DETECÇÃO DE VALORES EXTREMOS POR PROCESSO GRÁFICO EM DADOS DE PRECIPITAÇÃO PLUVIOMÉTRICA

Nilva Maria Prestes de Toledo [1]

INTRODUÇÃO

O estudo de valores atípicos, aberrantes, extremos ou, ainda, outliers, tem sido alvo de muitos trabalhos e estudos, desde o começo do nosso século.

Uma observação atípica, que aparece em uma massa de dados, é uma observação anômala, que foge ao comportamento da maioria dos dados, causando alterações na análise de variância do experimento, e que podem conduzir o pesquisador a conclusões errôneas.

Quando se obtém respostas (y_1, y_2, ..., y_n) a determinados tratamentos aplicados a um experimento, muitas vezes ocorre uma observação atípica, que pode ter como causa diversos fatores:

a) erros na tomada de dados, como contagem, pesagem, notas, medições etc.;
b) falhas na execução do experimento ;
c) variabilidade intrínseca dos dados.

No caso a, a detecção de valores aberrantes torna-se fácil, como também no caso b. Para manejar esse valor, usa-se normalmente tratá-lo como parcela perdida ou eliminá-lo da amostra, para que não interfira na análise de variância.

Mas, no c, é realmente difícil para o pesquisador detectá-lo. Ocorre com uma frequência muito pequena e reflete um fenômeno inexplicável pela lógica. Nesse caso, os modelos estatísticos normalmente adotados devem

---

[1] Instituto Agronômico, Campinas, SP. Com bolsa de suplementação do CNPq.

ser abandonados, como o modelo de F, e passa-se a usar inferência robusta.

Karl Pearson (citado por BARNETT, 1978) dá um exemplo de capacidade de crânios da raça Moriori, em ml, numa amostra de 17 crânios.

1230	1318	1380	1240	1630	1378
1348	1380	1470	1445	1360	1410
1540	1260	1364	1410	1545	

No bloco de dados acima, o valor 1630 discorda dos demais, mas não foi cometido nenhum erro grosseiro de medida.

Dentre as causas possíveis de sua ocorrência, existe a alternativa de que esse valor estaria refletindo uma modificação da raça. Ele poderia ser o que um arqueologista realmente esperava encontrar, indicando uma mistura de raças, provocando um desenvolvimento da capacidade craniana. Nesse caso, a ocorrência do outlier pode conduzir a descobertas e a conclusões não esperadas pelo pesquisador.

Então, a simples rejeição ou acomodação dos dados torna-se irrelevante.

DIXON (1950, 1951, 1953) discutiu longamente a ocorrência de valores atípicos, observando que, em primeiro lugar, deve-se selecionar esses valores, e tentar detectar as causas de sua ocorrência. Procurar descobrir se houve uma má condução do experimento, ou se está acontecendo uma ocorrência incomum, que o pesquisador poderá futuramente explorar.

Uma tal observação causa distorções na média da amostra, na variância e no coeficiente de variação, conduzindo a erros de interpretação.

DACHS (1980) recomenda utilizar-se, primeiramente de uma análise exploratória de dados, a fim de se ter uma idéia do comportamento dos dados, antes de se proceder à análise do experimento.

MOSTELER (s/data) recomenda, para um diagnóstico preliminar, a utilização de técnicas gráficas quando se tem uma grande quantidade de dados, e recomenda o uso de estimadores robustos, que são mais consistentes, como a mediana, em substituição à média aritimética de uso

BARNETT (1978) analisa longamente inúmeras técnicas de análise exploratória de dados e modelos alternativos para a análise estatística na presença de valores atípicos e, ainda, relata diversas técnicas para manejá-los, como rejeição, acomodação, incorporação ou identificação.

MATERIAL E MÉTODOS

Neste trabalho apresentamos uma técnica de análise exploratória de dados para detectar valores atípicos.

É uma técnica essencialmente gráfica, que permite "ver" e localizar os valores em questão.

Constitui uma fase preliminar, que procura estudar o comportamento das variáveis, na tentativa de buscar, como por exemplo, técnicas robustas, que são mais persistentes e confiáveis quando se tem valores anômalos presentes.

As técnicas utilizadas são:

1. **Ramo e folhas**

É uma forma simples e eficiente de trabalhar com grande número de dados, que nos chegam às mãos de forma desordenada. A técnica consiste em ordená-los, em ordem crescente e agrupá-los em intervalos pré-determinados. Essa técnica mostra com clareza quais são os valores que se destacam da maioria dos dados e nos dá uma idéia da distribuição que eles têm. A partir de uma linha vertical divisória, temos à esquerda dela o ramo, em escala pré-determinada e à direta dessa mesma linha, as folhas (quadros I, II, III e IV).

2. **Esquema dos cinco números**

A partir dos dados já ordenados, determinam-se cinco valores, que são:

2.1. **Mediana** (M)

É o valor correspondente ao elemento de ordem (N+1) /2 em um conjunto de N dados, arranjados em ordem crescente.

2.2. **Junta superior** (JS)

2.3. **Junta inferior** (JI)

É o valor calculado da mesma forma que a mediana, a partir dela e do menor valor do conjunto.

2.4. **Pontos extremos superior (PES) e inferior (PEI)**

Têm o valor de uma vez e meia a diferença das juntas superior e inferior

2.5. **Pontos soltos** (x)

São os valores maiores que os pontos extremos superiores e menores que os pontos extremos inferiores.

Os dados que foram utilizados neste trabalho, referem-se à precipitação pluviométrica mensal e anual, em mm, obtidos na sede do Instituto Agronômico de Campinas, SP, desde o ano de 1890 até o final do ano de 1983, num total de 94 anos. Foram desprezados os números após as vírgulas.

Portanto, são 94 totais anuais e 1128 totais mensais.

Os meses do ano foram agrupados em 4 períodos, visando agrupar em um mesmo período os meses chuvosos, secos e intermediários:

Período I: novembro - dezembro - janeiro
Período II: fevereiro - março - abril
Período III: maio - junho - julho
Período IV: agosto - setembro - outubro.

Nos quadros I, II, III e IV estão os ramos e folhas desses quatro períodos, em escala de 10 em 10 mm, com 282 dados em cada um.

Na figura 2 estão o ramo e folhas dos totais anuais de 94 anos.

A partir desses dados, já ordenados, foram calculados os valores do esquema dos cinco números, que são:

Período I:	PES	460
	JS	265
	M	197
	JI	135
	PEI	

QUADRO I - Ramo e folhas da precipitação pluviométrica mensal de Campinas, de 1890 a 1983.

Período I - nov/dez/jan

63:64		:	9
52:53	7		
50:51			
48:49		:	5
46:47	5	:	2
44:45	3		
42:43	6 9	:	0 1 2
40:41	2	:	1 6
38:39	0 8	:	1 4 6 8
36:37	6 9	:	3 5 7
34:35	6 8 8	:	4 7
32:33	3 5 6 7 9	:	5 5 9 9
30:31	2 2 7 8 8 9	:	0 2 8 9
28:29	2 3 5 5 8 8 8 9 9	:	2 4 4 7 7 8
26:27	3 5 5 6 9 9	:	3 7 8 8
24:25	0 0 1 2 2 3 3 5 5 5 5 6 6 7 9	:	1 3 4 4 7 7 9
22:23	1 2 3 3 3 3 4 5 7 9 9 9	:	1 2 4 9
20:21	1 1 2 3 4 4 6 6 6 6 8 8 9 9	:	0 0 1 2 2 3 3 3 4 5 6
18:19	1 3 3 4 4 5 7 7 8	:	0 0 1 1 1 4 5 6 7 7 7 8 8 9
16:17	0 2 2 7 7 7 8 9 9	:	0 0 1 1 1 2 3 3 5 8 9
14:15	0 0 0 1 1 3 3 3 4 4 5 6 6 7 7 9	:	1 1 1 2 3 4 5 5 6
12:13	0 2 3 4 6 6	:	0 0 2 2 4 4 5 5 5 8 8 8 9 9
10:11	2 3 3 3 4 5 5 5 6 7 9	:	0 0 1 2 4 4 4 5 6 7 9
8: 9	0 2 3 5 7 7 9 9	:	0 0 2 2 3 4 4 5 5 6 9 9
6: 7	3 3 4	:	0 2 4 5
4: 5	4 4 8	:	3 7 7 9
2: 3	1 7 9		
0: 1			
mm			

QUADRO II - Ramo e folhas da precipitação pluviométrica mensal de Campinas, de 1890 a 1983.

Período II - fev/mar/abr

52:53			
50:51			
48:49			
46:47			
44:45			
42:43			
40:41			
38:39	1 3		
36:37		:	3
34:35	2 3 5	:	3 7 9
32:33	0 1 8	:	4 5
30:31	0 1 2	:	8
28:29	3 6 6 9	:	9
26:27	4 7	:	1 7
24:25	0 1 2 4 6 6 9	:	1 5 7 8 8
22:23	2 4 6 9 9	:	0 0 2 3 6 9 9
20:21	0 0 2 2 3 4 4 6 9	:	0 1 1 2 2 5 8 8 9
18:19	2 3 3 4 4 6 7 8	:	1 1 2 2 3 4 4 4 6
16:17	0 1 2 3 3 3 5 8	:	0 0 0 1 2 2 4 4 4 7 7 8 9
14:15	0 2 3 5 5 6 7 8 8 9 9	:	0 0 0 0 2 2 3 5 6 7 7 9
12:13	1 2 2 5 5 8 8 8 9 9 9	:	0 2 2 3 3 3 5 7 8 8 9
10:11	0 3 3 5 6 6 7 8 8 9 9	:	0 0 1 1 1 3 3 3 4 4 5 5 6 6 8 9
8: 9	0 0 0 2 2 2 2 3 6 6 8 8	:	4 6 8 8 9 9
6: 7	3 4 4 4 7 7 8 9 9	:	0 1 2 2 4 4 5 5 6 6 9 9
4: 5	0 0 2 3 3 3 3 3 4 4 6 7	:	0 0 1 2 2 3 4 5 6 6 6 8 9
2: 3	0 1 1 3 4 5 5 6 6 7 7 8	:	0 0 1 2 4 4 4 6 6 8 8 9 9
0: 1	4 4 5 8 9	:	2 2 4 6 6 6 6 9
mm			

QUADRO III – Ramo e folhas da precipitação pluviométrica mensal de Campinas, de 1890 a 1983.

Período III – mai/jun/jul

36:37			
24:35			
32:33			
30:31	5		
28:29			
26:27			
24:25			
22:23	9		
20:21	8		
18:19	8 8	:	5 9
16:17			
14:15	1 4 6 6 9	:	6 8
12:13	0 3 7 8 9	:	7 8
10:11	1 2 4 4	:	0 0 1 2 3 4 8
8: 9	0 0 3 3 4 5 5 5 6 6 6 7 7 8	:	0 2 2 6 7
	9 9 9		
6: 7	0 0 2 3 3 3 3 4 4 4 5 5 6 9	:	0 2 2 2 4 4 6 7 8 8 8 8 9 9
4: 5	0 0 0 1 1 1 1 1 2 2 2 2 3 3	:	0 0 0 1 1 1 1 5 5 5 5 8 9 9
	4 4 4 4 5 5 5 7 8 9 9 9 9		
2: 3	0 0 0 0 0 1 1 1 1 1 2 3 3 4	:	0 0 0 0 1 1 1 1 1 1 2 2 2 2
	5 5 5 5 5 7 7 7 7 7 7 8		3 4 4 4 4 5 5 5 6 7 7 7 7 7
			8 8 9 9 9
0: 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0	:	0 0 0 0 0 0 0 1 1 1 1 1 2 4
	0 0 0 0 0 0 0 0 0 0 0 0 0 0		4 4 4 4 4 4 5 5 6 6 6 7 7 8
	1 1 1 1 2 2 2 2 3 3 3 4 4 5		8 8 9 9 9 9
	5 5 6 6 6 7 7 7 7 7 7 7 7 7		
	7 7 8 8 8 8 8 9 9		
mm			

QUADRO IV - Ramo e folhas da precipitação pluviométrica mensal de Campinas, de 1890 a 1983.

Período IV - ago/set/out

32:33			
30:31	4	:	0
28:29		:	4 5
26:27	4	:	7
24:25		:	1 7
22:23	1 7		
20:21	3 4 4 7	:	1 4 7
18:19	0 1 1 1 2 9	:	6 6
16:17	8 9	:	0 1 6
14:15	1 1 2 3 3 4 4 4 5 6 6 7 7 8	:	8 8 9
12:13	0 0 2 3 4 6 8 9	:	0 2 4 4 8 9
10:11	2 2 3 4 5 6 7	:	1 1 2 3 3 4 5 5 6 6 8 9 9
8: 9	0 0 1 1 1 3 3 4 5 5 5 6 7 8	:	0 0 0 1 1 1 2 2 4 6 6 6 6
	8 9 9		6 7 7 8
6: 7	2 2 3 3 4 4 4 4 4 6 7 7 7 8	:	0 1 1 2 2 2 3 3 3 3 4 4 5
	9 9 9		6 7 7 7 7
4: 5	0 1 1 2 2 2 3 4 4 6 7 7 7 9	:	0 0 1 2 3 3 3 4 5 5 5 5 6
	9 9		6 8 8 9
2: 3	0 0 0 2 2 2 3 4 4 5 5 5 6 6	:	0 1 2 2 2 3 4 4 4 5 5 6 6
	6 6 6 6 7 7 8 8 8 9 9		7 7 7 8 9
0: 1	0 0 0 0 0 0 0 0 0 0 0 0 0 0	:	0 0 0 1 1 2 2 3 3 4 5 5 5
	0 1 2 2 3 3 3 4 5 5 5 6 6 7		5 5 6 7 8 9 9
mm	7 8 8 8 9 9 9 9 9		

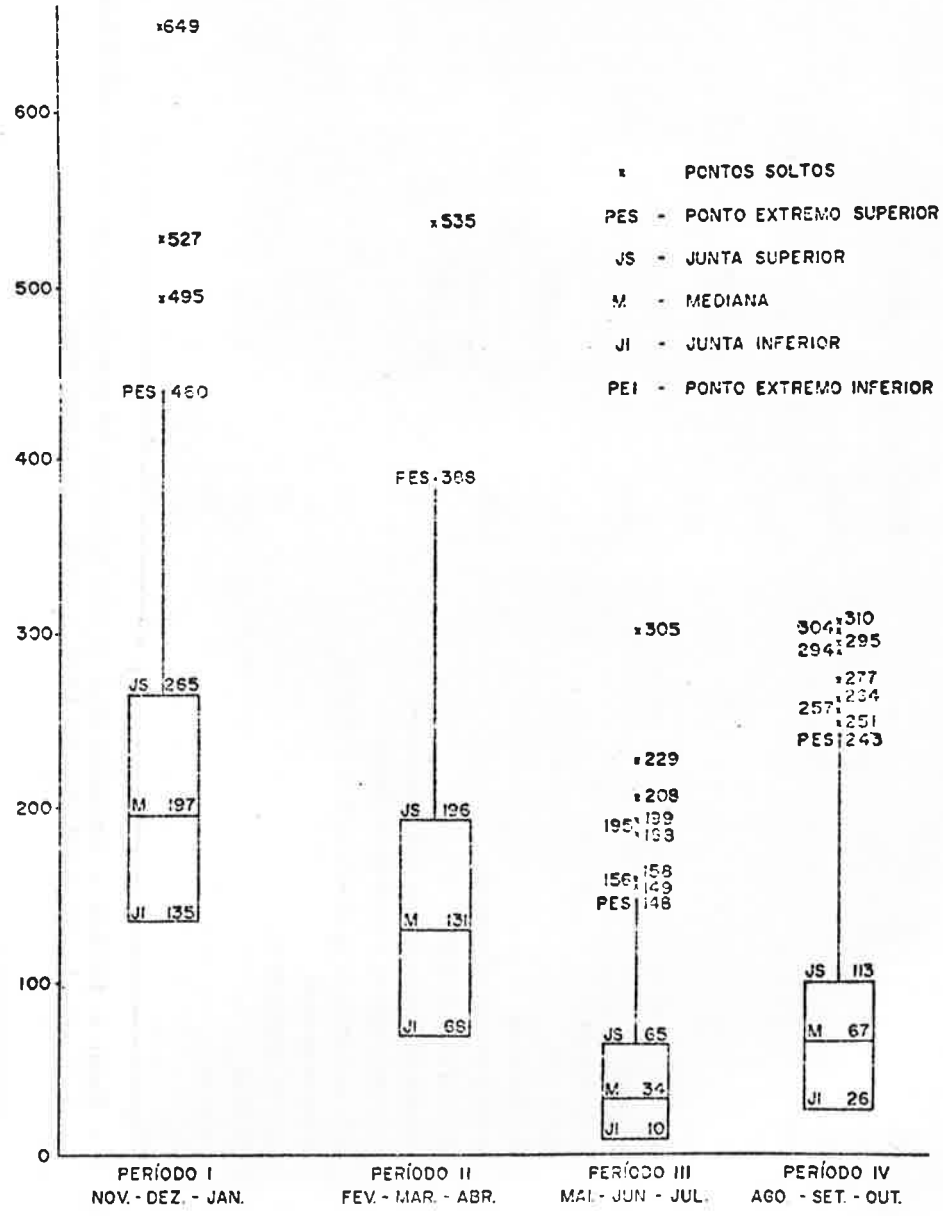

FIGURA 1 - Esquema dos cinco números da precipitação plu

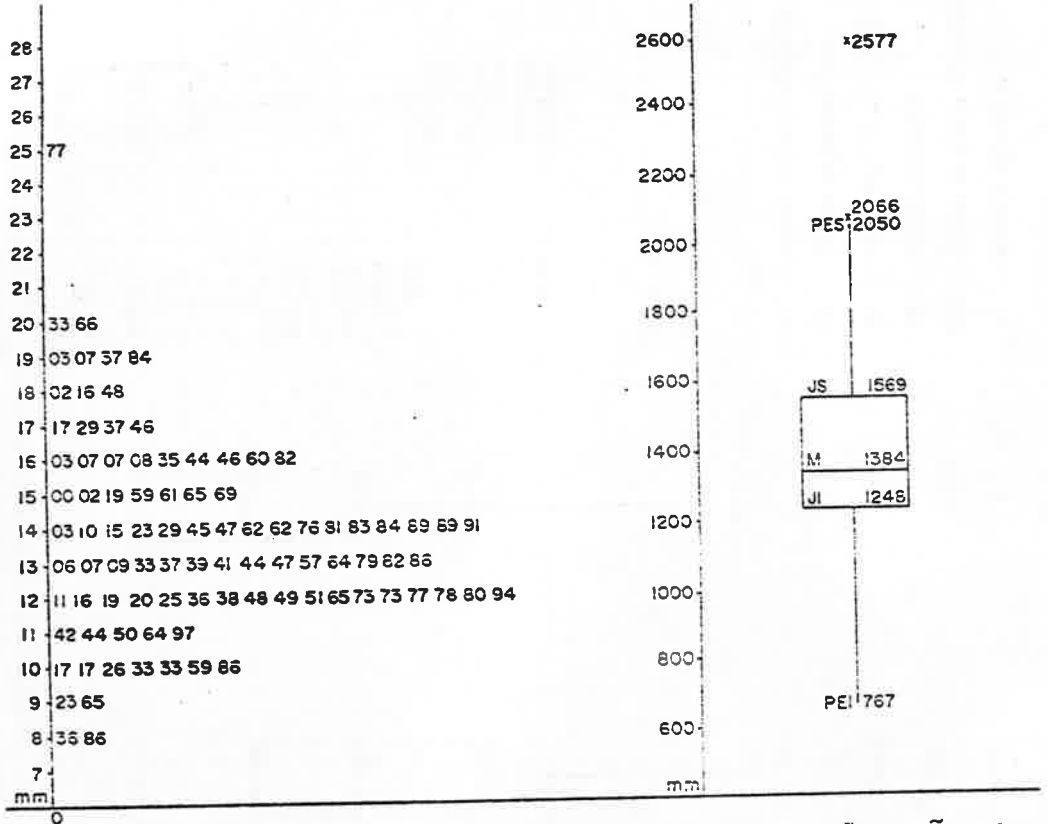

FIGURA 2 - Ramo e folhas e esquema dos cinco números da precipitação pluviométrica total anual, no período de 1890 a 1983, de Campinas.

Período II:	PES	388
	JS	196
	M	131
	JI	68
	PEI	
Período III:	PES	148
	JS	65
	M	34
	JI	10
	PEI	
Período IV:	PES	243
	JS	113
	M	67
	JI	26
	PEI	
Totais anuais:	PES	2050
	JS	1569
	M	1384
	JI	1248
	PEI	767

Em seguida, esses valores foram colocados em gráficos (figuras 1 e 2), onde foram também assinalados os pontos soltos (x) que são os prováveis valores atípicos da precipitação pluviométrica por período.

Nos gráficos não foram colocados os pontos extremos inferiores, em virtude de serem dados negativos, com exceção do gráfico da precipitação anual.

RESULTADOS E DISCUSSÃO

Período I: meses de novembro - dezembro - janeiro
Nesse período os pontos soltos (x) da figura 1 são:

485mm, dezembro de 1936
527mm, janeiro de 1899
649mm, janeiro de 1929.

De 282 dados, somente três estão além do limite es-

Período II: meses de fevereiro - março - abril
Nesse período, os pontos soltos (x) do gráfico são:

535mm, fevereiro de 1970

De 282 dados deste período, somente obtivemos um valor acima do ponto extremo superior.

Período III: meses de maio - junho - julho
Nesse período, os pontos soltos (x) da figura 1 são:

149mm, julho de 1976
156mm, maio de 1976
158mm, maio de 1929
188mm, maio de 1958
195mm, junho de 1982
199mm, junho de 1945
208mm, junho de 1919
229mm, junho de 1983
305mm, maio de 1983

Dos 282 dados deste período, foram obtidos nove valores acima do ponto extremo superior.

Período IV: meses de agosto - setembro - outubro
Nesse período, os pontos soltos (x) do gráfico são:

251mm, outubro de 1964
257mm, setembro de 1923
264mm, outubro de 1892
277mm, outubro de 1982
294mm, outubro de 1938
295mm, setembro de 1983
304mm, outubro de 1895
310mm, outubro de 1981.

Nesse período, foram obtidos oito valores acima do ponto extremo superior.

Totais anuais: de 1890 a 1983, os pontos soltos foram:

2066mm, ano de 1970
2577mm, ano de 1983

Dos 94 totais anuais, somente os dois anos acima

Geralmente, a previsão de chuvas é feita utilizando se dados de um determinado período e calculando-se a média deles.

Dessa forma, a previsão de chuva baseada em um período de 27 anos, seria:

janeiro	- 238,1mm
fevereiro	- 193,2mm
março	- 131,4mm
abril	- 68,6mm
maio	- 52,8mm
junho	- 53,8mm
julho	- 35,9mm
agosto	- 37,2mm
setembro	- 62,2mm
outubro	- 140,7mm
novembro	- 138,4mm
dezembro	- 218,5mm

Um ano atípico como o de 1983 teria essas médias muito alteradas nos meses secos, ao passo que a mediana não interfere nessa determinação.

Pelos gráficos apresentados, pode-se ter uma idéia bastante clara e precisa da ocorrência de chuvas e de como é a distribuição delas.

Note-se que nos meses mais secos, há maior ocorrência de pontos soltos, o que não acontece nos outros meses.

CONCLUSÕES GERAIS

Pode-se concluir que as duas técnicas apresentadas permitiram visualizar com facilidade a ocorrência de chuva num período de 94 anos, e a distribuição delas por períodos.

O ano de 1983 foi o mais chuvoso desse período, seguido pelo ano de 1970.

O mês mais chuvoso dos 1128 meses foi o de janeiro de 1929.

Todos os valores que saem da normalidade dos dados ficam claramente visíveis nos gráficos, podendo-se consi

SUMMARY

Outliers, or spurious, or unrepresentative, or mavericks observations in a set of data may distor the conclusions of the variance analysís, as the general mean, variance and variation coefficient.

Exploratory data analysis is a preliminary technique to detect outliers in great amount of data.

Stem-and-leaf and meadian, hinges, and hinges differences are the graphical techniques studied in this paper.

Data set were collected in Instituto Agronômico, Campinas, SP, of rain precipitation, in mm, from 1890 to 1983.

LITERATURA CITADA

BARNETT, V. & T. LEWIS, 1979. Outliers in Statistical Data, Wiley Series in Probability and Mathematical Statistics - Applied, New York, USA.

BARNETT, V., 1978. The study of outliers: purpose and model, Applied Statistics, vol. 27.

COOK, R.D., 1979. Influential observations in linear regression, JASA, vol. 74.

DACHS, J.N.W., 1980. Violência em algumas capitais brasileiras - Um exemplo de análise exploratória de dados. **Revista Brasileira de Estatística.**

DIXON, W.J., 1950. Annalysis of extreme values. **Annals of Math. Stat.**, vol. 21.

DIXON, W.J., 1951. Ratios involving extreme values. **Annals of Math. Stat.**, vol. 22

DIXON, W.J., 1953. Processing data for outliers. **Biometrics** vol. 9.

KEMPTHORNE, O., 1952. The design of analysis of experiments, John Wiley and Sons Inc, New York, USA.

LUND, R.E., 1975. Tables for outlier. **Technometrics**, vol. 17.

MOSTELLER, F. et al., s/data. **Statistics by example -exploring data.**