UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
CADERNOS DE MATEMÁTICA E ESTATÍSTICA
SÉRIE B: TRABALHO DE APOIO DIDÁTICO

ANÁLISE EXPLORATÓRIA DE DADOS UTILIZANDO O MINITAB

DINARA W. XAVIER FERNANDEZ
JOÃO RIBOLDI
CÉSAR EDUARDO DA SILVA DORNELES
CLÁUDIA ALGAYER DA ROSA

SÉRIE B, Nº 38
PORTO ALEGRE, NOVEMBRO DE 1997.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA
DEPARTAMENTO DE ESTATÍSTICA

ANÁLISE EXPLORATÓRIA DE DADOS UTILIZANDO O MINITAB

Dinara W. Xavier Fernandez

João Riboldi

César Eduardo da Silva Dorneles

Cláudia Algayer da Rosa

NOVEMBRO 97

ÍNDICE

ANÁLISE EXPLORATÓRIA DE DADOS UTILIZANDO O MINITAB

1. CONSIDERAÇÕES GERAIS

As técnicas de análise exploratória constituem um conjunto de métodos para sondagem de informações e fornecimento de evidências sobre os dados. Essas evidências não são suficientes para decisões. Para tal, é necessária a utilização dos métodos de inferência estatística que pressupõem modelos probabilísticos para a tomada de decisões.

A análise exploratória de dados constitui uma fase preliminar do processo decisório, onde se buscam indicações de possíveis modelos a serem utilizados para a inferência estatística. Serve para detectar se as suposições feitas sobre o comportamento probabilístico das variáveis consideradas são satisfeitas, pelo menos de forma aproximada.

As técnicas de análise exploratória auxiliam na detecção de valores aberrantes ("outliers"), de desvios nas suposições de normalidade, variância constante (homocedasticidade) e independência. Estas são suposições necessárias para a aplicação de técnicas clássicas de inferência estatística.

As técnicas de análise exploratória mais comumente empregadas são o diagrama de ramo e folhas (steam-and-leaf), o resumo de 5-números e o Box-Plot. Existem vários aplicativos disponíveis para obtenção desses procedimentos.

À medida em que os conteúdos forem apresentados neste material, serão indicados os respectivos comandos para o software MINITAB.

2. EXEMPLO

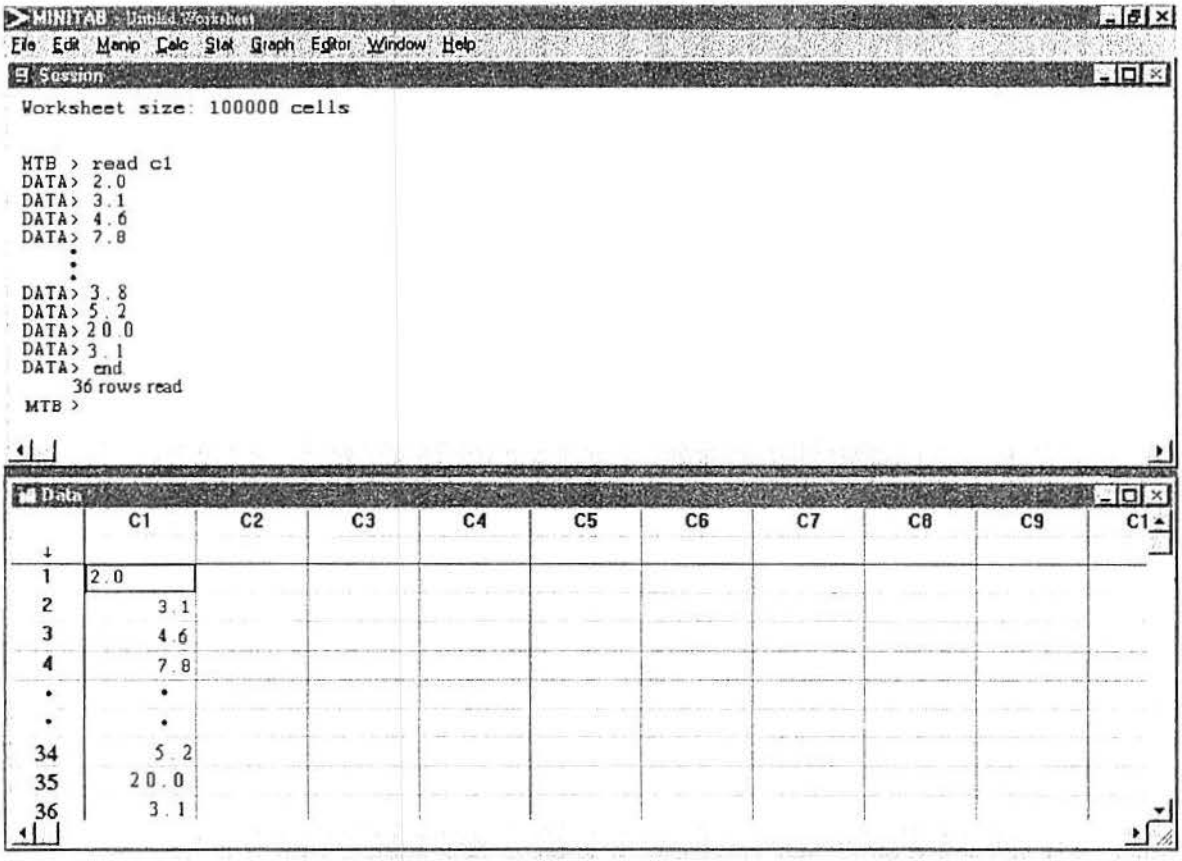
Para ilustração das diferentes técnicas de análise exploratória utilizar-se-á os seguintes dados, retirados de Alves e Sarries (1994).

Altura, em metros, de plantas de determinada cultura: 2,0 3,1 4,6 7,8 2,1 4,7 8,9 1,5 2,4 3,8 5,4 1,6 2,5 3,9 5,4 1,7 2,8 4,1 6,2 1,7 2,9 4,2 6,4 2,0 3,0 4,5 6,5 2,2 3,8 5,0 10,1 2,2 3,8 5,2 20,0 3,1

3 - CRIAR, LISTAR E SALVAR ARQUIVO DE DADOS ATRAVÉS DO MINITAB.

3.1- CRIAR ARQUIVO DE DADOS:

Para a entrada dos dados do EXEMPLO utiliza-se o comando READ ou se clica na coluna c1 e dá-se entrada nos dados, como segue na ilustração da tela abaixo.



3.2- LISTAR ARQUIVO DE DADOS.

Para listar os dados, basta pedir para exibi-los através do comando **PRINT** indicando a coluna a ser mostrada. Assim no EXEMPLO fica: **PRINT C1**.

3.3- SALVAR ARQUIVO DE DADOS.

Para salvar os dados na planilha utiliza-se o comando **SAVE** com a sintaxe: **SAVE** '(drive de trabalho):(nome do arquivo)'.

Assim para listar os dados da planilha e salvá-los tem-se a seguinte visualização na tela do MINITAB.


```
MTB > print c1

Data Display

C1
   2.0    3.1    4.6    7.8    2.1    4.7    8.9    1.5    2.4    3.8    5.4
   1.6    2.5    3.9    5.4    1.7    2.8    4.1    6.2    1.7    2.9    4.2
   6.4    2.0    3.0    4.5    6.5    2.2    3.8    5.0   10.1    2.2    3.8
   5.2   20.0    3.1

MTB > save 'a:testa'
Saving worksheet in file: a:testa.MTW
MTB >
```

4. DIAGRAMA DE RAMO E FOLHAS

Para analisar de forma adequada e eficiente um conjunto de informações é necessário que os dados estejam organizados de forma ordenada, visando sumarisar o comportamento dos dados e ao mesmo tempo fornecer uma idéia sobre o conjunto dos valores. Uma maneira simples de se atingir estes objetivos é através do Diagrama de Ramo e Folhas. Esse diagrama é uma variação das tabelas de freqüências, com uma forma de apresentação que facilita a observação de características importantes dos dados, tais como simetria, presença de valores discrepantes ("outliers"), concentração de observações e lacunas entre observações.

Considerando que os dados são constituídos de pelo menos dois dígitos, a construção do diagrama de ramo e folhas consiste em decompor os valores em duas partes: um ramo constituído por um ou mais dígitos principais e as folhas constituídas pelos dígitos restantes.

A maneira mais simples de construir o diagrama de ramo e folhas é colocar os dígitos principais (inteiros) como ramos e os dígitos secundários (decimais) como folhas.

Algumas variações do diagrama de ramo e folhas, são apresentados nas formas:

- compacta: onde considera-se como ramo dois ou mais ramos do caso mais simples
- desdobrada: onde cada ramo do caso mais simples se desdobra em dois ou mais ramos

- dividida: quando o caso mais simples forma muitos ramos com baixa concentração de folhas em alguns deles, é preferível trabalhar com ramos divididos.

Para o EXEMPLO, onde os inteiros constituem os ramos e os decimais as folhas, tem-se:

```
1. | 5 6 7 7
2. | 0 0 1 2 2 4 5 8 9
3. | 0 1 1 8 8 8 9
4. | 1 2 5 6 7
5. | 0 2 4 4
6. | 2 4 5
7. | 8
8. | 9
9. |
10. | 1
11. |
12. |
13. |
14. |
15. |
16. |
17. |
18. |
19. |
20. | 0
```

Com base no diagrama de ramo e folhas, pode-se, dentre outras, fazer as seguintes considerações:

a) O valor 20,0 se destaca dos demais, sendo seguramente uma observação discrepante ("outlier").
b) Os dados estão relativamente concentrados entre 1,5 e 6,5.
c) Há uma lacuna acentuada a partir do valor 10,1.
d) Há uma acentuada assimetria em relação aos valores maiores.
e) Com base no comportamento dos dados parece que não seria razoável admitir a distribuição normal para o estudo dos dados.

5 - **OBTENÇÃO DO RAMO E FOLHAS ATRAVÉS DO MINITAB**

Para a obtenção do diagrama de Ramo e Folhas se digita **STEM-AND-LEAF** e a coluna onde se encontram os dados. Com este processo, os dígitos principais (ramos) serão 0, 1 e 2; e os dígitos secundários (folhas) serão números de 0 a 9.

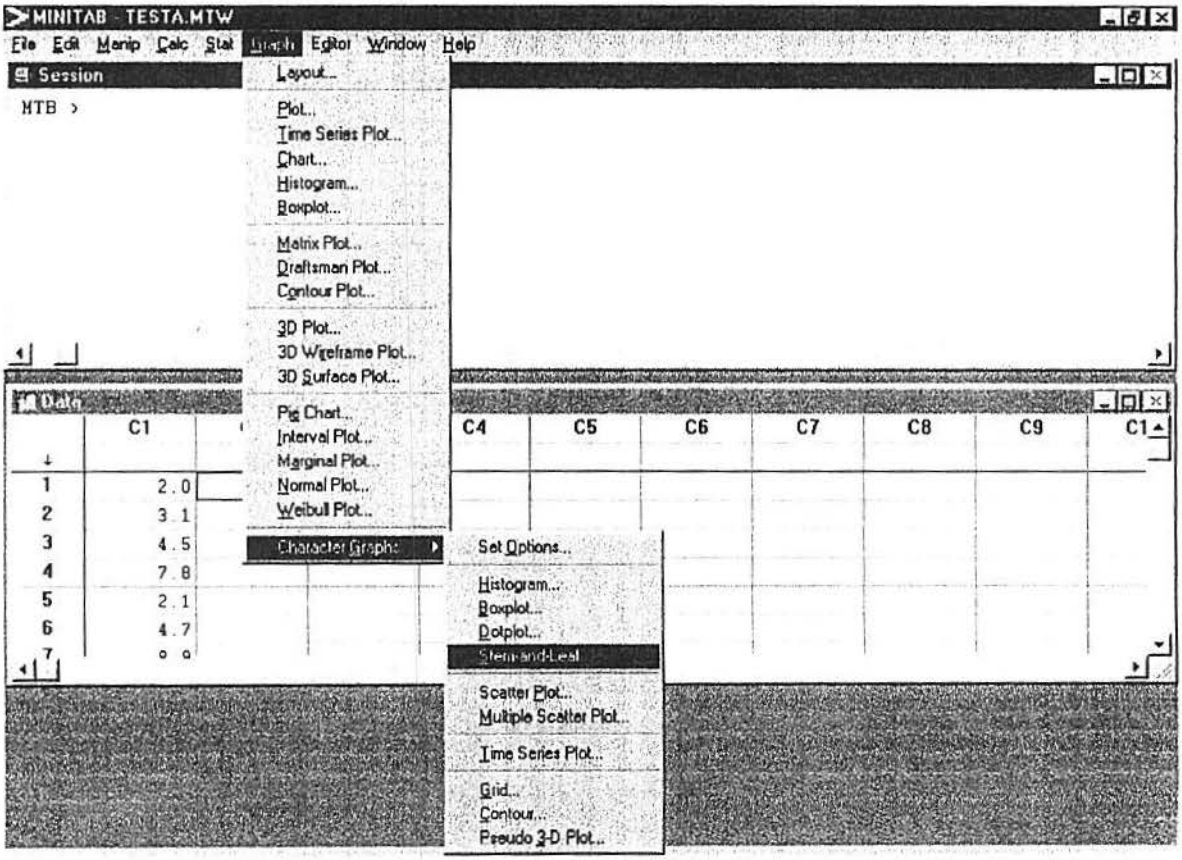
MINITAB - testa.MTW

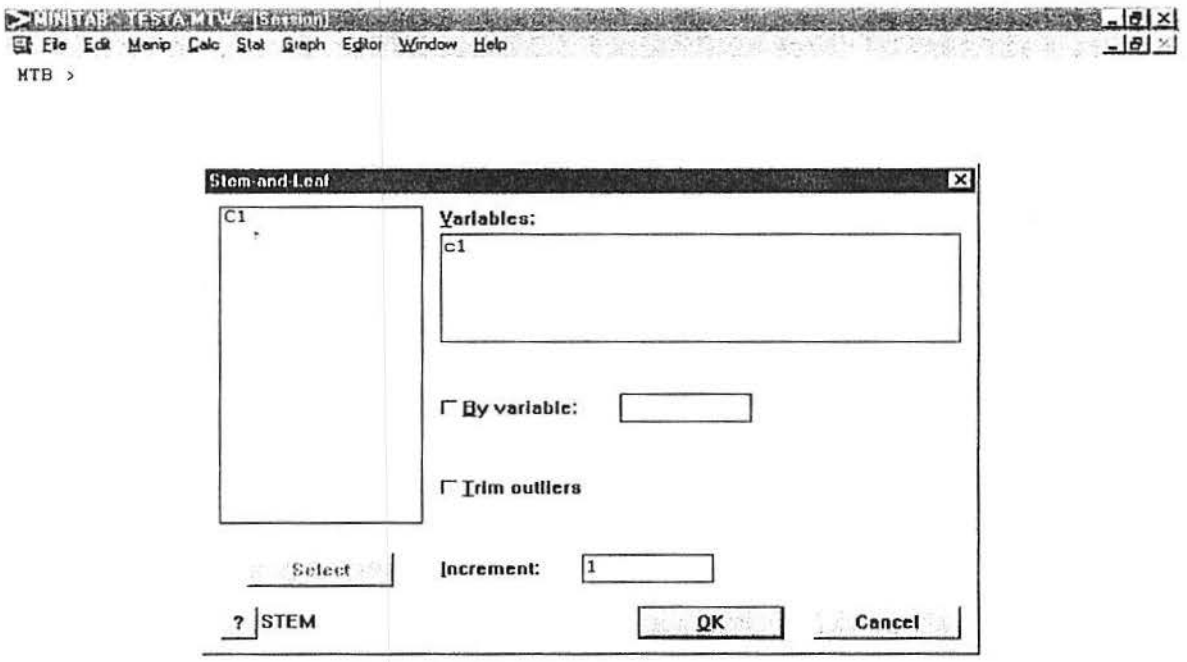
File Edit Manip Calc Stat Graph Editor Window Help

Session

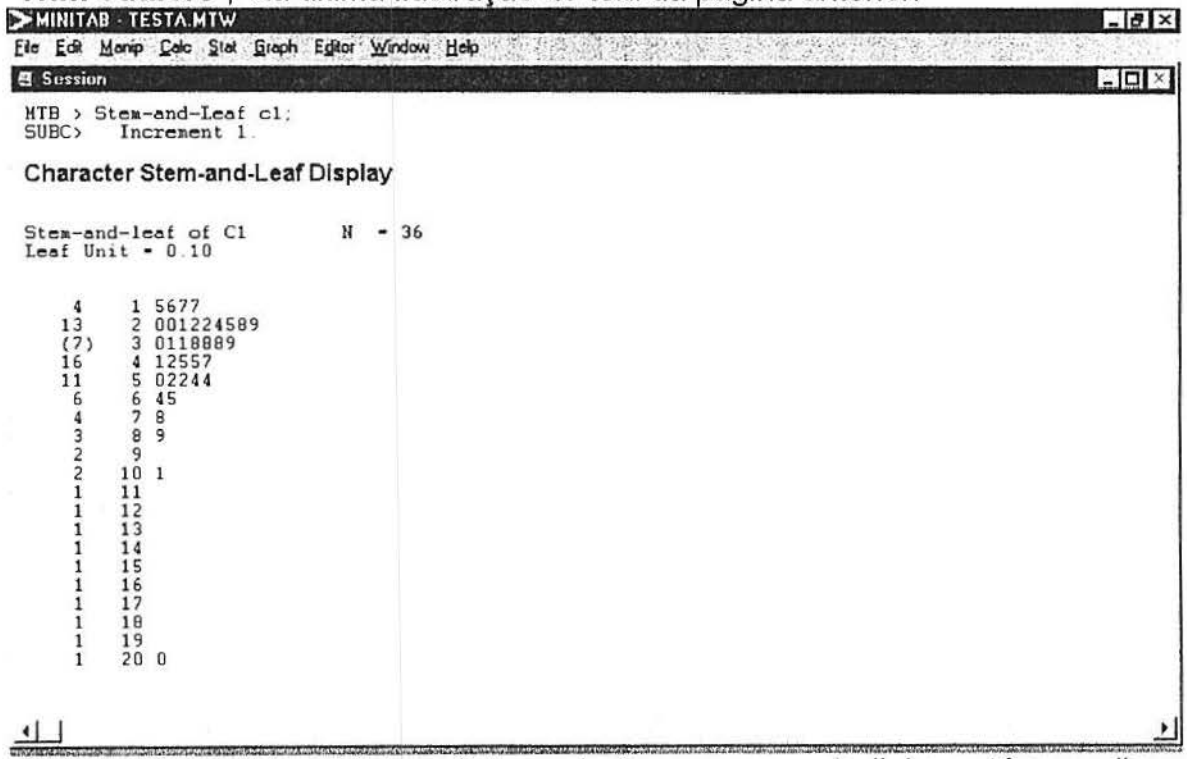
```
MTB > stem-and-leaf c1

Character Stem-and-Leaf Display

Stem-and-leaf of C1       N   = 36
Leaf Unit = 1.0

    4     0 1111
  (16)    0 2222222223333333
   16     0 4444455555
    6     0 667
    3     0 8
    2     1 0
    1     1
    1     1
    1     1
    1     1
    1     2 0

MTB >
```

- Ou, de outra forma, pode-se clicar na alternativa "**Graph**" do menu principal selecionando a opção "**Character Graphs**" e, após, na alternativa "**Stem-and-Leaf**", escrevendo c1 quando for solicitada a variável em questão. Como opção, pode-se selecionar em "**increment**" o valor 1 para se obter os inteiros como ramo e a primeira decimal como folhas. Seguem as telas do MINITAB:





- Após este processo será apresentado um diagrama de ramo e folhas mais detalhado que o anterior pois neste caso foi dada a opção de escolher o ramo e as folhas. Seria possível ainda excluir os valores discrepantes através da opção "**Trim outliers**", da última ilustração de tela da página anterior.



Observação: o valor entre parênteses indica que naquela linha está a mediana da distribuição.

6. RESUMO DE CINCO-NÚMEROS

Além de uma técnica de apresentação de forma ordenada para o conjunto dos dados é importante pensar em uma forma de resumir a informação contida nos dados. A forma costumeiramente empregada é através da média e do desvio padrão, que pode ser de pouca utilidade, pois são fortemente afetados por valores discrepantes, e nada informam quanto aos dados se distribuírem assimetricamente para valores grandes (pequenos).

Uma forma mais adequada de resumir os dados é através do resumo de cinco-números, composta de

n		
	Md	
Q_I		Q_S
L_I		L_S

onde n = nº de observações

Md = Mediana

Q_I = Quartil Inferior (25% das observações são inferiores ao 1º Quartil)

Q_S = Quartil Superior (75% das observações são inferiores ao 3º Quartil)

L_I = Extremo Inferior (menor valor)

L_S = Extremo Superior (maior valor)

Para o EXEMPLO tem-se

	Nº de obs.: 36		
Md	3,8		
Q_I	2,3	5,3	Q_S
L_I	1,5	20,0	L_S

A mediana, os extremos inferior e superior são identificados automaticamente no conjunto de dados ordenados.

Na determinação dos quartis:

- se o número n de observações for ímpar

  Q_I é o valor de ordem $\frac{n + 1}{4}$

  e Q_S é o valor de ordem $\frac{3(n + 1)}{4}$

- se o número n de observações for par

$Q_I = l_i + (L_i - l_i)\,p$ onde p é a parte fracionária da divisão $\frac{n+1}{4}$ e l_i e L_i são, respectivamente, valores de ordem inferior e superior a $\frac{n+1}{4}$

$Q_S = l_s + (L_s - l_s)\,q$ onde q é a parte fracionária da divisão $\frac{3(n+1)}{4}$ e l_S e L_S são, respectivamente, valores de ordem inferior e superior a $\frac{3(n+1)}{4}$

Para o EXEMPLO tem-se n=36,

$\frac{n+1}{4} = \frac{36+1}{4} = 9{,}25$ e p=0,25 e

$\frac{3(n+1)}{4} = \frac{3(36+1)}{4} = 27{,}75$ e q=0,75, então:

l_i = valor de ordem 9 = 2,2
L_i = valor de ordem 10 = 2,4
l_S = valor de ordem 27 = 5,2
L_S = valor de ordem 28 = 5,4

Assim

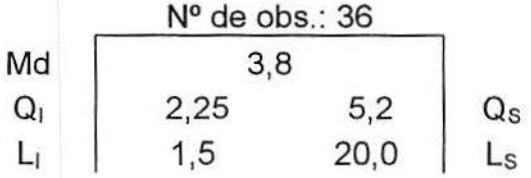
$Q_I = l_i + (L_i - l_i)p = 2{,}2 + (2{,}4 - 2{,}2)(0{,}25) = 2{,}3$
$Q_S = l_S + (L_S - l_S)q = 5{,}2 + (5{,}4 - 5{,}2)(0{,}75) = 5{,}3$

O diagrama de cinco-números fornece algumas informações úteis sobre o comportamento dos dados. Dentre estas, a diferença entre os extremos 20,0 - 1,5 = 18,5 é bem maior que a diferença entre os quartis 5,3 - 2,3 = 3,0. Este fato é um indicador da assimetria para valores grandes e do grande espalhamento dos dados.

7 - **OBTENÇÃO DO RESUMO DE CINCO NÚMEROS ATRAVÉS DO MINITAB.**

O MINITAB não fornece diretamente o Resumo de cinco-números. Através do comando "DESCRIBE" obtém-se todos os itens necessários para o Resumo de cinco-números, traçado em outro aplicativo (WORD, por exemplo), como ilustrado a seguir:

```
Worksheet size: 100000 cells

MTB > DESCRIBE C1

Descriptive Statistics

Variable    N    Mean   Median   TrMean    StDev   SEMean
C1         36   4.444    3.800    3.963    3.365    0.561

Variable    Min      Max      Q1      Q3
C1        1.500   20.000   2.250   5.200

MTB >
```



8. **BOX-PLOT**

O Box-Plot é uma variação gráfica do resumo de cinco-números, e é um excelente procedimento para:

a) Comparar variáveis de acordo com as características de sua distribuição

b) Identificar características importantes da distribuição dos dados, tal como assimetria.

c) Identificar pontos que se destacam no conjunto de dados.

Para a construção do Box-Plot precede-se da seguinte maneira:

i) Define-se uma quantidade $d_F = Q_S - Q_I$, onde Q_I e Q_S são, respectivamente, o quartil inferior e o quartil superior.

ii) Constrói-se uma caixa retangular com comprimento d_F, isto é, inicia em Q_I e termina em Q_S. Nesta caixa estará a maior concentração das observações (o "grosso" das observações).

iii) Constrói-se um traço transversal na caixa indicando a posição da mediana.

iv) Os limites críticos para as observações são definidos como LCI $=Q_I - 1,5\ d_F$ e LCS $= Q_S + 1,5\ d_F$ onde LCI e LCS são, respectivamente, limite crítico inferior e limite crítico superior.

v) A partir de Q_I (Q_S) é traçado uma linha até o limite crítico inferior (superior) ou até a menor (maior) observação.

vi) Valores que ultrapassam os limites críticos são registrados no gráfico.

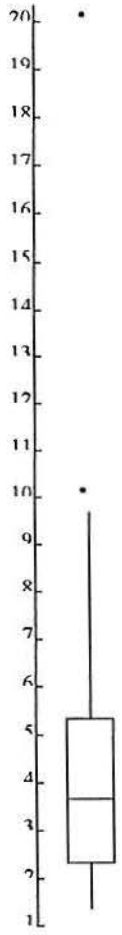
Para o EXEMPLO, tem-se: Q_I=2,3; Q_S=5,3; Mediana=3,8 e

$d_F = 5,3 - 2,3 = 3,0$

LCI=Q_I -1,5 d_F = 2,3 - 1,5 (3) =2,3 - 4,5 = -2,2 (até 1,5 que é a menor observação)

LCS=Q_S +1,5 d_F = 5,3 + 1,5 (3) = 5,3 + 4,5 = 9,8

O Box-Plot para o EXEMPLO é:

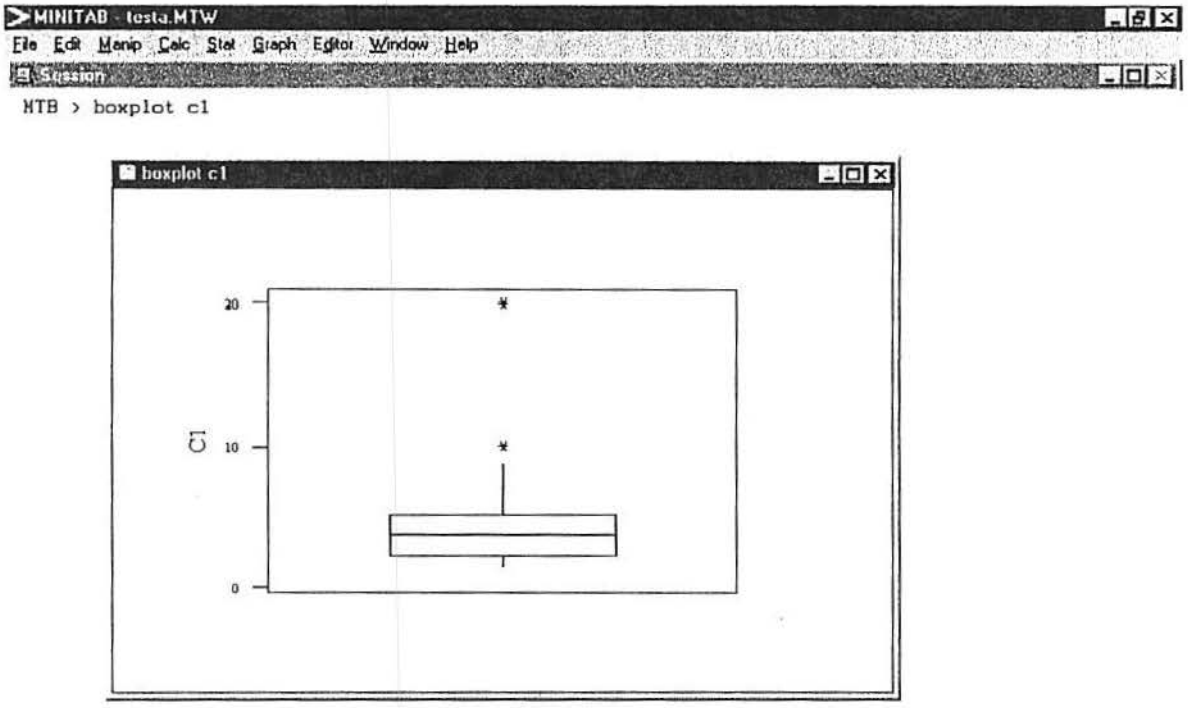


No EXEMPLO, o segmento de reta não atinge o limite crítico inferior, já que a menor observação é 1,5. No entanto o limite crítico superior está bem distante da maior observação. Esta observação é caracterizada agora com mais destaque como observação discrepante.

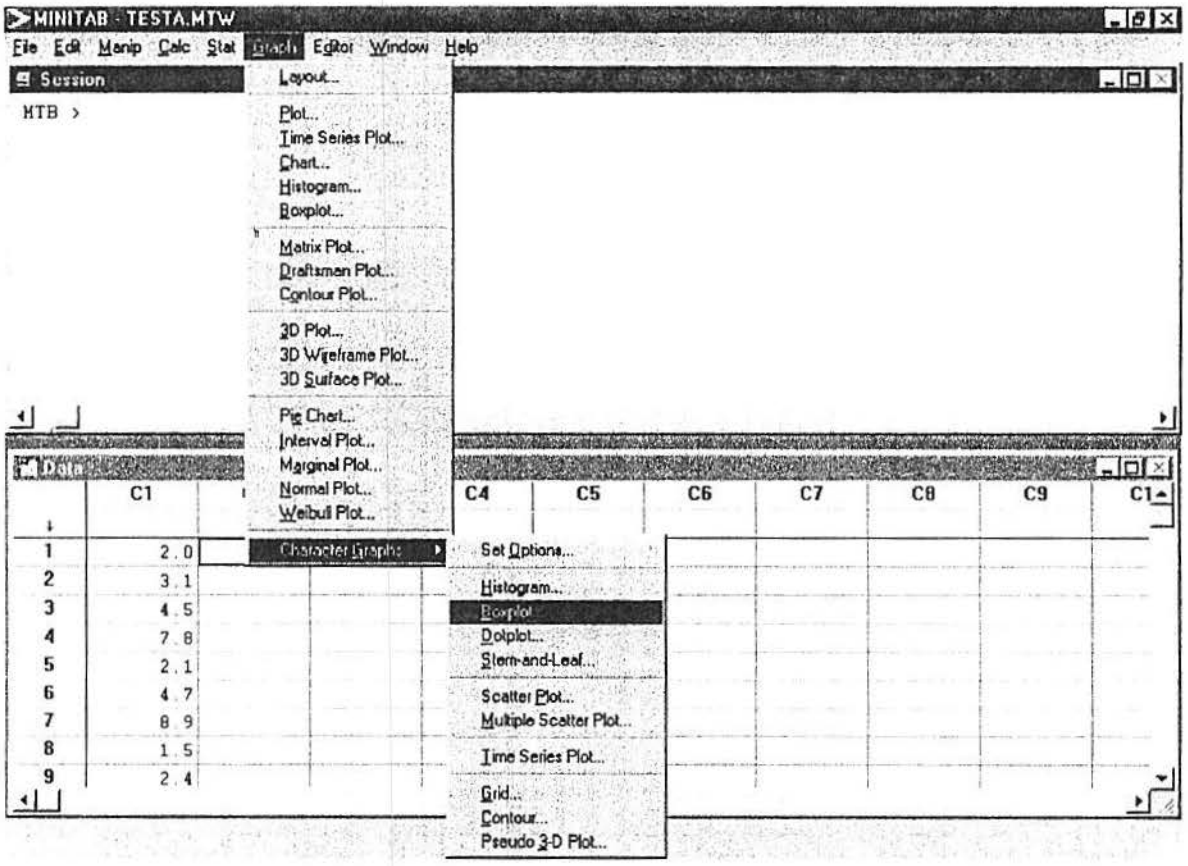
Os valores que forem superiores ao LCS ou inferiores ao LCI são considerados pontos discrepantes.

9 - **OBTENÇÃO DO BOXPLOT ATRAVÉS DO MINITAB**

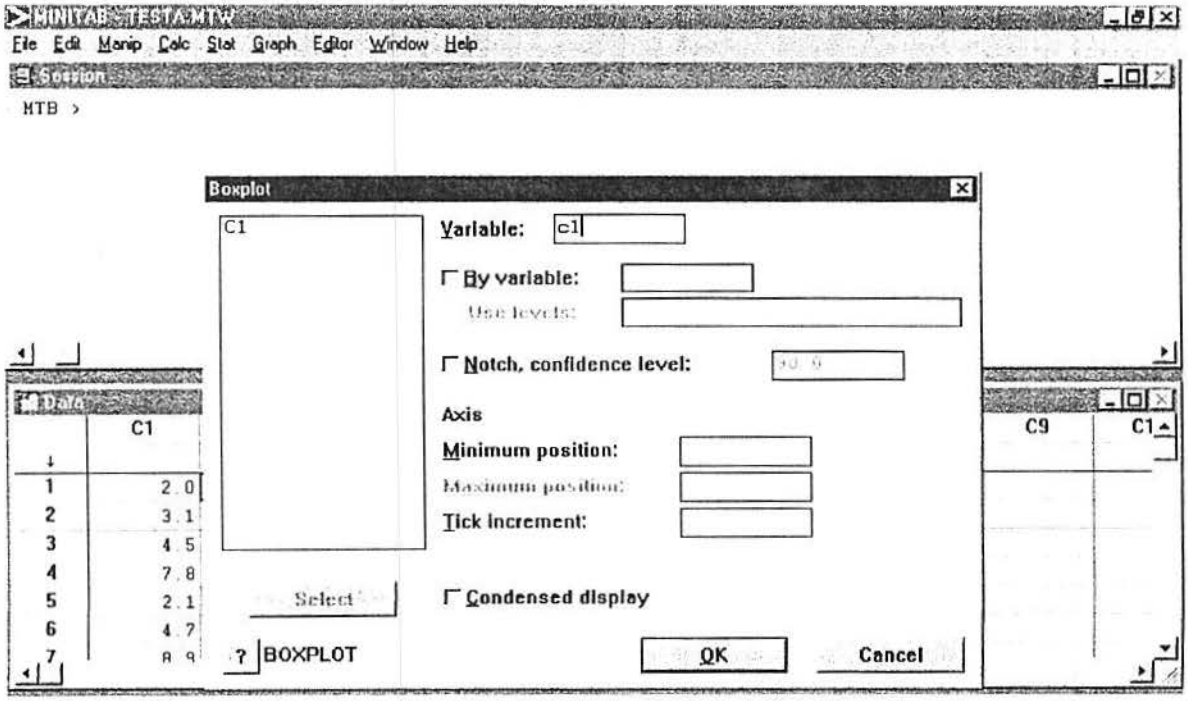
Inicialmente, para a obtenção do Boxplot, digita-se **BOXPLOT** e a coluna na qual estão os dados.



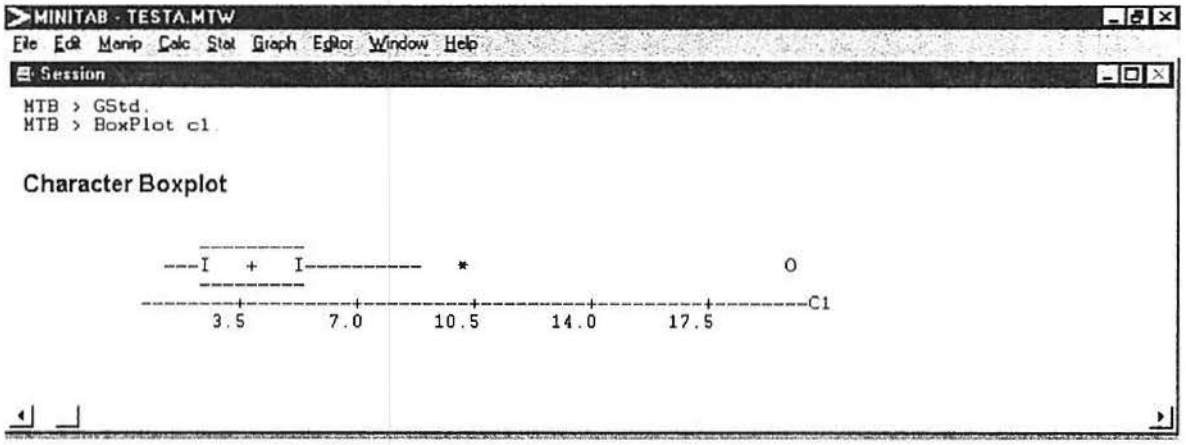
De outra forma, pode-se clicar na alternativa "**Graph**" do menu principal selecionando a opção "**Character Graphs**" e, após, na alternativa "**Boxplot**", como ilustrado a seguir:



Surgirá na tela o cursor pedindo para mostrar a variável, no caso, c1.



-Desta maneira, o boxplot será apresentado novamente, só que desta vez na horizontal e um pouco mais detalhado.



10 - **RE-EXPRESSÃO (transformações)**

Muitas vezes, para simplificar a análise dos dados, é necessário reexpressá-los através de outra escala (por exemplo, logarítmica ou raiz quadrada) que favoreça a homogeneidade, simetria ou aditividade.

Um dos melhores exemplos da conveniência de transformar dados é dada por McNeil (1977). Ele será aproveitado aqui na íntegra.

Uma discussão permanente quando alguém começa a estudar geografia é de se a Austrália é uma ilha ou um continente. Como, trabalhar com as áreas de todas as ilhas é impossível, estão listadas abaixo as áreas de "ilhas" com mais de 10.000 milhas quadradas. Além disto as ilhas pequenas, sejam poucas ou muitas, são irrelevantes para resolver a questão. Ásia e Europa estão listadas separadamente. Juntando as áreas de ambas em Eurásia não fará diferença.

Axelh.	16	Baffin	184	Banks	23	Devov	21
Melv.	16	P. of. W.	13	South.	16	Victória	82
Spits	15	Grabret.	84	Irlanda	33	Groenl.	840
Terra N.	43	T. d. F.	19	Cuba	43	Hisp.	30
Madagascar	227	Formosa	14	Hainan	13	Hohkaiko	30
Kyushu	14	N. Guiné	306	Nova Z(N)	44	Nova Z(S)	58
Mird	36	Sakh.	29	Tasmania	26	Vanconver	12
Celep	73	Java	49	Moluca	29	N. Br.	15
Timor	13	Elles	82	Nova Z.	32	Islandia	40
Ceilão	25	Houshu	89	Luzon	42	Borneo	280
Sumatra	183	Ásia	16988	África	11506	Am. Norte	9390
Am. Sul	6795	Europa	3745	Austrália	2968	Antártida	5500

As áreas estão em milhares de milhas quadradas. A Figura 1 apresenta um ramo e folhas em centenas de milhares de milhas quadradas.

Figura 1 - Ramo e folhas das áreas das "ilhas" com mais de 10.000 milhas quadradas, em centenas de milhares.

```
(41)   0 00000000000000000000000000000000000112238
  7    1
  7    2 9
  6    3 7
  5    4
  5    5 5
  4    6 7
  3    7
  3    8
  3    9 3
  2   10
  2   11 5
  1   12
  1   13
  1   14
  1   15
  1   16 9
```

Austrália

Este ramo e folhas é pouquíssimo revelador pois os pequenos valores estão muito "embolados". A distribuição é extremamente assimétrica. É necessário fazer uma transformação nos dados. Inicialmente será tentada a transformação raiz quadrada. O resultado está na Figura 2.

Figura 2 - Áreas das "ilhas" com mais de 10.000 milhas quadradas - raízes quadradas das áreas em milhares de milhas quadradas.

```
(35)   0 33333333444444555555556666667789999
 13    1 33567
  8    2 8
  7    3
  7    4
  7    5 4
  6    6 1
  5    7 4
  4    8 2
  3    9 6
  2   10 7
  1   11
  1   12
  1   13 0
```

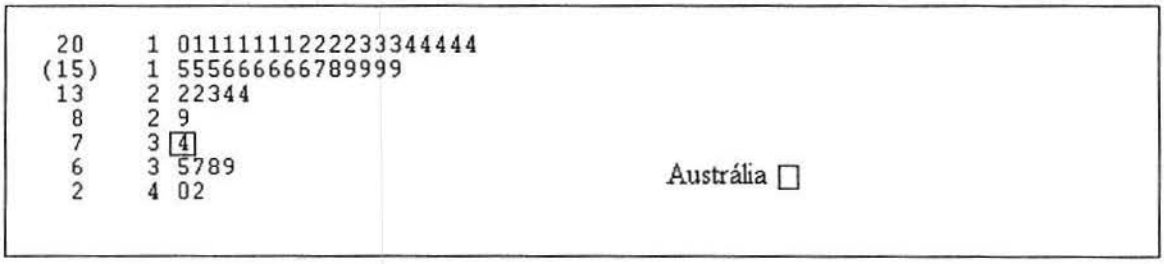
Austrália

A transformação utilizada não foi adequada para atenuar suficientemente ou eliminar a assimetria. Na Figura 3 são apresentados os logaritmos das áreas. É utilizada a base 10 e são tomados logaritmos das áreas em milhares de milhas quadradas.

Finalmente, agora, as coisas começam a ficar mais claras. Na realidade existem dois aglomerados distintos de valores. E estão a salvo os livros dos australianos.

Percebida a vantagem de fazer transformações é fundamental entender porque isto acontece.

Figura 3 - Área das "ilhas" com mais de 10.000 milhas quadradas - logaritmos em base 10 das áreas em milhares de milhas quadradas.

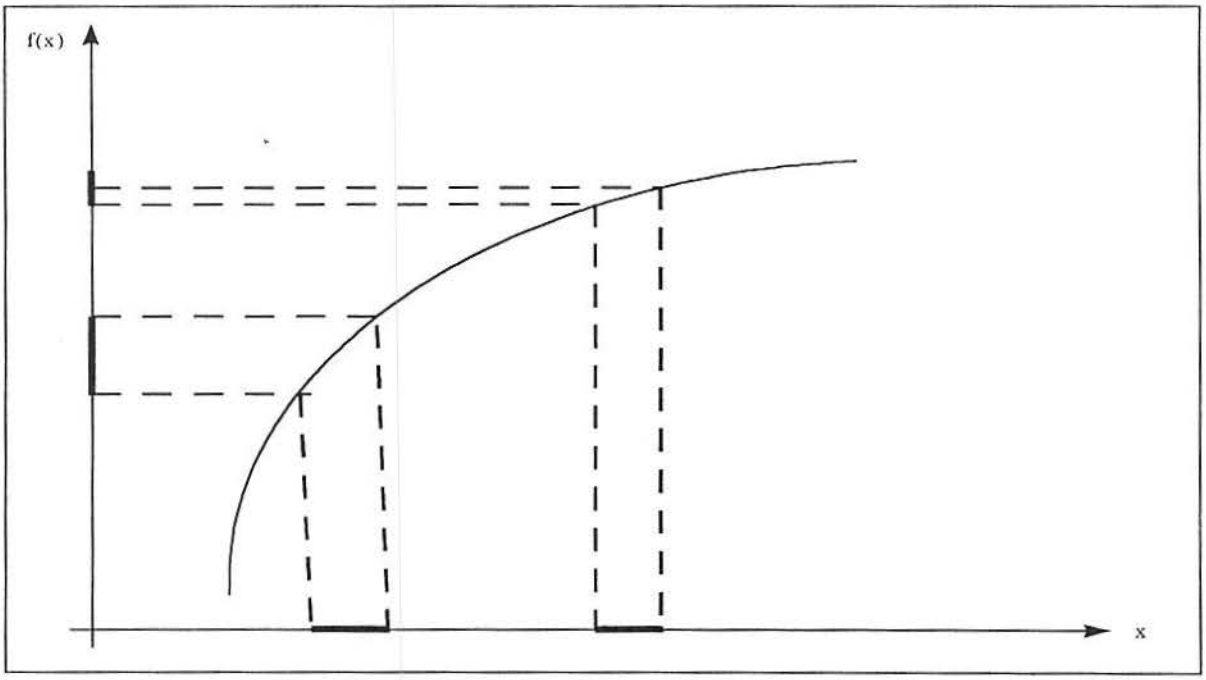


Outra razão da necessidade de fazer transformações está em que, ao compararmos dois ou mais conjunto de valores, se as dispersões são muito diferentes, a comparação é difícil , ou até impossível. Em geral, mesmo que não seja imprescindível, uma transformação melhora o entendimento que se tem de um conjunto de valores.

Vale a pena, também, chamar a atenção para a importância da simetria. Na realidade onde existem assimetrias muito acentuadas é impossível (pelo menos difícil) apresentar um valor que sirva como indicador de nível para um conjunto de valores. Senão por outras, pelo menos como forma de se atingir alguma simetria, as transformações são de suma importância.

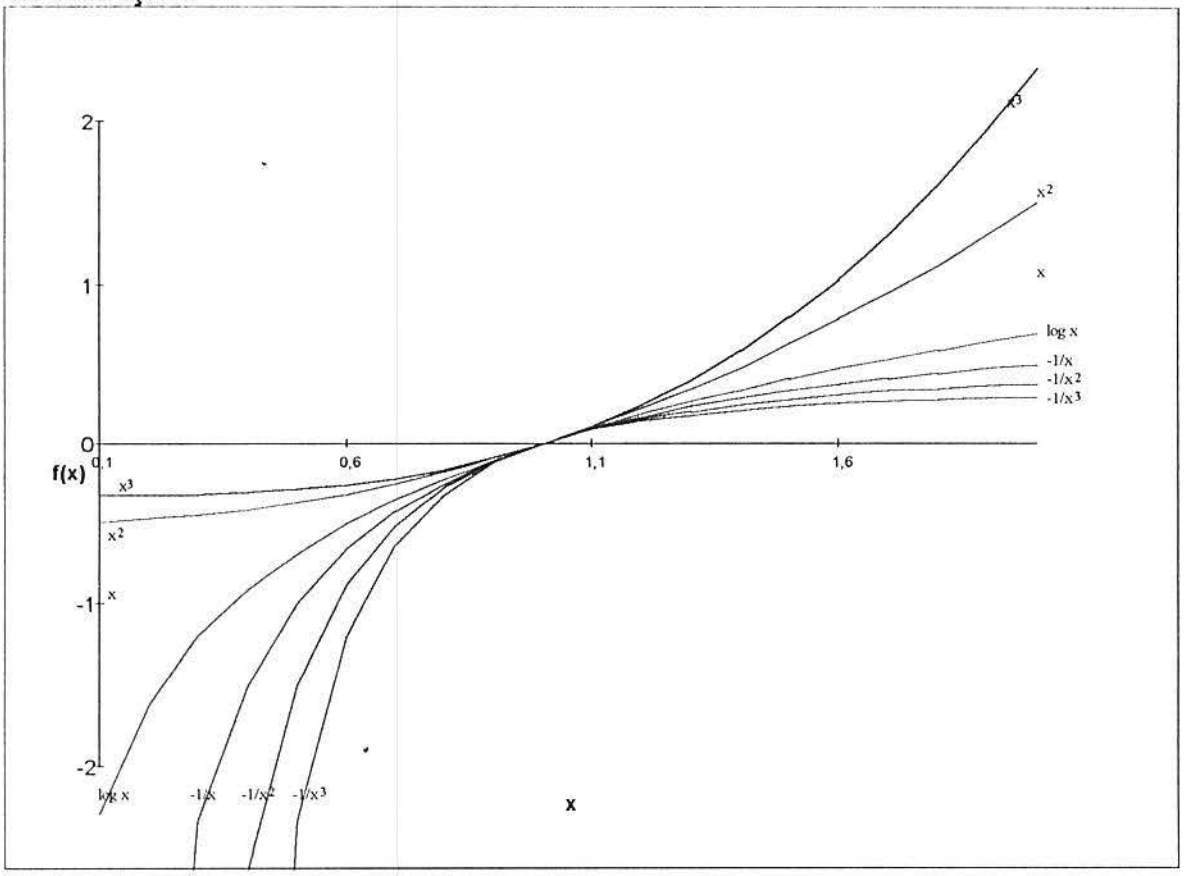
O que se busca é uma transformação do tipo representado no Gráfico 1 que pelo menos diminua as assimetrias do tipo que surge quando se tem uma ocorrência do gênero "lei anormal" dos grandes números. Como fácil observar, para valores grandes ocorre uma contração, quando comparada a valores pequenos.

Gráfico 1- Efeito de uma transformação sobre o comprimento de dois intervalos iguais - um de valores "grandes", outro de valores "pequenos".



Se a assimetria fosse em sentido contrário ter-se-ia que buscar uma transformação que tivesse o efeito oposto ao da que é apresentada no Gráfico 1. No Gráfico 2 são apresentadas diversas transformações, inclusive a não-transformação, isto é, a identidade. A seguir, no Gráfico 3, com a escala horizontal logarítmica, são mostradas novamente algumas das mesmas transformações e algumas outras. Aí pode-se ver o papel importante da transformação logarítmica como intermediária entre $\sqrt{x}$ e $-1/\sqrt{x}$. É importante ressaltar que o conjunto de transformações apresentado é em geral suficiente para resolver os problemas encontrados na prática, principalmente devido a que, com a imprecisão nos valores e com os limites possíveis e razoáveis de percepção, outras transformações intermediárias são um refinamento complicado, desnecessário e injustificado.

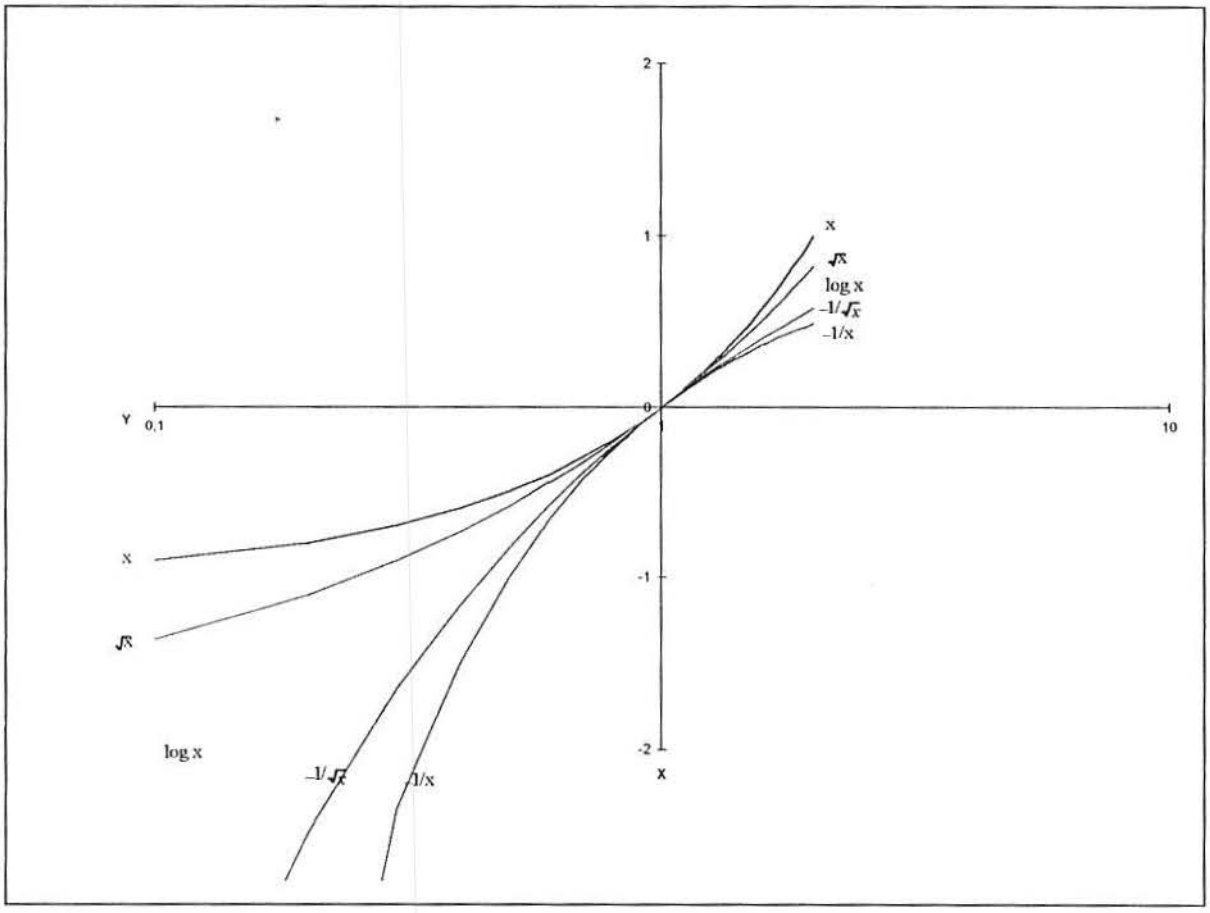
Gráfico 2- Algumas transformações - com escala e origem alteradas para melhor visualização.



Observe a correspondência entre as funções e sua representação no gráfico:

Função	No gráfico
x^3	$-1/3 + x^3/3$
x^2	$-1/2 + x^2/2$
x	$-1 + x$
$\log x$	$2{,}303 \log x$
$-1/x$	$1 + (-1/x)$
$-1/x^2$	$1/2 - 1/2(1/x^2)$
$-1/x^3$	$1/3 - 1/3(1/x^3)$

Gráfico 3 – Outras transformações: $x^1, x^{1/2}, \log x, x^{-1/2}, x^{-1}$.



Observe a correspondência entre as funções e sua representação no gráfico:

Função	No gráfico
x	$-1+x$
$\sqrt{x}$	$-2+2\sqrt{x}$
$\log x$	$2{,}303 \log x$
$-1/\sqrt{x}$	$2+2(-1/\sqrt{x})$
$-1/x$	$1+(-1/x)$

Pode-se dizer que o papel central de transformação logarítmica eqüivale a potência zero

A razão dos negativos nas transformações $-1/\sqrt{x}$, $-1/x$, $-1/x^2$ e $-1/x^3$ é de que assim se conserva a ordem dos valores com que se está trabalhando. Todas as transformações apresentadas são usadas quando os valores são todos positivos - essencialmente, pelo menos as

mais importantes ($\sqrt{x}, \log x, -1/\sqrt{x}$). O procedimento quando há valores negativos e positivos é mais complexo pode ser visto em Tukey (1977).

11 - **RE-EXPRESSÃO através do MINITAB.**

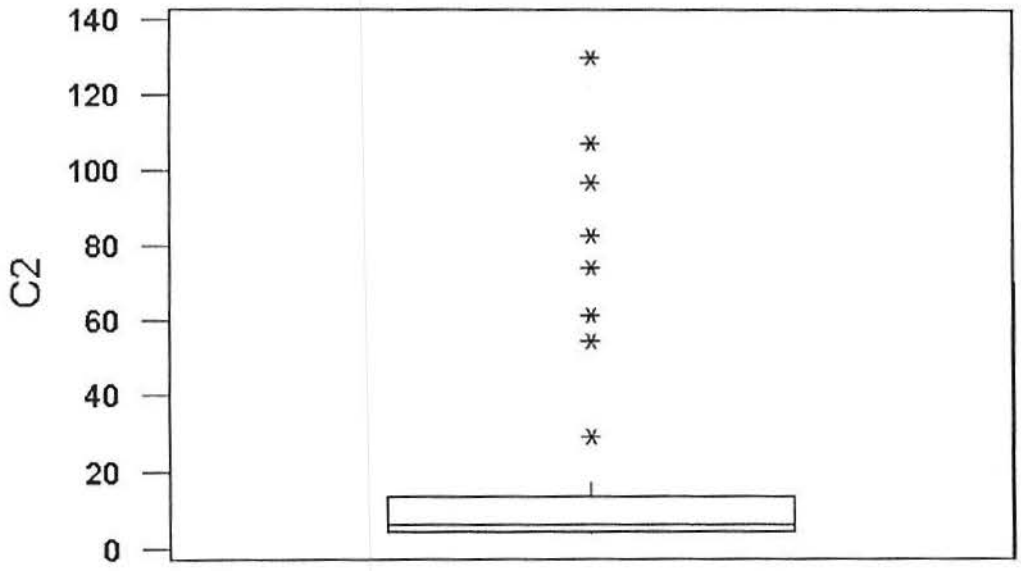
As transformações no MINITAB são feitas, diretamente, através dos operadores algébricos: SQRT (raiz quadrada), LOGTEN (logarítmo decimal), etc. Na seqüência solicita-se o diagrama desejado (Ramo-e-folhas, Boxplot ou Resumo de cinco-números). A seguir, três exemplos ilustram esta idéia.

- Transformação $\sqrt{x}$ e uso do Boxplot.

Considerando que os dados usados para o problema Austrália ilha ou continente (áreas das ilhas com mais de 10.000 milhas bem como dos continentes) estejam na coluna 1 (c1), o procedimento no MINITAB fica:

```
MTB> sqrt c1 c2          {raiz quadrada de c1 em c2}
MTB> boxplot c2          {boxplot de c2}
```

Que resulta em:



-Transformação log(x) e uso do diagrama de Ramo e folhas.
Considerando os mesmos dados na coluna 1 (c1), tem-se:

```
MTB > LET C3=LOGTEN(C1)
MTB > Stem-and-Leaf C3;
SUBC>   Increment 0.5.

 Character Stem-and-Leaf Display


Stem-and-leaf of C3        N   = 48
Leaf Unit = 0.10

  20     1 01111111222233344444
 (15)    1 555666666789999
  13     2 22344
   8     2 9
   7     3 4
   6     3 5789
   2     4 02

MTB >
```

- Transformação $1/\sqrt{x}$ e esquema de cinco-números.
Ainda estando os dados originais em c1, segue:

```
MTB > LET C4=1/(C1**(1/2))
MTB > DESCRIBE C4

 Descriptive Statistics


Variable   N     Mean   Median   TrMean   StDev   SEMean
C4        48   0.1516   0.1562   0.1521   0.0879   0.0127

Variable      Min      Max       Q1       Q3
C4         0.0077   0.2887   0.0738   0.2266

MTB >
```

A partir da identificação das estatísticas de interesse, constrói-se o Resumo de cinco-números em outro aplicativo. No caso utilizou-se o WORD.

	Nº de obs.: 48		
Md	0,1562		
Q_I	0,0738	0,2266	Q_S
L_I	0,0077	0,2887	L_S

12 – **UMA APLICAÇÃO**.

O Resumo de cinco-números de um grupo de dados e o Boxplot mostram muito da estrutura do grupo. De um Boxplot pode-se tomar as seguintes características de um grupo:

- posição;
- dispersão;
- assimetria;
- extensão da cauda (distribuição)
- "outliers".

Então, o Boxplot dá uma impressão visual de vários aspectos importantes da distribuição empírica de um grupo de dados.

É especialmente útil para comparação de vários grupos de dados. Extraindo-se um boxplot para cada grupo e arranjando-os paralelamente, pode-se comparar os grupos quanto às suas características. Nesta comparação pode-se perceber que os dados de diferentes grupos não estão todos ajustados dentro da mesma escala. Em particular, grupos localizados longe da origem podem estar mais dispersos que os grupos localizados próximos à origem.

Uma transformação apropriada pode aliviar esta dificuldade, fazendo a variação dos grupos mais homogêneos, o que os tornará mais comparáveis. Uma plotagem de dispersão versus nível (mediana) pode sugerir uma transformação potência que tenda a igualar a dispersão através de diferentes níveis.

Para ilustrar a potencialidade da Análise Exploratória serão utilizados os dados das populações dos maiores municípios dos maiores Estados do

Brasil, retirados do Anuário Estatístico do Brasil – 1985, FIBGE, Rio de Janeiro, 1986.

12.1 – "BOXPLOT" PARA UM ÚNICO GRUPO.

Inicialmente com um grupo formado pelos 15 maiores municípios do Estado de São Paulo (o maior em população), cujos dados são representados na Tabela 1.

Tabela 1 - População dos 15 maiores municípios do Estado de São Paulo, em 1985.

MUNICÍPIO	POPULAÇÃO (em 100.000)
São Paulo	100,99
Campinas	8,45
Guarulhos	7,18
Santo André	6,37
Osasco	5,94
São Bernardo do Campo	5,66
Santos	4,61
Ribeirão Preto	3,85
São José dos Campos	3,75
Sorocaba	3,29
Diadema	3,22
Jundiaí	3,15
Mauá	2,71
Carapicuaba	2,68
Piracicaba	2,53

Como primeiro passo da análise, constrói-se o "resumo de cinco-números", acrescido da dispersão-F, limites críticos para "outliers".

Figura 4 - "Resumo de cinco-números" para os 15 maiores municípios do Estado de São Paulo em 1985.

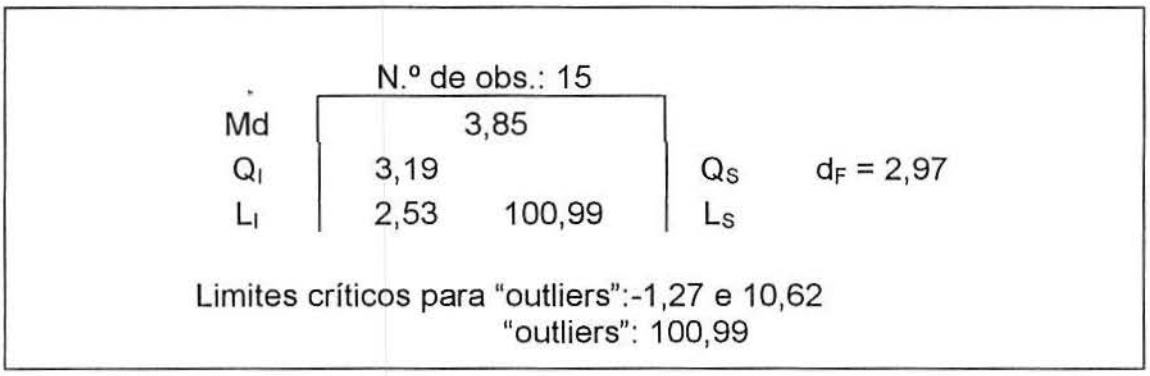
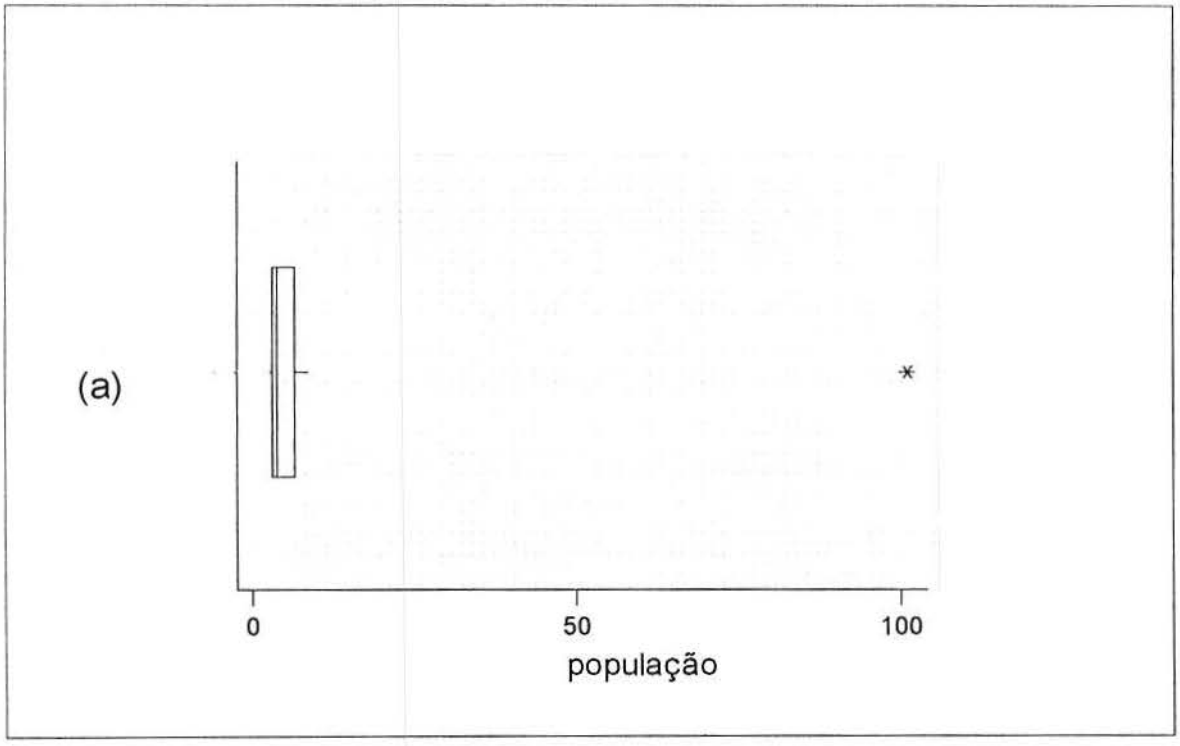
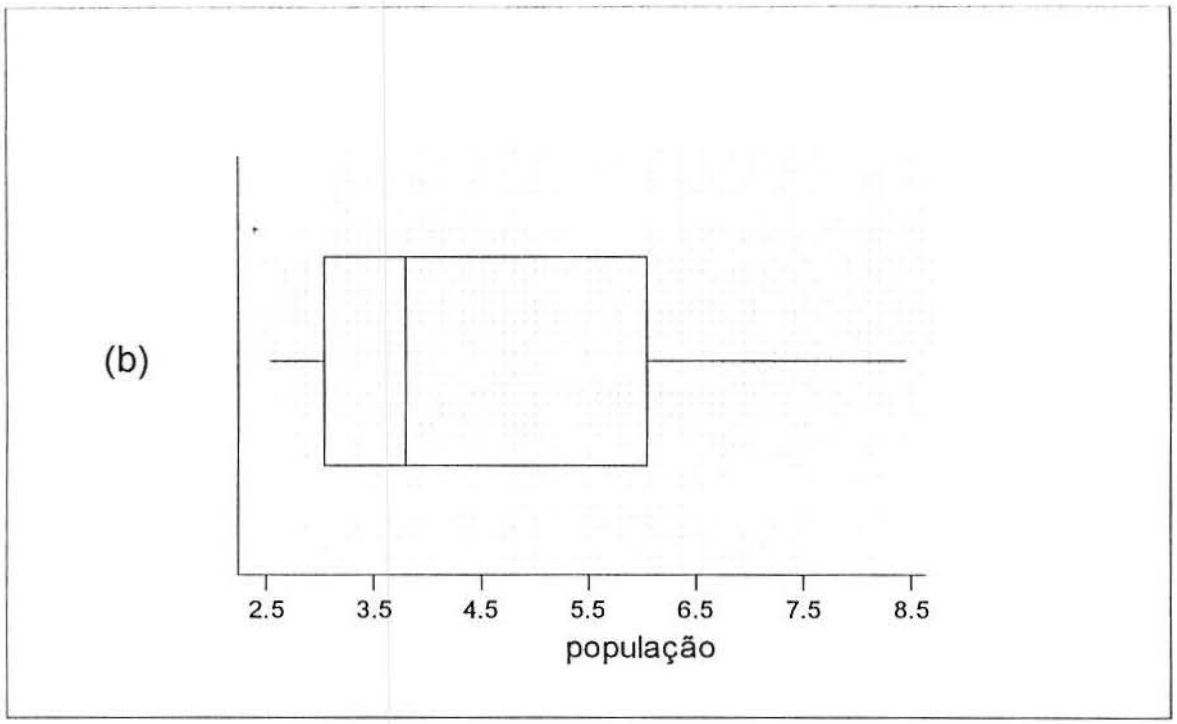


Figura 5 - "Boxplot" para os 15 maiores municípios do Estado de São Paulo, em 1985: (a) com "outlier" ; (b) sem "outlier".





Para construir o "Boxplot", primeiro toma-se uma caixa com extremidades nos quartos inferior e superior e uma faixa dividindo a caixa, que corresponde à mediana. A seguir, toma-se uma linha de cada extremidade da caixa para o mais remoto ponto que não é "outlier". A figura resultante representa esquematicamente o corpo dos dados, sem os "outliers". Estes são representados individualmente por pontos situados além dos limites críticos para 'outliers".

O "Boxplot" mostra, numa olhadela, a posição, dispersão, assimetria, comprimento da cauda e "outliers". A posição do grupo é resumida pela mediana, a faixa no interior da caixa. O comprimento da caixa mostra a dispersão dos dados, dada pela dispersão-F. Das posições da mediana e dos quartis pode-se ver um pouco de assimetria: a mediana está muito mais próxima do quartil inferior que do quartil superior, indicando que o grupo é positivamente assimétrico (uma situação comum em dados positivos ilimitados). A plotagem indica o comprimento da cauda pela linha estendida de Piracicaba a Campinas e pelo "outlier" (São Paulo).

As propriedades de resistência do "Boxplot" o tornam atrativo para uso em Análise Exploratória. Uma plotagem análoga poderia ser baseada na média

e desvio padrão amostrais, no entanto, estas medidas falhariam na resistência a um único valor estranho.

12.2 - BOXPLOT PARA COMPARAÇÃO DE GRUPOS

Continuando com os dados do exemplo, toma-se agora os 10 maiores municípios dos 10 maiores Estados do Brasil em 1985.

Tabela 2 - População dos 10 maiores municípios dos 10 maiores Estados do Brasil em 1985. (população em 100.000)

(1) GOIÁS		(2) CEARÁ	
Goiânia	9,28	Fortaleza	15,89
Anápolis	2,27	Juazeiro do Norte	1,60
Luziânia	1,01	Sobral	1,28
Araguaína	0,91	Itapipoca	1,09
Rio Verde	0,84	Caucaia	1,09
Itumbiara	0,80	Quixadá	1,00
Jatai	0,60	Crato	0,87
Formosa	0,55	Iguatu	0,86
Aparecida de Goiânia	0,54	Acaraú	0,77
Catalão	0,50	Morada Nova	0,75

(3) MARANHÃO		(4) PERNAMBUCO	
São luís	5,64	Recife	12,90
Imperatriz	2,37	Jabotão	4,11
Caxias	1,49	Olinda	3,36
Codó	1,19	Caruaru	1,91
Santa Luzia	1,17	Paulista	1,61
Barra do Corda	0,95	Petrolina	1,31
Timon	0,92	Cabo	1,22
Bacabal	0,88	Camaragibe	1,13
Monção	0,83	Vitória de Santo Antão	1,01
Pinheiro	0,77	Garanhuns	0,98

(5) BAHIA		(6) RIO GRANDE DO SUL	
Salvador	18,11	Porto Alegre	12,75
Feira de Santana	3,57	Pelotas	2,78
Vitória da Conquista	1,99	Caxias do Sul	2,68
Itabuna	1,79	Canoas	2,62
Juazeiro	1,54	Santa Maria	1,97
Ilhéus	1,46	Novo Hamburgo	1,68
Jequié	1,27	Rio Grande	1,65
Jacobina	1,21	Viamão	1,49
Alagoinhas	1,17	Gravataí	1,42
Camaçari	1,08	Passo fundo	1,38

(7) PARANÁ		(8) MINAS GERAIS	
Curitiba	12,85	Belo Horizonte	21,22
Londrina	3,48	Contagem	3,86
Ponta Grossa	2,24	Juiz de Fora	3,51
Cascavel	2,01	Uberlândia	3,14
Maringá	1,98	Uberaba	2,46
Foz do Iguaçu	1,83	Governador Valadares	2,17
Guarapuava	1,49	Montes Claros	2,15
Pitanga	1,00	Ipatinga	2,14
Paranaguá	0,97	Divinópolis	1,40
Toledo	0,96	Teófilo Otoni	1,26

(9) RIO DE JANEIRO		(1) SÃO PAULO	
Rio de janeiro	56,15	São Paulo	100,99
Nova iguaçu	13,25	Campinas	8,45
São Gonçalo	7,31	Guarulhos	7,18
Duque de Caxias	6,66	Santo André	6,37
São João do Meriti	4,59	Osasco	5,94
Niterói	4,43	São Bernardo do Campo	5,66
Campos	3,67	Santos	4,61
Petrópolis	2,75	Ribeirão Preto	3,85
Volta Redonda	2,20	São José dos Campos	3,75
Magé	2,00	Sorocaba	3,29

FONTE: Anuário Estatístico do Brasil - 1985. FIBGE

Os Estados estão ordenados pela população mediana dos municípios.

Figura 6 - "Resumo de 5-números", limites crítico para "outliers" e "outliers" para a população dos 10 maiores municípios dos 10 maiores Estados do Brasil, em 1985. População em unidade de 100.000.

# 10		GOIÁS		# 10		CEARÁ	
M	5,5	0,82		M	5,5	1,05	
F	3	0,55	1,01	F	3	0,86	1,28
	1	0,50	9,28		1	0,75	15,89
		d_F =0,46; 1,5 d_F =0,69				d_F =0,42; 1,5 d_F =0,63	
		Lim. Crít. p/ "outliers": -0,14 e 1,70				Lim. Crít. p/ "outliers": 0,23 e 1,91	
		"outliers": 2,27 e 9,28				"outliers": 15,89	

# 10		MARANHÃO		# 10		PERNAMBUCO	
M	5,5	1,06		M	5,5	1,46	
F	3	0,88	1,49	F	3	1,13	3,36
	1	0,77	5,64		1	0,98	12,90
		d_F =0,61; 1,5 d_F =0,92				d_F =2,23; 1,5 d_F =3,35	
		Lim. Crít. p/ "outliers": -0,04 e 2,41				Lim. Crít. p/ "outliers": -2,22 e 6,71	
		"outliers": 5,64				"outliers": 12,90	

# 10		BAHIA		# 10		R.G. SUL	
M	5,5	1,5		M	5,5	1,83	
F	3	1,21	1,99	F	3	1,49	2,68
	1	1,08	18,11		1	1,38	12,75
		d_F =0,78; 1,5 d_F =1,17				d_F =1,19; 1,5 d_F =1,79	
		Lim. Crít. p/ "outliers": 0,04 e 3,16				Lim. Crít. p/ "outliers": -0,3 e 4,47	
		"outliers": 3,57 e 18,11				"outliers": 12,75	

# 10		PARANÀ		# 10		M. GERAIS	
M	5,5	1,91		M	5,5	2,32	
F	3	1,00	2,24	F	3	2,14	3,51
	1	0,96	12,85		1	1,26	21,22
		d_F =1,24; 1,5 d_F =1,86				d_F =1,37; 1,5 d_F =2,06	
		Lim. Crít. p/ "outliers": -0,86 e 4,1				Lim. Crít. p/ "outliers": 0,08 e 5,57	
		"outliers": 12,85				"outliers": 21,22	

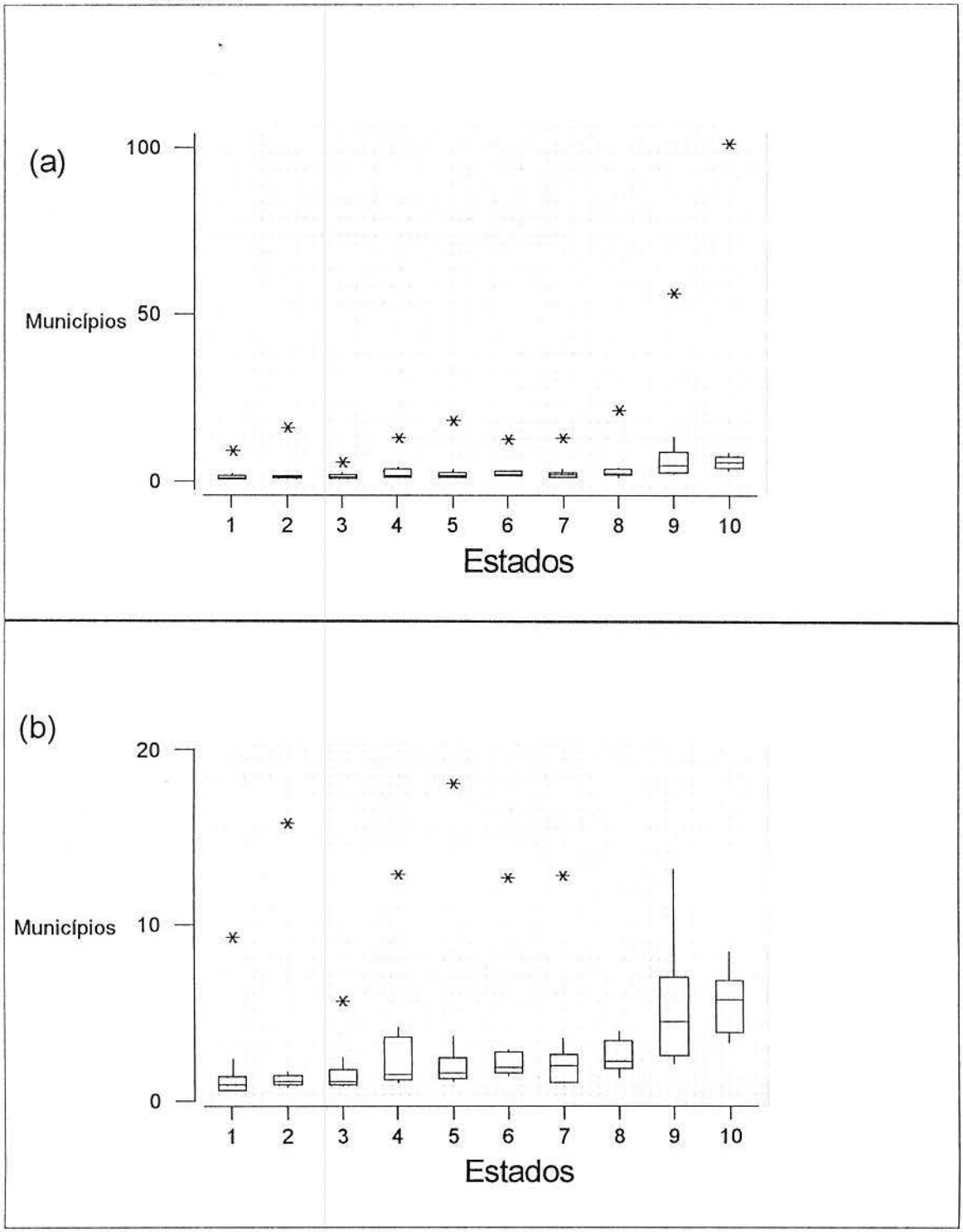
# 10		R. JANEIRO		# 10		S. PAULO	
M	5,5	4,51		M	5,5	5,8	
F	3	2,75	7,31	F	3	3,85	7,18
	1	2,00	56,15		1	3,29	100,99
		d_F =4,56; 1,5 d_F =6,84				d_F =3,33; 1,5 d_F =4,99	
		Lim. Crít. p/ "outliers": -4,09 e 14,15				Lim. Crít. p/ "outliers": -1,14 e 12,17	
		"outliers": 56,15				"outliers": 100,99	

As séries de dados dão origem a muitas questões sobre os municípios nesses 10 Estados. Como comparar a mediana populacional dos Estados? São os menores maiores municípios de São Paulo maiores que os maiores municípios dos outros Estados? Tendem os Estados com maiores municípios a mostrar maior variação nas populações dos municípios? Quais municípios são "outliers" em relação a outros em seu próprio Estado? Quanto de assimetria está presente nos vários grupos? Estas questões podem, em princípio ser respondidas usando o "resumo de cinco-números" da Figura 6 , mas na prática, um dispositivo de "Boxplots" paralelos para os 10 grupos dá a resposta a essas e a questões similares mais prontamente. Na Figura 7 apresenta-se este dispositivo.

Ordenando os Estados no dispositivo de acordo com a mediana populacional dos 10 maiores municípios, o dispositivo facilita a comparação de posição ou nível através dos Estados. Por exemplo, pode-se ver que os maiores municípios de São Paulo são maiores que aqueles dos outros Estados, excetuando alguns do Rio de Janeiro, que tem alguns municípios maiores que os de São Paulo, Nova Iguaçu, São Gonçalo e Duque de Caxias. Os maiores municípios de São Paulo e Rio de Janeiro são maiores que todos os municípios do Paraná, da Bahia, do Maranhão, do Ceará e de Goiás. Também pode-se comparar a dispersão dos 10 grupos usando o comprimento das caixas. As populações dos municípios do Ceará apresentam a menor dispersão, enquanto as do Rio de Janeiro apresentam a maior.

Usa-se este dispositivo paralelo também para comparar simetria, comprimento da cauda (distribuição) e "outliers" dos grupos. Mais da metade (7 Estados) indicam assimetria na direção dos maiores dos municípios, enquanto que 2 deles são assimétricos a esquerda, mas estes, São Paulo e Paraná, têm alguns municípios que são substancialmente maiores que os municípios representados pelas caixas. O Estado de Goiás demonstra simetria na caixa. O maior município para todos os Estados é o município de São Paulo, designado um "outlier". Todos os Estados têm pelo menos um "outlier", E os Estados de Goiás e Bahia têm dois. Todas as capitais são "outliers", o que era esperado, devido à crescente migração da população do interior para a capital nos últimos anos. Assim, nota-se duas características irregulares nesses dados: assimetria e muitos "outliers".

Figura 7 - "Boxplots" para os 10 maiores municípios dos 10 maiores Estados do Brasil, em 1985: (a) com todos os "outliers"; (b) com redução de três "outliers".



Por ter-se ordenado os Estados pela mediana, pode-se escolher outra característica comparativa: uma forte tendência da dispersão aumentar com os níveis (mediana). Uma análise complementar seria mais fácil se houvesse menos variação entre os grupos. Para promover igualdade de dispersão e reduzir a dependência de dispersão sobre o nível, pode-se tentar uma re-expressão ou transformação dos dados.

12.3 - DISPERSÃO X NÍVEL

Quando uma comparação de grupos mostra uma relação sistemática entre dispersão e nível (mediana), busca-se uma re-expressão ou transformação dos dados que reduza ou elimine essa dependência. Se tal transformação for encontrada, os dados re-expressos serão mais convenientes para exploração visual e para técnicas analíticas comuns de comparação de grupos. Por exemplo, a análise de variância usual com um fator (inteiramente casualizado) é mais simples e mais eficaz quando existe, pelo menos aproximadamente, igual variância entre os grupos.

Já que se deseja remover a relação entre dispersão e nível, procura-se entender essa relação através de uma plotagem da dispersão contra o nível. Pode-se supor que a dispersão dos quartis é proporcional a uma potência da mediana:

$$d_F = c\, M^b$$

$$\log d_F = \log c + b \log M\,, \qquad (1)$$

ou, tomando $\log c = k$,

$$\log d_F = k + b \log M \qquad (2)$$

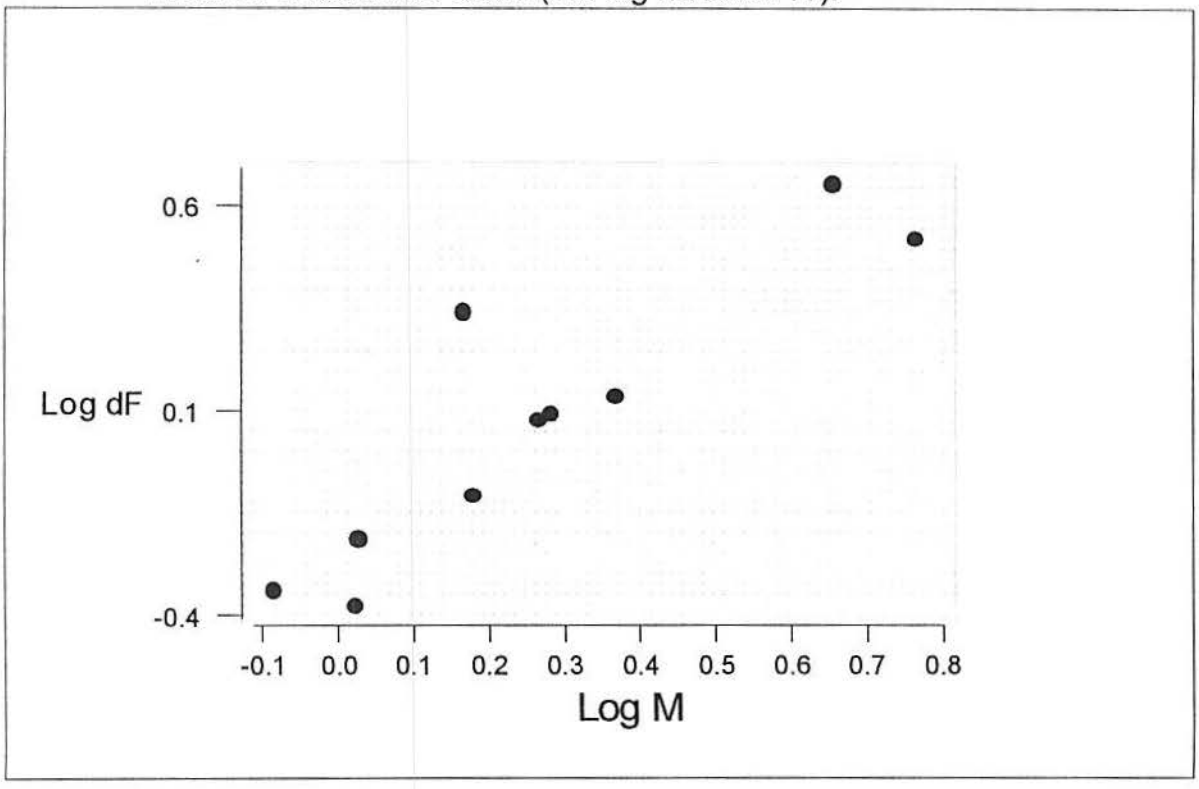
Assim, o logarítmo da dispersão-F e o logarítmo da mediana estão linearmente relacionados.

Pode-se tomar esta medida para os dados do exemplo em questão. Na Tabela 3 aparecem o logarítmo da mediana e o logarítmo da dispersão-F, retirados do "Resumo de cinco-números" dos 10 maiores Estados do Brasil, e na Figura 8 é mostrada a plotagem correspondente.

Tabela 3 - logaritmo da mediana e da dispersão-F para os 10 maiores municípios dos 10 maiores Estados do Brasil, em 1985.

Estado	Log M	Log dispersão-F
1	-0,09	-0,33
2	0,02	-0,37
3	0,03	-0,21
4	0,16	0,34
5	0,18	-0,10
6	0,26	0,07
7	0,28	0,09
8	0,37	0,13
9	0,65	0,65
10	0,76	0,52

Figura 8 - Plotagem da dispersão vs nível para os 10 maiores municípios nos 10 maiores Estados do Brasil (em log na base 10).



Realmente, a plotagem da Figura 8 indica uma tendência de crescimento do log de d_F à medida que cresce o log de M; além disso, numa primeira aproximação, a relação parece quase linear. O objetivo da plotagem é determinar o valor de b na equação (2). A transformação $z = x^{1-b}$ dos dados de x dá os valores re-expressos de z cuja dispersão-F não depende, ao menos aproximadamente, do nível da mediana.

Sob esta suposição, pode-se adotar o seguinte procedimento para diagnosticar a potência que irá estabilizar a dispersão.

Se b é a inclinação da plotagem, então p=1-b é o valor aproximado do expoente para a transformação potência de x para estabilizar a dispersão.

Neste sentido, são apresentadas na Tabela 4 as transformações potências mais freqüentemente usadas, as quais são os membros principais da "escada de potências" de Tukey.

Tabela 4 - Trasformações potências mais freqüentemente usadas.

Inclinação b	Potência p	TRANSFORMAÇÃO
-2	3	Cúbica
-1	2	Quadrada
0	1	Sem mudança
1/2	1/2	Raiz quadrada
1	0	Logaritmo
3/2	-1/2	Raiz quadrada recíproca
2	-1	Reciproca

Se for ajustado a olho uma linha para os pontos da Figura 8 poderá extrair-se uma linha cuja inclinação está próxima de 1.

Para comprovação, ajusta-se essa linha através do modelo de regressão linear e determina-se um b=1,15; mais próximo de 1 de que qualquer outra inclinação mostrada na Tabela 4. Assim, tomando p=1-b, implicará em p=1-1, portanto,

$$p \cong 0,$$

que conduz à transformação logarítmica para estabilizar a dispersão.

12.4 – RE-EXPRESSÃO EM ESCALA LOGARÍTMICA

Agora aplica-se a transformação em escala logarítmica aos dados e avalia-se seu efeito sobre os 10 grupos . Por serem as transformações potências monótonas para valores positivos, a estatística de ordem dos dados transformados igualarão a estatística de ordem original trasformada (exceto para os efeitos de arredondamento ou interpolação).

Então, para obter novos Boxplots, é preciso apenas aplicar a transformação no Resumo de 5-números. Daí recalcula-se a dispersão-F e os limites críticos para "outliers".

Por exemplo, para o Estado do Rio de Janeiro, tomando o "resumo de 5-números", obtem-se:

$$\log M = 0,65$$
$$\log F_S = 0,86$$
$$\log F_I = 0,44$$
$$d_F = 0,86-0,44 = 0,42$$
$$1,5\ d_F = 0,63$$

Limites críticos para "outliers": -0,19 e 1,49

A Figura 9 dá o log dos Resumos de 5-números e novos cálculos de "outliers" para os 10 Estados.

A Figura 10 mostra a nova plotagem dispersão-F vs nível, com os dados re-expressos em escala logarítmica, mostrando agora quase independência da dispersão quanto aos níveis das medianas dos Estados.

Note que se poderia tentar diversas transformações para esta plotagem, transformando apenas três números de cada conjunto: a mediana, o quartil superior e o quartil inferior, eliminando o trabalho de transformar os 100 dados em estudo. Na Figura 12 é apresentado, a título de ilustração, a mesma plotagem com os dados re-expressos em raiz quadrada e raiz cúbica, ambas com dados retirados das Tabelas 5 e 6 respectivamente.

Pode ser visto que nas duas plotagens que as potências 1/2 e 1/3 não eliminam a dependência da dispersão-F com o nível (mediana), confirmando, pois que a transformação adequada é a logarítmica.

Na Figura 11 são apresentadas os novos Boxplots paralelos, re-expressos em escala logarítmica.

Figura 9 - "Resumo de 5-números" dos dados dos 10 maiores municípios dos 10 maiores Estados do Brasil, re-expressos em log na base 10, com novos limites críticos para "outliers" e os "outliers" correspondentes.

# 10		GOIÁS	
M	5,5	-0,09	
F	3	0,26	0,00
	1	0,30	0,97

d_F =0,26; 1,5 d_F =0,39

Lim. Crít. p/ "outliers": -0,65 e 0,39

"outliers": 0,97

# 10		CEARÁ	
M	5,5	0,02	
F	3	-0,07	0,11
	1	-0,12	1,20

d_F =0,18; 1,5 d_F =0,27

Lim. Crít. p/ "outliers": -0,34 e 0,38

"outliers": 0,97

# 10		MARANHÃO	
M	5,5	0,03	
F	3	-0,06	0,17
	1	-0,11	0,75

d_F =0,23; 1,5 d_F =0,35

Lim. Crít. p/ "outliers": -0,41 e 0,52

"outliers": 0,75

# 10		PERNAMBUCO	
M	5,5	0,16	
F	3	0,05	0,53
	1	0,00	1,11

d_F =0,48; 1,5 d_F =0,72

Lim. Crít. p/ "outliers": -0,67 e 1,25

"outliers": nenhum

# 10		BAHIA	
M	5,5	0,18	
F	3	0,08	0,30
	1	0,03	1,26

d_F =0,22; 1,5 d_F =0,33

Lim. Crít. p/ "outliers": -0,25 e 0,63

"outliers": 1,26

# 10		R.G. SUL	
M	5,5	0,26	
F	3	0,17	0,43
	1	0,14	1,11

d_F =0,26; 1,5 d_F =0,39

Lim. Crít. p/ "outliers": -0,22 e 0,82

"outliers": 1,11

# 10		PARANÀ	
M	5,5	0,28	
F	3	0,00	0,35
	1	-0,02	1,11

d_F =0,22; 1,5 d_F =0,33

Lim. Crít. p/ "outliers": -0,53 e 0,88

"outliers": 1,11

# 10		M. GERAIS	
M	5,5	0,37	
F	3	0,33	0,55
	1	0,10	1,33

d_F =0,22; 1,5 d_F =0,33

Lim. Crít. p/ "outliers": 0,00 e 0,88

"outliers": 1,33

# 10		R.JANEIRO	
M	5,5	0,65	
F	3	0,44	0,86
	1	0,30	1,75

d_F =0,42; 1,5 d_F =0,63

Lim. Crít. p/ "outliers": -0,19 e 1,49

"outliers": 1,75

# 10		S. PAULO	
M	5,5	0,76	
F	3	0,59	0,86
	1	0,52	2,00

d_F =0,27; 1,5 d_F =0,41

Lim. Crít. p/ "outliers": 0,18 e 1,27

"outliers": 2,00

Figura 10 - Plotagem da dispersão vs nível (mediana) para o log da população dos 10 maiores municípios dos 10 maiores Estados do Brasil.

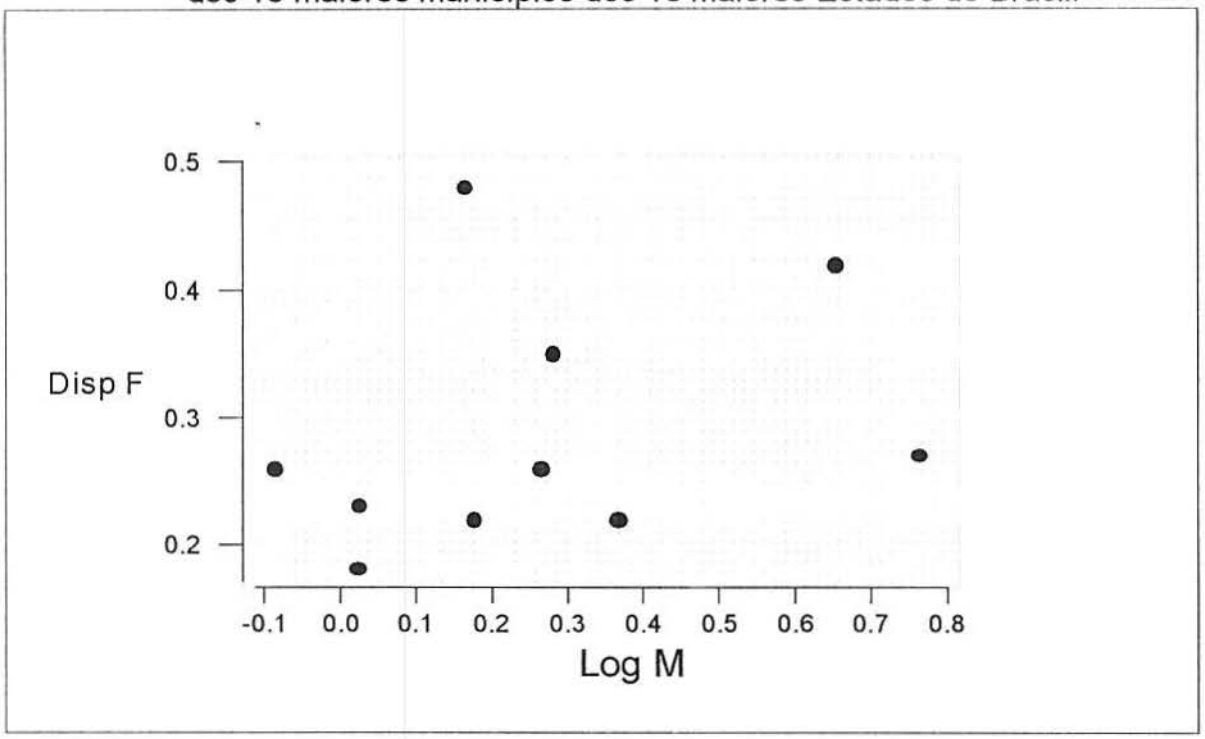


Tabela 5 - Raiz quadrada da mediana e da dispersão-F para os 10 maiores municípios dos 10 maiores Estados do Brasil, em 1985.

Estado	$(M)^{1/2}$	$(disp\text{-}F)^{1/2}$
1	0.90554	0.67823
2	1.02470	0.64807
3	1.02956	0.78102
4	1.20830	1.49332
5	1.22474	0.88318
6	1.35277	1.09087
7	1.38203	1.11355
8	1.52315	1.17047
9	2.12368	2.13542
10	2.40832	1.82483

Tabela 6 - Raiz cúbica da mediana e da dispersão-F para os 10 maiores municípios dos 10 maiores Estados do Brasil, em 1985.

Estado	$(M)^{1/3}$	$(disp\text{-}F)^{1/3}$
1	0.93599	0.77194
2	1.01640	0.74889
3	1.01961	0.84809
4	1.13445	1.30648
5	1.14471	0.92052
6	1.22316	1.05970
7	1.24073	1.07434
8	1.32382	1.11064
9	1.65219	1.65827
10	1.79670	1.49330

Figura 11 - Boxplots para os dados da população dos 10 maiores municípios dos 10 maiores Estados do Brasil, re-expressos em log na base 10.

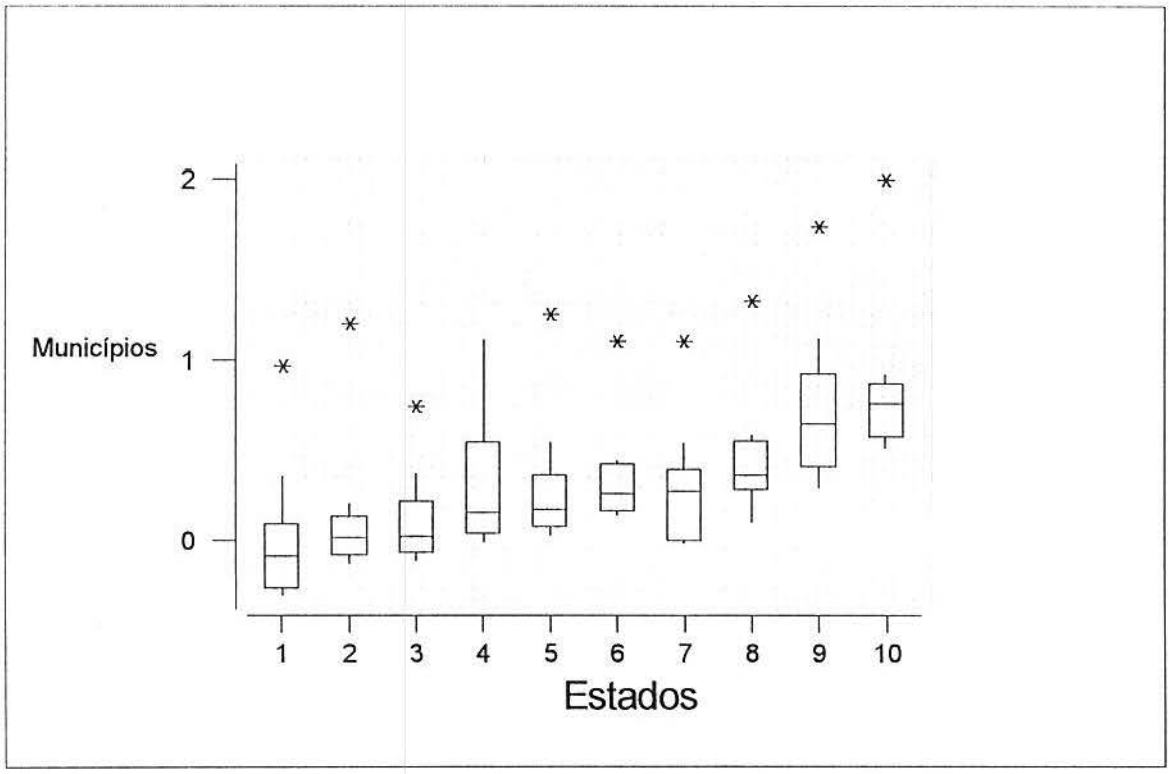
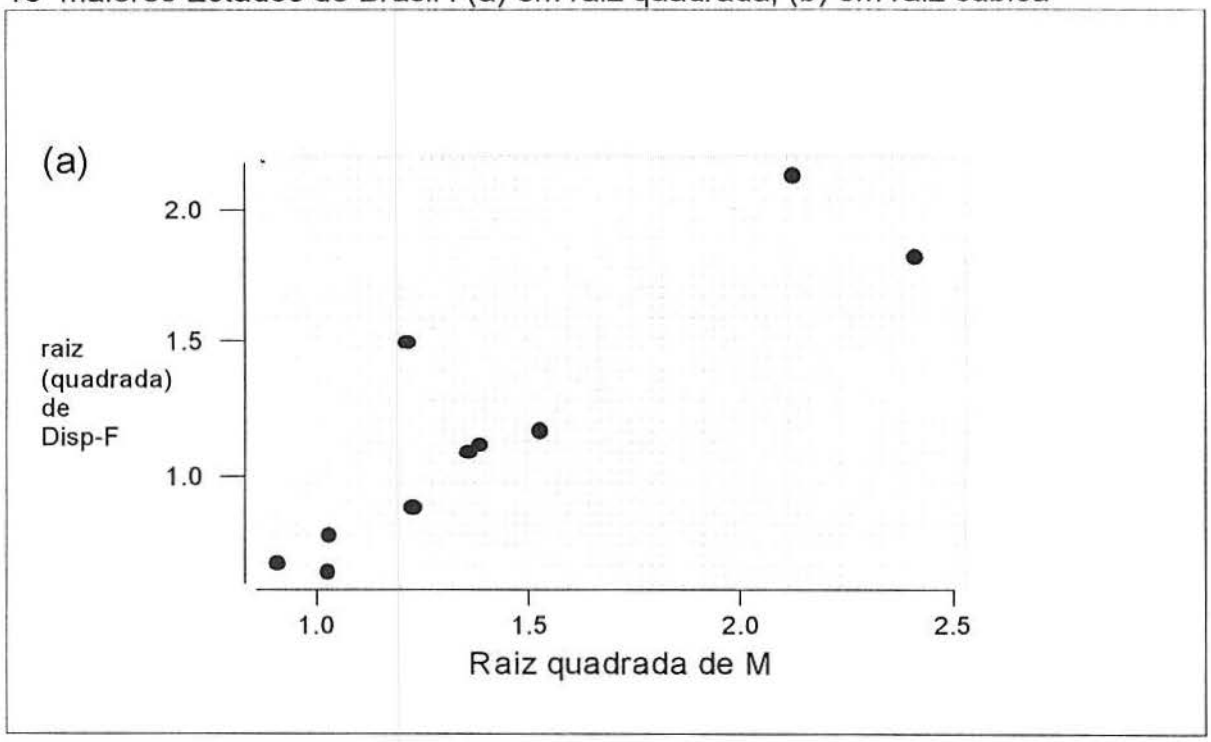
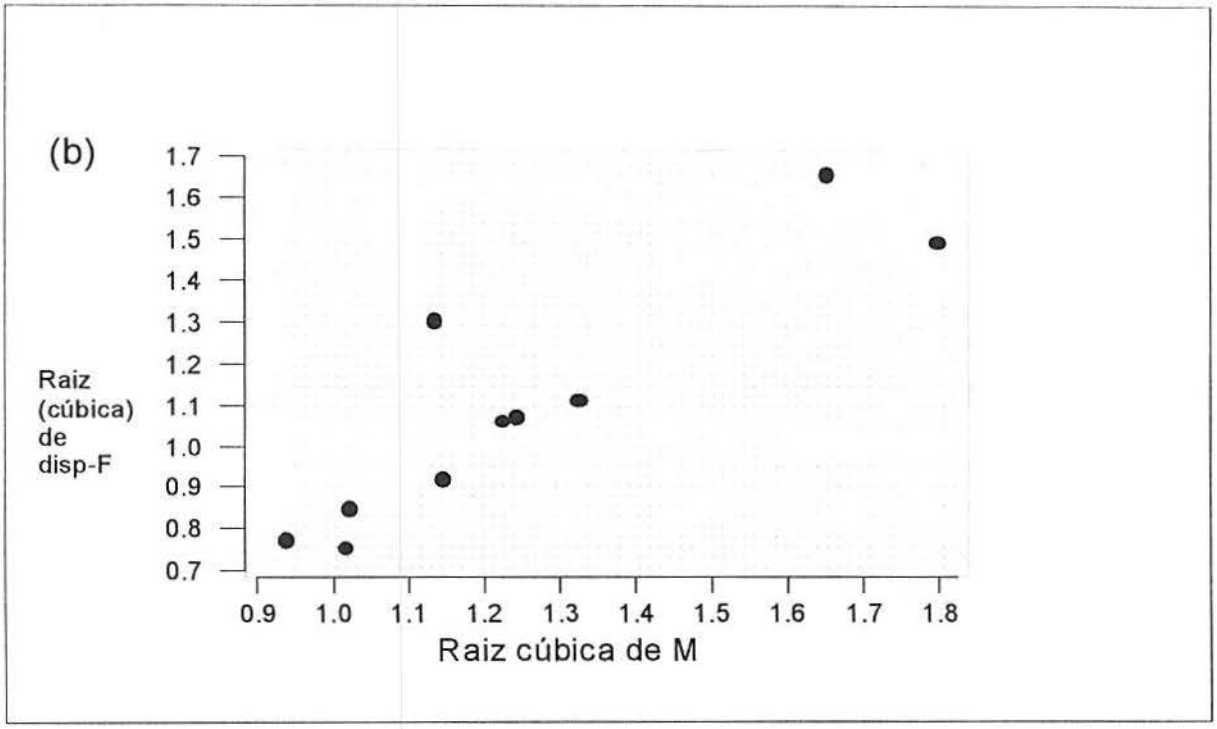


Figura 12 - Plotagem da dispersão vs nível para os 10 maiores municípios nos 10 maiores Estados do Brasil : (a) em raiz quadrada; (b) em raiz cúbica





Para ver o quanto a re-expressão em log melhorou a dispersão dos grupos, compare a Figura 7 com a Figura 11. As caixas agora estão mais similares em comprimento, e as desigualdades remanescentes parecem não estar relacionadas com nível.

A nova escala eliminou três "outliers". Dos 12 "outliers" na escala original, os 9 restantes não estão mais distantes dos limites superiores para "outliers". Note-se que a transformação permitiu incluir os "outliers" no gráfico sem produzir distorções.

Outra melhoria é que os novos Boxplots são muito mais fáceis de visualizar, e os Estados são vistos com os mesmos níveis de detalhes. Na escala original da Figura 7, Rio de Janeiro e São Paulo são mais fáceis de descrever que os outros Estados, enquanto Goiás, Ceará e Maranhão estão com visual confuso devido ao tamanho das caixas. Na Figura 11, os detalhes são vistos em igualdade para todos os Estados.

Este exame convence que a transformação melhorou os dados em muitos aspectos importantes, confirmando a suposição de vários autores de que a escala log é a mais apropriada para dados populacionais (visto que as populações têm crescimento exponencial).

O fato de ter-se tomado um processo aproximado para recomendar a transformação, conduz a sugerir uma replotagem dos pares [log mediana, log(dispersão-F)] para os dados re-expressos, apenas para checagem.

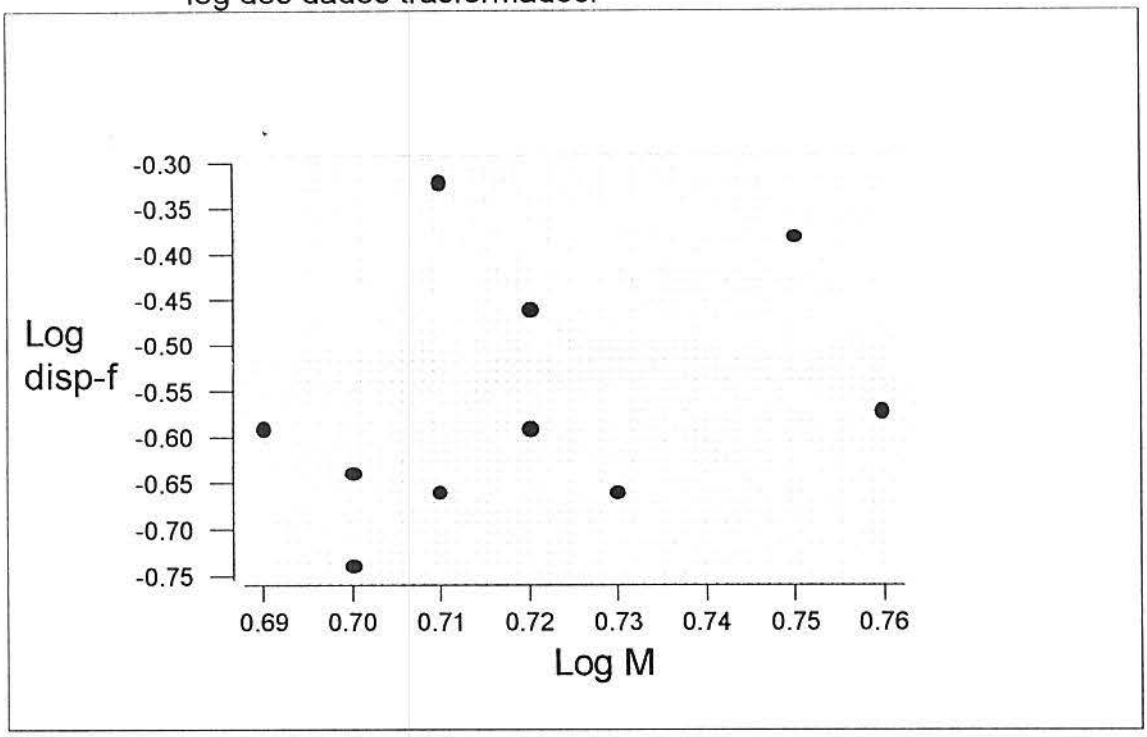
Toma-se então o log da mediana e acrescenta-se 5 (pois os dados originais foram divididos por 10^5) e a dispersão-F resultante dos dados da nova escala, e aplicamos o logarítmo nos novos pares. Os cálculos estão na Tabela 7.

Tabela 7 – Cálculos para plotagem de dispersão-F vs nível para log (população) dos 10 maiores municípios dos 10 maiores Estados do Brasil

ESTADOS	Mediana	d_F	log(M)	$\log(d_F)$
Goiás	4,91	0,26	0,69	-0,59
Ceará	5,02	0,18	0,70	-0,74
Maranhão	5,03	0,23	0,70	-0,64
Pernambuco	5,16	0,48	0,71	-0,32
Bahia	5,18	0,22	0,71	-0,66
Rio Grande do Sul	5,26	0,26	0,72	-0,59
Paraná	5,28	0,35	0,72	-0,46
Minas Gerais	5,37	0,22	0,73	-0,66
Rio de Janeiro	5,65	0,42	0,75	-0,38
São Paulo	5,76	0,27	0,76	-0,57

A Figura 13 mostra esta plotagem.

Figura 13 - Plotagem da dispersão vs nível para os dados dos maiores municípios dos maiores Estados do Brasil, re-expressos em log dos dados trasformados.



Esta checagem confirma que a transformação utilizada estabilizou a dispersão e retirou quase totalmente a dependência sobre o nível.

A plotagem dispersão-F vs nível é uma das muitas plotagens usadas em várias situações para direcionar para uma transformação apropriada para os dados. Discussão detalhada sobre a base matemática dessa plotagem pode ser vista em HOAGLIN *et alli* (1983), cap. 3, 4 e 8.

13 - **BIBLIOGRAFIA**

ALVES, M.I.F. **Introdução à análise exploratória de dados**. Seminário - CPG Estatística e Experimentação Agronômica, ESALQ / USP. Piracicaba. 1987.

ALVES, M.I.F. e SARRIES, G. A. **Utilização do SANEST (Sistema de Análise Estatística) na experimentação agronômica.** Apostila complementar. Piracicaba, ESALQ / USP, 1994, 41 p.

DACHS, J.N.W. **Análise de dados e regressão**. UNICAMP / IMECC. 1978. 125p.

FERNANDEZ, D.W.X. Estatística descritiva II. **Cadernos de Matemática e Estatística**. IM, UFRGS, Série B, Nº 24, 1994. 171p.

FIBGE. Anuário Estatístico do Brasil, 1985.

HOAGLIN, D. C.; MOSTELLER, F. e TUKEY, J. W. **Understanding Robust and Exploratory Data Analysis**. New York, John Wiley, 1983.

MACHADO, A. M. **Diagnóstico em regressão linear**. 3º SEAGRO, Lavras, ESAL / DCE, 1989. 73p.

MCNEIL, D. R. **Interactive Data Analysis** - A Pratical Primer. New York, John Wiley, 1977.

MONTGOMERY, D.C. **Introduction to statistical quality control**. 2ª edição, New York, John Wiley, 1985. 674p.

TUKEY, J.W. **Exploratory data analysis**. Reading, Addison-Wesley, 1977.