \frac{y^{(\lambda)}}{J(\lambda;Y)^{1/n}}$$

Como $J(\lambda;Y) = \prod_{i=1}^{n} y_i^{\lambda-1}$, para todo λ , teremos:

$$z^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}} & , \lambda \neq 0 \\ \\ \dot{y} \log y & , \lambda = 0 \end{cases} \qquad (5.2.3)$$

onde $\dot{y}$ é a média geométrica das observações da variável y .

É fácil mostrar que, para a transformação normalizada, $J(\lambda;z) = 1$ e portanto,

$$L_{max}(\lambda) = -\frac{n}{2} \log \hat{\sigma}^2_{\lambda}(Z) = -\frac{n}{2} \log \{\frac{S(\lambda;Z)}{n}\} \qquad (5.2.4)$$

onde $S(\lambda;Z) = Z^{(\lambda)'} A Z^{(\lambda)}$ é a soma de quadrados residual referente à variável normalizada.

Podemos repetir os cálculos de $\frac{d}{d\lambda} L_{max}(\lambda)$ para a variável normalizada mas também não é possível explicitar λ .

Para o desenvolvimento Baysiano, a expressão correspondente a (5.2.4) seria o logarítmo da distribuição a posteriori de λ , que é dada por:

$$L_B(\lambda) = -\frac{r}{2} \log \{\frac{S(\lambda;Z)}{r}\}$$

onde r representa os graus de liberdade do resíduo. As duas funções diferem somente pela substituição de r por n. Ambas são funções monotômicas de $S(\lambda;Z)$ e nos dois casos, o máximo da função ocorre quando $S(\lambda;Z)$ é mínima. Assim $L_{max}(\lambda)$ e $L_B(\lambda)$ são praticamente equivalentes e, só no caso em que r/n for bem menor que 1, é que haverá diferença entre os procedimentos.

Voltemos ao problema da estimação de λ . Procuramos o valor de λ que maximiza $L_{max}(\lambda)$ dado por (5.2.4) ou equivalentemente que minimiza $S(\lambda;Z)$, a soma de quadrados residual da variável normalizada (5.2.3).

BOX & COX (1964) sugerem a solução numérica mais simples possível. Para vários valores de λ num intervalo conveniente, calculamos $L_{max}(\lambda)$. Depois, construimos o gráfico de $L_{max}(\lambda)$ em função de λ e então determinamos graficamente o valor de λ que maximiza essa função. Este será o estimador de máxima verossimilhança ($\hat{\lambda}$), estando portanto determinada a transformação de BOX-COX.

Convém notar a equivalência em se maximizar $L_{max}(\lambda)$ dado por (5.2.1) e (5.2.4). Nos dois casos, o método descrito acima produz a mesma estimativa de λ e, portanto a mesma transformação mas, na prática a variável normalizada é mais utilizada.

A construção do intervalo de confiança para λ , baseia-se no fato de que $-2 \log \Lambda$ tem aproximadamente distribuição χ^2, onde Λ é a razão de verossimilhança generalizada (MOOD et alii, 1974, Capítulo 9). Trata-se portanto de um in

tervalo de confiança aproximado.

Primeiramente, construimos um intervalo para $L_{max}(\lambda)$, com coeficiente de confiança 1-α, que é dado por

$$L_{max}(\hat{\lambda}) - L_{max}(\lambda) < \frac{1}{2} \chi_1^2(\alpha) \qquad (5.2.5)$$

onde $\chi_1^2(\alpha)$ é o quantil de ordem 1-α de uma distribuição χ^2 com 1 grau de liberdade. Construído esse intervalo, fica determinado o intervalo de confiança para λ .

5.2.2 - Procedimento prático para estimar λ

Para a aplicação do procedimento de BOX-COX, DRAPER & SMITH (1981) sugerem que se tomem de 11 a 21 valores de λ , a princípio no intervalo [-2,2] ou talvez mesmo em [-1,1]. Depois, se for necessário, deve-se ampliar o intervalo considerado.

Nas proximidades do ponto de máximo pode-se tomar valores adicionais de λ para tornar a determinação de $\hat{\lambda}$ mais precisa. No entanto, na prática, nem sempre se utiliza o valor de $\hat{\lambda}$ obtido mas sim, um valor conveniente mais próximo da sequência ..., -2,$-1\frac{1}{2}$, -1, $-\frac{1}{2}$, 0, $\frac{1}{2}$, 1, $1\frac{1}{2}$, 2, ... Esse procedimento , em algumas situações, facilita a interpretação da variável transformada. Essa substituição é bastante razoável quando o valor adotado estiver contido no intervalo de confiança para λ.

O valor da média geométrica, necessário quando se utiliza a variável normalizada, pode ser obtido pelo comando GEOMETRIC

do BMDP ou facilmente calculado através do MINITAB (ver programa 2.2 do apêndice 2).

Os valores de $S(\lambda;Z)$, necessários para o cálculo de $L_{max}(\lambda)$ podem ser obtidos através dos comandos de Análise de Variância ou Regressão, conforme for o caso, dos "pacotes" disponíveis (BMDP, MINITAB, SPSS, etc). Como é esperado que após a transformação o modelo tenha a estrutura mais simples possível, $S(\lambda;Z)$ deve ser a soma de quadrados residual obtida sob esta condição. Por exemplo, num modelo cruzado a dois fatores, $S(\lambda;Z)$ deve ser a soma de quadrados residual do modelo sem interação (modelo aditivo).

DRAPER & SMITH (1981) sugerem uma maneira prática de se obter o intervalo de confiança para λ : traçar no gráfico de $L_{max}(\lambda)$ em função de λ uma linha horizontal no ponto $L_{max}(\hat{\lambda}) - \frac{1}{2}\chi^2_1(\alpha)$ do eixo das ordenadas. Essa reta corta a curva em dois pontos que correspondem a dois valores de λ que são os extremos do intervalo de confiança para λ.

5.3. - EXEMPLOS

Vamos ilustrar o método de determinação da transformação apresentado na secção anterior com dois exemplos extraídos do artigo de BOX & COX (1964). O primeiro é um problema de Análise de Variância em que a transformação é aplicada para se obter homogeneidade de variância. No segundo exemplo a transformação é utilizada com o objetivo de simplificar um modelo de Regressão.

EXEMPLO 1

A tabela abaixo fornece os tempos de sobrevivência de 48 animais (unidade = 10h), expostos a três diferentes substâncias tóxicas e sujeitos a quatro tratamentos distintos. O experimento foi conduzido segundo um planejamento fatorial 3 x 4, com 4 réplicas para cada combinação dos dois fatores considerados.

TÓXICO	TRATAMENTO			
	B_1	B_2	B_3	B_4
A_1	0,31	0,82	0,43	0,45
	0,45	1,10	0,45	0,71
	0,46	0,88	0,63	0,66
	0,43	0,72	0,76	0,62
A_2	0,36	0,92	0,44	0,56
	0,29	0,61	0,35	1,02
	0,40	0,49	0,31	0,71
	0,23	1,24	0,40	0,38
A_3	0,22	0,30	0,23	0,30
	0,21	0,37	0,25	0,36
	0,18	0,38	0,24	0,31
	0,23	0,29	0,22	0,33

No quadro a seguir, apresentamos a média ($\bar{y}$) e a variância (s_y^2) da variável analisada. O fato da maior variância (0,1131) ser cerca de 565 vezes maior que a menor variância (0,0002), é uma evidência forte da não homogeneidade de variância.

TÓXICO	ESTATÍSTICA	TRATAMENTO				TOTAL
		B_1	B_2	B_3	B_4	
A_1	$\bar{y}$	0,41	0,88	0,57	0,61	0,62
	s_y^2	0,0048	0,0259	0,0246	0,0127	0,0439
A_2	$\bar{y}$	0,32	0,82	0,38	0,67	0,64
	s_y^2	0,0057	0,1131	0,0032	0,0734	0,0837
A_3	$\bar{y}$	0,21	0,34	0,24	0,33	0,28
	s_y^2	0,0005	0,0022	0,0002	0,0007	0,0039
TOTAL	$\bar{y}$	0,31	0,68	0,39	0,53	0,48
	s_y^2	0,0105	0,1029	0,0279	0,0482	0,0639

Utilizando o procedimento de BOX-COX, procuramos uma transformação que torne o modelo homocedástico, normal e aditivo.

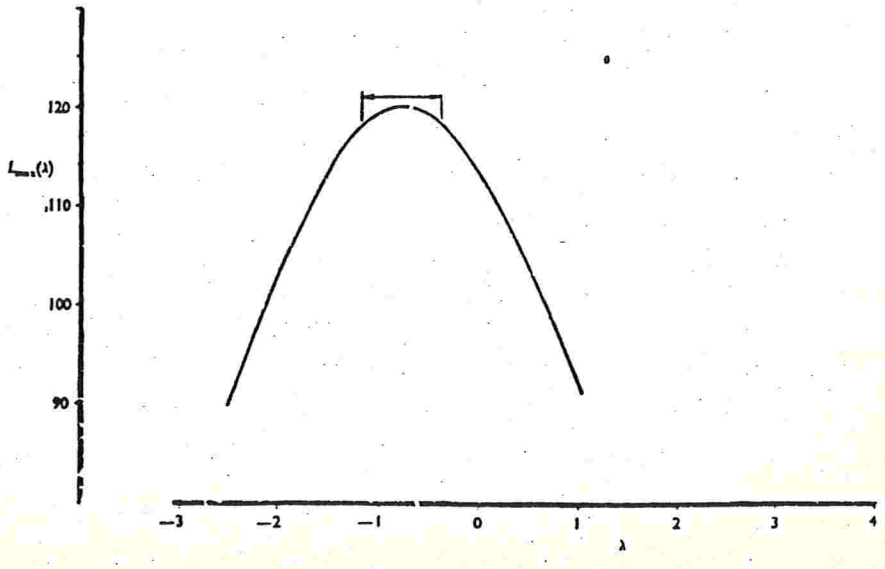
Consideremos então a família de transformação normalizada:

$$z^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}} & , \lambda \neq 0 \\ \dot{y} \log y & , \lambda = 0 \end{cases}$$

Para esse conjunto de dados a média geométrica é $\dot{y} = 0,42$. Para $\lambda = -1,0$, usamos a transformação $z^{(\lambda)} = (y^{-1,0} - 1)/ /-(0,42)^{-2,0}$; para $\lambda = 0$, $z^{(\lambda)} = 0,42 \log y$ e assim por diante.

Inicialmente calculamos, para vários valores de λ , a soma de quadrados residual ($S(\lambda;Z)$ para o modelo sem interação. A seguir, calculamos o máximo da função de verossimilhança , dado por $L_{max}(\lambda) = -24 \log \{ S(\lambda;Z)/48\}$, e então construimos o gráfico de $L_{max}(\lambda)$ em função de λ . Os resultados estão abaixo:

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
1,0	1,0509	91,72	-1,0	0,3331	119,29
0,5	0,6345	103,83	-1,2	0,3586	117,52
0,0	0,4239	113,51	-1,4	0,4007	114,86
-0,2	0,3752	116,44	-1,6	0,4625	114,43
-0,4	0,3431	118,58	-2,0	0,6639	102,74
-0,6	0,3258	119,82	-2,5	1,1331	89,91
-0,8	0,3225	120,07	-3,0	2,0489	75,69



A inspeção desse gráfico mostra que o estimador de máxima verossimilhança de λ é aproximadamente $\hat{\lambda} = -0,75$.

Um intervalo de confiança para λ é obtido através da desigualdade (5.2.5) da seção anterior. Traçando-se uma reta horizontal no ponto $L_{max}(\hat{\lambda}) - \frac{1}{2}\chi^2_1(0,05) = 118,19$ da escala vertical, a curva $L_{max}(\lambda)$ é cortada em dois pontos que correspondem aos seguintes valores de λ: -1,13 e -0,37. Assim, (-1,13; -0,37) é um intervalo aproximado para λ, com 95%, de confiança.

A transformação recíproca (λ=-1) tem um significado natural para a análise de tempo de sobrevivência, podendo ser interpretada como a taxa de mortalidade. Como o valor -1 está contido no intervalo construído acima, será adotada a transformação recíproca e não a transformação correspondente a $\hat{\lambda} = -0,75$.

Uma análise dos dados transformados mostra que a transformação recíproca produziu considerável redução da variabilidade das variâncias das caselas.

Os resultados da análise de variância para os dados originais e transformados encontram-se na tabela abaixo.

F.V.	G.L.	SEM TRANSFORMAÇÃO		TRANSFORMAÇÃO RECÍPROCA		CONCLUSÃO (α = 5%)
		Q.M.	F	Q.M.	F	
A	2	0,5165	23,27	0,5687	72,91	Significante
B	3	0,3071	13,83	0,2219	28,45	Significante
AB	6	0,0417	1,88	0,0085	1,09	Não Signif.
Resíduo	36	0,0222		0,0078		

Esses resultados mostram que a transformação:

(i) não produziu grande alteração no quadrado médio (Q.M.) do fator tóxico (A) e do tratamento (B).

(ii) reduziu o quadrado médio residual a quase um terço.

(iii) tornou o quadrado médio da interação (AB) bem mais próximo do quadrado médio residual.

EXEMPLO 2

Esse exemplo refere-se a um experimento têxtil em que a unidade experimental é um pedaço de fio torcido. O experimento consiste em aplicar-se uma carga ao fio, de tempos em tempos, e então anotar o número de trações até que ocorra o rompimento do fio (y).

Trata-se de um experimento fatorial 3^3, onde são considerados os seguintes fatores quantitativos:

X_1 : comprimento do fio (250, 300, 350 mm)

X_2 : período de carga (8,9, 10 min)

X_3 : intensidade da carga (40, 45, 50 g)

Foram obtidos os seguintes resultados:

PERÍODO DE CARGA	INTENSIDADE DE CARGA	COMPRIMENTO DO FIO		
		250	300	350
8	40	674	1 414	3 636
	45	370	1 198	3 184
	50	292	634	2 000
9	40	338	1 022	1 568
	45	266	620	1 070
	50	210	438	566
10	40	170	442	1 140
	45	118	332	884
	50	90	220	360

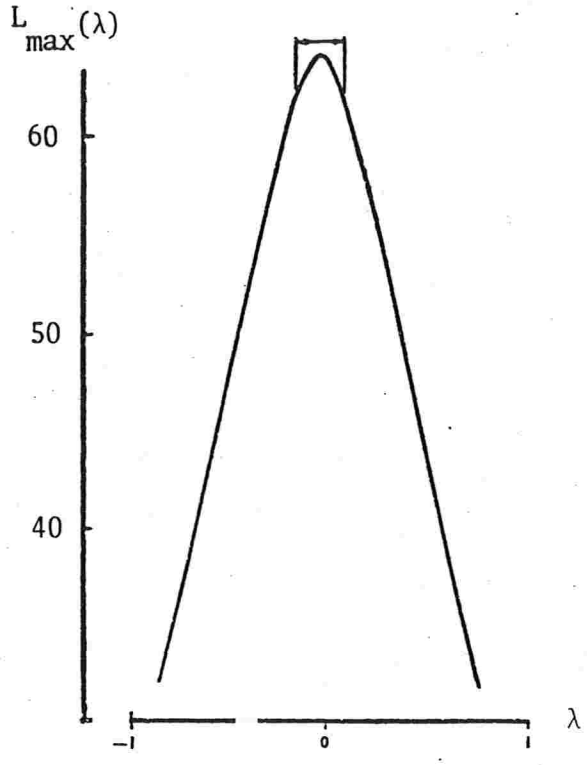
Para se estudar como a variável analisada (Y) é afetada por esses fatores considerados, adotou-se o seguinte modelo de superfície de resposta:

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_1X_2+\beta_5X_1X_3+\beta_6X_2X_3+\beta_7X_1^2+\beta_8X_2^2+\beta_9X_3^2+\varepsilon$$

A questão aqui é verificar se o modelo completo de 2º grau é necessário ou, se os dados convenientemente transformados podem ser representados pelo modelo linear de 1º grau (esse é o princípio da parcimônia, assim denominado por Tukey).

Para determinarmos a transformação de BOX-COX, calculamos a soma de quadrados residual ($S(\lambda;Z)$) e o respectivo valor da função de máxima verossimilhança ($L_{max}(\lambda)$) para vários valores de λ no intervalo $[-1,1]$. Neste caso, a média geométrica é $\dot{y} = 563,78$ e foi utilizada a transformação na forma normalizada. Como é esperado que após a transformação o modelo simplificado seja adequado, $S(\lambda;Z)$ é a soma de quadrados residual do modelo sem os termos do 2º grau. Os resultados encontram-se abaixo:

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
1,0	5,4810	21,52	-0,2	0,2920	61,11
0,8	2,9978	29,67	-0,4	0,5478	52,61
0,6	1,5968	38,17	-0,6	1,1035	43,16
0,4	0,8178	47,21	-0,8	2,1396	34,22
0,2	0,4115	56,48	-1,0	3,9955	25,79



Uma inspeção do gráfico de $L_{max}(\lambda)$ indica que a estimativa de λ é aproximadamente $\hat{\lambda} = -0,06$. Utilizando a relação (5.2.5), obtemos o seguinte intervalo aproximado, com 95% de confiança: (-0,18, 0,06).

Como $\hat{\lambda}$ é muito próximo de zero, será adotada a transformação logarítmica, que corresponde a $\lambda = 0$, e que está contido no intervalo construído acima.

Os resultados da análise de variância para os dados originais e transformados ($z^{(\lambda)} = 563,78 \log y$) são mostrados na tabela abaixo.

F.V.	G.L.	SEM TRANSFORMAÇÃO		TRANSFORMAÇÃO LOGARÍTIMICA	
		Q.M.	F	Q.M.	F
Linear	3	4 916 158,33	66,50	2 374 448,33	198,83
Quadrático	6	704 050,17	9,52	8 103,17	0,63
Resíduo	17	73 922,41		11 941,88	

Para os níveis usuais, o efeito quadrático no modelo é significante para os dados originais mas não é significante após a transformação logarítmica. Uma análise de resíduos tende a confirmar a adequacidade do modelo de 1º grau.

5.4 - COMENTÁRIOS ADICIONAIS

1 - Embora a família (5.1.2) tenha grande aplicabilidade, podem surgir situações em que nenhum membro dessa família forneça os resultados desejados. Neste caso, é sugerido na literatura o uso de transformações mais complexas, envolvendo mais de um parâmetro. BOX & COX (1964) propõem a seguinte família biparamétrica:

$$y^{(\lambda_1,\lambda_2)} = \begin{cases} \dfrac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1} & , \quad \lambda_1 \neq 0 \\ \log (y + \lambda_2) & , \quad \lambda_1 = 0 \end{cases} \qquad (5.4.1)$$

com $y > -\lambda_2$. O caso particular em que o parâmetro λ_2 é igual a zero, corresponde à família (5.1.2).

Uma outra aplicação da família (5.4.1) aparece quando algum valor de y é negativo ou nulo. Neste caso, λ_2 não é considerado como um parâmetro desconhecido mas sim como uma constante que torne $y + \lambda_2$ positivo para todo y.

A forma normalizada de (5.4.1) é dada por:

$$z^{(\lambda)} = \begin{cases} \dfrac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda_1 \dot{y}^{\lambda_1 - 1}} & , \quad \lambda_1 \neq 0 \\ \dot{y} \log (y + \lambda_2) & , \quad \lambda_1 = 0 \end{cases} \qquad (5.4.2)$$

onde $\dot{y}$ é a média geométrica da variável $y + \lambda_2$, ou seja,

$$\dot{y} = \prod_{i=1}^{n} (y_i + \lambda_2)^{1/n} .$$

O procedimento para estimar λ_1 e λ_2 é uma extensão do caso uniparamétrico e o intervalo é substituido pela região de confiança. Para a construção da região de confiança utiliza-se a mesma distribuição aproximada $(-2 \log \Lambda)$, agora uma χ^2 com 2 graus de liberdade. A idéia é exatamente a mesma, mas o procedimento é mais complicado. Um exemplo de aplicação, com breves comentários, é apresentado no artigo de BOX & COX (1964) e outro em JOHN & DRAPER (1980).

Muitas outras modificações da família (5.1.2) tem sido sugeridas na literatura. SCHLESSELMAN (1971) propõe a seguinte família:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - c^{\lambda}}{\lambda} , & \lambda \neq 0 \\ \log (y/c) , & \lambda = 0 \end{cases} \qquad (5.4.3)$$

onde c é uma constante. Sugere que c seja a média aritmética ou geométrica ou ainda a mediana amostral. Essa família, além de ser contínua, goza da propriedade de invariância da soma de quadrados residual, com relação à mudança de escala: $(y \xrightarrow{w} wy)$.

JOHN & DRAPER (1980) sugerem a família:

$$y^{(\lambda)} = \begin{cases} (\text{sinal de } y) \dfrac{(|y| + 1)^{\lambda} - 1}{\lambda} , & \lambda \neq 0 \\ (\text{sinal de } y) \log(|y| + 1) , & \lambda = 0 \end{cases} \qquad (5.4.4)$$

que é apropriada para o caso em que a distribuição dos erros é simétrica mas não normal. Se os dados são todos positivos, essa família é equivalente à família proposta por BOX & COX (1964)

2 - Se a variável estudada (y) é a "proporção de sucessos" , TUKEY (1960) sugere a família:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - (1-y)^{\lambda}}{\lambda} & , \quad \lambda \neq 0 \\ \log \dfrac{y}{1-y} & , \quad \lambda = 0 \end{cases} \qquad (5.4.5)$$

JOINER & ROSENBLATT (1971) estudam propriedades dessa família.

3 - O método de BOX-COX requer a repetição dos cálculos da análise de variância para vários valores de λ e a distribuição utilizada para a construção do intervalo de confiança para λ é assintótica. Além disso, $\hat{\lambda}$ também é afetado pela presença de "outliers".

Para contornar esses problemas mencionados, foram propostos na literatura outros métodos de inferência para o parâmetro λ . Os mais citados são: método exato de ANDREWS (1971); método de ATKINSON (1973), também baseado na função de verossimilhança; método de CARROLL (1980), baseado no processo de estimação robusta de HUBER (1977) e o método proposto por HERNANDEZ & JOHNSON (1980), baseado no número de informação de Kullback-Leibler.

Um estudo interessante sobre o efeito de um "outlier" na

escolha da transformação pode ser encontrado em ANDREWS (1971), ATKINSON (1973) e CARROLL (1980) sendo estabelecidas comparações entre o método de BOX-COX e estes três últimos métodos citados.

4 - Até agora comentamos sobre a aplicação do procedimento de BOX-COX no caso em que a variável resposta é transformada. Uma outra aplicação importante do método refere-se à transformação simultânea da variável resposta (y) e das variáveis explicativas $(x_1 \ldots, x_k)$, em problemas de Regressão.

Consideremos a transformação de y para $y^{(\lambda)}$ e de $x_1 \ldots, x_k$ para $x_1^{(\lambda_1)}, \ldots, x_k^{(\lambda_k)}$, respectivamente.

De forma análoga ao caso de transformação só de y, assumimos que para algum vetor de parâmetros desconhecidos $(\lambda, \lambda_1, \ldots, \lambda_k)$, valem as suposições usuais da Análise de Regressão.

Consideramos então o logarítmo do máximo da função de verossimilhança, que é dada por:

$$L_{max}(\lambda, \lambda_1, \ldots, \lambda_k) = -\frac{n}{2} \log \hat{\sigma}^2_{\lambda, \lambda_1, \ldots, \lambda_k}(Y) + \log J(\lambda; Y) \qquad (5.4.6)$$

onde $\hat{\sigma}^2_{\lambda, \lambda_1, \ldots, \lambda_k}(Y)$ é o estimador da máxima verossimilhança da variância dos erros em uma Análise de Regressão das variáveis transformadas e $J(\lambda; Y) = \prod_{i=1}^{n} \left| \frac{d}{dy_i} y_i^{(\lambda)} \right|$ é o jacobiano da transformação.

Procuramos então $(\hat{\lambda}, \hat{\lambda}_1, \ldots, \hat{\lambda}_k)$ que maximize a função de verossimilhança.

Teoricamente o desenvolvimento é uma extensão do caso em

que apenas y é transformado. Na prática, o procedimento consiste em se calcular $L_{max}(\lambda,\lambda_1,\ldots,\lambda_k)$ para valores convenientes de $(\lambda,\lambda_1,\ldots,\lambda_k)$ e então examinar as superfícies resultantes (especialmente nas proximidades do seu máximo) mas a representação gráfica só é possível se k = 1.

SPITZER (1978) apresenta um estudo dentro deste contexto, para o caso de amostras pequenas.

5 - Se o objetivo é a simplificação do modelo de Regressão e apenas as variáveis explicativas serão transformadas, o método de BOX-COX também pode ser utilizado.

Assim como no ítem anterior, na prática aparece a dificuldade operacional do método. Um processo alternativo, já comentado na seção 3.1, é proposto por BOX & TIDWELL (1962).

6 - Uma outra importante aplicação da transformação de BOX-COX aparece em Análise de Séries Temporais, conforme descrito por BOX & JENKINS (1976, Capítulo 9) e CHATFIELD & PROTHERO (1973).

7 - Finalmente, citamos outros possíveis desenvolvimentos do método de BOX-COX. Na seção 5.2 supusemos que os fatores da Análise de Variância são fixos e consideramos uma análise unidimensional.

Segundo BOX & COX (1964), o procedimento apresentado pode ser estendido a modelos com fatores aleatórios e a problemas de análise multivariada.

CAPÍTULO 6

APLICAÇÕES DO USO DE TRANSFORMAÇÃO

Neste capítulo apresentamos duas situações reais em que o uso de transformação é apropriado. Em cada uma delas, discutimos aspectos práticos importantes.

As tabelas e figuras referentes a essas aplicações encontram-se nos apêndices 3 e 4, respectivamente.

APLICAÇÃO 1

O experimento que descrevemos abaixo, refere-se a uma pesquisa realizada pelo Professor Dirceu do Nascimento, para tese de doutoramento, apresentada na Faculdade de Ciências Farmacêuticas da USP.

Um dos objetivos do experimento é verificar como a qualidade bacteriológica do leite é afetada pelo "manuseio" (ordenha, pasteurização e comercialização) e se esta varia segundo os meses do ano em que a pesquisa foi realizada.

Foram coletadas 153 amostras de leite no período de 30 de agosto de 1977 a 2 de fevereiro de 1978 na cidade de João Pessoa, Paraíba. Do total de amostras, 51 eram de leite cru, coletadas na plataforma de recepção da usina de pasteurização (PRUP), 51 de leite pasteurizado do tipo C, coletadas na câmara fria da usina de pasteurização (CFUP) e 51 de leite do tipo C já comercializado, coletadas nas padarias e supermercados da cidade (COM).

Para cada uma das 153 amostras, foram feitas *contagens/mℓ* dos seguintes tipos de bactérias, utilizados como indicadores da qualidade bacteriológica do leite: mesófilas, psicrófilas, coliformes e coliformes fecais. Os resultados dessas contagens para os três locais considerados (PRUP, CFUP, COM) encontram-se nas TABS. I, II e III, respectivamente.

Iniciamos por uma Análise Descritiva dos dados. Alguns resultados, para cada tipo bacteriológico, encontram-se nas TABS. IV a VII.

A partir desses resultados, verificamos que:

(i) Existe grande variação entre as amplitudes e entre os desvios padrão das caselas. Por exemplo, na TAB. VII, o desvio padrão da casela PRUP X OUTUBRO é cerca de 130 338 vezes maior que o desvio padrão da casela CFUP X NOVEMBRO.

(ii) É esperado que as amplitudes de variação e os desvios padrão variem bastante segundo local já que, após o processo de pasteurização, o leite torna-se mais homogêneo. Entretanto, fixado o local, pode-se observar ainda uma grande variação das amplitudes e dos desvios padrão dos períodos. Isto pode ser visto na TAB. VI em que as contagens do mês de outubro de amostras colhidas na CFUP apresentaram a amplitude de variação cerca de 1063 vezes maior que a amplitude de variação do mês de setembro.

(iii) Existem muitas caselas em que o desvio padrão é maior que a média sendo que na TAB. V isto não ocorre em apenas uma casela.

(iv) Em muitas caselas, a média e a mediana são bem distintas, indicando uma acentuada assimetria na distribuição das variá-

veis analisadas.

(v) Existem amostras que apresentam resultados que são bastante atípicos. Este é o caso da amostra número 25 da TAB.II cuja contagem do número de bactérias ~~mesófilas~~ coliformes foi 240 000, valor extremamente maior que os demais valores dessa variável.

Essas observações evidenciam que, para esses dados, as suposições de normalidade e homogeneidade de variância não são razoáveis. A continuação da análise com esses dados pode fornecer conclusões bastante distorcidas.

Com o objetivo de estabilizar a variância das caselas e tornar a distribuição dos erros mais próxima da normal, aplicaremos o método de BOX-COX para a escolha de uma transformação adequada a esses dados. Como esse método é sensível à presença de "outliers" , devemos detectar inicialmente tais observações. Para isso utilizamos uma técnica multivariada, que se baseia na distância de Mahalanobis. Essa técnica encontra-se descrita em AFIFI & AZEN (1972).

Como esta técnica exige que o vetor de observações tenha distribuição normal multivariada não devemos aplicá-la aos dados originais, visto que esta distribuição afasta-se muito da normal. Desta forma, a aplicação da técnica que detecta "outliers" aos dados originais produz uma quantidade injustificável de observações consideradas como "outliers" . Por experiências anteriores, sabemos que a transformação logarítmica aproxima a distribuição deste tipo de dados para a normal. Assim, aplicamos a técnica, para cada local, ao logarítmo das contagens de bactérias.

Fixando o nível de significância em 5%, foram detectados os seguintes casos: amostra número 15,24 e 46 da TAB. I; amostra número 25 da TAB. II e amostra número 3 e 49 da TAB. III. Decidimos que esses casos deveriam ser eliminados da análise. Assim, o tamanho da amostra ficou reduzido a 147.

Examinando a matriz de correlação e após ter sido discutido com o pesquisador os objetivos desse estudo, decidimos analisar as variáveis separadamente.

De acordo com os objetivos da pesquisa, a técnica apropriada é a Análise de Variância. Adotamos o modelo cruzado paramétrico:

$$y_{ijk} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk} \quad i=1,2,3; j=1,...5; k=1,..., n_{ij}$$

y_{ijk} : número de bactérias mesófilas da k-ésima amostra, coletada no i-ésimo local e no j-ésimo período.

μ: média geral

A_i: efeito do i-ésimo local

B_j: efeito do j-ésimo período

AB_{ij}: efeito da interação do i-ésimo local com o j-ésimo período (efeito de casela)

ε_{ijk}: desvio casual da observação y_{ijk}

Para as outras variáveis, o modelo é o mesmo.

Para aplicar o procedimento de BOX-COX, utilizamos o "pacote" SPSS (ver programa 2.1 do apêndice 2). Os resultados do procedimento de BOX-COX encontram-se nas TABS. VIII a XI e

FIGS. 1 a 4. Verificamos que, para as quatros variáveis analisadas, o valor de λ que maximiza a função de verossimilhança é aproximadamente zero. (que corresponde à transformação logarítmica). Entretanto para estimarmos mais precisamente o valor de λ, repetimos o procedimento de BOX-COX para valores de λ nas proximidades do zero. Os resultados encontram-se nas TABS. XII a XV. As estimativas de λ para as variáveis número de bactérias mesófilas, psicrófilas, coliformes e coliformes fecais são respectivamente $\hat{\lambda} = -0,05$, $\hat{\lambda} = 0$, $\hat{\lambda} = -0,02$ e $\hat{\lambda} = -0,06$.

Para verificarmos a efetividade da transformação, repetimos os cálculos da Análise Descritiva para os dados transformados, correspondentes aos valores estimados de λ e também para a transformação logarítmica. Para efeito de comparação dos dados originais com os dados transformados apresentamos o quadro abaixo que contém o quociente do maior desvio padrão das caselas pelo menor desvio padrão:

ESPECIFICAÇÃO	TIPO DE BACTÉRIA			
	Mesófilas	Psicrófilas	Coliformes	Coliformes fecais
Dados Originais	2 611,24	2 139,57	27 930,67	130 337,99
Transformação correspondente a λ	13,25	3,18	2,39	1,98
Transformação Logarítmica	16,29	3,18	2,42	2,48

Adotamos a transformação logarítmica, para as quatro variáveis analisadas, por facilidade de interpretação e, especialmente, porque nesta escala a ordenação das médias das

caselas e das marginais (por local e por período) foi razoavelmente preservada.

Os resultados da Análise Descritiva para os dados transformados estão resumidos nas TABS. XVI a XIX. A transformação reduziu consideravelmente a variabilidade das variâncias das caselas.

Os resultados da análise de variância para os dados transformados, obtidos através do comando ANOVA do SPSS, encontram-se nas TABS. XX a XXIII. Para as quatro variáveis analisadas, o efeito de interação não é significante, para os níveis usuais.

Nos casos em que o efeito do fator é significante, aplicamos o método de comparações múltiplas de Bonferroni (PERES & SALDIVA, 1982). As comparações de interesse são as seguintes: (i) leite pasteurizado com leite comercializado; (ii) cada período com o período seguinte.

O quadro abaixo resume as significâncias encontradas:

TIPO BACTERIOLÓGICO	COMPARAÇÕES SIGNIFICANTES
Mesófilas	Setembro x Outubro
Psicrófilas	Leite pasteurizado x Leite comercializado
Coliformes	-
Coliformes fecais	Novembro x Dezembro

APLICAÇÃO 2

Consideremos um estudo sobre a resistência de telhas onduladas da marca Eternit, realizado pelo Instituto de Pesquisas Tecno

lógicas do Estado de São Paulo.

Para o experimento , as telhas foram fixadas na 1ª e 5ª onda (ver FIG. 5) e variou-se o tamanho do vão compreendido entre as fixações. O teste de resistência consistiu em aplicar-se uma carga e medir-se a resistência(1). Os resultados obtidos encontram-se na TAB. XXIV.

O objetivo do experimento pode ser atingido através de uma Análise de Regressão, considerando-se como variável resposta a resistência (Y) e como variável explicativa o tamanho do vão (X).

A FIG. 6 mostra como a variância da variável resistência diminui à medida que o tamanho do vão aumenta. Com o objetivo de estabilizar as variâncias, aplicaremos o procedimento de BOX-COX para a escolha da transformação adequada. A utilização desse procedimento exige a especificação da relação que liga a variável transformada (y^{λ}) à variável explicativa (x). Em outras palavras, precisamos estabelecer a forma da função g tal que $y^{\lambda} = g(x)$, $\lambda \neq 0$.

De um modo geral, se g(x) é um polinômio em x, a função $y = g(x)^{1/\lambda}$, para diferentes valores de λ , descreve uma grande quantidade de curvas que explicam a relação entre y e x, em muitas situações práticas.

Desta forma, com a utilização desse procedimento, é possível obter um valor de λ tal que y^{λ} satisfaça as suposições básicas da Análise de Regressão e y como função de x represente bem o fenômeno estudado.

---

(1) carga em Kg até a ruptura.

Um problema prático que surge é a escolha do grau do polinômio g(x). Uma solução é adotar inicialmente o polinômio completo de maior grau possível e eliminar os termos não significantes.

Para a Regressão polinomial de 4º grau, aplicamos o procedimento de BOX-COX, utilizando o programa 2.2 do apêndice 2. Os resultados obtidos encontram-se na TAB. XXV e podem ser visualizados na FIG. 7.

A TAB. XXVI apresenta os resultados da análise de variância do modelo de Regressão correspondente ao valor de λ estimado ($\hat{\lambda} = -0,45$). Como o termo quadrático e o de 4º grau são não significantes para os níveis usuais, repetimos o procedimento de BOX-COX para o modelo que inclui apenas o efeito linear e o cúbico. Neste caso, o valor de λ estimado foi $\hat{\lambda}=-0,60$. Os resultados da análise de variância encontram-se na TAB. XXVII e a equação de Regressão estimada é dada por:

$$\hat{z} = 4330 - 43,0\ x_1 + 3,83\ x_3$$

onde $\hat{z}$ é o valor estimado da variável $(y^{-0,60}-1)/-0,60(140,931)^{-1,60}$, x_1 e x_3 são os polinômios ortogonais referentes ao efeito linear e cúbico, respectivamente.

Para verificar a efetividade da transformação, fizemos uma análise de resíduos do modelo acima e não foi detectado nenhum desvio das suposições básicas do modelo de Regressão.

APÊNDICE 1

TESTE DE TUKEY PARA NÃO-ADITIVIDADE

Consideremos o modelo cruzado paramétrico

$$y_{ij} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ij}, \quad i=1,\ldots,a, \; j=1,\ldots,b \qquad (A.1.1)$$

onde,

μ : média geral

A_i : efeito do i-ésimo nível de A

B_j : efeito do j-ésimo nível de B

AB_{ij} : efeito de interação

ε_{ij} : erro aleatório da observação y_{ij}.

Como é usual, supomos que $\varepsilon_{ij} \sim N(0,\sigma^2)$ e fazemos as seguintes restrições : $\sum_{i=1}^{a} A_i = 0$, $\sum_{j=1}^{b} B_j = 0$, $\sum_{i=1}^{a} AB_{ij} = 0$, $\sum_{j=1}^{b} AB_{ij} = 0$

Desejamos testar a hipótese

$$H_o : A_{11} = \ldots = A_{1b} = \ldots = A_{a1} = \ldots = A_{ab} = 0$$

mas o teste usual da interação não pode ser utilizado pois não há graus de liberdade suficiente. Uma maneira, sugeri-

da por Tukey, é construir um teste que imponha restrições convenientes sobre a interação AB_{ij}. Assumimos então que o efeito de interação para cada casela é uma função dos efeitos principais, segundo um polinômio do segundo grau, ou seja :

$$AB_{ij} = \alpha_o + \alpha_1 A_i + \beta_1 B_j + \alpha_2 A_i^2 + \gamma A_i B_j + \beta_2 B_j^2 \qquad (A.1.2)$$

onde $\alpha_o, \alpha_1, \beta_1, \beta_2$ e γ são constantes. Teremos :

$$AB_{i.} = \frac{\sum_{j=1}^{b} AB_{ij}}{b} = \alpha_o + \alpha_1 A_i + \frac{\beta_1}{b} \sum_{j=1}^{b} B_j + \alpha_2 A_i^2 + \frac{\gamma}{b} A_i \sum_{j=1}^{b} B_j + \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2 = 0$$

e

$$AB_{.j} = \frac{\sum_{i=1}^{a} AB_{ij}}{a} = \alpha_o + \frac{\alpha_1}{a} \sum_{i=1}^{a} A_i + \beta_1 B_j + \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2 + \frac{\gamma}{a} B_j \sum_{i=1}^{a} A_i + \beta_2 B_j^2 = 0$$

Como $\sum_{i=1}^{a} A_i = 0$ e $\sum_{j=1}^{b} B_j = 0$, essas expressões podem ser simplificadas :

$$AB_{i.} = \alpha_o + \alpha_1 A_i + \alpha_2 A_i^2 + \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2 = 0$$

e

$$AB_{.j} = \alpha_o + \beta_1 B_j + \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2 + \beta_2 B_j^2 = 0$$

Portanto,

$$\alpha_1 A_i + \alpha_2 A_i^2 = - \alpha_o - \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2$$

e

$$\beta_1 B_j + \beta_2 B_j^2 = - \alpha_o - \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2$$

Substituindo essas expressões em (A.1.2), teremos :

$$AB_{ij} = - \alpha_o - \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2 - \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2 + \gamma A_i B_j \qquad (A.1.3)$$

Mas $AB_{i.} = \frac{\sum_{j=1}^{b} AB_{ij}}{b} = \frac{1}{b} \sum_{j=1}^{b} (-\alpha_o - \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2 - \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2 + \gamma A_i B_j) =$

$$= - \alpha_o - \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2 - \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2 + \gamma \frac{A_i}{b} \sum_{j=1}^{b} B_j = - \alpha_o - \frac{\beta_2}{b} \sum_{j=1}^{b} B_j^2 - \frac{\alpha_2}{a} \sum_{i=1}^{a} A_i^2 = 0$$

pois, por hipótese, $\sum_{j=1}^{b} B_j = 0$ e $\sum_{j=1}^{b} AB_{ij} = 0$

Portanto, a expressão (A.1.3) se reduz a

$$AB_{ij} = \gamma A_i B_j \qquad (A.1.4)$$

Substituindo (A.1.4) em (A.1.1), obtemos o novo modelo :

$$y_{ij} = \mu + A_i + B_j + \gamma A_i B_j + \varepsilon_{ij}, \; i=1,\ldots,a \; , \; j=1,\ldots,b \qquad (A.1.5)$$

com a restrição de que $\sum_{i=1}^{a} A_i = 0$ e $\sum_{j=1}^{b} B_j = 0$.

Agora a hipótese de interesse é :

$$H_o : \gamma = 0 \qquad (A.1.6)$$

O próximo passo é definir uma estatística para testar essa hipótese. Para isso, vamos calcular primeiramente o estimador de mínimos quadrados de γ. A soma de quadrados dos erros é dada por $S = \sum_{i=1}^{a} \sum_{j=1}^{b} \varepsilon_{ij}^2 = \sum_{i=1}^{a} \sum_{j=1}^{b} (y_{ij} - \mu - A_i - B_j - \gamma A_i B_j)^2$.

Como ε_{ij} é função não linear de A_i e B_j, assumimos por um momento que esses parâmetros são conhecidos.

Derivando S com relação a γ, obtemos :

$$\frac{\partial}{\partial \gamma} S = -2 \sum_{i=1}^{a} \sum_{j=1}^{b} A_i B_j (y_{ij} - \mu - A_i - B_j - \gamma A_i B_j) = 0$$

Desenvolvendo o somatório, obtemos :

$$\sum_{i=1}^{a} \sum_{j=1}^{b} A_i B_j y_{ij} - \mu \sum_{i=1}^{a} \sum_{j=1}^{b} A_i B_j - \sum_{i=1}^{a} \sum_{j=1}^{b} A_i^2 B_j -$$

$$- \sum_{i=1}^{a} \sum_{j=1}^{b} A_i B_j^2 - \gamma \sum_{i=1}^{a} \sum_{j=1}^{b} A_i^2 B_j^2 = 0.$$

Como $\sum_{i=1}^{a}\sum_{j=1}^{b} A_i B_j = \sum_{i=1}^{a}\sum_{j=1}^{b} A_i^2 B_j = \sum_{i=1}^{a}\sum_{j=1}^{b} A_i B_j^2 = 0$, teremos que

$$\sum_{i=1}^{a}\sum_{j=1}^{b} A_i B_j y_{ij} - \gamma \sum_{i=1}^{a}\sum_{j=1}^{b} A_i^2 B_j^2 = 0, \text{ e portanto,}$$

$$\hat{\gamma} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b} A_i B_j y_{ij}}{\sum_{i=1}^{a} A_i^2 \sum_{j=1}^{b} B_j^2} \qquad (A.1.7)$$

A definição natural para a soma de quadrados da interação é $\sum_{i=1}^{a}\sum_{j=1}^{b} (\hat{\gamma} A_i B_j)^2 = \hat{\gamma}^2 \sum_{i=1}^{a} A_i^2 \sum_{j=1}^{b} B_j^2$. Substituindo (A.1.7) obtemos :

$$\frac{\left[\sum_{i=1}^{a}\sum_{j=1}^{b} A_i B_j y_{ij}\right]^2}{\sum_{i=1}^{a} A_i^2 \sum_{j=1}^{b} B_j^2}.$$

Como A_i e B_j, na verdade, não são conhecidos, substituimos pelos seus respectivos estimadores de mínimos quadrados, obtidos apartir do modelo (A.1.1): $\hat{A}_i = \bar{y}_{i.} - \bar{y}_{..}$ e $\hat{B}_j = \bar{y}_{.j} - \bar{y}_{..}$ onde $\bar{y}_{i.}$ e $\bar{y}_{.j}$ são respectivamente as médias das observações referentes a A_i e B_j e $\bar{y}_{..}$ é a média geral. Assim, a soma de quadrados associada à não-atividade é dada por :

$$SQ_{\bar{N}ADIT} = \frac{\left[\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{i.}-\bar{y}_{..})(\bar{y}_{.j}-\bar{y}_{..})\,y_{ij}\right]^2}{\sum_{i=1}^{a}(\bar{y}_{i.}-\bar{y}_{..})^2\sum_{j=1}^{b}(\bar{y}_{.j}-\bar{y}_{..})^2} \qquad (A.1.8)$$

Para construirmos a estatística que testa não-aditividade vamos enunciar o seguinte teorema :

TEOREMA : Seja $SQ_{RES} = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$ a soma de quadrados residual do modelo (A.1.1) e $SQ_{\bar{N}ADIT}$ dada por (A.1.8). Definimos $SQ_{NOVO\ RES} = SQ_{RES} - SQ_{\bar{N}\,ADIT}$. Então, sob a hipótese $H_o : \gamma = 0$, $SQ_{\bar{N}\,ADIT}/\sigma^2$ e $SQ_{NOVO\ RES}/\sigma^2$ são independentes e tem distribuição χ^2 com 1 e ab-a-b graus de liberdade, respectivamente.

DEMONSTRAÇÃO : Ver SCHEFFÉ (1959, pag.132) ou RAO (1973, pag.250 e 251).

A não-aditividade é testada pela estatística

$$F = \frac{SQ_{\bar{N}\,ADIT}}{SQ_{NOVO\ RES}/(ab-a-b)} \qquad (A.1.9)$$

que tem distribuição F com (1,ab-a-b) graus de liberdade.

Quando a razão F excede o valor tabelado, rejeita-se a hipótese de que o modelo aditivo é adequado.

Esse procedimento pode ser estendido para experimentos fatoriais com três ou mais fatores. No caso de três fatores, a soma de quadrados associada à não-aditividade é dada por :

$$SQ_{\bar{N}\ ADIT} = \frac{\left[\sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{c}(\bar{y}_{i..}-\bar{y}_{...})(\bar{y}_{.j.}-\bar{y}_{...})(\bar{y}_{..k}-\bar{y}_{...})\,y_{ijk}\right]^2}{\sum_{i}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2\sum_{j}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2\sum_{k}^{c}(\bar{y}_{..k}-\bar{y}_{...})^2}$$

APÊNDICE 2

PROGRAMAS PARA A ESTIMAÇÃO DE λ DA TRANSFORMAÇÃO DE BOX-COX

2.1. USO DO SPSS
(CARTÕES DE CONTROLE)

```
RUN NAME              ESTUDO BACTERIOLOGICO DO LEITE (MESOFILAS)
VARIABLE LIST         X1 TO X2, Y1 TO Y4
INPUT MEDIUM          CARD
INPUT FORMAT          FIXED(2F2.0, 2(2X,F8.0), 2(2X,F8.1))
N OF CASES            147
VAR LABELS            X1 LOCAL/X2 PERIODO/Y1 MESOFILAS/
                      Y2 PSICROFILAS/Y3 COLIFORMES/
                      Y4 COLIFORMES FECAIS
VALUE LABELS          X1 (1) PRUP (2) CFUP (3) COM/
                      X2 (1) SET (2) OUT (3) NOV (4) DEZ
                      (5) JAN + FEV
COMPUTE               Y1 = Y1/1000000
COMPUTE               MG = 0.450527
COMPUTE               LAMBDA = - 1.05
DO REPEAT             Z = Z1 TO Z20
COMPUTE               Z=((Y1**LAMBDA)-1.0)/(LAMBDA*MG**(LAMBDA-1.0))
COMPUTE               LAMBDA = LAMBDA + 0.1
END REPEAT
ANOVA                 Z1 TO Z5 BY X1(1,3) X2(1,5)/
                      Z6 TO Z10 BY X1(1,3) X2(1,5)/
                      Z11 TO Z15 BY X1(1,3) X2(1,5)/
                      Z16 TO Z20 BY X1(1,3) X2(1,5).
STATISTICS            1
READ INPUT DATA
(DADOS)
FINISH
? END JOB
```

OBSERVAÇÃO: Para a utilização desse programa é necessário o cálculo antecipado da média geométrica (MG).

2.2. USO DO MINITAB

(CARTÕES DE CONTROLE)

```
NOTE LEITURA DOS DADOS
NOTE C1 RESISTENCIA
SET C1
279 292 301 278 ....
....
NOTE VARIAVEIS DEPENDENTES SAO OS POLINOMIOS ORTOGONAIS
NOTE C2 LINEAR  C3 QUADRATICO
NOTE C3 CUBICO  C4 QUARTOGRAU
SET C2
(-2,-1, 0, 1, 2) 15
SET C3
(2, -1, -2, -1, 2) 15
SET C4
(-1, 2, 0,-2, 1) 15
SET C5
(1,-4, 6, -4, 1) 15
NOTE CALCULO DA MEDIA GEOMETRICA (K2)
LET C6 = LOGE (C1)
AVERAGE C3, K1
LET K2 = EXP (K1)
NOTE LAMBDA EM K3
LET K3 = - 1.05
NOTE PROCEDIMENTO DE BOX-COX
STORE
PRINT K3
BRIEF 2
LET K4 = K3*K2** (K3- 1.0)
LET K7 = (C1** K3-1.0)/ K4
REGRESS Y C7 4 C2 C3 C4 C5
LET K3 = K3 + 0.05
END
EXEC 20
STOP
?END JOB
```

TABELA I Contagem de bactérias em amostras colhidas na plataforma de recepção da usina de pasteurização (PRUP) em João Pessoa - Pb. Agosto de 1977 a fevereiro de 1978.

PERÍODO	AMOSTRA Nº	TIPO DE BACTÉRIA			
		Mesófilas	Psicrófilas	Coliformes	Coliformes fecais
Setembro	1	9 100 000	810 000	930 000	930 000
	2	11 100 000	5 300 000	1 200 000	230 000
	3	17 300 000	129 000	43 000	15 000
	4	16 800 000	1 880 000	9 300 000	1 500 000
	5	8 100 000	5 700 000	43 000	23 000
	6	10 100 000	350 000	2 300 000	21 000
	7	14 800 000	128 000	93 000	15 000
	8	8 100 000	790 000	230 000	230 000
	9	13 400 000	11 900 000	23 000	23 000
	10	8 300 000	7 900 000	230 000	93 000
	11	3 500 000	430 000	23 000	23 000
	12	3 200 000	1 190 000	15 000	4 300
	13	7 900 000	4 800 000	430 000	150 000
Outubro	14	7 900 000	770 000	93 000	1 500
	15	4 500 000	1 180 000	210 000	35
	16	13 800 000	2 100 000	2 400 000	35 000
	17	8 100 000	5 400 000	230 000	9 300
	18	8 100 000	1 140 000	2 400 000	21 000
	19	3 200 000	1 300 000	93 000	15 000
	20	6 200 000	250 000	930 000	93 000
	21	16 100 000	2 890 000	2 300 000	2 400 000
	22	8 200 000	310 000	93 000	43 000
	23	9 600 000	6 200 000	93 000	93 000
	24	950 000	0	15 000	930
	25	600 000	90 000	2 300	930
	26	1 340 000	60 000	43 000	43 000
Novembro	27	12 400 000	90 000	23 000	930
	28	4 000 000	1 300 000	430 000	21 000
	29	13 300 000	1 930 000	43 000	9 300
	30	3 700 000	410 000	2 100	430
	31	3 900 000	510 000	230 000	230 000
	32	1 090 000	30 000	43 000	930
	33	2 110 000	200 000	93 000	15 000
	34	760 000	550 000	7 500	1 200
	35	5 800 000	2 530 000	43 000	4 300
Dezembro	36	5 500 000	10 000	150 000	23 000
	37	510 000	150 000	2 800	2 800
	38	49 000 000	1 100 000	750 000	230 000
Janeiro e Fevereiro	39	13 300 000	8 400 000	2 400 000	210 000
	40	1 140 000	270 000	4 300	4 300
	41	61 000 000	3 200 000	2 400 000	930 000
	42	3 300 000	480 000	23 000	4 300
	43	5 800 000	1 410 000	93 000	7 500
	44	6 600 000	4 700 000	230 000	230 000
	45	28 300 000	1 400 000	4 600 000	280 000
	46	170 000	100 000	430	430
	47	6 700 000	6 100 000	93 000	7 300
	48	2 200 000	600 000	75 000	4 300
	49	11 300 000	1 230 000	150 000	93 000
	50	3 600 000	580 000	93 000	93 000
	51	760 000	260 000	43 000	4 300

FONTE - BUSSAB, W.O, 1981.

TABELA II Contagem de bactérias em amostras colhidas na câmara fria da usina de pasteurização (CFUP) em João Pessoa - Pb. Agosto de 1977 a fevereiro de 1978.

PERÍODO	AMOSTRA Nº	TIPO DE BACTÉRIA			
		Mesófilas	Psicrófilas	Coliformes	Coliformes fecais
Setembro	1	192 000	940	93	2
	2	300 000	35 000	4,3	0,3
	3	155 000	630	43	4,3
	4	176 000	1 350	230	7,5
	5	158 000	19 800	93	23
	6	119 000	1 050	230	7,5
	7	117 000	1 140	43	2,1
	8	196 000	5 300	21	9,3
	9	138 000	35 000	230	7,5
	10	147 000	6 600	150	23
	11	122 000	1 700	4,3	0,9
	12	116 000	3 300	93	0,7
	13	152 000	27 500	230	0,9
Outubro	14	151.000	730	23	0,3
	15	208 000	380	4,3	0,9
	16	148 000	71 000	230	230
	17	122 000	43 000	2 300	21
	18	75 000	10 200	93	93
	19	111 000	15 800	430	15
	20	33 000	510	23	9,3
	21	42 000	560	43	23
	22	18 100	450	93	0,3
	23	54 000	2 900	43	4,3
	24	75 000	2 520	75	4,3
	25	400 000	195 000	240 000	93
	26	15 000	1 160	93	4,3
Novembro	27	43 000	170	23	9,3
	28	18 800	300	4,3	0,4
	29	68 000	320	93	2,3
	30	226 000	1 600	93	0,9
	31	126 000	620	430	15
	32	23 100	410	2,3	0,3
	33	94 000	250	43	0,9
	34	57 000	440	430	4,3
	35	62 000	19 500	230	1,5
Dezembro	36	83 000	630	230	230
	37	64 000	6 600	430	230
	38	67 000	1 700	43	9,3
Janeiro e Fevereiro	39	74 000	5 900	43	1,5
	40	40 000	660	93	43
	41	102 000	10 100	930	930
	42	120 000	2 200	93	43
	43	170 000	500	93	23
	44	120 000	970	93	21
	45	202 000	570	150	150
	46	1 250 000	1 200	4,3	1,5
	47	800 000	2 100	230	9,3
	48	1 110 000	270	75	0,9
	49	197 000	3 700	43	1,5
	50	1 460 000	32 000	23	4,3
	51	1 880 000	31 000	930	4,3

FONTE - BUSSAB, W.O, 1981

TABELA III Contagem de bactérias em amostras de leite do tipo C colhidas nas padarias e supermercados da cidade, (COM), em João Pessoa - Pb. Agosto de 1977 a fevereiro de 1978.

PERÍODO	AMOSTRA Nº	TIPO DE BACTÉRIA			
		Mesófilas	Psicrófilas	Coliformes	Coliformes fecais
Setembro	1	121 000	890	23	1,5
	2	178 000	191 000	430	9,3
	3	162 000	10 500	2 300	2 300
	4	155 000	1 300	9,3	9,3
	5	145 000	1 760	43	0,4
	6	310 000	96 000	430	2,3
	7	103 000	36 000	230	93
	8	136 000	46 000	23	23
	9	147 000	1 340	230	1,5
	10	117 000	4 900	430	4,3
	11	110 000	470	9,3	2,3
	12	133 000	970	230	0,4
	13	102 000	11 000	150	1,5
Outubro	14	159 000	4 600	9,5	9,3
	15	298 000	53 000	23	0,7
	16	131 000	58 000	4,3	1,5
	17	3 150 000	1 140 000	9 300	93
	18	123 000	14 200	230	230
	19	115 000	44 000	930	7,5
	20	27 900	110	15	0,9
	21	125 000	2 820	230	2,3
	22	11 200	350	230	0,4
	23	430 000	61 000	1 500	0,9
	24	53 000	4 400	230	2,3
	25	24 200	7 900	93	0,9
	26	66 000	2 400	430	0,9
Novembro	27	129 000	1 500	93	2,1
	28	139 000	50	2,3	0,4
	29	100 000	23 400	93	0,4
	30	101 000	41 500	930	0,4
	31	78 000	1 220	93	9,3
	32	101 000	10 600	430	0,4
	33	164 000	63 000	23	2,3
	34	68 000	5 100	9 300	20
	35	4 620 000	2 800 000	21 000	75
Dezembro	36	45 000	360	4,3	0,7
	37	52 000	3 700	230	43
	38	89 000	1 300	210	23
Janeiro	39	83 000	10 300	75	23
e	40	60 000	2 100	430	23
Fevereiro	41	109 000	5 700	93	43
	42	72 000	7 500	150	23
	43	79 000	1 030	23	4,3
	44	209 000	970	93	4,3
	45	66 000	1 320	43	2,3
	46	117 000	3 100	43	15
	47	1 200 000	7 200	930	9,3
	48	700 000	1 800	43	0,4
	49	2 800 000	39 000	93 000	4,3
	50	1 140 000	11 000	93	2,3
	51	1 170 000	12 000	750	4,3

FONTE - BUSSAB, W.O, 1981.

TABELA IV - Estatísticas descritivas referentes à variável "número de bactérias mesófilas"

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Máximo	17 300 000	300 000	310 000	17 300 000
	Mínimo	3 200 000	116 000	102 000	102 000
	Amplitude	14 100 000	184 000	208 000	17 198 000
	Média	10 130 769,23	160 615,38	147 615,38	3 479 666,67
	Mediana	9 100 000	150 000	136 000	162 000
	Desvio padrão	4 462 320,89	49 885,69	84 057,85	5 384 281,17
	Tamanho da amostra	13	13	13	39
Outubro	Máximo	16 100 000	400 000	3 150 000	16 100 000
	Mínimo	600 000	15 500	11 200	11 200
	Amplitude	15 500 000	384 500	3 138 800	16 088 800
	Média	6 814 615,39	111 738,46	362 561,54	2 429 638,46
	Mediana	7 900 000	75 000	123 000	151 000
	Desvio padrão	4 751 853,35	104 263,12	845 695,86	4 151 821,29
	Tamanho da amostra	13	13	13	39
Novembro	Máximo	13 300 000	266 000	4 620 000	13 300 000
	Mínimo	760 000	18 800	68 000	18 800
	Amplitude	12 540 000	207 200	4 552 000	13 281 200
	Média	5 228 888,89	79 766,67	611 111,11	1 973 255,56
	Mediana	3 900 000	62 000	101 000	129 000
	Desvio padrão	4 600 764,73	64 203,35	1 503 631,31	3 572 412,98
	Tamanho da amostra	9	9	9	27
Dezembro	Máximo	49 000 000	83 000	89 000	49 000 000
	Mínimo	510 000	64 000	45 000	45 000
	Amplitude	48 490 000	19 000	44 000	48 955 000
	Média	18 336 666,67	71 333,33	62 000,00	6 156 666,67
	Mediana	5 500 000	67 000	52 000	83 000
	Desvio padrão	26 672 177,14	10 214,37	23 645,18	16 164 766,18
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Máximo	61 000 000	1 830 000	2 800 000	61 000 000
	Mínimo	170 000	40 000	60 000	40 000
	Amplitude	60 830 000	1 840 000	2 740 000	60 960 000
	Média	11 090 000,00	578 846,15	600 384,62	4 089 743,59
	Mediana	5 800 000	197 000	117 000	800 000
	Desvio padrão	16 783 699,43	638 981,98	808 703,65	10 697 544,33
	Tamanho da amostra	13	13	13	39
TOTAL	Máximo	61 000 000	1 880 000	4 620 000	61 000 000
	Mínimo	170 000	15 500	11 200	11 200
	Amplitude	60 830 000	1 864 500	4 608 800	60 988 800
	Média	9 147 647,06	235 245,10	394 574,51	3 259 155,56
	Mediana	6 700 000	120 000	121 000	170 000
	Desvio padrão	10 960 164,52	379 386,55	856 323,45	7 566 962,15
	Tamanho da amostra	51	51	51	153

TABELA V - Estatísticas descritivas referentes à variável "número de bactérias psicrófilas"

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Máximo	11 900 000	35 000	191 000	11 900 000
	Mínimo	128 000	630	470	470
	Amplitude	11 772 000	34 370	190 530	11 899 530
	Média	3 177 461,54	10 716,15	30 933,08	1 073 036,92
	Mediana	1 190 000	3 300	4 900	27 500
	Desvio padrão	3 685 102,00	13 528,85	55 477,98	2 561 656,87
	Tamanho da amostra	13	13	13	39
Outubro	Máximo	6 200 000	195 000	1 140 000	6 200 000
	Mínimo	0	380	110	0
	Amplitude	6 200 000	194 620	1 139 890	6 200 000
	Média	1 668 461,54	26 477,69	107 136,92	600 692,05
	Mediana	1 140 000	2 520	7 900	43 000
	Desvio padrão	2 027 279,98	54 903,21	311 251,54	1 384 043,73
	Tamanho da amostra	13	13	13	39
Novembro	Máximo	2 530 000	19 500	2 800 000	2 800 000
	Mínimo	30 000	170	50	50
	Amplitude	2 500 000	19 330	2 799 950	2 799 950
	Média	838 888,89	2 623,33	327 374,44	389 628,89
	Mediana	510 000	410	10 600	19 500
	Desvio padrão	884 723,75	6 343,33	927 484,14	792 837,57
	Tamanho da amostra	9	9	9	27
Dezembro	Máximo	1 100 000	6 600	3 700	1 100 000
	Mínimo	10 000	630	360	360
	Amplitude	1 090 000	5 970	3 340	1 099 640
	Média	420 000,00	2 976,67	1 786,67	141 587,78
	Mediana	150 000	1 700	1 300	3 700
	Desvio padrão	593 043,00	3 183,18	1 722,30	362 670,39
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Máximo	8 400 000	32 000	39 000	8 400 000
	Mínimo	100 000	270	970	270
	Amplitude	8 300 000	31 730	38 030	8 399 730
	Média	2 213 846,15	7 013,08	7 924,62	741 645,90
	Mediana	1 280 000	2 100	5 700	10 100
	Desvio padrão	2 630 736,71	11 213,24	10 154,15	1 815 075,72
	Tamanho da amostra	13	13	13	39
TOTAL	Máximo	11 900 000	195 000	2 800 000	11 900 000
	Mínimo	0	170	50	0
	Amplitude	11 900 000	194 830	2 799 950	11 900 000
	Média	1 971 313,73	11 906,47	95 091,37	692 770,52
	Mediana	810 000	1 600	5 100	11 000
	Desvio padrão	2 615 319,70	29 773,95	418 292,02	1 769 657,47
	Tamanho da amostra	51	51	51	153

TABELA VI - Estatísticas descritivas referentes à variável "número de bactérias coliformes"

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Máximo	9 300 000	230	2 300	9 300 000
	Mínimo	15 000	4,3	9,3	4,3
	Amplitude	9 285 000	225,7	2 290,7	9 299 995,7
	Média	1 143 076,92	112,66	349,05	381 179,54
	Mediana	230 000	93	230	230
	Desvio padrão	2 539 735,99	90,93	608,63	1 528 007,13
	Tamanho da amostra	13	13	13	39
Outubro	Máximo	2 400 000	240 000	9 300	2 400 000
	Mínimo	2 300	4,3	4,3	4,3
	Amplitude	2 397 700	239 995,7	9 295,7	2 399 995,7
	Média	684 792,31	18 726,95	1 017,28	234 845,52
	Mediana	93 000	93	230	430
	Desvio padrão	987 941,83	66 487,14	2 525,83	643 087,26
	Tamanho da amostra	13	13	13	39
Novembro	Máximo	430 000	430	21 000	430 000
	Mínimo	2 100	2,3	2,3	2,3
	Amplitude	427 900	427,7	20 997,7	429 997,7
	Média	101 622,22	149,84	3 551,59	35 107,89
	Mediana	43 000	93	93	430
	Desvio padrão	141 263,04	173,32	7 202,80	91 952,21
	Tamanho da amostra	9	9	9	27
Dezembro	Máximo	750 000	430	230	750 000
	Mínimo	2 800	43	4,3	4,3
	Amplitude	747 200	387	225,7	749 995,7
	Média	300 933,33	234,33	148,10	100 438,59
	Mediana	150 000	230	210	230
	Desvio padrão	395 806,28	193,54	124,94	248 550,04
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Máximo	4 600 000	930	93 000	4 600 000
	Mínimo	430	4,3	23	4,3
	Amplitude	4 599 570	925,7	92 977	4 599 995,7
	Média	784 979,23	215,41	7 366,62	264 187,08
	Mediana	93 000	93	93	230
	Desvio padrão	1 436 977,95	322,29	25 731,31	889 648,41
	Tamanho da amostra	13	13	13	39
TOTAL	Máximo	9 300 000	240 000	93 000	9 300 000
	Mínimo	430	2,3	2,3	2,3
	Amplitude	9 299 570	239 997,7	92 997,7	9 299 997,7
	Média	701 655,49	4 897,39	2 861,51	236 471,46
	Mediana	93 000	93	150	230
	Desvio padrão	1 553 810,96	33 581,29	13 309,98	950 540,30
	Tamanho da amostra	51	51	51	153

TABELA VII - Estatísticas descriticas referentes à variável "número de bactérias coliformes fecais"

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Máximo	1 500 000	23	2 300	1 500 000
	Mínimo	4 300	0,3	0,4	0,3
	Amplitude	1 495 700	22,7	2 299,6	1 499 999,7
	Média	250 561,54	6,85	188,37	83 585,58
	Mediana	23 000	4,3	2,3	9,3
	Desvio padrão	450 757,82	7,82	634,96	280 125,69
	Tamanho da amostra	13	13	13	39
Outubro	Máximo	2 400 000	230	230	2 400 000
	Mínimo	35	0,3	0,4	0,3
	Amplitude	2 399 965	229,7	229,6	2 399 999,7
	Média	211 976,54	38,36	26,97	70 680,62
	Mediana	21 000	9,3	1,5	21
	Desvio padrão	658 206,83	66,08	65,99	383 479,34
	Tamanho da amostra	13	13	13	39
Novembro	Máximo	230 000	15	200	230 000
	Mínimo	430	0,3	0,4	0,3
	Amplitude	229 570	14,7	199,6	229 999,7
	Média	31 454,44	3,88	32,26	10 496,86
	Mediana	4 300	1,5	2,1	9,3
	Desvio padrão	74 808,00	5,05	67,41	41 158,55
	Tamanho da amostra	9	9	9	27
Dezembro	Máximo	230 000	230	43	230 000
	Mínimo	2 800	9,3	0,7	0,7
	Amplitude	227 200	220,7	42,6	229 999,6
	Média	85 266,67	156,43	22,23	28 418,78
	Mediana	23 000	230	23	230
	Desvio padrão	125 749,01	127,42	21,10	75 940,80
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Máximo	930 000	930	43	930 000
	Mínimo	430	0,9	0,4	0,4
	Amplitude	929 570	929,1	42,6	929 999,6
	Média	143 740,77	94,87	12,19	47 944,15
	Mediana	7 500	9,3	4,3	23
	Desvio padrão	256 447,90	254,16	12,55	159 615,95
	Tamanho da amostra	13	13	13	39
TOTAL	Máximo	2 400 000	930	2 300	2 400 000
	Mínimo	35	0,3	0,4	0,3
	Amplitude	2 399 965	929,7	2 299,6	2 399 999,7
	Média	165 104,22	45,59	65,00	55 071,60
	Mediana	21 000	4,3	2,3	23
	Desvio padrão	419 809,15	138,95	322,41	253 114,70
	Tamanho da amostra	51	51	51	153

TABELA VIII - Soma de quadrados residual ($S(\lambda;z)$) e máximo da função de verossimilhança ($L_{max}(\lambda)$). $z = \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias mesófilas)/10^6 e $\dot{y}$ = 0,450527

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-1,05	877,169	-131,291	-0,05	30,014	116,774
-0,95	532,856	- 94,655	0,05	32,097	111,843
-0,85	330,151	- 59,470	0,15	39,119	97,301
-0,75	209,903	- 26,182	0,25	54,011	73,591
-0,65	136,338	5,534	0,35	83,007	42,006
-0,55	91,751	34,645	0,45	139,007	4,109
-0,45	64,267	60,813	0,55	248,846	- 38,691
-0,35	47,335	83,289	0,65	469,194	- 85,303
-0,25	37,179	101,040	0,75	921,717	-134,932
-0,15	31,718	112,716	0,85	1871,786	-187,000

TABELA IX - Soma de quadrados residual ($S(\lambda;z)$) e máximo da função de verossimilhança ($L_{max}(\lambda)$. $z = \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias psicrófilas + 1)/10^6 e $\dot{y}$ = 0,0211943

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-1,05	158,917	- 5,729	-0,05	0,206	482,918
-0,95	62,118	63,313	0,05	0,204	483,635
-0,85	25,225	129,551	0,15	0,266	464,130
-0,75	10,720	192,447	0,25	0,444	426,474
-0,65	4,804	251,442	0,35	0,879	376,276
-0,55	2,287	305,995	0,45	1,932	318,393
-0,45	1,167	355,446	0,55	4,542	255,564
-0,35	0,643	399,255	0,65	11,178	189,372
-0,25	0,387	436,573	0,75	28,482	120,625
-0,15	0,260	465,807	0,85	74,652	49,803

TABELA X - Soma de quadrados residual (S(λ;z)) e máximo da função de verossimilhança ($L_{max}(\lambda)$).

$z = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias coliformes) /10^4 e $\dot{y} = 0,102539$

λ	S(λ;Z)	$L_{max}(\lambda)$	λ	S(λ;Z)	$L_{max}(\lambda)$
-1,05	11 595,827	-321,046	-0,05	5,117	246,803
-0,95	4 412,974	-250,038	0,05	5,815	237,404
-0,85	1 718,007	-180,699	0,15	11,493	187,329
-0,75	686,541	-113,281	0,25	32,247	111,500
-0,65	282,804	- 48,093	0,35	106,838	23,455
-0,55	120,716	14,479	0,45	387,948	- 71,327
-0,45	53,762	73,931	0,55	1 503,328	-170,888
-0,35	25,228	129,542	0,65	6 142,738	-274,346
-0,25	12,699	179,995	0,75	26 273,479	-381,162
-0,15	7,164	222,070	0,85	116 975,644	-490,928

TABELA XI - Soma de quadrados residual (S(λ;z) e máximo da função verossimilhança ($L_{max}(\lambda)$). $z = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias coliformes fecais + 1)/10^3 e $\dot{y} = 0,105790$

λ	S(λ;Z)	$L_{max}(\lambda)$	λ	S(λ;Z)	$L_{max}(\lambda)$
-1,05	982,486	-139,625	-0,05	3,998	264,941
-0,95	494,886	- 89,236	0,05	5,737	238,397
-0,85	253,009	- 39,910	0,15	14,680	169,340
-0,75	131,467	8,208	0,25	51,432	77,188
-0,65	69,612	54,941	0,35	208,266	- 25,606
-0,55	37,670	100,075	0,45	922,332	-134,981
-0,45	20,915	143,322	0,55	4 377,279	-249,441
-0,35	11,989	184,224	0,65	21 989,595	-368,080
-0,25	7,199	221,712	0,75	115 788,257	-490,178
-0,15	4,745	252,351	0,85	663 871,618	-618,533

TABELA XII - Soma de quadrados residual ($S(\lambda;z)$) e máximo da função de verossimilhança ($L_{max}(\lambda)$). $z = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias mesófilas)/10^6 e $\dot{y} = 0,450527$

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-0,10	30,417	115,794	-0,05	30,014	116,774
-0,09	30,264	116,165	-0,04	30,042	116,706
-0,08	30,147	116,450	-0,03	30,109	116,542
-0,07	30,066	116,647	-0,02	30,214	116,286
-0,06	30,021	116,757	-0,01	30,358	115,937

TABELA XIII - Soma de quadrados residual ($S(\lambda;z)$) e máximo da função de verossimilhança ($L_{max}(\lambda)$.

y = (número de bactérias psicrófilas + 1)/10^6 e $\dot{y} = 0,0211943$

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-0,04	0,204	483,635	0,01	0,199	485,459
-0,03	0,202	484,359	0,02	0,200	485,090
-0,02	0,200	485,090	0,03	0,201	484,724
-0,01	0,199	485,459	0,04	0,202	484,359
0,00	0,198	485,829	0,05	0,204	483,635

TABELA XIV - Soma de quadrados residual ($S(\lambda;z)$ e máximo da função de verossimilhança ($L_{max}(\lambda)$. $z = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias coliformes)/10^4 e $\dot{y} = 0,102539$

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-0,10	5,804	237,544	-0,05	5,117	246,803
-0,09	5,616	239,964	-0,04	5,055	247,699
-0,08	5,453	242,129	-0,03	5,019	248,224
-0,07	5,316	243,999	-0,02	5,010	248,356
-0,06	5,203	245,578	-0,01	5,028	248,093

TABELA XV - Soma de quadrados residual ($S(\lambda;z)$ e máximo da função de verossimilhança ($L_{max}(\lambda)$. $z = \frac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda-1}}$, y = (número de bactérias coliformes fecais + 1)/10^3 e $\dot{y} = 0,105790$

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-0,10	4,153	262,145	-0,05	3,998	264,941
-0,09	4,085	263,359	-0,04	4,029	264,373
-0,08	4,034	264,282	-0,03	4,084	263,377
-0,07	4,002	264,867	-0,02	4,164	261,951
-0,06	3,990	265,088	-0,01	4,274	260,034

TABELA XVI - Estatísticas descritivas referente à variável log (número de bactérias mesófilas)

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Amplitude	1,6876	0,9502	1,1116	5,1335
	Média	16,0211	11,9506	11,8453	13,3099
	Desvio padrão	0,5233	0,2678	0,3035	2,0167
	Tamanho da amostra	13	13	12	38
Outubro	Amplitude	3,2896	2,5967	5,6392	7,2707
	Média	15,5236	11,1045	11,6182	12,6403
	Desvio padrão	1,0055	0,8438	1,4286	2,2410
	Tamanho da amostra	11	12	13	36
Novembro	Amplitude	2,8622	2,4867	4,2186	6,5617
	Média	15,0908	11,0216	11,9904	12,7009
	Desvio padrão	0,9756	0,7823	1,2878	2,0296
	Tamanho da amostra	9	9	9	27
Dezembro	Amplitude	4,5652	0,2600	0,6820	6,9929
	Média	15,4566	11,1686	10,9899	12,5384
	Desvio padrão	2,2832	0,1388	0,3593	2,4772
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Amplitude	4,3853	3,8501	2,9957	7,3297
	Média	15,5767	12,5677	12,2325	13,4349
	Desvio padrão	1,2611	1,2928	1,2412	1,9484
	Tamanho da amostra	12	13	12	37
TOTAL	Amplitude	4,7842	4,7982	6,0222	8,6027
	Média	15,5860	11,6944	11,8537	13,0182
	Desvio padrão	1,0693	1,0453	1,1259	2,0916
	Tamanho da amostra	48	50	49	147

TABELA XVII - Estatísticas descritivas referente à variável log (número de bactérias psicrófilas + 1)

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Amplitude	4,5323	4,0174	6,0073	10,1393
	Média	14,1141	8,3426	8,6945	10,4282
	Desvio padrão	1,5352	1,4943	2,0658	3,1676
	Tamanho da amostra	13	13	12	38
Outubro	Amplitude	4,6380	5,2303	9,2461	10,9396
	Média	13,6035	7,9012	9,1663	10,1004
	Desvio padrão	1,5658	1,8573	2,4206	3,1036
	Tamanho da amostra	11	12	13	36
Novembro	Amplitude	4,4348	4,7424	10,9331	10,9331
	Média	12,9446	6,4379	9,1943	9,5256
	Desvio padrão	1,4502	1,4365	3,0459	3,3943
	Tamanho da amostra	9	9	9	27
Dezembro	Amplitude	4,7005	2,3491	2,3300	8,0247
	Média	11,6799	7,5604	7,0920	8,7774
	Desvio padrão	2,3593	1,1785	1,1657	2,6188
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Amplitude	3,4376	4,7751	2,5154	10,3453
	Média	14,0729	7,7547	8,2150	9,9531
	Desvio padrão	1,1928	1,5429	0,9584	3,1494
	Tamanho da amostra	12	13	12	37
TOTAL	Amplitude	7,0817	6,0346	10,9331	12,3800
	Média	13,6153	7,6940	8,6959	9,9615
	Desvio padrão	1,5807	1,6468	2,1404	3,1477
	Tamanho da amostra	48	50	49	147

TABELA XVIII - Estatísticas descritivas referente à variável log (número de bactérias coliformes)

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Amplitude	6,4297	3,9795	3,8338	14,5869
	Média	12,1983	4,1419	4,5007	7,0114
	Desvio padrão	2,0058	1,4091	1,5019	4,1243
	Tamanho da amostra	13	13	12	38
Outubro	Amplitude	6,9503	6,2820	7,6792	13,2324
	Média	12,2082	4,3675	5,0145	6,9969
	Desvio padrão	2,1294	1,5874	2,1753	4,0094
	Tamanho da amostra	11	12	13	36
Novembro	Amplitude	5,3218	5,2309	9,1194	6,9078
	Média	10,5978	3,9799	5,5061	6,6946
	Desvio padrão	1,6287	1,8856	2,8578	3,5688
	Tamanho da amostra	9	9	9	27
Dezembro	Amplitude	5,5905	2,3026	3,9795	12,0692
	Média	11,1279	5,0877	4,0813	6,7656
	Desvio padrão	2,8778	1,1906	2,2717	3,8222
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Amplitude	6,9751	5,3766	3,6997	13,8830
	Média	11,9677	4,5141	4,7387	7,0043
	Desvio padrão	2,0473	1,4226	1,1873	3,8142
	Tamanho da amostra	12	13	12	37
TOTAL	Amplitude	8,3958	6,9078	9,1194	15,2126
	Média	11,7759	4,3204	4,8543	6,9328
	Desvio padrão	2,0425	1,5041	1,9385	3,8518
	Tamanho da amostra	48	50	49	147

TABELA XIX - Estatísticas descritivas referente à variável log (número de bactérias coliformes fecais + 1)

PERÍODO	ESTATÍSTICAS	LOCAL			TOTAL
		PRUP	CFUP	COM	
Setembro	Amplitude	5,8546	4,3395	5,4489	15,4249
	Média	11,0535	1,6218	1,6553	4,8590
	Desvio padrão	1,7591	0,9845	1,2475	4,7195
	Tamanho da amostra	13	13	12	38
Outubro	Amplitude	7,8558	6,6421	6,3544	15,8950
	Média	10,2026	2,2941	1,6305	4,4710
	Desvio padrão	2,1197	1,6175	1,6190	4,2369
	Tamanho da amostra	11	12	13	36
Novembro	Amplitude	6,2820	3,9120	6,2146	13,5498
	Média	8,4720	1,1961	1,7375	3,8019
	Desvio padrão	2,0002	0,8875	1,8811	3,7333
	Tamanho da amostra	9	9	9	27
Dezembro	Amplitude	4,4085	3,2081	4,1179	12,7025
	Média	10,1090	4,4057	2,4976	5,6707
	Desvio padrão	2,2048	1,7957	1,7302	3,8121
	Tamanho da amostra	3	3	3	9
Janeiro e Fevereiro	Amplitude	5,3766	6,9405	4,6775	14,6592
	Média	10,4210	2,6730	2,1792	5,0257
	Desvio padrão	2,0480	1,8540	1,0474	4,1431
	Tamanho da amostra	12	13	12	37
TOTAL	Amplitude	8,6272	8,0392	6,3544	15,8950
	Média	10,1573	2,1469	1,8437	4,6615
	Desvio padrão	2,0976	1,5991	1,4339	4,2084
	Tamanho da amostra	48	50	49	147

TABELA XX - Resultados da análise de variância da variável log (número de bactérias mesófilas)

FONTE DE VARIAÇÃO	G.L.	S.Q.	Q.M.	F	PROBABILIDADE DE SIGNIFICÂNCIA
A	2	468,674	234,337	223,196	0,000
B	4	17,566	4,392	4,183	0,003
AB	8	11,876	1,485	1,414	0,196
Resíduo	132	138,589	1,050	-	-
Total	146	638,715	-	-	-

TABELA XXI - Resultados da análise de variância da variável log (número de bactérias psicrófilas + 1)

FONTE DE VARIAÇÃO	G.L.	S.Q.	Q.M.	F	PROBABILIDADE DE SIGNIFICÂNCIA
A	2	977,112	488,556	156,036	0,000
B	4	27,453	6,863	2,192	0,073
AB	8	29,471	3,684	1,177	0,318
Resíduo	132	413,299	3,131	-	-
Total	146	1 446,603			

TABELA XXII - Resultados da análise de variância da variável log (número de bactérias coliformes)

FONTE DE VARIAÇÃO	G.L.	S.Q.	Q.M.	F	PROBABILIDADE DE SIGNIFICÂNCIA
A	2	1 681,009	840,504	242,569	0,000
B	4	4,554	1,139	0,329	0,858
AB	8	25,363	3,170	0,915	0,506
Resíduo	132	457,382	3,465	-	-
Total	146	2 166,109	-	-	-

TABELA XXIII - Resultados da análise de variância da variável log (número de bactérias coliformes fecais +1)

FONTE DE VARIAÇÃO	G.L.	S.Q.	Q.M.	F	PROBABILIDADE DE SIGNIFICÂNCIA
A	2	2 154,471	1 077,235	395,789	0,000
B	4	36,267	9,067	3,331	0,012
AB	8	35,237	4,405	1,618	0,125
Resíduo	132	359,270	2,722	-	-
Total	146	2 585,793	-	-	-

TABELA XXIV - Resistência(1) de telhas onduladas Eternit

Nº DE ORDEM	TAMANHO DO VÃO (mm)				
	1080	1690	2300	2910	3520
1	279	244	138	118	75
2	292	205	141	92	78
3	301	197	129	91	86
4	278	215	133	103	80
5	286	180	140	98	79
6	263	164	145	103	83
7	270	225	112	114	83
8	312	220	133	98	82
9	219	187	141	94	88
10	257	208	121	101	88
11	281	191	121	102	86
12	284	196	112	85	83
13	264	192	144	100	82
14	232	195	117	85	81
15	208	178	138	90	78

(1) Carga em Kg até a ruptura

TABELA XXV - Soma de quadrados residual da variável $z = \frac{y^{\lambda}-1}{\lambda \dot{y}^{\lambda-1}}$, y: resistência (1), $\dot{y}$ = 140,931

λ	$S(\lambda;Z)$	$L_{max}(\lambda)$	λ	$S(\lambda;Z)$	$L_{max}(\lambda)$
-1,05	12 372,2	-83,15	-0,55	11 141,9	-81,45
-1,00	12 170,0	-82,88	-0,50	11 114,4	-81,41
-0,95	11 985,8	-82,64	-0,45	11 104,4	-81,39
-0,90	11 819,4	-82,41	-0,40	11 112,2	-81,40
-0,85	11 670,6	-82,20	-0,35	11 138,0	-81,44
-0,80	11 539,3	-82,02	-0,30	11 182,0	-81,50
-0,75	11 425,3	-81,86	-0,25	11 244,8	-81,60
-0,70	11 328,6	-81,72	-0,20	11 326,5	-81,71
-0,65	11 249,1	-81,60	-0,15	11 427,7	-81,86
-0,60	11 186,8	-81,51	-0,10	11 548,9	-82,03

(1) Carga em Kg até a ruptura.

TBAELA XXVI - Tabela de análise de variância da variável

$$z = \frac{y^{-0,45}-1}{(-0,45).(140,931)^{-1,45}}, \quad y: \text{resistência(1)}$$

FONTE DE VARIAÇÃO	G.L.	S.Q.	Q.M.	F
Linear	1	276 608,5	276 608,5	1 744,06
Quadrático	1	310,4	310,4	1,96
Cúbico	1	2 102,6	2 102,6	13,26
Quarto grau	1	195,2	195,2	1,23
Resíduo	70	11 104,5	158,6	
Total	74	290 321,2	-	-

(1) Carga em Kg até a ruptura.

TABELA XXVII - Tabela de análise de variância da variável

$$z = \frac{y^{-0,60} - 1}{(-0,60)(140,931)^{-1,60}}$$, y: resistência (1)

FONTE DE VARIAÇÃO	G.L.	S.Q.	Q.M.	F
Linear	1	277 987,1	277 987,1	1 795,41
Cúbico	1	2 196,6	2 196,6	13,90
Resíduo	72	11 374,6	158,0	-
Total	74	291 558,3	-	-

(1) Carga em Kg até a ruptura

FIGURA 1 - Máximo da função de verossimilhança (mesófilas).

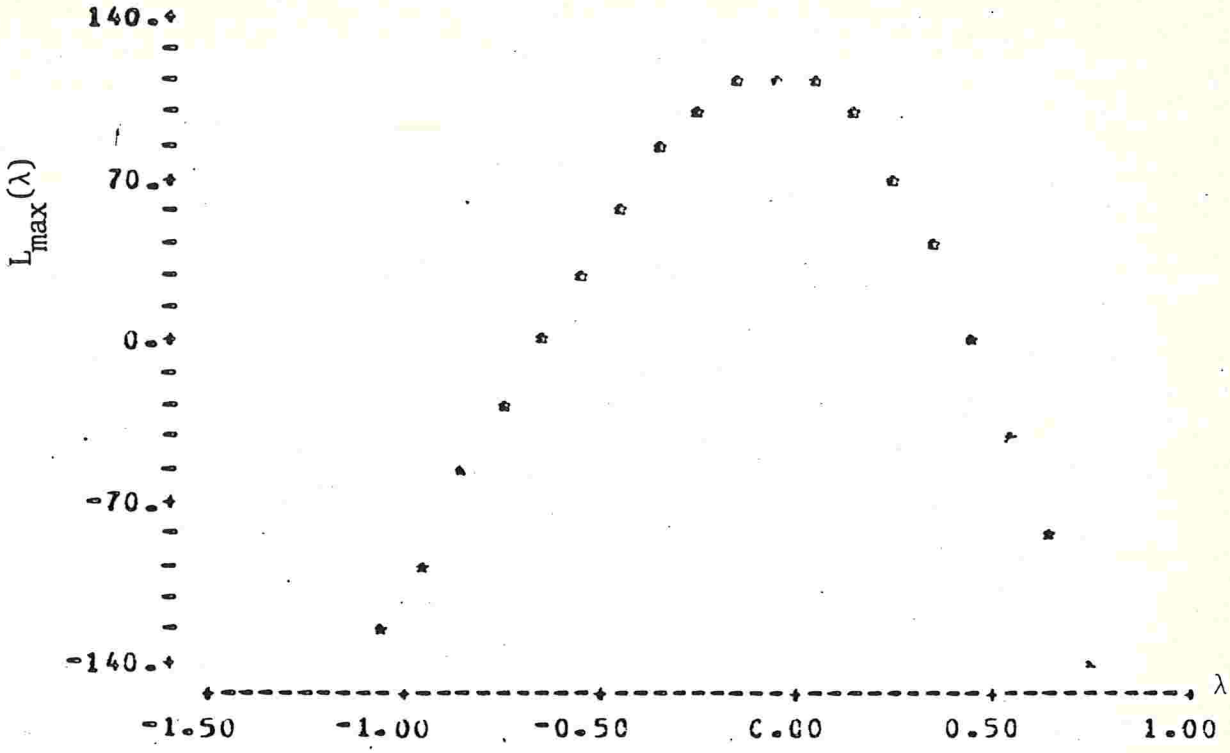


FIGURA 2 - Máximo da função de verossimilhança (psicrófilas).

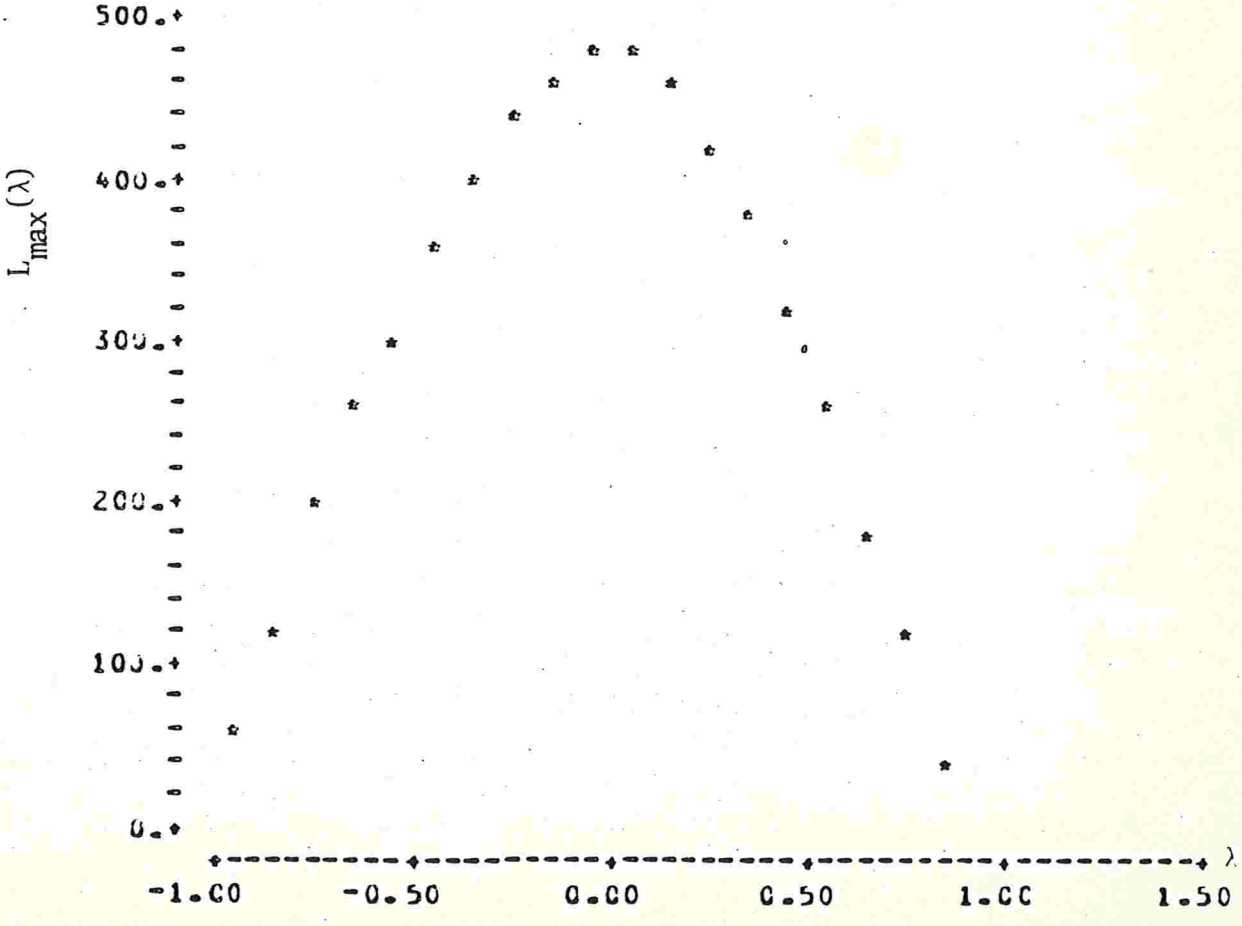


FIGURA 3 - Máximo da função de verossimilhança (coliformes).

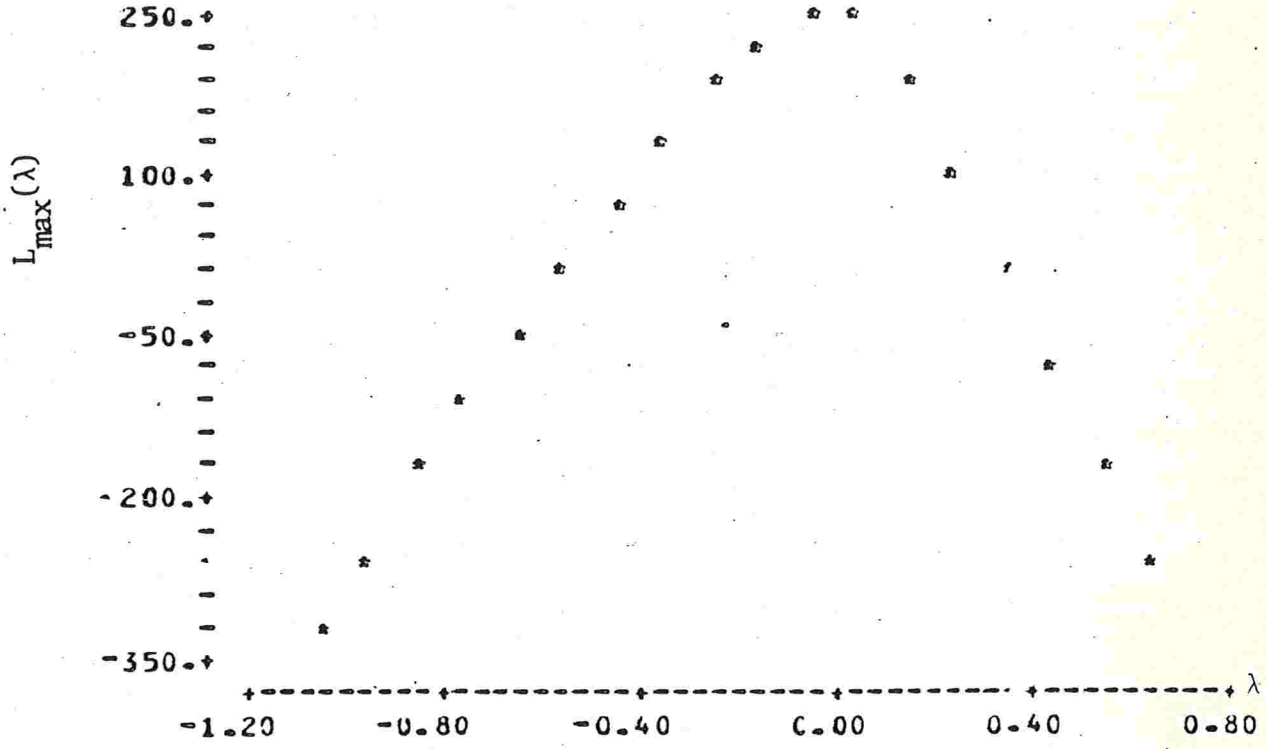


FIGURA 4 - Máximo da função de verossimilhança (coliformes fecais).

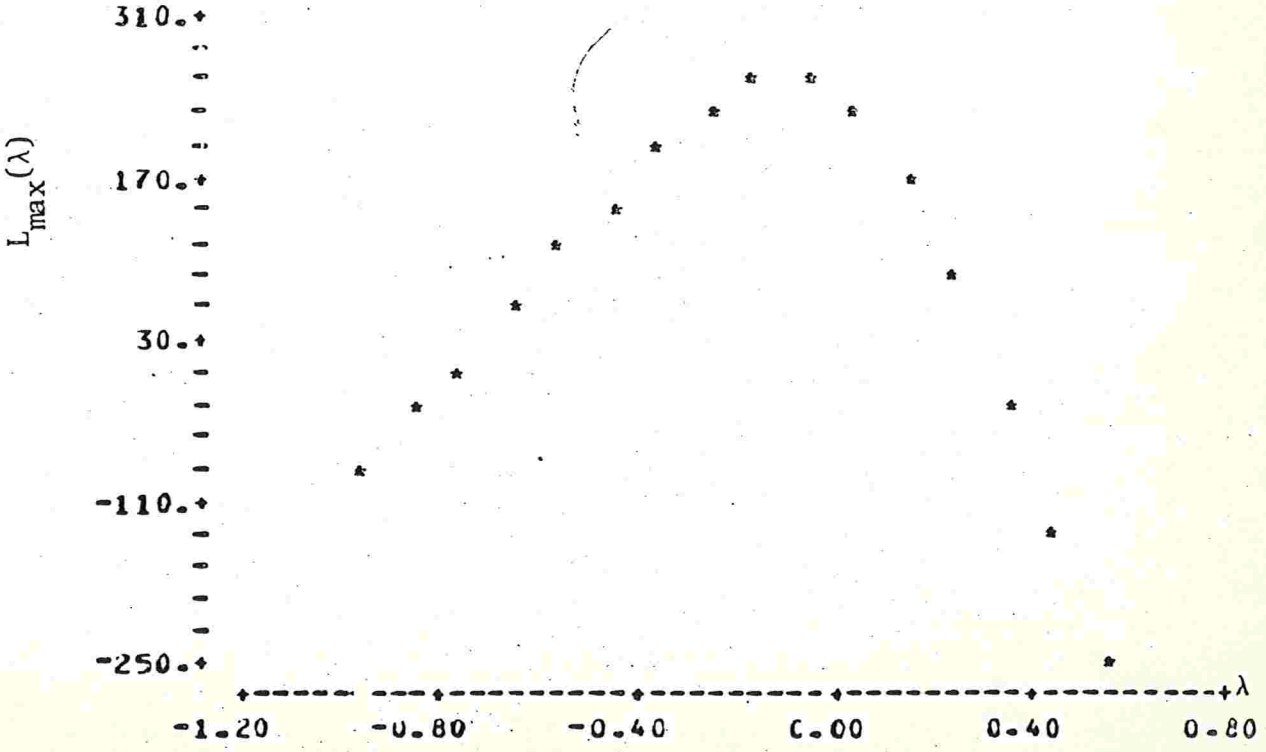
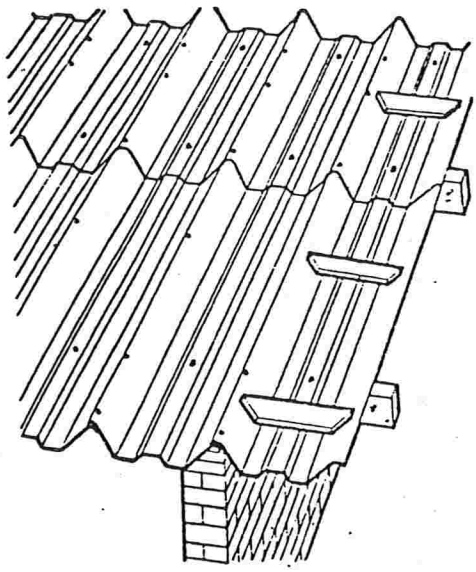
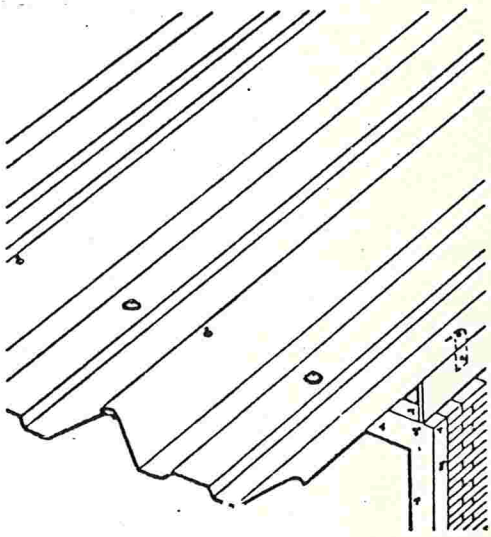


FIGURA 5 - Fixação de telhas onduladas.





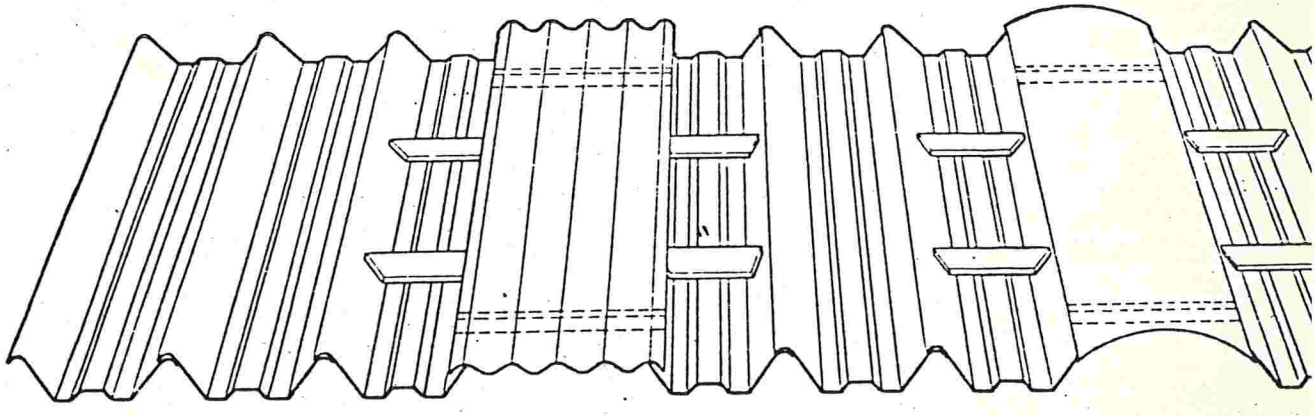
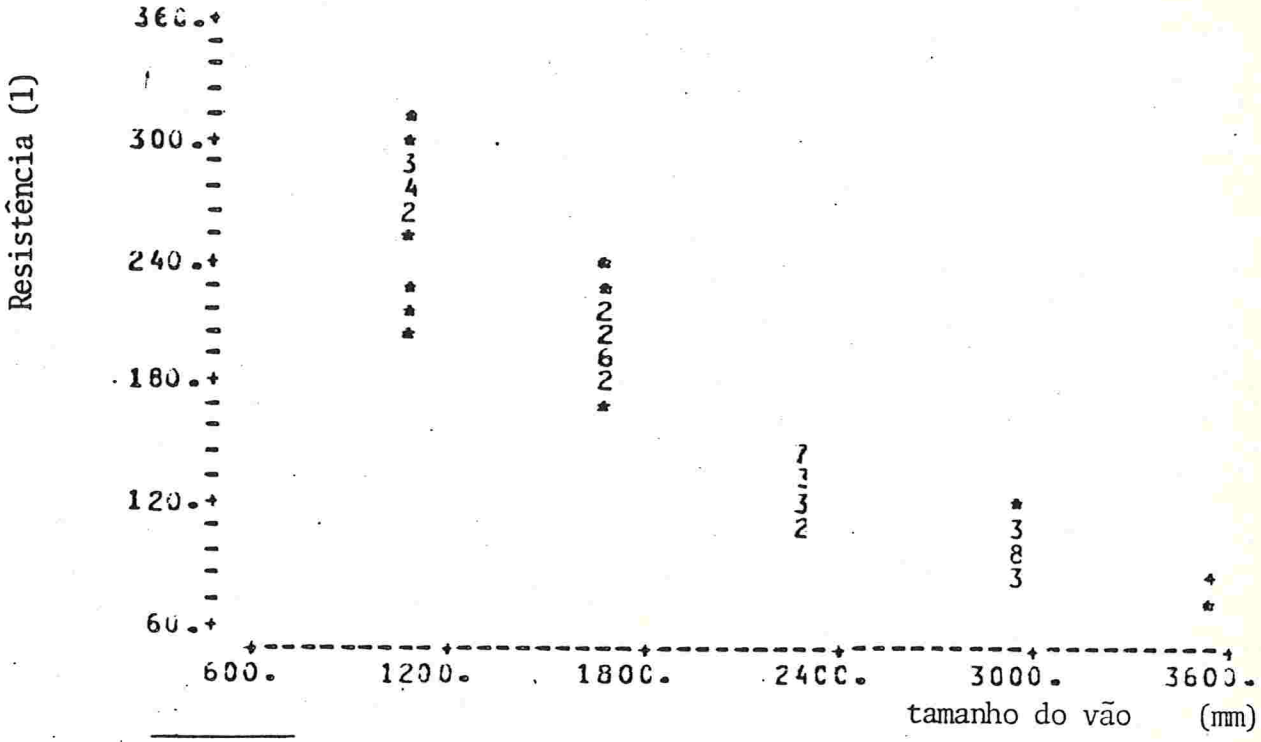
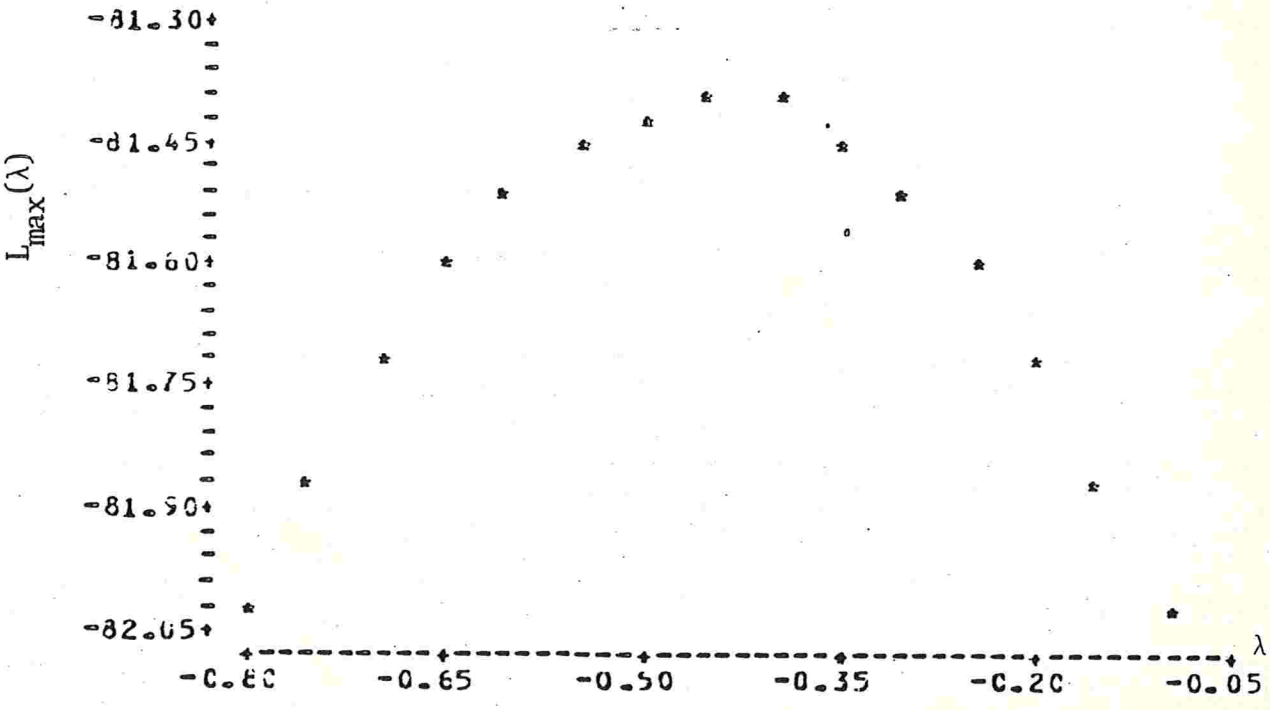


FIGURA 6 - Resistência em função do tamanho do vão.



(1) carga em Kg até a ruptura

FIGURA 7 - Máximo da função de verossimilhança (resistência).



REFERÊNCIAS BIBLIOGRÁFICAS

1 AFIFI, A.A. & AZEN, S.P. *Statistical analysis: a computer oriented approach*. New York, Academic Press, 1972. 366 p.

2 ANDERSON, V.L. Restriction errors for linear models. *Biometrics*, 26:255-68, 1970.

3 ANDERSON, V.L. & MCLEAN, R.A. *Designs of experiments: a realist approach*. New York, Marcel Dekker, 1974.418 p.

4 ANDREWS, D.F. A note on the selection of data transformations. *Biometrika*, 58(2): 249-54, 1971.

5 ANSCOMBE, F.J. The transformation of poisson, binomial and negative binomial data. *Biometrika*, 35: 246-54, 1948.

6 _____. Rejection of outliers. *Technometrics*, 2(2): 123-47, 1960.

7 ANSCOMBE, F.J. & TUKEY, J.W., 1954 apud (TUKEY, J.W. On the comparative anatomy of transformations. *Ann-Math. Stat.*, 28: 602-32, 1957).

8 _____. The examination and analysis of residuals. *Technometrics*, 5(2): 141-60, 1963.

9 ARMITAGE, P. *Statistical methods in medical research*. New York, John Wiley, 1977. 504 p.

10 ATKINSON, A.C. Testing transformations to normality. *J.R. Stat. Soc. B*, 35: 473-9, 1973.

11 _____. Regression diagnostics, transformations and constructed variables. *J.R. Stat. Soc. B*, 44(1): 1-36, 1982.

12 BARTLETT, M.S. The square root transformation in analysis of variance. *J.R. Stat. Soc.*, 3(1): 68-78, 1936.

13 _____. The use of transformations. *Biometrics*, 3: 39-52. 1947.

14 BEALL, G. The transformation of data from entomological field experiments so that the analysis of variance becames applicable. *Biometrika*, 32: 243-62, 1942.

15 BERKSON, J. Application of the logistic function to biossay. *J. Am. Stat. Ass.*, 39: 357-65, 1944.

16 BHATTACHARYYA, G.K. & JOHNSON, R.A. *Statistical concepts and methods*. New York, John Wiley, 1977. 639 p.

17 BLISS, C.I. *Statistical in biology*. New York, McGraw-Hill, 1967. v.1.

18 BOX, G.E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. I. Effects of inequality of variance in the one-way classification. *Ann. Math. Stat.*, 23: 290-302, 1954a.

19 _____. Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and correlation between errors in the two-way classification. *Ann. Math. Stat.*, 25: 484-98, 1954b.

20 BOX, G.E.P. & ANDERSEN, S.L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J.R. Stat. Soc.*, 26(1): 1-34, 1955.

21 BOX, G.E.P. & COX, D.R. An analysis of transformations. *J.R. Stat. Soc.* B, 26: 211-52, 1964.

22 BOX,G.E.P. & JENKINS, G.M. *Time series analysis, forecasting and control*. 2. ed. São Francisco, Holden-Day, 1976. 575 p.

23 BOX, G.E.P. & TIDWELL, P.W. Transformation of the independent variables. *Technometrics*, 4(4): 531-50, 1962.

24 BUSSAB, W.O. *Estudos bacteriológicos de amostras de leite tipo C consumido em João Pessoa, antes e após a pasteurização*. São Paulo, Setor de Estatística Aplicada do Departamento de Estatística do IME/USP, 1981. 10 p. (relatório RAE-SEA-8101).

25 CARROLL, R.J. A robust method for testing transformations. to achieve approximate normality. *J.R. Stat. Soc. B*, 42(10): 71-8, 1980.

26 CHATFIELD, C. & PROTHERO, D.L. Box - Jenkins seasonal forecasting: problems in a case-study. *J.R. Stat. Soc. A*, 136 (3): 295-352, 1973.

27 COCHRAN, W.G. The analysis of variance when experimental errors follow the poisson or binomial laws. *Ann. Math. Stat.*, 11: 335-47, 1940.

28 _____. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 3(1): 22-38, 1947.

29 COCHRAN, W.G. & COX G.M. *Experimental designs*. 2. ed. New York, John Wiley, 1957. 611 p.

30 CONOVER, W.J. *Pratical nonparametric statistics*. 2. ed. New York, John Wiley, 1980. 493 p.

31 CUNHA, A.M.S. *Não normalidade e testes de hipóteses sobre variâncias e médias*. São Paulo, Instituto de Matemática e Estatística da USP, 1978.95 p. (Tese, Mestre em Estatística).

32 CURTISS, J.H. On transformations used in the analysis of variance. *Ann. Math. Stat.*, 14: 107-22, 1943.

33 DRAPER, N.R. & SMITH, H. *Applied regression analysis*. 2. ed. New York, John Wiley, 1981. 709 p.

34 EFRON, B. Transformation theory: how normal is a family of distributions? Stanford, Stanford University, Department of Statistics, 1981. 40 p (Technical report, 69).

35 EISENHART, C. The assumptions underlying the analysis of variance. *Biometrics*, 3(1): 1-21, 1947 .

36 FINNEY, D.J. On the distribution of a variate whose logaritm is normally distributed. *J.R. Stat. Soc.* (7): 155-61, 1941.

37 _____. *Statistical methods in biological assay*. 2.ed. New York, Hafner, 1964.

38 _____. *Probit analysis*. 3. ed. Cambridge, Cambridge University Press, 1971.

39 _____. Transformation of observations for statistical analysis. *Cotton Grow. Rev.*, 50: 1-14, 1973.

40 FISHER, R.A. Gene frequencies in a cline determined by selection and diffusion. *Biometrics*, 6: 353-61, 1950.

41 FISHER, R.A. & YATES, F. *Tabelas estatísticas para Biologia Medicina e Agricultura*. Edição Brasileira. São Paulo, Polígono, 1971. 150 p.

42 HERNANDEZ, F. & JOHNSON, R.A. The large sample behavior of transformations to normality. *J. Am. Stat. Ass.*, 75 (372): 855-61, 1980.

43 HINKLEY, D.V. On power transformations to symmetry. *Biometrika*, 62(1): 101-11, 1975.

44 HUBER, P.J. *Robust statistical procedures*. Philadelphia, Soc. Indus. Appl. Math., 1977.

45 JOHN, J.A. & DRAPER, N.R. An alternative family of transformations. *Appl. Stat.*, 29(2): 190-7, 1980.

46 JOHNSON, N.L. & LEONE, F.C. *Statistical and experimental design*. New York. John Wiley, 1964. v.2. 399 p.

47 JOINER, B.L. & ROSENBLATT, J.R. Some properties of the range in samples from Tukey's symmetric lambda distributions. *J. Am. Stat. Ass.*, 66(334): 394-9, 1971.

48 KEMPTHORNE, O. *Design and analysis of experiments*. New York, John Wiley, 1952. 631 p.

49 MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. *Introduction to the theory of statistics*. 3. ed. New York, McGraw-Hill, 1974. 564 p.

50 MOORE, P.G. Transformation to normality using fractional powers of the variable. *J. Am. Stat. Ass.*, 52:237-46, 1957.

51 MOORE, P.G. & TUKEY, J.W. Answer to query no 112. *Biometrics*, 10: 562-8, 1954.

52 NARULA, S.C. & WELLINGTON, J.F. Absolute error regression: a state of the art survey. *Internat. Stat. Rev.*, 1982.

53 NETER, J. & WASSERMAN, W. *Applied linear statistical models*. Homewood (Illinois), Richard D. Irwin, 1974. 842 p.

54 NEYMAN, J. & SCOTT, E. Correction for bias introduced by a transformation of variables. *Ann. Math. Stat.*, 31: 643-55, 1960.

55 PERES, C.A. Testing the effect of blocking in a randomized complete block design (RCBD). *Commun. Stat. Theor. Meth.*, A,10 (23). 2447-59, 1981.

56 PERES, C.A. & SALDIVA, C.D. *Planejamento de experimentos*. In: SIMPÓSIO DE PROBABILIDADE E ESTATÍSTICA, 5. São Paulo, 1982. São Paulo, Instituto de Matemática e Estatística da USP, 1982, 98 p.

57 RAO, C.R. *Linear statistical inference and its application*. 2. ed. New York, John Wiley, 1973. 625 p.

58 SCHEFFÉ, H. *The analysis of variance*. New York, John Wiley, 1959. 477 p.

59 SCHLESSELMAN, J. Power families: a note on the Box and Cox transformation. *J.R. Stat. Soc.*, B, 33: 307-11,1971.

60 SEARLE, S.R. *Linear models*. New York, John Wiley, 1971. 532 p.

61 SNEDECOR, G.W. *Statistical methods*. 5. ed. Ames, Iowa State University Press, 1956. 534 p.

62 SNEDECOR, G.W. & COCHRÀN, W.G. *Statistical methods*. 7.ed. Iowa State University Press, 1980. 507 p.

63 SPITZER, J.J. A Monte Carlo investigation of the Box-Cox transformation in small samples. *J.Am. Stat. Ass.*, 73(363): 488-95, 1978.

64 STEEL, R.G. & TORRIE, J.H. *Principles and procedures in statistics*. New York, McGraw-Hill, 1960. 481 p.

65 THÖNI, H. Transformations of variables used in the analysis of experimental and observational data - a review. Ames. Statistical Laboratory (Iowa State University), 1978. 60 p (Technical report, 7).

66 TUKEY, J.W. One degree of freedom for non-additivity. *Biometrics*, 5: 232-42, 1949.

67 _____. Dyadic anova, an analysis of variance for vectors. *Hum. Biol.*, 21: 232-42, 1950.

68 _____. On the comparative anatomy of transformations. *Ann. Math. Stat.*, 28: 602-32, 1957.

69 _____. The practical relationship between the common transformations of percentages fractions and of amounts. Princeton, Statistical Research Group, 1960. (Technical Report, 36).

70 YATES, F. Incomplete latin squares. *J. Agric. Scie.*, 26: 301-15, 1936.

71 YATES, F. & HALE, R.W. The analysis of latin squares when two or more rows, columns or treatments are missing. *J. R. Stat. Soc.*, 6: 67-79, 1939.

72 WASOW, W. On the asymptotic transformation of certain distributions into the normal distribution. *Proc. Symp. Appl. Math.*, 6: 251-9, 1956.

73 WILK, M.B. & KEMPTHORNE, O. Non- additivities in a latin square design. *J. Am. Stat. Ass.*, 52: 218-36, 1957.

74 WINER, B.J. *Statistical principles in experimental design.* London, McGraw-Hill, 1970. 671 p.

75 WINSOR, C.P. & CLARKE, G.L., 1940 apud (SNEDECOR, G.W. & COCHRAN, W.G. *Statistical methods*. 7. ed. Iowa State University Press, 1980, pag. 291).

76 WOODING, W.M. The computation and used of residuals in the analysis of experimental data. *J. Qual. Technol.*, 1 (3): 175-88, 1969.