UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
IMD – INSTITUTO METRÓPOLE DIGITAL
Disciplina: IMD0033 – PROBABILIDADE – 24N12
Professora: Ismenia

ROTEIRO 7

Aula 07 – Resumir a informação

1. Objetivo

Trabalhar com a base de dados **utilizados da Tabela 2.1 do Livro de Bussab e Morettin (2014)** no RStudio, de forma a consolidar conceitos relacionados a variáveis do tipo quantitativas/ qualitativas e suas características.

2. Expectativa

Apresentar os passos básicos para a análise da base de dados utilizados da Tabela 2.1 do Livro de Bussab e Morettin (2014) no R/Rstudio, para que o aluno possa construir conceitos acerca de medidas resumo, medidas de posição e medidas de tendência central, além de gráficos para diversos tipos de variáveis

3. Teoria

Ver o arquivo: Atividade Complementar 02

4. Passos

Execute os comandos abaixo em R ou RStudio (ou o equivalente em outro software de sua preferência) e observe os seus resultados. Aproveite a sequencia obtida para criar uma análise resumida desta base de dados.

```
> library(readr)
> dados <- read_delim("~/Tab2-1.csv", ";",
+     escape_double = FALSE, trim_ws = TRUE)
```

ou

```
> dados <- read.csv2("/media/iblavatsky/Tab2-1.csv")
```

```
> names(dados)
[1] "N"  "estado_civil"    "grau_instrucao"  "n_filhos"  "salario"     "idade_anos"
[7] "idade_meses"     "reg_procedencia"
```

Tabelas

sintaxe:

```
table(dados)
```

Exemplo:

Dados utilizados da tabela 2.1 de Bussab e Morettin (2010).

```
> table(dados$reg_procedencia)
 capital interior    outra
      11       12       13
```

```
> table(dados$reg_procedencia,dados$estado_civil)
           casado solteiro
  capital       7        4
  interior      8        4
  outra         5        8

> table(dados$reg_procedencia,dados$estado_civil,dados$grau_instrucao)
, ,  = ensino fundamental

           casado solteiro
  capital       2        2
  interior      1        2
  outra         2        3

, ,  = ensino médio

           casado solteiro
  capital       4        1
  interior      6        1
  outra         2        4

, ,  = superior

           casado solteiro
  capital       1        1
  interior      1        1
  outra         1        1
```

Tabela de proporções

Mostra os dados em formato de tabela usando proporções:

sintaxe:

```
prop.table(tabela)
```

Exemplo:

```
> prop.table(table(dados$grau_instrucao))
ensino fundamental       ensino médio           superior
         0.3333333          0.5000000          0.1666667
```

Summary (Resumo)

Resume a variável quantitativa em: mínimo, máximo, média, mediana, 1º.quartil, 3º. quartil e dados não preenchidos. Caso a variável seja qualitativa, é informado o número de observações para cada nível.

sintaxe:

```
summary(variável)
```

Exemplo:

Resumo da variável salário

```
> summary(dados$salario)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   7.553  10.165  11.122  14.060  23.300
```

Resumo da variável salário apenas para casados

```
> summary(dados$salario[dados$estado_civil=="casado"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.560   8.742  11.925  12.123  15.030  23.300
```

Resumo da variável salário apenas para solteiros

```
> summary(dados$salario[dados$estado_civil=="solteiro"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   7.258   9.045   9.871  11.693  18.750
```

Resumo da variável qualitativa origem

```
> summary(dados$reg_procedencia)
 capital interior    outra
      11       12       13
```

Observação: Caso a variável desejada seja qualitativa numérica, é possível que o R interprete-a como sendo uma variável quantitativa. Para evitar que isso aconteça, utilize a função as.factor().

Média

sintaxe:

mean(variável)

Opções:

na.rm: TRUE, calcula a média considerando apenas os dados existentes, ignora os dados faltantes.

FALSE, calcula a média apenas se todos os valores estiverem preenchidos, caso contrário retorna NA.

Exemplo:

```
> mean(dados$n_filhos)
[1] NA
> mean(dados$n_filhos,na.rm=TRUE)
[1] 1.65
```

Variância

sintaxe:

var(variável)

Opções:

na.rm: TRUE, calcula a variância considerando apenas os dados existentes, ignora os dados faltantes.

FALSE, calcula a variância apenas se todos os valores estiverem preenchidos, caso contrário retorna NA.

Exemplo:

```
> var(dados$n_filhos)
[1] NA
> var(dados$n_filhos,na.rm=TRUE)
[1] 1.607895
```

Desvio Padrão

sintaxe:

sd(variável)

Opções:

na.rm: TRUE, calcula o desvio padrão considerando apenas os dados existentes, ignora os dados faltantes.

FALSE, calcula o desvio padrão apenas se todos os valores estiverem preenchidos, caso contrário retorna NA.

Exemplo:

```
> sd(dados$n_filhos)
[1] NA
> sd(dados$n_filhos,na.rm=TRUE)
[1] 1.268028
```

Mediana

Calcula a mediana do conjunto de dados.

median(variável)

Opções:

na.rm = TRUE calcula a mediana considerando apenas os dados existentes, ignora os dados faltantes.

FALSE calcula a mediana apenas se todos os valores estiverem preenchidos, caso contrário retorna NA.

Exemplo:

```
> median(dados$n_filhos)
[1] NA
> median(dados$n_filhos,na.rm=TRUE)
[1] 2
```

Aplica funções

Aplica a função desejada na variável escolhida segundo cada nível de um determinado fator.

sintaxe:

tapply(variável, fator, função)

Exemplo:

```
> tapply(dados$salario,dados$estado_civil,summary)
$casado
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.560   8.742  11.925  12.123  15.030  23.300

$solteiro
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.000   7.258   9.045   9.871  11.693  18.750

> tapply(dados$salario,dados$estado_civil,var)
  casado solteiro
24.12800 15.53465
```

Divisão de dados

Divide os dados em faixas determinadas.

sintaxe:

cut(variável, faixas, rótulos, opções)

Opções:

right: TRUE faz com que o intervalo seja fechado na direita e aberto na esquerda.

FALSE faz com que o intervalo seja aberto na direita e fechado na esquerda.

Exemplo:

```
> cut(dados$idade_anos, breaks=c(20,30,40,50), labels=c("A","B","C"), right=FALSE)
 [1] A B B A C A C C B A B A B C B B B B A B B B C A B B C A C B B B C B C C
Levels: A B C

> table(cut(dados$idade_anos,right=F, breaks=c(20,30,40,50),labels=c("20|-30","30|-40","40|-50")))
 20|-30 30|-40 40|-50
      8     18     10

> table(cut(dados$salario,breaks=c(4,8,12,16,20,24), labels=c("4|-8","8|-12","12|-16","16|-20","20|-24"), right=F))
  4|-8  8|-12 12|-16 16|-20 20|-24
    10     12      8      5      1
```

Observação: É possível a utilização da função cut() dentro da função table() diretamente. O resultado será uma tabela com a freqüência de cada intervalo determinado pela função cut().

Gráficos

Os gráficos nos permitem analisar uma grande quantidade de informações de forma rápida, sem que seja necessário olhar tabelas e medidas de resumo. O R possui uma enorme capacidade para gerar diversos tipos de gráficos de alta qualidade totalmente configuráveis, desde cores e tipos de linhas, até legendas e textos adicionais.

A grande maioria das funções gráficas faz uso de opções comuns, ou seja, é extremamente fácil personalizar qualquer tipo de gráfico pois muitas das opções são iguais. As opções comuns a todos os gráficos serão abordadas aqui, e em cada seção seguinte as opções específicas àquele determinado tipo de gráfico serão apresentadas.

Opções:

- xlim:(inicio,fim) dupla contendo os limites do eixo X.
- ylim:(inicio,fim) dupla contendo os limites do eixo Y.
- xlab:rótulo para o eixo X.
- ylab:rótulo para o eixo Y.
- main:título principal do gráfico.
- col:cor de preenchimento do gráfico, podendo ser um vetor. A lista das cores disponíveis pode ser obtida através do comando colors().

locator

Permite localizar uma coordenada clicando com o mouse no gráfico. Se não for definida a opção type, retorna apenas as coordenadas do ponto clicado. Útil para inserir textos e outros elementos em gráficos já prontos.

sintaxe:

```
locator()
```

Opções

n: Número máximo de pontos a localizar.

type: p: cria pontos no gráfico com as coordenadas indicadas pelo mouse.
l: cria linhas no gráfico com as coordenadas indicadas pelo mouse.

text

Insere um texto nas coordenadas definidas.

sintaxe:

```
text(x, y, labels, cex, col)
```

Opções

x: Posição relativa a abscissa (eixo X).

y: Posição relativa a ordenada (eixo Y).

labels: Texto (ou vetor com textos) a ser inserido nas coordenadas definidas por x e y.

cex: Proporção relativa ao tamanho dos caracteres do texto (padrão: 1).

col: Cor do texto a ser inserido (padrão: preto).

Gráfico de barras

Gráfico de freqüências para variáveis qualitativas.

opções:

space: espaço deixado antes de cada barra

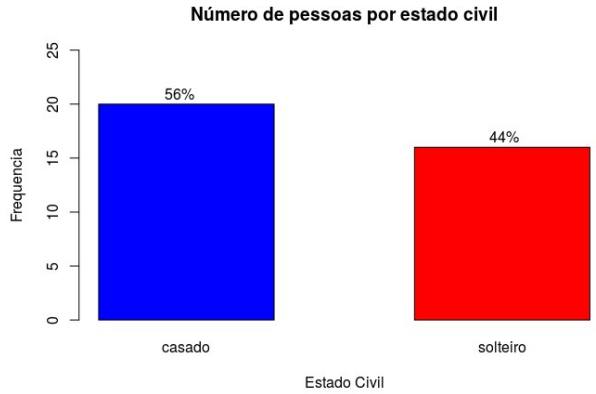
width: vetor contendo a largura relativa de cada barra com relação as demais. Valores iguais para todas as barras não terão efeito, pois a relação entre elas será 1.

sintaxe:

```
barplot(dados,opções)
```

Exemplo:

```
> barplot(table(dados$estado_civil), col=c("blue","red"),
            ylim=c(0,25),
            space=.8, width=c(.2,.2),
            main="Número de pessoas por estado civil",
            xlab="Estado Civil", ylab="Frequencia")
> text(locator(n=2),c("56%","44%"))
```



Observação:

• Caso não seja especificada a opção xlim, os valores da opção width não serão interpretados como valores absolutos, mas como valores relativos as demais barras.

Ex: barplot(table(dados$estado_civil), width(0.2,0.2)) tem o mesmo efeito que
barplot(table(dados$estado_civil), width(2,2))

• Após o comando text(locator(n=2), c("56%","44%")), são necessários dois cliques em pontos do gráfico de barras onde serão inseridos os textos com os percentuais relativos a cada barra.

Histograma

Gráfico de distribuição de freqüências para variáveis quantitativas.

opções:

prob: T plota a densidade.
F plota a freqüência absoluta.

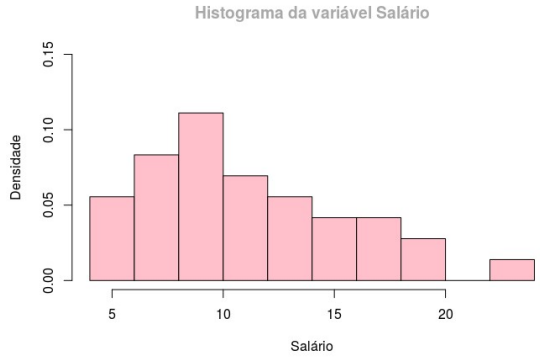
breaks: vetor contendo os pontos de definição das larguras das barra do histograma.

sintaxe:

hist(dados, opções)

Exemplo:

```
> hist(dados$salario, main="Histograma da variável Salário", prob=T, xlab="Salário", ylab="Densidade",col=c("pink"), ylim=c(0,0.15), col.main="darkgray")
```



Boxplot

sintaxe:

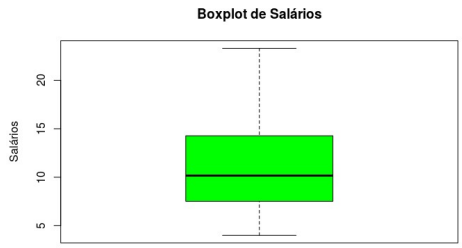
boxplot(dados, opções)

opções:

outline: T plota os outliers.
F não plota os outliers.

Exemplo:

```
> boxplot(dados$salario, main="Boxplot de Salários",
        ylab="Salários", col=("green"))
```




```
> boxplot(dados$salario ~ dados$grau_instrucao,
        main="Boxplot de Salários por grau de Instrução",
        xlab="Grau de Instrução", ylab="Salários",
        col=c("yellow","orange","red"))
```

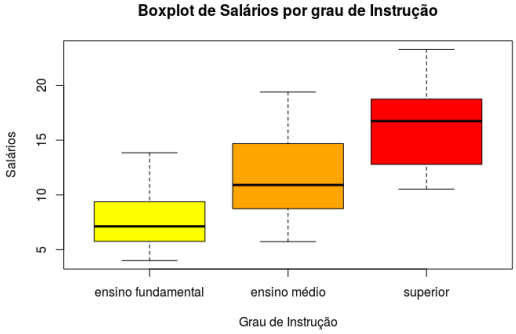


Gráfico de Pizza

sintaxe:

pie(dados, opções)

opções:

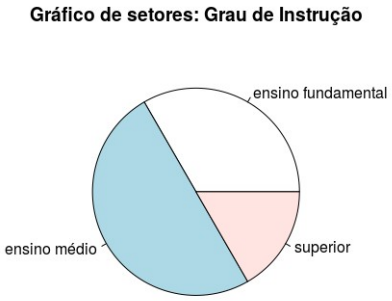
labels: vetor contendo os rótulos de cada fatia.

radius: raio da circunferência da pizza. (padrão=1)

col: vetor contendo as cores das fatias.

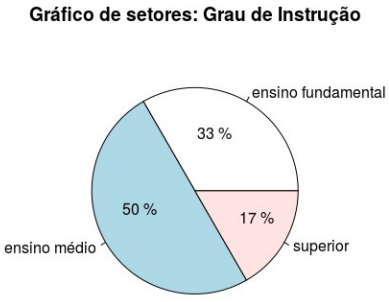
Exemplo:

```
> pie(table(dados$grau_instrucao),
     main="Gráfico de setores: Grau de Instrução")
```



Para incluir as porcentagens dentro de cada fatia, execute as linhas abaixo clicando na fatia branca, azul e rosa, nesta sequência.

```
> text(locator(n=1),
      paste(round(prop.table(table(dados$grau_instrucao))[1],
                  digits=2)*100,"%"))
> text(locator(n=1),
      paste(round(prop.table(table(dados$grau_instrucao))[2],
                  digits=2)*100,"%"))
> text(locator(n=1),
      paste(round(prop.table(table(dados$grau_instrucao))[3],
                  digits=2)*100,"%"))
```



Sugestão: Copie os comandos em um arquivo texto e abra este arquivo dentro do RStudio para executar passo a passo. Observe os comandos que por ventura dêm erro em sua execução, e compare os resultados com os obtidos neste Roteiro.

5. Referências

MORETTIN, Pedro Alberto; BUSSAB, Wilton de Oliveira. **Estatística básica.** 6. ed. rev. atual. São Paulo: Saraiva, São Paulo: Saraiva, c2014.