Análise descritiva em R

MAT021 - Estatística I - B

Rodney Fonseca

9/12/2024

Lendo banco de dados em R

- Para ler um arquivo, salvamos o script do RStudio numa pasta, e colocamos o arquivo de dados na mesma pasta.
- Para indicar ao R onde o arquivo está, podemos clicar na seguinte sequência de botões:

Session -> Set Working Directory -> To Source File Location

- Para ler um arquivo, salvamos o script do RStudio numa pasta, e colocamos o arquivo de dados na mesma pasta.
- Para indicar ao R onde o arquivo está, podemos clicar na seguinte sequência de botões:

Session -> Set Working Directory -> To Source File Location

- Os nossos dados estão num arquivo CSV. Usaremos a função `read.csv()` para ler o arquivo

```
dados_mb = read.csv('empresa_MB.csv',
                    header = TRUE, sep = ',', dec = ',')
```

- Para ler um arquivo, salvamos o script do RStudio numa pasta, e colocamos o arquivo de dados na mesma pasta.
- Para indicar ao R onde o arquivo está, podemos clicar na seguinte sequência de botões:

Session -> Set Working Directory -> To Source File Location

- Os nossos dados estão num arquivo CSV. Usaremos a função `read.csv()` para ler o arquivo

```
dados_mb = read.csv('empresa_MB.csv',
                    header = TRUE, sep = ',', dec = ',')
```

- Para visualizar os dados, basta digitar o comando `View(dados_mb)`

Relembrando

Variáveis na planilha

```
names(dados_mb)
```

```
## [1] "N"                "estado_civil"    "grau_instrucao" "
## [5] "salario"          "idade"           "regiao"
```

Relembrando

Variáveis na planilha

```
names(dados_mb)
```

```
## [1] "N"                "estado_civil"    "grau_instrucao" "
## [5] "salario"          "idade"           "regiao"
```

Checando os valores de uma das variáveis

```
dados_mb$grau_instrucao[1]
```

```
## [1] "ensino fundamental"
```

Relembrando

Variáveis na planilha

```
names(dados_mb)
```

```
## [1] "N"                "estado_civil"   "grau_instrucao" "
## [5] "salario"          "idade"          "regiao"
```

Checando os valores de uma das variáveis

```
dados_mb$grau_instrucao[1]
```

```
## [1] "ensino fundamental"
```

Tabela de frequências para uma variável qualitativa

```
tab_instrucao <- table(dados_mb$grau_instrucao)
tab_instrucao
```

```
##
## ensino fundamental       ensino médio           superio
##                 12                 18                 6
```

Gráfico para uma variável qualitativa

```
barplot(tab_instrucao,
        main = 'Gráfico de colunas',
        ylab = 'frequência')
```

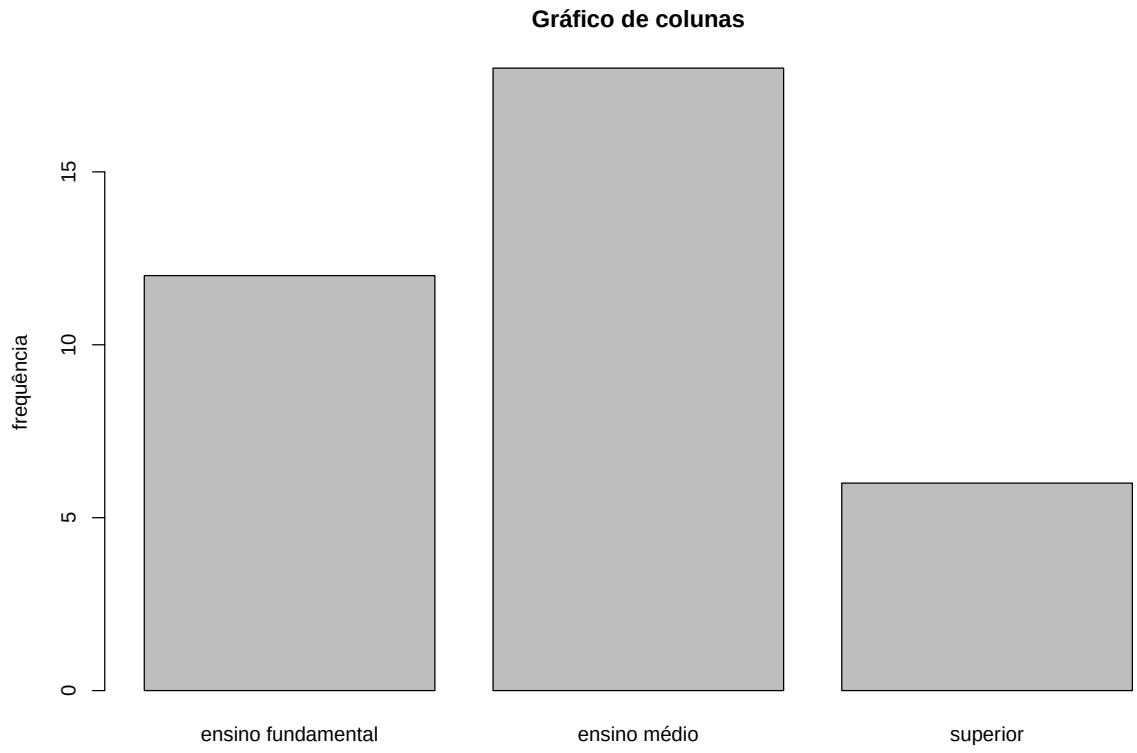
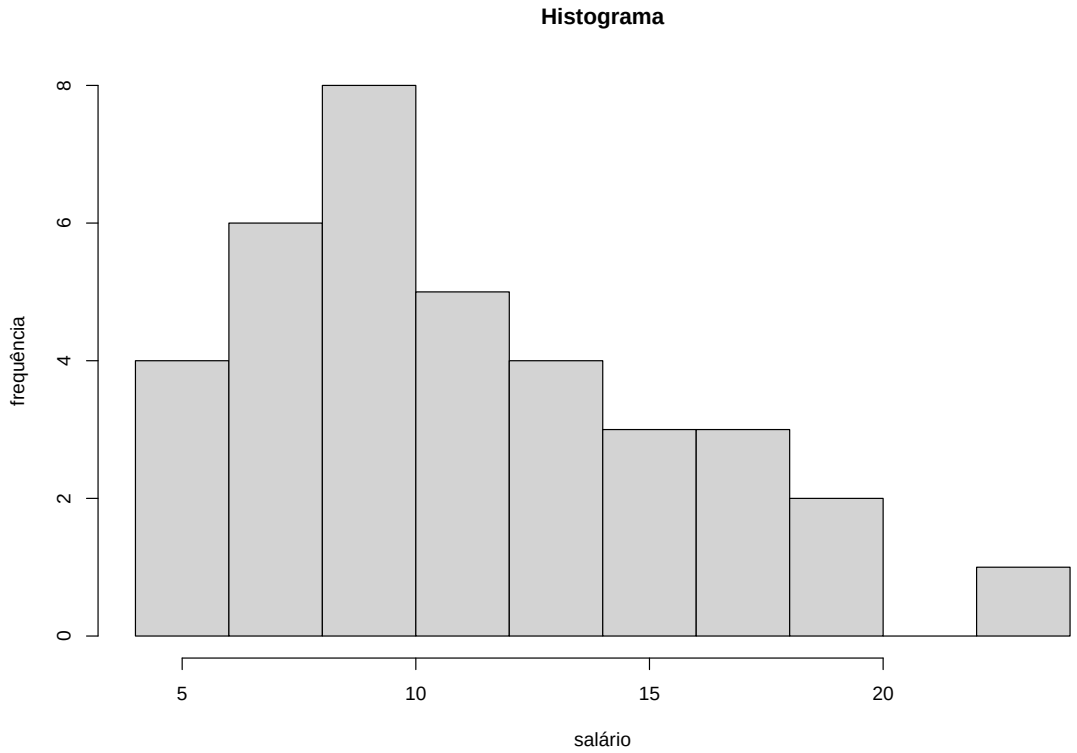


Gráfico para uma variável quantitativa

```
hist(dados_mb$salario, main = 'Histograma',
     ylab = 'frequência', xlab = 'salário')
```



Medidas de posição central

Média aritmética

- Sejam $x_1, x_2, \ldots, x_n$ os n valores observados. A média amostral é definida como

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Média aritmética

- Sejam $x_1, x_2, \ldots, x_n$ os n valores observados. A média amostral é definida como

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Em R usamos a função `mean()`

```
salario = dados_mb$salario
mean(salario)
```

```
## [1] 11.12222
```

Mediana

- Valor que ocupa a posição central nos dados ordenados, de tal forma que 50% dos valores são menores do que a mediana

Mediana

- Valor que ocupa a posição central nos dados ordenados, de tal forma que 50% dos valores são menores do que a mediana
- Em R usamos a função `median()`

```
median(salario)
```

```
## [1] 10.165
```

Mediana

- Valor que ocupa a posição central nos dados ordenados, de tal forma que 50% dos valores são menores do que a mediana
- Em R usamos a função `median()`

```
median(salario)
```

```
## [1] 10.165
```

- Como o número de observações `n`=36 é par, se usássemos a fórmula teríamos:

```
n = length(salario)
(salario[n/2] + salario[(n/2) + 1])/2
```

```
## [1] 10.165
```

Moda

- É o valor da variável que ocorre com *maior frequência* nos dados observados.

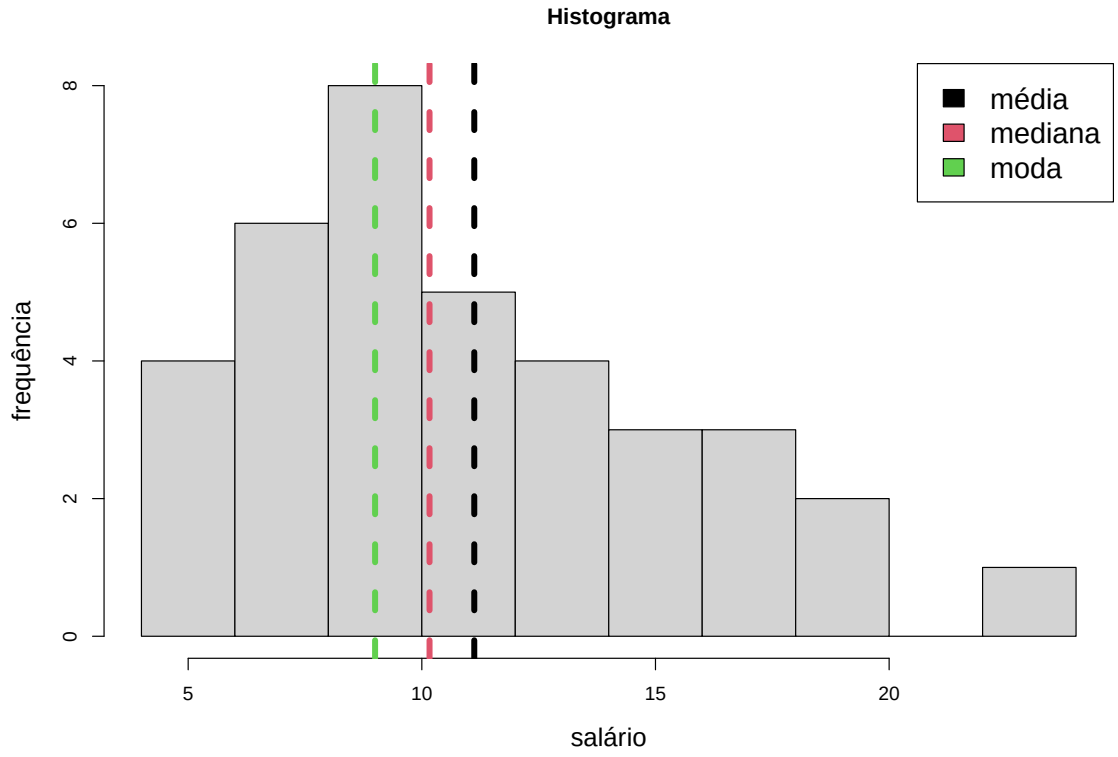
Moda

- É o valor da variável que ocorre com *maior frequência* nos dados observados.
- O R base não possui uma função para a moda, mas podemos calculá-la usando frequências obtidas pela função que faz o histograma

```
hist_salario <- hist(salario, plot = FALSE)
id_max <- which.max(hist_salario$counts)
hist_salario$mids[id_max]
```

```
## [1] 9
```

Comparação da média, mediana e moda



Quartis e boxplot

Definição de quartil

- O quantil de ordem $i\%$ é um número P_i tal que $i\%$ dos dados são menores que tal valor
- Quantis especiais
  - Primeiro quartil: percentil 25%
  - Segundo quartil: percentil 50% (mediana)
  - Terceiro quartil: percentil 75%

Definição de quartil

- O quantil de ordem $i\%$ é um número P_i tal que $i\%$ dos dados são menores que tal valor
- Quantis especiais
  - Primeiro quartil: percentil 25%
  - Segundo quartil: percentil 50% (mediana)
  - Terceiro quartil: percentil 75%
- Em R usamos a função `quantile()`

Primeiro quartil

```
quantile(dados_mb$salario, 0.25)
```

```
##     25%
## 7.5525
```

Segundo quartil (mediana)

```
quantile(dados_mb$salario, 0.5)
```

```
##     50%
## 10.165
```

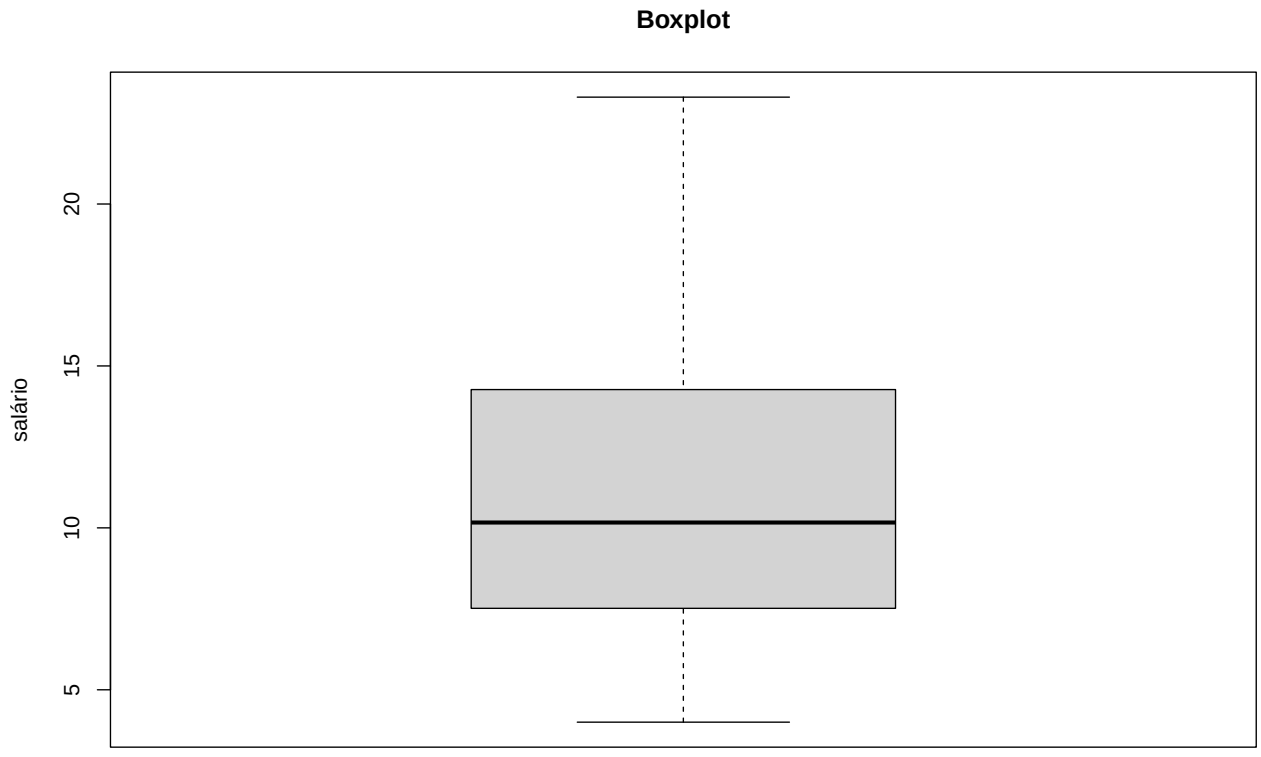
Terceiro quartil

```
quantile(dados_mb$salario, 0.75)
```

```
##    75%
## 14.06
```

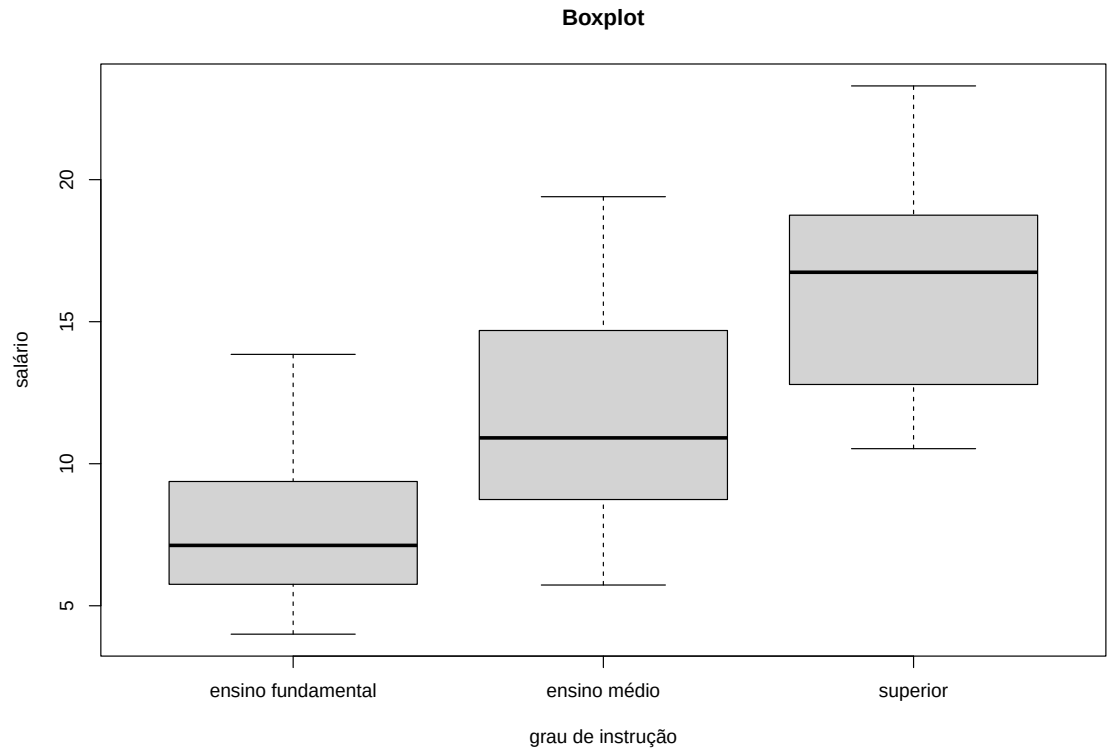
Boxplot

```
boxplot(salario, main = 'Boxplot', ylab = 'salário')
```



Boxplot com variáveis quantitativas e qualitativas

```
boxplot(dados_mb$salario ~ dados_mb$grau_instrucao,
        main = 'Boxplot', ylab = 'salário',
        xlab = 'grau de instrução')
```



Medidas de dispersão

Amplitude total

- Diferença entre valores máximo e mínimo nos dados

$$AT = X_{max} - X_{min}$$

Amplitude total

- Diferença entre valores máximo e mínimo nos dados

$$AT = X_{max} - X_{min}$$

- Para calcular amplitude total no R, basta usar as funções `max()` e `min()`

```
max(salario) - min(salario)
```

```
## [1] 19.3
```

Variância

- Dado um conjunto de dados $x_1, \ldots, x_n$, seja $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a média amostral. A *variância amostral* é então calculada como

$$S^2 = \frac{1}{n}\Big((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\Big) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

Variância

- Dado um conjunto de dados $x_1, \ldots, x_n$, seja $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ a média amostral. A *variância amostral* é então calculada como

$$S^2 = \frac{1}{n}\Big((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\Big) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

- Em R usamos a função `var()` para calcular variância

```
var(salario)
```

```
## [1] 21.04477
```

Desvio padrão

- O *desvio padrão* é a raiz quadrada da variância:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

- Em R, usamos `sd()` para calcular o desvio padrão

```
sd(salario)
```

```
## [1] 4.587458
```

Desvio padrão

- O *desvio padrão* é a raiz quadrada da variância:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

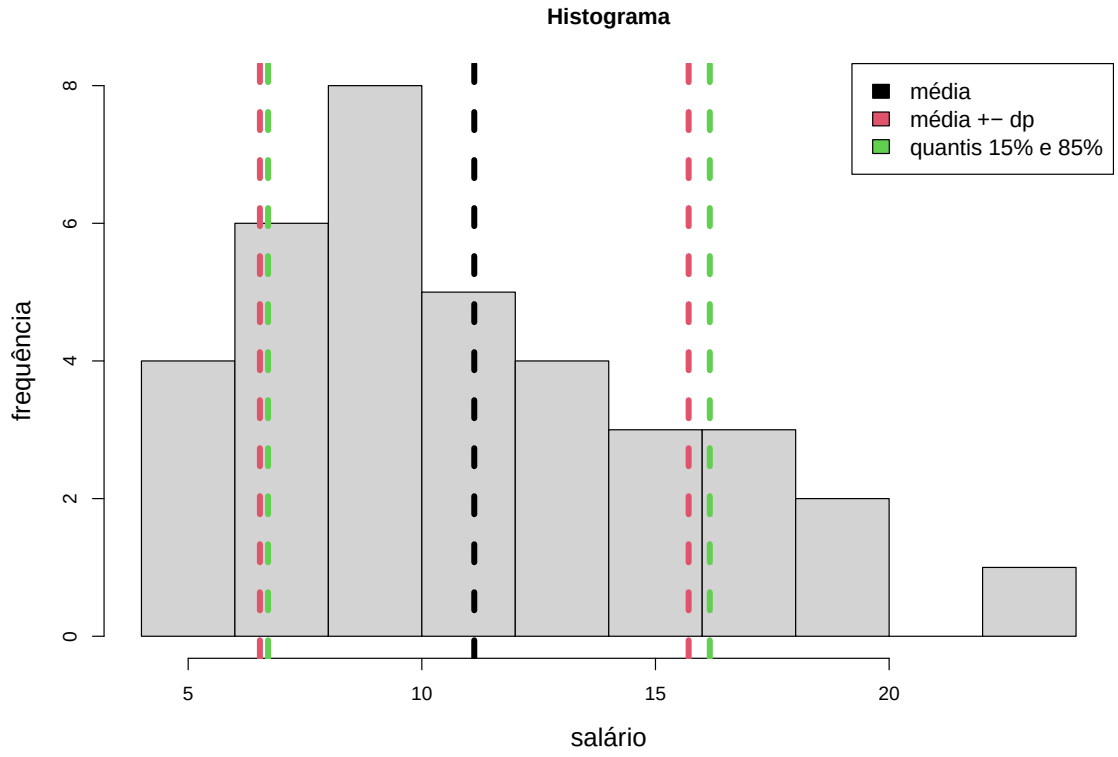
- Em R, usamos `sd()` para calcular o desvio padrão

```
sd(salario)
```

```
## [1] 4.587458
```

- Normalmente, em torno de 70% dos dados ficam concentrados entre $\bar{x} - S$ e $\bar{x} + S$

Interpretação do desvio padrão



Coeficiente de variação

- É definido como a razão do desvio padrão e da média amostral

$$CV = \frac{S}{\bar{X}},$$

em que $\bar{X}$ é a média e S é o desvio padrão

Coeficiente de variação

- É definido como a razão do desvio padrão e da média amostral

$$CV = \frac{S}{\bar{X}},$$

em que $\bar{X}$ é a média e S é o desvio padrão

- Em R, usamos as funções `mean()` e `sd()` para calcular o CV

```
sd(salario)/mean(salario)
```

```
## [1] 0.4124587
```

- O coeficiente de variação permite comparar a variabilidade de dados com valores de magnitudes diferentes

- O coeficiente de variação permite comparar a variabilidade de dados com valores de magnitudes diferentes
- Por exemplo, vamos comparar o **desvio padrão** do salário para cada grau de instrução:

```
by(dados_mb$salario, dados_mb$grau_instrucao, sd)
```

```
## dados_mb$grau_instrucao: ensino fundamental
## [1] 2.956464
## ------------------------------------------------------------
## dados_mb$grau_instrucao: ensino médio
## [1] 3.715144
## ------------------------------------------------------------
## dados_mb$grau_instrucao: superior
## [1] 4.502438
```

- Obtemos uma conclusão diferente se usarmos o *coeficiente de variação*:

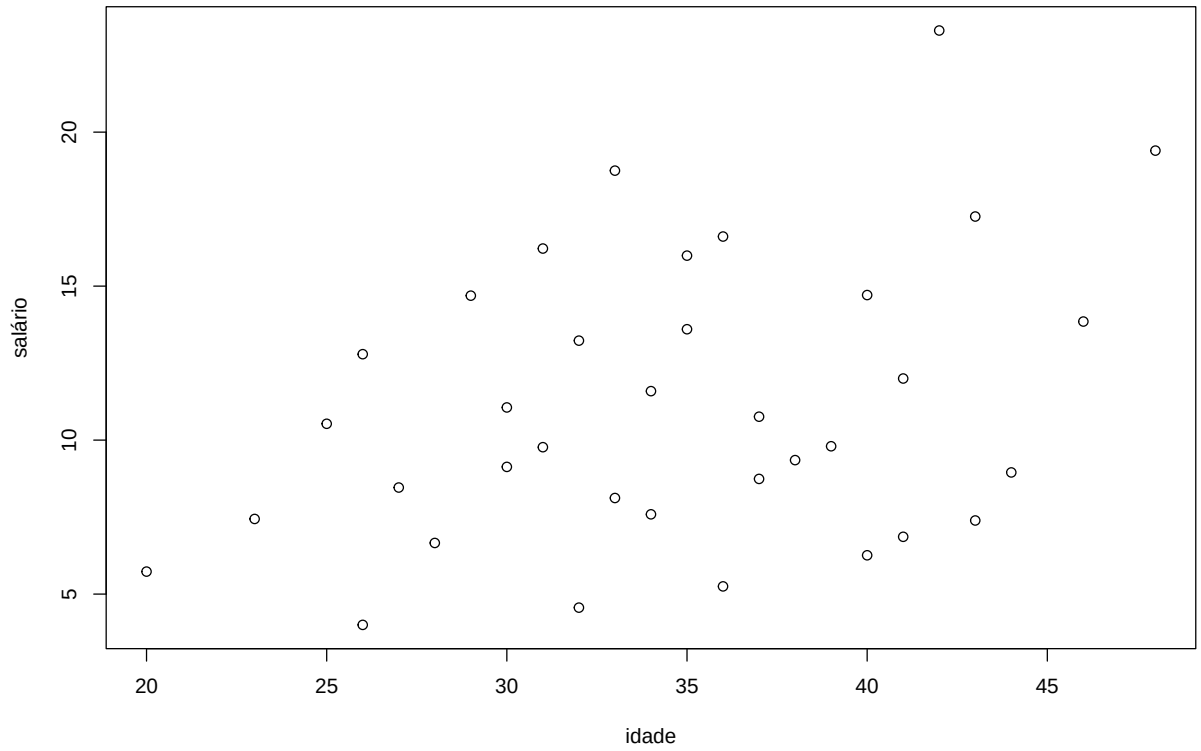
```
by(dados_mb$salario, dados_mb$grau_instrucao,
   function(x) sd(x)/mean(x))
```

```
## dados_mb$grau_instrucao: ensino fundamental
## [1] 0.3772604
## ------------------------------------------------------------
## dados_mb$grau_instrucao: ensino médio
## [1] 0.322262
## ------------------------------------------------------------
## dados_mb$grau_instrucao: superior
## [1] 0.2732891
```

Análise bivariada

diagrama de dispersão

```
idade = dados_mb$idade
plot(salario ~ idade, ylab = 'salário',xlab = 'idade')
```



Coeficiente de correlação

- Dado um conjunto de pares de dados $(x_1, y_1), \ldots, (x_n, y_n)$, o *coeficiente de correlação linear* é calculado como

$$\operatorname{corr}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_X} \right) \left(\frac{y_i - \bar{y}}{S_Y} \right),$$

Coeficiente de correlação

- Dado um conjunto de pares de dados $(x_1, y_1), \ldots, (x_n, y_n)$, o *coeficiente de correlação linear* é calculado como

$$\text{corr}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_X} \right) \left(\frac{y_i - \bar{y}}{S_Y} \right),$$

- Em R usamos a função `cor()`

```
cor(salario, idade)
```

```
## [1] 0.3633622
```