**ANÁLISE DESCRITIVA E EXPLORATÓRIA DE DADOS COM O *INSTAT***

**Algumas dicas**

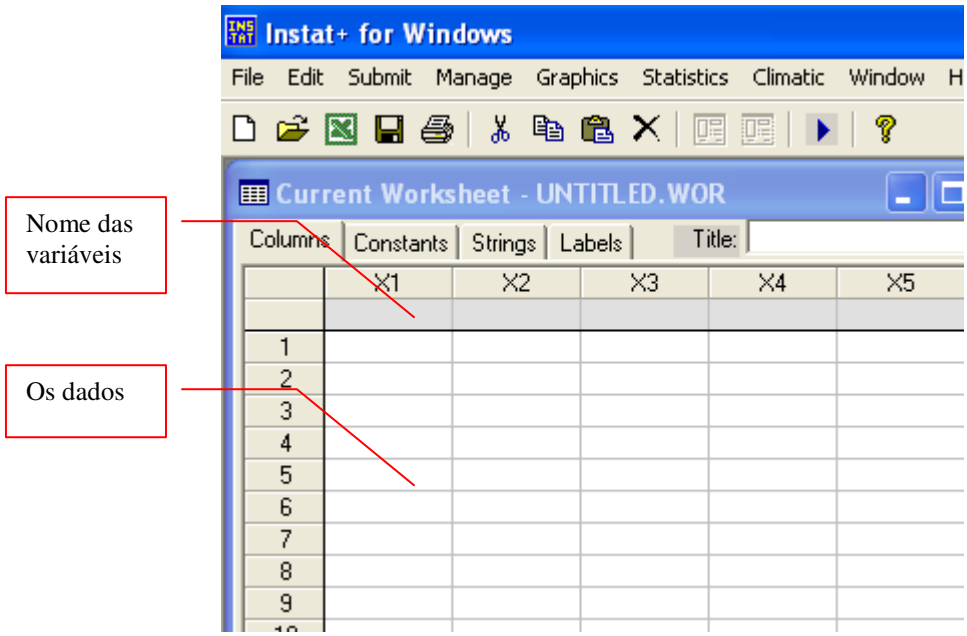
Pedro Alberto Barbetta

O INSTAT é um software para análise estatística da University of Reading, no Reino Unido e é gratuito para uso acadêmico, e pode ser baixado de: http://www.rdg.ac.uk/ssc/software/download.html. Essas notas baseiam-se na versão do Instat 3.36.

## 1 - CRIAÇÃO DE UM ARQUIVO DE DADOS

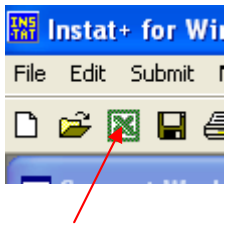
Na forma usual, ao abrir o INSTAT, aparece uma janela, onde no lado esquerdo tem uma planilha, e do lado direito um espaço para texto, onde aparecerão os resultados de suas análises.

Na planilha, colocamos os nossos dados. No cabeçalho (linha mais escura e não numerada) escrevemos o nome das variáveis de interesse e nas linhas abaixo os dados.



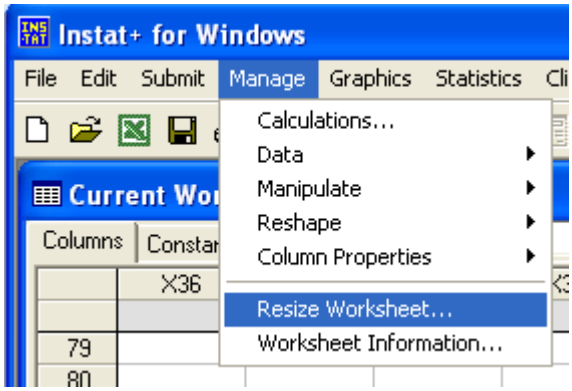

Se os dados estão em planilhas do Excel, podemos copiar (CTRL – C) e colar (CTRL – V) no Instat. Se o arquivo do Excel já tem o nome das variáveis, você deverá deixar o cursor no cabeçalho da planilha do Instat para que os nomes fiquem no lugar correto.

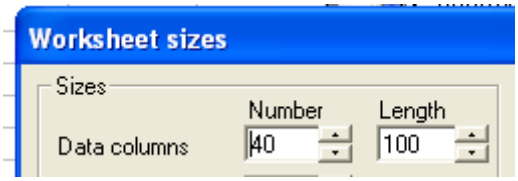
Opcionalmente, você pode importar dados de alguns outros softwares. Em especial, do Excel pode ser feito pelo atalho:




**TAMANHO DA PLANILHA**

A planilha do Instat está desenhada para 40 variáveis (colunas) e 100 observações ou casos (linhas). Se seu arquivo for maior, você deve ir em *Manage, Resize Worksheet* e colocar o tamanho desejado (o software <u>não</u> amplia a planilha automaticamente, ao digitar ou colar os dados):




**EXEMPLO**

Como exemplo, vamos importar os dados do arquivo SacoGrande.xls, que pode ser obtido em:

http://www.inf.ufsc.br/~barbetta/livro1/Livro1_Dados.zip

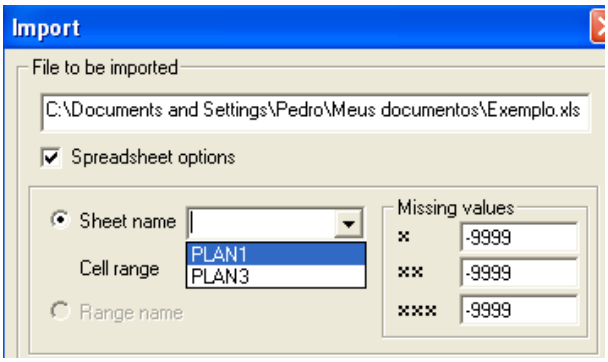
Esses dados referem-se a uma amostra de 120 famílias de um bairro de Florianópolis, em que foram coletadas as variáveis:

- local (local de residência: 1 = Monte Verde, 2 = Parque da Figueira, 3= encosta do morro);
- p.a.p (uso de programa de alimentação popular: 1 = sim, 2 = não);
- instr. (grau de instrução do chefe da casa: 1 = nenhum, 2 = fundamental, 3 = médio ou superior);
- tam. (número de moradores no domicílio);
- renda (renda familiar em quantidade de salários mínimos).

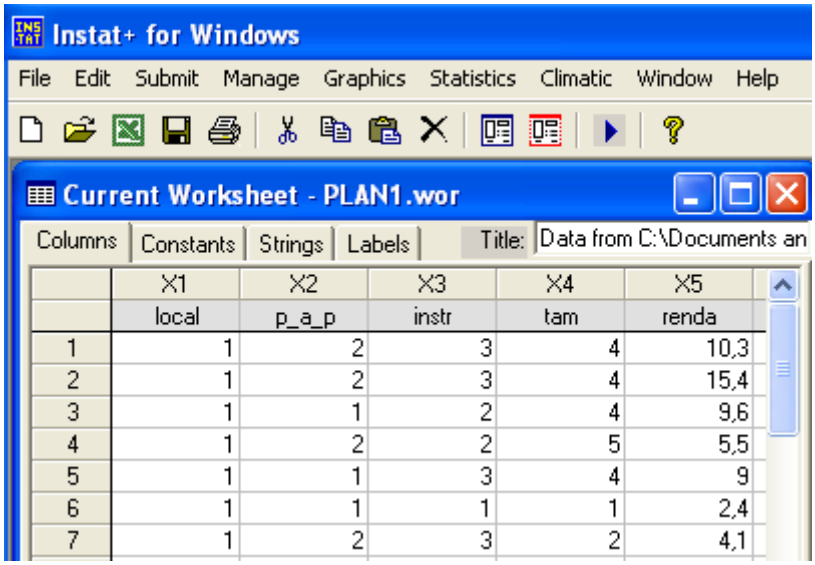
Observar que as três primeiras variáveis são qualitativas (categóricas). É conveniente identificar as categorias com códigos formados por números naturais: 1, 2, 3, ... No arquivo original (SacoGrande.xls), o código de "não" em "p.a.p" era 0 (zero). Embora o arquivo pudesse ser importado assim, o uso de rótulos naturais

facilita a colocação de rótulos, como veremos. Por isto, alteramos o código de "0" para "2".

Ao pedir a importação dos dados, é necessário informar a planilha do arquivo Excel em que estão os dados:



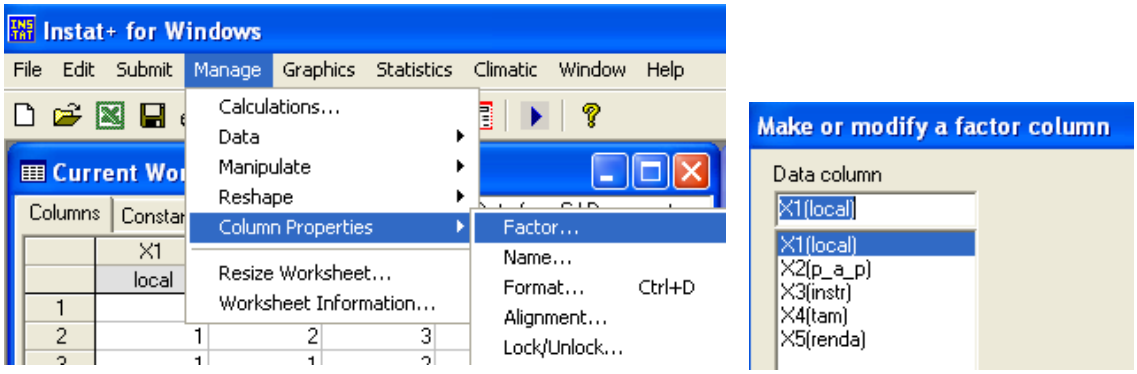

O arquivo de dados do INSTAT resultante (PLAN1.wor):



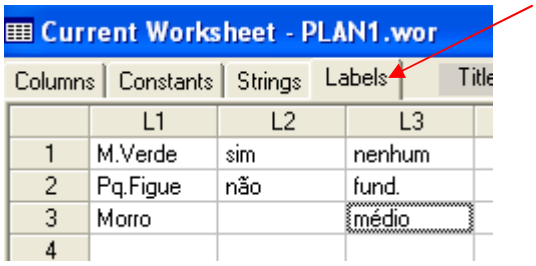

| | X1 | X2 | X3 | X4 | X5 |
|---|---|---|---|---|---|
| | local | p_a_p | instr | tam | renda |
| 1 | 1 | 2 | 3 | 4 | 10,3 |
| 2 | 1 | 2 | 3 | 4 | 15,4 |
| 3 | 1 | 1 | 2 | 4 | 9,6 |
| 4 | 1 | 2 | 2 | 5 | 5,5 |
| 5 | 1 | 1 | 3 | 4 | 9 |
| 6 | 1 | 1 | 1 | 1 | 2,4 |
| 7 | 1 | 2 | 3 | 2 | 4,1 |

**VARIÁVEIS QUALITATIVAS (CATEGÓRICAS)**

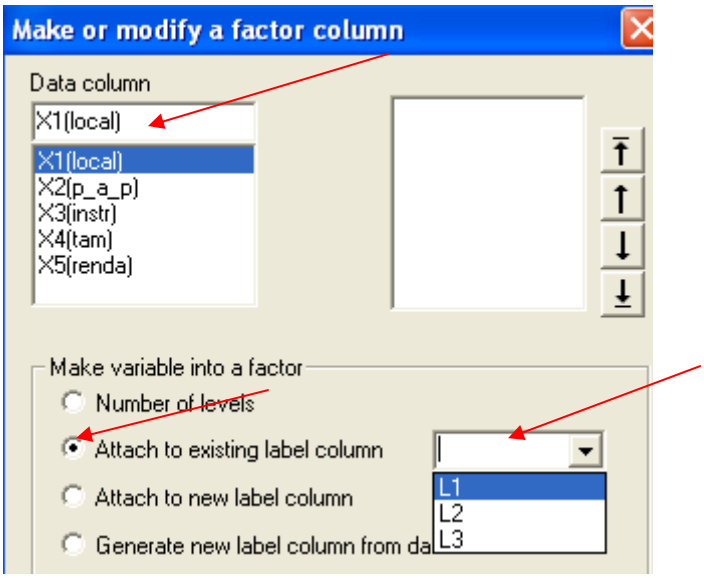
No INSTAT as variáveis qualitativas são denominadas de FACTOR (fator). Como padrão o software considera as variáveis como quantitativas, então precisamos identificar as qualitativas em *Manage*, *Column Properties*, *Factor*:



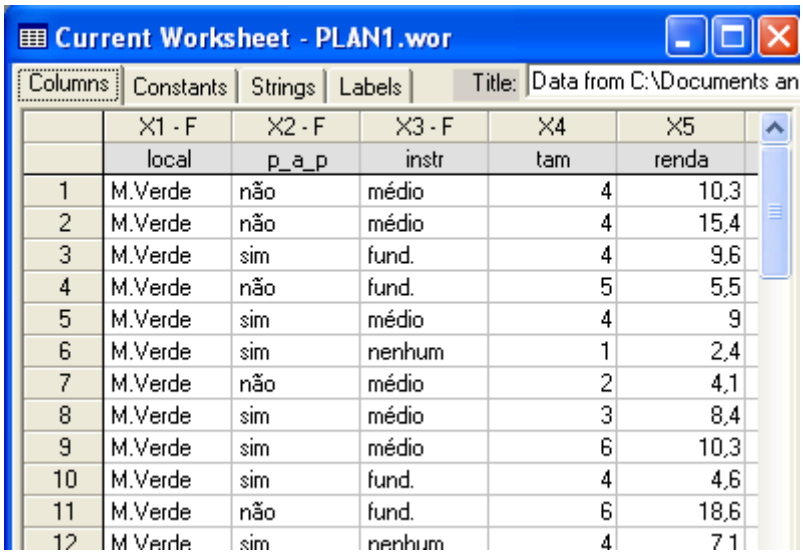

Opcionalmente, podemos identificar os rótulos (labels) associados a cada código numérico. Para isso, usamos a aba *Labels*:




A ordem dos rótulos deve corresponder à ordem dos números naturais dos códigos de cada variável. Ao passar cada variável para *Factor*, podemos identificar o seu conjunto de *labels*. Por exemplo, para "local" identificamos "L1", abaixo:




Repetindo o processo para "p.a.p" e "instr.", temos a planilha [note que X1, X2 e X3 estão identificadas com um "F" (de Factor)]:



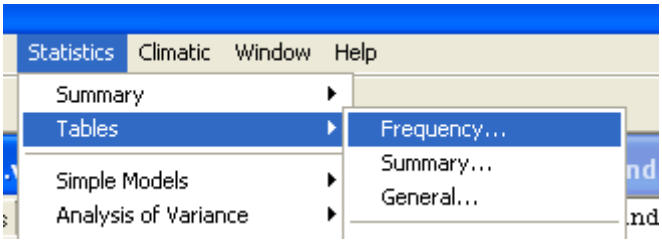

| | X1 - F | X2 - F | X3 - F | X4 | X5 |
|---|---|---|---|---|---|
| | local | p_a_p | instr | tam | renda |
| 1 | M.Verde | não | médio | 4 | 10,3 |
| 2 | M.Verde | não | médio | 4 | 15,4 |
| 3 | M.Verde | sim | fund. | 4 | 9,6 |
| 4 | M.Verde | não | fund. | 5 | 5,5 |
| 5 | M.Verde | sim | médio | 4 | 9 |
| 6 | M.Verde | sim | nenhum | 1 | 2,4 |
| 7 | M.Verde | não | médio | 2 | 4,1 |
| 8 | M.Verde | sim | médio | 3 | 8,4 |
| 9 | M.Verde | sim | médio | 6 | 10,3 |
| 10 | M.Verde | sim | fund. | 4 | 4,6 |
| 11 | M.Verde | não | fund. | 6 | 18,6 |
| 12 | M.Verde | sim | nenhum | 4 | 7,1 |

Outras manipulações, como cálculo de valores transformados por alguma função matemática, recodificação, criação de variáveis indicadoras etc., podem ser feitas em *Manage* (menu principal).

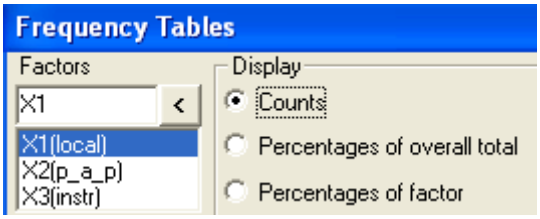
## 2 - DISTRIBUIÇÃO DE FREQÜÊNCIAS DE VARIÁVEIS QUALITATIVAS – CLASSIFICAÇÃO SIMPLES

Tendo o arquivo de dados, podemos partir para a realização de análises descritivas e exploratórias. Para se ter uma distribuição de freqüências de uma variável qualitativa (ou mesmo uma quantitativa discreta), podemos fazer uma tabela ou gráfico da distribuição de freqüências

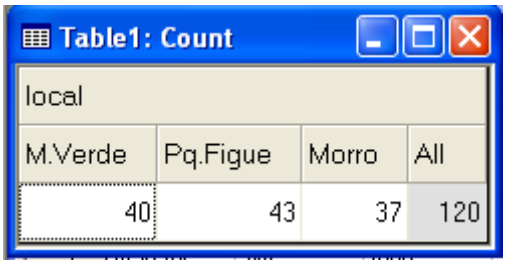
Para construir tabela de freqüências (contagem das categorias):



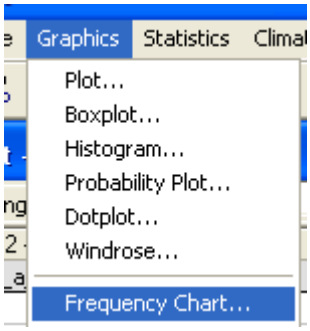

Exemplificando com a variável "local":




Resultando:

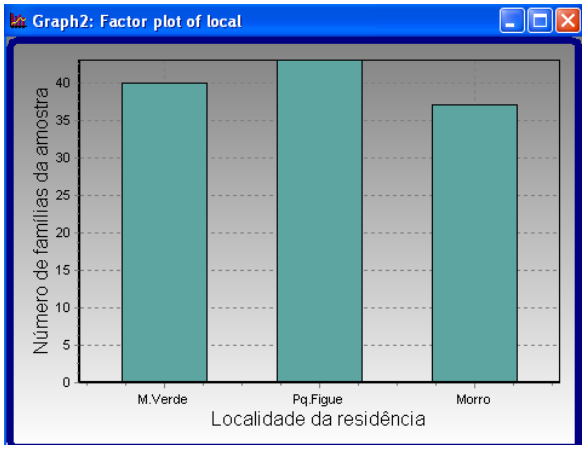



Note que a tabela apresenta as contagens de cada categoria. Se houver interesse em porcentagens, basta marcar uma das duas opções de *Percentages* na etapa anterior. Observe que a tabela de freqüências está em forma de planilha, o que permite copiar (CTRL-C) e colar (CTRL-V) numa planilha eletrônica (Excel, p. ex.) para se fazer os ajustes.




Alternativamente, os resultados poderiam ser apresentados num gráfico (de barras, colunas ou setores). Para isto, com a planilha de dados ativa, entrar em: *Graphics, Frequency chart,* e preencher os campos. No gráfico apresentado a seguir, escolhemos a variável "local", tipo *Bar* e *display number of cases*.

Quando o Instat apresenta o gráfico, aparece um novo menu, que permite editá-lo. Nós mudamos o título dos eixos (em *Chart, Axes*); e a cor e formato (em *Chart, Series*, aba *theme*, opção *Business*.

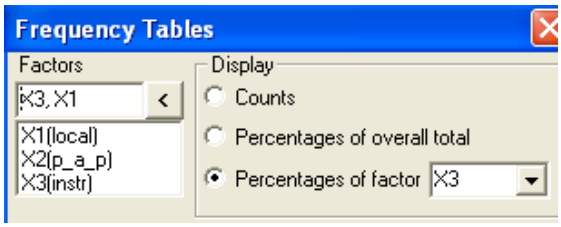



Os gráficos podem ser colocados num editor de texto como figuras. Com o gráfico ativo no Instat, clicar em *Edit, Copy* (CTRL-C). No editor: *Editar, colar* (CTRL-V).

## 3 - DISTRIBUIÇÃO DE FREQÜÊNCIAS DE VARIÁVEIS QUALITATIVAS – CLASSIFICAÇÃO DUPLA

Para fazer uma tabela com a contagem de freqüências conjunta de duas variáveis qualitativas (tabela de dupla entrada ou de contingência), seguimos os mesmos passos da seção anterior (*Statistics, Tables, Ferquencies*), colocando as duas variáveis.

Como exemplo, construímos uma tabela com a porcentagem de famílias em cada nível de instrução do chefe da casa, por localidade:



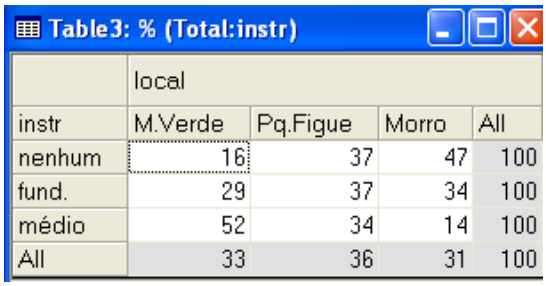
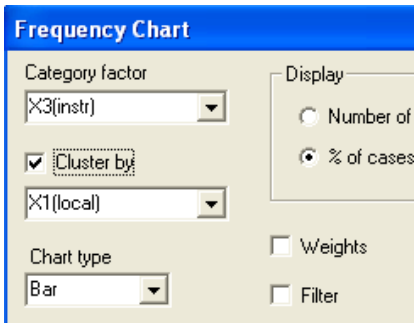
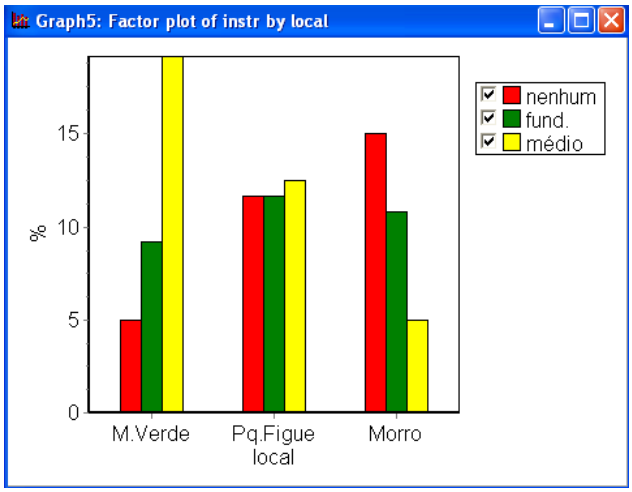



Table3: % (Total:instr)

| | local | | | |
|---|---|---|---|---|
| instr | M.Verde | Pq.Figue | Morro | All |
| nenhum | 16 | 37 | 47 | 100 |
| fund. | 29 | 37 | 34 | 100 |
| médio | 52 | 34 | 14 | 100 |
| All | 33 | 36 | 31 | 100 |

Se houver interesse em representar essa distribuição de freqüências de forma gráfica, basta, com a planilha de dados ativa, entrar em *Graphics, Frequency chart* e preencher como abaixo:

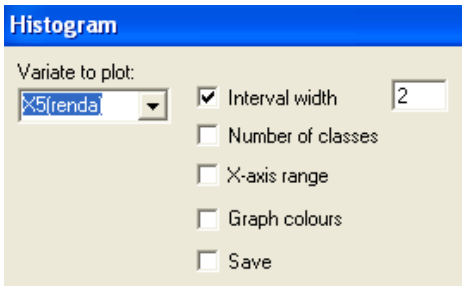


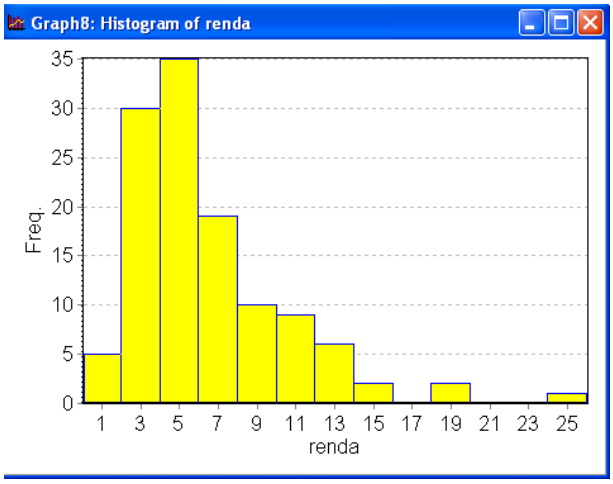




## 4 - DISTRIBUIÇÃO DE FREQÜÊNCIAS DE VARIÁVEIS QUANTITATIVAS

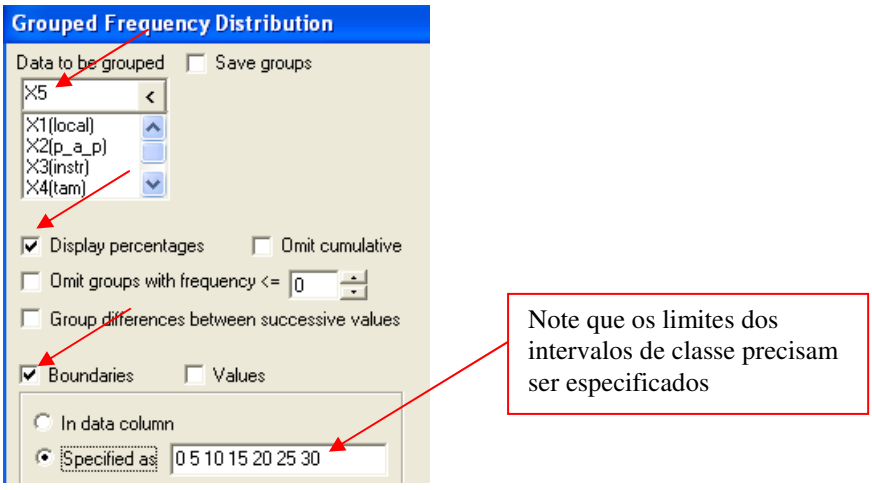
Vamos aqui considerar variáveis quantitativas contínuas. Para variáveis discretas (como "tam." em nosso arquivo exemplo), a construção da distribuição de freqüências pode ser feita como na seção 2. Para variáveis discretas com muitos valores distintos (como, p. ex., número de habitantes de municípios), podemos analisá-la como se fosse contínua.

A distribuição de freqüências desse tipo de variável é, usualmente, representada por um histograma de freqüências. No Instat: *Graphics*, *Histogram*. Observe que nas opções podemos escolher o número de classes ou a amplitude dessas. Abaixo exemplificamos com a variável "renda" e escolhemos a amplitude do intervalo igual a 2 (dois), observando que essa escolha só deve ser feita depois de uma primeira observação na distribuição dos valores.




Tabelas de freqüências com dados grupados (ou não) podem ser construídas usando: *Statistics, Summary, Group*. Exemplificamos com a variável "renda":




Os resultados, em formato texto (em *Commands and Output*), são:

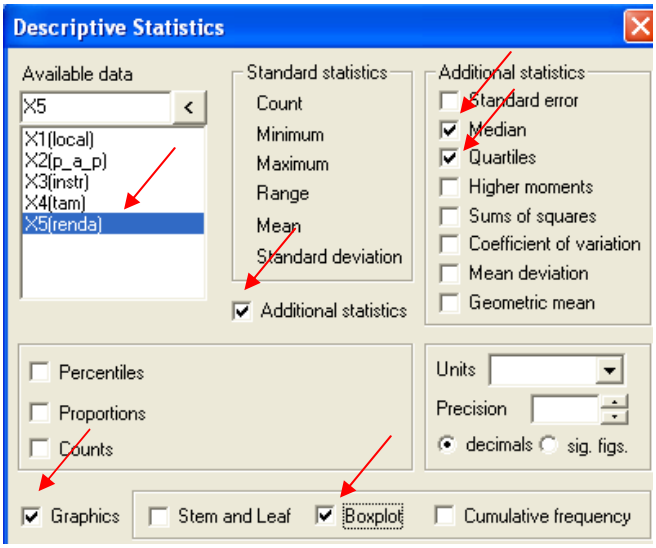
```
Grouped Frequency Distribution
GROup X5;BOU 0 5 10 15 20 25 30;PER
         1 values missing
Values       Count         %       Cum. %

<=   0           0      0.00         0.00
to   5          55     46.22        46.22
to  10          44     36.97        83.19
```

```
to 15          16       13.45      96.64
to 20           3        2.52      99.16
to 25           0        0.00      99.16
to 30           1        0.84     100.00
 > 30           0        0.00     100.00
_____________________________________________
Total         119      100.00
```

## 5 – MEDIDAS DESCRITIVAS

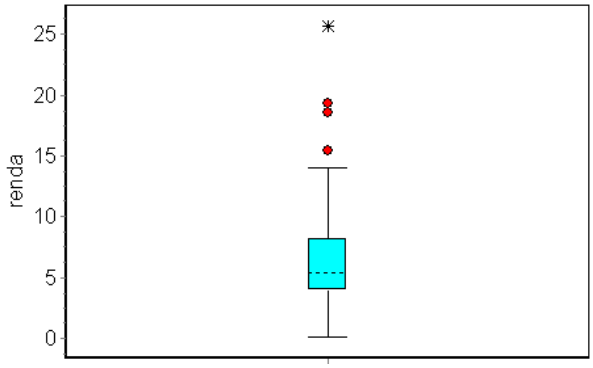
Medidas descritivas de uma variável quantitativa podem ser obtidas clicando na seqüência: *Statistics, Summary, Describe*.

Na janela que se abre, exemplificamos com a variável "renda". Optamos por incluir em *Additional Statistics* a *Median* (mediana) e os *Quartiles*; e o gráfico *Box-plot*.




Os resultados são apresentados em formato texto no lado direito (na janela *Commands and Output*) e o gráfico numa janela de formato gráfico. Vejam os resultados:

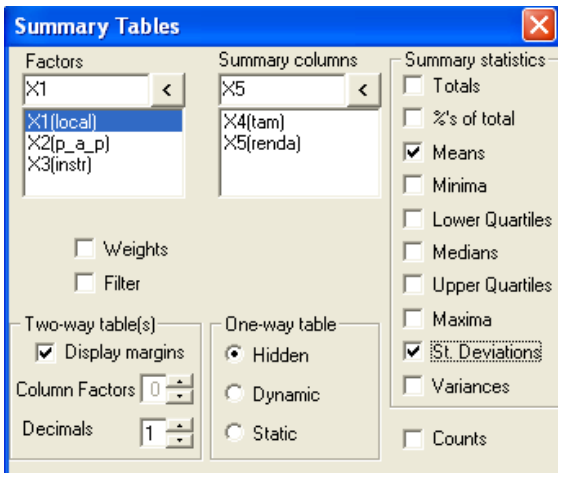
```
Descriptive Statistics

DES X5;LQU;UQU;IQU

Column                          renda

 No. of observations              120
 No. not missing                  119
 Minimum                          0.1
 Maximum                         25.7
 Range                           25.6
 Mean                          6.3403
 Std. deviation                4.0323
 Median                           5.4
 Lower Quartile                   3.9
 Upper Quartile                   8.4
 Quartile Deviation              2.25
```

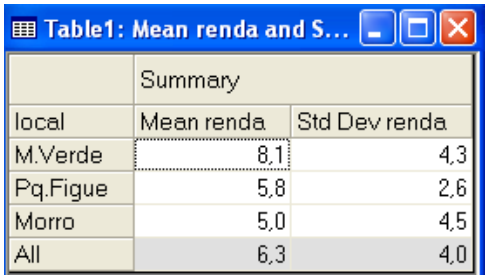



## 6 – MEDIDAS DESCRITIVAS EM SUBGRUPOS DA AMOSTRA

É muito comum termos interesse em se ter medidas descritivas de uma variável quantitativa em cada categoria de uma variável qualitativa ou em subgrupos da amostra ou população.

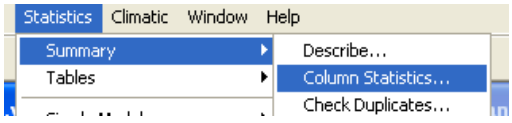
UMA MANEIRA: *Statistics, Tables, Summary*, conforme exemplificamos para o cálculo de medidas descritivas de "renda" por "local":



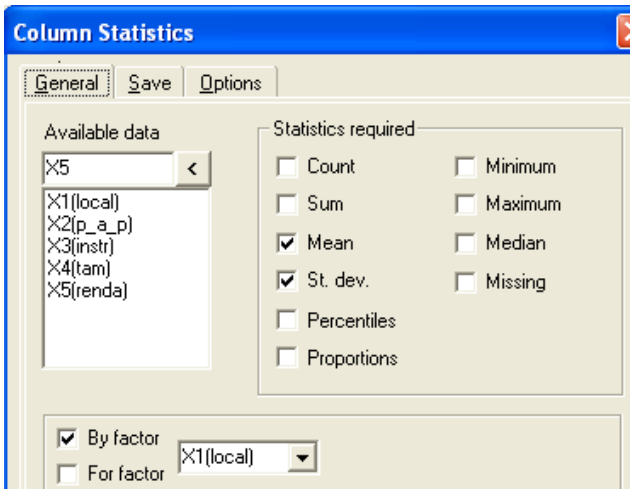




Na janela que se abriu, escolhemos apenas *Means* (média aritmética) e *St. Deviation* (desvio padrão). Os resultados estão na figura da direita, acima.

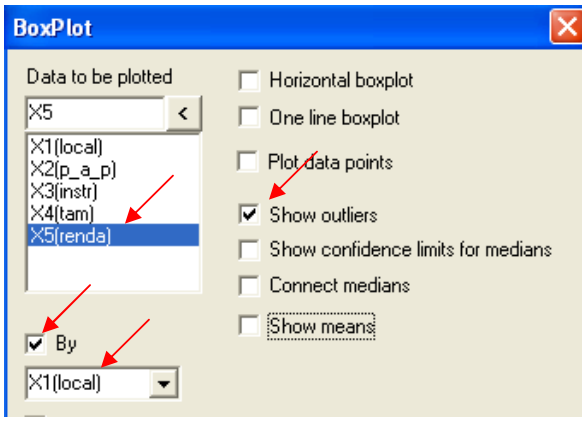
OUTRA MANEIRA: *Statistics, Summary, Column Statistics*, preenchendo a janela que se abre (*Available factor*, *By factor* e as estatísticas de interesse). Neste caso, os

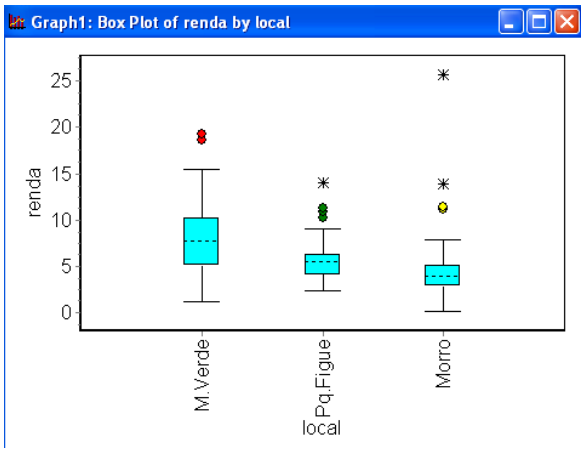


resultados são apresentados em formato texto, podendo também serem salvos na planilha de dados (usar *Save* na aba superior).




GRÁFICOS: Algumas medidas descritivas costumam ser representadas em diagramas de caixas. Seguindo o mesmo exemplo, clicar em *Graphics, Box-plot*, conforme mostrado a seguir:



## 7 – DIAGRAMAS DE DISPERSÃO

Clicar em *Graphics, Plot*. A seguir, exemplificado para "tam." E "renda":