Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Exercícios de Estatística para Ciências Sociais Utilizando R Commander

Monitoria PMG
Maria Eduarda da Silva Vilas Bôas
Pedro Henrique Chaves Cardoso
Pedro Henrique de Matos Araujo
e Adrian Luna(orientador)

Conteúdo

1 Introdução **2**

2 Amostragem **3**

3 Estatística Descritiva **11**

4 Probabilidade **20**

5 Regressão Linear **27**

6 Mineração de Textos **38**

Capítulo 1

Introdução

O filósofo Lakatos já dizia que princípio básico para verificação científica é a base matemática, e nas ciências sociais não é diferente. Utilizando princípios lógicos intuitivos com base estatística é possível verificar inúmeras informações e não somente isto, com a base dessa ferramenta poderosa podemos interpretar inúmeros dados importantes. Como saber se a taxa de mortalidade infantil difere por classe social? Como saber se a proporção de desempregados é preocupante para o país? Essas e outras perguntas podem ser muito bem respondidas com um bom auxílio da estatística descritiva e inúmeros outros métodos. Mesmo sendo uma ciência trabalhosa que por muitas vezes tem uma grande quantidade de cálculos, podemos utilizar várias ferramentas poderosas como programação para facilitar o aprendizado.

Pensando nisso, este material foi elaborado com o objetivo de elucidar e apresentar uma ferramenta do R mais amigável ao usuário, que dipensa o conhecimento prévio sobre programação. Para tanto, preparamos resoluções de alguns exercícios, sendo eles de autoria própria e outros retirados da apostila "Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais". Utilizando o pacote "Rcmdr" do software R, que em si é uma versão do R mais simples para sintetizar análises. Ao abrir o pacote, teremos a interface do R commander possibilitando vários tipos de análise. Dessa forma, trouxemos esses exercícios de forma a exemplificar e mostrar na prática como fazer algumas das análises usadas pelas Ciências Sociais utilizando essa extensão gráfica do R.

Capítulo 2

Amostragem

A amostragem pode ser entendida como um conjunto de métodos estatísticos para, como o nome já presupõe, retirarmos uma amostra do conjunto universo de dados, que consiste em uma parcela menor, porém ainda sim significativa desse conjunto total. Algumas formas de realizar amostragem que são abarcadas pelo R Commander são as seguintes:

1. Amostragem Aleatória: amostra onde cada um dos elementos da população possuem a mesma probabilidade de serem selecionados na amostra;

2. Amostragem Estratificada: consiste em dividir e estratificar a população em subgrupos que não se sobrepõem e então extrair uma amostra de cada estrato. Esse tipo, de amostragem é muito utilizado para aplicar diferentes métodos de coleta de dados para diferentes partes da população.

Utilizando Ferramenta de Amostragem no Rcmdr

Considerando a priori que já tenha instalado o pacote "RcmdiPlugin.sampling", vá à barra de comando e clique em Ferramentas → carregar plug-in(s) do Rcmdr → OK. Como mostrado nas próximas figuras, esse passo inicial irá carregar o pacote para assim ser possível a sua utilização.

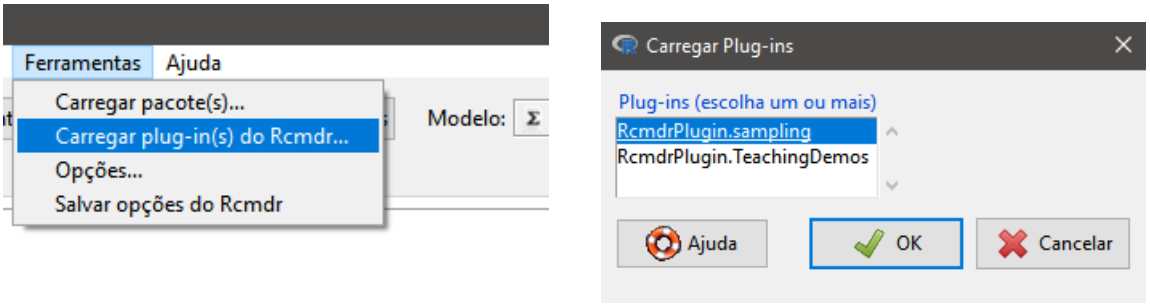


Figura 2.1: Seguindo esse caminho, devemos configurar as opções da forma mostrada

Note que a barra de comando mudou. Além disso, haverá uma mensagem para confirmação de reinício do R commander para a utilização do plugin. Para exemplificar, utilizaremos um banco de dados desse mesmo plugin, seguindo os passos abaixo:

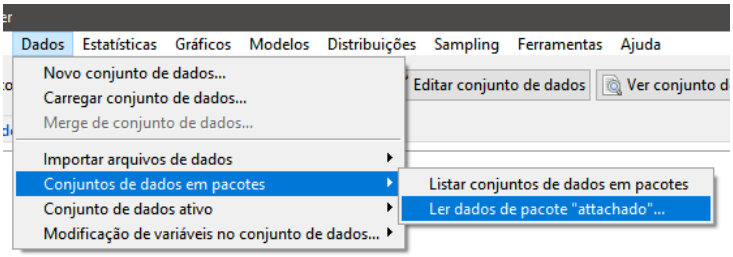


Figura 2.2: Seguindo os passos

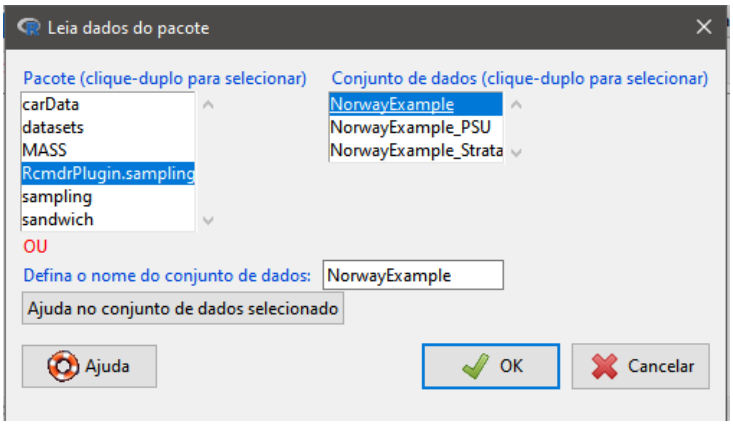


Figura 2.3: Seguindo os passos

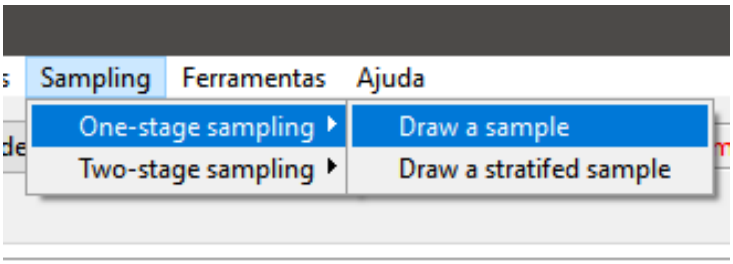


Figura 2.4: Seguindo os passos

Após a reprodução dos passos acima, o R Commander solicitará na tela para que insira o tamanho da sua nova amostra e dê "ok". A mérito de exemplo, foi escolhido o tamanho de amostra "10", como mostrado abaixo:

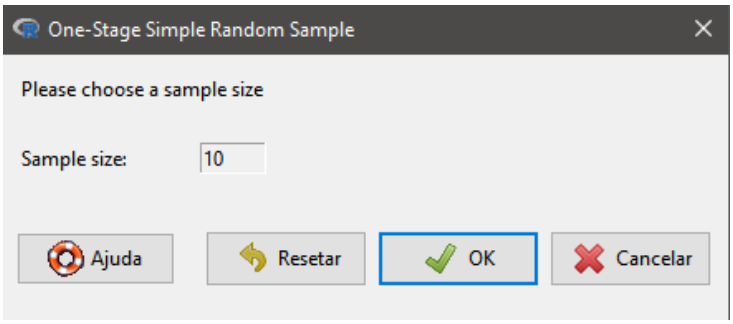


Figura 2.5: Escolhendo o tamanho de amostra

Ao dar "ok" no tamanho da nova amostra, os códigos para execução aparecerão no script, como vê-se abaixo:

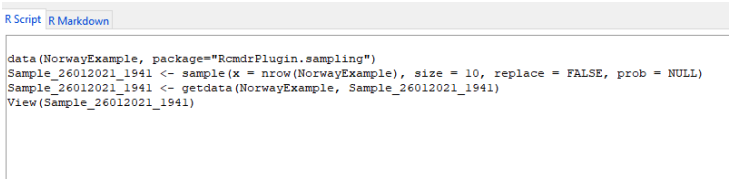


Figura 2.6: Códigos

Podemos verificar o novo conjunto de dados nomeado, nesse caso, "Sample_26012021_1941" no RStudio, o banco de dados estará aberto, porém nossa reamostragem possui um problema no código gerado por esse pacote. Para corrigir esse erro, basta colocar o seguinte código na aba RScript e submeter as 3 linhas, separadamente:

Sample_26012021_1941 ≪- sample(x = nrow(NorwayExample), size = 10, replace = FALSE, prob = NULL)
Sample_26012021_1941 ≪- data.frame
View(Sample_26012021_1941)

É importante lembrar que o cursor deve estar no início da linha ou na linha anterior para submeter o código corretamente. Além disso, devemos clicar no botão submeter algumas vezes, pois ele roda cada linha de forma individual.

Caso haja dúvida quanto o código acima , "size" será o tamanho da reamostragem, "replace" se deseja substituir a amostra (caso sim (TRUE), caso não (FALSE)) e prob é a probabilidade de algum elemento especial ser escolhido. Como fizemos um evento aleatório, a probabilidade utilizada foi "NULL" (nula).

Exercício 2.1

Utilizando o banco de dados AirPassangers do pacote "datasets", realize uma amostragem de tamanho 15, de forma aleatória, ou seja $n = 15$. Após isso calcule o que se pede: (Utilize nível de significância de 5%)

1. Realize um teste de normalidade utilizando o método Shapiro-Wilk na população e comente;
2. Realize um teste de normalidade utilizando o método de Shapiro-Wilk na amostra coletada e comente;
3. Comente a diferença encontrada, caso ela exista.

Solução

1- Primeiramente, devemos carregar o banco de dados citado, para tanto temos:

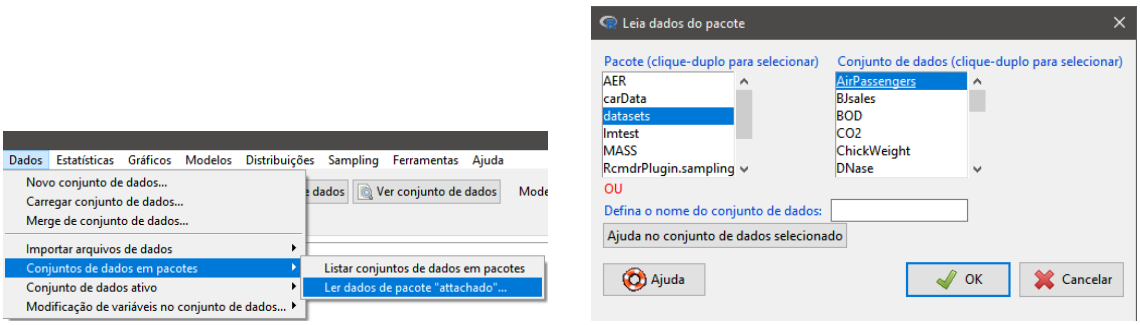


Figura 2.7: Siga os passos acima

Agora, devemos fazer o teste de normalidade pedido no enunciado. Para isto, vá à barra de comandos e clique em Estatística → Resumos → Test of normality, como mostrado abaixo, seguindo o teste de hipóteses $H0 : p-\text{valor} \geq 0.05$ ou $H1 : p-\text{valor} < 0.5$

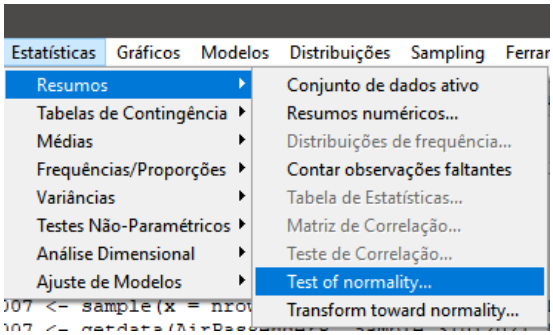


Figura 2.8: Siga os passos

```
> normalityTest(~x, test="shapiro.test", data=AirPassengers)

        Shapiro-Wilk normality test

data:  x
W = 0.95196, p-value = 0.00006832
```

Figura 2.9: Obtemos o resultado

Comentário: A partir do teste acima, com um poder (w) de 0.95196 e um p-valor extremamente pequeno, quando comparado a um nível de significância de 5%, logo rejeita-se a normalidade (H0). Portanto, não há evidências que a amostra segue uma distribuição normal.

2- Agora, devemos realizar a amostragem da nossa população e o mesmo teste de normalidade do item anterior:

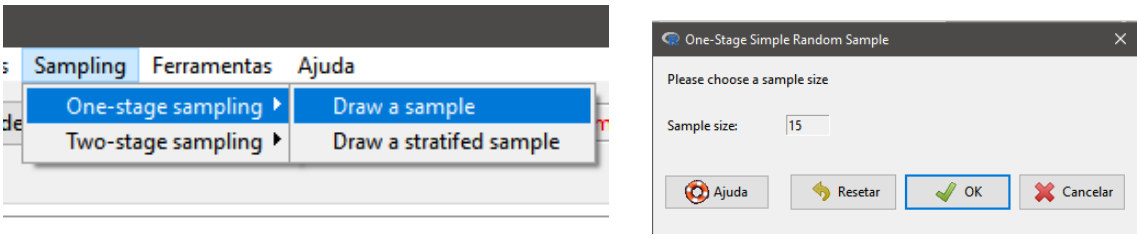


Figura 2.10: Siga os passos acima

Devemos ressaltar que o código apresenta uma falha, então para isso realiza-se uma correção que está mais bem explicada no "Apêndice". Assim, é necessário alterar o código na aba do RScript da seguinte forma:

```
Sample_31012021_1020 <- sample(x = nrow(AirPassengers), size = 20, replace = FALSE, prob = NULL)
Sample_31012021_1020 <- data.frame(AirPassengers[Sample_31012021_1020, ])
View(Sample_31012021_1020)
```

Figura 2.11: Siga os passos acima

Agora, deve-se submeter todas as 3 linhas de código e, para que sejam submetidas corretamente, mantenha o cursor no início de cada linha. Posteriormente, selecione a amostra gerada, clicando em "AirPassangers":

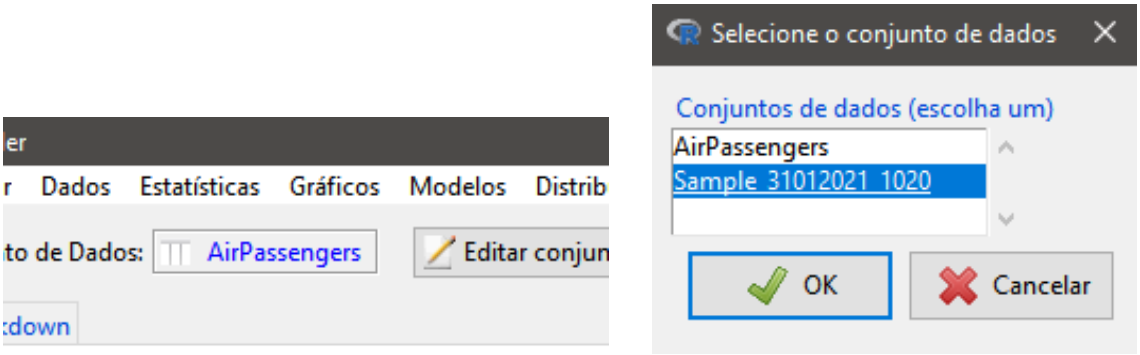


Figura 2.12: Siga os passos acima

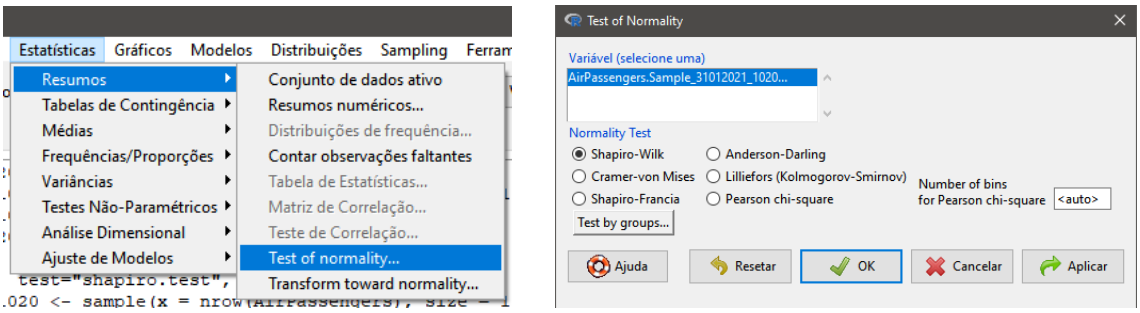


Figura 2.13: Siga os passos acima

```
> normalityTest(~AirPassengers.Sample_31012021_1020..., test="shapiro.test", data=Sample_31012021_1020)

        Shapiro-Wilk normality test

data:  AirPassengers.Sample_31012021_1020...
W = 0.92456, p-value = 0.1213
```

Figura 2.14: Obtemos o resultado

Comentário: A partir do teste acima com um poder (w) de 0,92456 e um p-valor alto quando comparado a um nível de significância de 5%, logo não rejeita-se a normalidade (H0), portanto há evidências que a amostra segue uma distribuição normal.

3- A diferença encontrada é que no segundo teste, quando feito com uma amostra bem maior se comparado ao primeiro teste, aceita-se a normalidade, o que já era esperado pelo Teorema Central do Limite (grandes amostras tendem a uma distribuição Normal), enquanto, por outro lado, a primeira amostra adquirida de forma aleatória, mas com o tamanho relativamente baixo, rejeitou a normalidade.

Exercício 2.2

Utilizando banco de dados "airquality" do pacote "datasets", realize uma amostragem de tamanho n = 25, em que cada mês terá 5 amostras. Uma vez feita essa amostra, calcule a média e o desvio padrão da amostra das demais variáveis:

Solução

Primeiramente, devemos carregar o banco de dados citado:

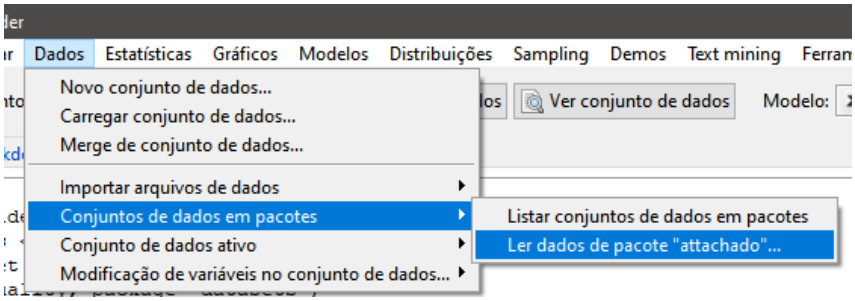


Figura 2.15: Siga o Caminho Mostrado Acima

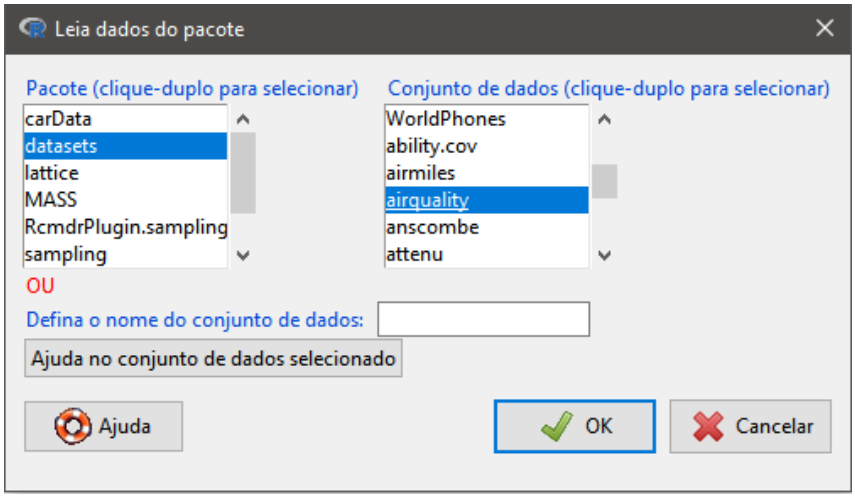


Figura 2.16: Selecione o Banco de Dados

Com os dados já carregados, devemos agora amostrar de forma estratificada, utilizando os meses como base. Para isso, siga os passos das figuras a seguir:

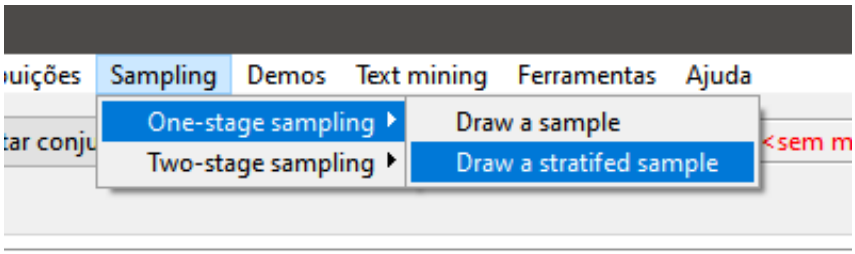


Figura 2.17: Selecione a Amostra Estratificada

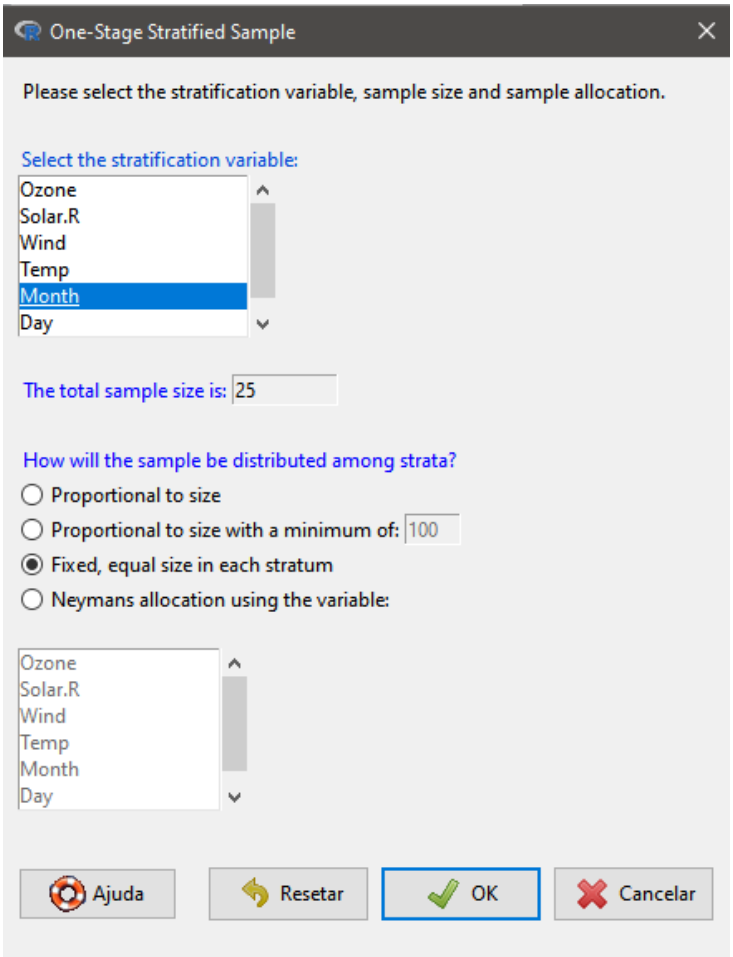


Figura 2.18: Insira a Caixa de Opções da Seguinte

Agora, devemos calcular o que se pede, a média e o desvio padrão da amostra:

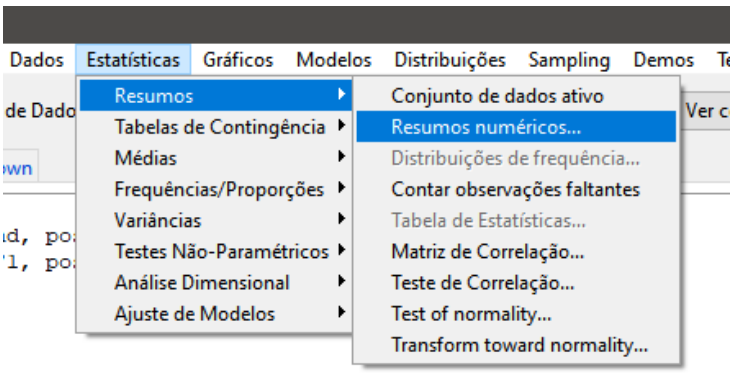


Figura 2.19: Agora Selecione os Resumos Numéricos

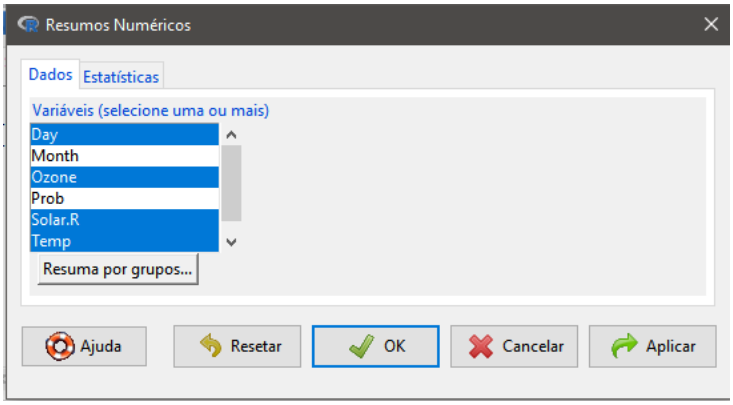


Figura 2.20: Selecione Todas as Variáveis com Exceção Month e Prob

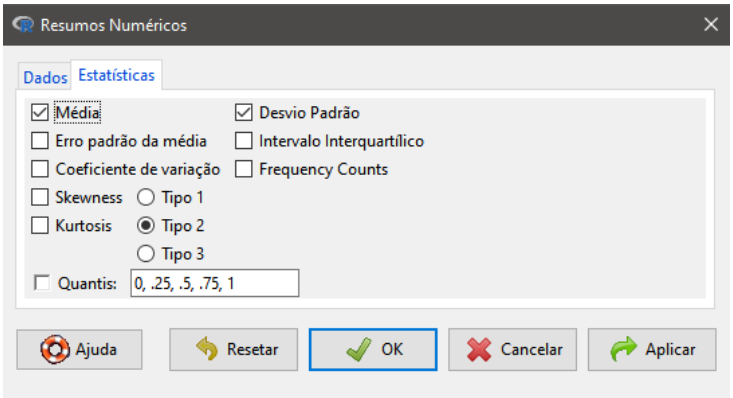


Figura 2.21: Agora Selecione os Dados ao Serem Calculados

```
> numSummary(Sample_06022021_1549[,c("Day", "Ozone", "Solar.
         mean   sd  n NA
Day      15.3  9.7 25  0
Ozone    61.7 49.7 18  7
Solar.R 174.8 94.0 22  3
Temp     78.9  9.5 25  0
Wind      8.7  4.2 25  0
```

Figura 2.22: Os Valores Pedidos estão na Tabela Gerada pelo R

Capítulo 3

Estatística Descritiva

A estatística descritiva aplica várias técnicas com o objetivo de descrever, organizar e sumarizar dados sobre a população. Para descrever um conjunto de dados, são usadas medidas de tendência central e medidas de variabilidade.

As medidas de tendência central incluem:

1. **Média:** é entendida como o valor que demonstra a concentração dos dados de uma distribuição; como a soma dos valores dividido pelo número de dados apresentados.
2. **Mediana:** é o valor que separa a metade maior e a metade menor de uma amostra ou população; o valor do meio de um conjunto de dados.
3. **Moda:** é o valor que ocorre com maior frequência ou o valor mais comum em um conjunto de dados.

As medidas de variabilidade incluem:

1. **Amplitude Total:** é uma medida simples, mas não muito adequada pois só considera os extremos do conjunto de dados, representada pela diferença entre o valor máximo e o valor mínimo.
2. **Variância:** é uma medida de dispersão que indica o quão longe, em geral, os valores se encontram do valor esperado.
3. **Desvio Padrão:** é a raiz quadrada da variância.

Dependendo do volume de dados, torna-se complexo tirar interpretá-los e, assim, coloca-los em uma ordem (rol) seria um facilitador para identificar valores maiores e menores, ou mesmo a concentração destes. Assim, temos:

1. **Frequência Absoluta:** o número de vezes em que uma determinada variável assume um valor.
2. **Frequência Relativa:** a razão entre a frequência absoluta e o número total de observações.
3. **Frequência Absoluta Acumulada:** soma da frequência absoluta do valor considerado às frequências absolutas anteriores a este valor.
4. **Frequência Relativa Acumulada:** soma da frequência relativa do valor considerado às frequências relativas anteriores a este valor.

O **Histograma** é uma representação gráfica para a organização e apresentação da distribuição de dados numéricos, exibindo a frequência em que uma determinada amostra de dados ocorre. A área da barra representa a frequência relativa da classe, de forma que a soma da área de todas as barras soma 100%.

Exercício 3.1 (3.0.1 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

A idade dos 10 ingressantes num certo ano no curso de pós-graduação em jornalismo de uma universidade foi a seguinte:

22,23,23,25,21,24,20,23,20,21

1. Ache as frequências absolutas, relativas, e relativas acumuladas.
2. Grafique o histograma.
3. Ache a média, moda, mediana, variância e desvio padrão da idade.
4. Construa o boxplot.

Solução

1. Levando a priori que já tenha baixado R ou R Studio, tenha instalado o pacote R Commander e o tenha aberto a solução será da seguinte forma:

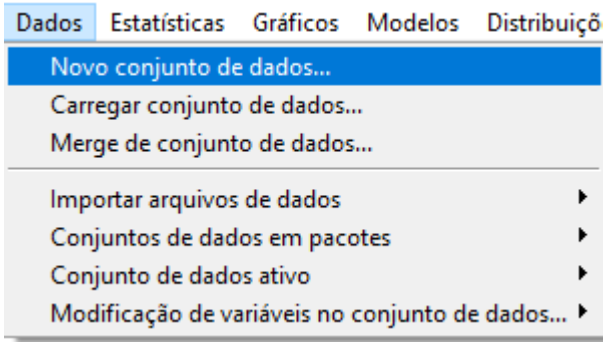


Figura 3.1: Inserindo Banco de Dados

Logo após digite o nome do seu conjunto de dados, e clique em ok. Como sugestão sempre defina um nome objetivo, para assim facilitar um possível acesso futuro.

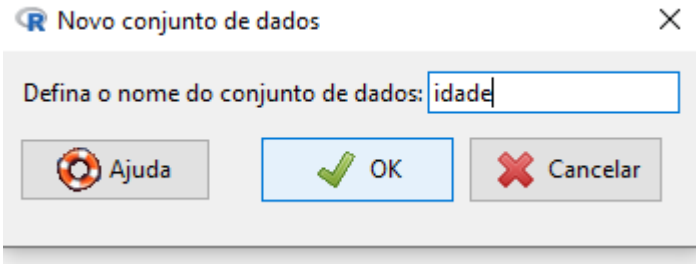


Figura 3.2: Nomear o Banco de Dados

Agora insira os dados, sendo a primeira linha os títulos das colunas, ao dar enter no teclado, irá adicionar uma linha automaticamente. Quando finalizar a inserção dos dados digite no Script o nome do arquivo e clique em submeter, assim aparecerá os seus dados como pode-se observar na figura abaixo:

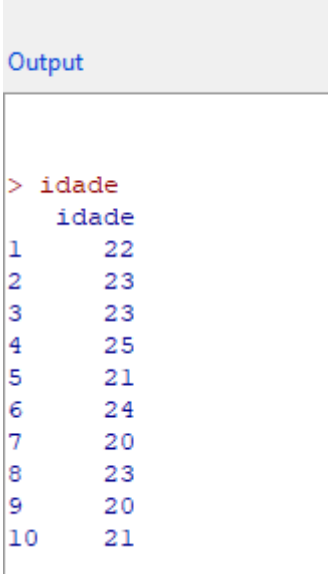


Figura 3.3: Banco de Dados

Com a confirmação dos dados agora é possível fazer análises estatísticas, conforme a figura abaixo:

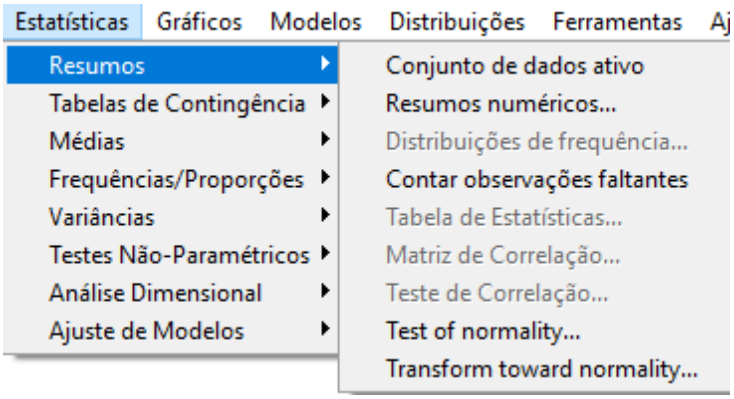


Figura 3.4: Vá em Estatística → Resumos → Resumos numéricos

Em seguida abrirá uma interface e você terá que selecionar quais medidas são requisitadas na questão:

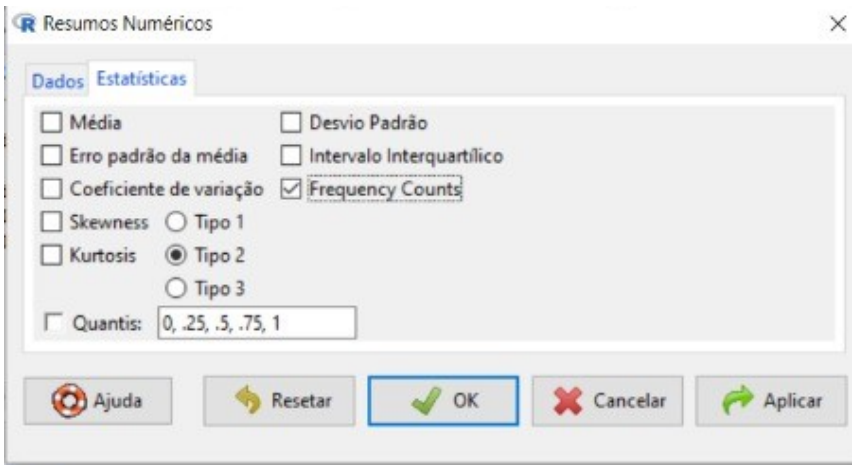


Figura 3.5: Primeiramente, marque apenas a opção “Frequency Counts”

Abaixo, é possível verificar algumas medidas como a frequência absoluta (count) e frequência relativa (percent):

```
> library(abind, pos=16)

> library(e1071, pos=17)

> discreteCounts(idade[,"idade", drop=FALSE])
Distribution of idade
      Count Percent
20        2      20
21        2      20
22        1      10
23        3      30
24        1      10
25        1      10
Total    10     100
```

Figura 3.6: Resultado das Frequências

Contudo, o R commander não disponibiliza de uma ferramenta para a grafia da frequência relativa acumulada, portanto, podemos calcular esses valores manualmente ou através do código abaixo:

cumsum(prop.table(table(idade)))

Demos à nossa base de dados o nome de "idade". Para rodar o código, basta digitá-lo na seção R Script e com o cursor ainda nessa linha clicar em submeter. Teríamos, então, a seguinte saída:

```
Output

> cumsum(prop.table(table(idade)))
 20  21  22  23  24  25
0.2 0.4 0.5 0.8 0.9 1.0
```

Figura 3.7: Frequência Relativa Acumulada

2. Com o conjunto de dados já identificado, clique na aba em gráficos e selecione histograma como na figura abaixo . Após dar enter, o histograma será gerado.

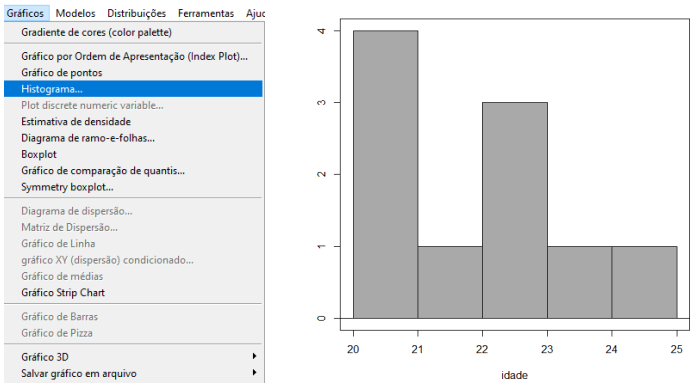


Figura 3.8: Seleção da opção Histograma e gráfico gerado

Devemos ressaltar que o Histograma gerado considera como a primeira classe do 20 ao 21 com intervalo

fechado, ou seja, conta-se tanto os valores de 20 quanto de 21 na geração desse gráfico.

3. Agora, marque as opções "Média", "Desvio Padrão" e "Quantis":

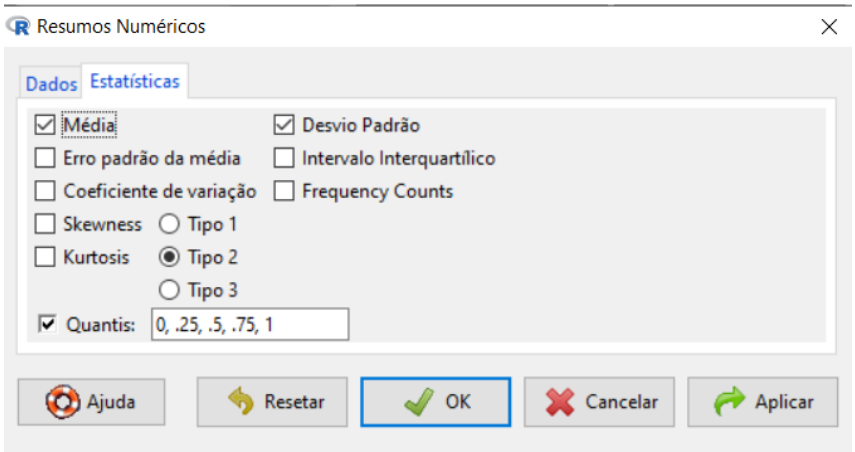


Figura 3.9: A interface deve ter essa configuração

Abaixo, é possível verificar algumas medidas como a média aritmética (mean), desvio padrão (sd), tamanho da amostra (n) e a frequência quartílica (0% 25% 50% 75% 100%).

```
> numSummary(idade[,"idade", drop=FALSE], statistics=c("mean", "sd", "IQR",
+   "quantiles"), quantiles=c(0,.25,.5,.75,1))
 mean       sd IQR 0% 25%  50% 75% 100%  n
 22.2 1.686548   2 20  21 22.5  23   25 10
```

Figura 3.10: Os Resultados gerados no Output

É importante lembrar que mediana e 2º quartil (50%) têm o mesmo significado. Novamente, não temos uma função "built-in" no R para o cálculo da moda, portanto, isso ainda é possível através de uma função customizável abaixo inserida no console do R:

mode = function(x)
freq = count(x)[[2]]
result = which.max(freq)
print(count(x)[result,1])

Após isso, inserimos o comando mode (idade) e teremos como saída o seguinte resultado:

```
> mode(idade)
[1] 23
```

Figura 3.11: A moda calcula via função customizável

Uma observação importante é que a função count() pertence ao pacote "plyr", que deve ser instalado separadamente.

Por outro lado, a variância possui uma função própria já disponibilizada pelo R, apesar de que não há opção para ela no Rcommander. Para se encontrar a mediana, basta entrar com o seguinte comando no R Script:

var(idade)

E teremos como saída o seguinte resultado:

```
> var(idade)
          idade
idade 2.844444
```

Figura 3.12: A variância calcula via comando do R

4. Com o conjunto de dados já identificado, clique na aba em gráficos e selecione boxplot como na figura abaixo . Após dar enter, o boxplot será gerado.

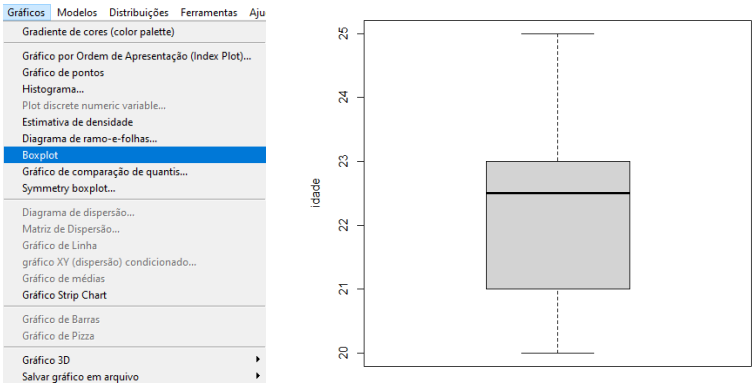


Figura 3.13: Seleção da opção Boxplot e gráfico gerado

Exercício 3.2 (3.0.3 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

Numa pesquisa sobre o tempo de trabalho dos estagiários de duas empresas A e B obteve-se os seguintes dados amostrais:

Empresa A	Empresa B
1 2 4 4 5 5 6 6 8 9	0 1 1 2 2 2 3 3 4 5

Qual a mediana em ambos os dados? Comente o significado da diferença entre as medianas. Compare a distância interquartílica (DI) entre as duas amostras e descreva o significado desta DI em ambos os gráficos.

Solução

Primeiramente para ser possível a análise de dados o Rcommander é necessário a inserção da amostra. Para isso basta clicar em dados → novo conjunto de dados. Após isto basta definir o nome do conjunto de dados e clicar em ok, como na imagem abaixo:

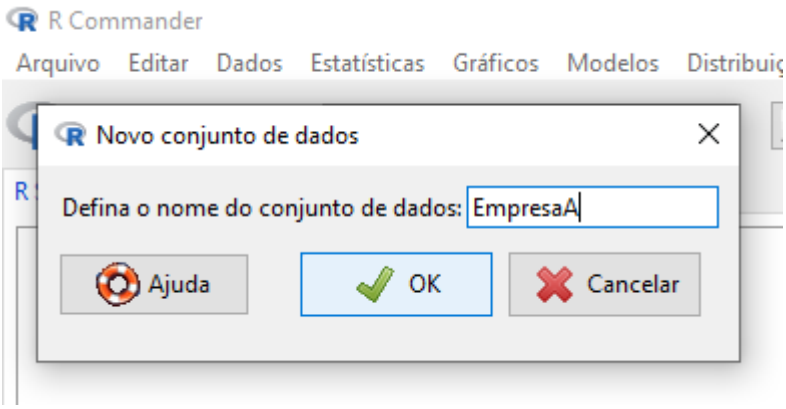
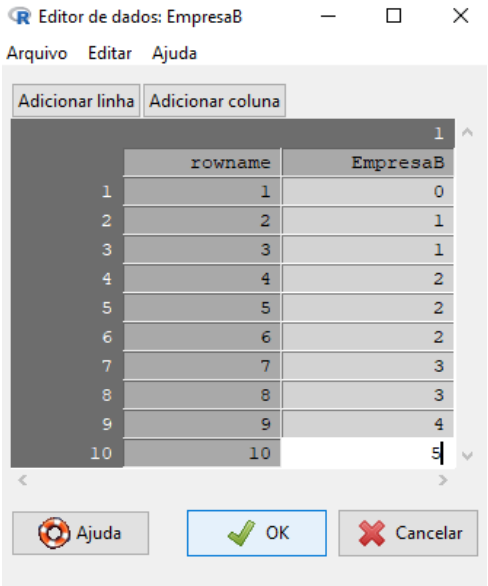


Figura 3.14: Criando o Banco de Dados

Após essa etapa basta colocar os dados em disposições de linhas, linha 1 observação 1, linha 2 observação 2 e assim por diante. A imagem a seguir foi feita a inserção dos dados amostrais na empresa B.



	rowname	EmpresaB
1	1	0
2	2	1
3	3	1
4	4	2
5	5	2
6	6	2
7	7	3
8	8	3
9	9	4
10	10	5

Figura 3.15: Inserindo os Dados

Para facilitar a análise de dados, iremos gerar os boxplots da empresa A e B. Logo abaixo, encontra-se a forma de gerar os boxplots. Assumindo que seu conjunto de dados ativos é da empresa A, basta acessar a aba de gráficos e clicar em Boxplot.

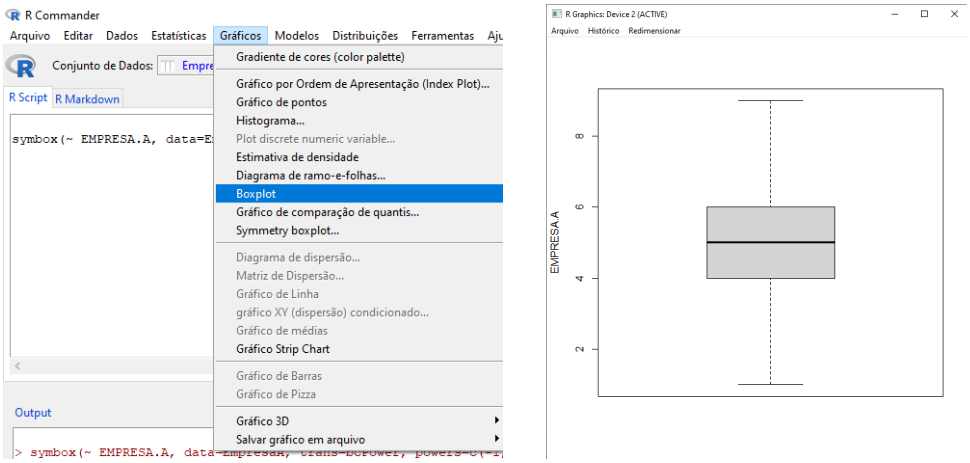


Figura 3.16: Seleção da opção Boxplot e gráfico gerado

De maneira análoga, para gerar o boxplot da empresa B, basta substituir o conjunto de dados ativos:

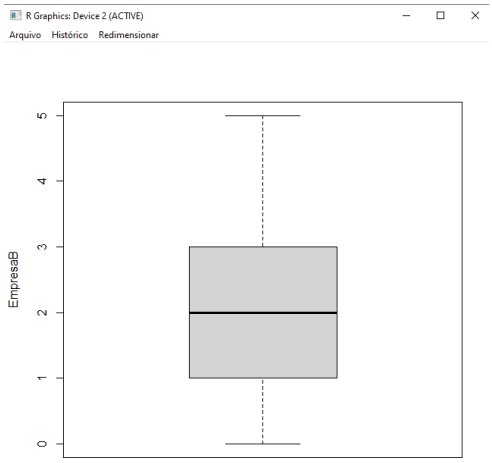


Figura 3.17: Boxplot da Empresa A

1. Qual a mediana em ambos os dados? Comente o significado da diferença entre as medianas.

Para achar a mediana, devemos ordenar o conjunto de dados em ordem crescente. A mediana é o valor que separa a metade maior e a metade menor da amostra, em termos mais simples, é o valor do meio de um conjunto. Para encontrar a mediana, devemos seguir a regra: se o número de elementos na amostra for par, então a mediana é a média dos dois valores centrais. Se o número de elementos na amostra for ímpar, a mediana é o valor central. Como obtivemos 10 dados nas empresas A e B, devemos, então, fazer a média entre os valores centrais (que se encontram na 5ª e 6ª posições).

Empresa A: Os valores centrais são 5 e 5, a média entre eles é igual a $\frac{5+5}{2} = 5$.

Empresa B: Os valores centrais são 2 e 2, a média entre eles é igual a $\frac{2+2}{2} = 2$.

2. Compare a distância interquartílica (DI) entre os dois gráficos e descreva o significado desta DI em ambos os gráficos.

Empresa A: O primeiro quartil é o valor em que 25% dos números do conjunto ordenado estão abaixo dele. Para determinar sua posição basta calcular quanto é 25% de 10 (pois o conjunto numérico possui 10 números) e arredondar para cima, caso a posição não seja exata. Daí, o primeiro quartil é o número

cuja posição no conjunto ordenado é $0,25 * 10 = 2.5$ mas, como arredondamos para cima, como dito, sua posição é o número 3, ou seja, o terceiro número no conjunto ordenado: 1, 2, 2, 4, 5, 5, 6, 6, 8, 9.
O terceiro quartil é o valor em que 75% dos números do conjunto ordenado estão abaixo dele. Para determinar sua posição basta calcular quanto é 75% de 10 (pois o conjunto numérico possui 10 números) e arredondar para cima, caso a posição não seja exata. Daí, o terceiro quartil é o número cuja posição no conjunto ordenado é $0,75 * 10 = 7,5$ mas, como arredondamos para cima, como dito, sua posição é o número 8, ou seja, o oitavo número no conjunto ordenado: 1, 2, 2, 4, 5, 5, 6, 6, 8, 9.

Empresa B: Para a empresa B, repetiremos o mesmo processo, visto que também apresenta uma amostra de 10 valores. O primeiro quartil é o terceiro número no conjunto ordenado: 0, 1, 1, 2, 2, 2, 3, 3, 4, 5. Já o terceiro quartil é o oitavo número no conjunto ordenado: 0, 1, 1, 2, 2, 2, 3, 3, 4, 5.
Assim, a distância interquartílica é a distância entre o primeiro quartil e o terceiro quartil, a distância interquartílica da empresa A é $6 - 4 = 2$, enquanto a distância interquartílica da empresa B é $3 - 1 = 2$. Ou seja, a distância interquartílica entre as empresas é igual.

Capítulo 4

Probabilidade

A probabilidade é uma área da matemática que quantifica as chances de ocorrência em experimentos . No livro " Probabilidade : um curso moderno com aplicações I Sheldon Ross" escrito por Sheldon Ross traz que a sua importância é dupla. Ross refere-se que, em primeiro lugar, estamos constantemente curiosos em calcular probabilidades quando temos conhecimento parcial sobre o resultado de algum experimento, neste caso a probabilidade é condicional. Ele refere-se também à situação onde não há conhecimento parcial sobre o resultado de um experimento, mas a necessidade de se calcular sua probabilidade . Pensando nisso, abaixo seguirá métodos para cálculo de probabilidade utilizando o R Commander.

Exercício 4.1 (4.0.1 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

Numa pesquisa feita com $n = 10$ amantes do cinema de uma cidade do interior, encontramos que 5 deles preferem um filme de drama, 1 de ação e 4 de terror.

1. Encontre a probabilidade p de preferir um filme de drama.
2. Numa exibição especial, nessa cidade, do filme "Que horas ela volta?" se venderam 20 ingressos. Qual a probabilidade de no máximo 3, dos que irão assistir o filme, tenham preferência pelo drama?

Solução

1. Encontre a probabilidade p de preferir um filme de drama.

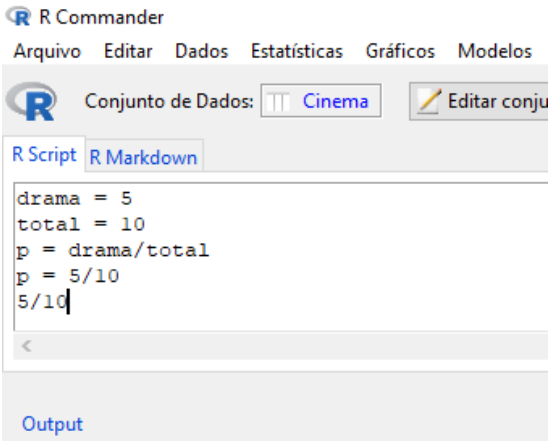


Figura 4.1: Insira os Dados como na Janela a seguir

```
> drama = 5

> total = 10

> p = drama/total

> p = 5/10

> 5/10
[1] 0.5
```

Figura 4.2: Clique em Submeter e terá os Resultados Embaixo

2. Numa exibição especial, nessa cidade, do filme "Que horas ela volta?" Se venderam 20 ingressos, qual a probabilidade de no máximo 3 dos que iram assistir o filme tenham preferência pelo drama.

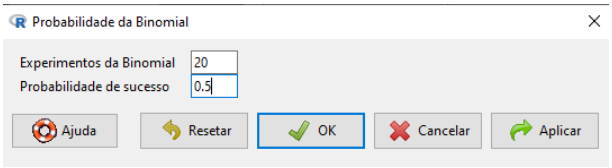
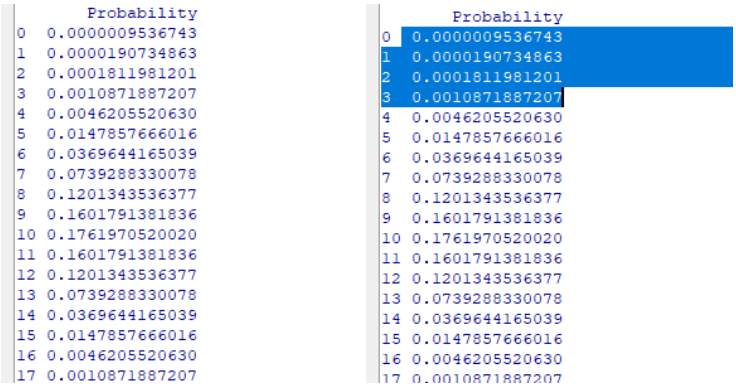


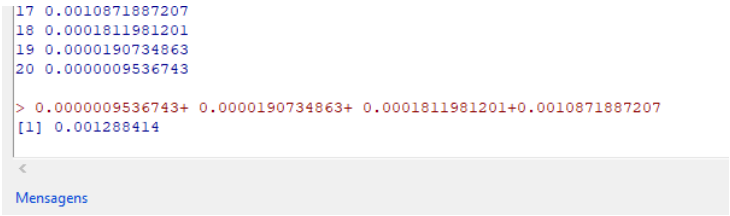
Figura 4.3: Insira os Dados como na Janela a seguir



```
       Probability
0  0.0000009536743
1  0.0000190734863
2  0.0001811981201
3  0.0010871887207
4  0.0046205520630
5  0.0147857666016
6  0.0369644165039
7  0.0739288330078
8  0.1201343536377
9  0.1601791381836
10 0.1761970520020
11 0.1601791381836
12 0.1201343536377
13 0.0739288330078
14 0.0369644165039
15 0.0147857666016
16 0.0046205520630
17 0.0010871887207
```

Figura 4.4: As Probabilidades da Quantidade de cada Número de Sucessos

Figura 4.5: Cole os valores selecionados na aba R Script somando-os



```
17 0.0010871887207
18 0.0001811981201
19 0.0000190734863
20 0.0000009536743

> 0.0000009536743+ 0.0000190734863+ 0.00018119812 01+0.0010871887207
[1] 0.001288414
```

Mensagens

Figura 4.6: Ao submeter os valores chegamos a nossa resposta

Exercício 4.2 (4.0.2 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

O tempo diário de ocupação por pessoa da biblioteca de uma faculdade tem distribuição Gaussiana com média 2,5 horas e desvio padrão 0,8. Calcule a probabilidade de que as pessoas que usam a biblioteca fiquem nela entre 1,5 e 2,5 horas.

Solução

Para fazer análises da distribuição Gaussiana (Normal), no R Commander clique na aba distribuições → Distribuições contínuas → Gráfico da distribuição normal como mostrados a seguir:

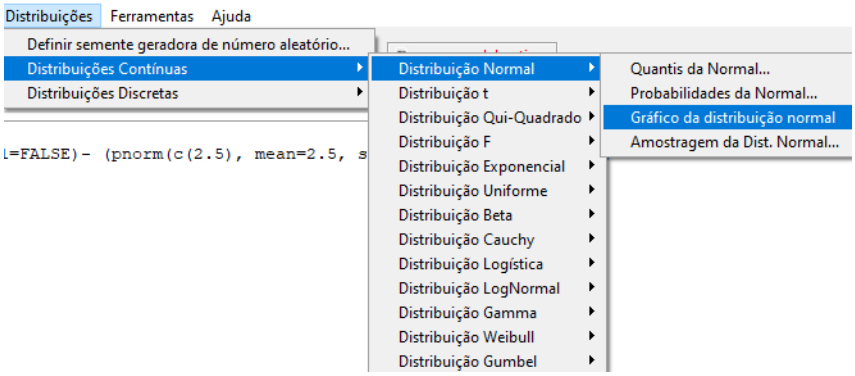


Figura 4.7: Siga esse Caminho

Após o passo acima, basta inserir os dados de média, desvio padrão da região de interesse e clicar em ok.

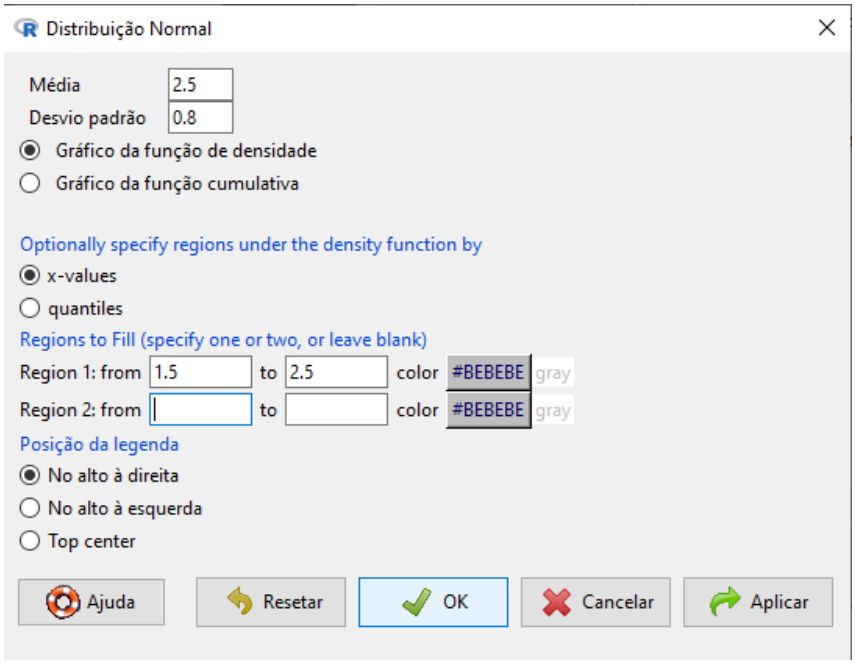


Figura 4.8: Insira os Dados como mostrado

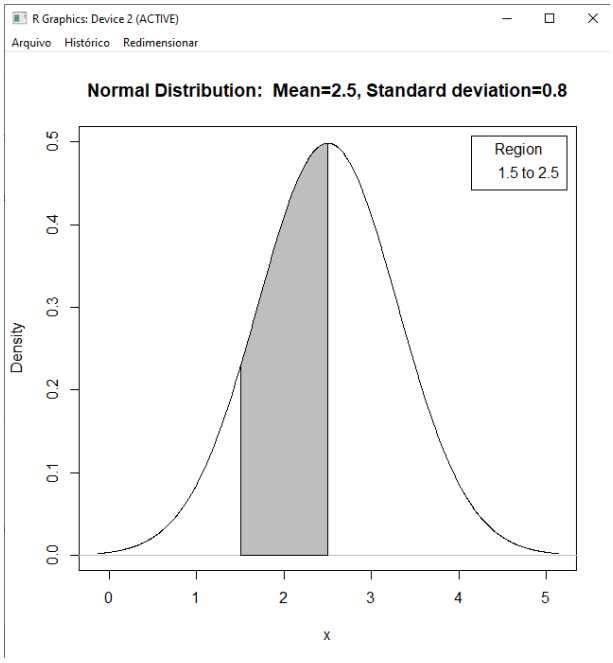


Figura 4.9: Essa é a Distribuição Proposta

De modo análogo, o cálculo da probabilidade segue os mesmos passos da criação do gráfico, ao invés de clicar em gráfico da distribuição normal, clique em probabilidade da normal.

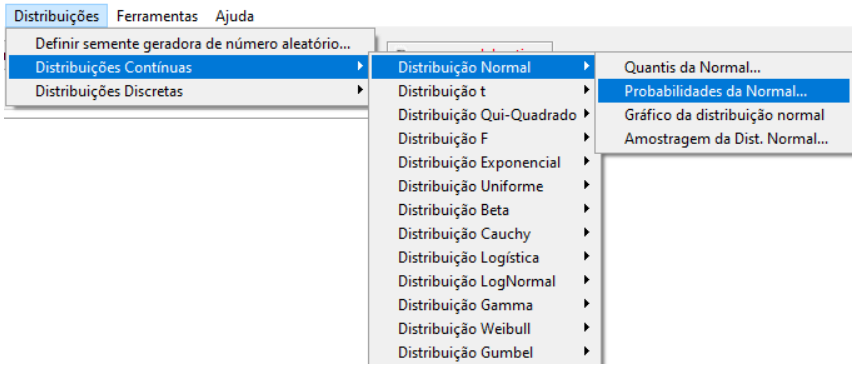


Figura 4.10: De Forma Análoga

Após esse primeiro passo, insira os limites inferior e superior da curva como abaixo:

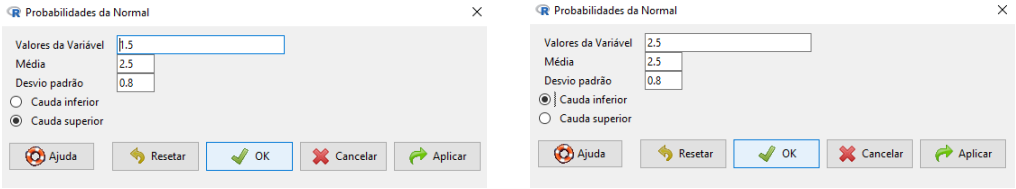


Figura 4.11: Inserir os valores da Máximo e Mínimo

Para o cálculo da probabilidade final de interesse, basta fazer a probabilidade da cauda superior menos a inferior.

```
[1] 0.5

> pnorm(c(1.5), mean=2.5, sd=0.8, lower.tail=FALSE)- (pnorm(c(2.5), mean=2.5, sd=0.8, lower.tail=TRUE))
[1] 0.3943502
```

Figura 4.12: Assim teremos o valor desejado

Exercício 4.3 (4.0.3 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

Em um curso de graduação em uma faculdade, sabemos que o número de favoráveis ao trabalho de campo é igual a 70% e 30% são contra.

1. Apresente a distribuição de probabilidade do número de alunos favoráveis quando selecionamos uma amostra aleatória (simples) de 6 alunos.
2. Chamando de X o número de alunos favoráveis nessa amostra de tamanho igual a 6, ache a esperança de X.

Solução

1. Apresente a distribuição de probabilidade do número de alunos favoráveis quando selecionamos uma amostra aleatória (simples) de 6 alunos.

Para fazer análises da distribuição binomial, no R Commander clique na aba Distribuições → Distribuições Discretas → Distribuição Binomial → Gráfico da distribuição binomial.

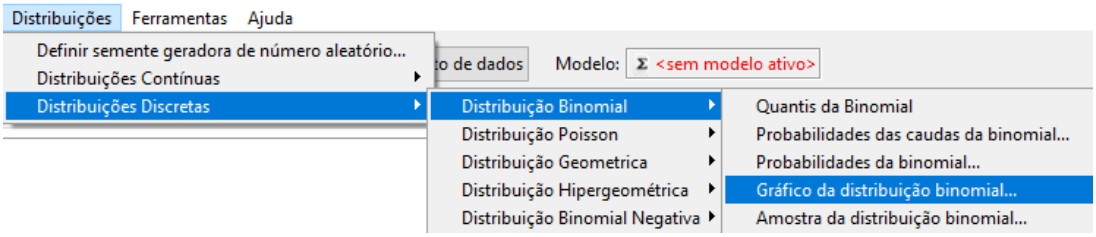


Figura 4.13: Seguindo as Instruções Temos

Insira a quantidade de experimentos da binomial (amostra aleatória) e a probabilidade de sucesso (ser favorável), como feito abaixo:

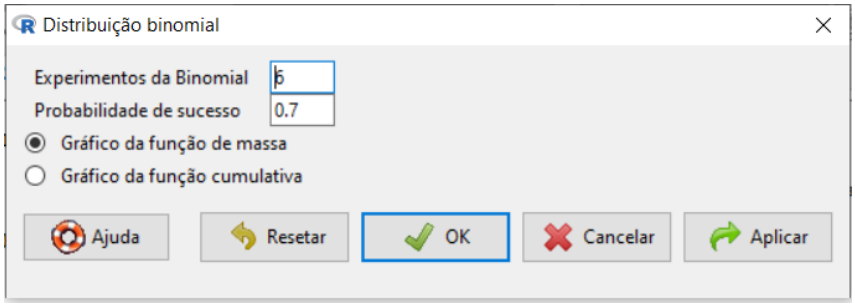


Figura 4.14: Insira as Informações

O resultado será dado pelo gráfico abaixo:

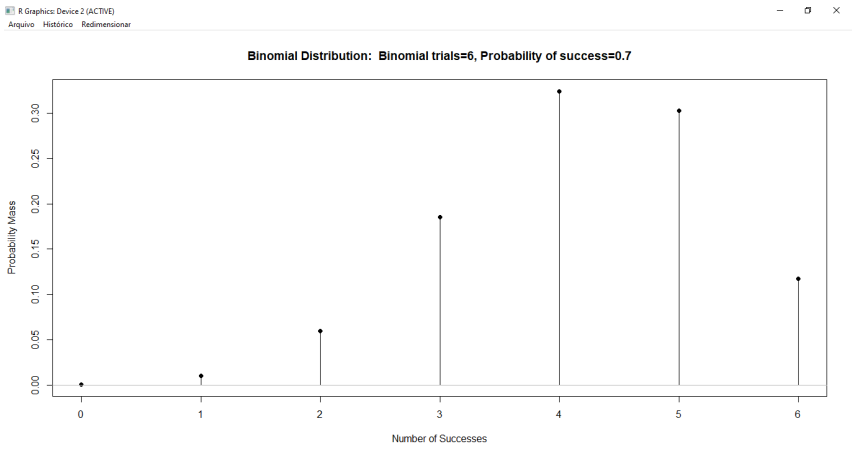


Figura 4.15: O Gráfico Gerado

2. Chamando de X o número de alunos favoráveis nessa amostra de tamanho igual a 6, ache a esperança de X.

Para encontrar a esperança de X, clique na aba Distribuições → Distribuições Discretas → Distribuição Binomial → Probabilidades da binomial.

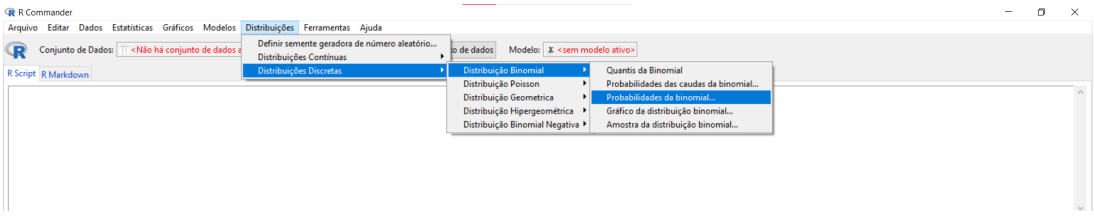


Figura 4.16: Seguindo as Instruções

Insira a quantidade de experimentos da binomial (amostra aleatória) e a probabilidade de sucesso (ser favorável), como feito abaixo:

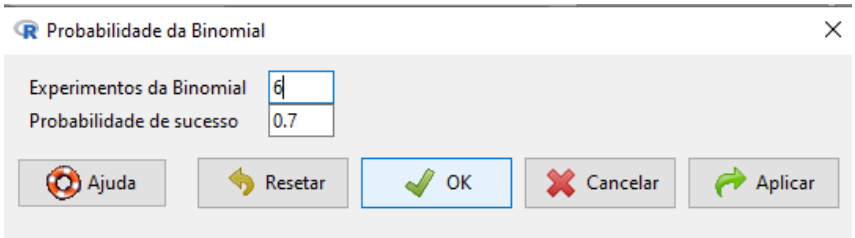


Figura 4.17: Inserindo os Dados

Abaixo se encontra o resultado:

```
> local({
+   .Table <- data.frame(Probability=dbinom(0:6, size=6, prob=0.7))
+   rownames(.Table) <- 0:6
+   print(.Table)
+ })
  Probability
0    0.000729
1    0.010206
2    0.059535
3    0.185220
4    0.324135
5    0.302526
6    0.117649
```

Figura 4.18: A Tabela com os Resultados

Porém, é importante lembrar que a esperança de X da binomial é dada por $n*p$, como feito abaixo:

```
Output

> ESPERANÇA

> 6*0.117649
[1] 0.705894
```

Figura 4.19: Resultado Final

Capítulo 5

Regressão Linear

A regressão linear é o processo de estimar uma equação, na qual a variável Y está condicionada a uma (regressão linear simples) ou mais variáveis X (regressão linear múltipla). Dessa forma, a regressão linear se mostra um modelo robusto, pois é capaz de estimar equações diversas, desde que seus parâmetros sejam lineares, como seu nome sugere. Para realizarmos essa estimação, existem diversos métodos, contudo, aquele que utilizaremos através de software será o Método dos Mínimos Quadrados, que consiste na minimização dos quadrados das diferenças entre os valores observados e os estimados. Ademais, nem sempre teremos uma modelo de regressão linear útil, pois podemos utilizar desse método para regredir duas variáveis que não possuem correlação. Para avaliarmos a qualidade da nossa regressão é comum utilizarmos o Coeficiente de Determinação (R^2), que avalia a porcentagem da variação de Y que está sendo explicada pelo modelo.

Exercício 5.1 (6.0.1 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

Acredita-se que um bom desempenho em matemática resulta em um bom desempenho em português. Para testar isso, uma prova de matemática e uma prova de português foram aplicadas a 10 alunos de uma turma. Os resultados obtidos por cada aluno são mostrados a seguir:

Aluno	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Português	9	9	8	5	9	8	6	4	8	8
Matemática	7	8	5	6	8	5	8	9	4	3

1. Supondo que as notas de português e matemática seguem uma distribuição normal, grafique a curva esperada.
2. Faça um diagrama de dispersão com os pontos observados.
3. Encontre o coeficiente de variação das variáveis notas de **português** e **matemática**

Solução

1. Supondo que as notas de português e matemática seguem uma distribuição normal, grafique a curva esperada.

Primeiramente, para ser possível a análise dos dados no R commander é necessário a sua inserção, como nos exercícios anteriores.

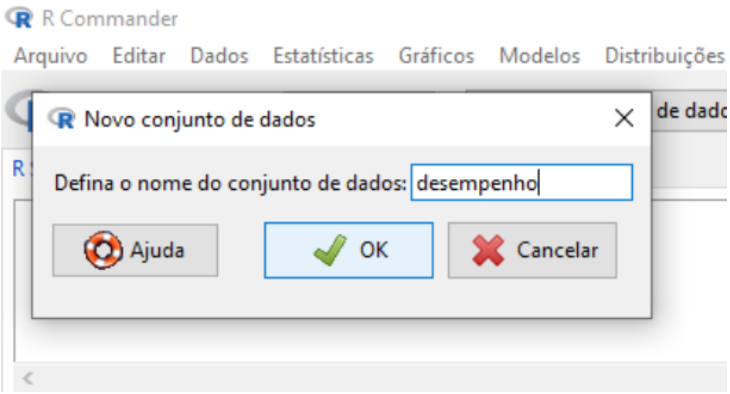
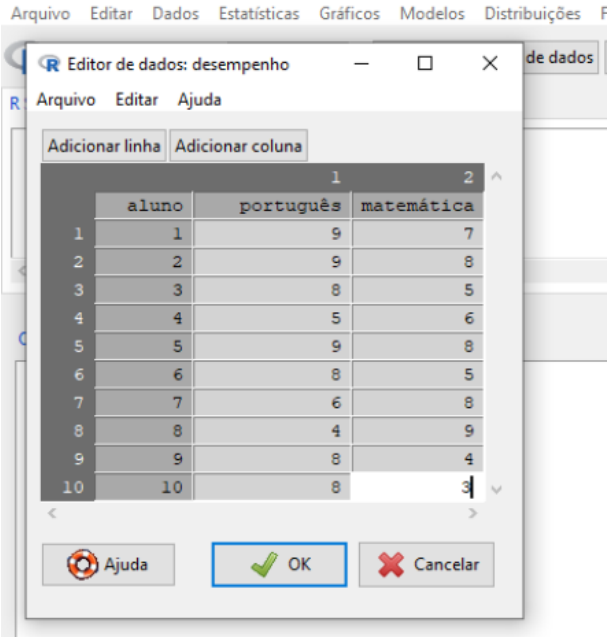


Figura 5.1: Seguindo as instruções temos

A diferença é que, neste caso, serão analisados dois dados simultaneamente, as notas de portugues e matemática. Para isso, basta a inserção de uma nova coluna, como pode-se observar na imagem abaixo:



	aluno	português	matemática
1	1	9	7
2	2	9	8
3	3	8	5
4	4	5	6
5	5	9	8
6	6	8	5
7	7	6	8
8	8	4	9
9	9	8	4
10	10	8	3

Figura 5.2: Inserindo ambos os dados

Após a inserção dos dados, basta clicar em ok, de forma que o conjunto de dados ficará ativo. Posteriormente a esta etapa de iniciação, é necessário a inserção das medidas resumo, para assim obter-se as medidas necessárias para a síntese do gráfico da normal (média e desvio padrão).Para isso, basta seguir os passos: Estatísticas → Resumos → Resumos numéricos.

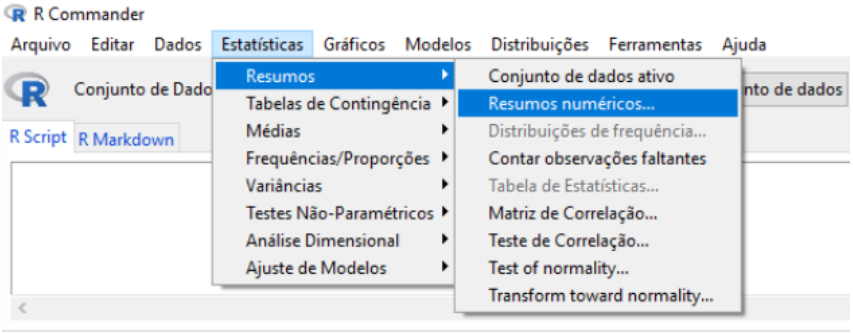


Figura 5.3: Seguindo as instruções temos

Como cada dado gera uma média ponderada e um desvio padrão diferentes, é necessário as medidas de resumo do conjunto "português" e "matemática", de forma a realizar a síntese dos gráficos.

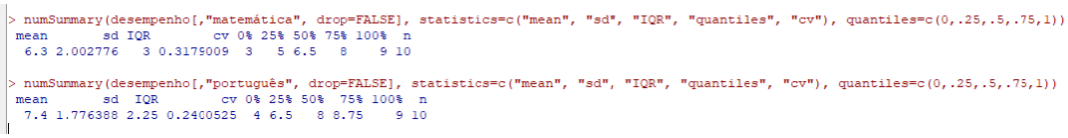


Figura 5.4: Resultado

Após a obtenção das medidas resumo, é possível a graficação da curva normal. Para isto, basta acessar Distribuições → Distribuições contínuas → Distribuição normal → Gráfico da distribuição normal.
Com isto, basta completar os dados da média e desvio padrão obtidos das medidas de resumo e clicar em "ok".

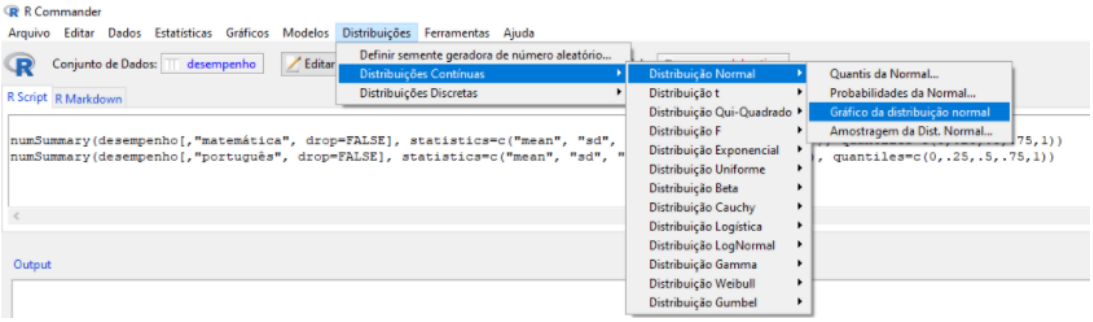


Figura 5.5: Seguindo as instruções temos

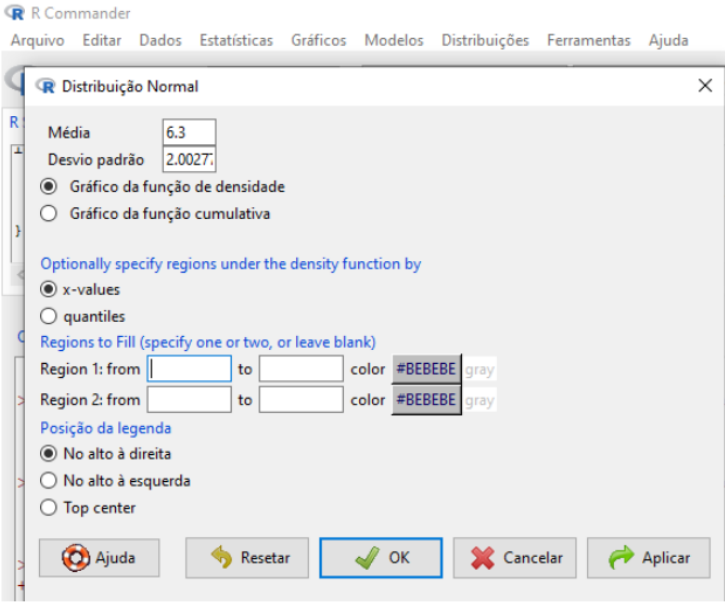


Figura 5.6: Inserindo os dados

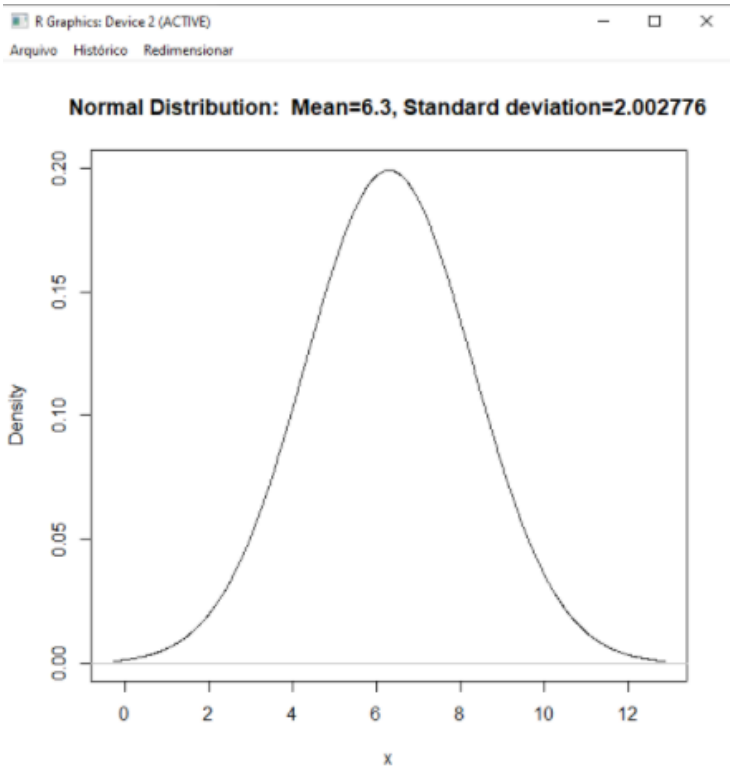


Figura 5.7: Resultado

De maneira análoga, para obtenção da curva da normal referente as notas de portugues, basta a substituição da média e desvio padrão, o que pode ser observado abaixo:

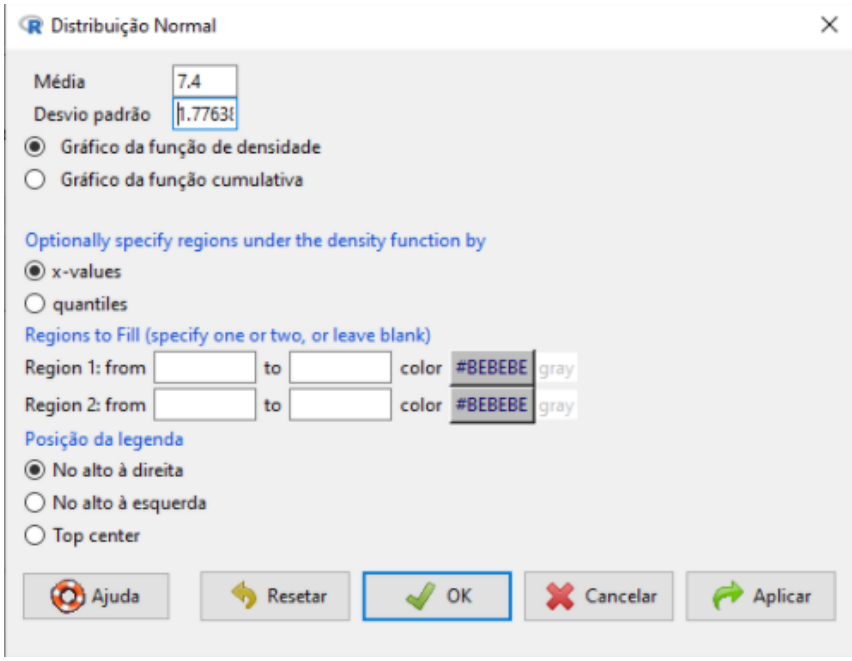


Figura 5.8: Inserindo os dados

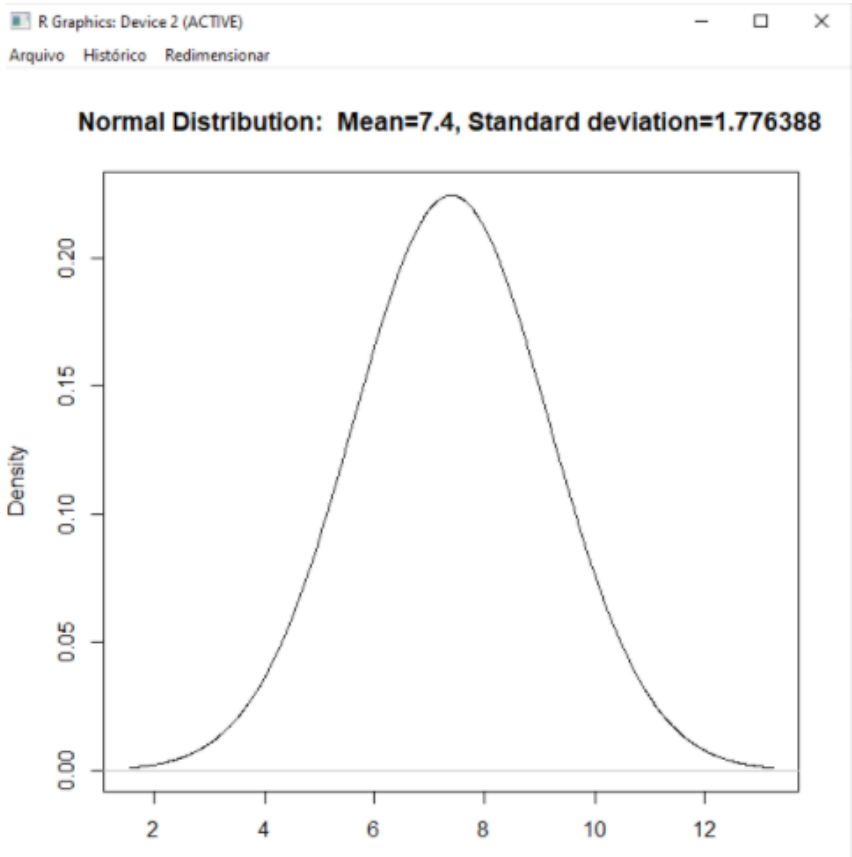


Figura 5.9: Resultado

2. Faça um diagrama de dispersão com os pontos observados.

Assumindo que o conjunto de dados que está ativo no R commander é o "desempenho", para a síntese do diagrama de dispersão basta acessar a barra de comandos Gráfico → Diagrama de dispersão.

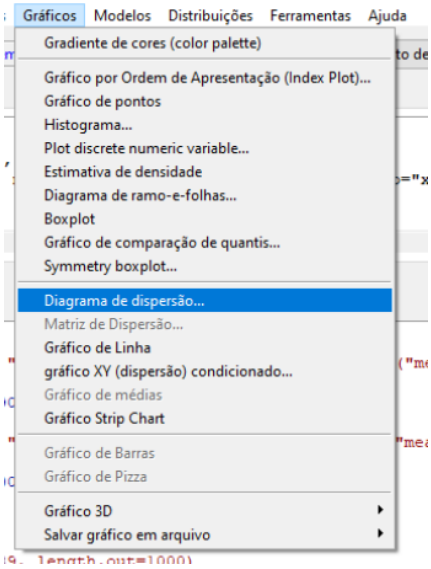


Figura 5.10: Seguindo as instruções temos

O R Commander gerará opções de escolha para caracterizar o eixo x e y como mostrado abaixo. Após a escolha basta clicar em ok.

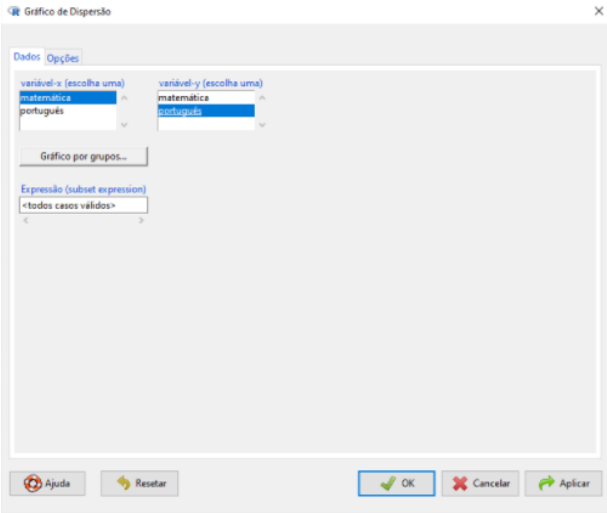


Figura 5.11: Seguindo as instruções temos

Após os passos acima, será gerado na tela o diagrama de dispersão, como mostrado abaixo:

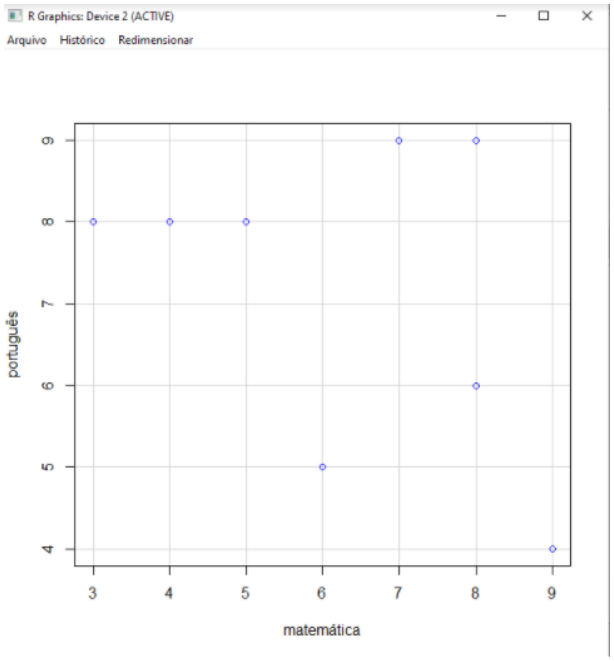


Figura 5.12: Seguindo as instruções temos

3. Encontre o coeficiente de variação das variáveis notas de português e matemática.

Os coeficientes de variação foram calculados juntamente com as médias de resumo no tópico 1. A abreviatura "cv" é utilizada para se referir a coeficiente de variação.

Coeficiente de variação - Português = 0.2400525

Coeficiente de variação - Matemática = 0.3179009

Exercício 5.2 (6.0.2 do Coleção de Exercícios de Estatística e Probabilidade para Ciências Sociais)

Após uma reestruturação na empresa, a produtividade média da empresa mudou. A empresa fez uma medição da produtividade de cada trabalhador (antes e depois da reestruturação), numa escala de 20 a 40 pontos. Considere as variáveis, X = "Produtividade Anterior" e Y = "Produtividade Posterior".

Funcionário	João	Maria	José	Pedro	Rita	Joana	Flávio	Paula	Catarina	Felipe
Antes	22	21	28	30	33	33	26	24	31	22
Depois	25	28	25	36	32	39	28	33	37	27

1. Faça um diagrama de dispersão para X e Y.
2. Faça uma regressão linear simples para Y como função de X e represente ela no gráfico da dispersão
3. Ache o coeficiente de correlação r.

Solução

1 - Inicialmente, devemos inserir os dados no R commander, analogamente aos exercícios anteriores:

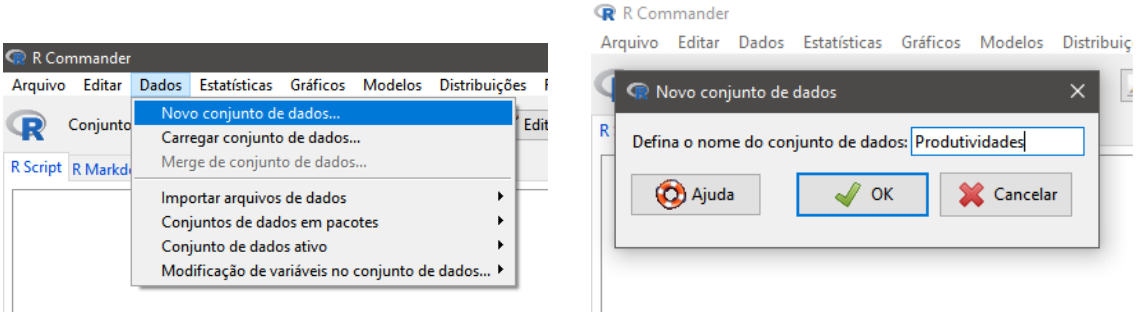
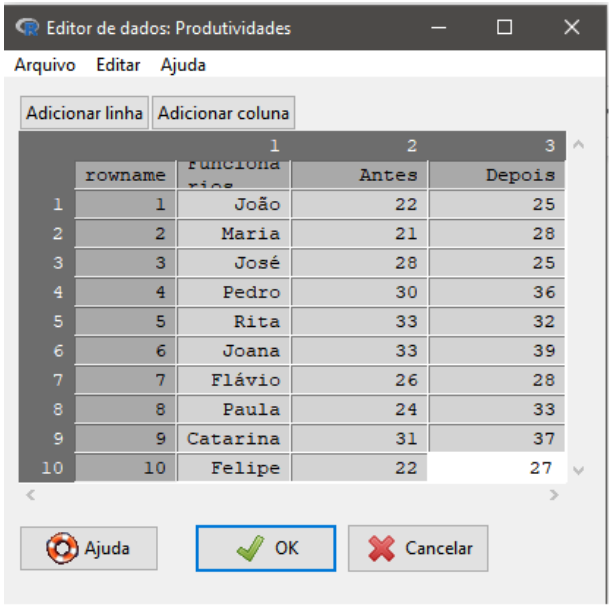


Figura 5.13: Começando com a Criação e a Nomeação do Banco de Dados



	1	2	3
rowname	funcionários	Antes	Depois
1	João	22	25
2	Maria	21	28
3	José	28	25
4	Pedro	30	36
5	Rita	33	32
6	Joana	33	39
7	Flávio	26	28
8	Paula	24	33
9	Catarina	31	37
10	Felipe	22	27

Figura 5.14: Monte o Banco de Dados da Seguinte Forma

Partindo disso, podemos fazer nosso gráfico de dispersão:

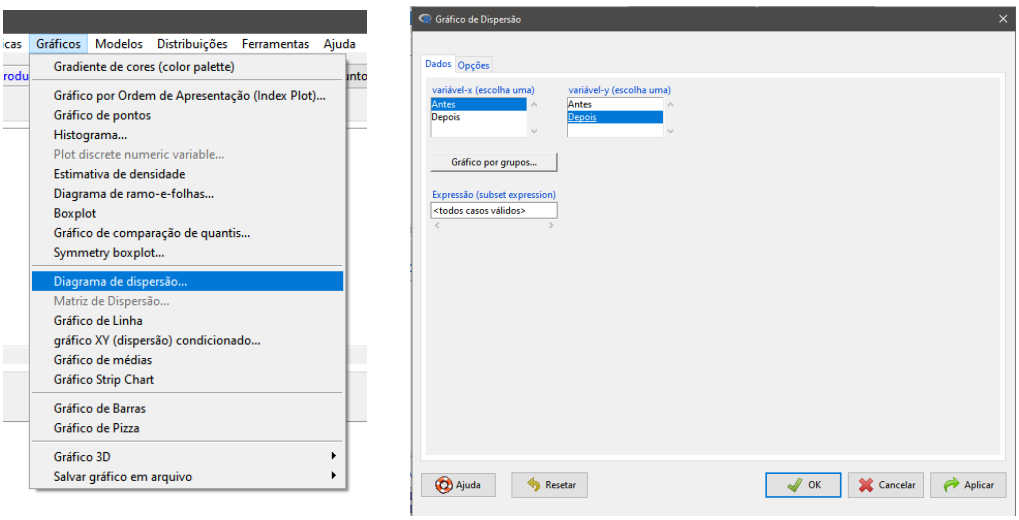


Figura 5.15: Começando com a Criação e a Nomeação do Banco de Dados

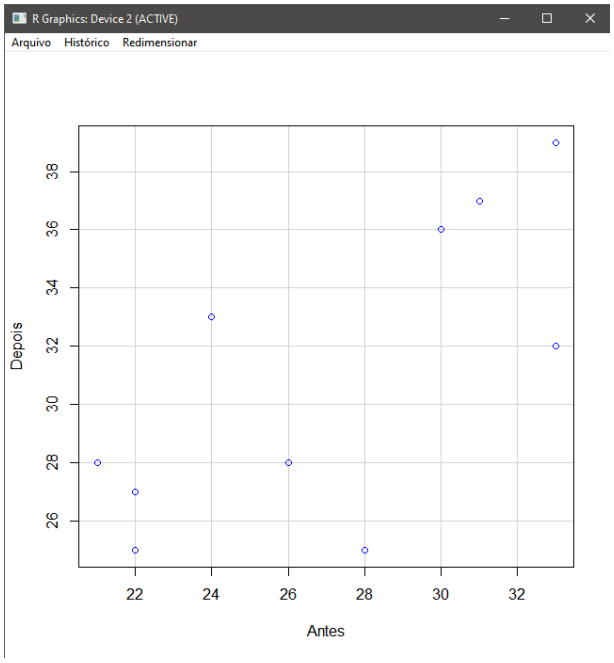


Figura 5.16: Temos assim, o Gráfico de Dispersão

2 - Agora, para fazermos a regressão devemos usar ir no seguinte menu abaixo:

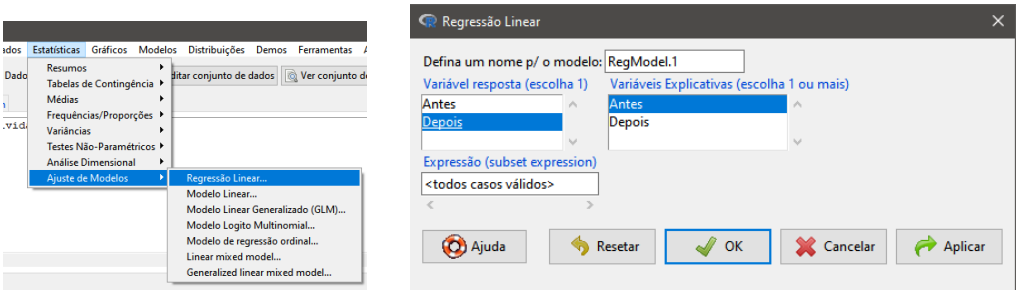


Figura 5.17: Seguindo esse caminho, devemos configurar as opções da forma mostrada

```
> RegModel.1 <- lm(Depois~Antes, data=Produtividades)

> summary(RegModel.1)

Call:
lm(formula = Depois ~ Antes, data = Produtividades)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7835 -2.1830  0.8093  2.8119  4.3505

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.8454     7.5083   1.311   0.2262
Antes         0.7835     0.2745   2.855   0.0213 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.823 on 8 degrees of freedom
Multiple R-squared:  0.5046, Adjusted R-squared:  0.4427
F-statistic:  8.15 on 1 and 8 DF,  p-value: 0.02132
```

Figura 5.18: Teremos esse Output

Você tem como estimativa de $B_0 = 9.84$ e de $B_1 = 0.78$, agora, devemos plotar esse modelo gerado no gráfico de dispersão:

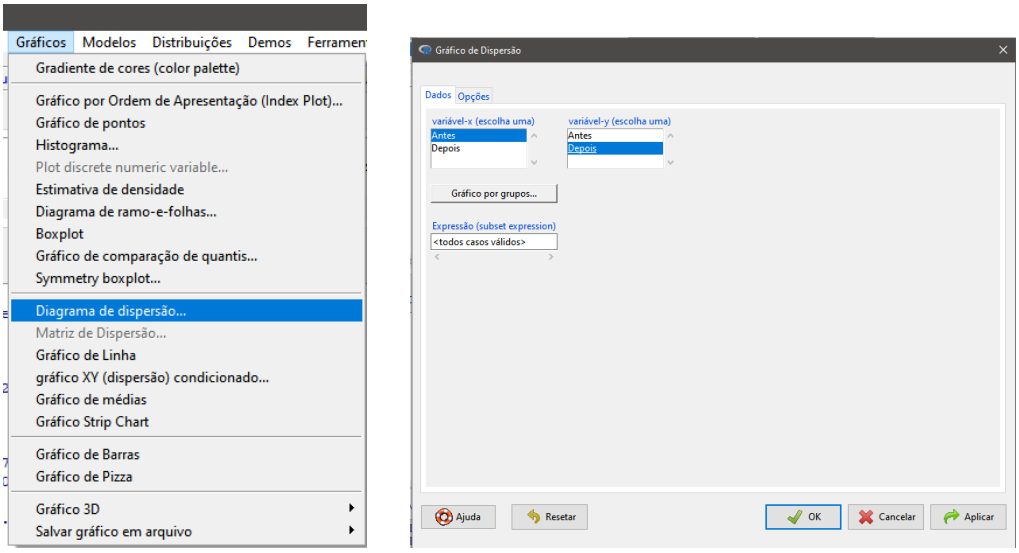


Figura 5.19: De forma análoga ao item anterior

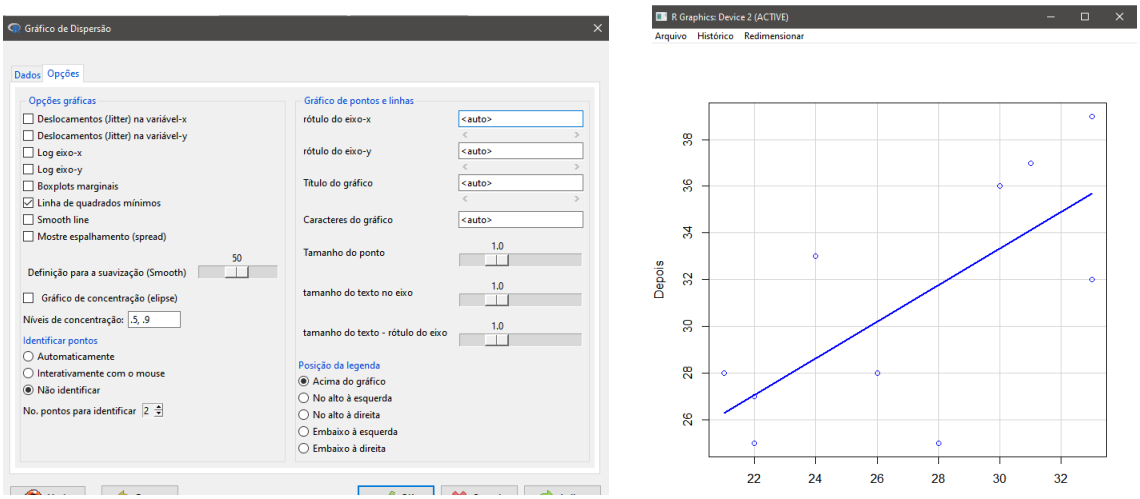


Figura 5.20: Devemos ir em opções e marcar a opção: Linha de Quadrados Mínimos. Então, teremos nosso gráfico plotado

3. Para tanto, devemos seguir os seguintes passos:

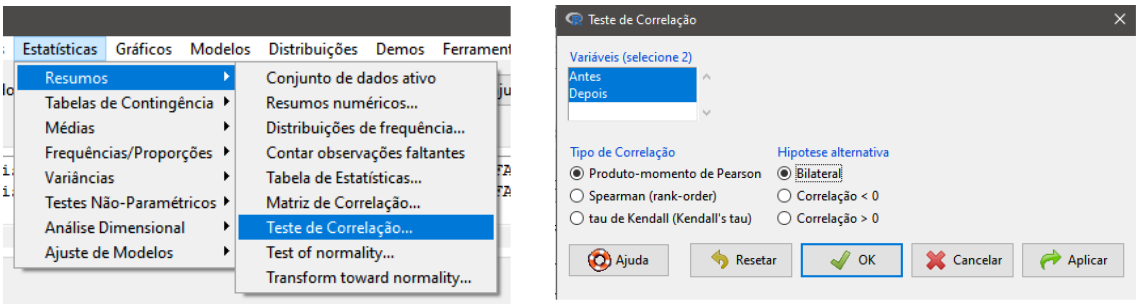


Figura 5.21: Devemos seguir esse caminho e selecionar as opções acima

```
        Pearson's product-moment correlation

data:  Antes and Depois
t = 2.8547, df = 8, p-value = 0.02132
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1460878 0.9258811
sample estimates:
      cor
0.7103735
```

Figura 5.22: Teremos esse Output

Capítulo 6

Mineração de Textos

A mineração textual é entendida como o processo de explorar um conjunto de dados, em busca de padrões consistentes, reduzindo o texto à termos significativos, e subsequentemente, ajudando na descoberta de conhecimentos, que poderão ser apresentados de diversas formas, como: hipóteses, árvores de decisão, dendrogramas, agrupamentos, grafos ou regras. Uma das possibilidades do campo é a mineração de textos (ou mineração de dados textuais), que refere-se ao processo de obtenção de informações relevantes de um texto. Normalmente, essas informações são obtidas pela composição de padrões e tendências, por meios como o padrão estatístico de aprendizagem. O processo, geralmente, inclui: a estruturação do texto de entrada (com a adição de algumas características linguísticas e a retirada de outras, e a inserção em um banco de dados); a derivação de padrões em meio a estrutura de dados; e a avaliação e interpretação do resultado. Ademais, para ciências sociais e humanidades de modo geral a mineração de textos é de suma importância pois ajuda no nivelamento e escolha de textos, já que com o uso desta ferramenta poderosa as escolhas ficam mais práticas com palavras chaves a priori estabelecidas, assim trazendo maximização do tempo e otimizando processo de pesquisa. Abaixo encontra-se um pequeno exemplo de mineração de texto utilizando o pacote RcmdrPlugin.temis.

Exemplo de Mineração Textual

Assumindo que tenha instalado e o aberto o pacote RcmdrPlugin.temis a seguir tome a título de exemplo uma forma para mineração de textos.

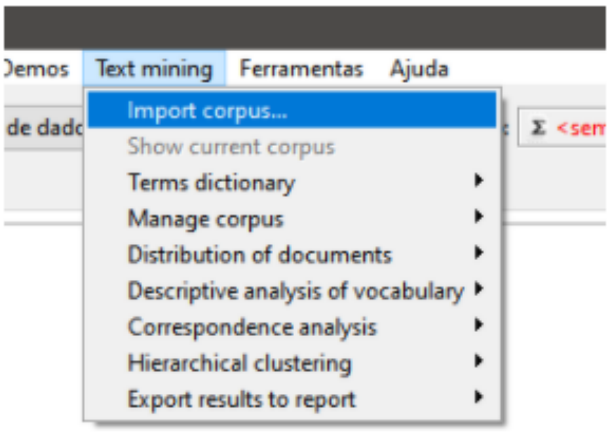


Figura 6.1: Selecione a seguinte opção

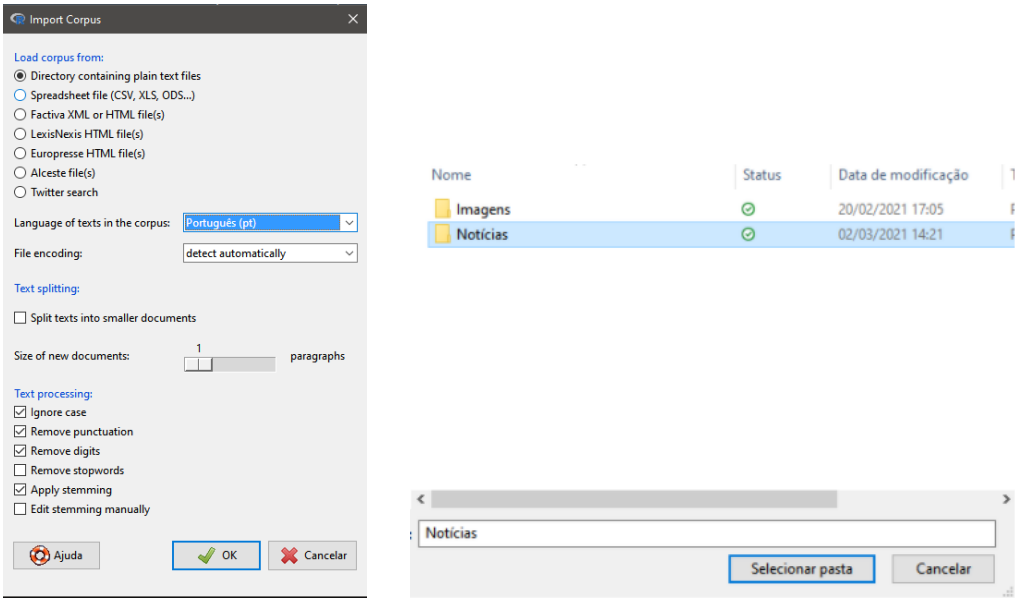


Figura 6.2: Configure da seguinte forma

Os arquivos de texto devem estar dentro de uma pasta, o programa vai importar todos os documentos .txt dentro da pasta.

Faremos uma limpeza inicial no documento tirando termos não significativos como os seguintes: “de”, “da”, “das”, “do”, “dos”, “em”, “que”, “na”, “no”, “nas” e “tem”. Para isso, clique em: Text mining → Manage corpus → Select or exclude terms.

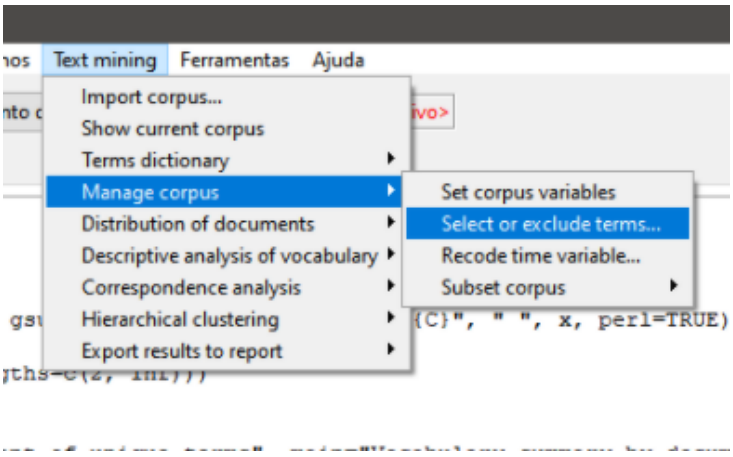


Figura 6.3: Selecione uma pasta com os arquivos de texto

Posteriormente, clique em “Exclude these terms” e escreva no espaço os termos que deseja excluir da análise e clique em “OK”.

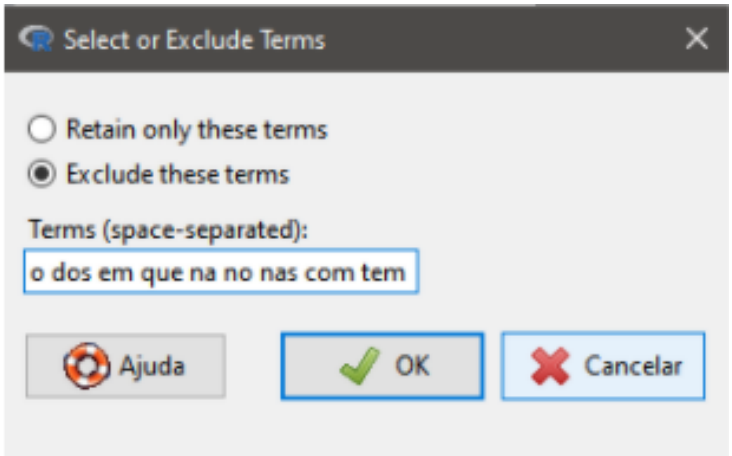


Figura 6.4: Insira os termos a serem excluídos

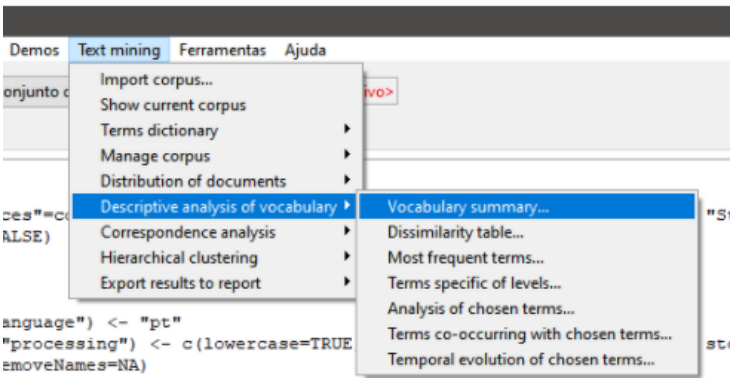


Figura 6.5: Siga as Instruções

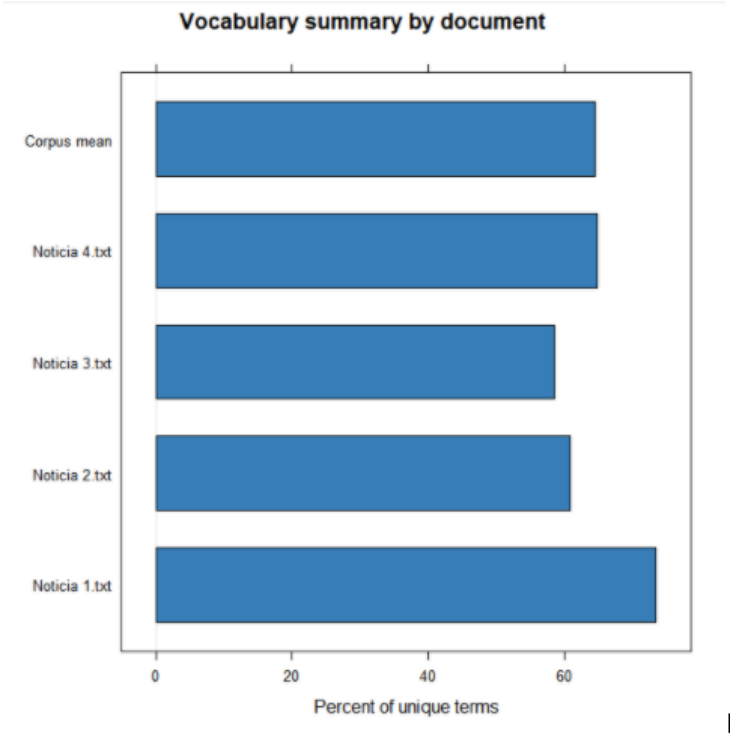


Figura 6.6: Teremos o seguinte Gráfico com o Resultado

	Noticia 1.txt	Noticia 2.txt	Noticia 3.txt	Noticia 4.txt	Corpus mean	Corpus total
Number of terms	105.0	307.0	132.0	155.0	174.8	699.0
Number of unique terms	80.0	192.0	79.0	103.0	113.5	295.0
Percent of unique terms	76.2	62.5	59.8	66.5	66.3	42.2
Number of hapax legomena	64.0	135.0	49.0	72.0	80.0	174.0
Percent of hapax legomena	61.0	44.0	37.1	46.5	47.1	24.9
Number of words	131.0	370.0	168.0	188.0	214.2	857.0
Number of long words	47.0	122.0	55.0	51.0	68.8	275.0
Percent of long words	35.9	33.0	32.7	27.1	32.2	32.1
Number of very long words	12.0	18.0	13.0	12.0	13.8	55.0
Percent of very long words	9.2	4.9	7.7	6.4	7.0	6.4
Average word length	5.6	5.3	5.2	5.2	5.3	5.3

Figura 6.7: E também a seguinte Tabela com o Resultado

Para a mineração de dados em sumo de textos uma forma eficiente de analisar é utilizando a análise descritiva. Para isto vá a barra de comandos Text mining > Descriptive analysis of vocabulary, a título de exemplo as imagens a seguir mostraram uma análise descritiva dos termos mais frequentes. Assim, clique posteriormente em "Most frequent terms".

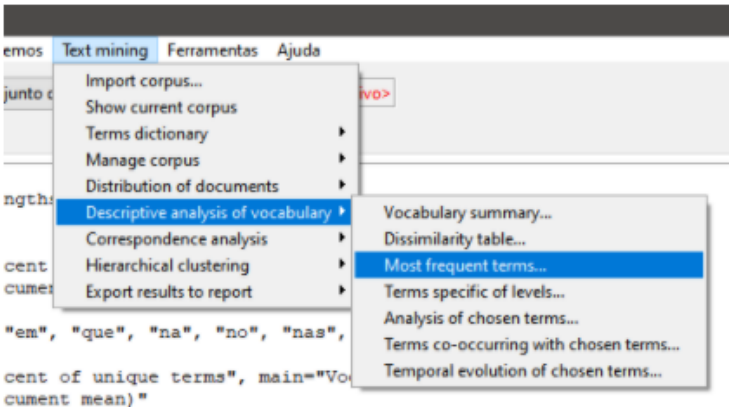


Figura 6.8: Agora a análise dos termos mais frequentes

Em "Number of terms to show"refere-se a quantidade de palavras mais frequentes que deseja analisar, neste caso analisaremos as 10 palavras mais frequentes.

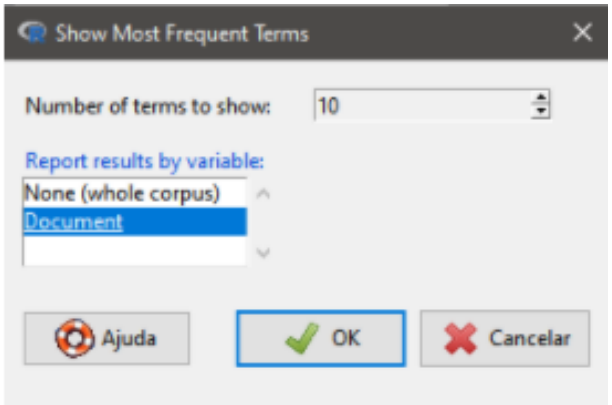


Figura 6.9: Selecione descritivo por documento

Logo após o passo acima o R Commander gerará os 10 termos mais frequentes, sendo eles separados por texto, como são 4 textos (notícias) gerou se 4 plots cada um referente as 10 palavras mais citadas de cada notícia.

$`Noticia 1.txt`

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
kalil	3.9	15	3.87	4	26	-0.19	0.575
par	3.9	19	3.12	4	21	0.26	0.398
paran	3.9	31	1.93	4	13	1.17	0.120
pesquis	3.9	24	2.53	4	17	0.68	0.250
aprov	2.9	30	1.49	3	10	0.90	0.183
levant	2.9	50	0.89	3	6	1.66	0.048
mostr	2.9	50	0.89	3	6	1.66	0.048
apo	2.0	67	0.45	2	3	1.54	0.062
bel	2.0	18	1.64	2	11	-0.04	0.516
carg	2.0	50	0.60	2	4	1.22	0.111

Figura 6.10: Resultado da Análise da Notícia 1

$`Noticia 2.txt`

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
govern	2.4	50	2.08	7	14	0.214	0.415
kalil	2.4	27	3.87	7	26	-1.581	0.057
bel	2.0	55	1.64	6	11	0.429	0.334
horizont	2.0	55	1.64	6	11	0.429	0.334
min	2.0	55	1.64	6	11	0.429	0.334
par	2.0	29	3.12	6	21	-1.207	0.114
prefeit	2.0	43	2.08	6	14	0.214	0.585
nã	1.7	50	1.49	5	10	0.093	0.463
pesquis	1.7	29	2.53	5	17	-0.958	0.169
ao	1.4	80	0.74	4	5	1.184	0.118

Figura 6.11: Resultado da Análise da Notícia 2

$`Noticia 3.txt`

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
kalil	5.5	27	3.87	7	26	0.807	0.2097
par	4.7	29	3.12	6	21	0.864	0.1938
pesquis	3.9	29	2.53	5	17	0.814	0.2079
prefeit	3.9	36	2.08	5	14	1.237	0.1080
bh	3.1	50	1.19	4	8	1.675	0.0469
paran	3.1	31	1.93	4	13	0.764	0.2224
arma	2.3	100	0.45	3	3	2.469	0.0068
govern	2.3	21	2.08	3	14	-0.043	0.5170
pt	2.3	75	0.60	3	4	1.990	0.0233
secret	2.3	100	0.45	3	3	2.469	0.0068

Figura 6.12: Resultado da Análise da Notícia 3

$`Noticia 4.txt`

	% Term/Level	% Level/Term	Global %	Level	Global	t value	Prob.
kalil	5.4	31	3.87	8	26	0.868	0.193
par	3.4	24	3.12	5	21	-0.018	0.507
disput	2.7	50	1.19	4	8	1.436	0.075
apoi	2.0	100	0.45	3	3	2.307	0.010
entre	2.0	43	1.04	3	7	0.902	0.184
govern	2.0	21	2.08	3	14	0.325	0.628
min	2.0	27	1.64	3	11	0.126	0.450
nã	2.0	30	1.49	3	10	0.296	0.384
os	2.0	43	1.04	3	7	0.902	0.184
paran	2.0	23	1.93	3	13	-0.183	0.572

Figura 6.13: Resultado da Análise da Notícia 4

Como pode-se observar em $\frac{3}{4}$ dos textos a palavra "govern"referente ao radical de "governo" apareceu, o que demonstra uma possível tendência desse arquivo.

Aqui, escolhemos fazer a análise hierárquica de cluster, caracterizado por um algoritmo capaz de fornecer mais de um tipo de forma de separar os dados. Neste exercício, faremos a partição dos dados por meio de métodos divisivos, em que todos os dados começam juntos em um único cluster e vão sendo separados, até que cada elemento seja seu próprio cluster.

Para isso, clique em: Text mining → Hierarchical clustering → Run clustering.

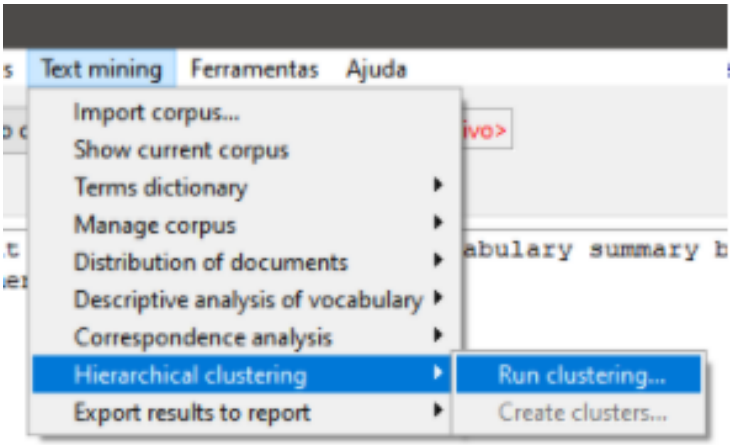


Figura 6.14: Agora, vamos calcular os Clusters

Ao abrir a nova janela, clique em "Use full document-term matrix" e coloque no espaço o valor desejado (no caso, 75.0) e clique em "OK".

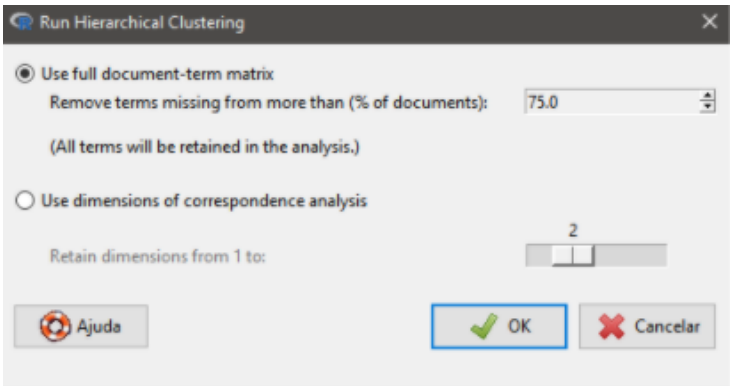


Figura 6.15: Configure da seguinte forma

Uma nova janela será aberta. Escolhemos o número máximo de clusters (4) e, posteriormente, clicamos em "OK".

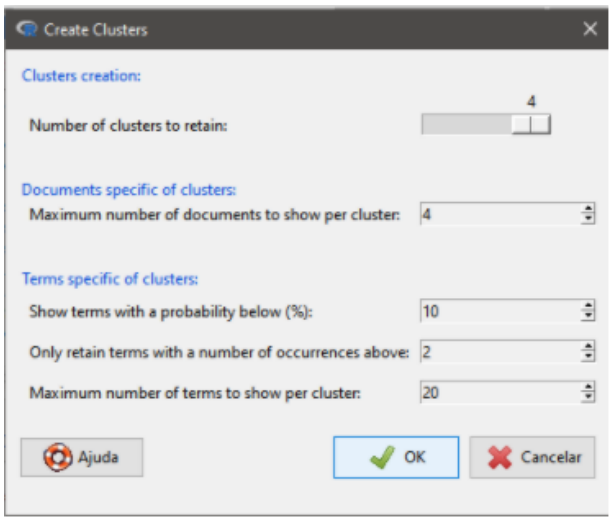


Figura 6.16: Configure da seguinte forma

Será gerado um dendrograma, em que é possível verificar, passo a passo, o processo de clusterização. É possível analisar, abaixo, os níveis de similaridade (distância) entre cada agrupamento (ramos) que é formado. Pode-se notar que, no caso em questão, a distância entre o cluster inicial e a primeira ramificação é de, aproximadamente, 0.45. Porém, a distância entre essa ramificação é a seguinte se reduz drasticamente, aproximando-se de 0.1.

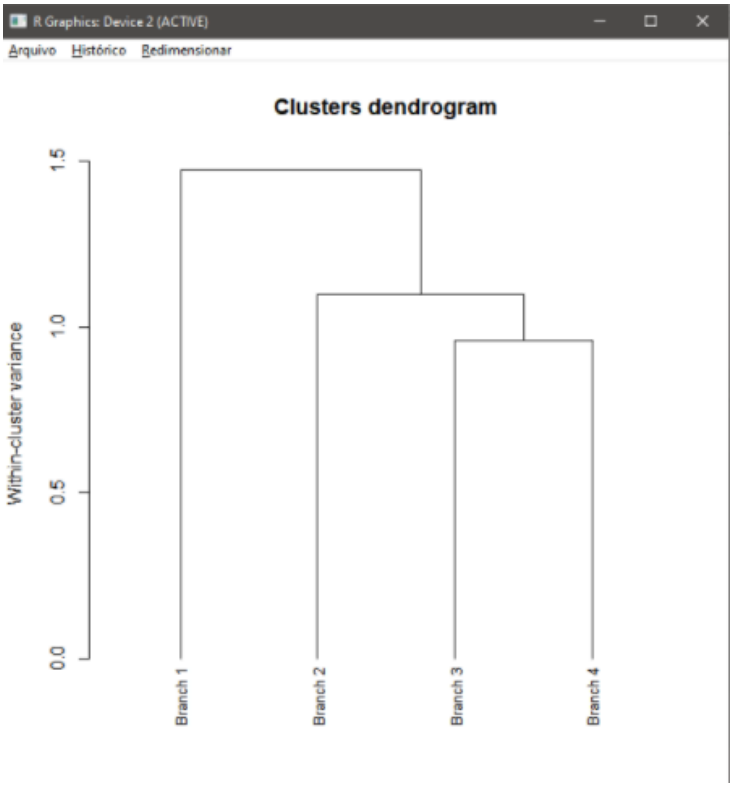


Figura 6.17: Você terá o seguinte resultado

Outra possibilidade interessante para mineração é a correspondência analítica. A título de exemplo, seguem as imagens abaixo:

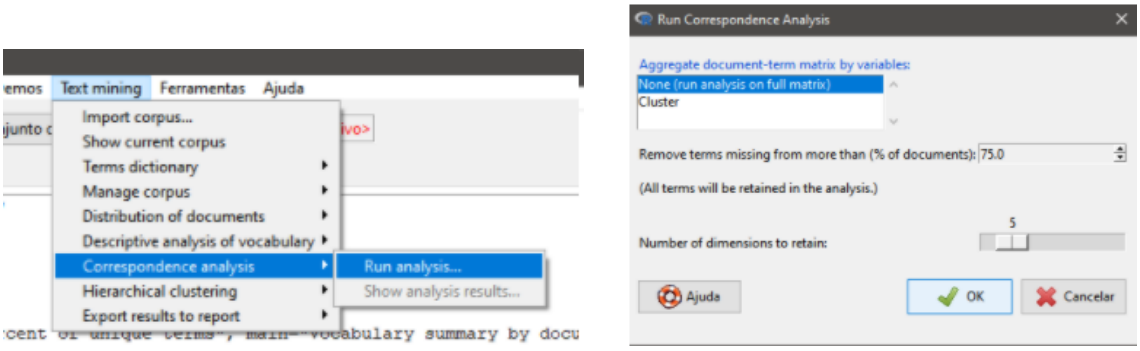


Figura 6.18: Configure da seguinte forma

Em Documents, coloque o número de texto que irá submeter e em Terms a quantidade de termos que deseja observar.

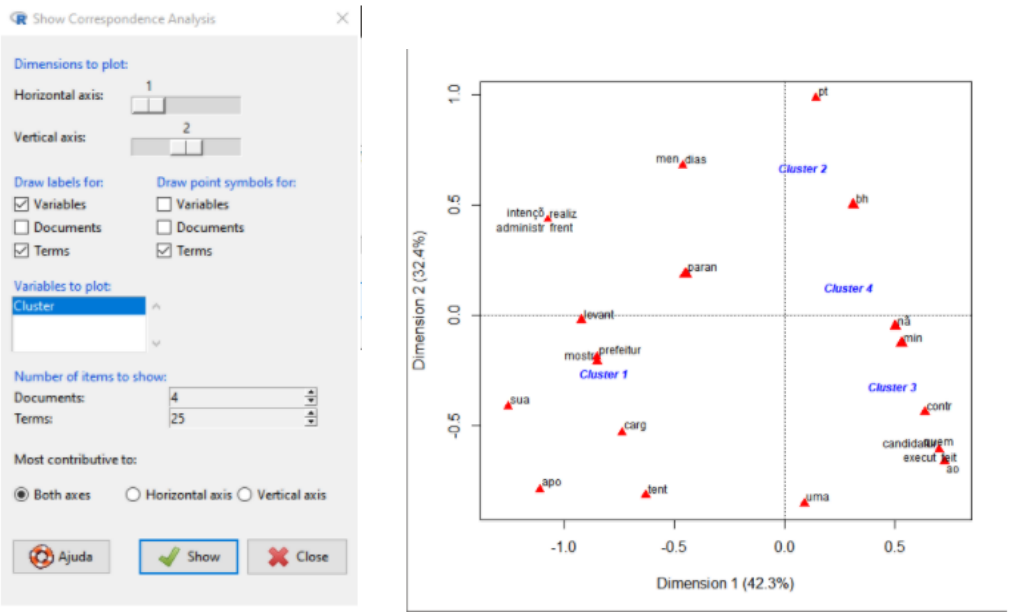


Figura 6.19: Configure dessa forma e obterá os seguintes resultados

Acima encontram-se as 25 palavras mais frequentes em cada texto, mas não apenas isso. O processo também considera o tamanho dos textos para a abordagem das palavras selecionadas com maior frequência (por isto a importância de excluir conectivos comuns no início do processo, para assim ser possível uma leitura mais precisa durante o processo de mineração).

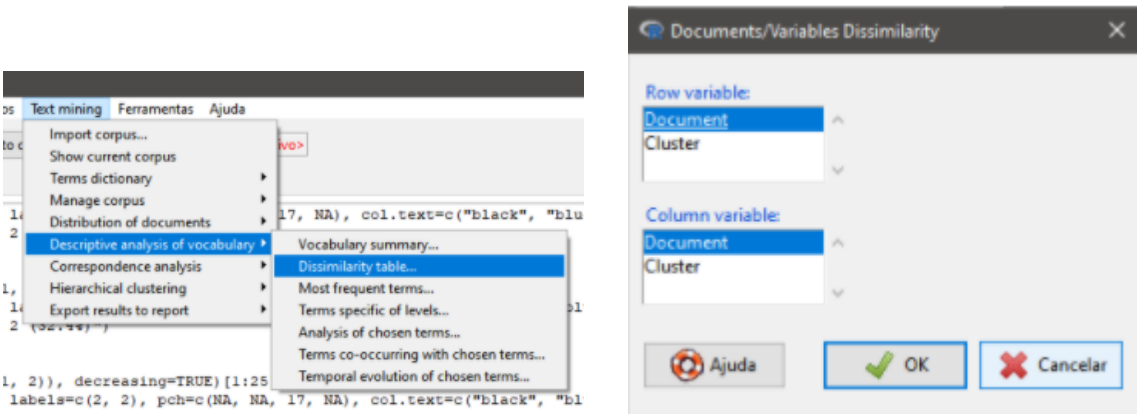


Figura 6.20: Siga as Instruções e Configure dessa forma

```
> diss
             Noticia 1.txt Noticia 2.txt Noticia 3.txt
Noticia 2.txt           2.0
Noticia 3.txt           2.0           1.7
Noticia 4.txt           2.1           1.7           1.8
```

Figura 6.21: Você terá o seguinte resultado

É possível fazer pesquisa de palavras e identificar a sua frequência de maneira individual nos textos. Abaixo encontram-se os passos a serem tomados.

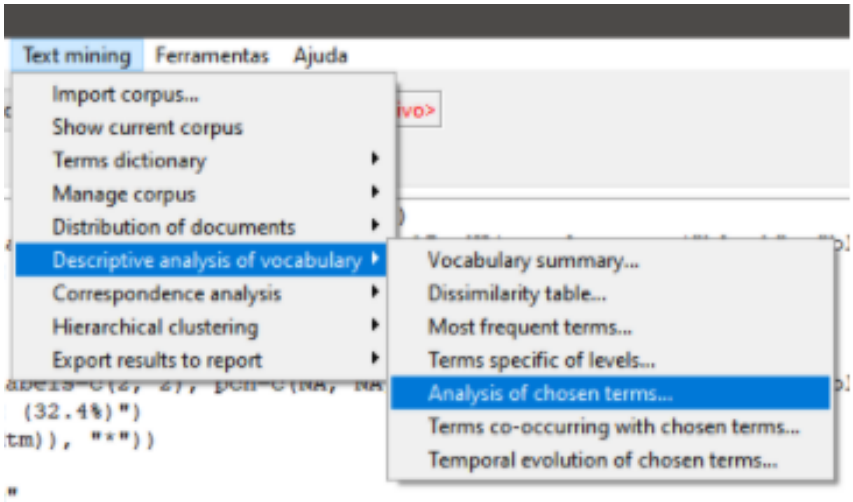


Figura 6.22: Siga as seguintes instruções

Digite no primeiro espaço a palavra de interesse e selecione as opções que tenha interesse.

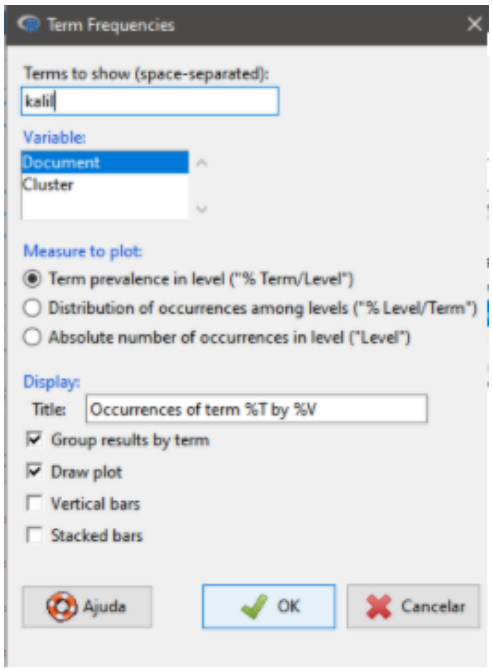


Figura 6.23: Configure dessa forma

```
> termFreqs
, , kalil

              % Term/Level % Level/Term Global % Level Global t value Prob.
Noticia 1.txt          3.9          15      3.9     4     26   -0.19 0.575
Noticia 2.txt          2.4          27      3.9     7     26   -1.58 0.057
Noticia 3.txt          5.5          27      3.9     7     26    0.81 0.210
Noticia 4.txt          5.4          31      3.9     8     26    0.87 0.193
```

Figura 6.24: Você terá o seguinte resultado

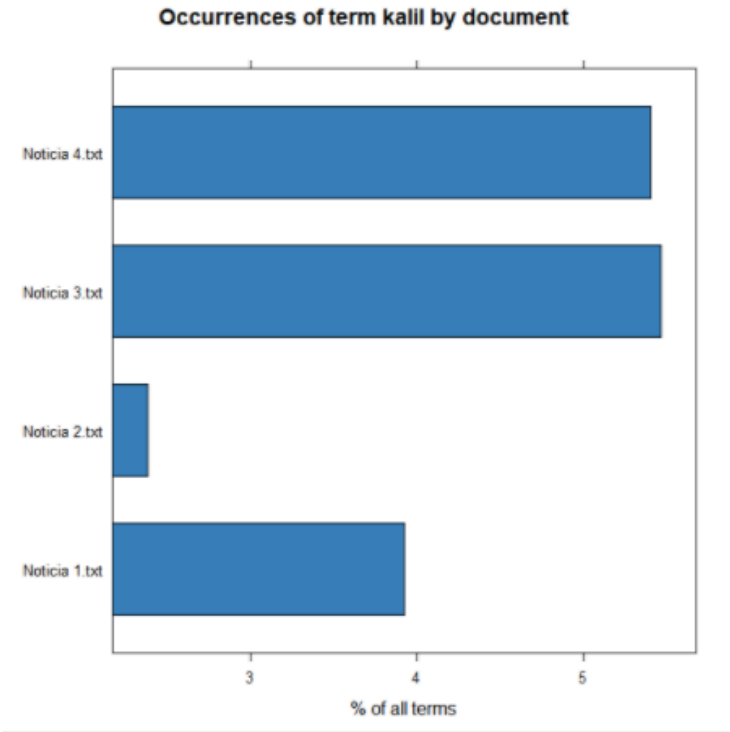


Figura 6.25: Você terá também seguinte resultado

Observando os plots acima pode-se afirmar que o termo “Kalil” é mais frequente no texto 1, mas está presente em todos os textos o de pode levar a um assunto em comum.

Bibliografia

Pedro Barbetta. *Estatística Aplicada às Ciências Sociais*. Editora da UFSC, 7ª edição edition, 2008.

Gabriel Barbosa. Paraná pesquisas: Kalil (psd) tem apoio popular para renunciar e disputar o governo de minas gerais em 2022. URL `https://www.ocafezinho.com/2021/02/28/parana-pesquisas--psd-tem-apoio-popular-para-renunciar-e-disputar-o-governo-de- -gerais-em-2022/`.

Milan Bouchet-Valat and Gilles Bastin. Rcmdrplugin. temis, a graphical integrated text mining solution in r. *The R Journal*, 5(1):188–196, 2013.

Esmael. Arma secreta do pt, prefeito de bh tem 75% de aprovação. URL `https://www.esmaelmorais.com.br/2021/02/arma-secreta-do-pt-prefeito-de-bh-tem-75-de-aprovacao/`.

John Fox. Getting started with the r commander: a basic-statistics graphical user interface to r. *J Stat Softw*, 14(9):1–42, 2005.

John Fox, Milan Bouchet-Valat, Liviu Andronic, Michael Ash, Theophilius Boye, Stefano Calza, Andy Chang, Philippe Grosjean, Richard Heiberger, Kosar Karimi Pour, et al. Package 'rcmdr'. 2020.

Lindenberg. Maior parte da população aprova a candidatura de kalil ao governo de mg. URL `https://blogdolindenberg.com.br/maior-parte-da-populacao-aprova-a-candidatura-de-kalil-ao-governo-de-mg/`.

Maquiavel. População avalia bem a gestão de kalil, mas maioria apoia renúncia em 2022. URL `https://veja.abril.com.br/blog/maquiavel/populacao-avalia-bem-a-gestao-de-kalil-mas-maioria-apoia-renuncia-em-2022`.

João Ismael D. Pinheiro. *Probabilidade e Estatística. Quantificando a incerteza.*

Sheldon Ross. *Probabilidade: um curso moderno com aplicações*. Bookman Editora, 2009.

Amanda Xavier, Gustavo Narimatsu, and Adrian(orientador) Luna. Coleção de exercícios de estatística e probabilidade para ciências sociais. pages 1–65, 2020.