Relatório Descritivo do Conjunto de Dados Companhia MB

20 de abril de 2019

Introdução

Neste trabalho analisa-se o conjunto de dados Companhia MB. Esse conjunto de dados foi obtido da seguinte referência:

Bussab, W. O. e Morettin, P. A. **Estatística Básica**. 8 ed. 2013

O conjunto de dados refere-se à uma pesquisa feita com empregados da fictícia companhia MB. Se tratam, portanto, de dados artificiais criados para fins didáticos.

Para facilitar o estudo chamaremos os dados por CMB:

```
CMB <- CompanhiaMB
```

Para realizar essas análises será necessária a seguinte biblioteca:

```
library("e1071")
```

Além disso vamos definir a seguinte função para calcular a variância populacional quando necessária:

```
pvar <- function(x) sum((x - mean(x))**2) / length(x)
```

O conjunto de dados possui as seguintes dimensões:

```
dim(CMB)
```

```
[1] 36  8
```

Portanto, esse conjunto de dados corresponde à 36 observações de 8 variáveis. As variáveis consideradas são as seguintes:

```
str(CMB)
```

```
'data.frame':   36 obs. of  8 variables:
 $ Estado.Civil     : Factor w/ 2 levels "casado","solteiro": 2 1 1 2 2 1 2 2 1 2 ...
 $ Grau.de.Instrução: Ord.factor w/ 3 levels "fundamental"<..: 1 1 1 2 1 1 1 1 2 2 ...
 $ Filhos           : int  NA 1 2 NA NA 0 NA NA 1 NA ...
 $ Salário          : num  4000 4560 5250 5730 6260 6660 6860 7390 7590 7440 ...
 $ Anos             : int  26 32 36 20 40 28 41 43 34 23 ...
 $ Meses            : int  3 10 5 10 7 0 0 4 10 6 ...
 $ Idade            : num  26.2 32.8 36.4 20.8 40.6 ...
 $ Procedência      : Factor w/ 3 levels "capital","interior",..: 2 1 1 3 3 2 2 1 1 3 ...
```

Note que as variáveis ** e *Meses* foram usadas no cálculo da variável *Idade* por meio do seguinte comando:

```
round(CMB$Anos+CMB$Meses/12,2)
```

```
 [1] 26.25 32.83 36.42 20.83 40.58 28.00 41.00 43.33 34.83 23.50 33.50
[12] 27.92 37.42 44.17 30.42 38.67 31.58 39.58 25.67 37.33 30.75 34.17
[23] 41.00 26.08 32.42 35.00 46.58 29.67 40.50 35.83 31.42 36.33 43.58
[34] 33.58 48.92 42.17
```

```
CMB$Idade
```

```
 [1] 26.25 32.83 36.42 20.83 40.58 28.00 41.00 43.33 34.83 23.50 33.50
[12] 27.92 37.42 44.17 30.42 38.67 31.58 39.58 25.67 37.33 30.75 34.17
[23] 41.00 26.08 32.42 35.00 46.58 29.67 40.50 35.83 31.42 36.33 43.58
[34] 33.58 48.92 42.17
```

Portanto vamos ignorar as variáveis *Anos* e *Meses* em nosso estudo e nos focar apenas na variável *Idade*.

Desta forma as variáveis de interesse podem ser classificadas da seguinte forma:

- *Estado Civil*: Categórica Nominal
- *Grau de Instrução*: Categórica Ordinal
- *Número de Filhos*: Quantitativa Discreta
- *Salários*: Quantitativa Contínua
- *Idade*: Quantitativa Contínua
- *Região de Procedência*: Categórica Nominal

Análise Univariada

Variável *Estado Civil*

A seguir vemos a distribuições de frequências absolutas para a variável *Estado Civil*:

```
FreqAbs <- table(CMB$Estado.Civil)
FreqAbs
```

```
  casado solteiro
      20       16
```

E a correspondente distribuição de frequências relativas:

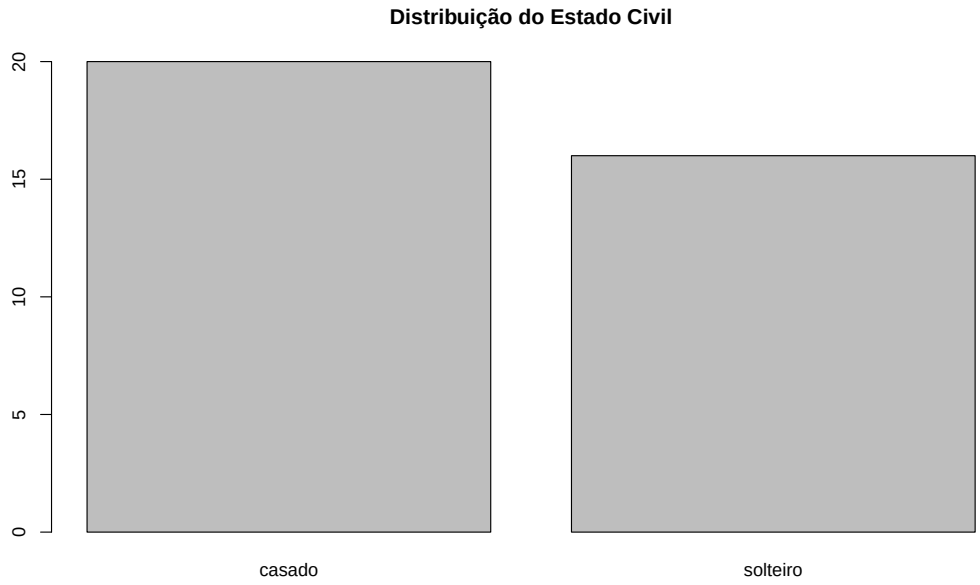
```
FreqRel <- prop.table(FreqAbs)
FreqRel
```

```
   casado  solteiro
0.5555556 0.4444444
```

Note que a maioria dos entrevistados é casado, correspondendo a uma proporção de 0.56 para 1.

Um gráfico adequado para essa variável é o gráfico de barras:

```
barplot(FreqAbs, main = "Distribuição do Estado Civil")
```



Variável *Grau de Instrução*

As distribuições de frequências absolutas e relativas para a variável *Grau de Instrução* são:

```
FreqAbs <- table(CMB$Grau.de.Instrução)
FreqAbs
```

```
fundamental       médio    superior
         12          18           6
```

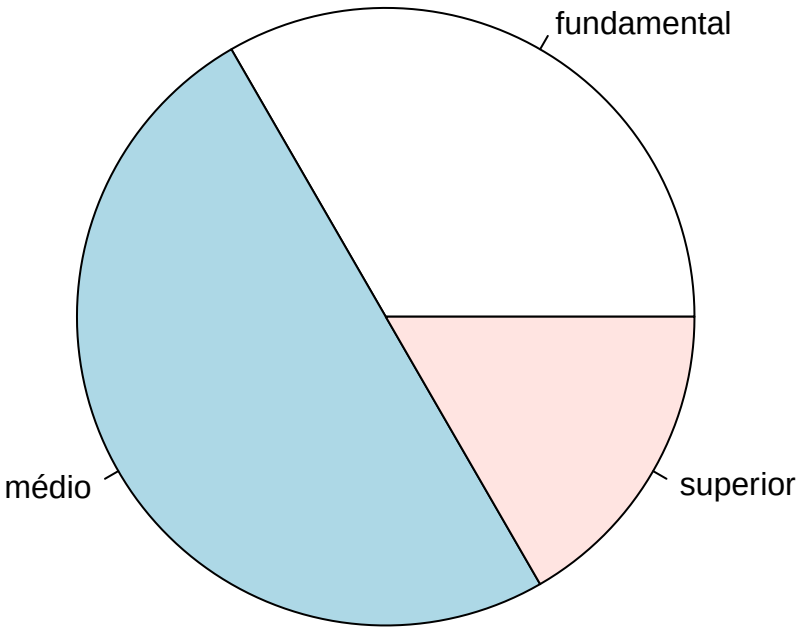
```
FreqRel <- prop.table(FreqAbs)
FreqRel
```

```
fundamental       médio    superior
  0.3333333   0.5000000   0.1666667
```

Note que a maioria dos entrevistados é casado, correspondendo a uma proporção de 0.5 para 1.

Um gráfico adequado para essa variável é o gráfico de setores:

```
pie(FreqAbs, main = "Distribuição do Grau de Instrução")
```



Como visto na distribuição de frequências o gráfico de setores apresenta o maior setor para o ensino médio e o menor para ensino superior.

Variável *Número de Filhos*

Um breve sumário sobre essa variável pode ser obtido pelo comando:

```
summary(CMB$Filhos)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    1.00    2.00    1.65    2.00    5.00      16
```

Note que o número de filhos varia de 0 à 5. O número médio de filhos é 1.65 com mediana igual à 2. Note que a moda do número de filhos é médio, havendo 18 funcionários com essa quantidade de filhos. Isso justifica o fato da mediana ser igual ao terceiro quartil nesse caso.

A seguir vemos a distribuição de frequências absolutas e relativa para a variável *Número de Filhos*:

```
FreqAbs <- table(CMB$Filhos)
FreqAbs
```

```
0 1 2 3 5
4 5 7 3 1
```

A grande maioria dos funcionários tem menos que 2 filhos. Note que apenas os filhos dos funcionários casados foram computados, por isso temos que a somatória do total de entrevistados não é 36.

```
sum(FreqAbs)
```

```
[1] 20
```

Há apenas 20 funcionários casados como observado anteriormente.

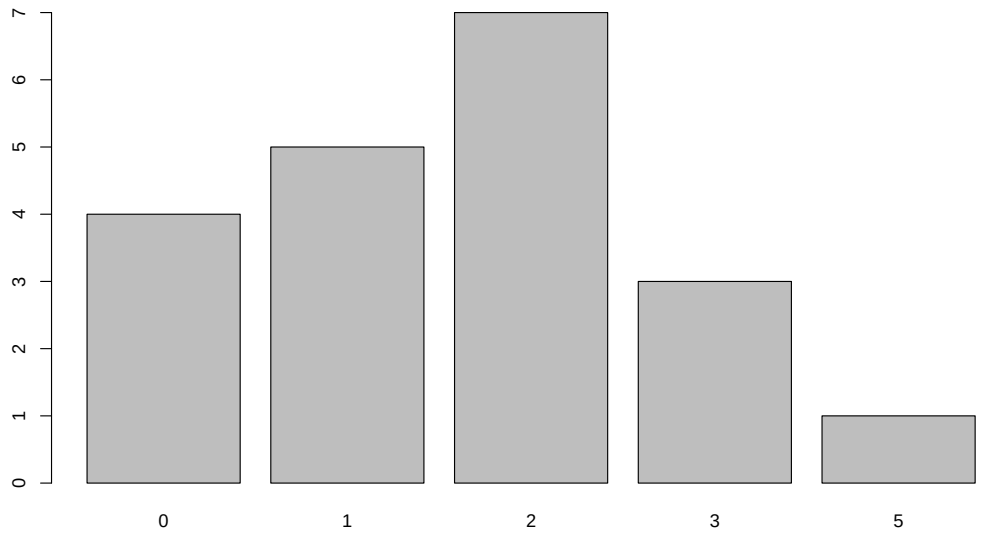
A distribuição de frequências relativas revela as proporções de funcionários que tem uma determinada quantidade de filhos:

```
FreqRel <- prop.table(FreqAbs)
FreqRel
```

```
   0    1    2    3    5
0.20 0.25 0.35 0.15 0.05
```

Um gráfico adequado para essa variável é o gráfico de barras:

```
barplot(FreqAbs)
```



Variável *Salário*

Um sumário preliminar dessa variável é obtido pelo comando

```
summary(CMB$Salário)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4000    7552   10160   11120   14060   23300
```

Note que os salários nessa companhia estão entre R$4000 e R$23000. A amplitude, o desvio interquartil e o desvio padrão são dados respectivamente por:

```
diff(range(CMB$Salário))
```

```
[1] 19300
```

```
IQR(CMB$Salário)
```

```
[1] 6507.5
```

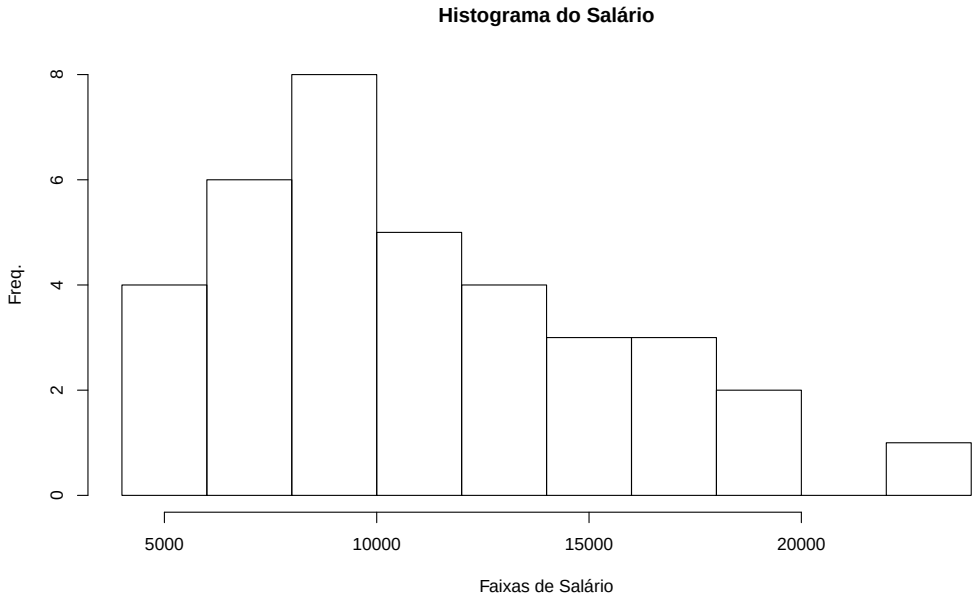
```
sd(CMB$Salário)
```

```
[1] 4587.458
```

Desse modo a distância do menor ao maior salário é de R$19300, a porção das 50% estatísticas de ordem centrais se dispersam num raio de R$6507,50 e em média os salários se afastam da média por aproximadamente R$4587,45.

Um gráfico adequado é o histograma:

```r
hist(CMB$Salário,
     main = "Histograma do Salário",
     ylab = "Freq.",
     xlab = "Faixas de Salário")
```



Pode-se observar que a faixa salarial de maior freqüência é de 9000 à 10000. Além disso a distribuição apresenta assimetria à direita. O coeficiente de assimetria é dado por:

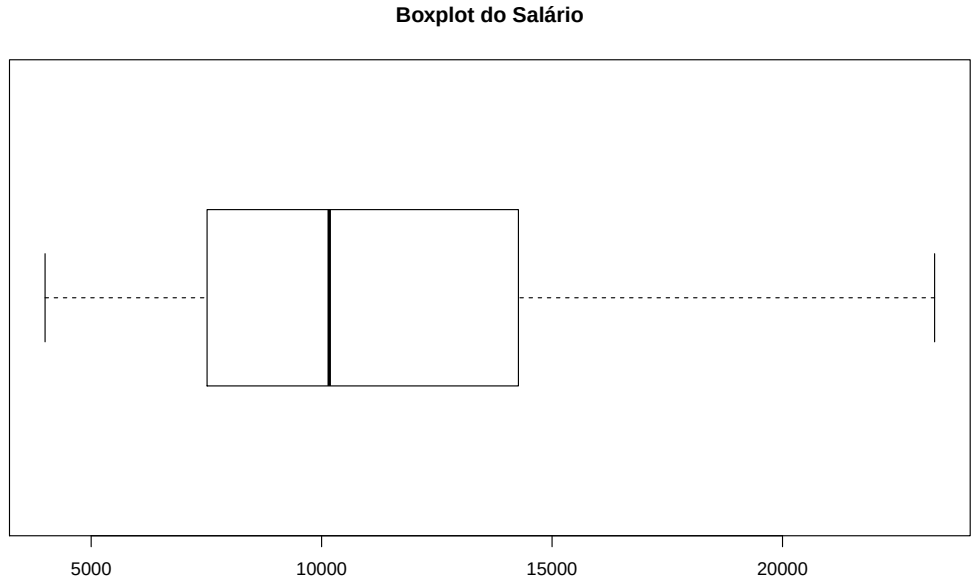
```r
skewness(CMB$Salário)
```

```
[1] 0.5997938
```

Como o coeficiente de assimetria é positivo, fica confirmada a assimetria à direita.

Um outro gráfico útil é o boxplot:

```r
boxplot(CMB$Salário,
        main = "Boxplot do Salário",
        horizontal = TRUE)
```



Veja que os dados não apresentam valores extremos. A distância entre o terceiro quartil e o máximo é próxima à distância do terceiro quartil ao mínimo indicando que as 25% maiores estatísticas de ordem se distribuem numa amplitude próxima que as 75% menores.

Variável *Idade*

Um sumário preliminar dessa variável é obtido pelo comando.

```
summary(CMB$Idade)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  20.83   30.67   34.92   35.05   40.52   48.92 
```

As idades dos funcionários variam de 20,83 à 48,92. Note que os valores da média e da mediana são muito próximos o que é uma característica de dados com distribuições simétricas. O primeiro e o terceiro quartis estão mais próximos da mediana do que do mínimo e do máximo respectivamente o que indica que os 50% das estatísticas de ordem centrais se concentram mais intensamente em torno da mediana.

A amplitude, o desvio interquartil e o desvio padrão são dados respectivamente por:

```
diff(range(CMB$Idade))
```

```
[1] 28.09
```

```
IQR(CMB$Idade)
```

```
[1] 9.8525
```

```
sd(CMB$Idade)
```

```
[1] 6.705046
```

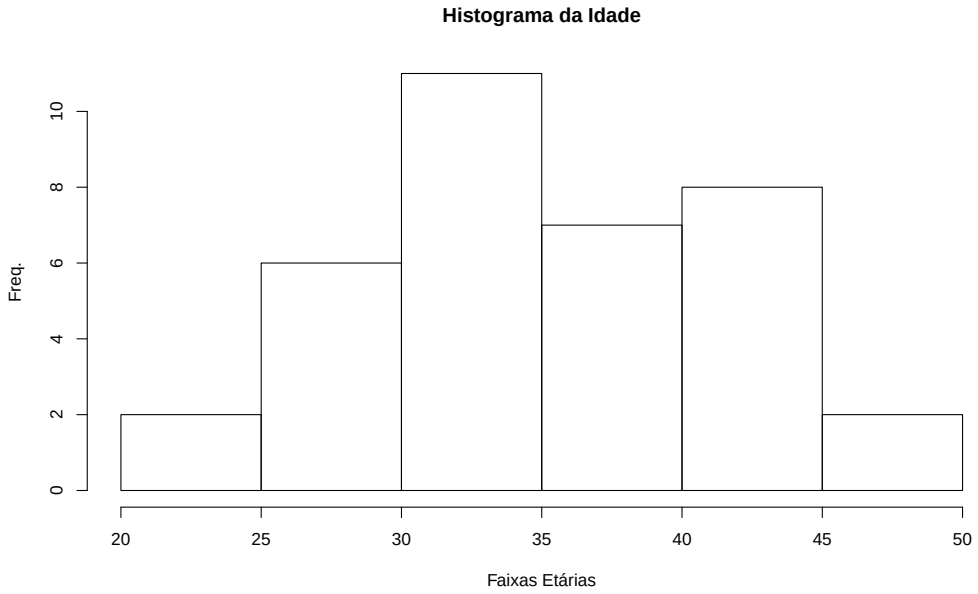
O funcionário mais velho da companhia tem 28,09 anos de idade a mais que o mais jovem, a porção das 50% estatísticas de ordem centrais das idades se dispersam num raio de 9,85 anos e em média as idades se afastam da média por aproximadamente 6,7 anos. O coeficiente de assimetria é muito próximo de zero, indicando que os dados se distribuem de forma aproximadamente simétrica em torno da média/mediana:

```
skewness(CMB$Idade)
```

```
[1] -0.03598018
```

O histograma dos dados confirma essa simetria:

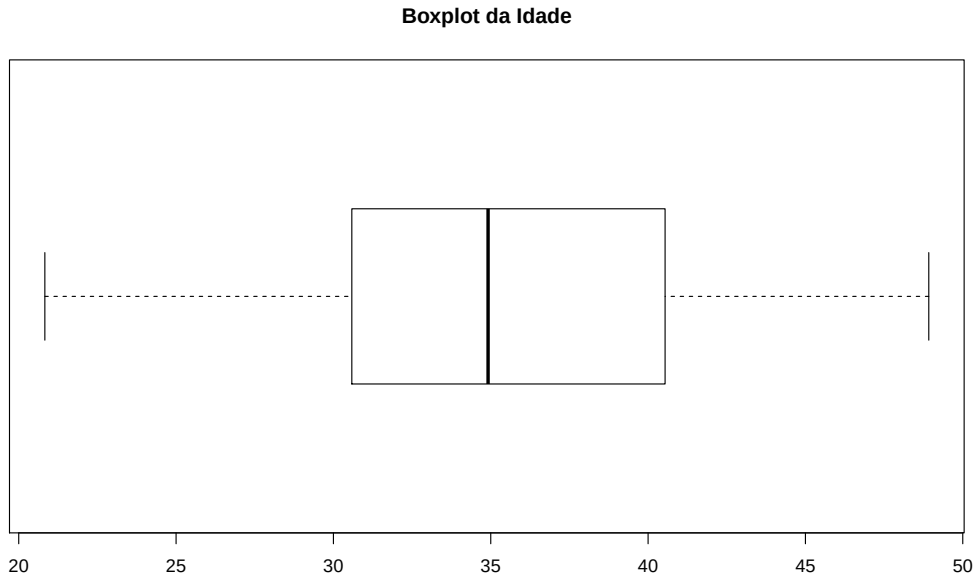
```
hist(CMB$Idade,
     main = "Histograma da Idade",
     ylab = "Freq.",
     xlab = "Faixas Etárias")
```



A faixa etária com a maior quantidade de funcionários é de 30 à 35 anos. Há apenas dois funcionários com menos de 25 anos e, simetricamente, apenas dois funcionários com mais de 45 anos.

O boxplot para essa variável é o seguinte:

```
boxplot(CMB$Idade,
        main = "Boxplot da Idade",
        horizontal = TRUE)
```



Como apontado anteriormente, o primeiro e o terceiro quartis estão mais próximos da mediana do que, respectivamente, do mínimo e do máximo indicando que as 50% estatísticas de ordem centrais estão muito concentradas em torno da mediana.

Variável *Região de Procedência*

A distribuição de frequências absolutas e relativas para a variável *Região de Procedência* são dadas respectivamente por:

```
FreqAbs <- table(CMB$Procedência)
FreqAbs
```

```
 capital interior    outra
      11       12       13
```

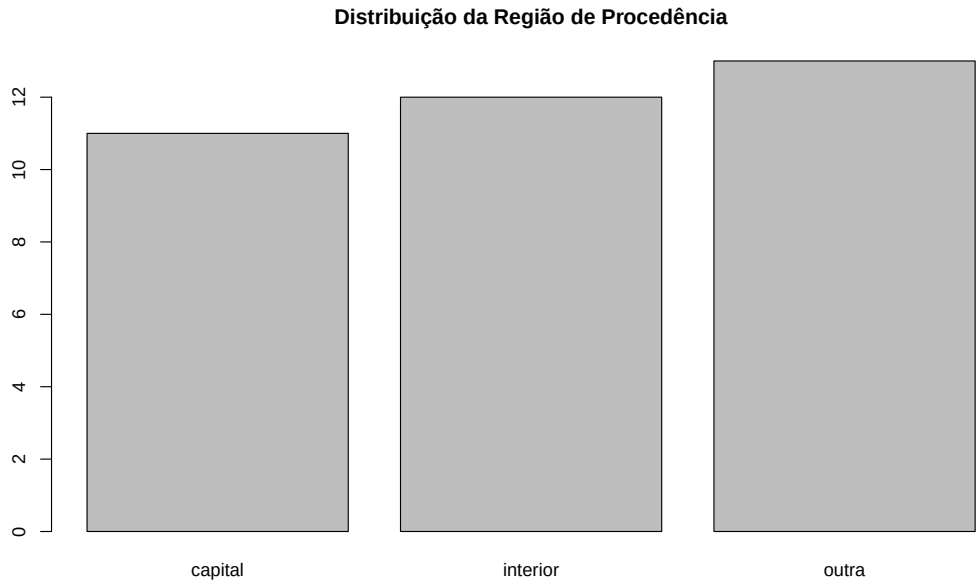
```
FreqRel <- prop.table(FreqAbs)
FreqRel
```

```
  capital  interior     outra
0.3055556 0.3333333 0.3611111
```

As proporções das três Regiões de Procedência são bastante próximas.

Um gráfico adequado para essa variável é o gráfico de barras:

```
barplot(FreqAbs, main = "Distribuição da Região de Procedência")
```



Como esperado as barras tem alturas similares. A *capital* é a região com menor número de representantes e *outra* é a com maior número de representantes.

Análise Bivariada

Variáveis *Grau de Instrução* e *Região de Procedência*

Uma tabela de contingência para essas variáveis é a seguinte:

```
TC <- table(CMB$Grau.de.Instrução,CMB$Procedência)
TC
```

```
             capital interior outra
  fundamental       4        3     5
  médio             5        7     6
  superior          2        2     2
```

Essa tabela apresenta a distribuição conjunta dessas variáveis. As distribuições marginais para cada variável podem ser encontradas nas seções de análise univariada.

A tabela de distribuição conjunta as proporções relativas ao total geral pode ser obtida da seguinte forma

```
prop.table(TC)
```

```
                capital   interior      outra
  fundamental 0.11111111 0.08333333 0.13888889
  médio       0.13888889 0.19444444 0.16666667
  superior    0.05555556 0.05555556 0.05555556
```

Aqui podemos obter várias informações como a de que 19.4% dos entrevistados tem ensino médio e são da capital, correspondendo à maior proporção obtida dentre todas as combinações possíveis. As menores proporções obtidas foram para os funcionários com nível superior, a distribuição das proporções de funcionários com nível superior foi uniforme com relação à região de procedência.

A distribuição das proporções relativas ao total das linhas pode ser obtido da seguinte forma

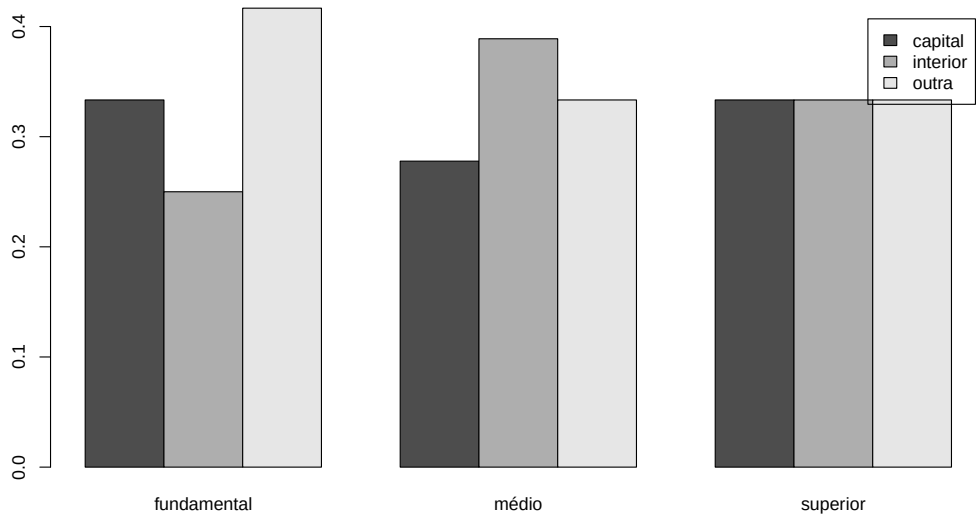
```
PTC1 <- prop.table(TC,margin = 1)
PTC1
```

```
               capital  interior     outra
  fundamental 0.3333333 0.2500000 0.4166667
  médio       0.2777778 0.3888889 0.3333333
  superior    0.3333333 0.3333333 0.3333333
```

Cada linha dessa tabulação apresenta a proporção de pessoas com um determinado nível de instrução em cada uma das três regiões de procedência. Podemos notar que dentre as pessoas com ensino fundamental a maioria provem de outra região (41,67%), dentre as pessoas com ensino médio a maioria provém do interior (38,89%) e as pessoas com ensino superior se distribuem de modo uniforme nas três regiões de procedência consideradas.

Essa tabela pode ser representada por meio de um gráfico de barras:

```
barplot(t(PTC1),legend.text = colnames(PTC1),beside = T)
```



Note que as barras das regiões de procedência apresentam alturas próximas para cada um dos graus de instrução indicando pouca evidência sobre a associação entre as variáveis *Região de Procedência* e *Grau de Instrução*.

A distribuição das proporções relativas ao total das linhas pode ser obtido da seguinte forma

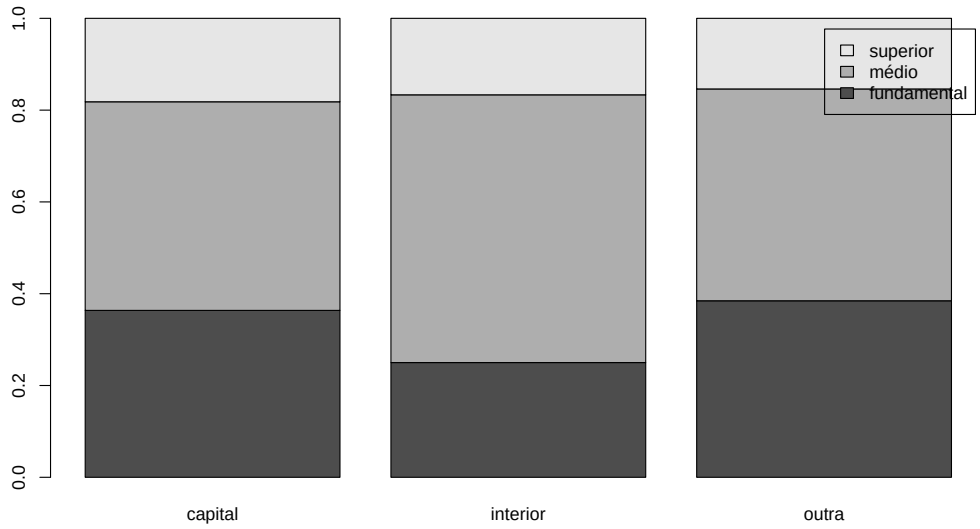
```
PTC2 <- prop.table(TC,margin = 2)
PTC2
```

```
               capital  interior     outra
fundamental 0.3636364 0.2500000 0.3846154
médio       0.4545455 0.5833333 0.4615385
superior    0.1818182 0.1666667 0.1538462
```

Cada coluna dessa tabulação apresenta a proporção de pessoas provenientes de uma determinada região em cada um dos três graus de instrução considerados. Podemos notar que em todas as regiões de procedência a proporção de pessoas com ensino médio foi sempre a maior: 45,45% na capital, 58,33% no interior e 46,15% em outras regiões.

Essa tabela também pode ser representada por meio de um gráfico de barras empilhadas:

```
barplot(PTC2,legend.text = rownames(PTC2))
```



Note, novamente, que as barras tem proporções similares entre si, o que indica que as variáveis consideradas podem não ter associação.

O coeficiente Qui-quadrado pode ser obtido da seguinte forma:

```
TesteQui2 <- summary(TC)
CoefQui2 <- TesteQui2$statistic
CoefQui2
```

```
[1] 0.6614219
```

O coeficiente de contingência pode ser obtido da seguinte forma:

```
CoefC <- sqrt(CoefQui2/(CoefQui2+sum(TC)))
CoefC
```

```
[1] 0.1343181
```

Finalmente, o coeficiente de contingência T pode ser obtido da seguinte forma:

```
CoefT <- sqrt(CoefQui2/(sum(TC)*(nrow(TC)-1)*(ncol(TC)-1)))
CoefT
```

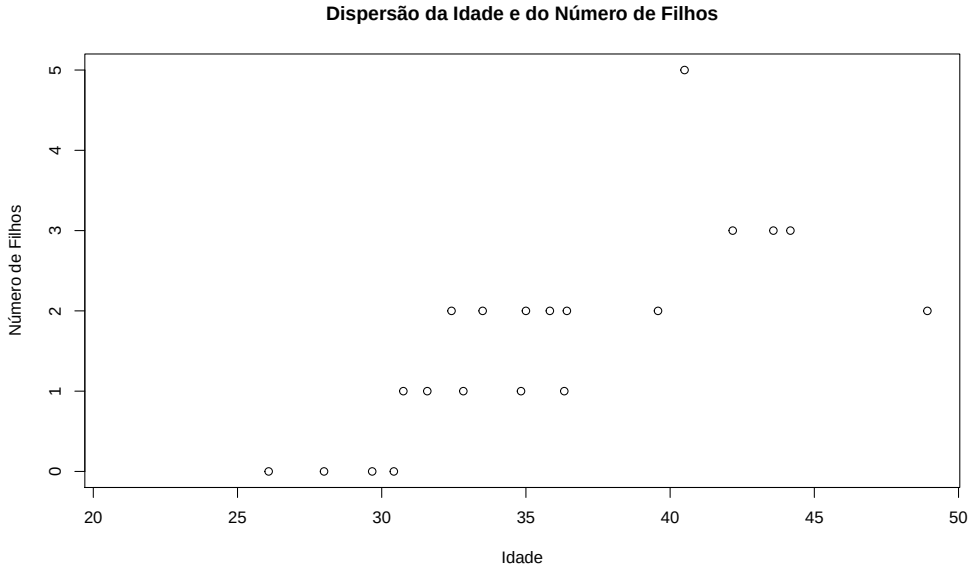
```
[1] 0.06777321
```

Os três coeficiente apresentam valores muito próximos de zero, o que indica um grau bastante fraco de associação entre as variáveis *Região de Procedência* e *Grau de Instrução*.

Variáveis *Idade* e *Número de Filhos*

Note que o gráfico de dispersão entre as variáveis *Idade* e *Número de Filhos* indica uma dependência linear direta entre as duas variáveis, uma vez que os dados aparentam se agrupar em torno de uma reta crescente.

```
plot(x = CMB$Idade,
     y = CMB$Filhos,
     xlab = "Idade",
     ylab = "Número de Filhos",
     main = "Dispersão da Idade e do Número de Filhos")
```



O coeficiente de correlação entre essas duas variáveis é relativamente bem próximo de um o que indica uma correlação linear direta entre as duas variáveis.

```
cor(x = CMB$Filhos, y = CMB$Idade, use = "complete.obs")
```

```
[1] 0.740927
```

Note que a covariância amostral entre as duas variáveis é dada por:

```
cov(x = CMB$Filhos, y = CMB$Idade, use = "complete.obs")
```

```
[1] 5.587526
```

A covariância é relativamente próxima de zero, indicando uma baixa dispersão dos dados.

A reta de regressão que minimiza os erros quadráticos médios é obtida pelo comando

```
reta <- lm(formula = Filhos~Idade, data = CMB)
reta$coefficients
```

```
(Intercept)       Idade
 -3.9784986   0.1579752
```

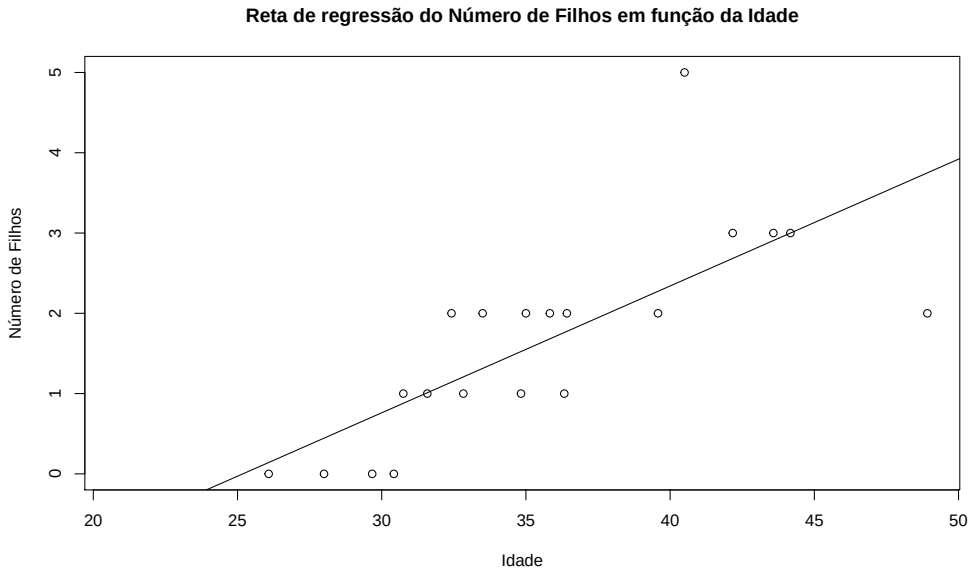
Essa reta terá a seguinte equação

$$\hat{y} = 0,158x - 3,978.$$

onde x é um valor para a idade e $\hat{y}$ é o correspondente valor predito pelo modelo linear para o número de filhos.

```
plot(x = CMB$Idade,
     y = CMB$Filhos,
     xlab = "Idade",
```

```
     ylab = "Número de Filhos",
     main = "Reta de regressão do Número de Filhos em função da Idade")
abline(reta)
```



Para usarmos o modelo para predizer o valor médio do número de filhos para as idades de 30, 35 e 40 anos basta:

```
PredizerFilhos <- data.frame(Idade = c(30,35,40))
PredicoesObtidas <- predict(reta,PredizerFilhos)
Sumario <- data.frame(PredizerFilhos,PredicoesObtidas)
Sumario
```

```
  Idade PredicoesObtidas
1    30        0.7607576
2    35        1.5506336
3    40        2.3405096
```

Portanto, em média os funcionários com 30 anos possuem 0,76 filhos, com 35 anos os funcionários tem em média 1,55 filhos e com 40 anos de idade os funcionários acumulam uma média de 2,34 filhos. Note agora que as idades dos funcionarios estão no intervalo:

```
range(CMB$Idade)
```

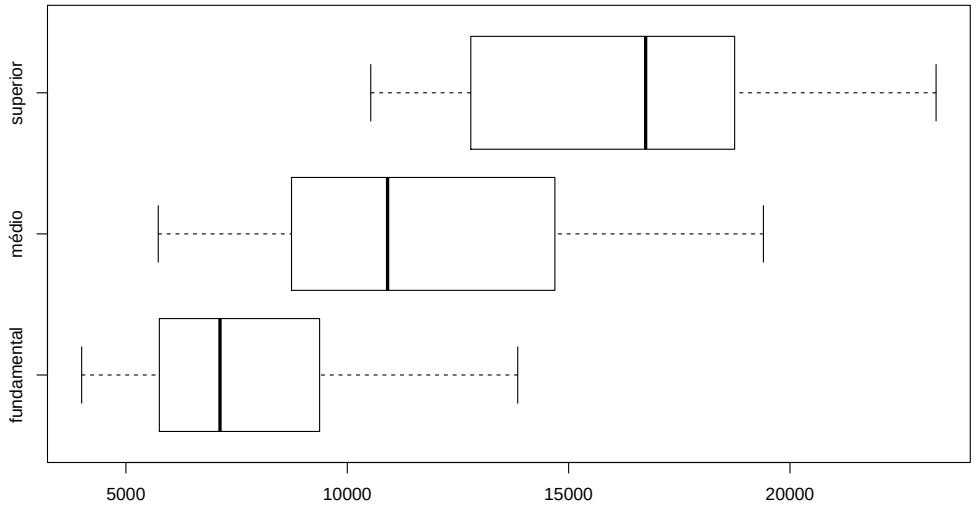
```
[1] 20.83 48.92
```

Portanto, não podemos usar esse modelo para predizer o número de filhos dos funcionários usando idades fora desse intervalo.

Variáveis *Grau de Instrução* e *Salários*

Um gráfico adequado para analisar esse par de variáveis é o seguinte:

```
boxplot(Salário~Grau.de.Instrução, data = CMB, horizontal = TRUE)
```



O gráfico deixa bem claro que a distribuição da variável *Salários* muda significativamente em função do *Grau de Instrução*.

Algumas medidas resumo para a variável *Salários* segundo o *Grau de Instrução*. podem ser obtidas pelo comando:

```
aggregate(Salário~Grau.de.Instrução, FUN = summary, data=CMB)
```

```
  Grau.de.Instrução Salário.Min. Salário.1st Qu. Salário.Median
1       fundamental         4000            6008           7125
2             médio         5730            8838          10910
3          superior        10530           13650          16740
  Salário.Mean Salário.3rd Qu. Salário.Max.
1         7837            9162        13850
2        11530           14420        19400
3        16480           18380        23300
```

Todas as estatísticas aumentam consideravelmente a medida em que o grau de instrução se eleva indicando, mais uma vez, uma clara associação entre essas variáveis. O mesmo ocorre com os valores da variância populacional

```
Variancias <- aggregate(Salário~Grau.de.Instrução, FUN = pvar, data=CMB)
Variancias
```

```
  Grau.de.Instrução  Salário
1       fundamental  8012289
2             médio 13035503
3          superior 16893292
```

Para calcular o valor do coeficiente de explicação entre essas variáveis basta:

```
pesos <- tabulate(CMB$Grau.de.Instrução)
1 - weighted.mean(x = Variancias$Salário,w = pesos)/pvar(CMB$Salário)
```

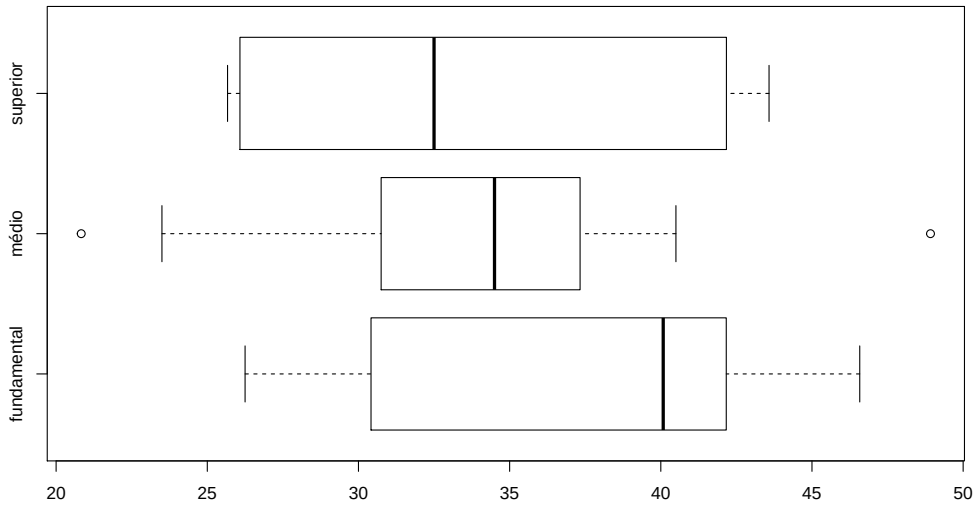
```
[1] 0.4132966
```

O valor desse coeficiente indica que 41,33% da variação total da variável *Salários* é explicada pela variável *Grau de Instrução*.

Variáveis *Grau de Instrução* e *Idade*

Um gráfico adequado para analisar esse par de variáveis é o seguinte:

```
boxplot(Idade~Grau.de.Instrução, data = CMB, horizontal = TRUE)
```



O gráfico mostra que as distribuições da variável *Idade* não são iguais para os diferentes valores da variável *Grau de Instrução*.

Algumas medidas resumo para a variável *Idade* segundo o *Grau de Instrução*. podem ser obtidas pelo comando:

```
aggregate(Idade~Grau.de.Instrução, FUN = summary, data=CMB)
```

```
  Grau.de.Instrução Idade.Min. Idade.1st Qu. Idade.Median Idade.Mean
1       fundamental      26.25         31.62        40.08      37.30
2             médio      20.83         30.96        34.50      33.98
3          superior      25.67         27.42        32.50      33.75
  Idade.3rd Qu. Idade.Max.
1         41.58      46.58
2         37.08      48.92
3         40.02      43.58
```

Em geral as estatísticas correspondentes divergem bastante para os diferentes valores da variável *Grau de Instrução*, o que nos permite afirmar que existe algum grau de dependência entre essas variáveis. O mesmo ocorre com os valores da variância populacional:

```
Variancias <- aggregate(Idade~Grau.de.Instrução, FUN = pvar, data=CMB)
Variancias
```

```
  Grau.de.Instrução    Idade
1       fundamental 44.19541
2             médio 36.35387
3          superior 49.51640
```

Para calcular o valor do coeficiente de explicação entre essas variáveis basta:

```
pesos <- tabulate(CMB$Grau.de.Instrução)
1 - weighted.mean(x = Variancias$Idade,w = pesos)/pvar(CMB$Idade)
```

```
[1] 0.05827983
```

O valor desse coeficiente indica que 5,83% da variação total da variável *Idade* é explicada pela variável *Grau de Instrução*. Logo há uma maior relação entre as variáveis *Salários* e *Grau de Instrução* do que

entreas variáveis *Idade* e *Grau de Instrução*.

Conclusão

???