Introdução ao R Commander

Análise exploratória de dados

Marcelo Lauretto

Análise Exploratória de Dados (AED)

- Objetivo: examinar os dados previamente à aplicação de qualquer técnica estatística.
  - Desta forma o analista consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas.
- Etapas típicas:
  - Preparação dos dados
  - Exame das variáveis individuais (distribuição, estatísticas padrões)
  - Exame de relações entre as variáveis
  - Identificação de possíveis casos atípicos (outliers)
  - Identificação e avaliação da presença de dados ausentes (missing);
  - Avaliar, se necessário, algumas suposições básicas, como simetria, homocedasticidade, entre outras.

Variáveis Quantitativas

- Medidas de posição:
  - Máximo (max): a maior observação
  - Mínimo (min): a menor observação
  - Moda (mo): é o valor (ou atributo) que ocorre com maior freqüência.
  - Média ($\bar{X}$): soma de todos os valores da variável dividida pelo número de observações.
  - Mediana (Md): valor que deixa 50% das observações à sua esquerda
  - Quartis: divide um conjunto de valores dispostos em forma crescente em quatro partes.
    - Primeiro Quartil (Q1): valor que deixa 25% das observações à sua esquerda.
    - Segundo Quartil (Q2): corresponde à mediana
    - Terceiro Quartil (Q3): valor que deixa 75% das observações à sua esquerda.

Variáveis Quantitativas

- Medidas de dispersão:
  - Amplitude: diferença entre os valores máximo e mínimo
  - Intervalo-Interquartil: É a diferença entre o terceiro e o primeiro quartil, ou seja, Q3 - Q1
  - Variância: média dos quadrados dos desvios em relação à média aritmética
  - Desvio Padrão (s): mede a variabilidade independente do número de observações e com a mesma unidade de medida da média
  - Coeficiente de Variação: mede a variabilidade numa escala percentual independente da unidade de medida ou da ordem de grandeza da variável:

$$CV = s \div \bar{X}$$

Data sets utilizados nesta aula:

- *Prestige*: Canadian occupational prestige data

Variable	**Values**
education	average years of education of occupational incumbents
income	average annual income of occupational incumbents, in dollars
prestige	average prestige rating of the occupation (0–100 scale)
women	percentage of occupational incumbents who were women
census	the Census occupation code
type	bc, blue-collar; wc, white-collar; prof, profess/technical/manager

Data sets utilizados nesta aula:

- *Prestige*: Canadian occupational prestige data (cont)
  - Disponível no Pacote *car:*
    - *Data > Data in packages > Read data from an attached package*
  - Após a carga:
    - Reordenar as classes da variável *type* para: bc, wc, prof:
      Data > Manage variables in active data set > Reorder factor levels
    - Converter a variável *census* para *factor:*
      Data > Manage variables in active data set > Convert numeric variables to factors

Data sets utilizados nesta aula:

- *Adler* data set:
  - Experimento destinado a analisar como as expectativas dos pesquisadores podem influenciar os dados por eles coletados.
  - “Pesquisadores assistentes” deveriam mostrar fotos de profissionais para entrevistados e pedir àqueles que atribuíssem uma nota de prestígio ao profissional da foto
  - Variáveis:
    - Expectation: para alguns pesquisadores, Adler informou que deveriam esperar altas notas, enquanto para outros, informou que deveriam esperar notas baixas
    - Instruction: alguns pesquisadores receberam instrução para coletarem “bons” dados; outros deveriam coletar dados “científicos”; para um 3º grupo, não foi dada nenhuma instrução de como coletar os dados.
  - Disponível no pacote *car*
  - Após a carga, reordenar os níveis do fator *instruction*

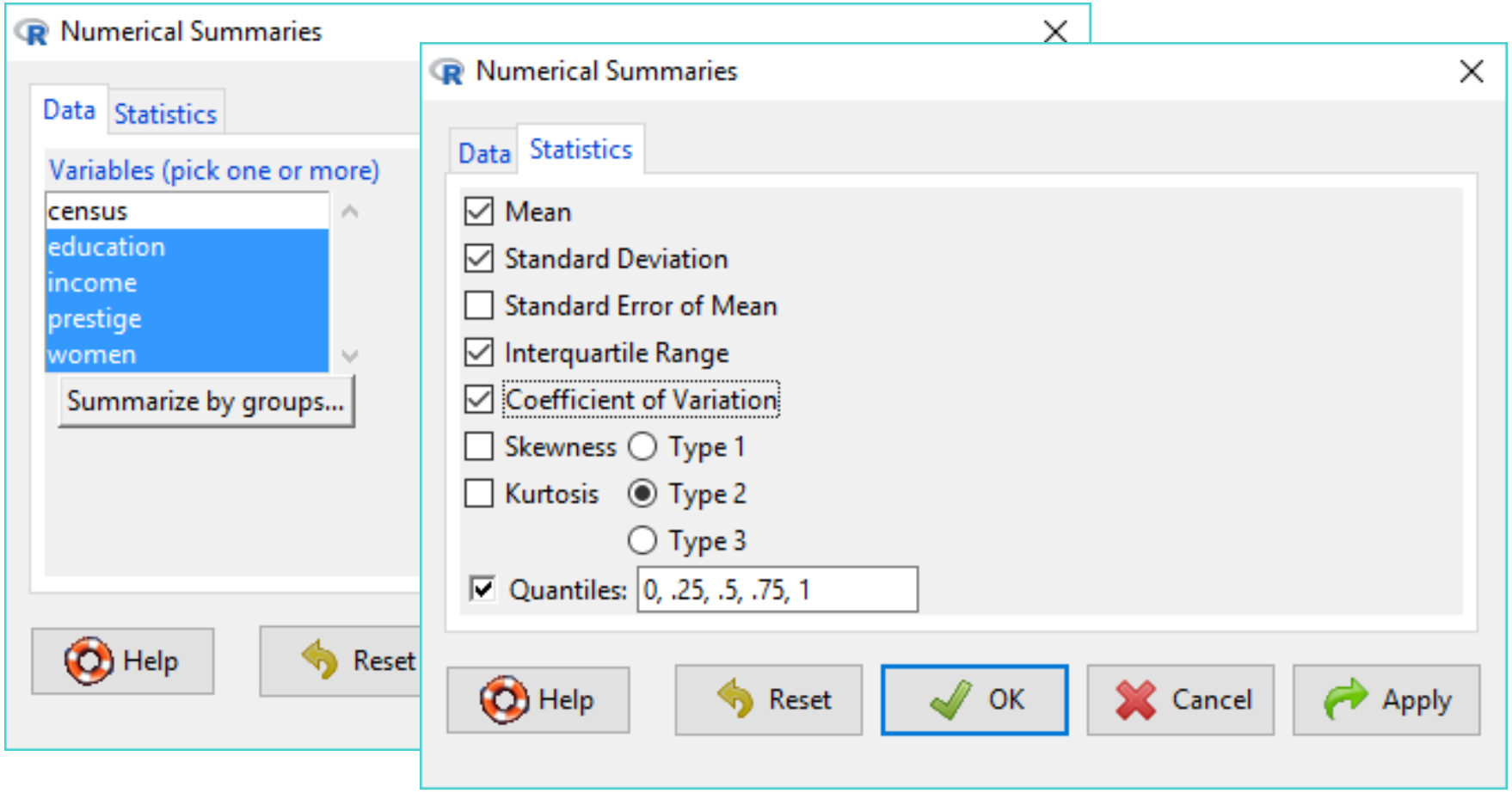
Sumários numéricos simples

- Sumário geral:
  - Data set *Prestige:*
  - *Statistics > Summaries > Active data set*

```
> summary(Prestige)
   education          income           women          prestige
 Min.   : 6.380   Min.   :  611   Min.   : 0.000   Min.   :14.80
 1st Qu.: 8.445   1st Qu.: 4106   1st Qu.: 3.592   1st Qu.:35.23
 Median :10.540   Median : 5930   Median :13.600   Median :43.60
 Mean   :10.738   Mean   : 6798   Mean   :28.979   Mean   :46.83
 3rd Qu.:12.648   3rd Qu.: 8187   3rd Qu.:52.203   3rd Qu.:59.27
 Max.   :15.970   Max.   :25879   Max.   :97.510   Max.   :87.20
     census       type
 Min.   :1113   bc  :44
 1st Qu.:3120   wc  :23
 Median :5135   prof:31
 Mean   :5402   NA's: 4
 3rd Qu.:8312
 Max.   :9517
```

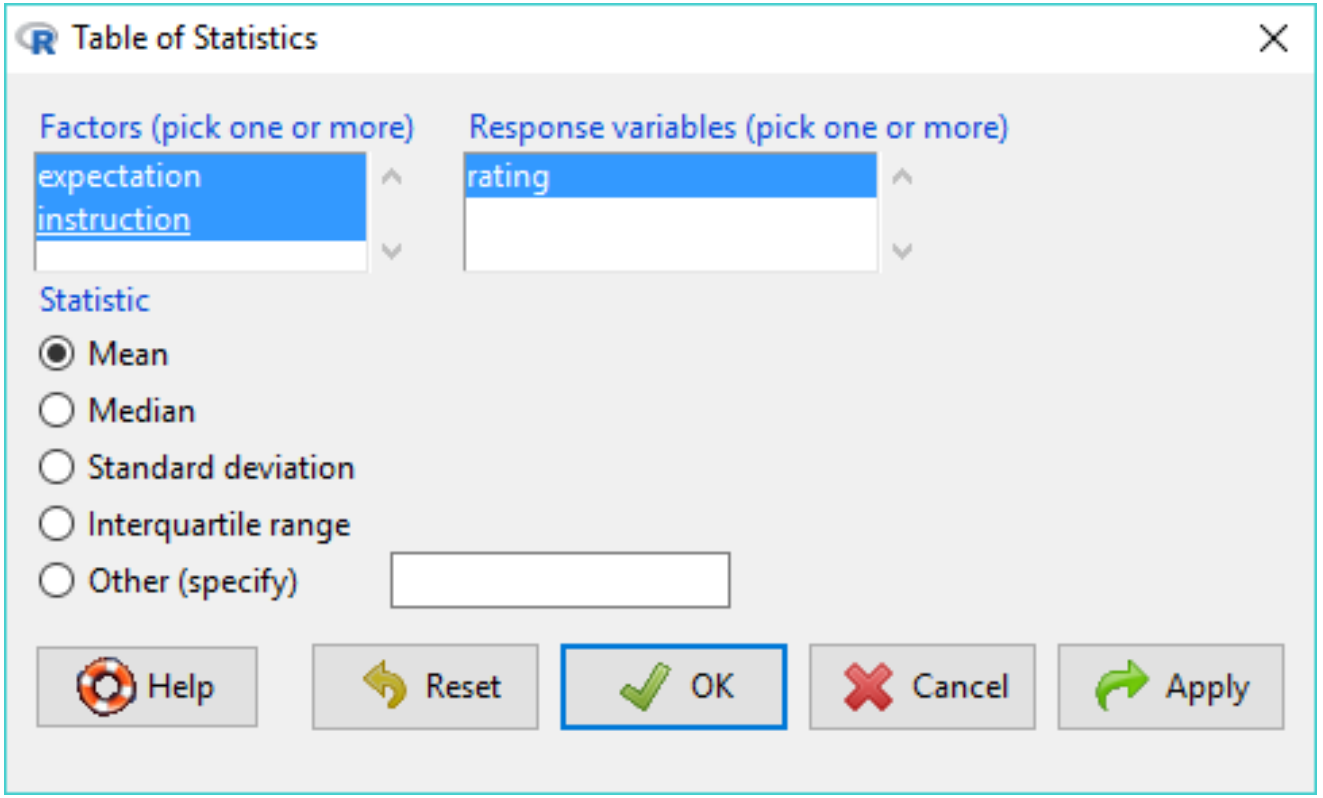
Sumários numéricos simples

- Sumário das variáveis numéricas:
  - *Statistics > Summaries > Numerical summaries*



Sumários numéricos simples

- Tabela de estatísticas:
  - Data set *Adler:*
  - *Statistics > Summaries > Table of statistics*
    - Executar com média, em seguida com desvio padrão



Sumários numéricos simples

- Outros sumários no submenu *Statistics > Summaries*:
  - *Frequency Distributions*
  - *Count missing observations*
  - *Correlation Matrix*