



Universidade Federal do ABC

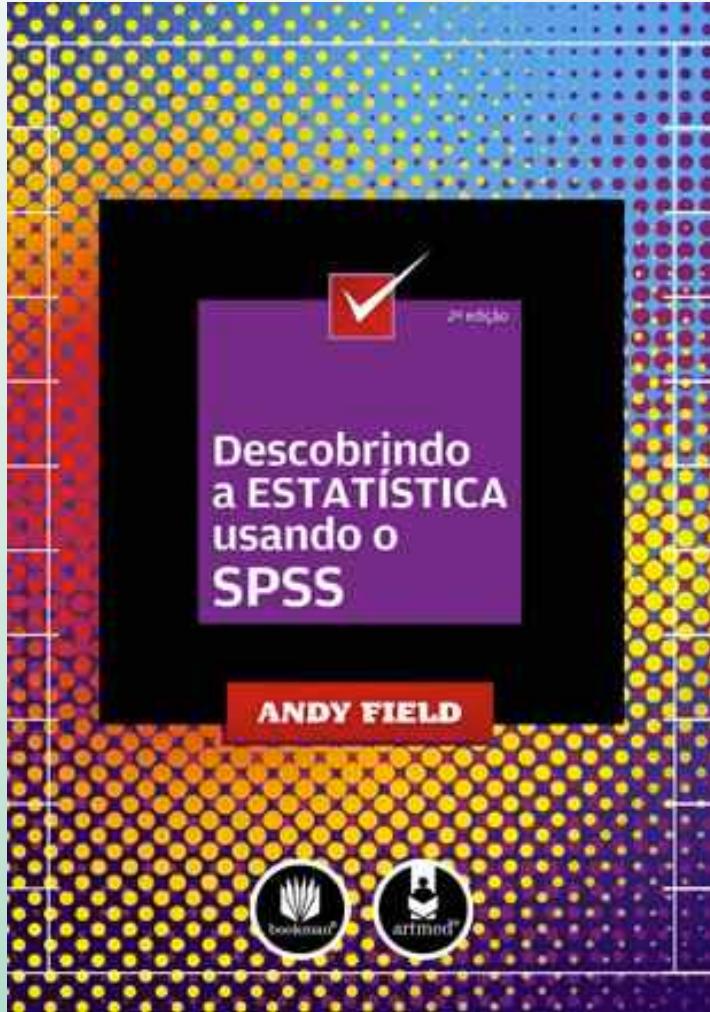
Análise Exploratória de Dados no SPSS

Gráficos e Estatísticas Descritivas

Flávia F. Feitosa

BH1350 – Métodos e Técnicas de Análise da Informação para o Planejamento
Junho de 2015

Leitura de Referência



Capítulo 1 – p. 47-59

Capítulo 3
Explorando Dados
(p. 85 – 124)

Inferência Estatística

Método científico para tirar conclusões sobre os parâmetros da **população** a partir da coleta, tratamento e análise dos dados de uma **amostra** recolhida dessa população.

Inferência Estatística

PARAMÉTRICA: Admite que a distribuição da população tem uma forma matemática conhecida, embora contendo um ou mais parâmetros desconhecidos.

NÃO-PARAMÉTRICA: Pretende-se conhecer a forma da distribuição

Inferência Estatística

PARAMÉTRICA: Admite que a distribuição da população tem uma forma matemática conhecida, embora contendo um ou mais parâmetros desconhecidos.

Em muitos casos, uma distribuição normal.

Normalidade dos Dados

Assume-se que os dados foram obtidos de uma ou mais populações normais.

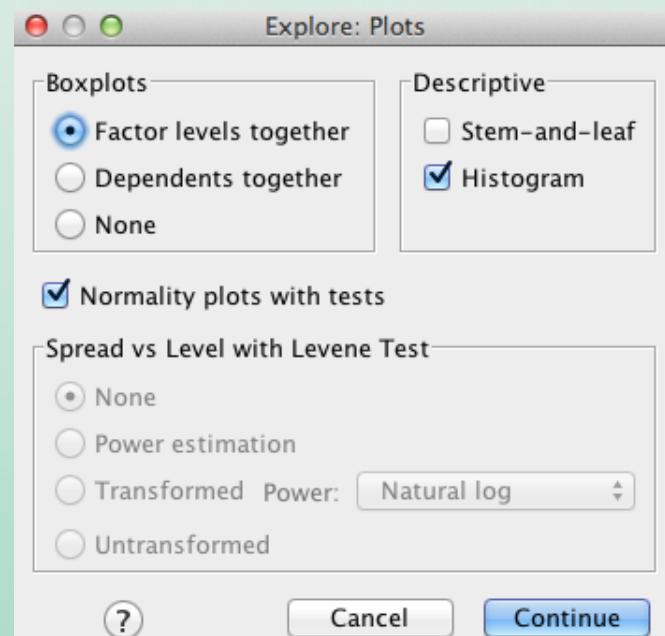
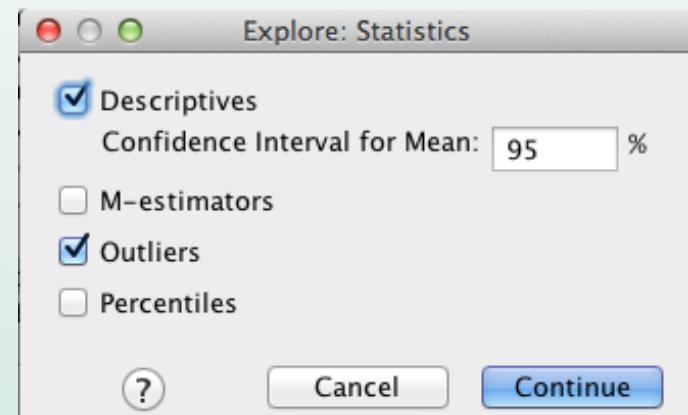
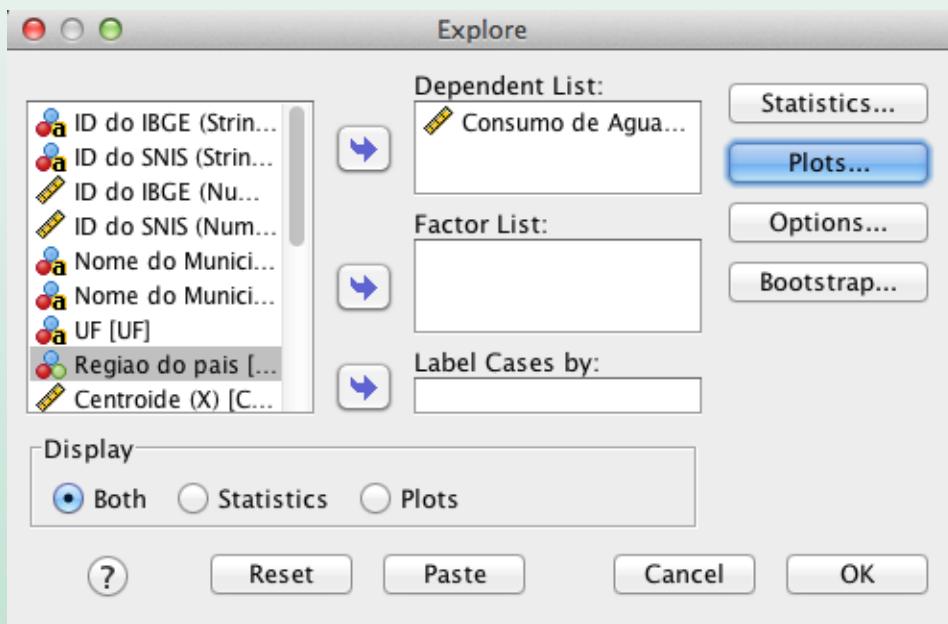
Pesquisadores verificam suas amostras (histograma e outros testes) e se a amostra assemelha-se a uma normal, assume-se que a população também o é.

Explorando Dados no SPSS

- 1. Distribuição dos Escores e Valores Atípicos
(verificação da normalidade dos dados)**
- 2. Corrigindo problemas nos dados**
- 3. Transformando dados**

Explorando Dados

- Abra o arquivo “Agua2010_SNIS.sav”
- Analyze > Descriptive > Explore > Statistics... > Plots...



Explorando Dados

A Distribuição é Normal?

Descriptives

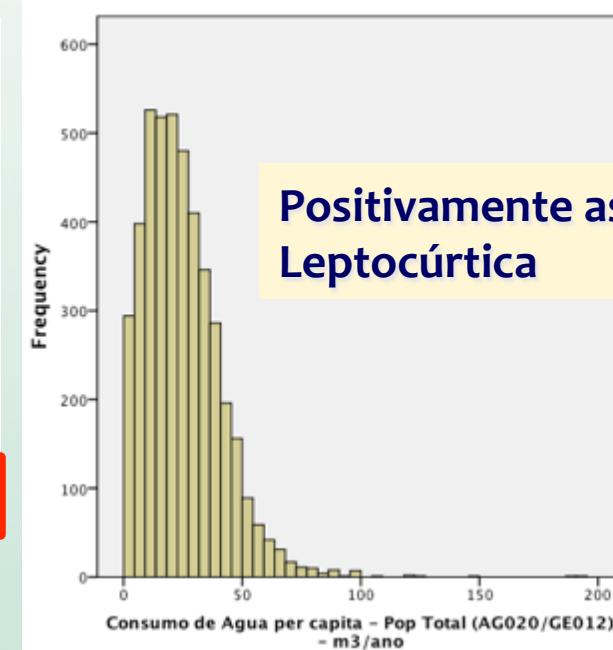
		Statistic	Std. Error
Consumo de Agua per capita - Pop Total (AG020/GE012) - m3/ano	Mean	24.7688911	.247318170
	95% Confidence Interval for Mean	24.2840235	
	Lower Bound	24.2840235	
	Upper Bound	25.2537587	
	5% Trimmed Mean	23.7000167	
	Median	22.3544550	
	Variance	270.171	
	Std. Deviation	16.4368929	
	Minimum	.000502008	
	Maximum	193.493151	
	Range	193.492649	
	Interquartile Range	21.0574510	
	Skewness	1.532	.037
	Kurtosis	7.097	.074

Uma distribuição normal deverá ter assimetria e curtose nulas. Será possível que a distribuição da população assemelhe-se a uma normal?

Converter assimetria e curtose em escores-z

$$Z_{Assimetria} = \frac{S - 0}{EP_{Assimetria}}$$

$$Z_{Curtose} = \frac{S - 0}{EP_{Curtose}}$$



Positivamente assimétrica
Leptocúrtica

Explorando Dados

A Distribuição é Normal?

Descriptives

		Statistic	Std. Error
Consumo de Agua per capita - Pop Total (AG020/GE012) - m3/ano	Mean	24.7688911	.247318170
	95% Confidence Interval for Mean	Lower Bound	24.2840235
		Upper Bound	25.2537587
	5% Trimmed Mean	23.7000167	
	Median	22.3544550	
	Variance	270.171	
	Std. Deviation	16.4368929	
	Minimum	.000502008	
	Maximum	193.493151	
	Range	193.492649	
	Interquartile Range	21.0574510	
	Skewness	$z_s = (1.532 - 0) / .037 = 41.4$	1.532
	Kurtosis	$z_k = (7.097 - 0) / .074 = 95.9$.037
			.074

Curtose e Assimetria Significativa

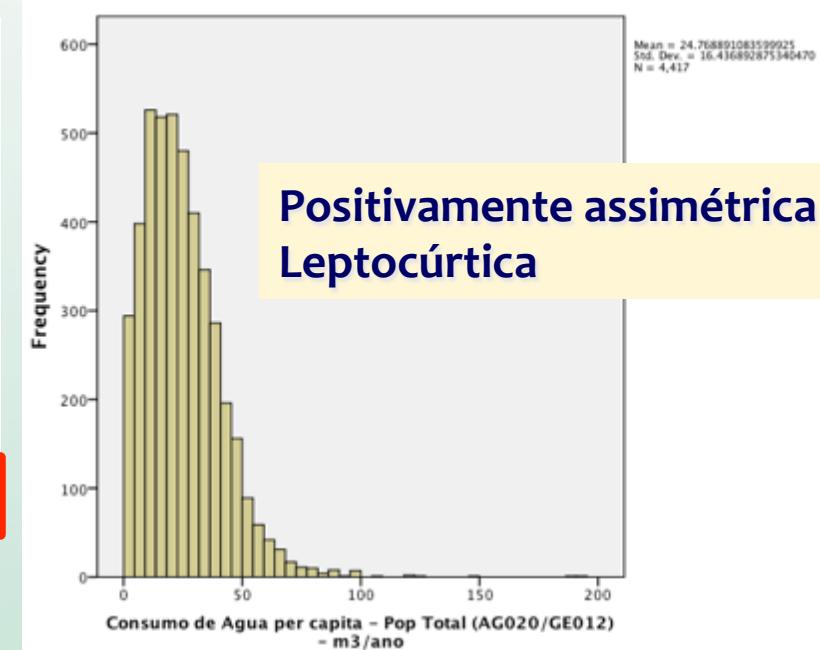


Uma distribuição normal deverá ter assimetria e curtose nulas. Será possível que a distribuição da população assemelhe-se a uma normal?

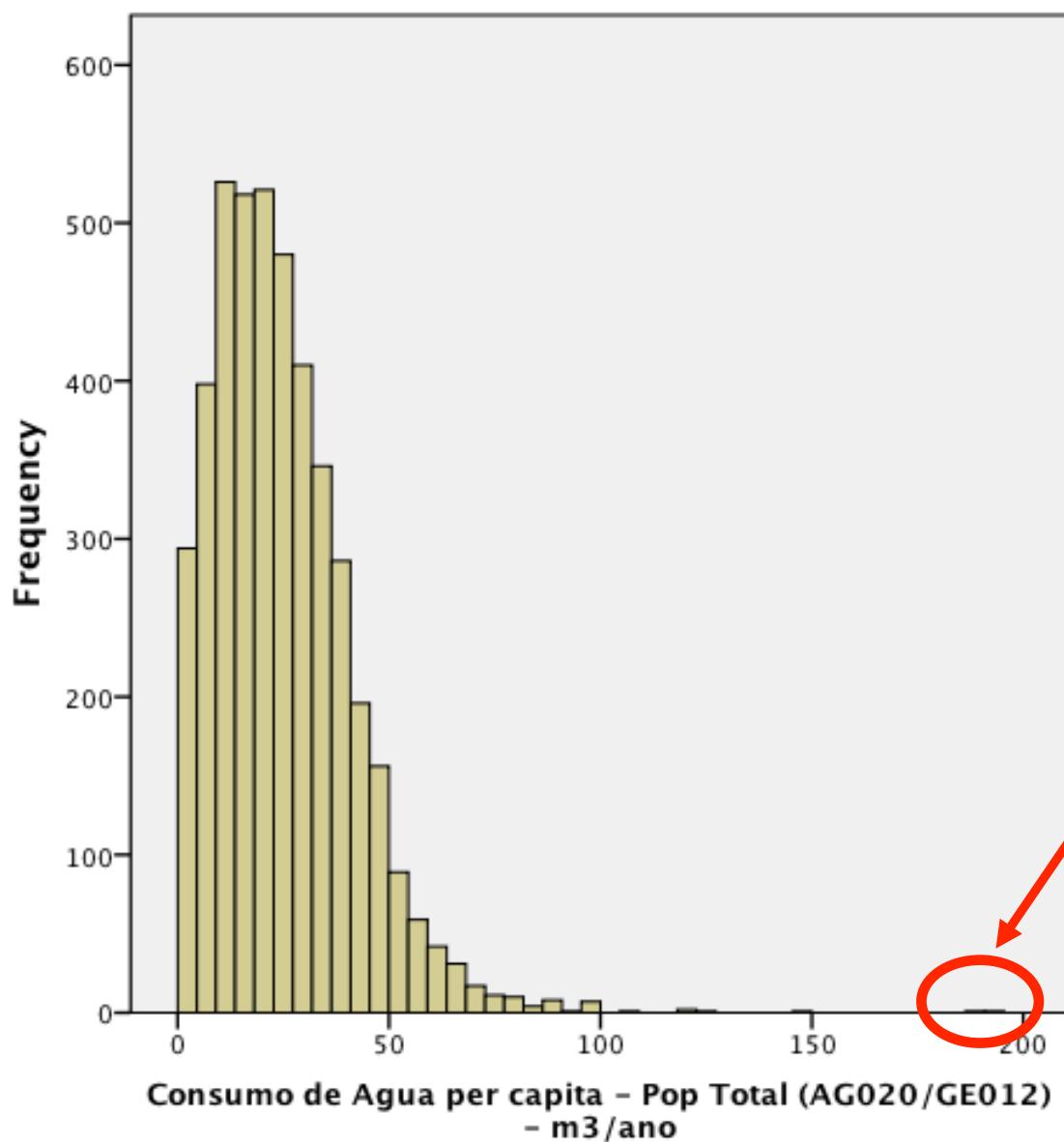
Converter assimetria e curtose em escores-z

$$Z_{Assimetria} = \frac{S - 0}{EP_{Assimetria}}$$

$$Z_{Curtose} = \frac{S - 0}{EP_{Curtose}}$$



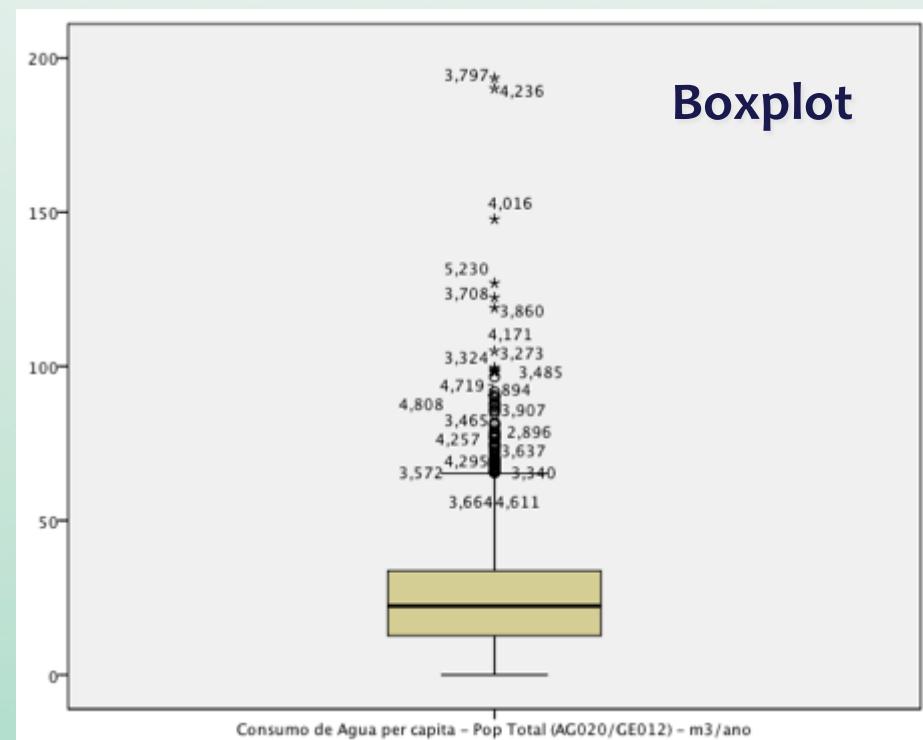
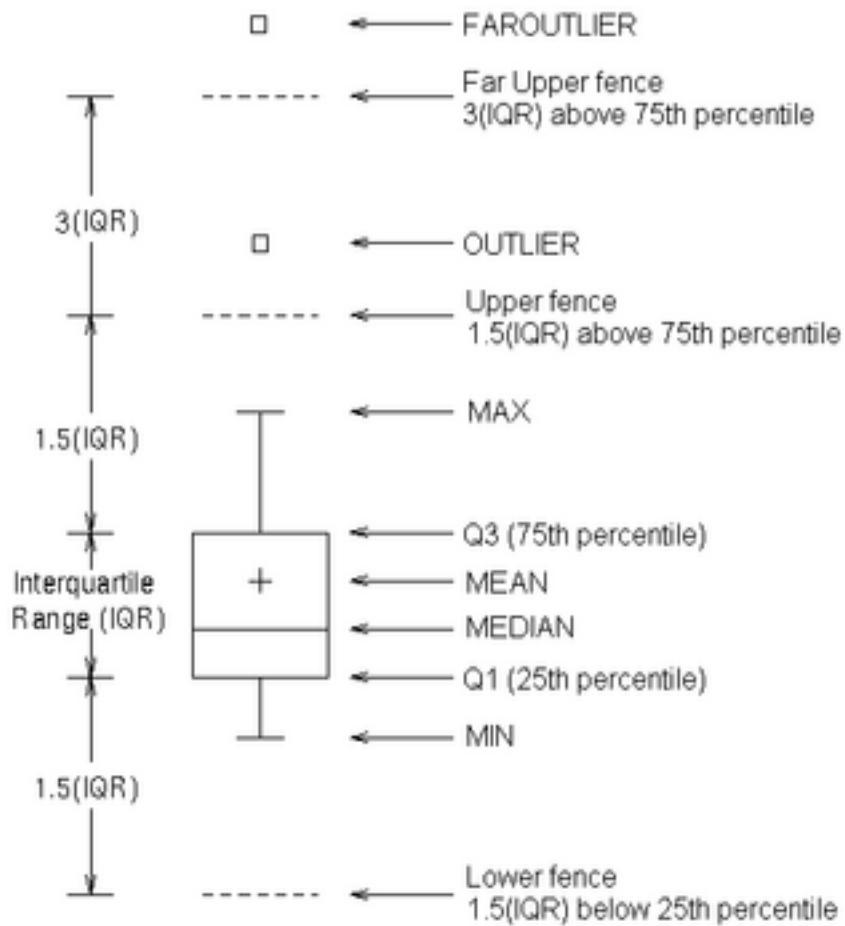
Histograma - Outliers



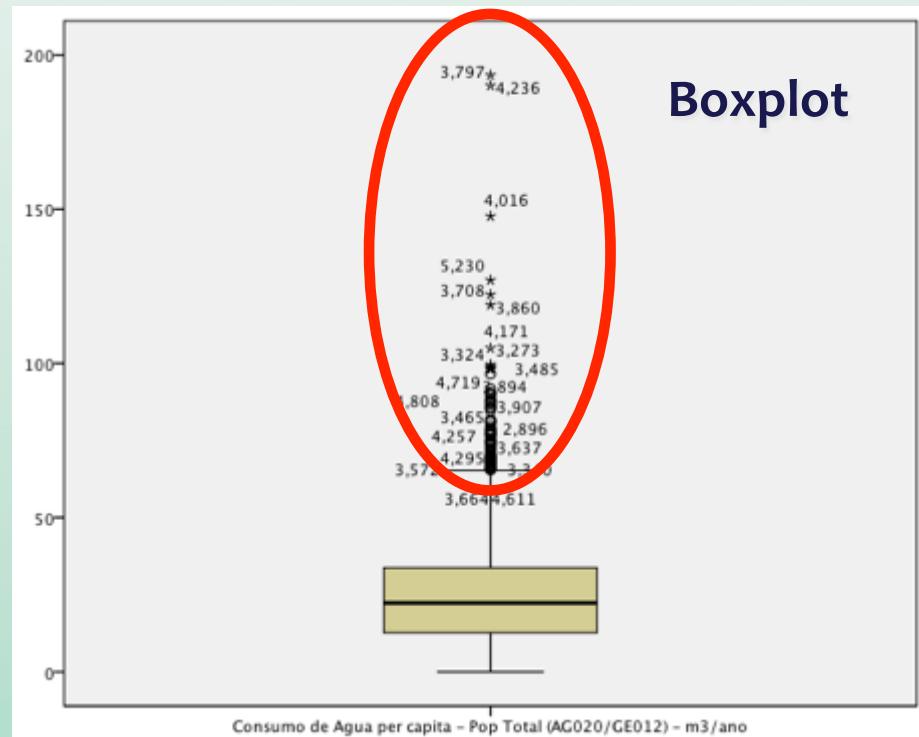
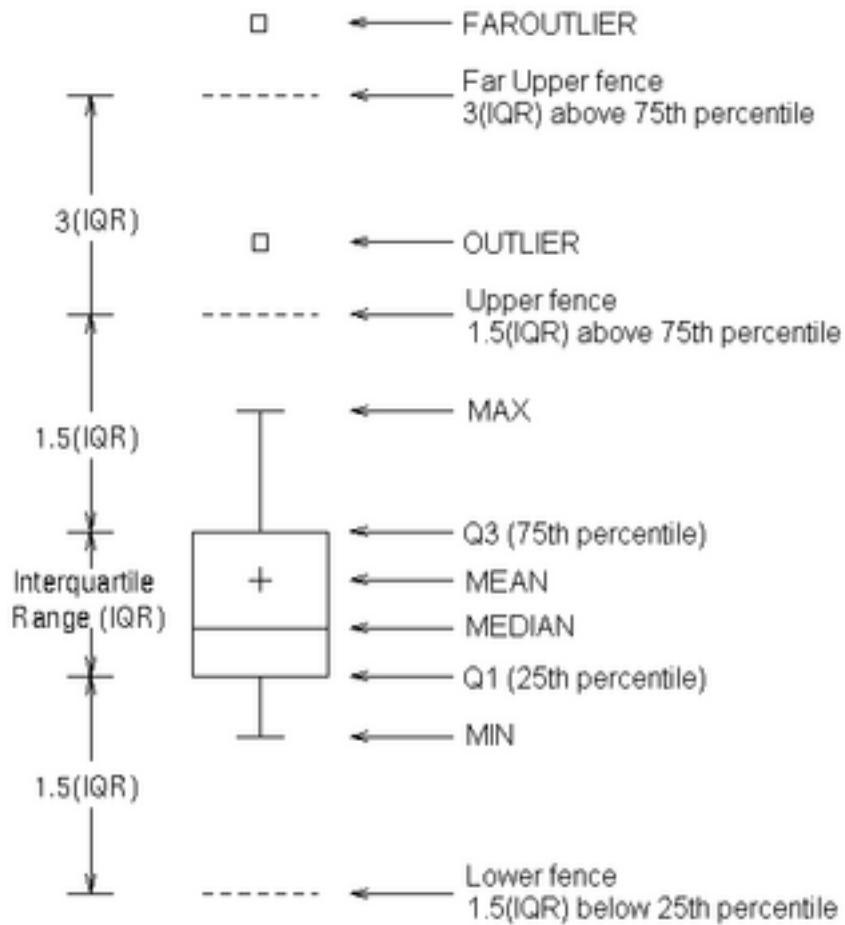
Importante para a detecção de erros, valores atípicos e observação da forma da distribuição dos dados

Valores atípicos (outliers)
Distorcem a média e inflacionam o desvio padrão

Explorando Dados – Box Plot

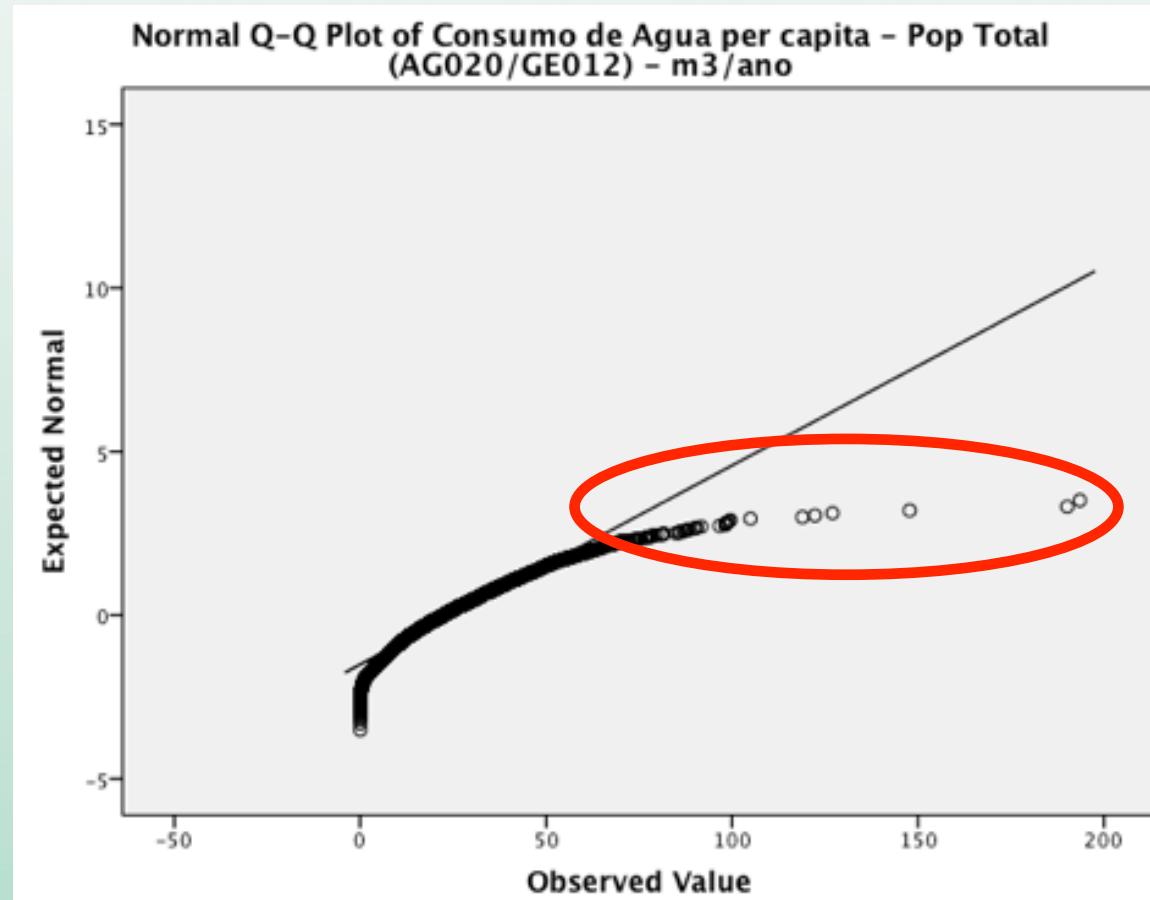


Explorando Dados – Box Plot



Explorando Dados – Q-Q Plot (quantil-quantil plot)

Valores esperados
caso a variável
tenha uma
distribuição
normal



Valores Observados na Amostra

Explorando Dados por Regiões

■ Graphs > Histogram

Sat 9:44 AM Flavia

Graphs Utilities Add-ons Window Help

Legacy Dialogs

IAO	CENTRX	CENTRY	RENDAPITA	GINI	PIB
N	-62.3893	-13.0113	467.72	.59	167212.4
N	-62.9573	-9.9519	672.87	.55	613103.3
N	-60.6398	-13.4749	446.58	.52	41686.8
N	-61.3247	-11.3013	718.79	.59	582804.4
N	-61.2609	-13.2036	553.47	.51	86885.2
N	-60.5502	-13.1592	507.70	.52	109115.0
N	-61.0916	-12.9266	401.70	.49	58447.8
N	-64.0588	-12.1473	352.67	.52	49736.8

Chart Builder...
Graphboard Template Chooser...

Bar...
3-D Bar...
Line...
Area...
Pie...
High-Low...
Boxplot...
Error Bar...
Population Pyramid...
Scatter/Dot...
Histogram...

Histogram

Variable: Consumo de Agua per capita - Pop T...
 Display normal curve

Panel by

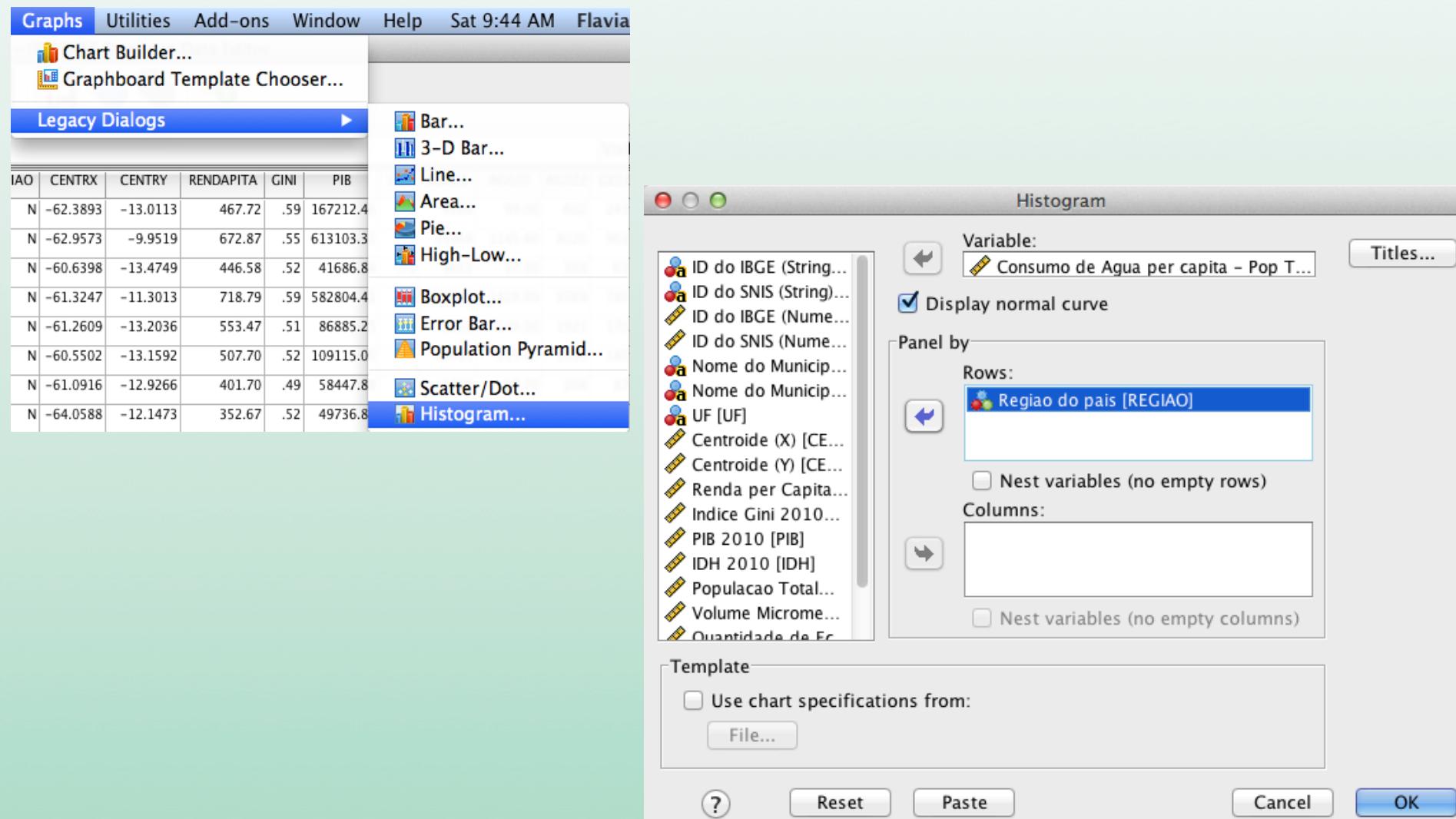
Rows: Regiao do pais [REGIAO]
 Nest variables (no empty rows)

Columns:
 Nest variables (no empty columns)

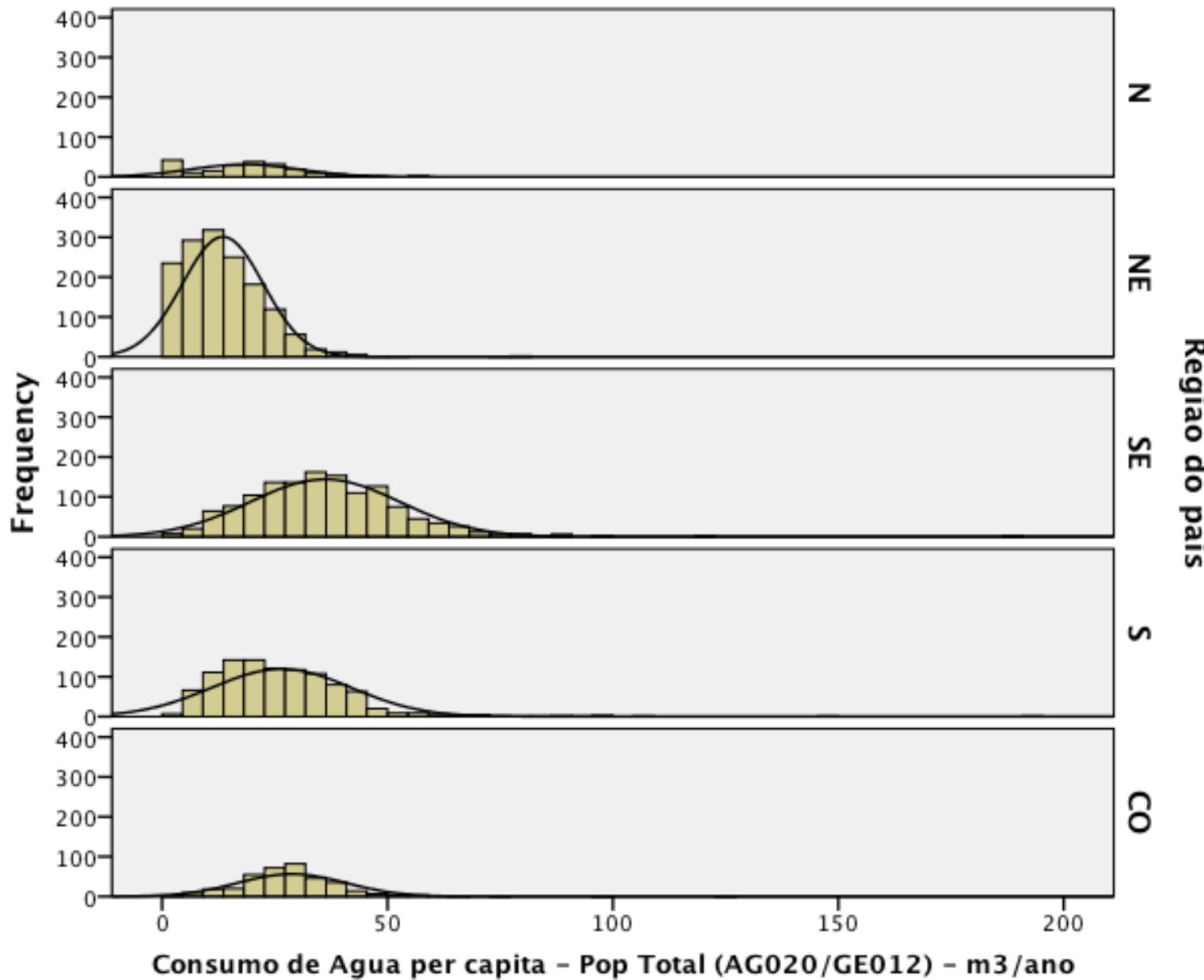
Template

Use chart specifications from:
File...

? Reset Paste Cancel OK

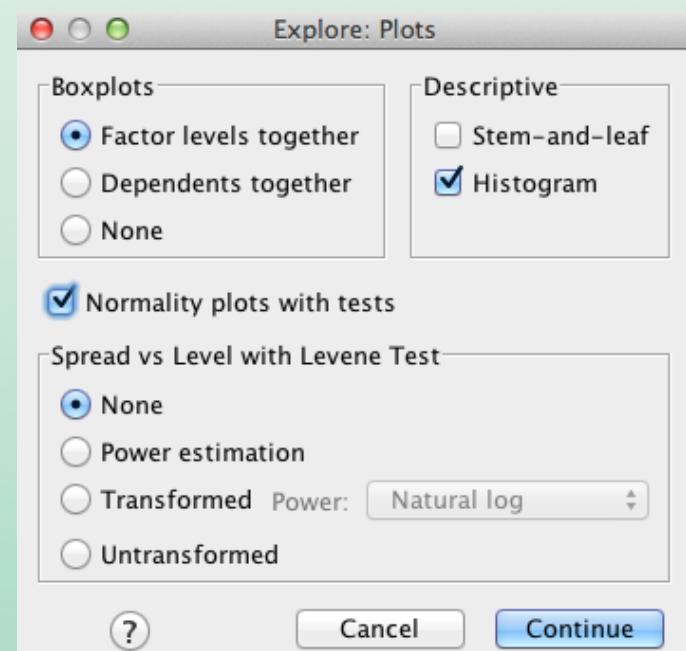
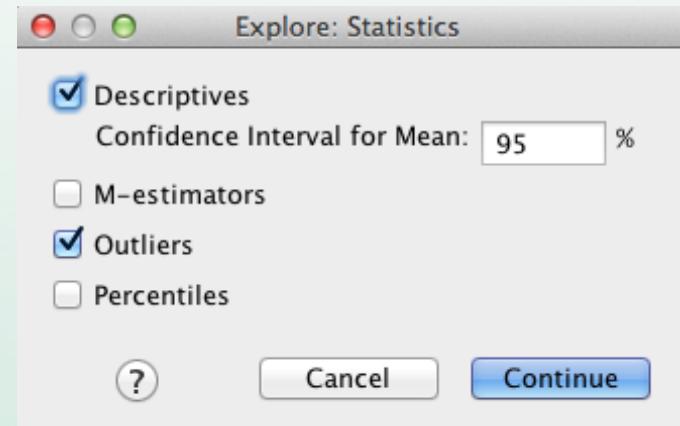
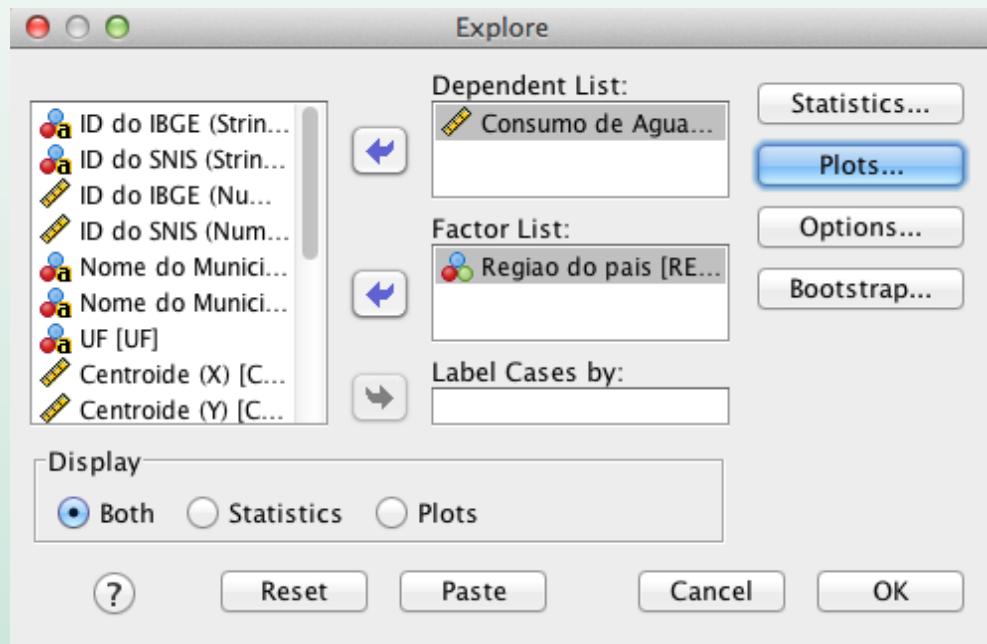


Histograma por Regiões

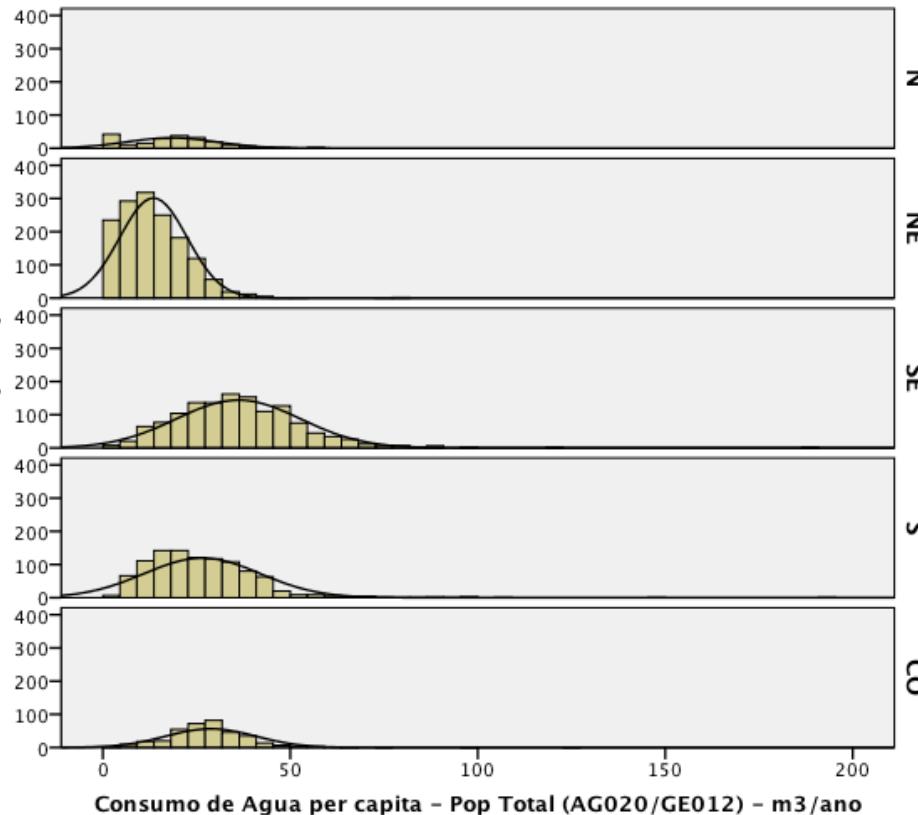


Estatísticas por Regiões

- Analyze > Descriptive > Explore

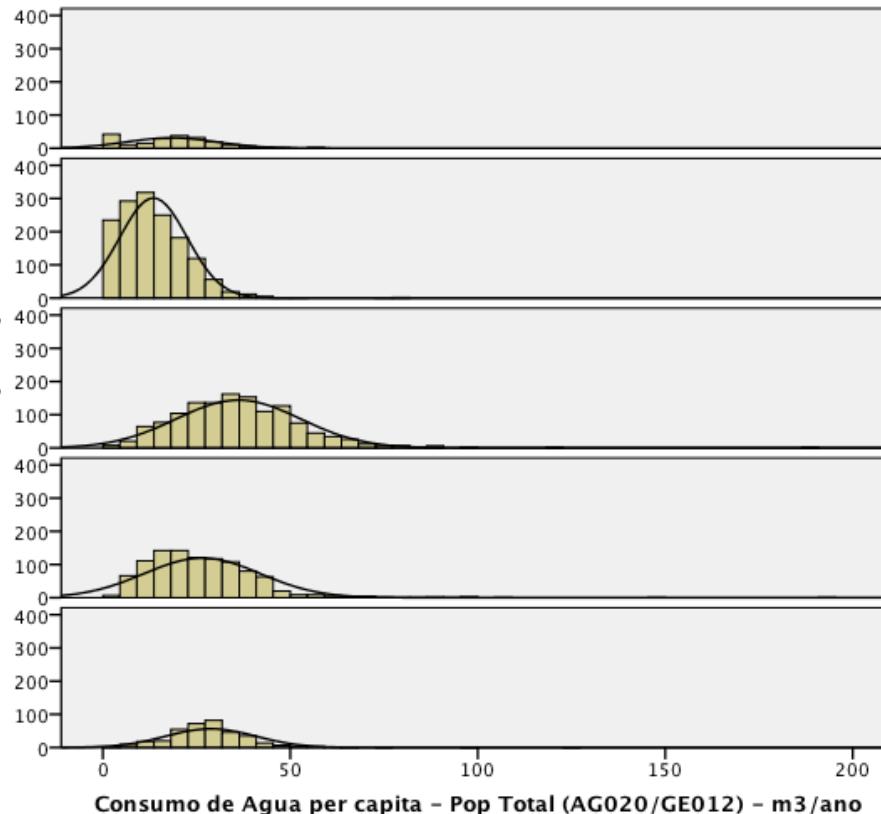


Frequency

N
NE
SE
S
CO

N	Mean	18.2785556	.836912047
	95% Confidence Interval for Mean	Lower Bound Upper Bound	16.6286843 19.9284268
	5% Trimmed Mean	17.7925417	
	Median	18.8827130	
	Variance	147.089	
	Std. Deviation	12.1280078	
	Minimum	.002217690	
	Maximum	57.2839930	
	Range	57.2817753	
	Interquartile Range	!!! 17.4335258	
	Skewness	$z=.294/.168= 1.75$.294 .168
	Kurtosis	$z=-.077/.334= 0.23$	-.077 .334
NE	Mean	13.4795386	.232355658
	95% Confidence Interval for Mean	Lower Bound Upper Bound	13.0237601 13.9353171
	5% Trimmed Mean	12.9620307	
	Median	12.3720860	
	Variance	80.606	
	Std. Deviation	8.97807347	
	Minimum	.000502008	
	Maximum	81.8104110	
	Range	81.8099090	
	Interquartile Range	11.3442080	
	Skewness	$Z=1.339/.063=21.25$	1.339 .063
	Kurtosis	$z=-5.378/.127=42.35$	5.378 .127
SE	Mean	36.1390013	.457337247
	95% Confidence Interval for Mean	Lower Bound Upper Bound	35.2418129 37.0361898
	5% Trimmed Mean	35.4049234	
	Median	34.9633810	
	Variance	275.879	
	Std. Deviation	16.6095922	
	Minimum	.244042000	
	Maximum	190.120100	
	Range	189.876058	
	Interquartile Range	21.7176040	
	Skewness	$Z=1.225/.067=18.3$	1.225 .067
	Kurtosis	$z=-6.499/.135=48.1$	6.499 .135

Frequency



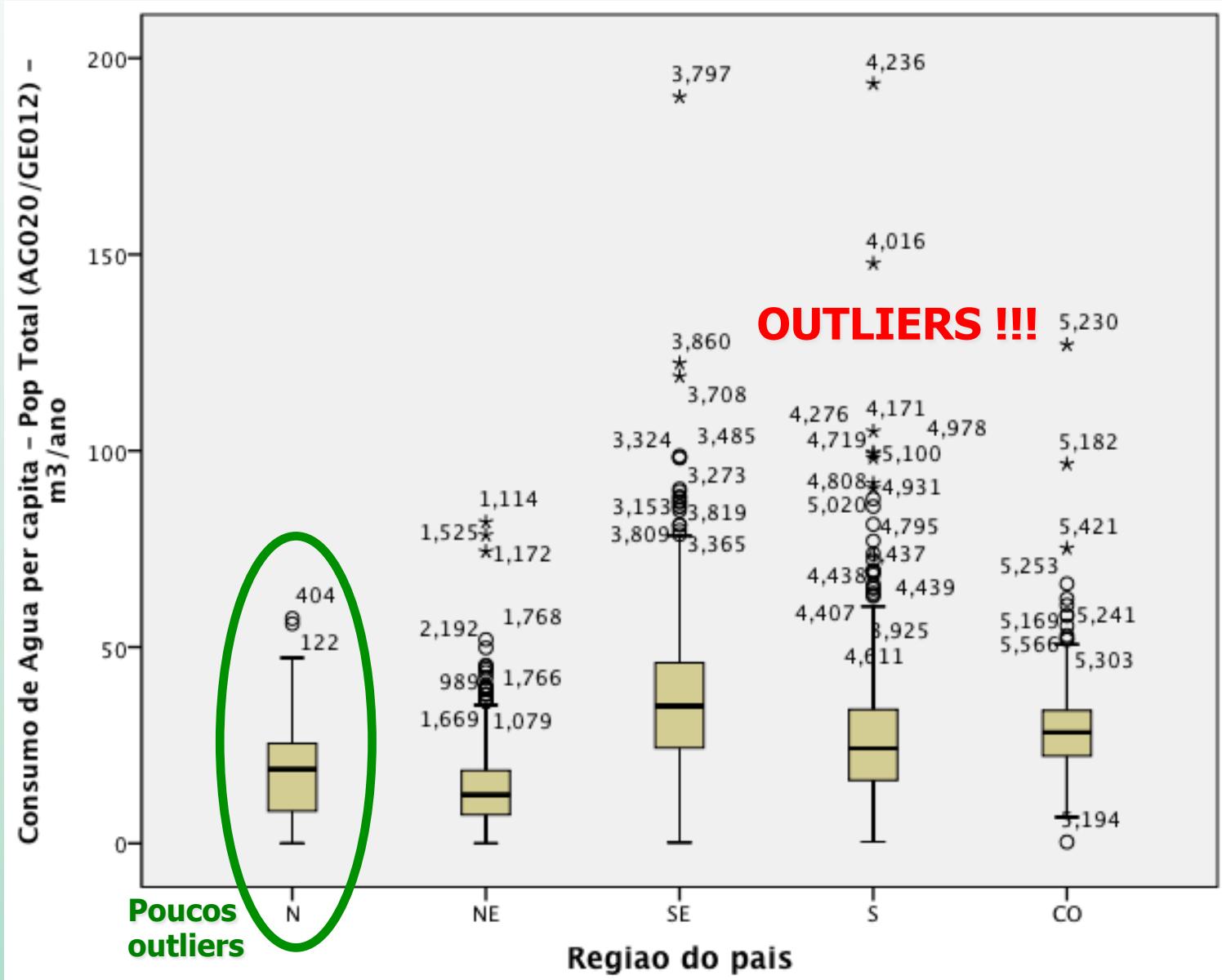
Região do país

S	Mean	26.4722730	.485837396	
	95% Confidence Interval for Mean	Lower Bound Upper Bound	25.5189202 27.4256259	
	5% Trimmed Mean	25.2431869		
	Median	24.2355860		
	Variance	241.467		
	Std. Deviation	15.5392036		
	Minimum	.419118000		
	Maximum	193.493151		
	Range	193.074033		
	Interquartile Range	18.0550220		
	Skewness	$z=2.671/.076=35.1$	2.671	.076
	Kurtosis	$z=18.517/.153=121$	18.517	.153

CO	Mean	28.7426507	.620833392	
	95% Confidence Interval for Mean	Lower Bound Upper Bound	27.5218571 29.9634444	
	5% Trimmed Mean	28.0669352		
	Median	28.2475860		
	Variance	143.381		
	Std. Deviation	11.9742008		
	Minimum	.345988000		
	Maximum	126.990144		
	Range	126.644156		
	Interquartile Range	11.5080482		
	Skewness	$z=2.315/.126=18.4$	2.315	.126
	Kurtosis	$z=14.757/.252=58.6$	14.757	.252

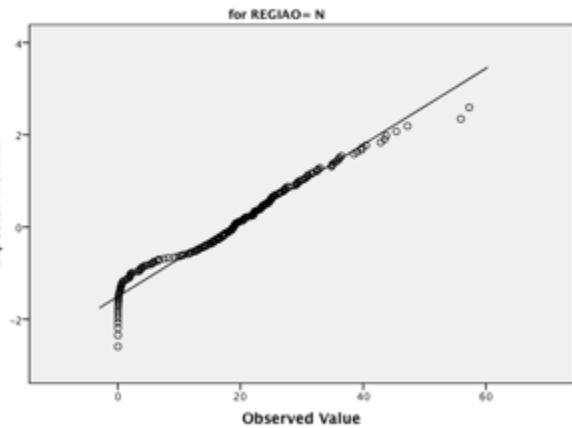
No histograma parece simétrica, mas nas estatísticas não. Pq?

Boxplot (Caixa e Bigodes)

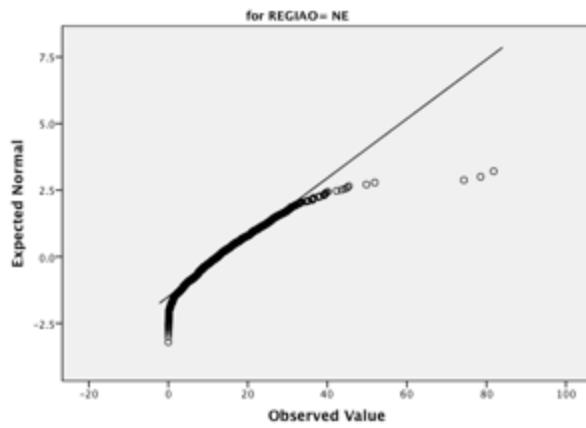


Q-Q Plots

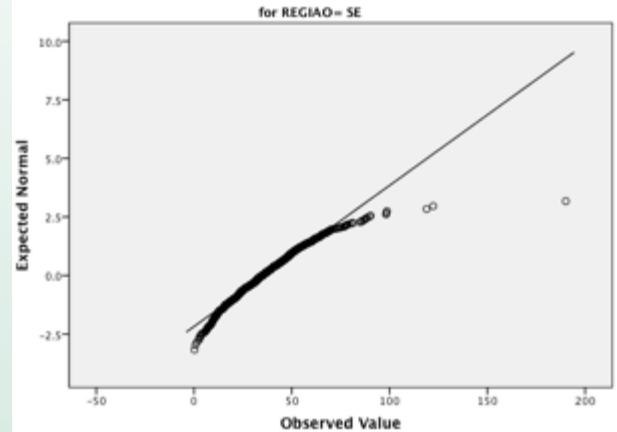
Normal Q-Q Plot of Consumo de Agua per capita - Pop Total
(AG020/GE012) - m³/ano



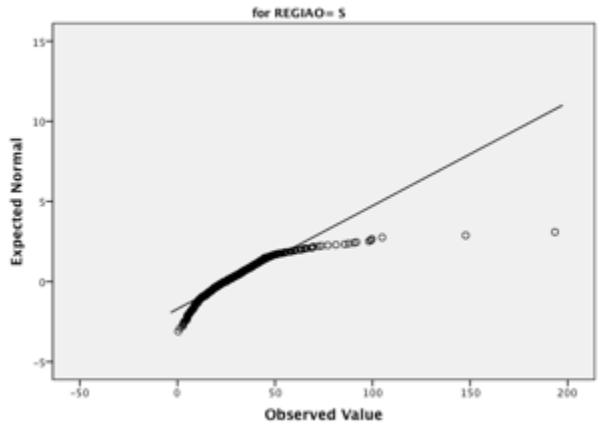
Normal Q-Q Plot of Consumo de Agua per capita - Pop Total
(AG020/GE012) - m³/ano



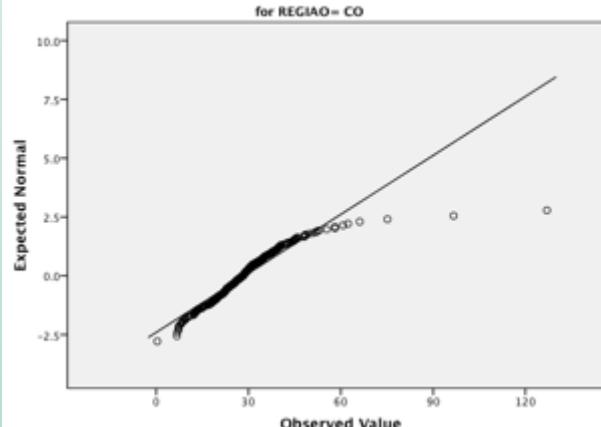
Normal Q-Q Plot of Consumo de Agua per capita - Pop Total
(AG020/GE012) - m³/ano



Normal Q-Q Plot of Consumo de Agua per capita - Pop Total
(AG020/GE012) - m³/ano



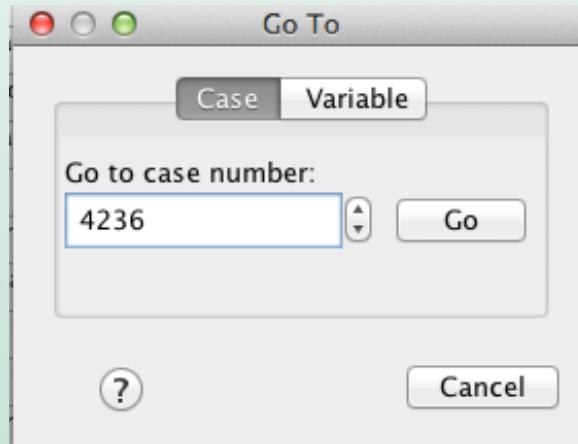
Normal Q-Q Plot of Consumo de Agua per capita - Pop Total
(AG020/GE012) - m³/ano



CORRIGINDO PROBLEMAS NOS DADOS

Observou algum erro de digitação que gostaria de corrigir?

- Vá para o editor de dados, em “go to case”



- Edite o valor desejado

Reduzindo o Impacto de Outliers

1. Remover o caso

Só deve ser feito se tiver uma boa razão para acreditar que esse valor não é representante da população.

2. Transformar os dados

Deverá ser feito no caso de termos uma distribuição não normal. Costumam reduzir o impacto de outliers.

3. Substituir o valor

- O próximo escore mais alto adicionado de 1
- Inverter o valor do escore-z (adicionar o triplo do desvio padrão à média e substituir o valor atípico por esse)
- A média mais dois desvios padrão (variação do método acima)

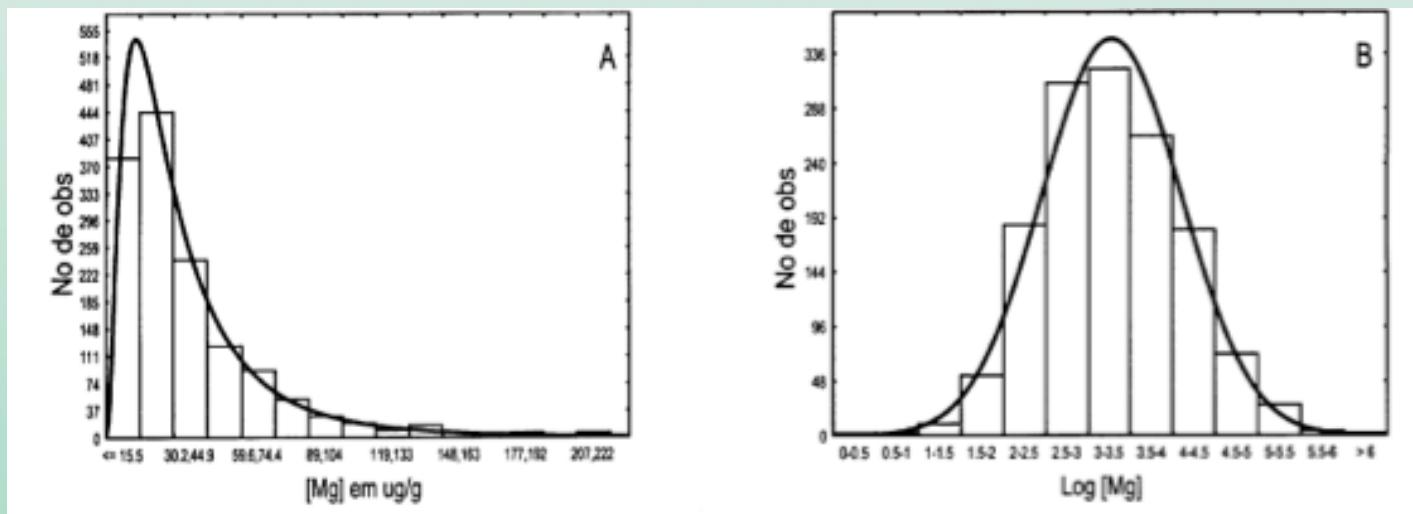
Transformação dos Dados

Para corrigir problemas relacionados à não-normalidade da distribuição ou valores atípicos (outliers)

a. Transformação logaritmica ($\log(X_i)$)

Tomar o logaritmo de um conjunto de números “esmaga” a cauda direita da distribuição. É uma boa maneira de reduzir uma assimetria positivo.

Atenção: Não podemos obter logaritmo de zero ou valores negativos. Se tiver zero nos dados, faça $\log(X_i+1)$



Transformação dos Dados

Para corrigir problemas relacionados à não-normalidade da distribuição ou valores atípicos (outliers)

b. Transformação por radiciação ($\text{sqr}(X_i)$)

Tomar a raiz quadrada de valores grandes tem efeito maior do que extrair a raiz de efeitos pequenos. Útil para dados com assimetria positiva.

Problemas com números negativos.

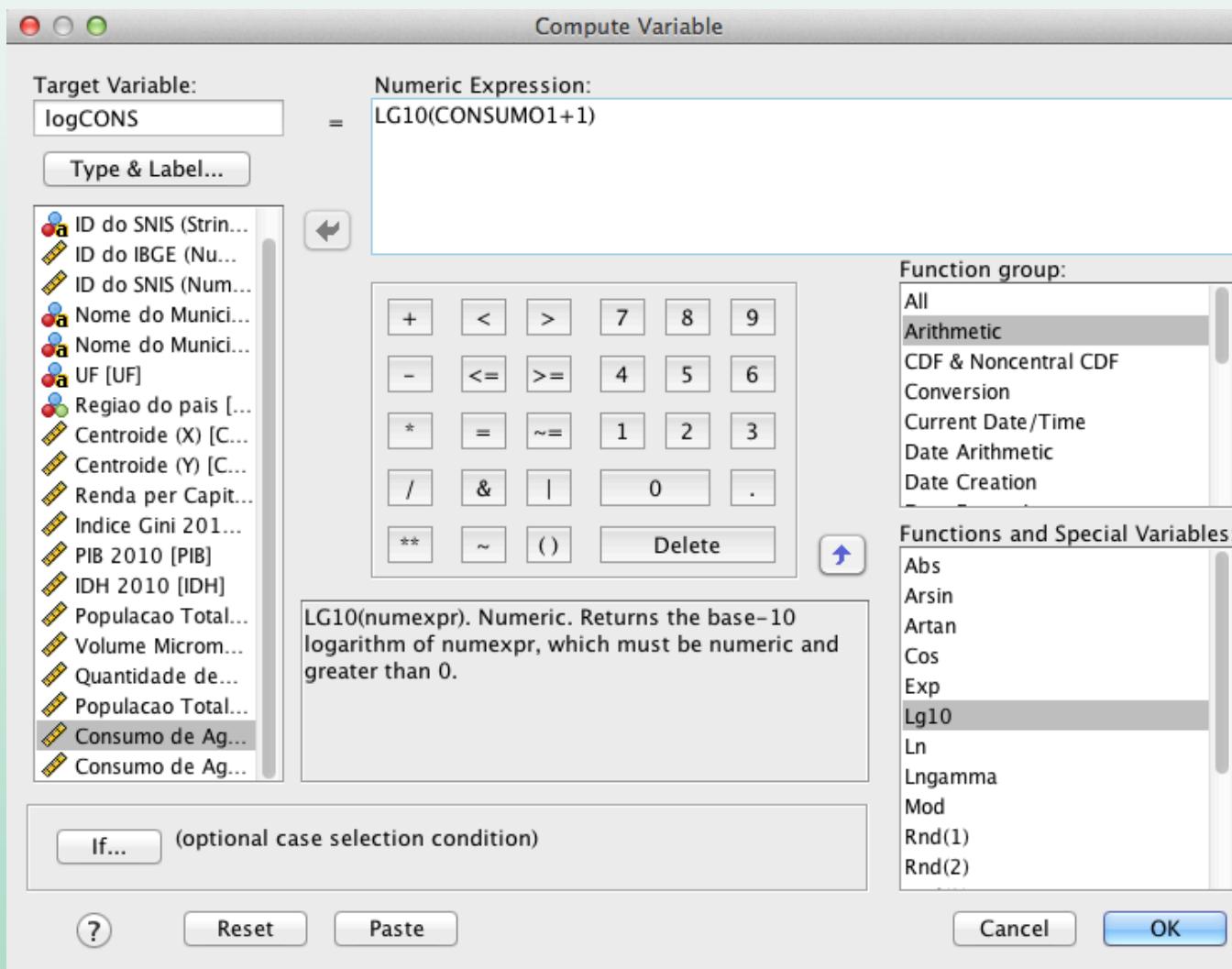
c. Transformação recíproca ($1/X_i$)

Dividir 1 por cada escore reduz o impacto dos grandes valores. A variável transformada terá um limite inferior de zero (grandes valores ficarão próximos de zero).

Atenção: Este tipo de transformação reverte os escores (valores grandes se tornarão pequenos e vice-versa)

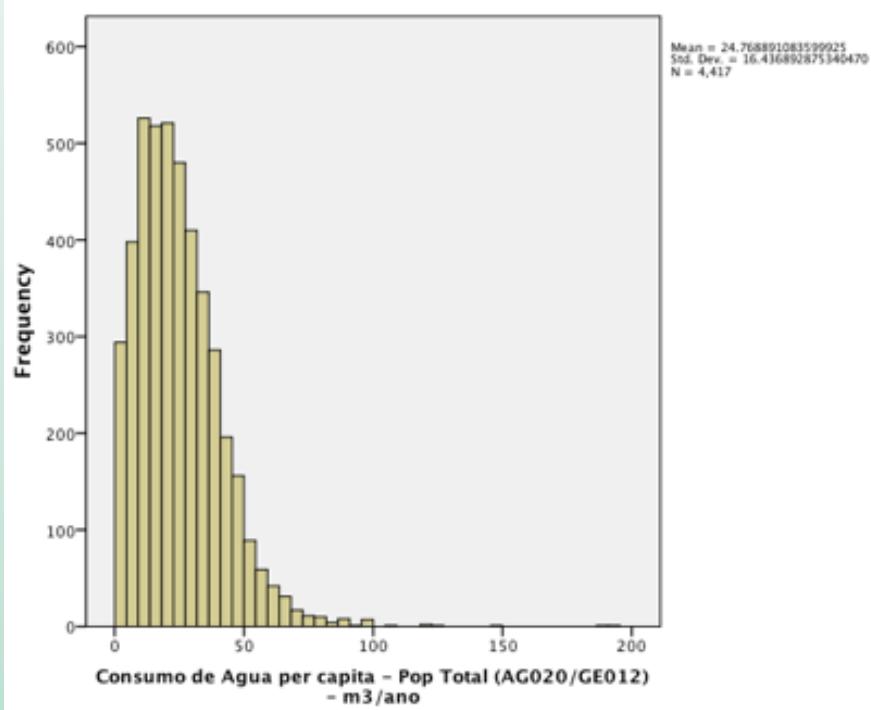
Transformando Dados no SPSS

Transform > Compute

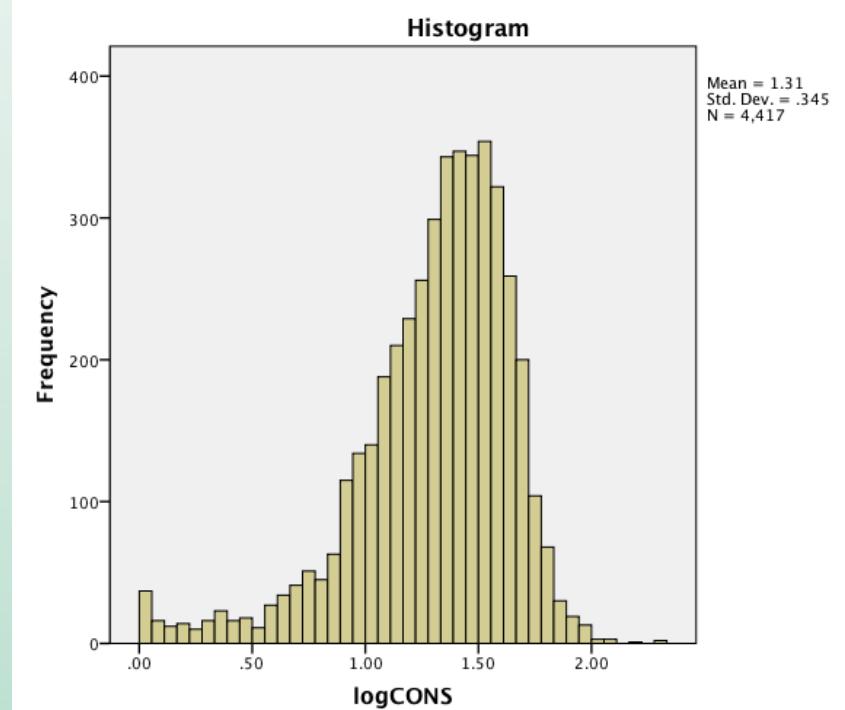


Transformando Dados no SPSS

CONSUMO



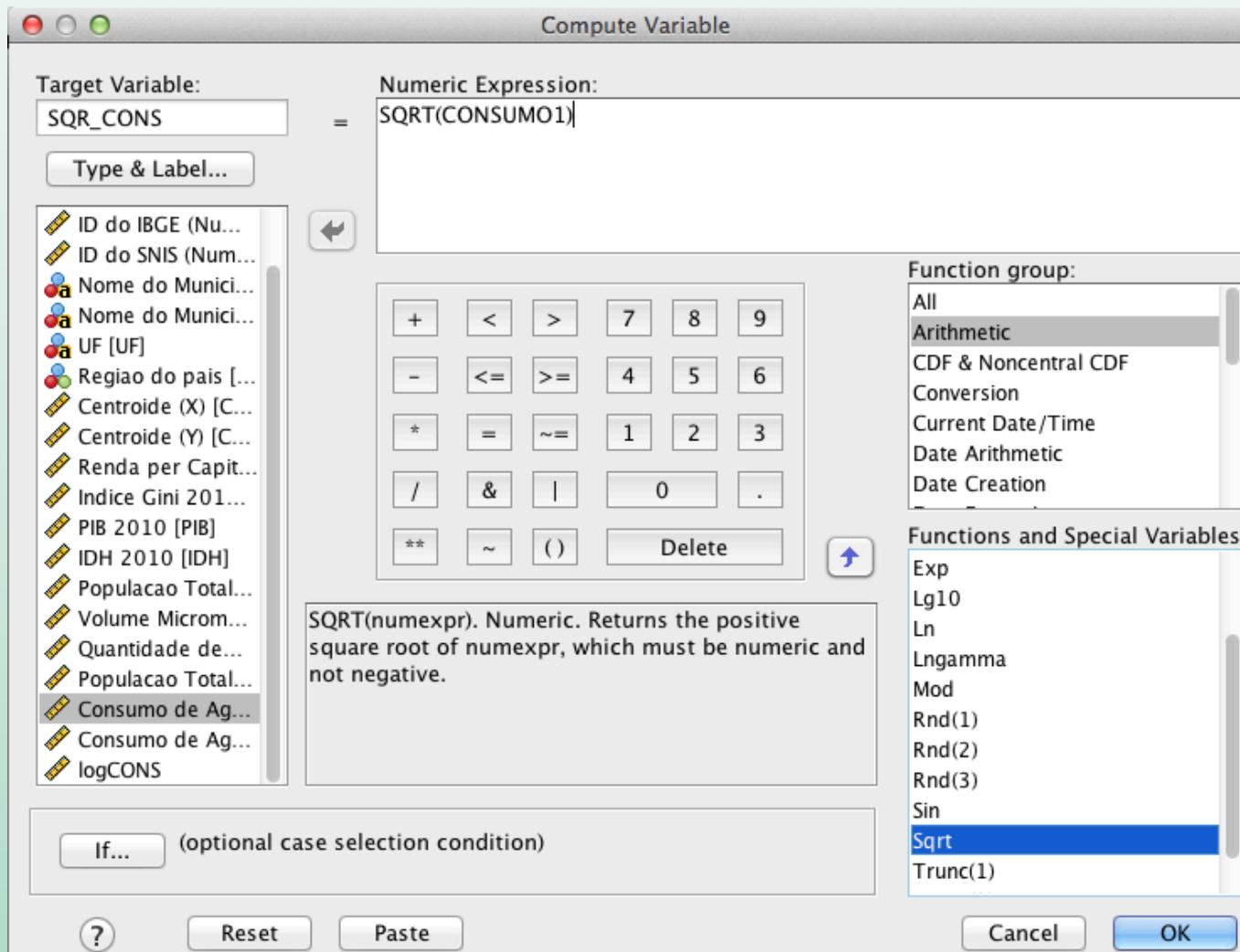
LOG(CONSUMO + 1)



Skewness	-1.195	.037
Kurtosis	2.058	.074

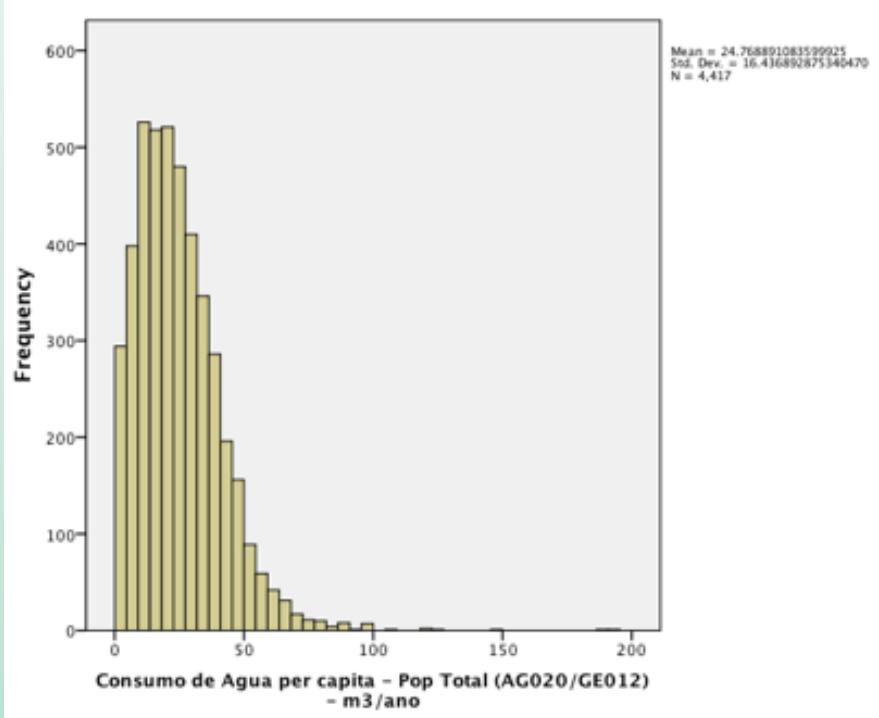
Transformando Dados no SPSS

Transform > Compute

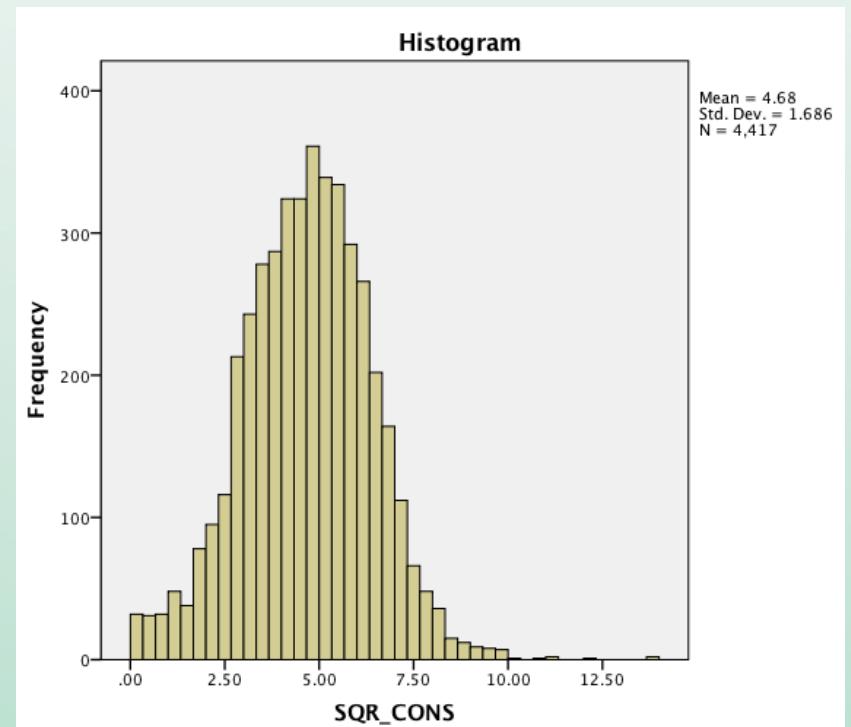


Transformando Dados no SPSS

CONSUMO



SQR(CONSUMO)



Skewness	.007	.037
Kurtosis	.528	.074

Transformando Dados no SPSS

Vocês podem usar o comando “Transform > Compute” para realizar as mais diversas transformações nos dados!

Por exemplo: Normalizar os dados, calcular taxas e proporções, etc.

PARTE II

**Realizar análises exploratórias sobre
os dados do trabalho do curso!!!**

Para importar um arquivo .csv ou .dbf para o SPSS, vá em:

File > Read Text Data...

DICA: Vocês podem importar o arquivo .dbf que compõe o arquivo vetorial (shapefile)

Atividade

A ser entregue no dia 30/06 (Pode ser realizado em grupo)

Utilizando os dados que pretende usar no trabalho final da disciplina:

- a. Calcule as estatísticas descritivas de uma variável de sua escolha. Explique cada uma delas.
- b. Apresente o histograma, box-plot e Q-Q Plot. Explique.
- c. Realize alguma(s) transformação(ões) em uma ou mais variáveis selecionadas (ex: log, raiz quadrada...). Explique.
- d. Se houver grupos distintos (bairros, distritos), repita os itens (a) e (b) para cada grupo. Compare intervalos de confiança da média de uma variável de interesse. Há sobreposições? O que isso significa? Interprete!

CAPRICHE NAS ANÁLISES!!!

Já estamos elaborando o trabalho final da disciplina!!!