Técnicas Computacionais em Probabilidade e Estatística I

Aula IV

Chang Chiann
MAE 5704- IME/USP
1º Sem/2008

Transformações

Problema 1: em muitas situações de interesse, a distribuição da amostra é assimétrica e pode conter valores atípicos → a suposição de normalidade não está satisfeita.

Procedimento utilizado: fazer uma transformação das observações de modo a obter uma distribuição mais simétrica e próxima da normal.

Uma transformação bastante usada é:

$$x^{(p)} = \begin{cases} x^{p}, p > 0 \\ \ln(x), p = 0 \\ -x^{p}, p < 0 \end{cases}$$

p = 0,5 → $(x)^{1/2}$

P = -1 → -1/x

A razão para a mudança de sinal quando p<0 é a de assegurar que os dados transformados tenham a mesma ordem relativa que os originais.

Assimetria à direita → P<1

Assimetria à esquerda → P>1

Na prática, consideramos valores de p na seqüência

-3, -2, -1, -1/2, -1/3, -1/4, 0, ¼, 1/3, ½, 1, 2, 3

• Para cada valore de p, obtemos gráficos apropriados: histograma, box-plot e gráficos de simetria para os dados originais e transformados, de modo que podemos escolher o valor mais adequado para p.

• para cada valor de p na seqüência se calcule a média, mediana e um estimador de escala(desvio padrão ou algum estimador robusto) e então se escolher o valor que minimiza

$$d_p = (\text{média-mediana})/(\text{medida de escala})$$

que pode ser vista como uma medida de assimetria. Numa distribuição simétrica, $d_p=0$.

Problema 2: estabilizar a variância às vezes é mais importante do que tornar a distribuição aproximadamente normal.

Suponha X uma v.a. com

$E(x) = \mu, \quad Var(x) = h^2(u)\,\sigma^2$ (não é constante)

Idéia: $X \rightarrow g(X)$ tal que $Var[g(X)] = cte.$

Considere uma expansão de Taylor de g(X) até primeira ordem, ao redor de g(μ):

$$g(X) \approx g(\mu) + (X-\mu)g'(\mu)$$

Então:

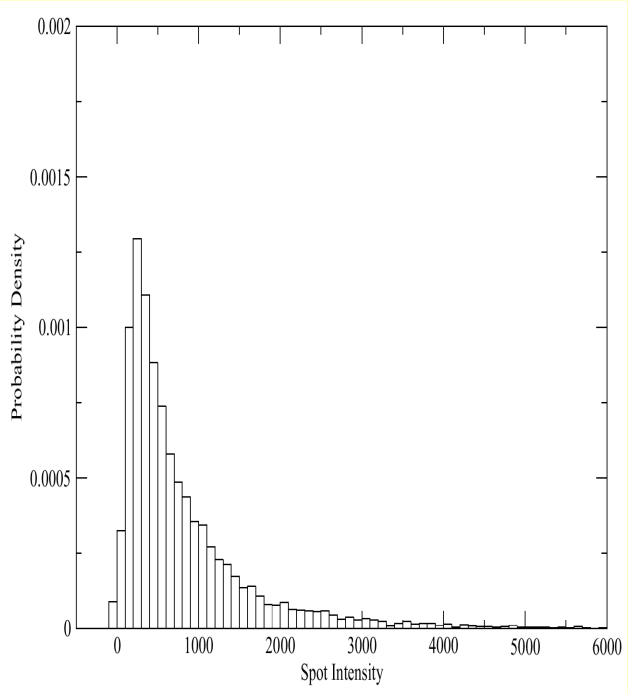
$$Var(g(X)) \approx [g'(\mu)]^2 Var(X) = g'(\mu)^2 h^2(\mu)\sigma^2$$

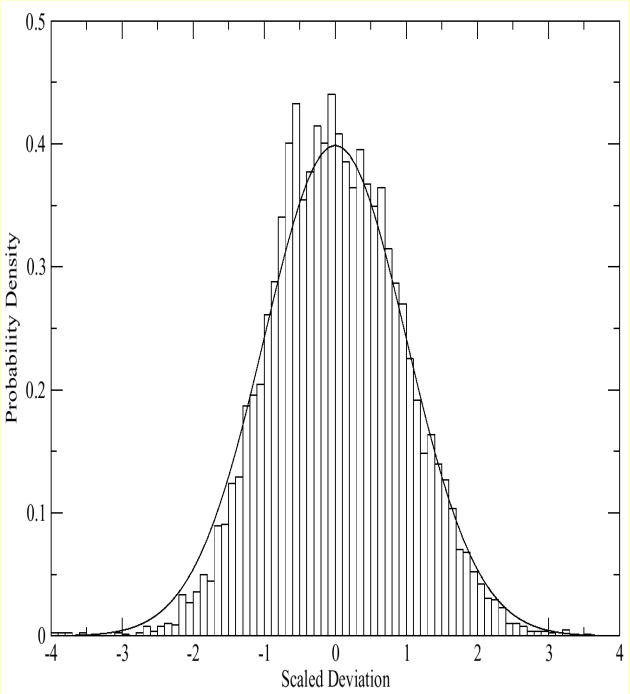
Assim,

$$Var(g(X)) \approx cte \rightarrow g'(\mu) = 1/h(\mu)$$

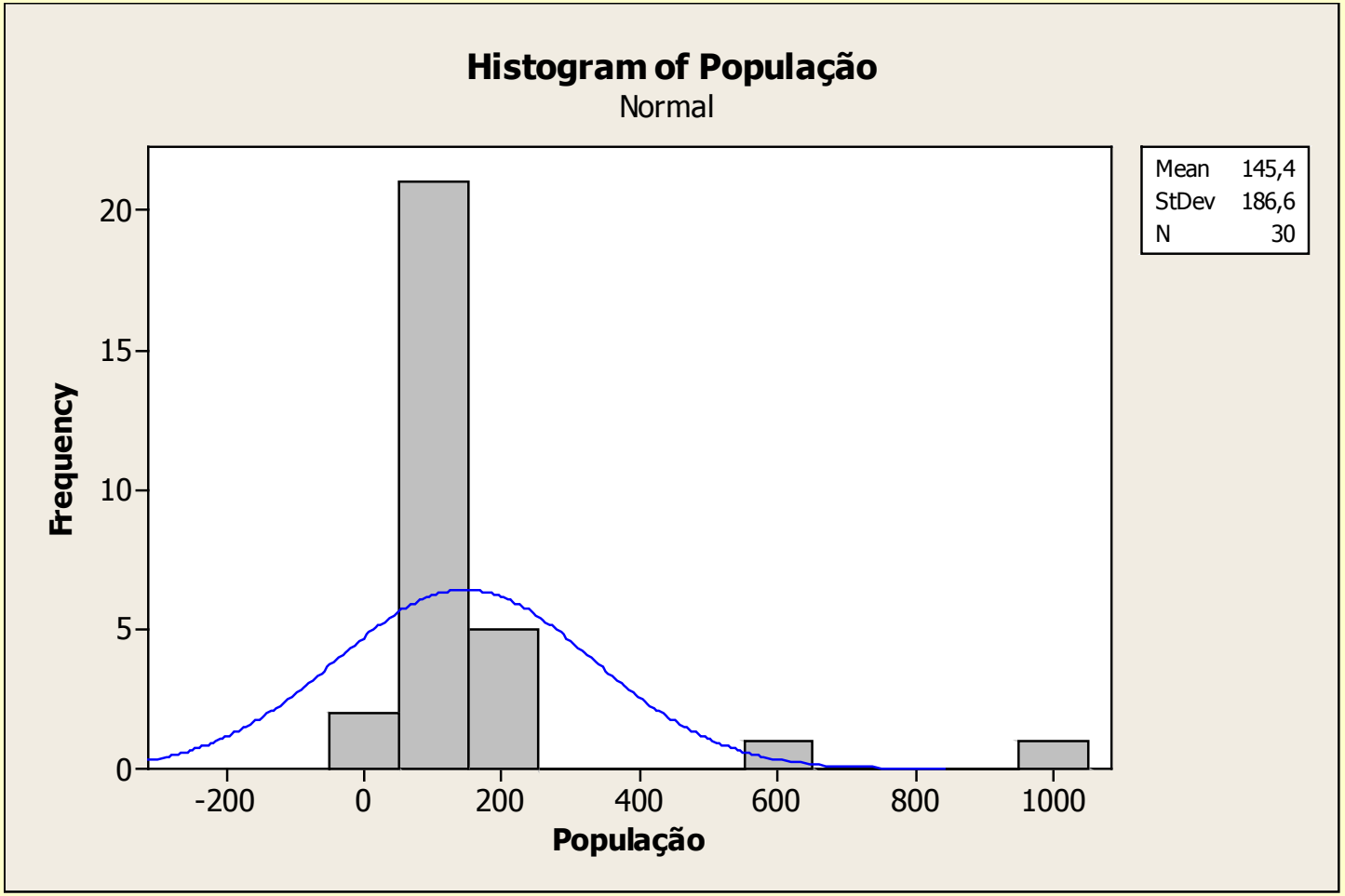
Transformação Log

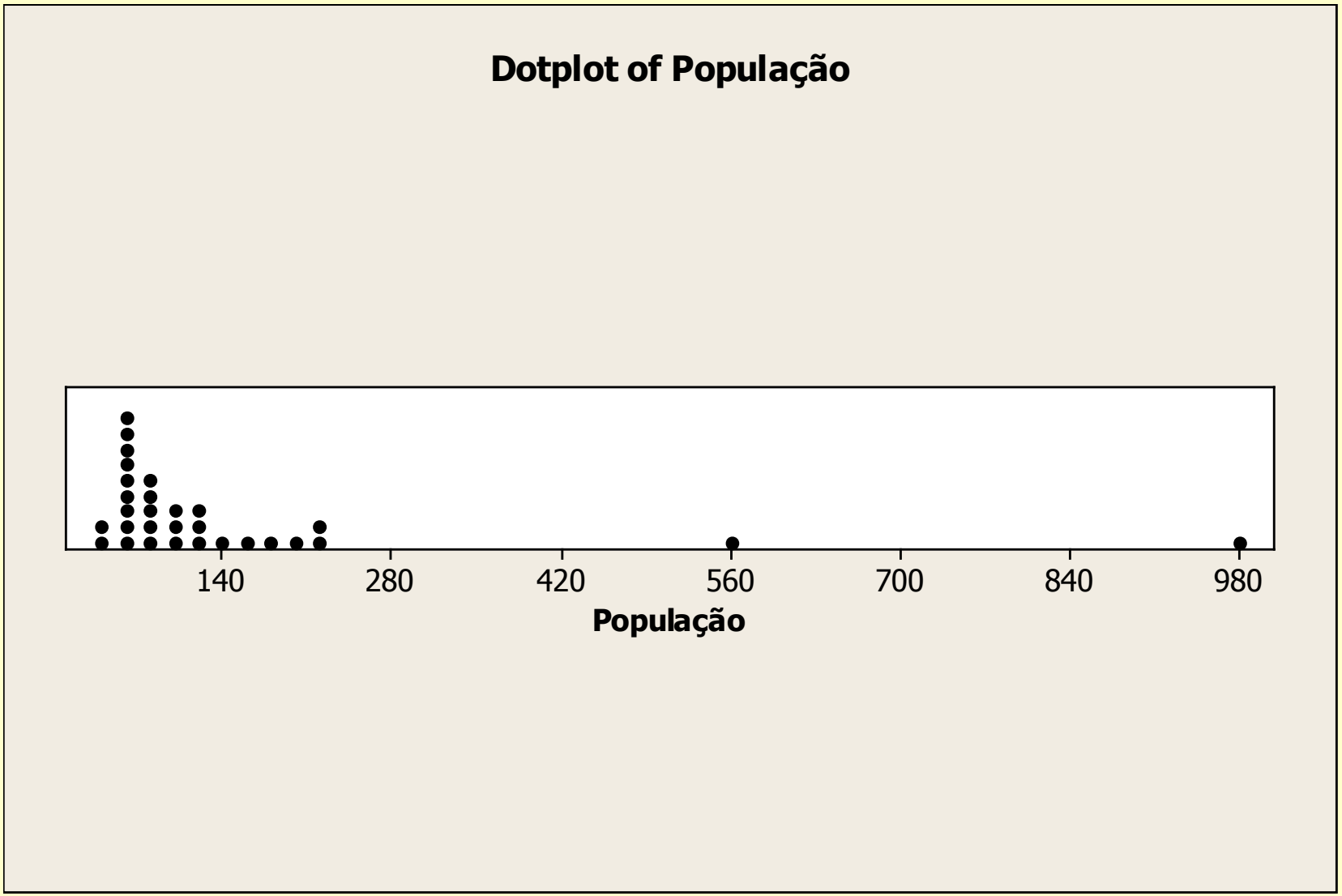
- Modelo de Efeitos Multiplicativos $\left(y_j = \lambda e^{\beta C_j} \right)$ $\Rightarrow$ Linearizar
- Distribuição Assimétrica Positiva $\Rightarrow$ Normalizar
- Estabilizar Variância

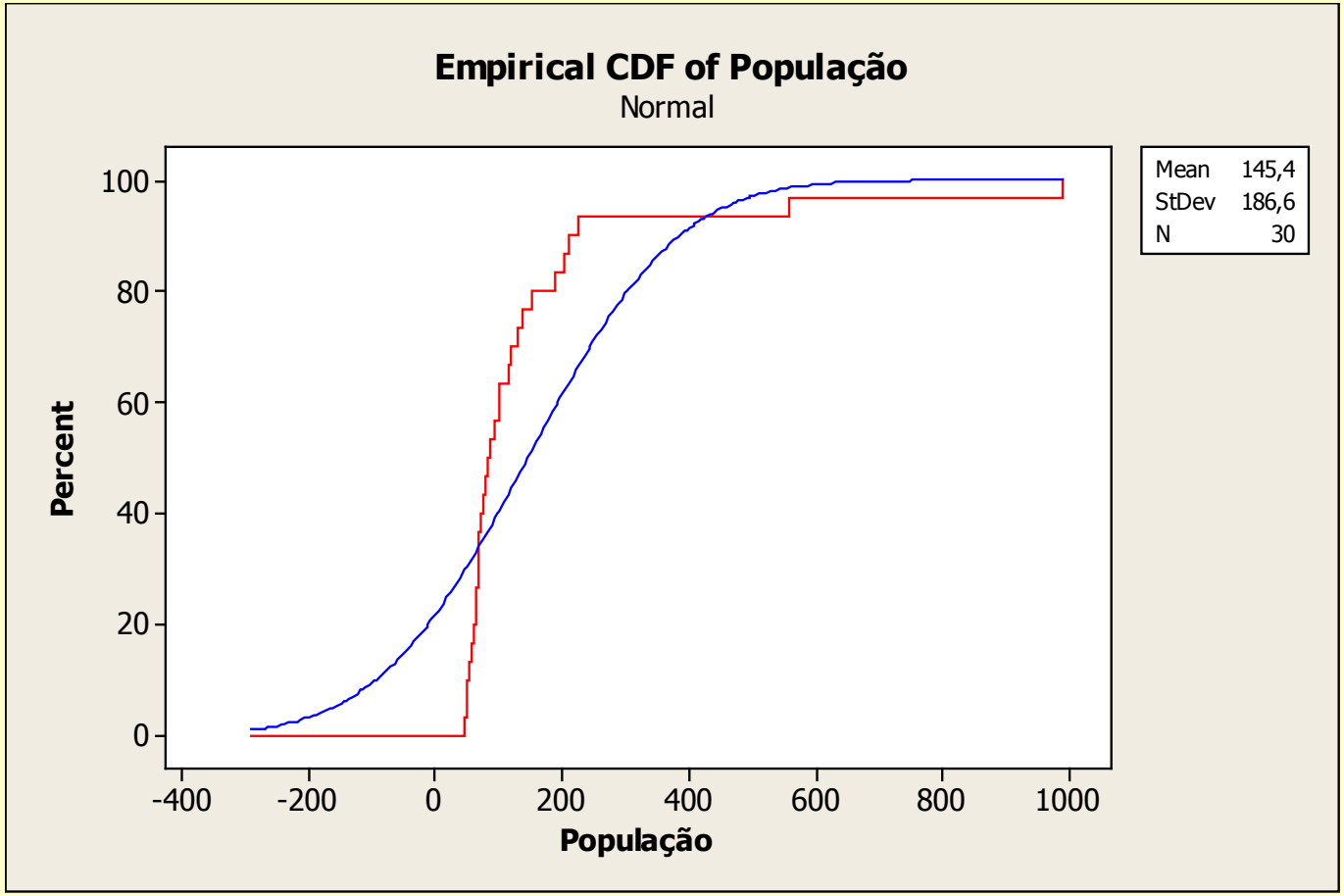


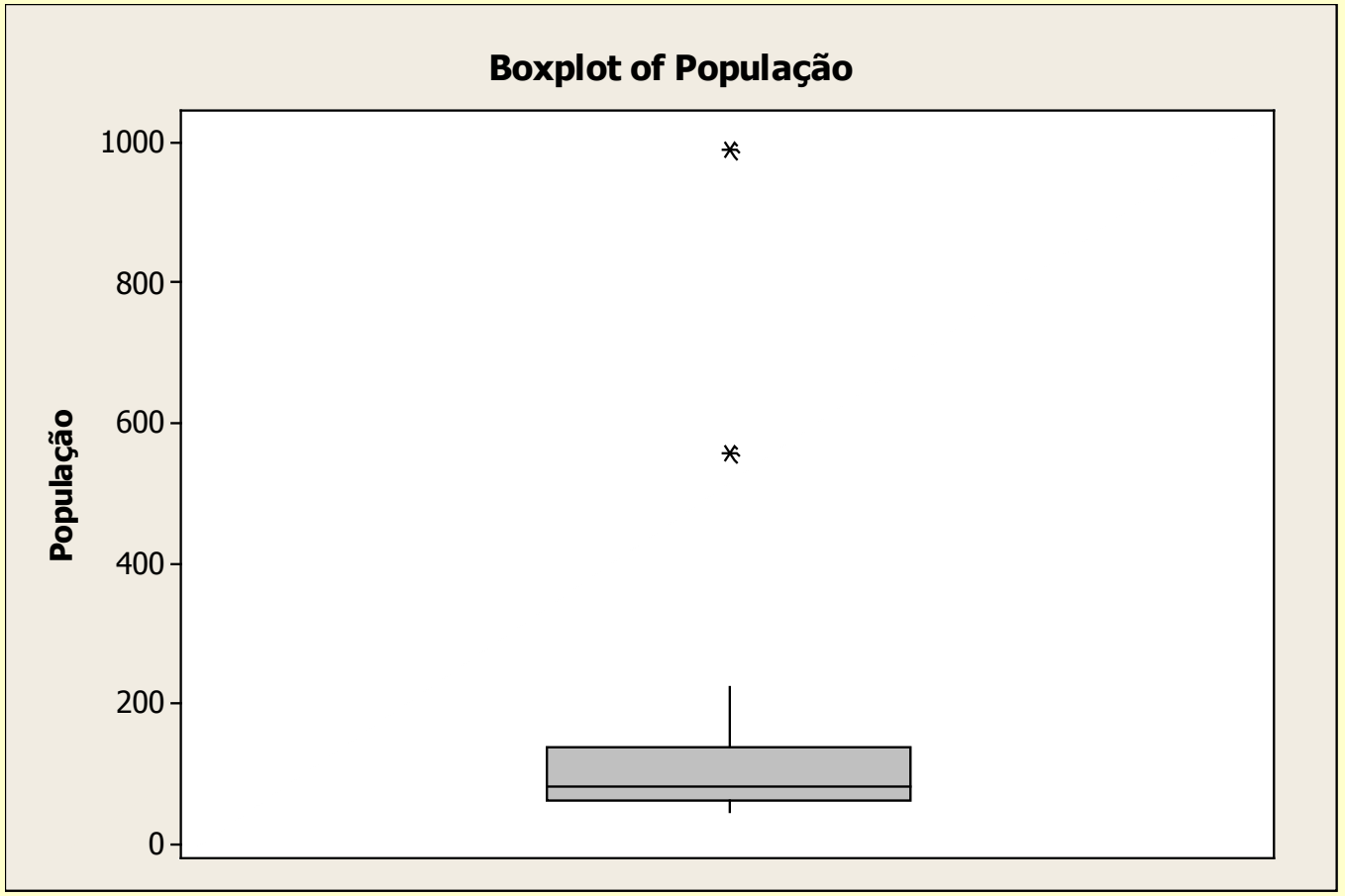


Populações(em 10.000 habitantes) dos 30 municípios mais populosos do Brasil			
Município	**População**	**Município**	**População**
São Paulo	988,8	Nova Iguaçu	83,9
Rio de Janeiro	556,9	São Luis	80,2
Salvador	224,6	Maceió	74,7
Belo Horizonte	210,9	Duque de Caxias	72,7
Fortaleza	201,5	São Bernardo do Campo	68,4
Brasilia	187,7	Natal	66,8
Curitiba	151,6	Teresina	66,8
Recife	135,8	Osasco	63,7
Porto Alegre	129,8	Santo André	62,8
Manaus	119,4	Campo Grande	61,9
Belem	116,0	João Pessoa	56,2
Goiania	102,3	Jaboatão	54,1
Guarulhos	101,8	Contagem	50,3
Campinhas	92,4	São Jose dos Campos	49,7
São Gonçalo	84,7	Ribeirão Preto	46,3


Histogram of População
Normal
Mean 145,4
StDev 186,6
N 30
Frequency
20
15
10
5
0
-200
0
200
400
600
800
1000
População

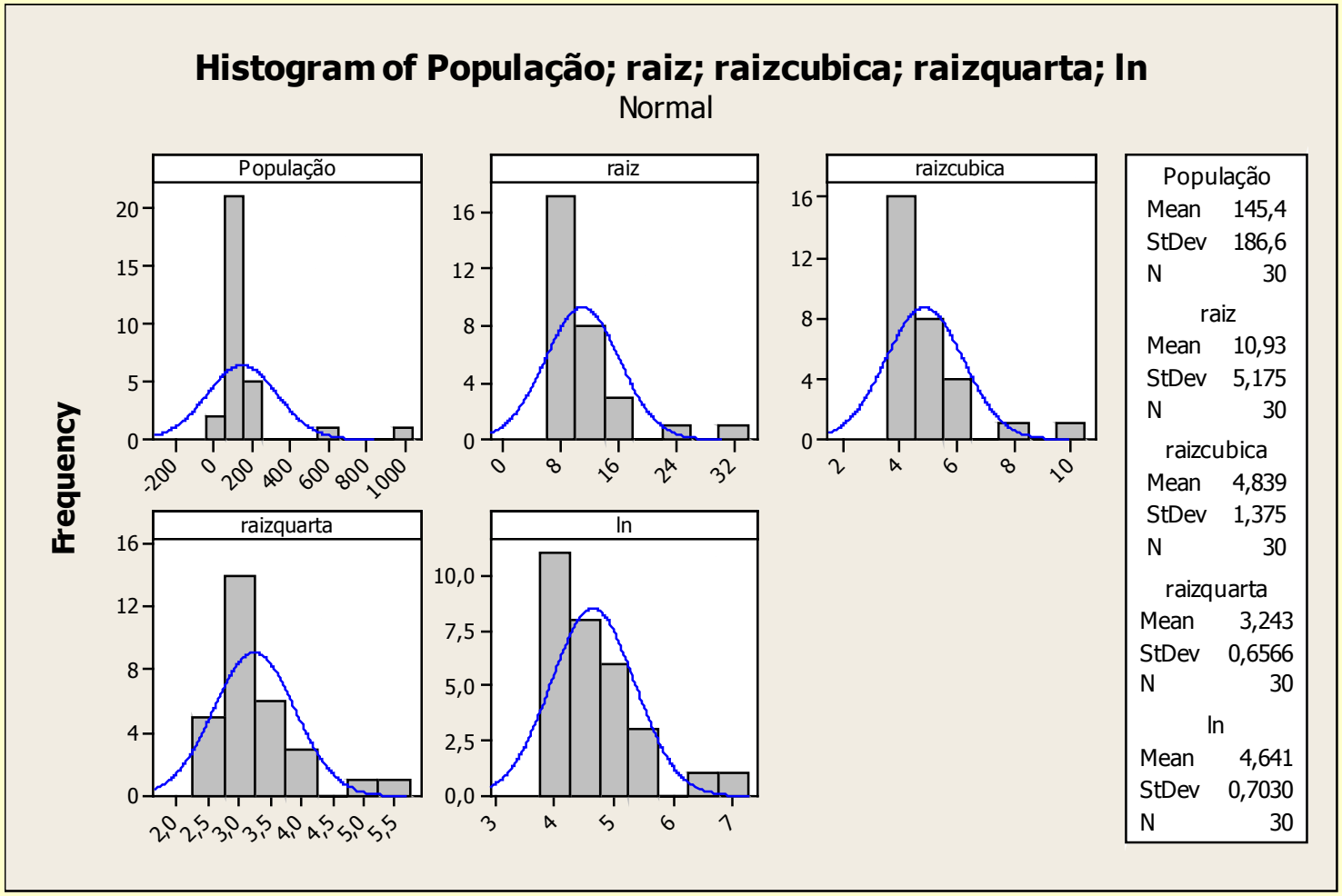

Dotplot of População
140
280
420
560
700
840
980
População

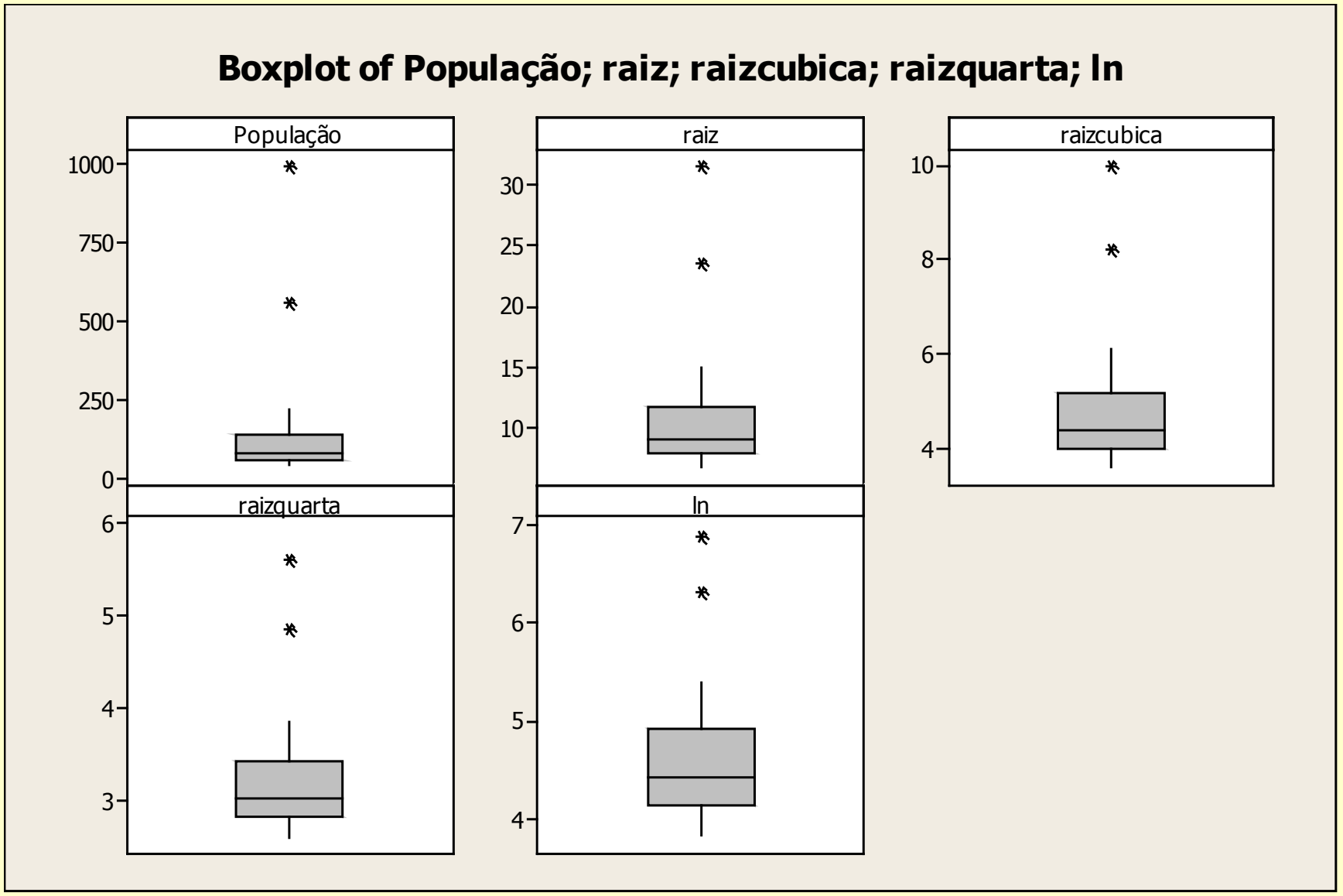

Empirical CDF of População
Normal
Mean 145,4
StDev 186,6
N 30
Percent
100
80
60
40
20
0
-400
-200
0
200
400
600
800
1000
População


Boxplot of População
População
1000
800
600
400
200
0

Descriptive Statistics: População

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
População	30	0	145,4	34,1	186,6	46,3	63,5	84,3	139,8	988,8


Histogram of População; raiz; raizcubica; raizquarta; ln
Normal
Frequency
População
raiz
raizcubica
raizquarta
ln
População
Mean 145,4
StDev 186,6
N 30
raiz
Mean 10,93
StDev 5,175
N 30
raizcubica
Mean 4,839
StDev 1,375
N 30
raizquarta
Mean 3,243
StDev 0,6566
N 30
ln
Mean 4,641
StDev 0,7030
N 30


Boxplot of População; raiz; raizcubica; raizquarta; ln
População
raiz
raizcubica
raizquarta
ln
1000
750
500
250
0
30
25
20
15
10
10
8
6
4
6
5
4
3
7
6
5
4

Descriptive Statistics: População; raiz; raizcubica; raizquarta; ln

Variable	N	N*	Mean	SE Mean	StDev	CoefVar	Minimum	Q1	Median
População	30	0	145,4	34,1	186,6	128,31	46,3	63,5	84,3
raiz	30	0	10,933	0,945	5,175	47,33	6,804	7,967	9,181
raizcubica	30	0	4,839	0,251	1,375	28,41	3,591	3,989	4,385
raizquarta	30	0	3,243	0,120	0,657	20,25	2,609	2,823	3,030
ln	30	0	4,641	0,128	0,703	15,15	3,835	4,151	4,434

Variable	Q3	Maximum	Skewness	Kurtosis
População	139,8	988,8	3,76	15,40
raiz	11,818	31,445	2,76	8,78
raizcubica	5,188	9,963	2,38	6,62
raizquarta	3,438	5,608	2,18	5,60
ln	4,939	6,896	1,61	3,00