Universidade Federal de Minas Gerais

Instituto de Ciências Exatas

Departamento de Estatística

Manual Rcommander

Monitoria PMG

Amanda Xavier, Gustavo Narimatsu, Larissa Carolina,

Matheus Gonzaga e Adrian Luna(orientador)

Sumário

0.1 Introdução

O R (R Development Core Team, o nome é a inicial dos creadores Ross Ihaka e Robert Gentleman, 1995) como linguaje de programação preferido para o trabalho estatístico com dados nas Ciências Sociais tem ido afirmado-se cada vez mais. Num principio a principal ferramenta nesta área estava dominado pelo uso pacote SPSS (Statistical Package for the Social Sciences, 1968), também pelo Stata (statistics and data, 1985) em menor medida. Essa situação mudou apreciavelmente nos últimos anos, agora com a popularidade da mineração de dados e o uso do Python o uso do R tem-se convertido em um padrão. Nas Ciências Sociais a adoção do R tem sido lenta, pela dificuldade do uso do R diretamente na linha de comando, como pseudocódigo. Esta situação tem mudado muito recentemente com o desenvolvimento de interfaces mais apelativas e naturais como por exemplo o Rstudio. Mesmo assim a dificuldade tem persistido pela necessidade do que o usuário tenha um conhecimento prévio de programação. A necessidade de uma interface mais amigável tem sido continua, especialmente para quem tem que lidar com dados e que provem de outra cultura com as coisas humanas. Ao encontro desta dificuldade é que está sendo apresentado o pacote R commander ("The R Commander: A basic-statistics graphical user interface to R", John Fox, 2005), que visa suprir a interação, com os dados e modelos, que o SPSS vinha (ou ainda está) desempenhado. O presente manual foi elaborado pelos Monitores de Graduação como parte das atividades do Programa de Monitoria da Graduação (PMG) no Departamento de Estatística, a partir do texto "Using the R Commander: A Point-and-Click Interface for R" John Fox, um dos principais desenvolvedores desta interface. Também usamos alguns dos dados e exemplos, gentilmente autorizados por Pedro Barbetta, autor, do livro "Estatística Aplicada às Ciências Sociais" . Aproveitamos para agradecer o apoio, para a conclusão deste manual, da Chefa do Departamento de Estatística, Professora Glaura Franco, e também ao financiamento das bolsas dos monitores pelo Programa de Monitoria da Graduação (PMG) da Pró-Reitoria de Graduação (Prograd).

No capítulo 1 (Instalações) mostramos um passo a passo da instalação dos softwares R e Rstudio, e a instalação do pacote R Commander. No capítulo 2 (Conhecendo o R commander) apresentamos o pacote R commander e sua interface, mostrando como iniciá-lo no R e suas principais características, além de mostrar como fazer algumas alterações que serão úteis durante a utilização da interface. No capítulo 3 (Gerenciamento de dados) mostramos como ler dados externos e de outros pacotes do R. Nesse capítulo também mostramos como fazer alterações nos nosso bancos de dados. No capítulo 4 (Estatísticas descritivas) apresentamos como usar o Rcommander para fazer análise descritiva de diferentes tipos de dados, mostramos como fazer resumos numéricos(média, mediana, etc...), tabelas de contingência e diversos tipos de gráficos. No capítulo 5 (Inferência) apresentamos, além da ideia geral de um teste de hipóteses, exemplos de testes para a média (pareados ou não), testes para a proporção, teste qui-quadrado, teste de correlação e suas respectivas resoluções utilizando o software R Commander. Incluímos, além das explicações, imagens que auxiliam no uso do software para que o aprendizado fique ainda mais completo. No capítulo 6 (Modelos Lineares) apresentamos, além da explicação generalizada do que é um modelo linear, como fazer uma regressão linear simples e múltipla no R Commander assim como dicas de boas práticas de análise, como fazer uma análise descritiva antes, interpretar o problema para identificar qual é a variável dependente e a independente, interpretar cada parâmetro da equação e entre outras. No capítulo 7 (simulações) é apresentado o conceito de

números aleatórios, como fazer simulações de distribuições de probabilidade, a explicação da lei dos grandes números e do TCL-Teorema Central do Limite e como testá-los no R Commander. Por fim, apresentamos um método de fazer teste de aderência por meio do qqplot (quantil-quantil) no software referido.

Capítulo 1

Instalações

1.1 Instalação Windows

O software R e os pacotes do R (assim como o pacote R-comander) estão disponíveis na Internet no CRAN (a Rede abrangente de arquivamentos R | consulte o Capítulo 1) em https://cran.r-project.org.

Ao inserir este link no seu navegador, o usuário será direcionado para a seguinte página inicial:

Figura 1.1: Página inicial do CRAN

O usuário Windows deve selecionar a opção "Download R for Windows". E algumas informações seão disponibilizadas na tela.

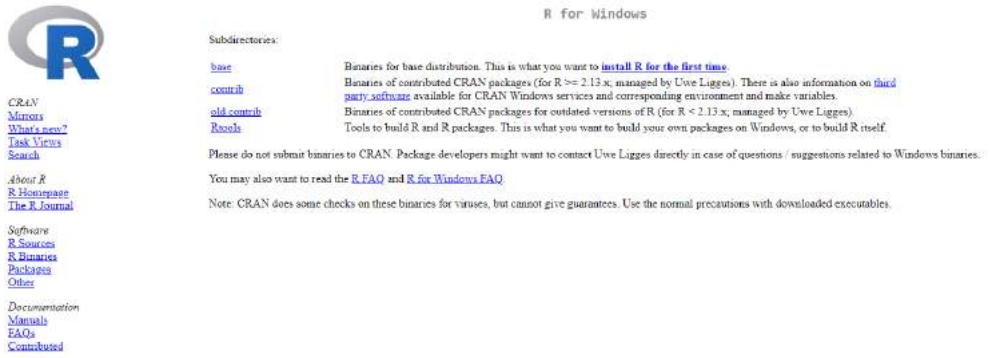


Figura 1.2: Informações adicionais

Para baixar o arquivo de instalação o usuário deve clicar em "Download R3.6.1 for Windows". s

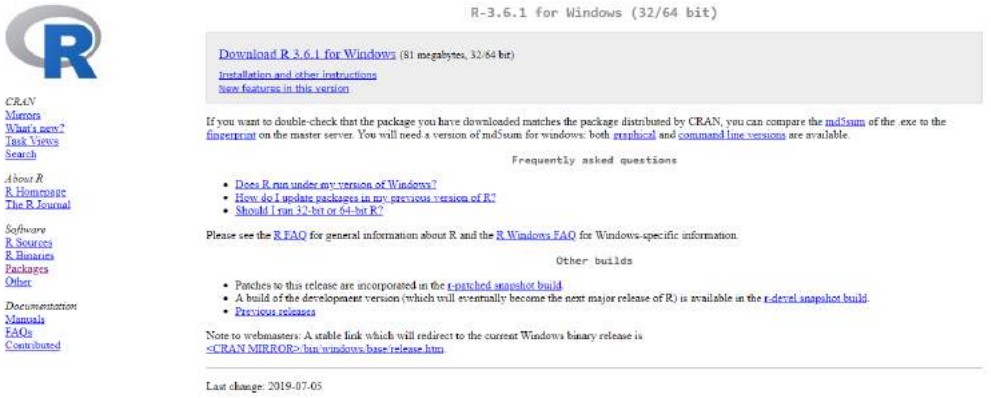


Figura 1.3: Link para iniciar o Download

Um arquivo executável aparecerá sob a sua barra de tarefas.

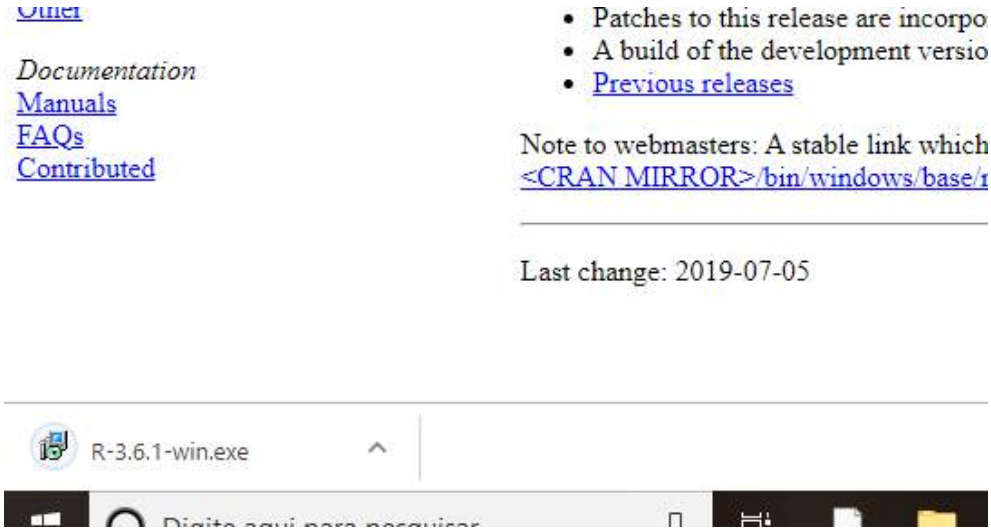


Figura 1.4: Arquivo executável

O próximo passo será clicar neste executável e iniciar a instalação do R.

Uma janela de execução será aberta e o usuário deve clicar em executar.

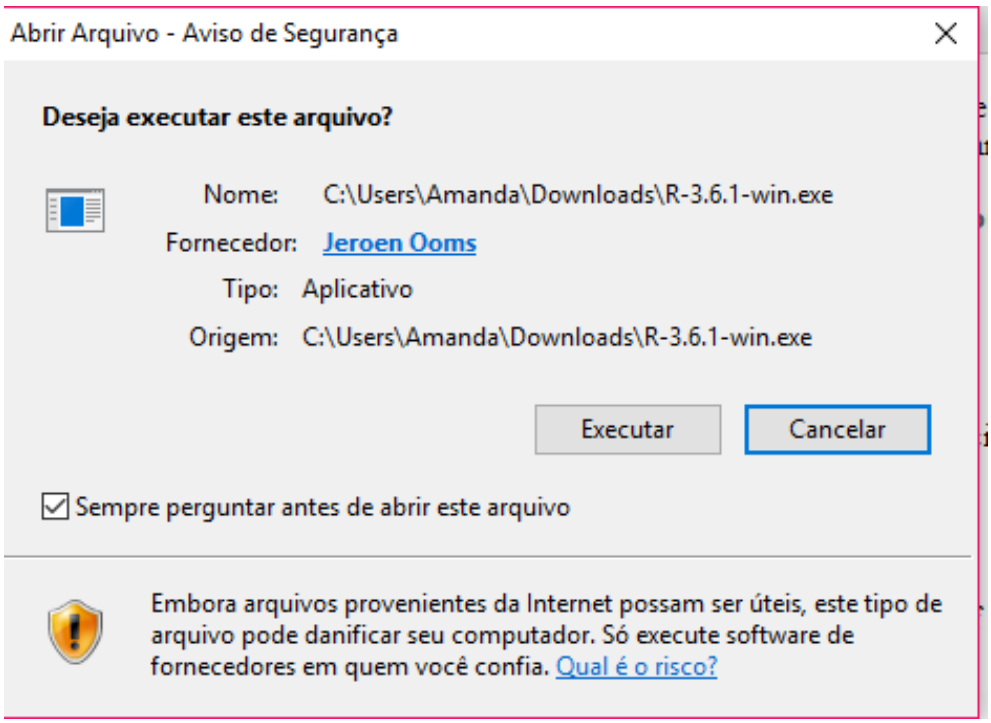


Figura 1.5: Janela de execução

O próximo passo será, selecionar o idioma do instalador (escolha o de sua preferência).

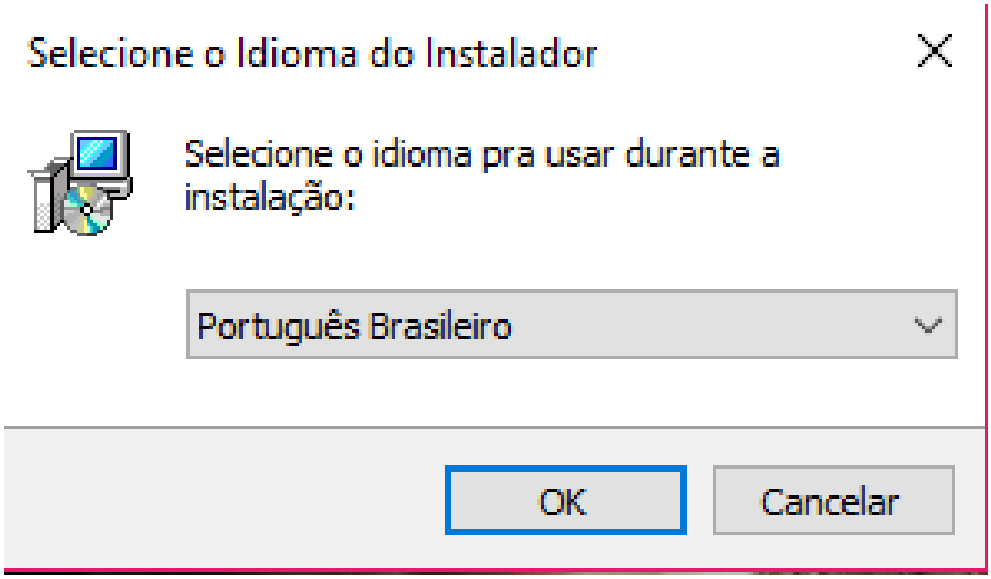


Figura 1.6: Selecionando o idioma

Algumas informações importantes sobre a licença de uso do software serão disponibilizadas na tela. Quando as informações estiverem entendidas, o usuário deve clicar em "Próximo".

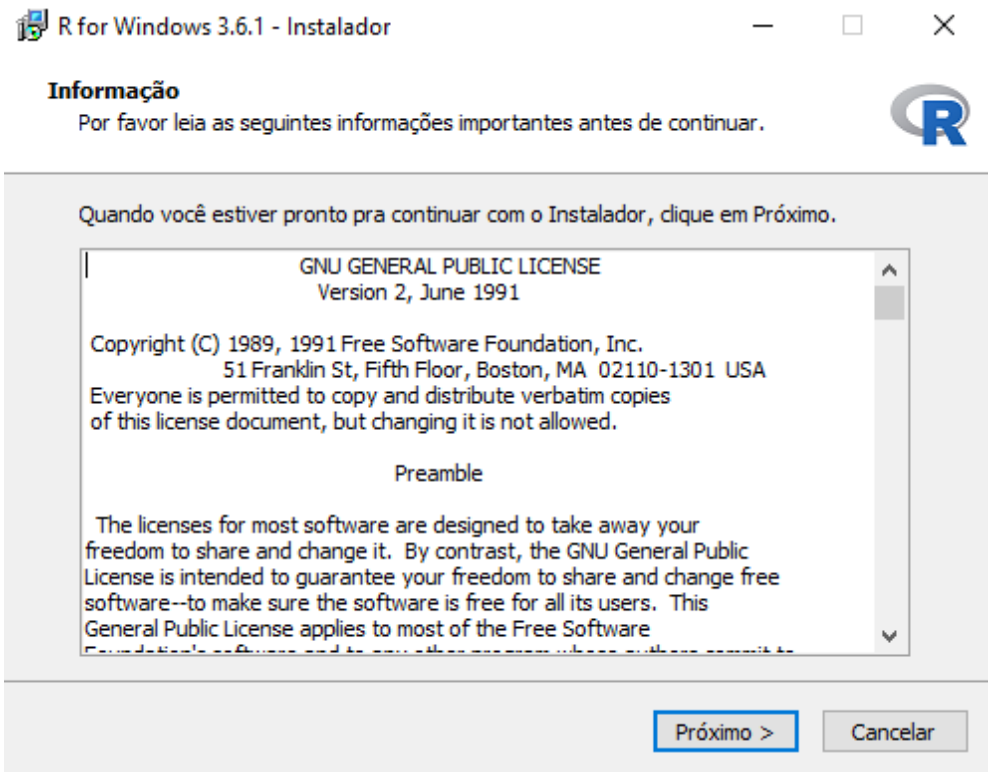


Figura 1.7: Informações Importantes

O próximo passo é selecionar a pasta onde o R será instalado. Uma vez selecionada a pasta de preferência, o usuário deve clicar em "Próximo".

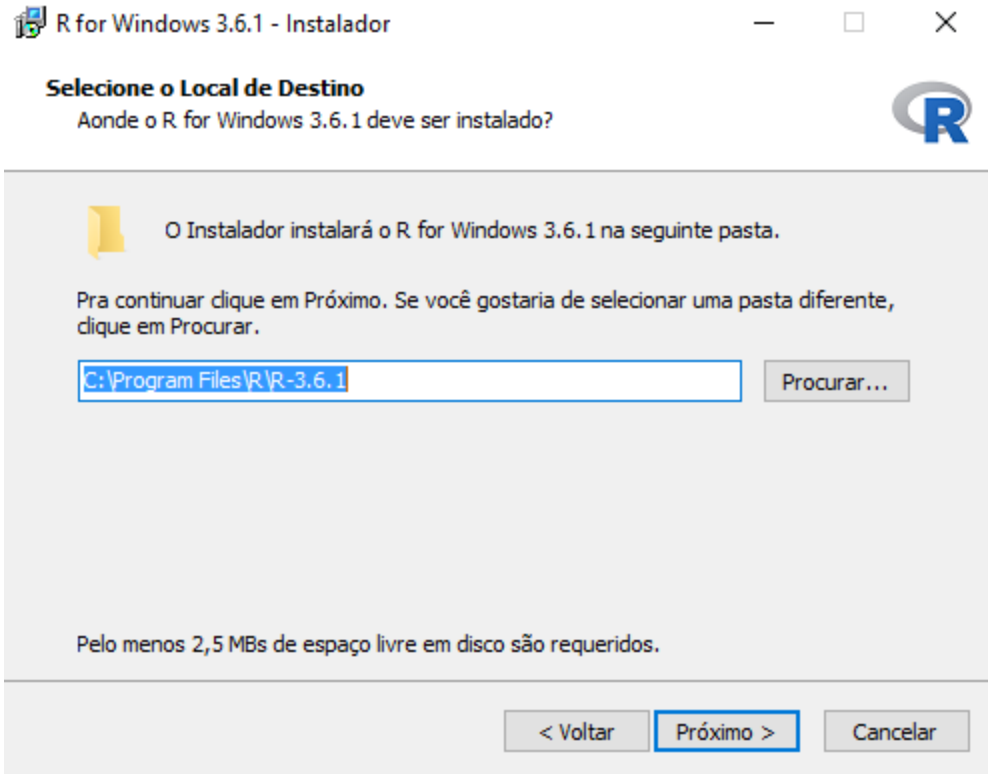


Figura 1.8: Pasta de instalação

O próximo passo será selecionar os componentes a serem instalados. Sugere-se marcar todos os componentes.

No entanto, é necessário se atentar a quantidade de espaço disponível no seu computador.

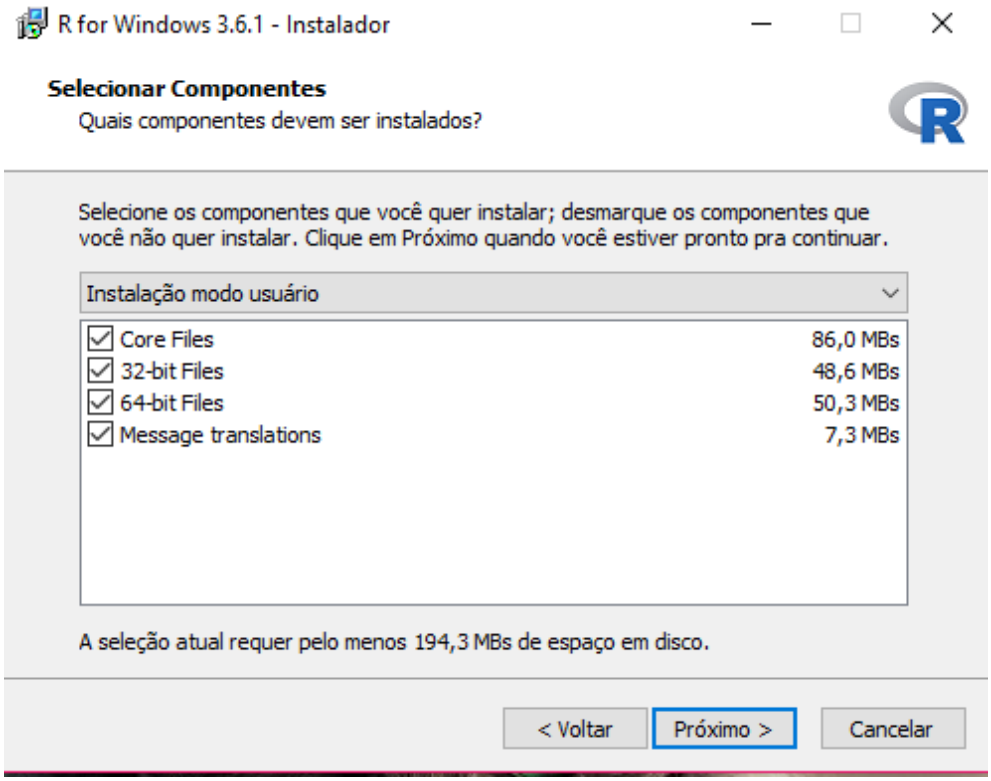


Figura 1.9: Componentes de instalação

A seguinte aba irá pergunta-lo sobre as opções para personalizar a inicialização. Para usuários iniciantes, sugerimos utilizar a padrão.

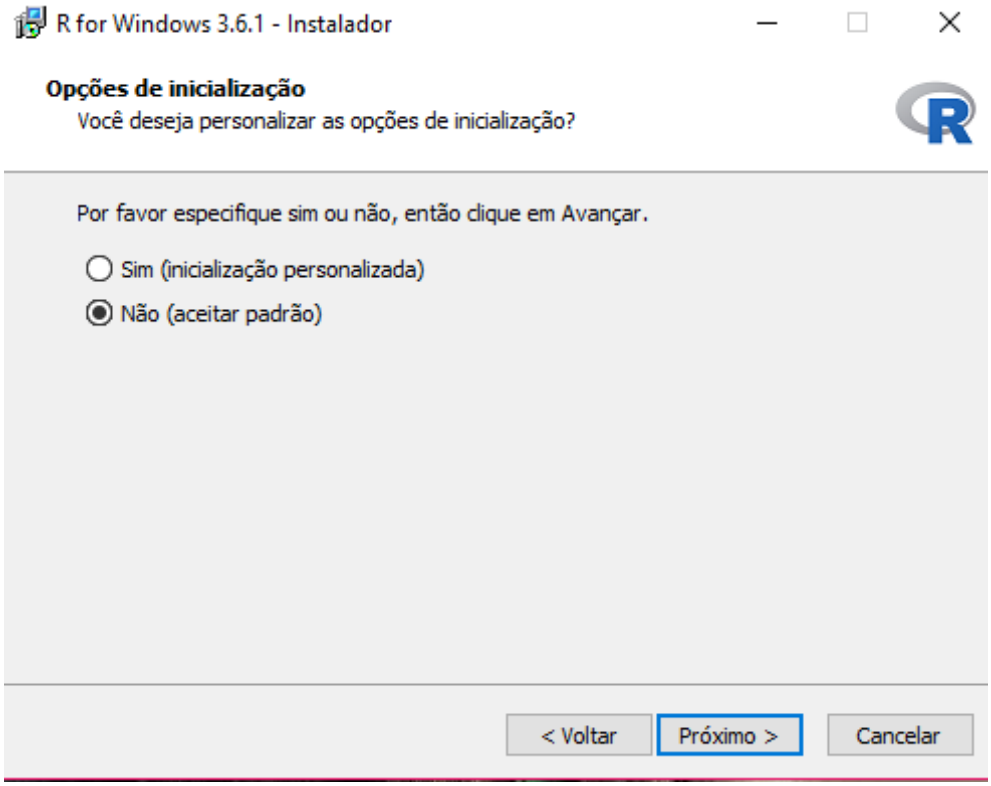


Figura 1.10: Opções de inicialização

O próximo passo será indicar a pasta onde será criado o atalho do programa.Caso queira, o usuário selecionar uma pasta diferente da indicada. Feito isso, o usuário deve clicar em próximo.

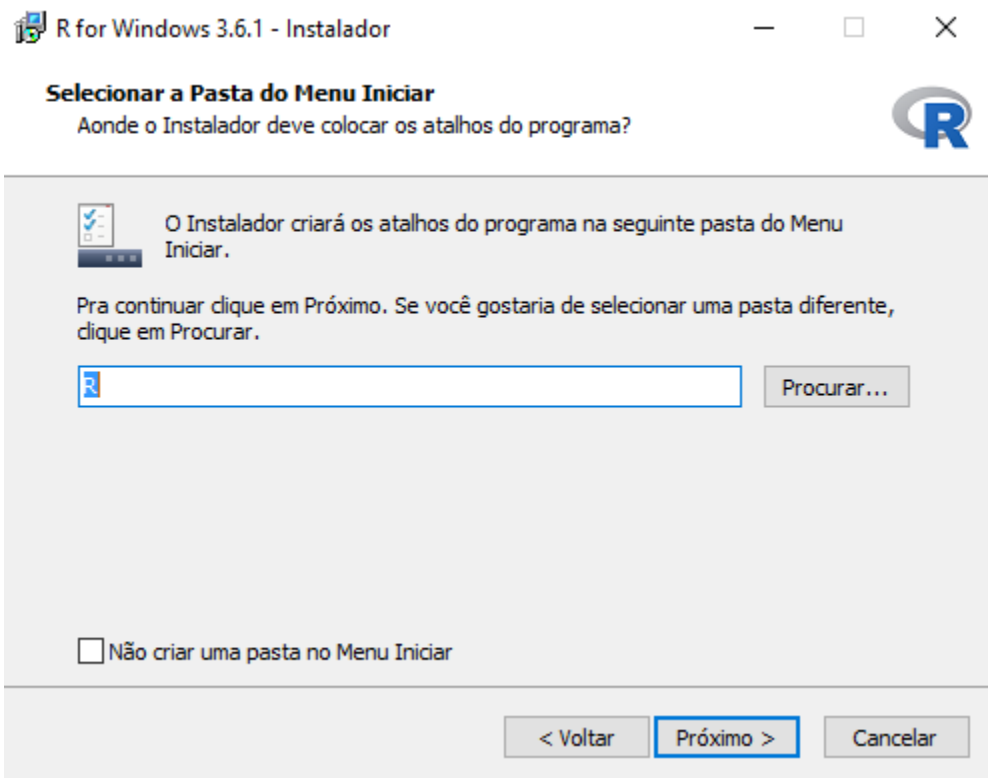


Figura 1.11: Pasta de atalho

A próxima aba perguntará ao usuário quais tarefas adicionais devem ser excutadas durante a instalação do R.

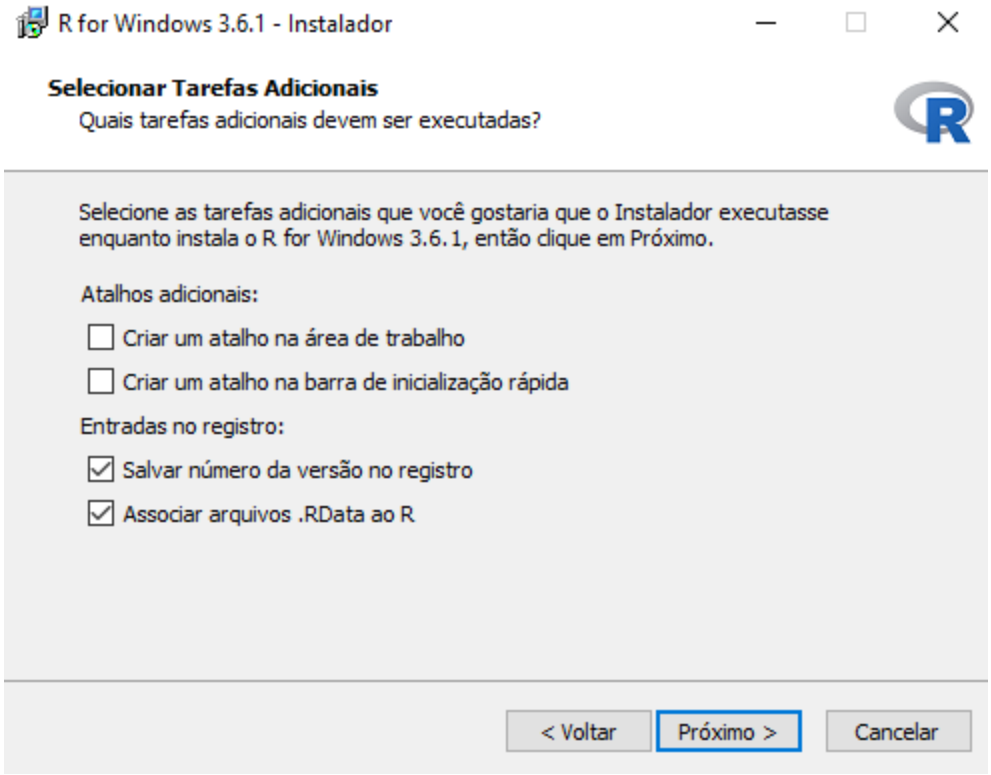


Figura 1.12: Tarefas adicionais

O próximo e ultimo passo será concluir a instalação. Para isso, o usuário deve clicar em "Concluir".

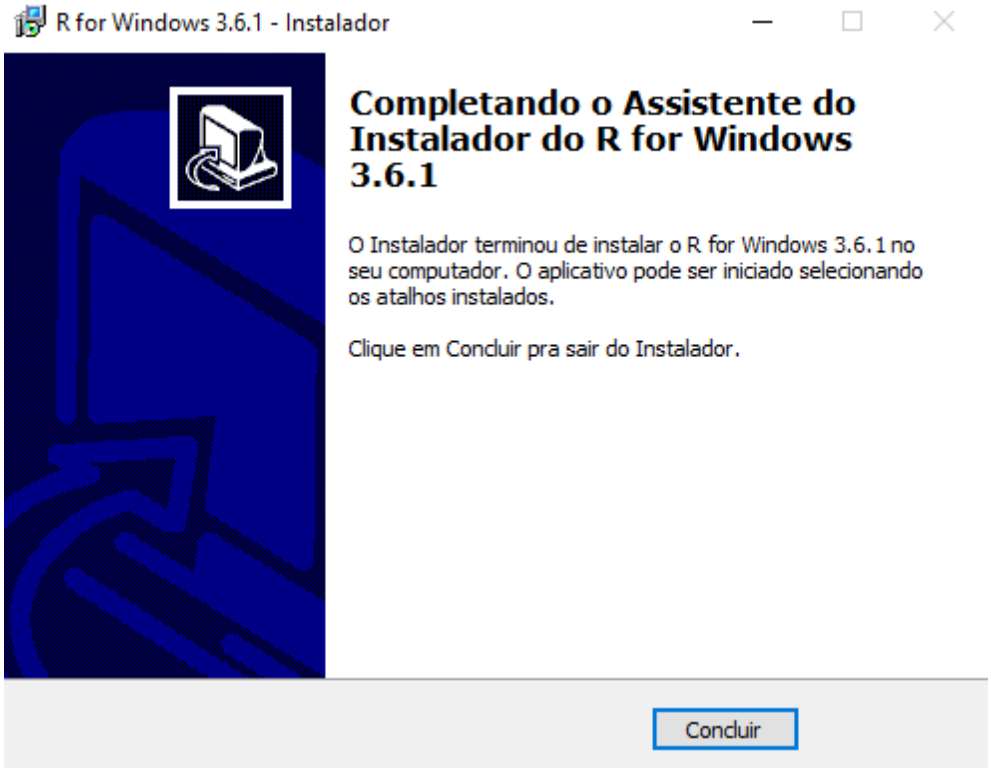


Figura 1.13: Tarefas adicionais

Ao clicar no arquivo R, presente na pasta indicada para instalação uma tela inicial será aberta. Pronto o software R já está pronto para o uso no seu computador.

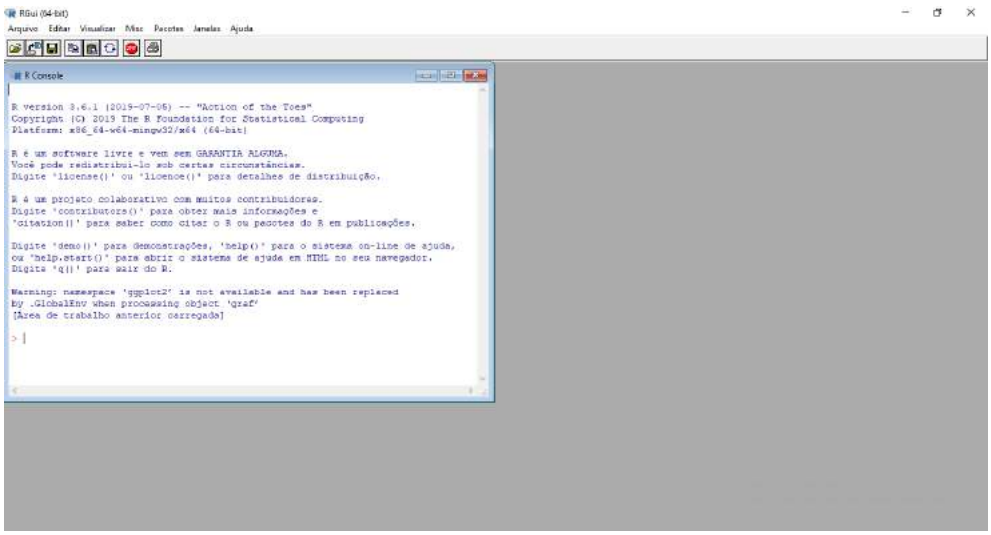


Figura 1.14: Tarefas adicionais

1.1.1 Instalando o RStudio

O RStudio é um software livre de ambiente de desenvolvimento integrado para R, uma linguagem de programação para gráficos e cálculos estatísticos.

O RStudio está disponível na internet a partir do link www.rstudio.org.

Ao inserir este link no seu navegador, o usuário será direcionado para a seguinte página inicial:

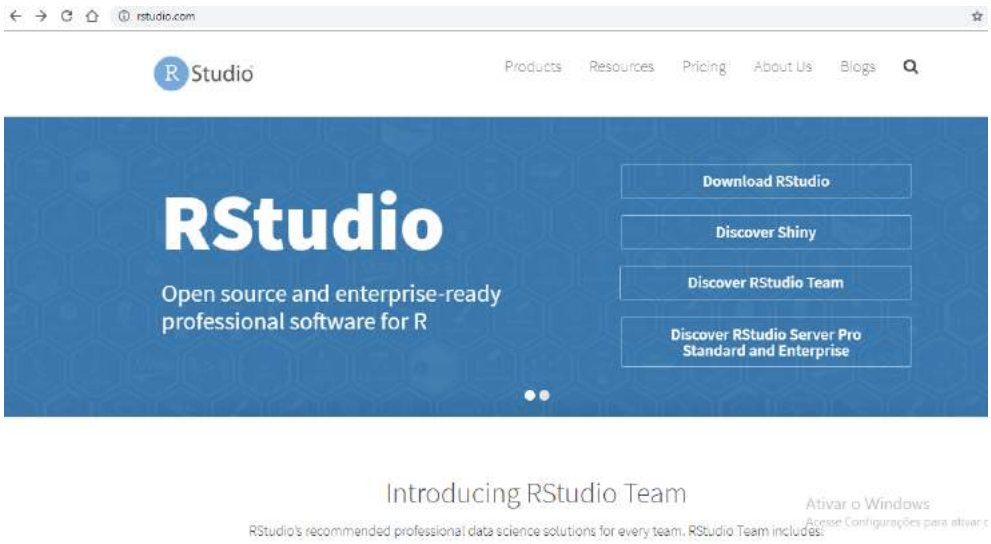


Figura 1.15: Página inicial

O usuário deve clicar em "Download RStudio".

Duas opções de arquivos de versão livre do RStudio serão disponibilizados. Sugere-se escolher a versão Desktop.

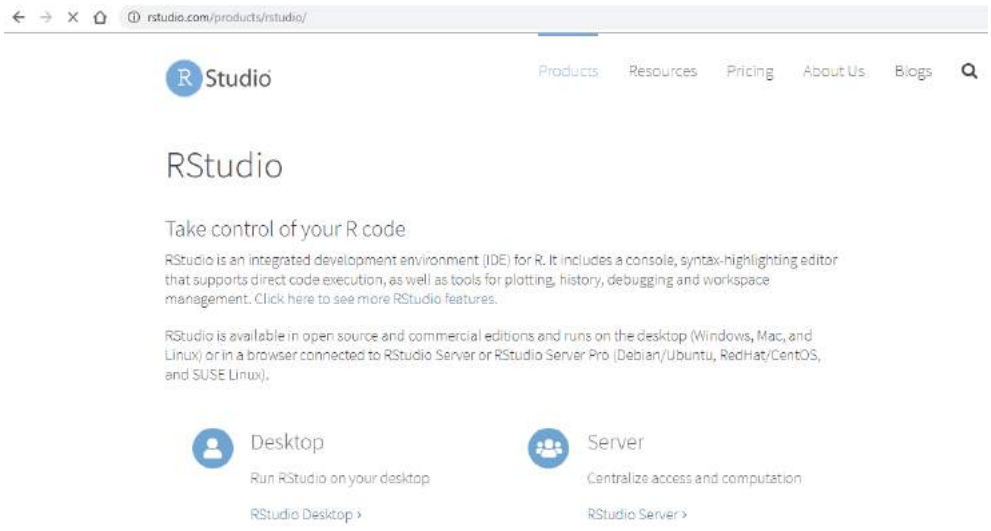


Figura 1.16

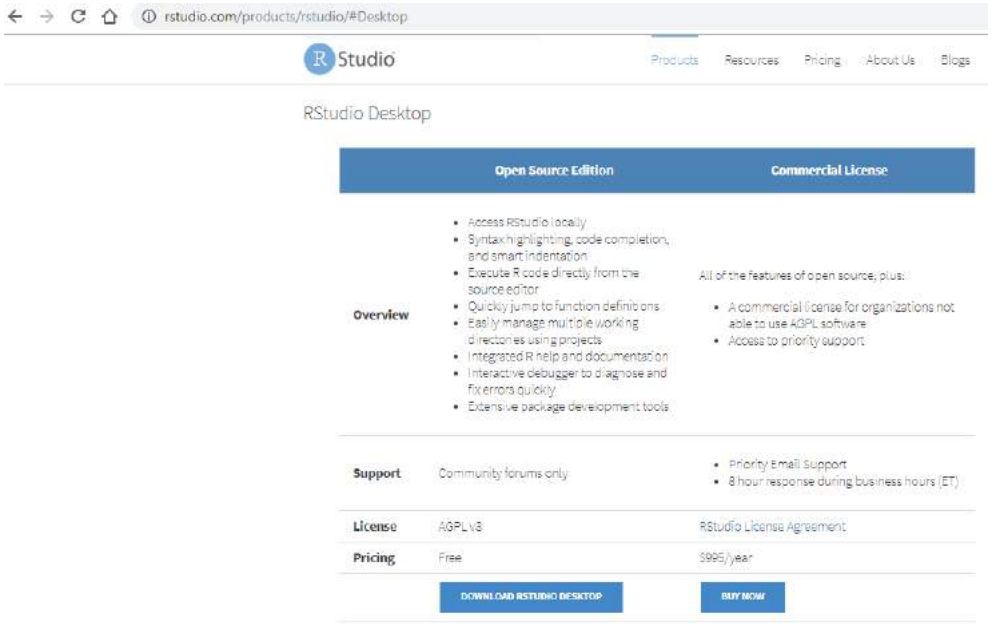


Figura 1.17

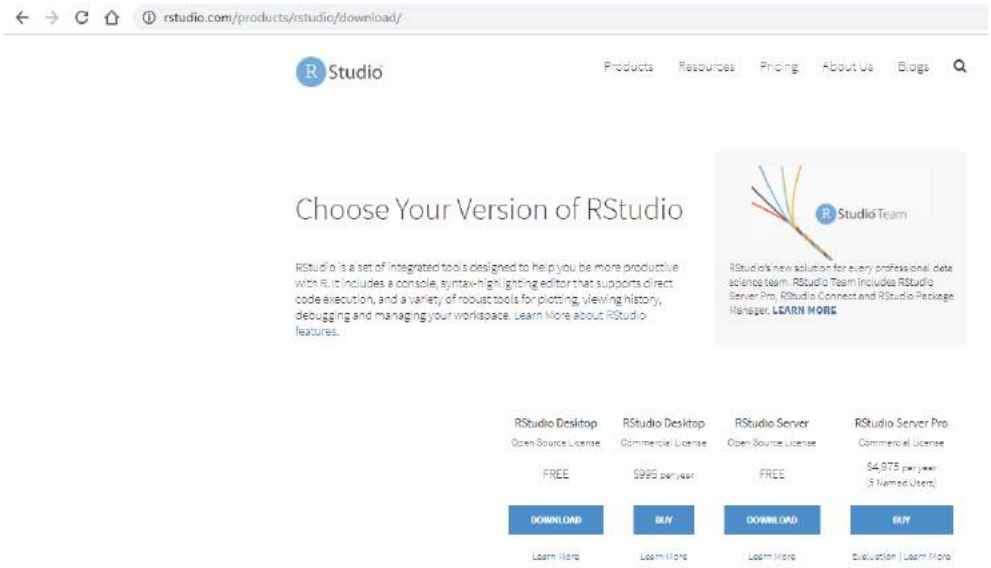


Figura 1.18

O próximo passo será selecionar o arquivo instalador a ser baixado. Para usuários, o arquivo selecionado deve ser o primeiro que aparece na imagem abaixo.

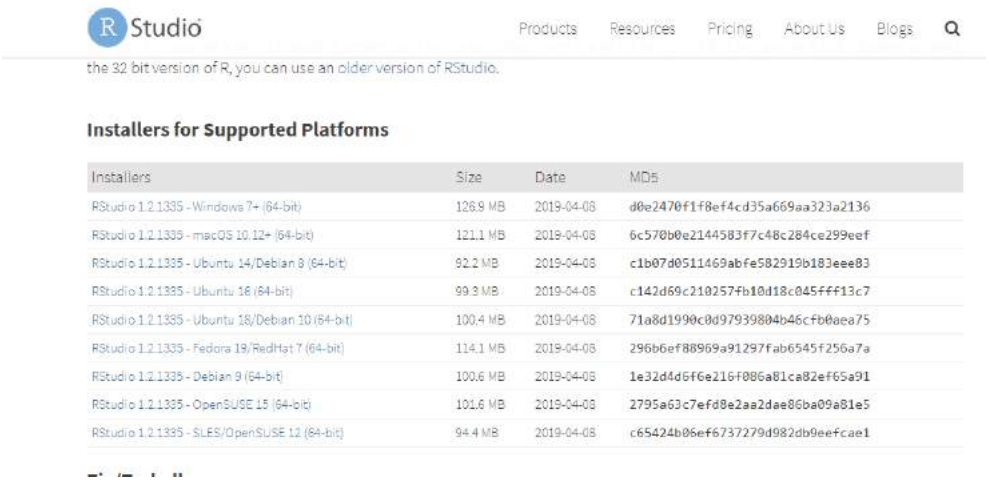


Figura 1.19

Um arquivo executável aparecerá sob a sua barra de tarefas.

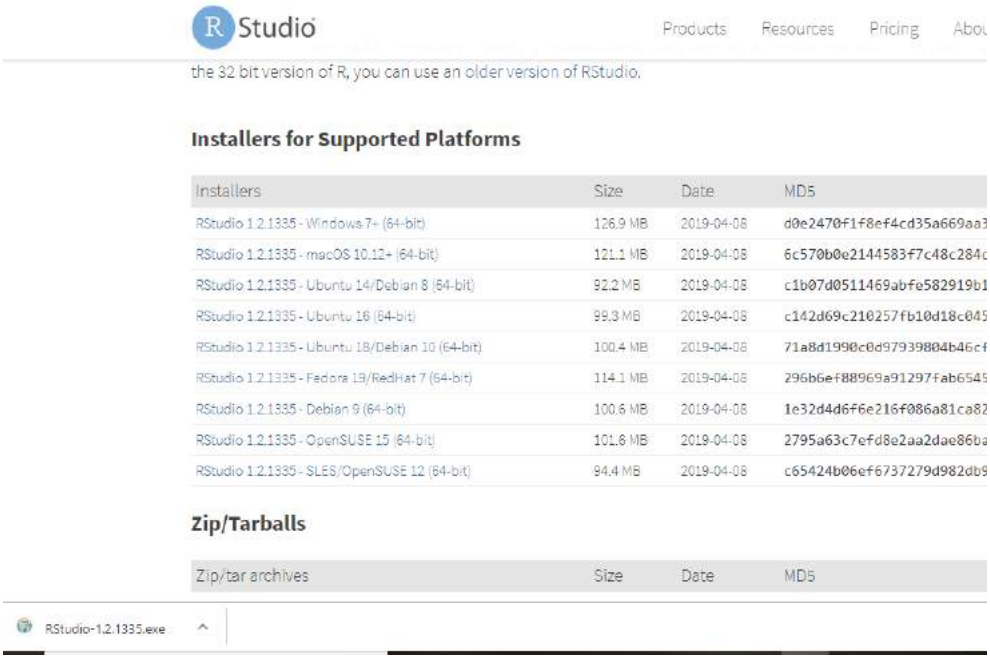


Figura 1.20: Arquivo executável

O próximo passo será clicar neste executável e iniciar a instalação do RStudio.

Uma janela de execução será aberta e o usuário deve clicar em "Próximo".

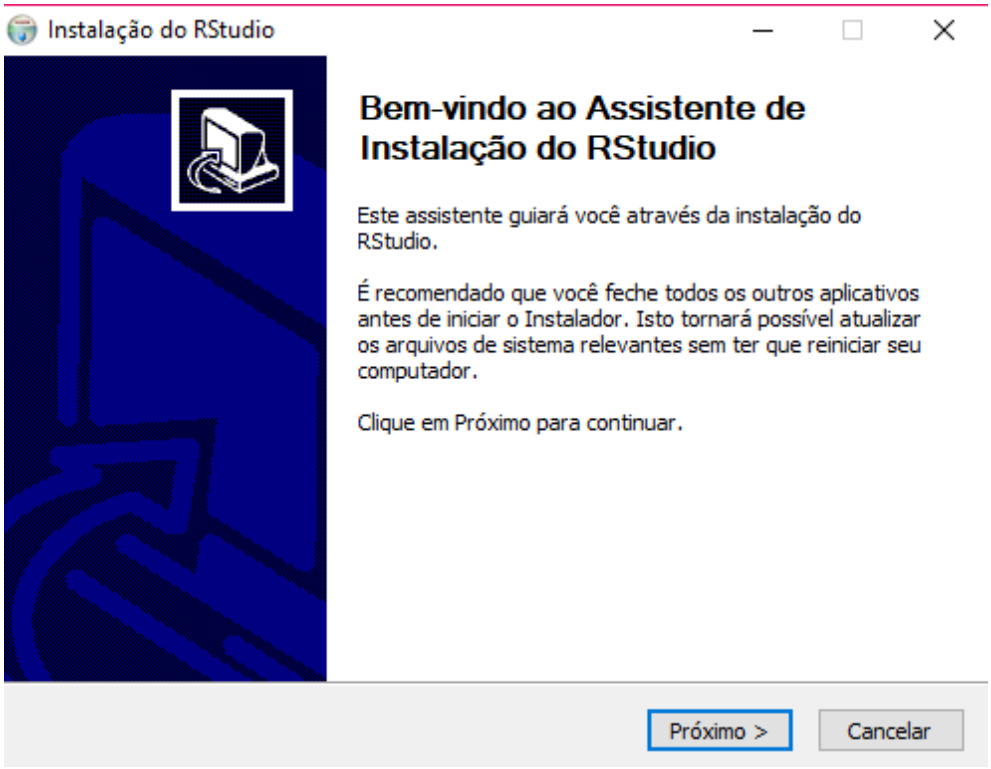


Figura 1.21: Instalação

O próximo passo é selecionar a pasta onde o RStudio será instalado. Uma vez selecionada a pasta de preferência, o usuário deve clicar em "Próximo".

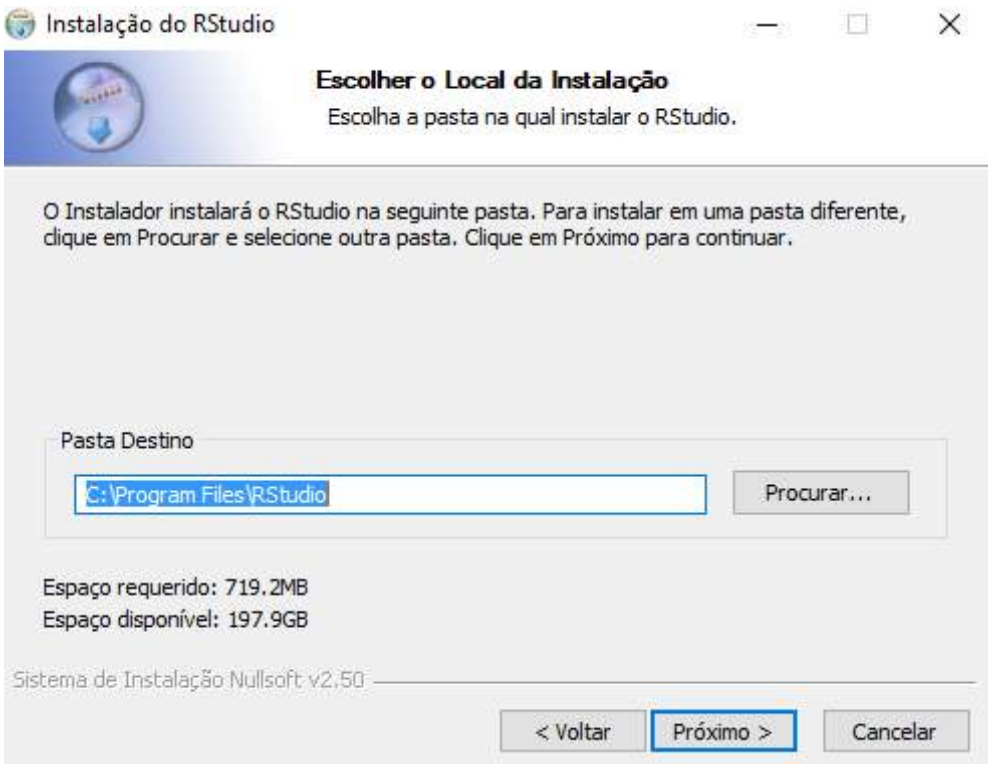


Figura 1.22: Pasta de instalação

O próximo passo será indicar a pasta onde serão criados os atalhos do programa.Caso queira, o usuário selecionar uma pasta diferente da indicada. Feito isso, o usuário deve clicar em "Instalar".

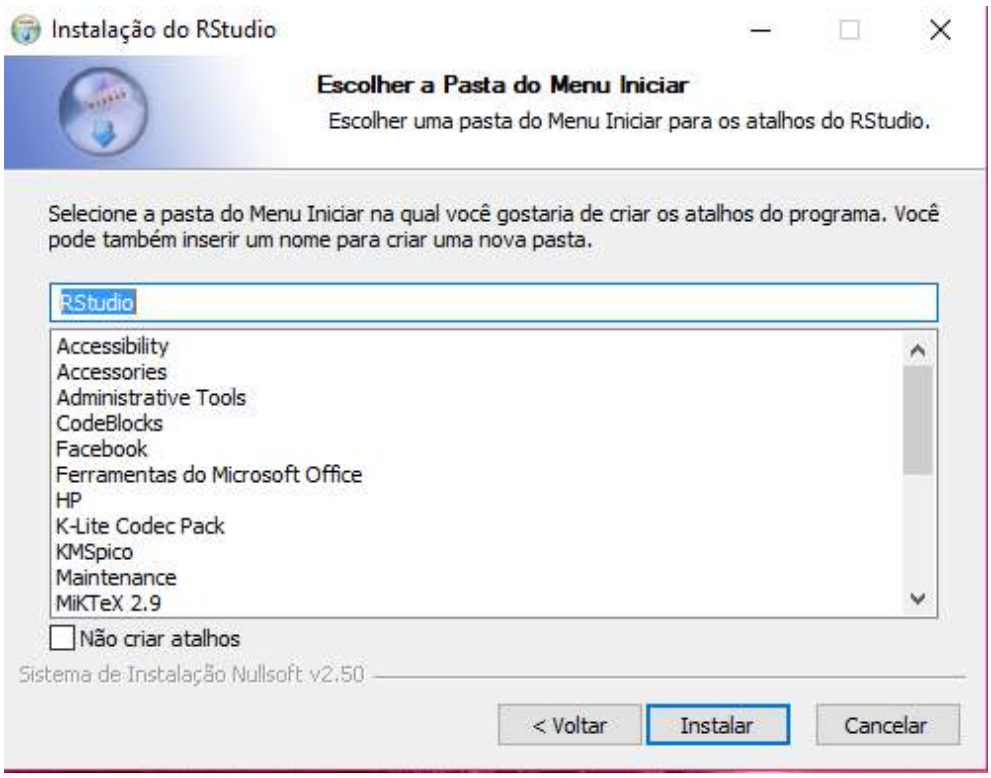


Figura 1.23: Pasta de atalhos

Feito isso, o usuário deve aguardar a instalação do programa. Ao fim, aparecerá a seguinte janela e o usuário deve clicar em "Terminar".

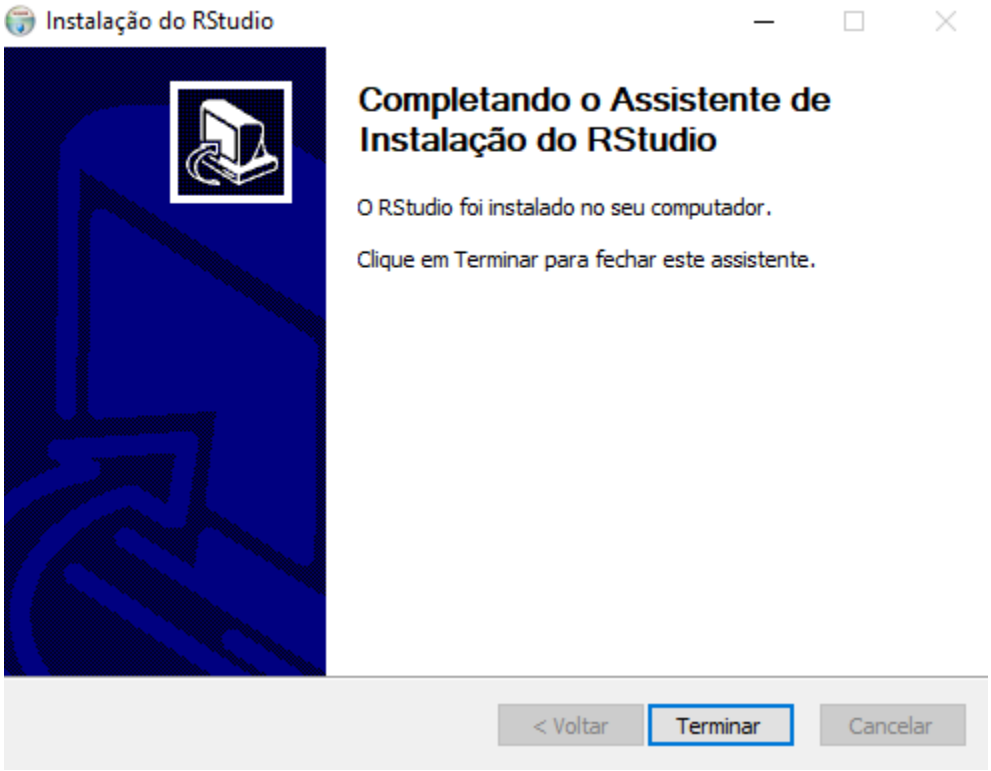


Figura 1.24: Conclusão da instalação

Ao clicar no arquivo RStudio, presente na pasta indicada para instalação uma tela inicial será aberta. Pronto o software RStudio já está pronto para o uso no seu computador.

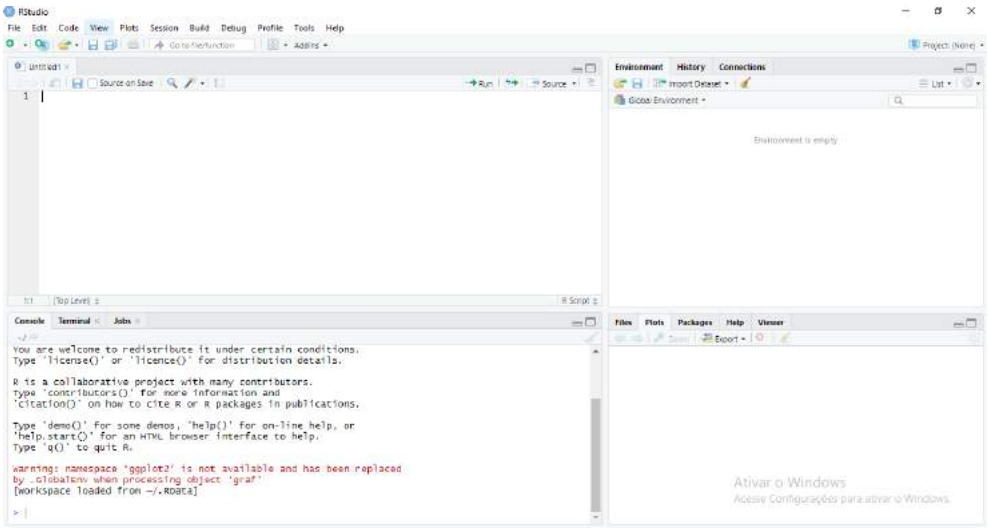


Figura 1.25: Tela inicial do software

1.2 Intalando o R Commander

Esta seção descreverá como baixar e instalar o R Commander.

1.2.1 Instalando o R Commander no R

Para baixar e instalar o pacote Rcmdr no R ou no RStudio, abra o software e digite no console "install.packages ("Rcmdr")". Aparecerá uma tela pedindo que selecione um "CRAN mirror"e escolha um deles. Caso a instalação não funcione, feche o R/Rstudio e tente novamente, dessa vez com outro CRAN .

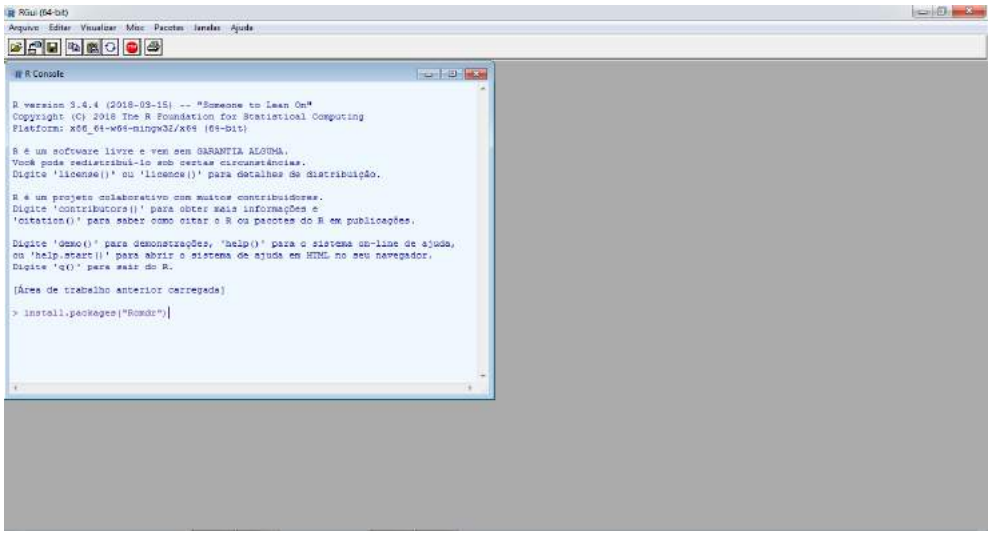


Figura 1.26: Instalação do Pacote Rcmdr no R

1.2.2 Instalando o R Commander no RStudio

Na janela direita inferior do R studio, selecione a opção "Packages", e em seguida na aba install.

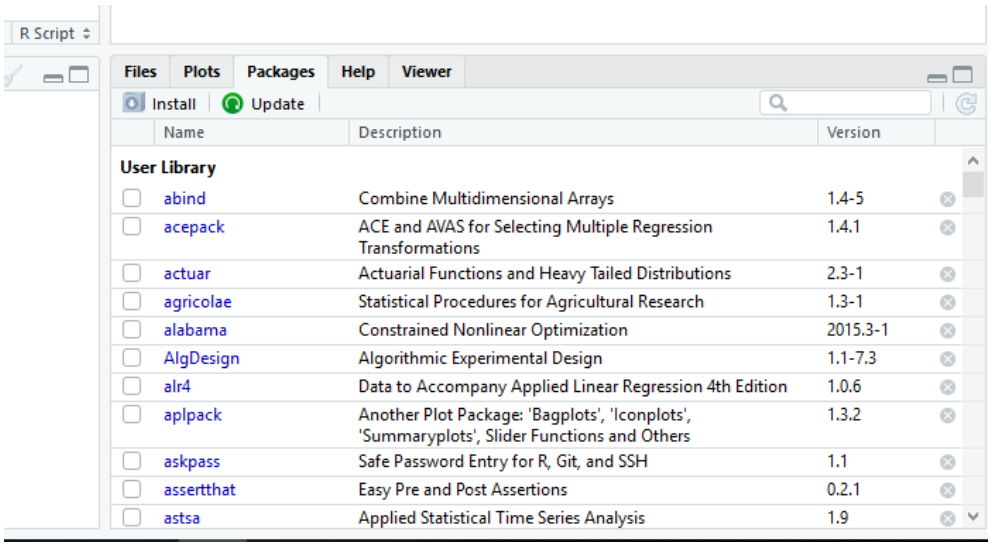


Figura 1.27: Instalação do Pacote Rcmdr no R

Após selecionar a aba install, o Rstudio abrirá a seguinte janela

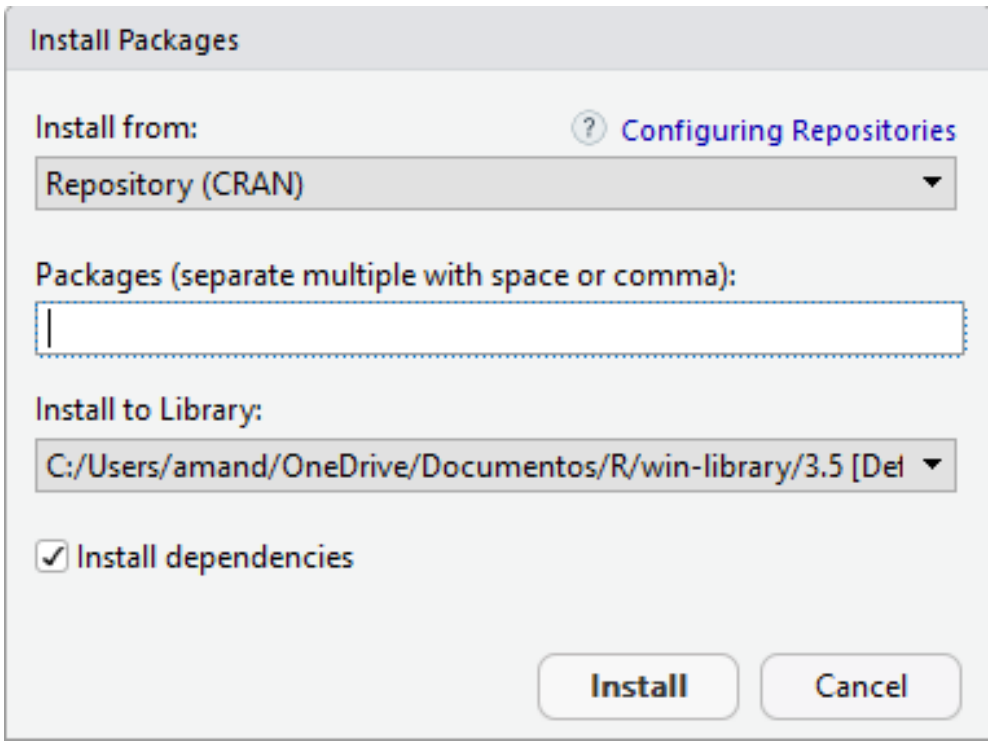


Figura 1.28: Instalação do Pacote Rcmdr no R

Digite "Rcmdr"e em seguida clique em install.

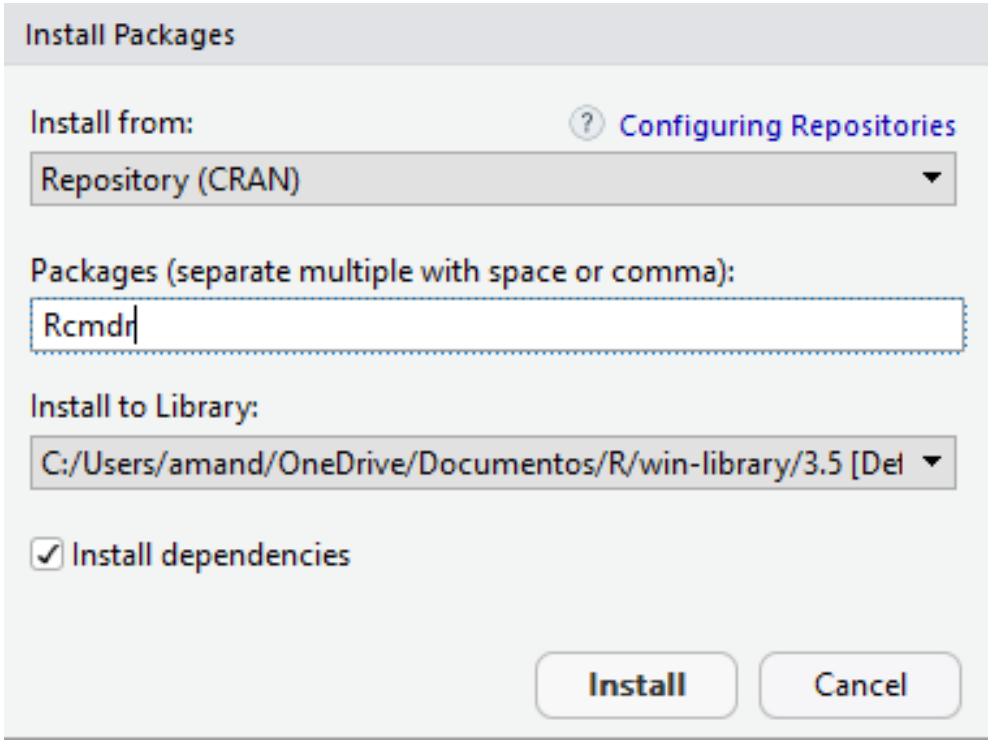


Figura 1.29: Instalação do Pacote Rcmdr no R

Capítulo 2

Conhecendo o R commander

Este capítulo apresenta a interface gráfica do usuário (GUI) do R Commander demonstrando seu uso para um problema simples: construir uma tabela de contingência para examinar o relacionamento entre duas variáveis categóricas. No desenvolvimento do exemplo, explicaremos como iniciar o R Commander e descreveremos a estrutura de sua interface , mostraremos como ler dados e prepará-los para análise, como fazer gráficos, calcular resumos numéricos de dados, criar um relatório impresso do seu trabalho, editar e reexecutar comandos gerados pelo R Commander e como encerrar sua sessão - em resumo, o fluxo de trabalho típico da análise de dados usando tal pacote. Também explicamos como personalizar a interface do R Commander.

Nesta fase inicial o intuito é mostrar o funcionamento da interface, mostrando uma visão geral de sua operação. Mais adiante, voltaremos com detalhes em muitos tópicos mostrados neste capítulo.

2.1 inicializando o R commander

Assumindo que você já tenha instalado o R e baixado o pacote "Rcmdr". Para inicializar o R commander, vá ao console do R, ou Rstudio, e digite: library(rcdmr)

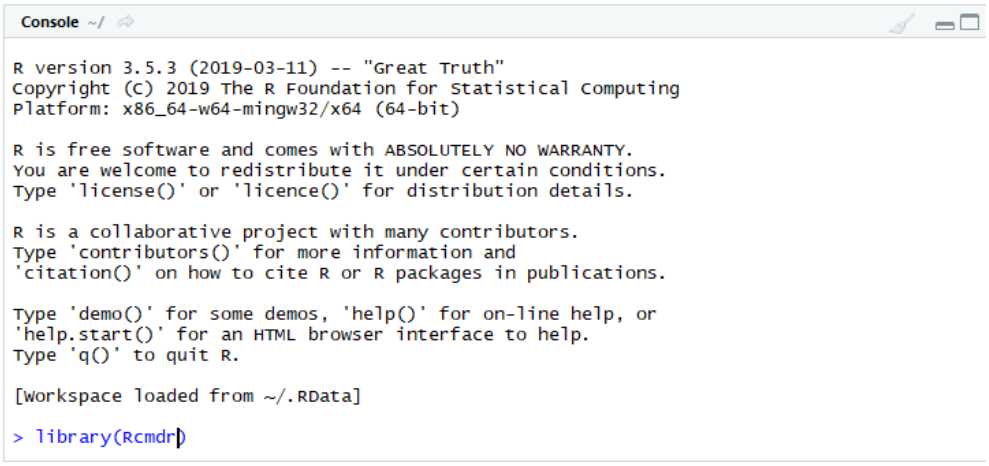


Figura 2.1: Carregando o Pacote Rcmdr

Após digitar a tecla enter, o pacote será carregado, caso precise da atualização ou instalação de algum outro

pacote, o R abrirá uma caixa de diálogo, onde você deverá selecionar a opção instalar. Caso todos os pacotes necessários estejam atualizados, a interface do Rcmdr será finalmente aberta.

Outra opção, que serve para quando a janela do Rcommander ja foi fechada alguma vez é utilizar o comando "Commander()". Ele também irá abrir a janela inicial do Rcommander.

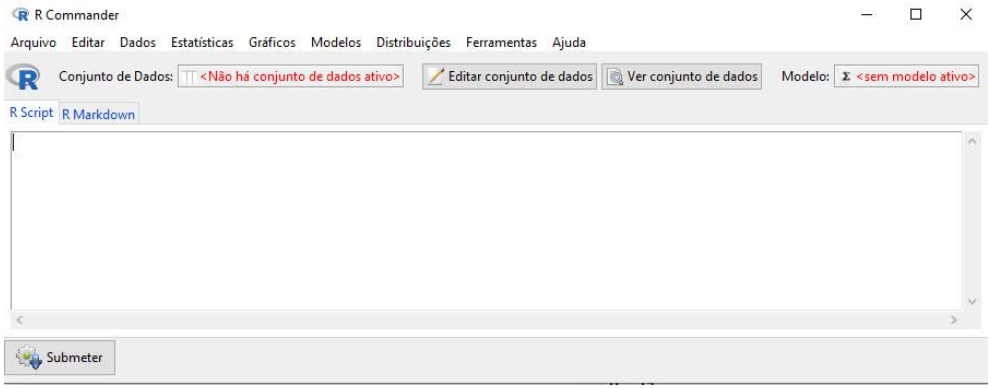


Figura 2.2: Interface inicial do R commander

2.2 A interface do R commander

Na parte superior da janela do R Commander, há uma barra de menus com os seguintes itens:

Arquivo: Parte onde estarão todas as opções relacionada a arquivos, como abrir e salvar vario tipos de arquivo, além de alterar o diretório de trabalho atual(Diretório é a pasta em que se sistema procurara e escreverá os arquivos)

Editar: Seção destinada a edição de textos, contém opções como "copiar"e "colar"além de opções de edição do arquivo R markdown (Isso será discutido posteriormente)

Dados: Seção onde encontraremos menus e submenus para importação, exportação e manipulação de dados.

Estatísticas: Contém submenus para várias análises estatísticas e ajustes de modelo.

Gráficos: Contém menus e submenus para criação de diversos gráficos.

Modelos: Seção com vários menus e submenus para executar varias operações em modelos que foram ajustados.

Distribuições: Seção onde pode-se gerar números aleatórios, ou provenientes de diversas distribuições de probabilidade

Ferramentas: Contém opções para carregar pacotes , plug-in(s) do R commander , além de outros softwares necessários.

Ajuda: contém itens de menu para obter informações sobre o R Commander e R, incluindo links para um breve manual introdutório e os sites R Commander e R; informações sobre o conjunto de dados ativo; e um link para um site com instruções detalhadas por usar o R Markdown para criar relatórios

2.3 Preparando o Rcommander

Na primeira vez que abrimos o Rcommander, não temos disponível a opção "Output"Que é onde veremos os resultados enquanto trabalhamos no Rcommander. Para ativar essa aba bata clicar em: Ferramentas>Opções. Na

janela de opções podemos personalizar vários aspectos do Rcommander, para nosso objetivo atual, selecionaremos a aba "Output". O primeiro passo, é desmarcar a opção: "Envie resultados para o Console do R"e em seguida, ajustar as alturas e larguras dos espaços. Recomenda-se as seguintes configurações:

Largura da janela do script(caracteres) = 80

Altura da janela do script(linhas) = 25

Altura da janela do output(linhas) = 30

Altura da janela de msgs(linhas) = 3

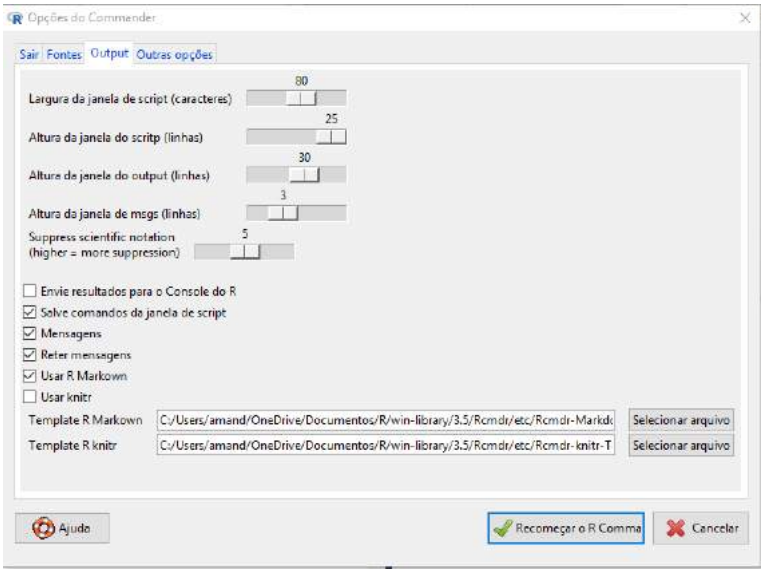


Figura 2.3: Janela de opções

Após ajustar as configurações necessárias, selecione a opção Recomeçar o Rcommander. O Rcommander então fechará e abrirá novamente com a nova página inicial.

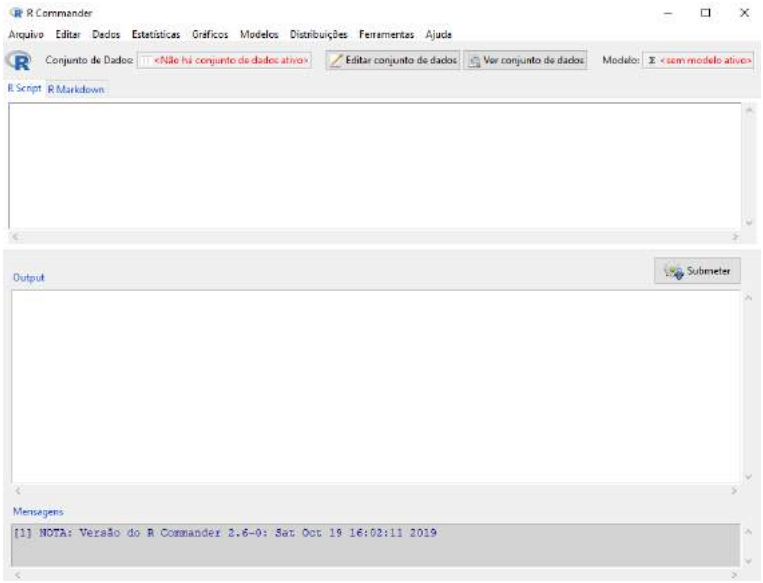


Figura 2.4: Interface inicial do R commander

Obs: A configurações acima são apenas sugestões, o usuário pode configura-la de forma que achar mais confortável para o uso.

Capítulo 3

Gerenciamento de dados

3.1 Lendo dados no R commander

A análise de dados estatísticos no R Commander é baseada em um conjunto de dados ativo, geralmente , os conjuntos de dados são organizados de forma que cada coluna represente uma variável e cada linha uma observação das variáveis. As colunas podem ser de diversos tipos, numéricas, caracteres, lógicos(T pra verdadeiro e F para falso) e fatores, que são usados para representar variáveis categóricas. Um conjunto de dados no R/ R commander é tratado como data frame. È possível trabalhar com quantos conjuntos de dados você precisar dentro do R.

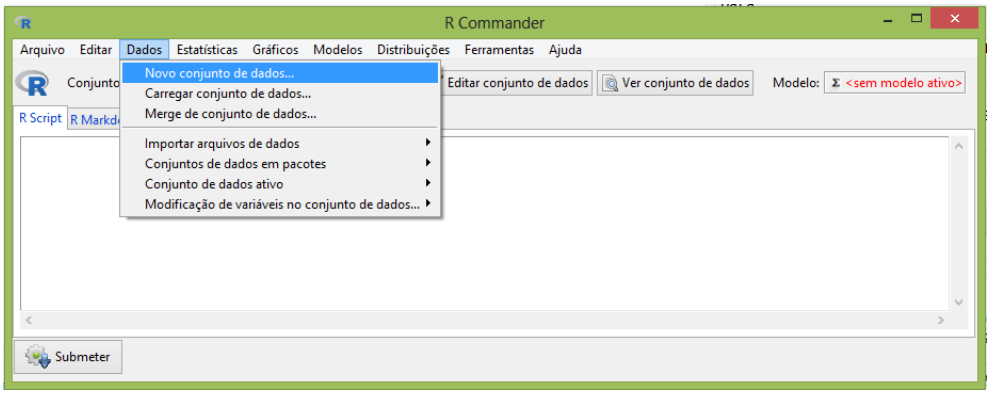
Através dos menus do Rcommander é possível ler conjuntos de dados de diversos formatos(txt, excel, SAS e outros), também é possível carregar bancos de dados contidos dentro de pacotes do R. Além de conseguir criar seu banco de dados digitando os valores.

3.1.1 Digitando os dados

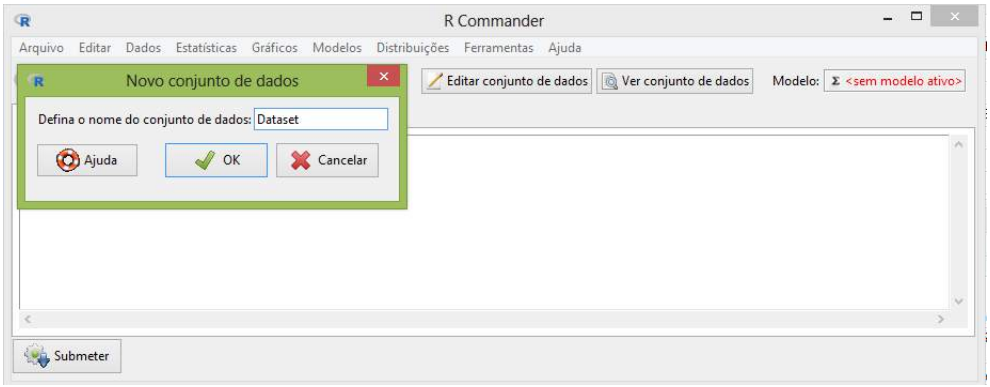
Para ilustrar essa maneira de inserir dados, iremos adotar o seguinte conjunto de dados fictícios.

Id	X	Y
1	4	7
2	7	6
3	6	3
4	5	2
5	7	1

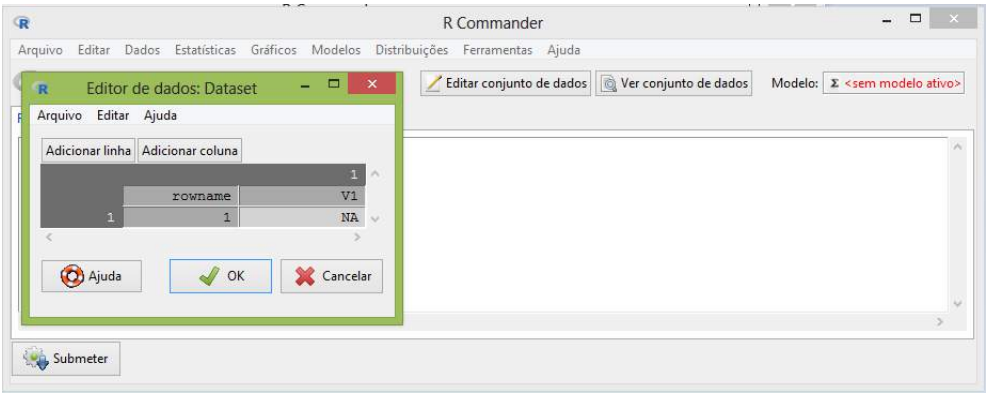
O passo a passo será descrito a seguir.



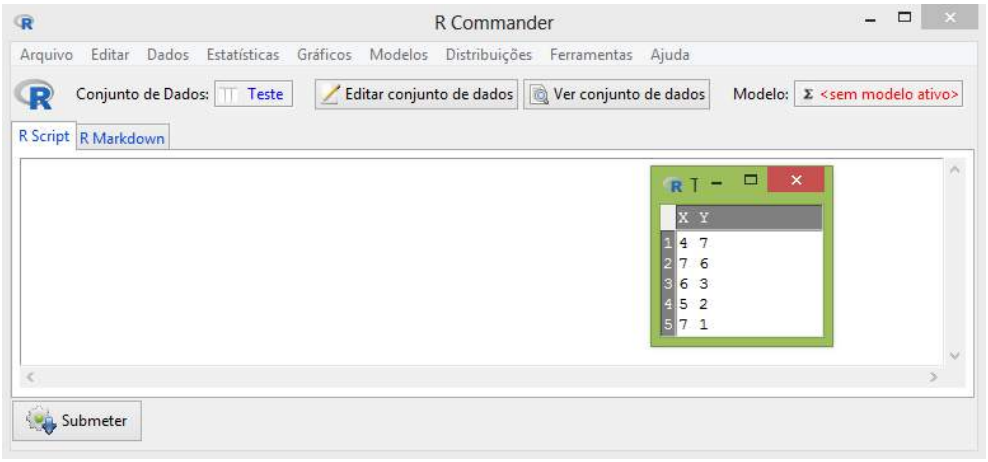
Na tela inicial do Rcomander, clique em dados e logo em seguida em "Novo conjunto de dados".



Insira o nome do conjunto de dados.



Nesta parte, edite o banco de acordo com a adequação. Inserindo novas linhas e colunas e até mesmo renomeando as colunas. Ao fim, clique em "OK".



Ao final do processo, clique em "Ver conjunto de dados"para visualizar os dados no formato inserido.

A seguir destacamos algumas observações importantes:

- Se uma coluna inserida no editor de dados consistir inteiramente em números, ela se tornará uma variável numérica. Por outro lado, uma coluna contendo dados não numéricos (com o exceção do indicador de dados ausentes NA) se tornará um fator. Dados de texto com os espaços em branco incorporados devem estar entre aspas (por exemplo, "concordo totalmente");
- O menu Editar no editor de dados suporta algumas ações adicionais, como excluir linhas ou colunas, e o menu Ajuda fornece acesso a informações sobre o uso do editor;
- Depois de digitar um conjunto de dados dessa maneira, é sempre uma boa idéia pressionar o botão View;

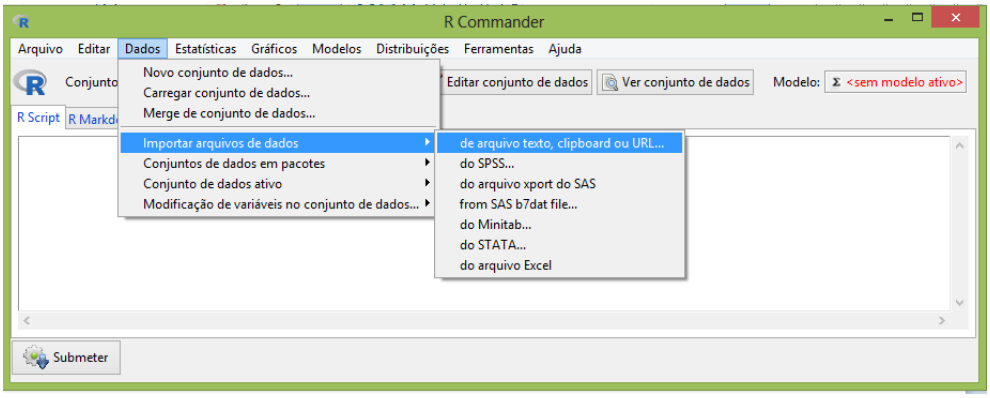
3.1.2 Importando dados em outros formatos para o R

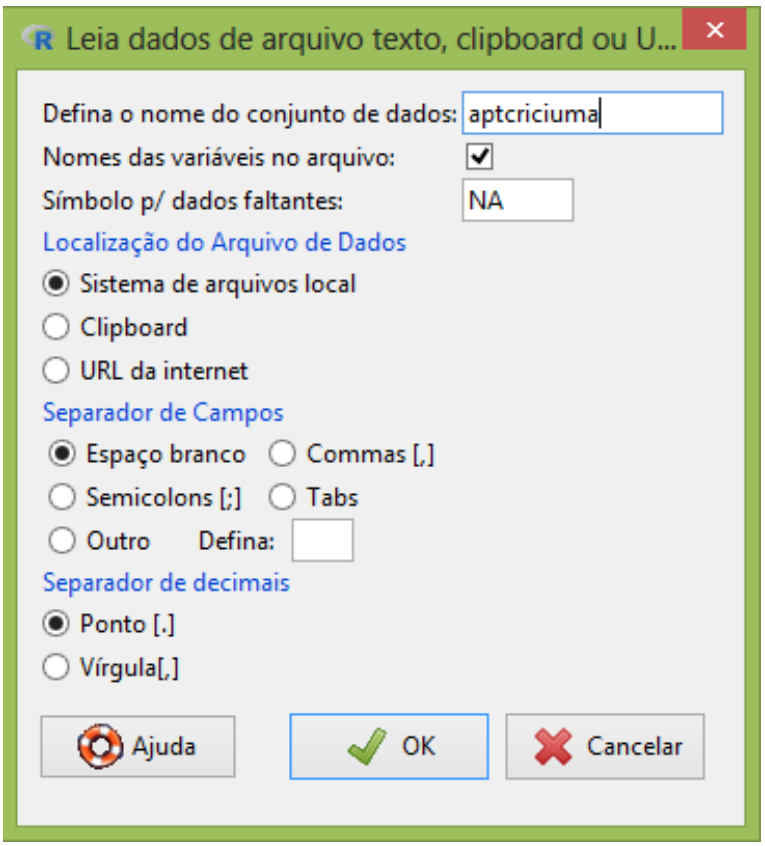
Importando dados em formato ".txt"

Para importar arquivos com em formato ".txt", na tela inicial do Rcomander, clique em Dados -> Importar arquivo de dados -> Arquivo de texto.

Feito isso, uma janela será aberta para que você possa editar as especificações do arquivo. Ao final clique em "OK"e você será direcionado para encontrar o arquivo do banco de dados em seu computador.

As imagens desse passo a passo são apresentadas a seguir:



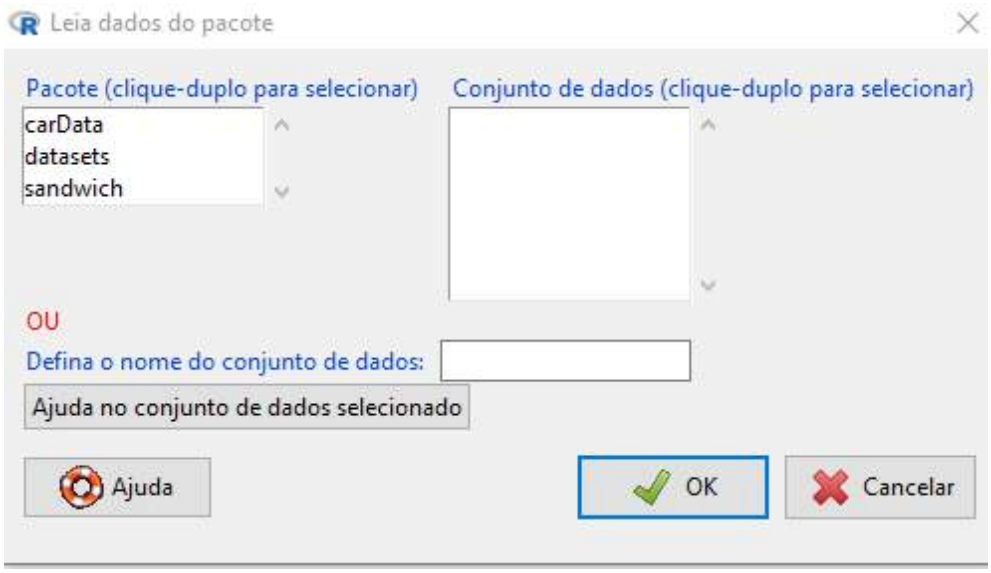


Importando dados em formato ".csv"

Para importar arquivos em formato em formato Excel utilize os mesmos passos anteriores, adaptando o tipo de arquivo clicando em "arquivos em excel".

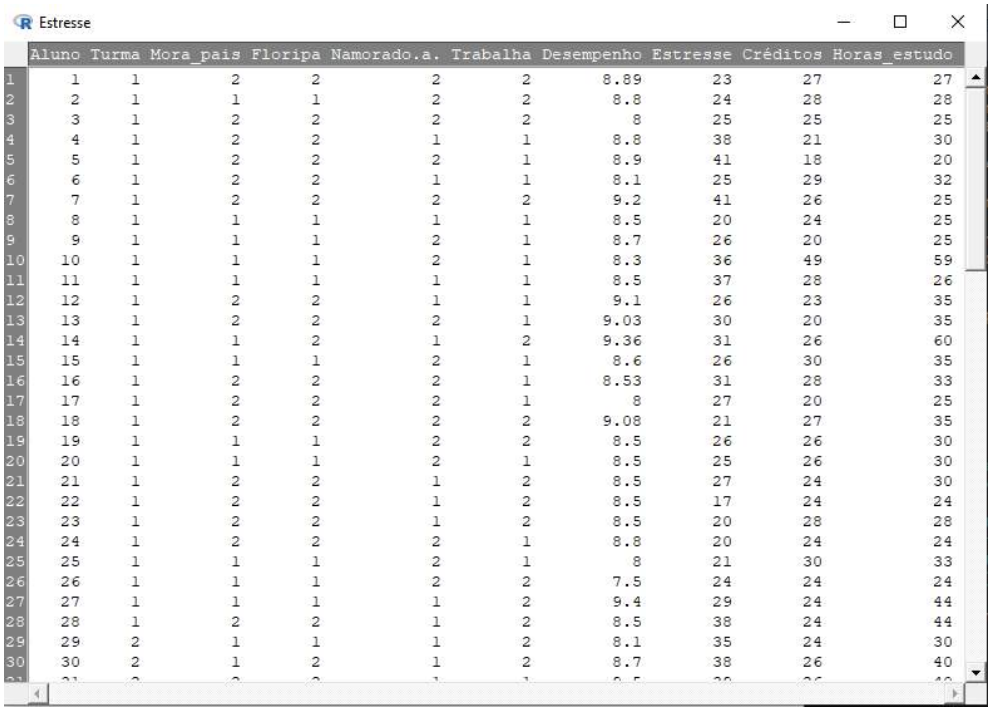
Carregando dados de pacotes do R

Para ler dados que estão em algum pacote do R, na tela inicial do Rcommander clique em: Dados > Conjuntos de dados em Pacotes > Ler dados de pacote "attachado". Na janela aberta, selecione o nome do pacote em que os dados estão presentes em "Pacote(clique-duplo para selecionar)"e em seguida na aba "Conjunto de dados(clique-duplo para selecionar") selecione o banco de dados desejado e em seguida clique em OK.



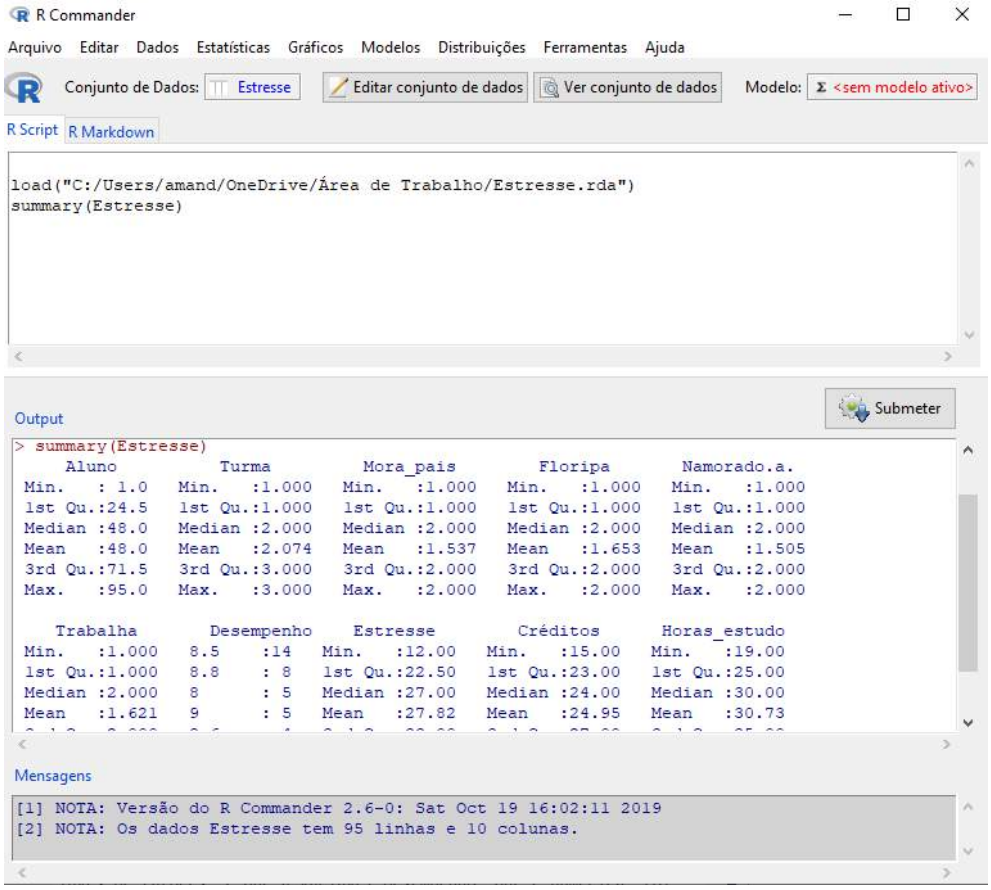
3.2 Examinando e re-codificando variáveis

Quando estamos lendo arquivos no R é recomendável que se olhe com calma para os dados antes de começar a usa-los, para que se tenha certeza de que eles foram lidos de maneira correta. Para exemplificar utilizaremos o banco de dados "questionarioEstresse"a ser disponibilizado junto a este manual. Para carregar o banco de dados reveja a seção "Importando dados em outros formatos para o R". Após carregar o banco de dados, na pagina inicial o Rcommander selecione a opção "Ver conjunto de dados, observe que nesta etapa, é aberta uma outra janela, onde espera-se que o banco de dados carregados, contenha os indivíduos nas linhas e as variáveis nas colunas. Ao abrir a janela de visualização dos dados, não é necessário fecha-la para trabalhar com os dados, caso a janela fique aberta, as alterações nos dados serão atualizadas na janela de visualização.

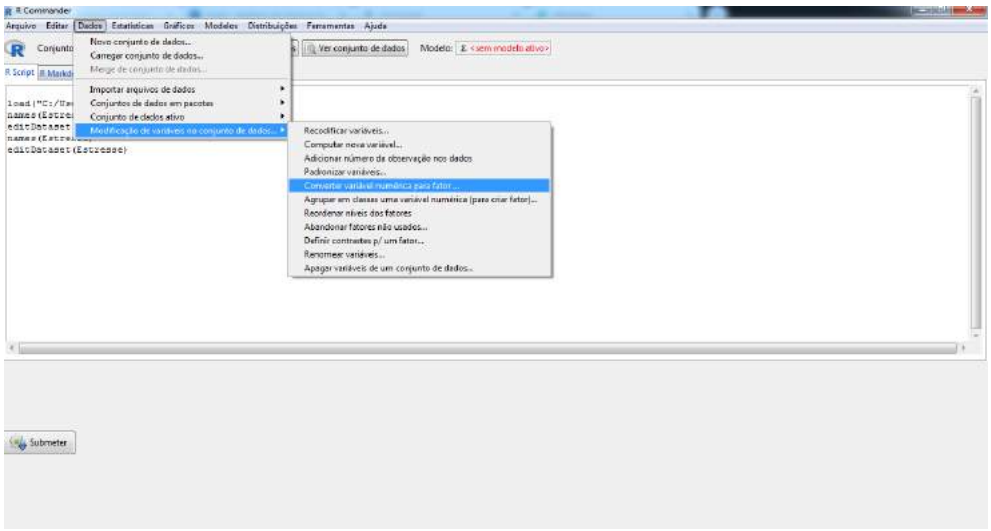

Estresse

	Aluno	Turma	Mora_pais	Floripa	Namorado.a.	Trabalha	Desempenho	Estresse	Créditos	Horas_estudo
1	1	1	2	2	2	2	8.89	23	27	27
2	2	1	1	1	2	2	8.8	24	28	28
3	3	1	2	2	2	2	8	25	25	25
4	4	1	2	2	1	1	8.8	38	21	30
5	5	1	2	2	2	1	8.9	41	18	20
6	6	1	2	2	1	1	8.1	25	29	32
7	7	1	2	2	2	2	9.2	41	26	25
8	8	1	1	1	1	1	8.5	20	24	25
9	9	1	1	1	2	1	8.7	26	20	25
10	10	1	1	1	2	1	8.3	36	49	59
11	11	1	1	1	1	1	8.5	37	28	26
12	12	1	2	2	1	1	9.1	26	23	35
13	13	1	2	2	2	1	9.03	30	20	35
14	14	1	1	2	1	2	9.36	31	26	60
15	15	1	1	1	2	1	8.6	26	30	35
16	16	1	2	2	2	1	8.53	31	28	33
17	17	1	2	2	2	1	8	27	20	25
18	18	1	2	2	2	2	9.08	21	27	35
19	19	1	1	1	2	2	8.5	26	26	30
20	20	1	1	1	2	1	8.5	25	26	30
21	21	1	2	2	1	2	8.5	27	24	30
22	22	1	2	2	1	2	8.5	17	24	24
23	23	1	2	2	1	2	8.5	20	28	28
24	24	1	2	2	2	1	8.8	20	24	24
25	25	1	1	1	2	1	8	21	30	33
26	26	1	1	1	2	2	7.5	24	24	24
27	27	1	1	1	1	2	9.4	29	24	44
28	28	1	2	2	1	2	8.5	38	24	44
29	29	2	1	1	1	2	8.1	35	24	30
30	30	2	1	2	1	2	8.7	38	26	40

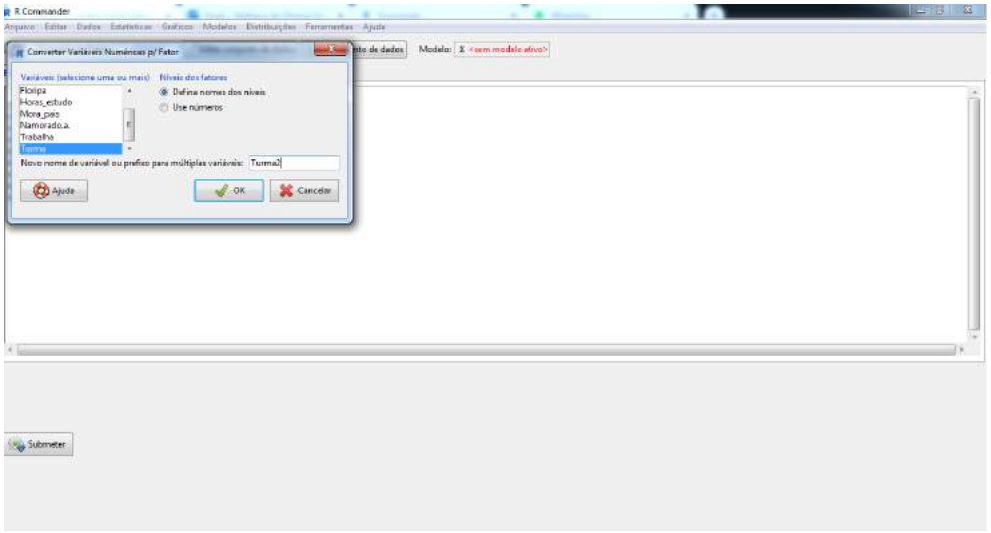
Desejamos observar então as descrições básicas das variáveis, para isso, selecione Estatísticas>Resumos>Conjunto de dados ativo. Note, na aba Output, que a maioria dos dados nos da informações do tipo: Min, moda, média... Isso por que os R, leu tais variáveis como variáveis numéricas. A Variável "Desempenho"foi lida como fator, por isso nela são mostradas as frequências, e não as medidas que são mostradas nas variáveis numéricas.

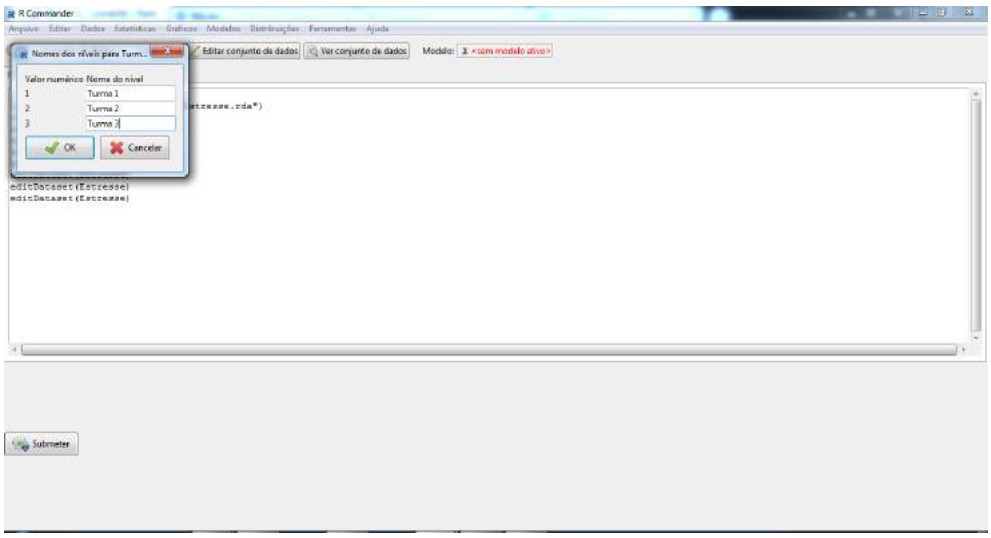


Olhando para a descrição do banco de dados, podemos ver que algumas variáveis que são categóricas, foram classificadas como numéricas, ao invés de fatores, e que a variável Desempenho, que é numérica, foi classificada como fator. Devemos então mudar a classe destas variáveis para que possamos usa-las de maneira adequada. Para transformar uma variável lida como numérica em fator (caso, por exemplo da variável Turma), devemos ir no menu "Dados>Modificação de Variáveis no conjunto de dados>Converter variável numérica p/fator"como pode ser visto na figura a seguir.



Será aberta uma janela, onde se deve selecionar a variável a ser transformada. Podemos escolher usar fatores numericos ou definir nomes para cada fator. Neste caso, transformaremos a variável Turma e definiremos os nomes "Turma 1","Turma 2", Turma 3"para 1,2 e 3 , respectivamente, e salvaremos isso nova variável "TurmaB"(Caso queira sobrescrever a anterior, devemos nomear a variável transformada com o mesmo nome da original). As imagens a seguir mostram o procedimento:





Agora para transformar uma variável lida como categórica em numérica, no caso, a variável Desempenho.

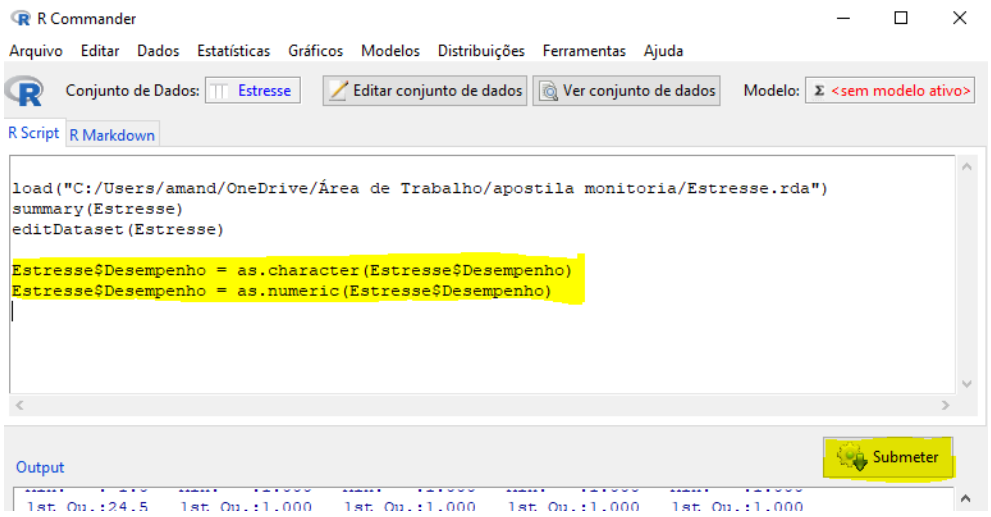
Nos casos de transformar as variáveis de fator para numérico, o método mais fácil será transformando-as em characteres e depois em numéricas, pois, o R lê cada fator como um numero, fazendo a conversão direta, perderíamos os valores originais e teríamos os valores dos fatores. Outro ponto importante é que esta conversão, mesmo no Rcommander, é mais facilmente feita por linha de comando. Observe que é possível escrever na parte do script, e é nesta parte que colocaremos os comandos

para acessar uma variável de um banco de dados , utilizamos a segunte extrutura: Nomedobanco$Variáveldesejada.

os comandos para transformar em charactere e em numéricos são respectivamente: as.character() e as.numeric(). Basta então atribuir os valores as variáveis da seguinte maneira:

Estresse$Desempenho = as.character(Estresse$Desempenho)

Estresse$Desempenho = as.numeric(Estresse$Desempenho)



Após escrever os comandos, selecione as linhas escritas e clique em submeter! Tendo feito isso, as variáveis que haviam sido lidas como categóricas, foram convertidas em numéricas. Para checar, olhe para as medidas de resumo dos dados(assim como foi feito anteriormente).

Fica como exercício para o leitor re-codificar as variáveis do banco para as seguintes especificações:

- **Numéricas:** Horas_estudo, Créditos, Estresse e Desempenho;
- **Aluno:** Use os próprios números como níveis;
- **Turma:** 1= 2007_2, 2 = 2008_1, 3 = 2008_2;
- **Mora_pais:** 1= Sim, 2= Não;
- **Floripa:** 1 = Natural, 2= Outra_Cidade;
- **Namorado.a.:** 1= Sim, 2= Não;
- **Trabalha:**1= Sim, 2= Não.

Capítulo 4

Estatísticas descritivas

Neste capítulo iremos tratar de fazer estatísticas descritivas básicas para conjuntos de dados,como resumos numéricos, tabelas de contingência e resumos numéricos simples.

Continuaremos usando o banco de dados de estresse, editado anteriormente, para apresentar as análises descritivas.

4.1 Resumo das variáveis

Através do menu Resumos é possível obter os mais diversos tipos de resumos para as variáveis.

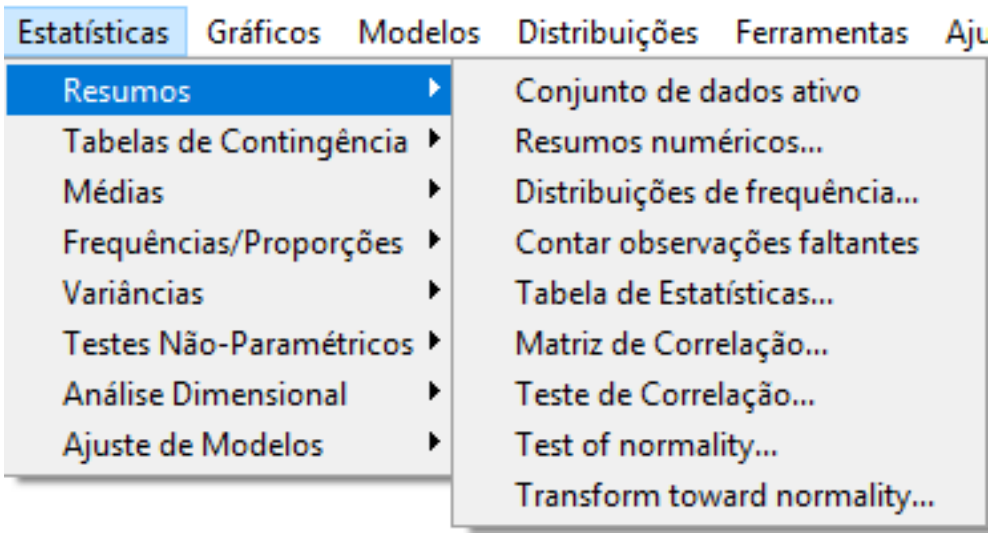


Figura 4.1: Menu de resumos

Quando começamos a analisar os dados é interessantes olhar para o resumo das variáveis numéricas, Selecionando: Estatística>Resumos> Resumos numéricos é aberta uma janela com duas seções, a seção Dados onde podemos selecionar as variáveis(pode-se selecionar só uma, ou várias clicando e arrastando com o mouse), e a seção Estatísticas onde podemos selecionar os resumos que desejamos observar para as variáveis numéricas, por exemplo: média, quantis, desvio padrão, intervalo interquantilico...

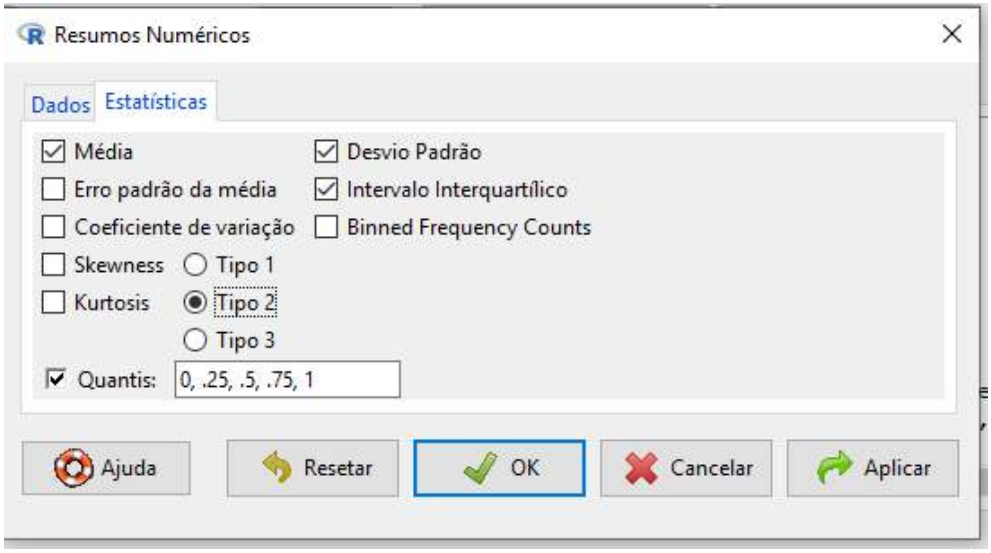


Figura 4.2: Janela de opções dos resumos numéricos

Ao selecionar as medidas desejadas, e clicar em OK, o Rcmdr exibirá as medidas solicitadas na janela de output. onde temos as saidas:

mean: A média da variável.

sd: O desvio padrão da variável.

IQR: O intervalo interquantilico(A diferença entre o terceiro e o primeiro quartil).

0%: È o Quantil 0% ou seja o mínimo da variável.

25%: È o quantil 25% ou seja o 1° quartil.

50%: È o quantil 50% ou seja a mediana.

75%: È o quantil 75% ou seja a o 3° quartil.

100%: È o quantil 100% ou seja o máximo da variável.

n: É a quantidade de respostas da variável

NA: É a quantidade dados faltantes na variável.

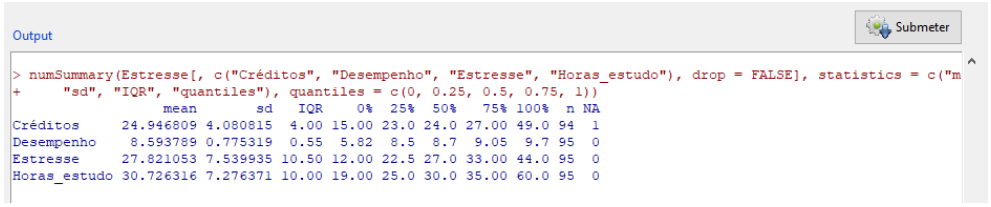


Figura 4.3: Output dos resumos numéricos

Para os fatores é interessante olhar para a distribuição de frequência dos níveis dos fatores, para isso vamos em : Estatísticas > Resumo > Distribuições de Frequência. Após isso será aberta a janela para selecionar a(s) Variável(is) de interesse:

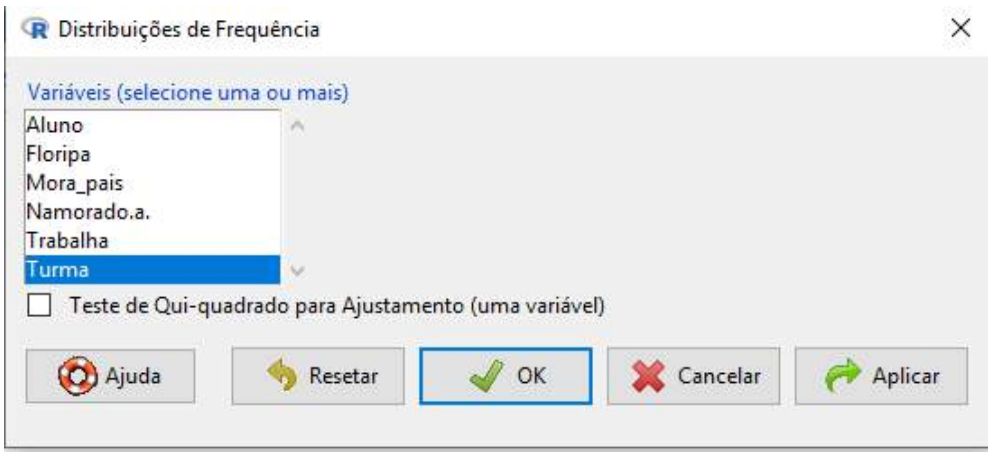


Figura 4.4: Janela de seleção de variáveis

Ao selecionar, por exemplo para a variável turma, o Rcommander nos mostra como saída a quantidade de indivíduos em cada nível do fator(frequência absoluta) e a porcentagem que cada fator tem do total(frequência relativa).

Figura 4.5: Distribuição de frequência da variável Turma

Outra opção interessante é observar os resumos das variáveis numéricas dentro de cada níveis dos fatores. por exemplo, seria interessante observar a média de horas de estudo em cada turma. Para isso, iremos em: Estatísticas > Resumos > Tabela de estatísticas. Onde será aberta a janela de seleção das variáveis e medidas:

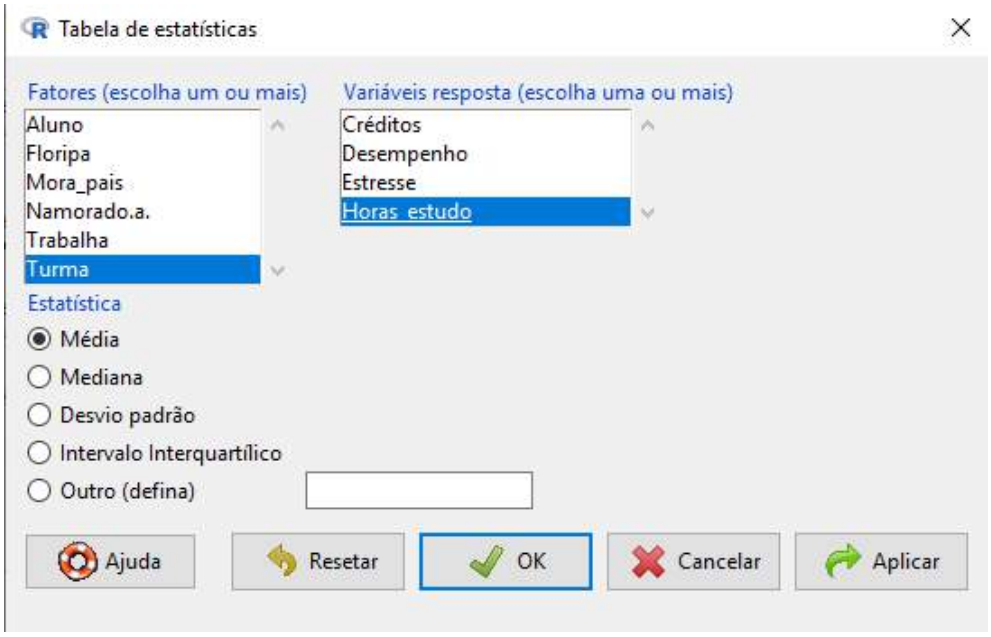


Figura 4.6: Tabela de estatísticas

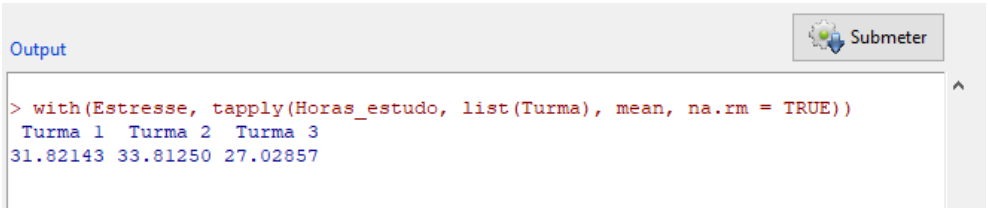


Figura 4.7: média das horas de estudo por turma

De modo geral, podemos ir através do Menu, em Estatística>Resumos>Conjunto de dados ativos(como ja foi feito em capítulos anteriores), para obter os resumos mais simples de todas as variáveis, onde para variáveis numéricas serão exibidos: mínimo,2º quartil, mediana, média, 3º Quartil e Máximo das variáveis e para os fatores, será exibida a distribuição da frequência dos níveis do fator, incluindo a contagem de NA.

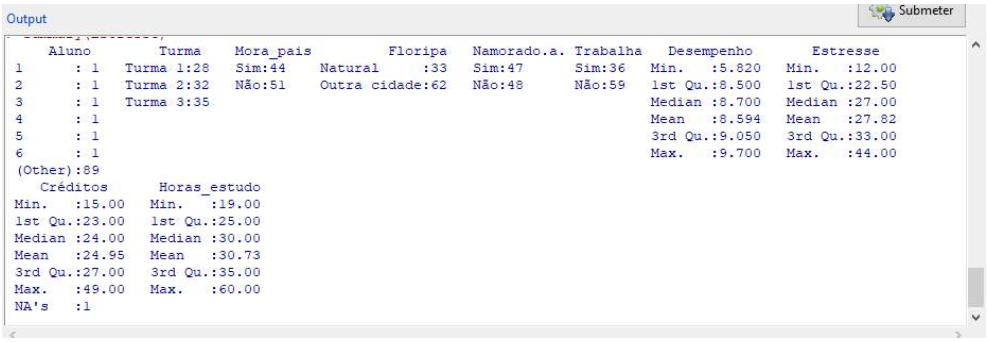


Figura 4.8: Output dos resumos em geral

4.2 Tabelas de contingência

Uma tabela de contingência é a tabela que calcula a quantidade de observações por múltiplas variáveis categóricas. As linhas e colunas das tabelas correspondem a essas variáveis categóricas.

O menu Estatísticas> Tabelas de contingência possui itens para construindo tabelas bidimensionais e multidimensionais a partir do conjunto de dados ativo.

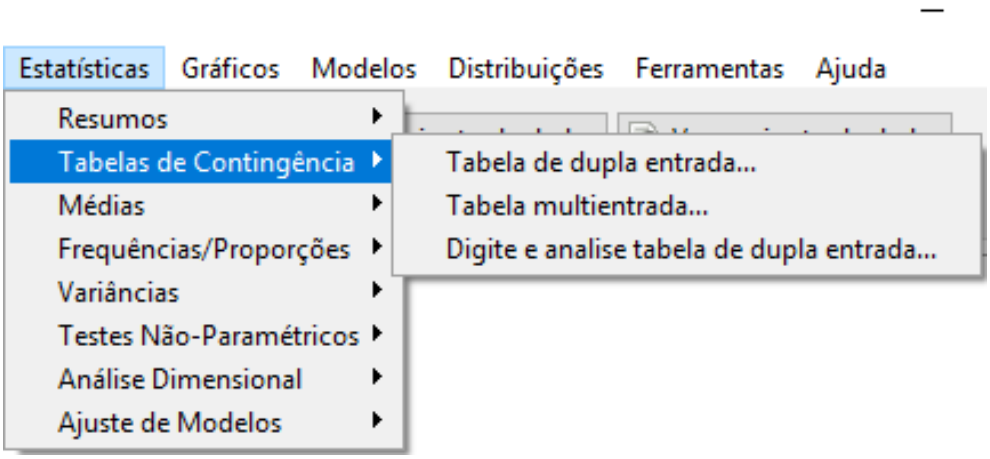


Figura 4.9: Opções de tabelas de contingência

4.2.1 Tabelas bidimensionais

Para montar uma tabela de contingência de duas variáveis do banco, iremos em Estatísticas> Tabelas de Contingência > Tabelas de dupla entrada. Após selecionar estas opções, será aberta uma janela para seleção das variáveis.

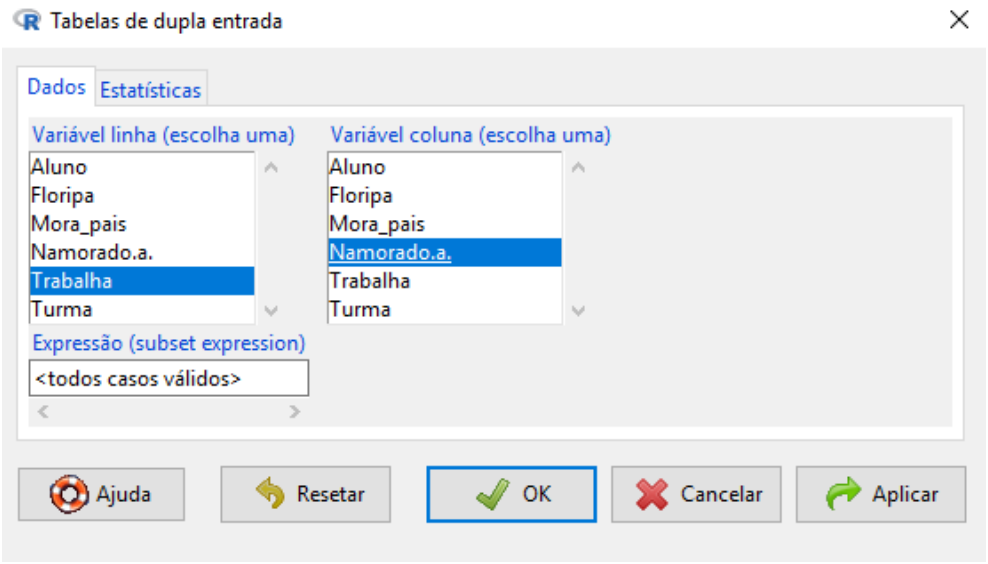


Figura 4.10: Opções de tabelas de contingência

Apenas selecionando as variáveis e clicando em ok,teremos como saída a tabela e a saída do teste Qui-Quadrado (será discutido posteriormente)

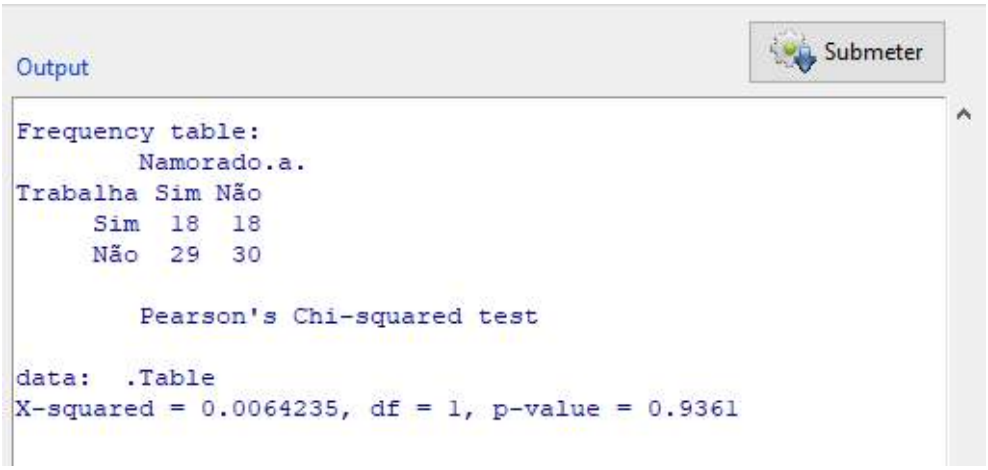


Figura 4.11: Tabela de contingência: Namora por Trabalho

No entanto, podemos também obter além da tabela com os valores, obter também os percentuais de linha, coluna ou totais. Além de fazer outros testes de hipóteses além do Qui-Quadrado. Para isso, seguiremos os mesmos passos iniciais da tabela simples: Estatísticas> Tabelas de Contingência > Tabelas de dupla entrada. Após abrir a janela de opções, selecionaremos a aba Estatísticas. E nela selecionaremos as especificações e testes que desejarmo obter na tabela.

Para exemplificar vamos gerar uma tabela que exibe os percentuais do total, sem fazer nenhum teste de hipóteses.

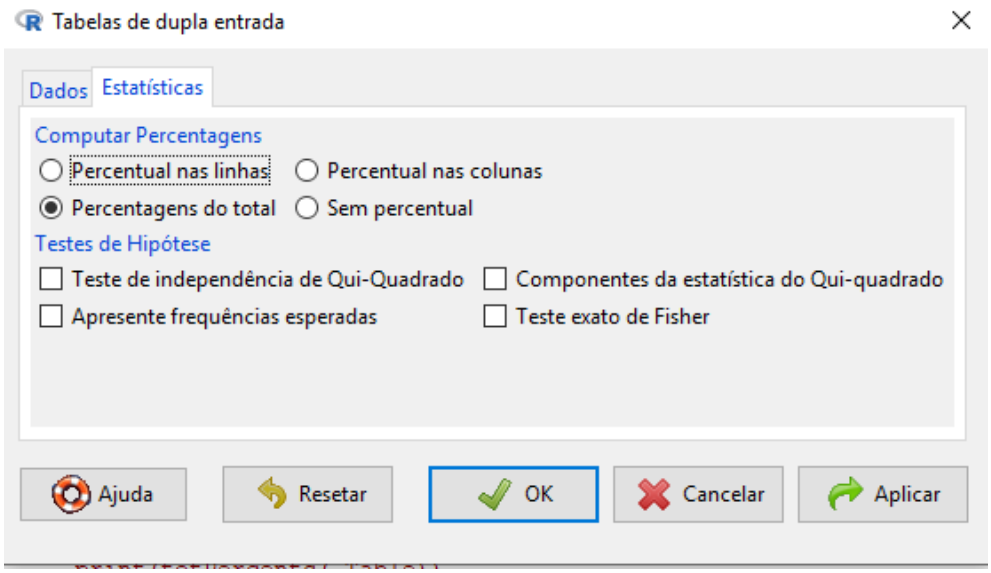


Figura 4.12: Opções a serem selecionadas

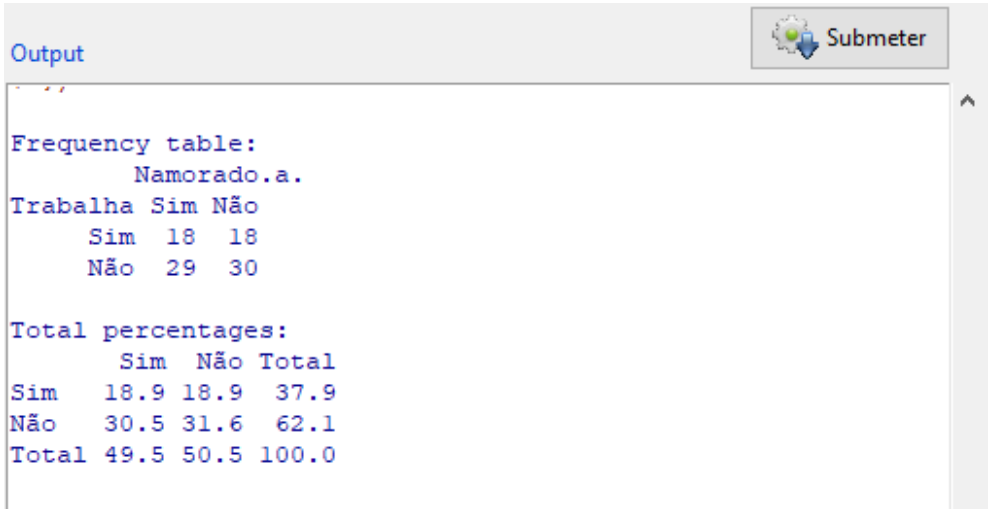


Figura 4.13: Saída da tabela com as proporções

4.2.2 Tabelas Multidimensionais

Pode ser de interesse olhar a tabela de contingência de mais de duas categorias, Neste caso teremos a variável da linha, a da coluna e as variáveis controle. Quando tiver uma variável de controle, será gerado uma tabela para cada nível da variável controle. Quando quisermos mais de uma variável de controle, será gerada uma tabela de contingência para cada combinação dos níveis das variáveis de controle.

Para obter as tabelas de contingência seguiremos iremos em: Estatísticas> Tabelas de Contingência> Tabela Multientrada. Na janela aberta, selecionaremos as opções de variável pra linha, coluna e a variável de controle:

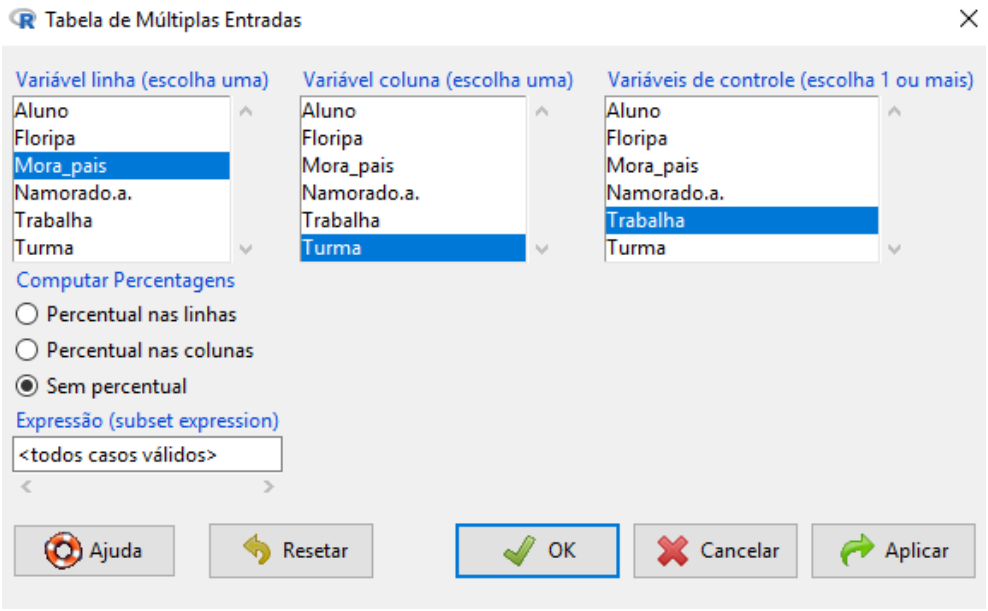


Figura 4.14: Opções para tabela multidimensional com 3 variáveis

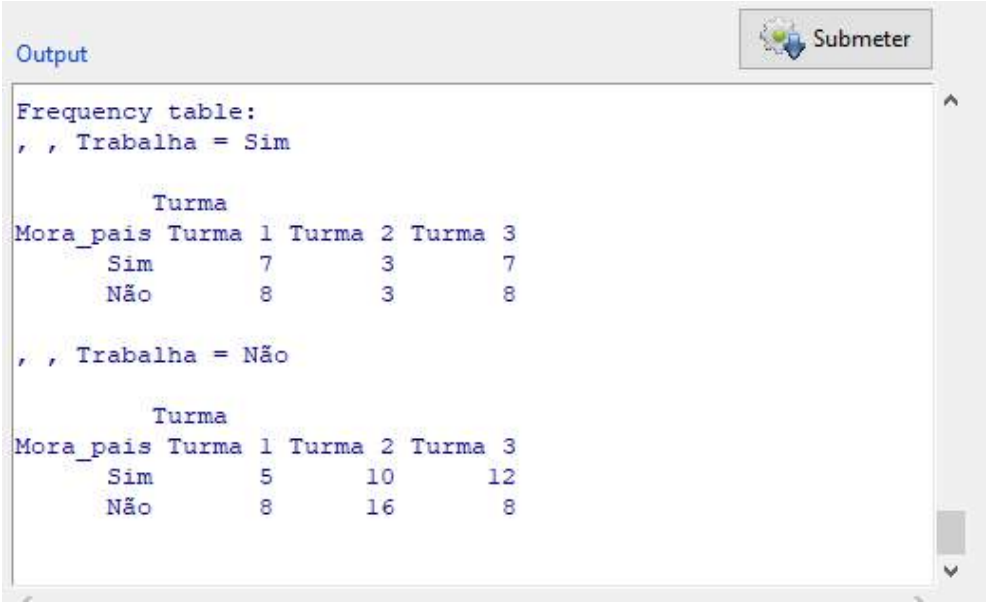


Figura 4.15: Output tabela com 3 variáveis

Para obter a tabela com mais de uma variável de controle, basta seguir os mesmos passos e selecionar mais de uma variável controle(Para selecionar mais de uma, segure a tecla Ctrl e clique nas variáveis)

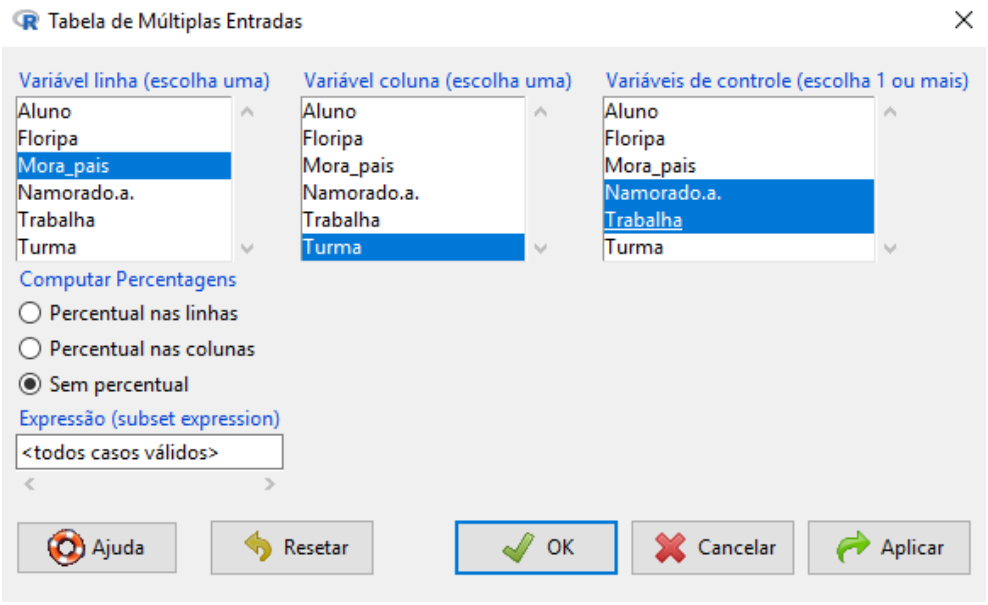


Figura 4.16: Opções para tabela multidimensional com mais de 3 variáveis

Obteremos a seguinte saída:

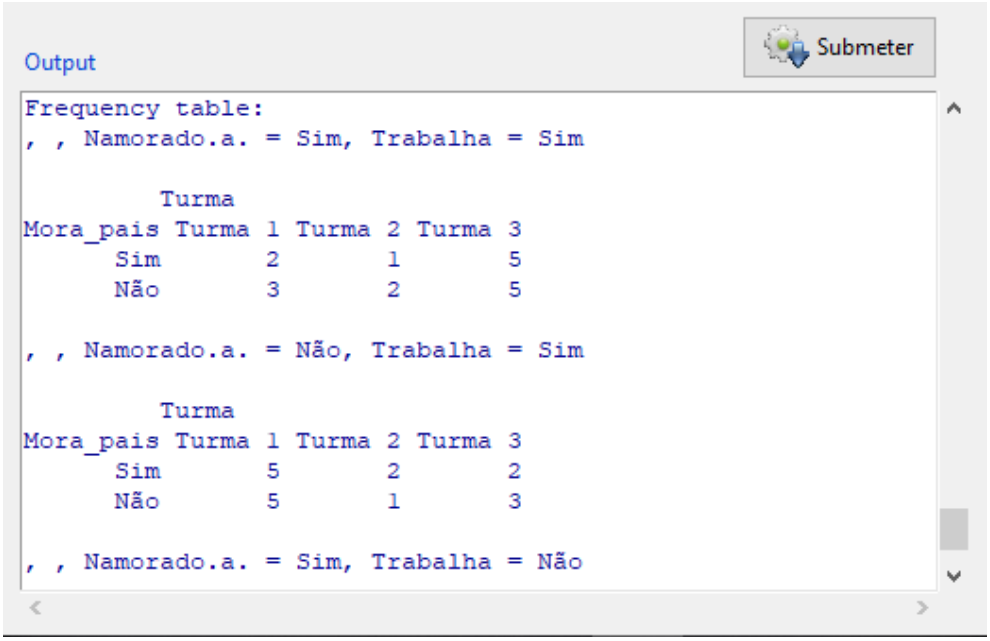


Figura 4.17: inicio do output da tabela com mais de 3 variáveis

Caso deseje-se ver os percentuais, basta selecionar a opção na mesma janela em que se seleciona as variáveis.

4.2.3 Digitando e analisando a tabela

Caso Deseje analisar uma tabela que não vem do banco de dados, podemos digita-la diretamente e analisá-la. para isso , vá em: Estatísticas > Tabelas de Contingência>Digite e analise tabela de dupla entrada.

A título de exemplo vamos criar uma tabela de turma(T1,T2,T3) do trabalho(Sim, Não)

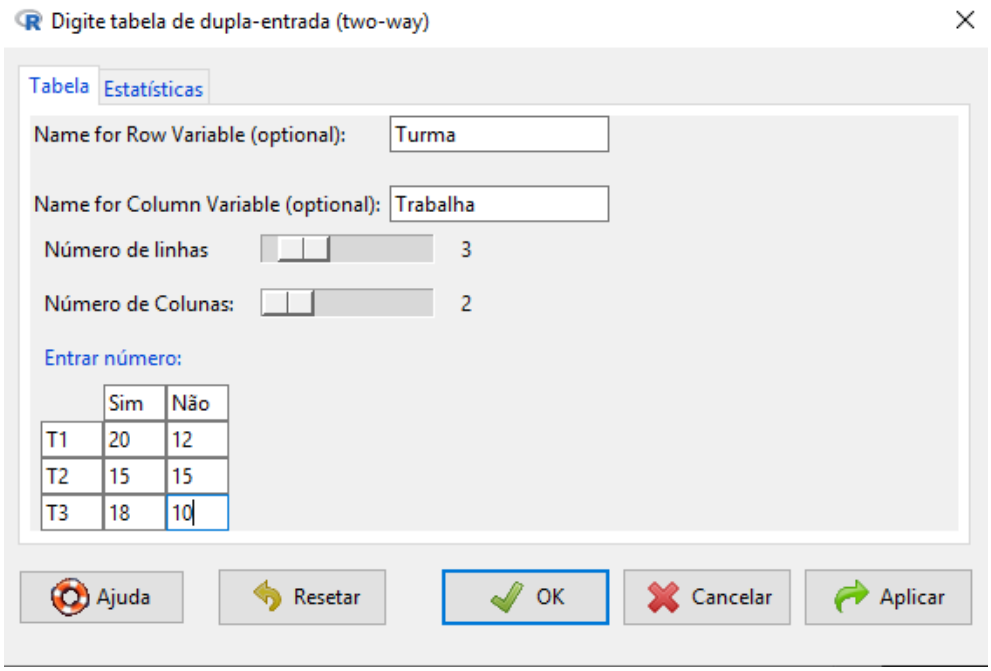


Figura 4.18: Janela para criar uma tabela de contingência

Na aba Estatísticas, podemos selecionar as especificações e testes que queremos fazer. No exemplo vamos pegar uma tabela simples, fazendo o teste Qui-quadrado

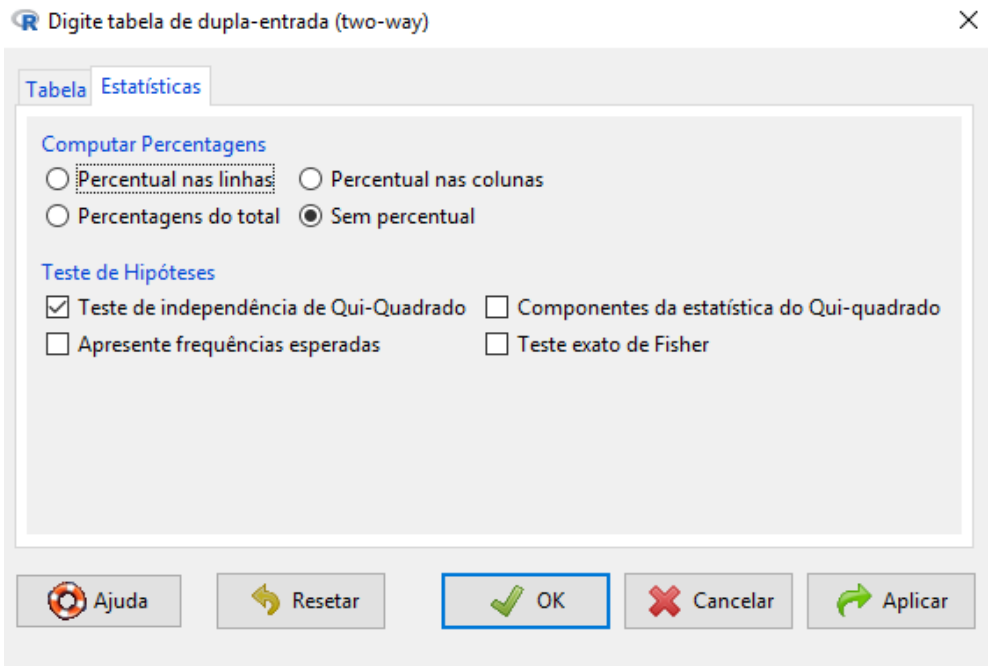


Figura 4.19: Janela de opções

Obtendo então a seguinte saída:

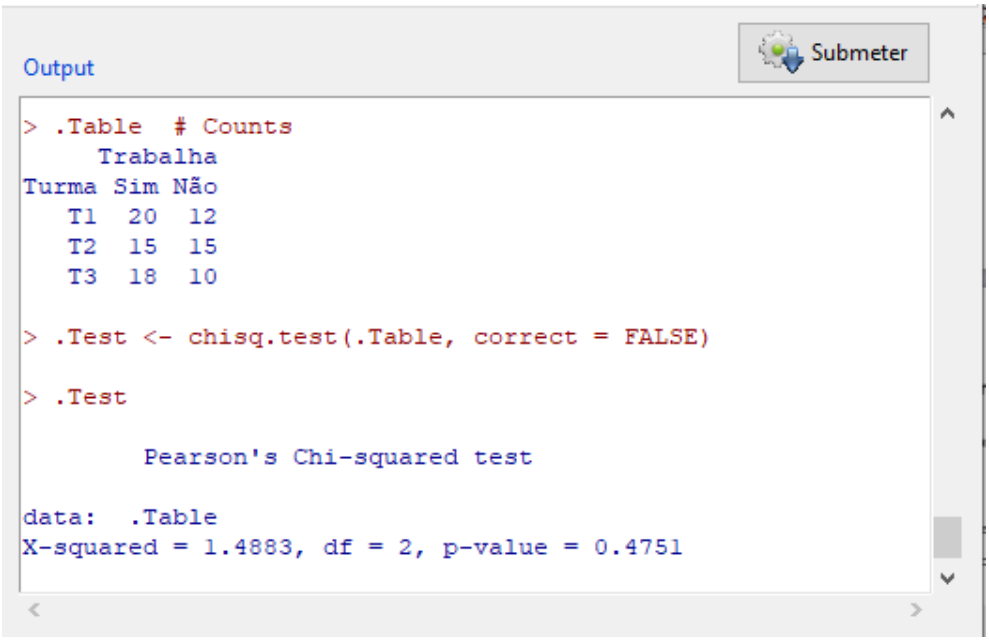


Figura 4.20: Saída da tabela criada

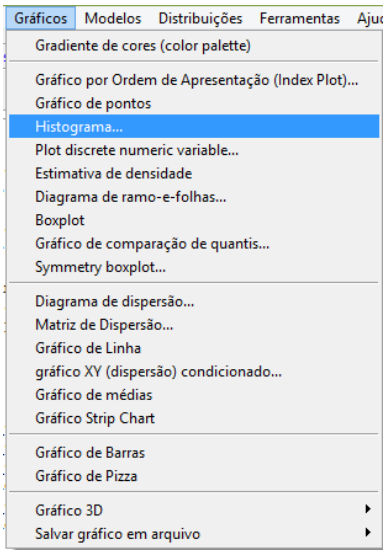
4.3 Construção de gráficos

4.3.1 Gráficos para variáveis numéricas

Histograma

A construção de histograma é adequada para variáveis numéricas. Dessa forma selecionaremos a variável "horas de estudos"do banco Estresse e construiremos um histograma para a mesma. A sequência de passos é apresentada a seguir.

Ao clicar em gráficos, selecione "Histograma".



Feito isso, selecione a variável para a qual se deseja construir o histograma.

Na janela "opções", o usuário poderá editar alguns aspectos do gráfico. Configure de acordo com a sua preferência e ao fim clique em OK.

A seguinte saída será produzida:

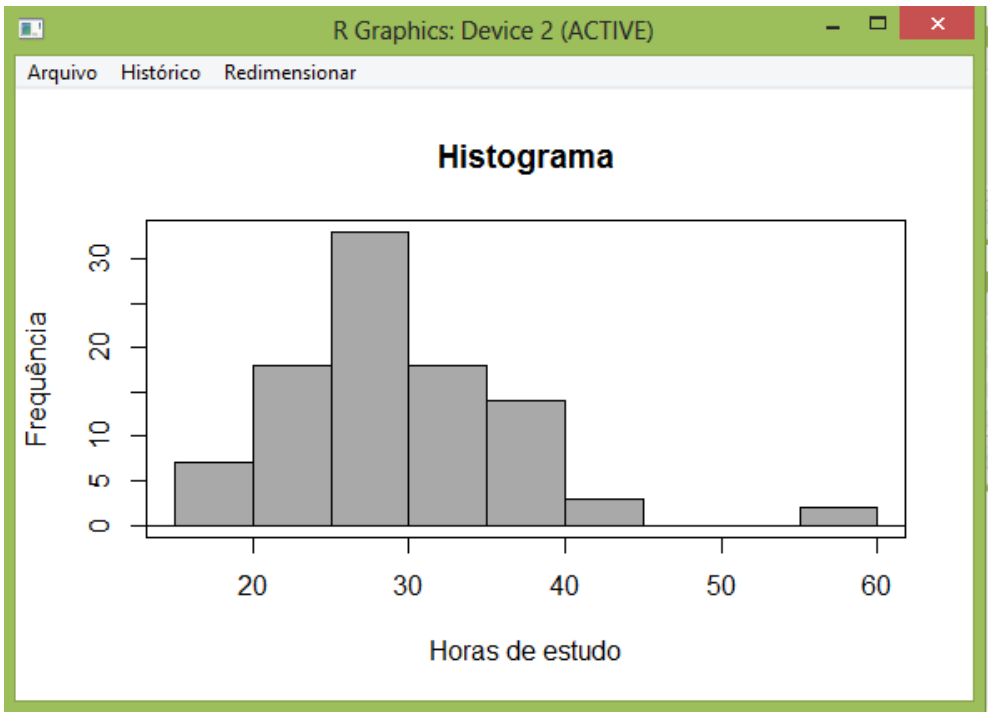


Figura 4.21: Output do gráfico histograma

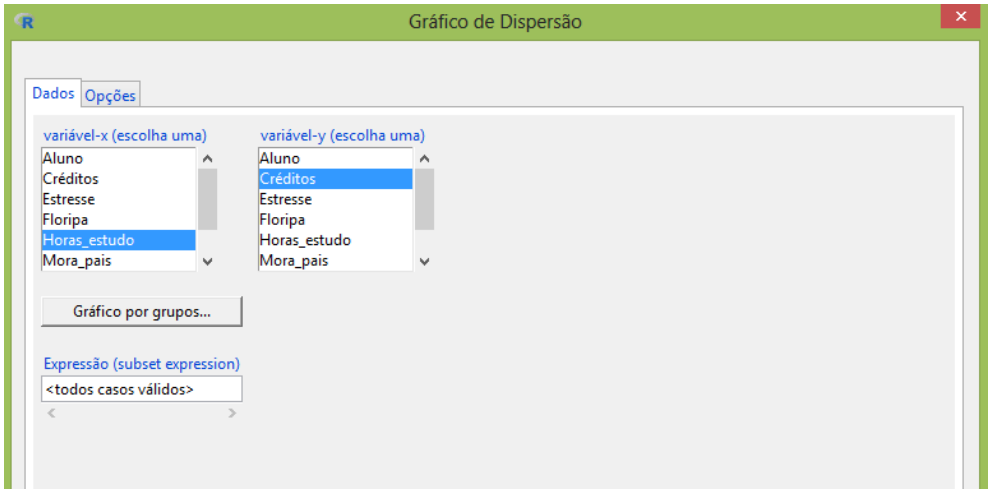
O histograma produzido nos mostra que a maioria dos alunos fazem entre 25 e 30 horas de estudo por semana.

Para salvar o gráfico produzido clique em "Arquivo", selecione o formato desejado e escolha o arquivo para o qual o arquivo com o gráfico deve ser mandado.

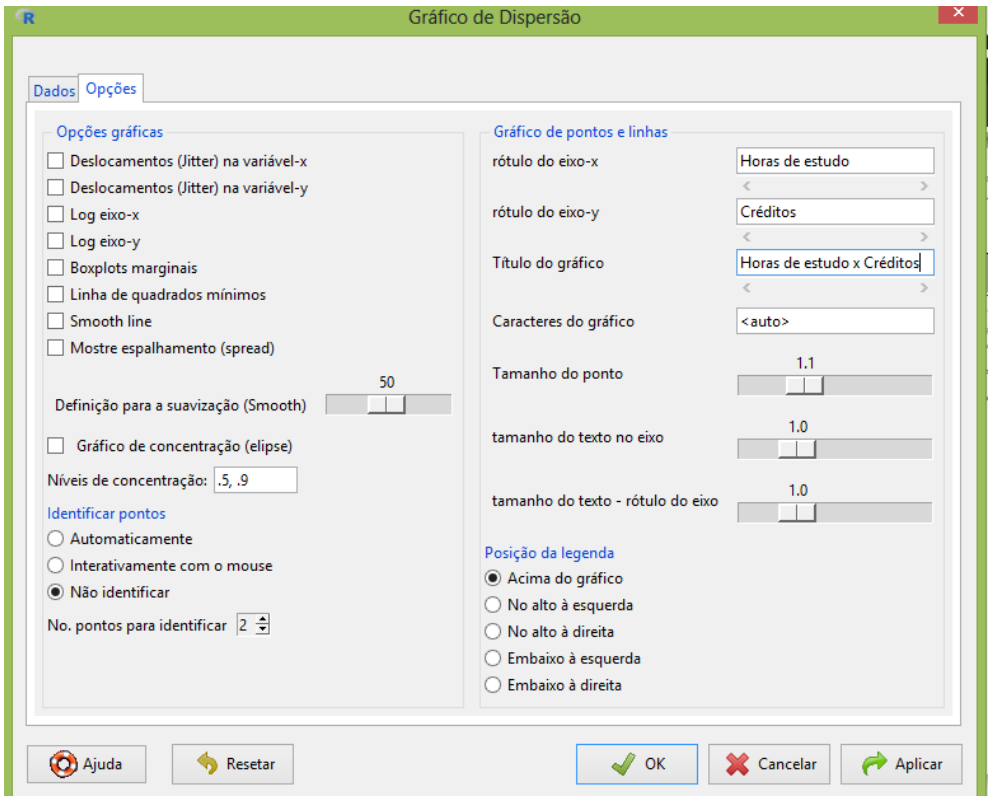
Gráfico de dispersão considerando duas variáveis

O gráfico de dispersão relacionando duas variáveis é construído de forma semelhante.

Clique em gráficos e selecione a opção gráfico de dispersão. Feito isso, uma janela será aberta para que o usuário posso selecionar as variáveis de interesse.



Clicando em "Opções"o usuário poderá editar a saída do gráfico de acordo com o seu interesse.



Uma vez realizadas todas as escolhas de edição desejadas, clique em "OK".

A saída para o nosso exemplo ficou assim:

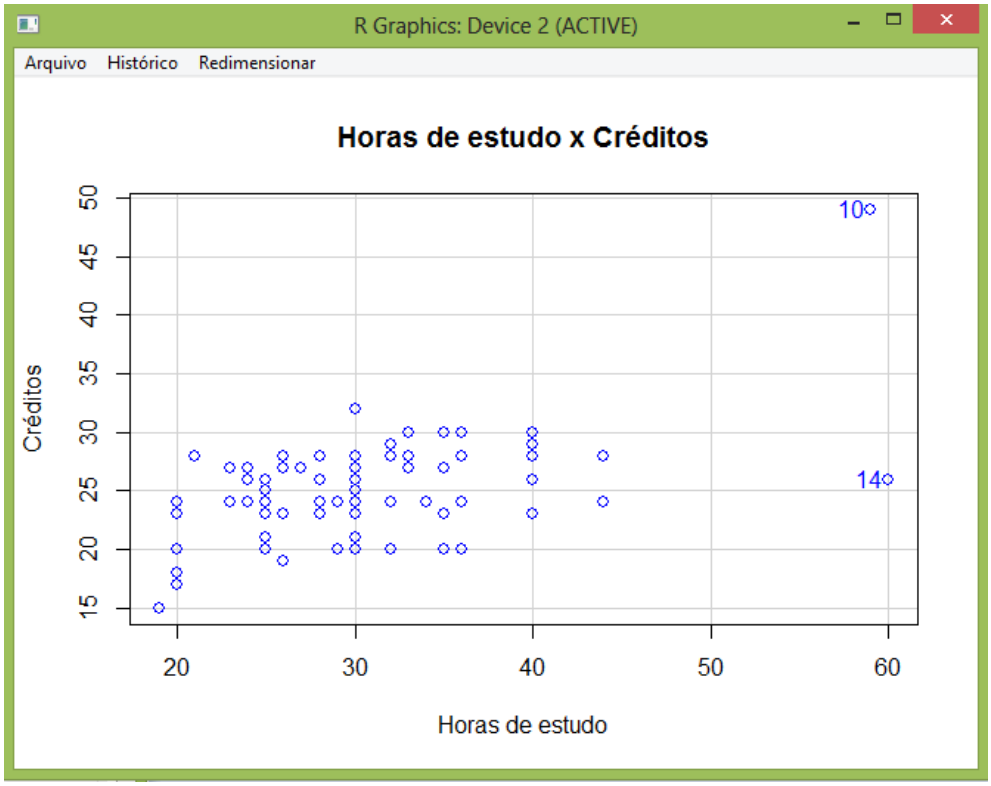


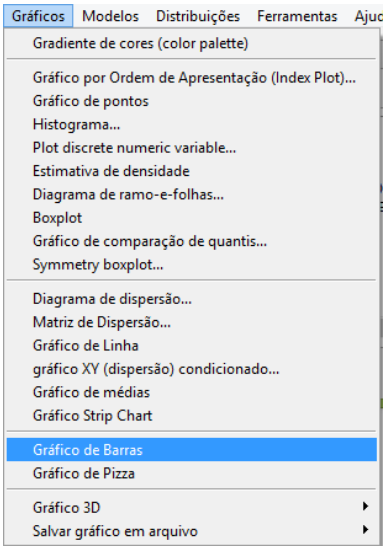
Figura 4.22: Output do gráfico de dispersão entre as variáveis "Créditos"e "Horas de Estudo".

Muitas outras opções de gráficos podem ser construídas para variáveis numéricas. Deixaremos como exemplo somente esses dois e o leitor pode criar outras baseadas no caminho indicado.

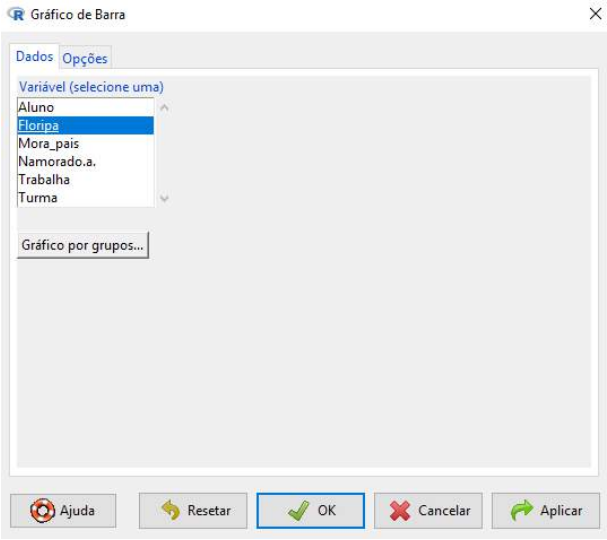
4.3.2 Gráficos para variáveis categóricas

Gráfico de Barras

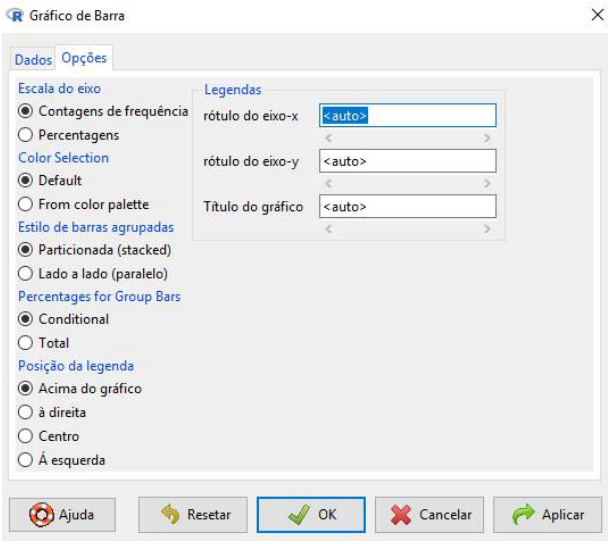
Para construir um gráfico de barras para alguma variável categórica clique em "Gráficos"e selecione a opção "Gráfico de Barras".



Feito isso, será aberta uma janela onde o usuário irá selecionar a variável de interesse.



Uma vez selecionada a variável, o usuário deve clicar em "Opções"para que possa acessar as opções de edição do gráfico, assim como feito na construção dos gráficos anteriores.



Ao final do processo de edição, clique em "OK". Uma janela com o gráfico criado será aberta. Para salvar, clique em "Arquivo", selecione o formato desejado e indique a pasta de destino do arquivo.

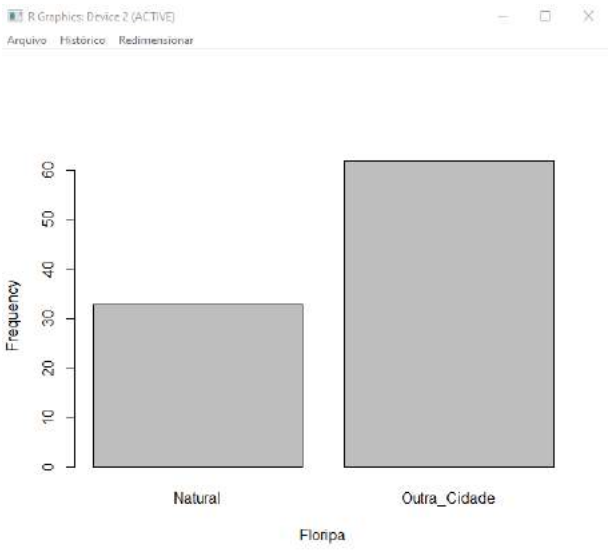
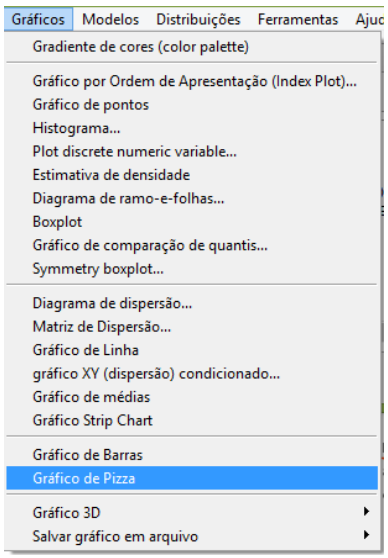
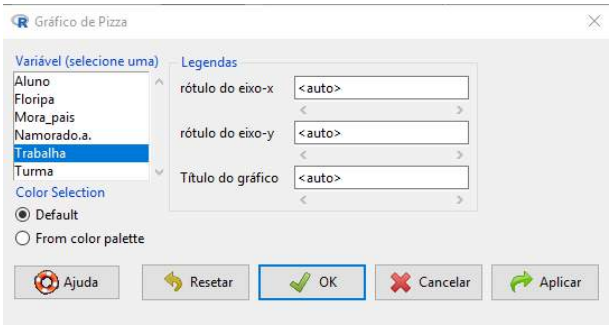


Gráfico de Pizza

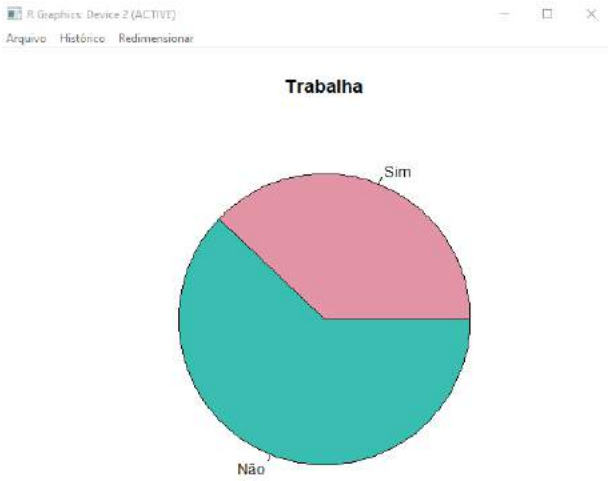
A construção do gráfico de pizza é muito semelhante à construção do gráfico de barras. Clique em "Gráficos"e selecione a opção "Gráfico de Barras".



Feito isso, será aberta uma janela onde o usuário irá selecionar a variável de interesse, e especificar as opções desejadas.



Ao final do processo de edição, clique em "OK". Uma janela com o gráfico criado será aberta. Para salvar, clique em "Arquivo", selecione o formato desejado e indique a pasta de destico do arquivo.



Capítulo 5

Inferência

Neste capítulo, apresentaremos métodos de se fazer inferência usando o rcommander, em especial, focaremos em testes de hipóteses. Um teste de hipótese nada mais e que uma regra com a qual decidimos se há evidências para dizer algo sobre um parâmetro (característica) da população ou não. Um teste é feito a partir de duas hipóteses: a nula, que usualmente afirma que o parâmetro é igual a algo e a alternativa, que afirma o contrário. Então observamos a amostra e vemos se há evidências ou não para rejeitar a hipótese nula em favor da alternativa para dado nível de confiança. Evidentemente, há dois tipos possíveis de erro a se cometer: rejeitar a hipótese nula quando ela é verdadeira (erro tipo I) e não rejeitá-la quando a mesma é falsa (erro tipo II). O nível de significância de um teste (o nível de confiança subtraído de um), é a probabilidade de erro do tipo I. Para avaliarmos se iremos rejeitar ou não a hipótese nula, utilizaremos o p-valor (que nada mais é que a probabilidade de cometer o Erro Tipo I caso rejeitemos a hipótese nula). Caso ele seja menor que o nível de significância escolhido, rejeitamos a hipótese nula em favor da alternativa. Um teste de hipótese pode ser bilateral(hipótese nula diz que a média é igual a algum valor e a alternativa que é diferente do mesmo) ou unilateral (hipótese nula diz que a média é menor ou igual/maior ou igual a algum valor e a alternativa diz que é maior/menor que o mesmo). Um exemplo possível de teste se trata de pegar dados sobre a altura de 30 alunos de uma turma e testar se há evidências para dizer que a altura média da turma completa é menor ou igual a 170cm (hipótese nula), ou se ela é maior que 170cm (hipótese alternativa). Então pegaremos a média amostral da altura dos alunos, e a partir de uma regra, veremos se ela é indicativa de que a altura média da turma como um todo é maior que 170cm. Nas seções a seguir veremos diversas regras para realizar testes como esse e muitos outros.

5.1 Testes para média

5.1.1 Teste t para média de uma população

O teste t para média de uma população deve ser usado quando desejamos testar se a média é igual/menor ou igual/maior ou igual a um valor especificado. Para usá-lo é necessário que a suposição a seguir seja atendida:

(i) A distribuição de probabilidade dos dados é normal.

A exemplificação será feita utilizando o banco de dados de dados Questionário_Estresse. Para executar o teste no rcommander, deve-se ter o conjunto de dados carregado, ir no menu superior Estatísticas>Médias>Teste t para uma amostra. A janela a seguir será aberta:

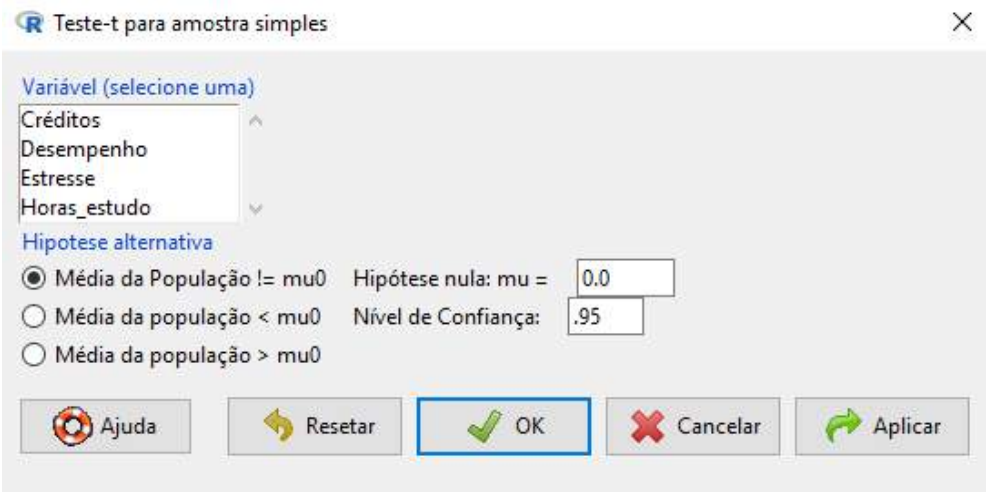


Figura 5.1: Janela do teste t

Suponha que desejamos testar com 0.95 de confiança se a média do indicador de estresse é igual a 30 (hipótese nula) ou se é diferente (hipótese alternativa) e que ele atenda aos pressuposto de normalidade. Portanto, selecionamos Estresse como variável, mu0=30, a hipótese alternativa será do tipo "média da população!=mu0"(pois se trata de um teste bilateral) e o nível de confiança será de 0.95.

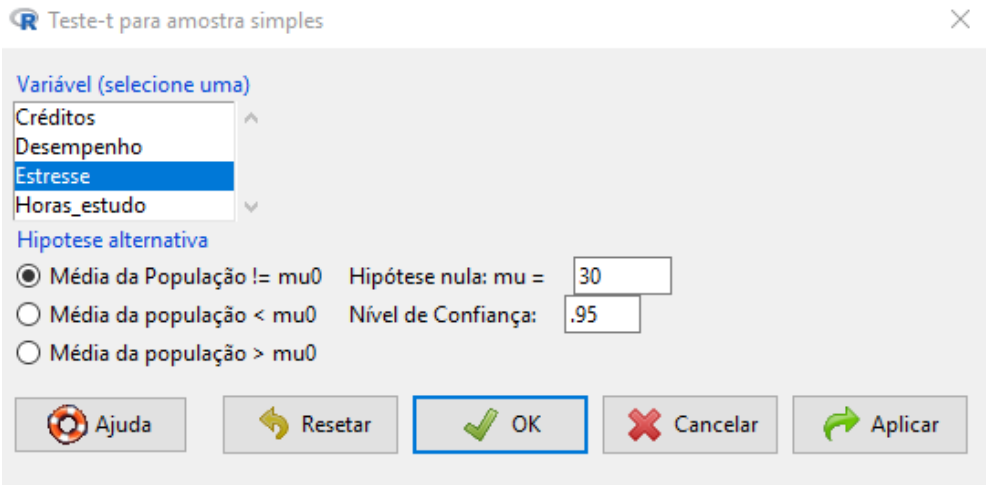


Figura 5.2: Janela do teste t com as opções

Após colocarmos essas informações e clicarmos em "ok", aparecerá o texto a seguir na janela de output:

```
        One Sample t-test

data:  Estresse
t = -2.8167, df = 94, p-value = 0.005913
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 26.28509 29.35702
sample estimates:
mean of x
 27.82105
```

Figura 5.3: Resultados do teste

Se considerarmos um nível de significância de 5% ($\alpha = 0.05$) ao olhar para o p-valor (destacado em amarelo) observamos que ele é menos menor que o α, portanto nós rejeitamos a hipótese nula, isto é, dizemos que não há evidência amostral para afirmarmos que a média populacional do indicador de estresse é igual a 30.

Para fazer um teste unilateral, devemos alterar a marcação no checkbox da hipótese alternativa. Por exemplo, caso desejemos testar se a média do indicador de estresse é maior ou igual a 30 (hipótese nula) ou não (hipótese alternativa), devemos selecionar a hipótese alternativa como "Média da população <mu0".

5.1.2 Teste t para média de amostras independentes

O teste t para a média de amostras independentes, é usado quando desejamos comparar a médias entre dois grupos. O uso de tal teste é condicionado á suposição de normalidade, ou seja, devemos usá-lo se houver evidências para dizer que os dois grupos são provenientes de uma distribuição normal. Para a realização do teste é necessário verificar se a variância das duas populações é igual ou diferente, o teste para tal verificação será apresentado adiante neste manual.

Suponhamos que desejamos testar se o desempenho dos alunos que não trabalham é superior ao desempenho dos alunos que trabalham. Neste caso nossa hipótese nula é:

$$H_0 : \mu_{trabalha} \leq \mu_{Naotrabalha}$$

No R commander, com o banco de dados carregado, basta ir no menu superior:

Estatísticas > Médias > Teste t para amostras independentes. Na janela aberta há duas abas. na aba dados devemos selecionar a variável que especifica os Grupos e a Variável Resposta, que é a variável com os valores a serem testados.

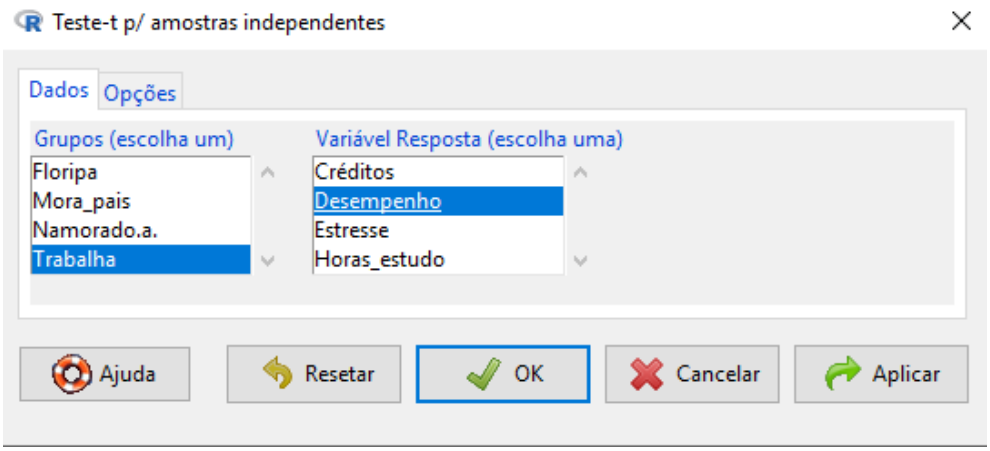


Figura 5.4: Aba de seleção dos dados do teste

Na aba opções é possível selecionar o teste a ser feito, o nível de confiança e especificar se as variâncias são iguais ou não. Por se tratar da comparação entre duas médias as nossas hipóteses podem ser definidas com base na diferença entre as médias. neste caso testaríamos se a diferença é maior/igual, menor/igual ou diferente de 0. No topo da aba é especificada a diferença usada, e com base nessa diferença, especificamos nossa hipótese alternativa.

No nosso exemplo a hipótese alternativa é de que a média de quem trabalha é maior do que a de quem não trabalha, portanto a diferença entre a média de quem trabalha e de quem não trabalha seria maior que 0. $H_1 : Diferença > 0$, utilizaremos um nível de significância de 0.05, portanto o nível de confiança é de 0.95. Assumiremos também que as variâncias são iguais:

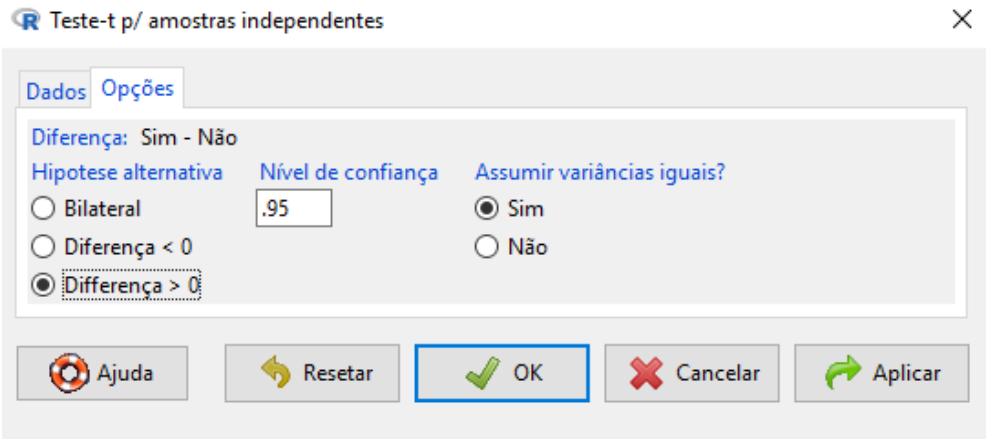


Figura 5.5: Aba de opções do teste

Após selecionarmos as opções e clicar em "ok", será exibido o resultado na janela de output:

```
        Two Sample t-test

data:  Desempenho by Trabalha
t = -0.52345, df = 93, p-value = 0.699
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.3596418        Inf
sample estimates:
mean in group Sim mean in group Não
         8.540278          8.626441
```

Figura 5.6: Resultado do teste

O resultado é obtido de forma bem detalhada, ele descreve vários aspectos do teste, inclusive o intervalo de confiança. A conclusão do teste pode ser feita tanto pelo p-valor quanto pelo intervalo de confiança. A conclusão pelo p-valor é dada pela comparação do p-valor com o nível de significância. A conclusão a partir co intervalo de confiança é feita observando se o valor 0 está contido no intervalo. O p-valor é igual a 0.699(maior que 0.05) e o intervalo de confiança vai de -0.3596 até ∞ (Contém o valor 0), portanto não há evidências para se afirmar que o desempenho dos alunos que trabalham é inferior ao desempenho dos alunos que não trabalham.

5.1.3 Teste t para a média de amostras pareadas

Uma amostra Pareada é a amostra onde temos pares de observações. Por exemplo, o peso de uma pessoa no inicio e no final da dieta, a renda dos dois membros de um casal e a nota da primeira e da ultima prova de um aluno. Quando testamos a médias de amostras independentes, estamos comparando a média geral de um grupo, com a média geral de outro. Mas quando tamos tratando de uma amostra pareada, estamos testando se as observações tem diferença em cada indivíduo. Para isso, subtraímos a primeira observação da segunda observação de cada indivíduo, se não houver diferença entre as duas observações de um mesmo indivíduo, é esperado que a diferença seja igual a 0. Então, tendo a diferença de todos os indivíduos é realizado o teste para verificar a média das diferenças em relação ao valor 0 (pode ser bilateral ou unilateral). Para realização deste teste, novamente temos a suposição de normalidade da variável de interesse.

Suponhamos que desejamos testar se um treinamento feito em uma empresa foi eficaz. Antes do treinamento, foi avaliada a produtividade de cada um dos funcionários da empresa, e depois do treinamento, a produtividade foi avaliada novamente.

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

Figura 5.7: Produtividade dos funcionários de uma empresa

Suponha que saibamos que a produtividade dos funcionários é proveniente de uma distribuição normal. Como desejamos testar se o tratamento foi efetivo, estamos interessados em testar se a produtividade aumentou após o tratamento, por consequência, se o valor da diferença entre a produtividade antes e depois do treinamento é menor que 0.

Nossas hipóteses serão da seguinte forma

$$H_0 : Produtividade_{Depois} = ProdutividadeAntes$$
$$H_1 : produtividade_{Depois} > ProdutividadeAntes$$

Que é equivalente a seguinte hipótese:

$$H_0 : \mu_{diferencas} = 0$$
$$H_1 : \mu_{diferencas} < 0$$

Para este exemplo carregamos o banco de dados "Produtividade".

NO R commander, com o banco de dados carregados, basta ir no menu superior:

Estatísticas > Médias > Teste t (dados Pareados). Na janela aberta, há duas abas, Na aba dados, vamos especificar a primeira e a segunda variável a serem consideradas pelo teste, lembrando que o teste fara a diferença entre a primeira e a segunda:

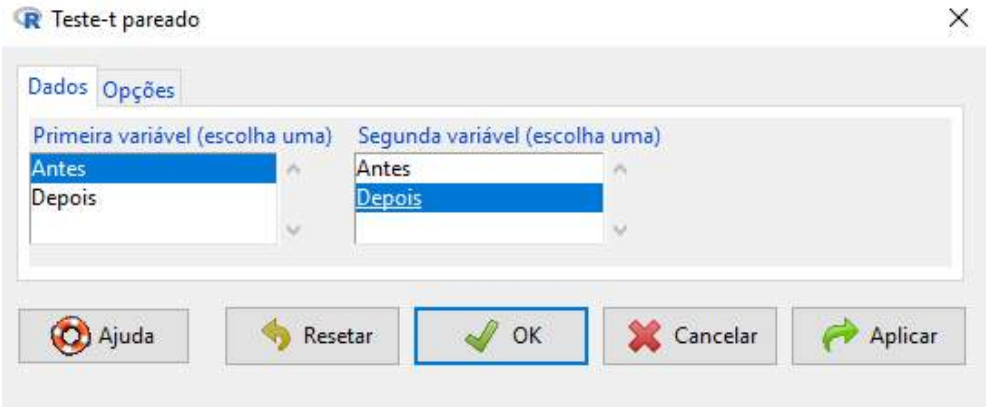


Figura 5.8: Aba "dados"do teste pareado

Na aba opções, selecionamos qual a hipótese alternativa do nosso teste, e o nível de confiança se ser utilizado. Ne nosso exemplo, estamos interessados em testar de a Diferença é menor que 0, portanto, selecionaremos tal opção.

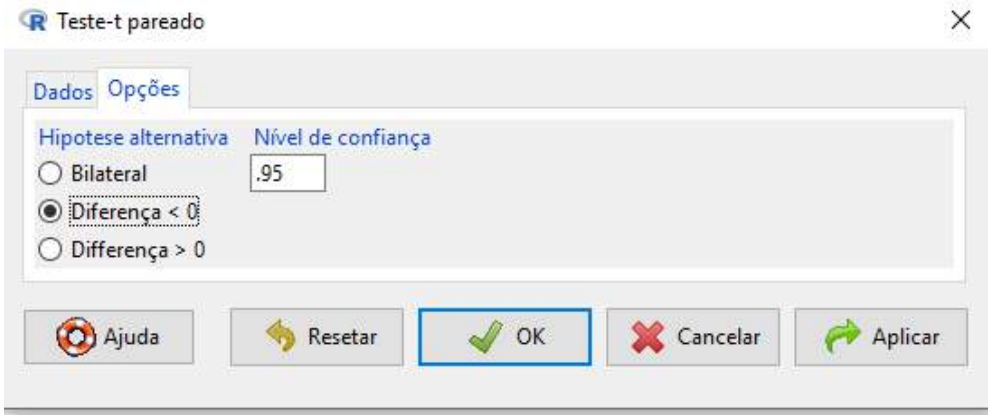


Figura 5.9: Aba "opções"do teste pareado

Após selecionarmos as opções e clicarmos em "ok", o resultado será exibido na janela de output:

```
        Paired t-test

data:  Antes and Depois
t = -2.8246, df = 9, p-value = 0.009948
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.193486
sample estimates:
mean of the differences
                  -3.4
```

Figura 5.10: Resultado do Teste

O resultado é exibido de forma bem detalhada, informado, o valor t, os graus de liberdade, p-valor, intervalo de confiança e valor estimado.A conclusão pelo p-valor é de que rejeitamos a hipótese nula, pois 0.0099 é menor que 0.05(1 - 0.95(nível de confiança)). Portanto, com 5% de significância, concluímos que houve um aumentos na produtividade após o treinamento.

5.2 Teste para proporção

5.2.1 Pela Distribuição Normal

Para que o entendimento do que será explicado aqui fique mais claro recomendamos fortemente que leia e tente compreender o TCL - Teorema Central do Limite, o qual explicamos na parte de simulações.

O teste da proporção é um teste de hipóteses cujo objetivo é decidir probabilisticamente se uma proporção populacional é igual ou não a um número especificado pelo pesquisador (lembre-se que proporção é uma medida representada por números entre 0 e 1). Por exemplo, imagine que você está conduzindo uma pesquisa e deseja verificar se a metade da população é contra as decisões do atual governo. Nesse sentido, você está estabelecendo duas hipóteses:

H0: $p = 0.5$

H1: $p \neq 0.5$

Sendo que p é a proporção populacional de opositores (ou de favoráveis, neste caso tanto faz). Lembre-se também que $0.5 = 50\%$

É importante ressaltar que nesse caso você não tem acesso à toda população e, por isso, a decisão de rejeitar ou não uma hipótese depende do estudo de uma amostra, ou seja, você vai tomar a decisão com base na proporção de opositores encontrada na amostra coletada.

Sendo assim, você deve estar pensando: "Ok, mas se eu encontrar uma amostra cuja proporção de opositores é igual a 0,55 (55%), não necessariamente a proporção populacional de opositores (p) será 0,55 ou até mesmo igual a 0,5 (50%)."Este questionamento está correto e é plausível que se pense nessa questão. É por isso que usamos uma distribuição de probabilidade para determinar o que vamos considerar como provavelmente correto: H0 ou H1.

Para realizarmos um teste de hipóteses devemos primeiramente supor que H0 seja verdade. Fazemos isso porque H0 contém a hipótese mais restrita: a igualdade. Neste exemplo, quando supomos que a proporção populacional de opositores é igual a 0,5 (p = 0.5) conseguimos construir a distribuição de probabilidade da proporção amostral cuja média é igual a 0.5, uma vez que estamos supondo que H0 seja verdadeira (se não conseguiu entender por que conseguimos construir a distribuição de probabilidade da proporção amostral, leia atentamente à explicação do Teorema Central do Limite - TCL). Nesse sentido, a distribuição de probabilidade da proporção amostral supondo que H0 seja verdadeira é uma distribuição Normal com média igual a 0.5 e variância igual a $\frac{p(1-p)}{n} = \frac{0.5(1-0.5)}{n}$:

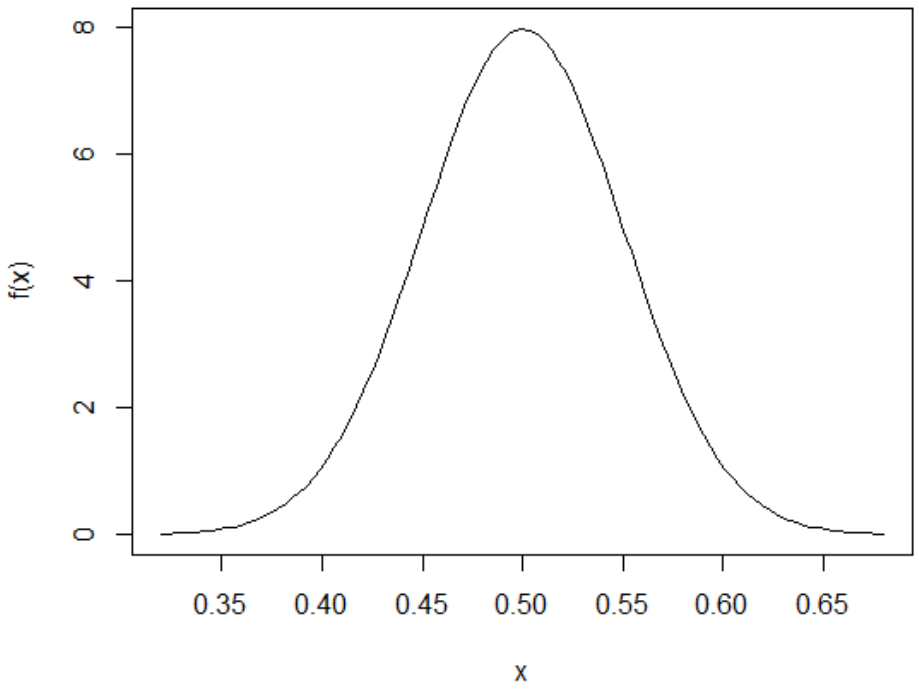


Figura 5.11: Distribuição normal com média 0.5

Se você leu o Teorema Central do Limite (TCL) deve ter percebido que o exemplo dado servia para a média. Neste exemplo, estamos tratando da proporção (daí o nome "Teste da Proporção"). Entretanto, é importante você perceber que a proporção é uma média. Por exemplo, se considerarmos que opositores sejam representados pelo número 1 e os favoráveis sejam representados pelo número 0, a proporção de opositores em um grupo de n pessoas é a soma de todos os zeros e uns dividido pelo total de integrantes (n). Exemplo: seja um grupo de 10 pessoas: 0,1,1,1,0,0,0,1,0,1. Proporção de opositores: $\frac{0+1+1+1+0+0+0+1+0+1}{10} = \frac{5}{10} = 0,5$ (50%). Como a proporção é uma média, o TCL nos informa que a distribuição de probabilidade da proporção amostral ($\hat{p}$) é uma normal com média igual a p (proporção populacional) e variância igual a $\frac{p(1-p)}{n}$. Como estamos supondo que H0 seja verdadeira, substituímos p por 0.5.

Já que sabemos qual é a distribuição de probabilidade da proporção amostral supondo que H0 seja verdadeira, conseguimos determinar a probabilidade de observarmos a proporção que encontramos em nossa amostra se H0 for verdadeira. Essa probabilidade é calculada determinando onde a nossa observação amostral (proporção amostral) se encontra nessa distribuição. Por exemplo, se descobrirmos que a probabilidade de encontrarmos uma proporção mais atípica do que encontramos, dado que H0 é verdadeira, é igual a 0.03 (3%), significa que é muito improvável encontrar uma amostra que apresenta essa proporção e, por isso, é uma decisão inteligente admitir que H0 não é verdadeira ao invés de admitir que nossa amostra realmente é atípica, pois admitindo nesse contexto que a nossa amostra é atípica e que H0 é verdadeira temos apenas 0,03 (3%) de chance de acertarmos nessa decisão. Vale lembrar que a probabilidade de encontrarmos uma proporção mais atípica do que encontramos, dado que H0 é verdadeira, é chamada de p-valor e, neste exemplo, ela é igual a 0.03 (3%). Para que possamos ter um critério para decidir se rejeitamos ou não H0, estabelecemos um valor mínimo aceitável para o p-valor. Esse valor mínimo aceitável é arbitrário (o pesquisador é quem escolhe) e ele é chamado de significância. Então, no exemplo que acabamos de dar, rejeitaríamos H0 se o p-valor fosse menor que a significância (significância = 0.05, por exemplo) e não rejeitaríamos se o p-valor fosse maior ou igual a significância (significância $= 0.02$, por exemplo). Dessa forma, conseguimos avaliar, com critérios probabilísticos, a veracidade de H0.

Se encontrarmos, por exemplo, 42 opositores em uma amostra de 70 pessoas, isto é, uma proporção de opositores igual a 0.6 (60%), devemos primeiramente calcular qual é o p-valor. Para tal, encontramos primeiro a estatística de teste z, ou seja, encontramos qual é a distância entre nossa observação (0.6) e a média da distribuição a qual ela pertence. Neste caso, a média da distribuição a qual a nossa observação ($\hat{p} = 0.6$) pertence é igual a 0.5, pois estamos supondo que H0 seja verdadeira. Porém, essa distância deve ser determinada em número de erros padrão (erro padrão é desvio padrão da distribuição - ou a raiz da variância da distribuição, ou seja, o desvio padrão das possíveis proporções de amostras do mesmo tamanho n (n = 70, neste caso)), pois os valores de z que se encontram na tabela da distribuição normal padrão estão todos na mesma unidade de medida: o erro padrão. Assim,

$$z = \frac{0.6-0.5}{\sqrt{\frac{0.5(1-0.5)}{70}}}$$

Vamos considerar que a significância seja igual 0.05 (5%). Então a área rachurada em vermelho no gráfico abaixo é chamada de região crítica e ela representa o valor da significância, sendo que ela está dividida em duas partes: a área rachurada abaixo da média (média $= 0.5$) e a área rachurada acima da média (média $= 0.5$), ambas representando

0.025 (2.5%) da área total da distribuição, ou seja, a probabilidade de encontrarmos uma proporção amostral dentro dessa área vermelha é igual a 0.05 (5%) e, caso isto aconteça, como já explicado, rejeitamos H0. Podemos também, como dito, comparar o p-valor com essa área. Se o p-valor for menor do que essa área, podemos notar que nossa observação (proporção amostral) está localizada dentro da região crítica e, assim, rejeitamos H0 caso isso aconteça.

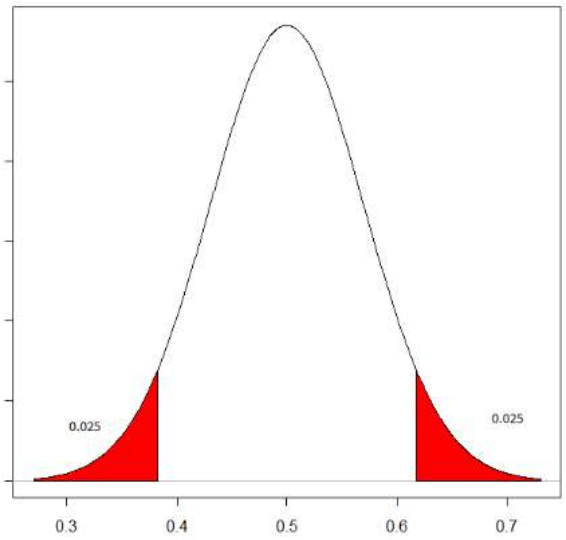


Figura 5.12: Região crítica: representação gráfica da significância

Ora, como encontramos $z = 1.67$, então basta olharmos na tabela Z (Normal padrão) a probabilidade de uma observação ser maior que 1.67. O p-valor será duas vezes essa probabilidade, uma vez que estamos comparando o p-valor com a significância e, esta, está construída dos dois lados da distribuição (veja a figura acima), já que é um teste bilateral (veja a hipótese alternativa H1, ser diferente pode ser tanto menor quanto maior). Então, $p - valor = 2.P(Z > 1.67) = 0.09492$. Veja o p-valor desenhado abaixo:

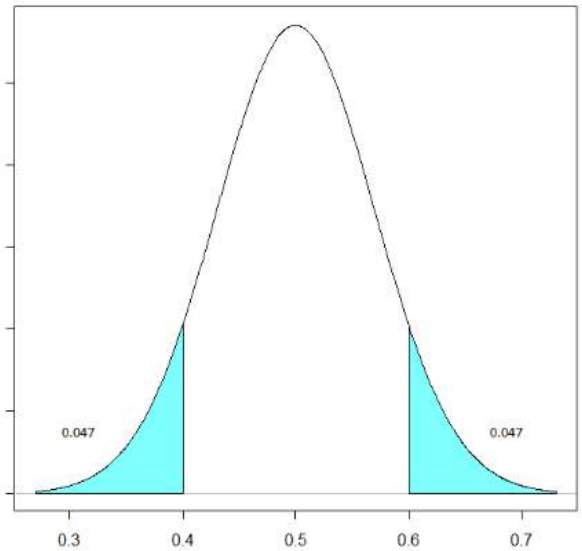


Figura 5.13: Representação gráfica do p-valor

Obs: encontramos um valor $z = 1.67$, mas se nossa observação estivesse abaixo da média, ou seja, se a proporção amostral encontrada estivesse abaixo de 0.5, esse valor z seria negativo. Assim, o p-valor seria calculado da seguinte

forma: $p - valor = 2.P(Z < z)$.

Como a área azul (p-valor) é maior que a área vermelha (significância), não rejeitamos H0.

Além disso, podemos notar que nossa observação não está dentro da região crítica. Para encontrar os valores que delimitam essa região, basta calcular quais são as proporções, abaixo e acima da média, que deixam uma área de 0.025 (2.5%) abaixo e acima delas, respectivamente. 0.025 (2.5%) porque $0.025 + 0.025 = 0.05$, já que adotamos uma significância igual a 0.05 (5%). Sabemos, ao olhar na tabela Z, que 1.96 é o valor que deixa uma área de 0.025 (2.5%) acima dele. Como a distribuição é simétrica, -1.96 é o valor que deixa uma área de 0.025 (2.5%) acima dele. Mas quais são as proporções, acima e abaixo da média (0.5), que estão a 1.96 erros padrão distantes da média (0.5) ? Basta resolver as equações: $1.96 = \frac{x1-0.5}{\sqrt{\frac{0.5(1-0.5)}{70}}}$ e $-1.96 = \frac{x2-0.5}{\sqrt{\frac{0.5(1-0.5)}{70}}}$;
Para a primeira: $x1 = 1.96.\sqrt{\frac{0.5(1-0.5)}{70}} + 0.5$, assim, $x1 = 0.6171$; para a segunda: $x2 = -1.96.\sqrt{\frac{0.5(1-0.5)}{70}} + 0.5$, assim, $x2 = 0.3828$. Veja a figura abaixo:

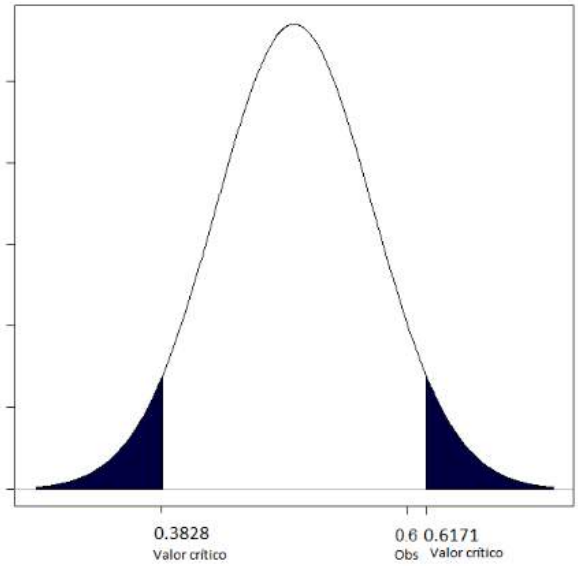


Figura 5.14: Representação da região crítica delimitada por valores não padronizados

Dessa forma, se nossa observação (proporção amostral) não estiver abaixo de 0.3828 ou acima de 0.6171 ela não pertence à região crítica e, assim, não rejeitamos H0 caso isso aconteça. Em nosso exemplo isso de fato não aconteceu, já que 0.6 não está abaixo de 0.3828 ou acima de 0.6171. Confirmamos este fato também pelo p-valor, pois este se apresentou maior do que a significância neste nosso exemplo ($0.09492 > 0.05$).

É importante ressaltar que toda vez que o p-valor se apresenta maior do que a significância nossa observação nunca estará na região crítica. Além disso, não faz-se necessário verificar tanto a presença de nossa observação na região crítica quanto o p-valor, pois o p-valor e a região crítica são métodos distintos para rejeitar ou não H0, eles nunca vão se contradizer e eles foram apresentados em conjunto aqui nesta explicação apenas para efeito didático. Em um exercício, pode-se resolver tanto por um método quanto do outro, ambos chegarão à mesma conclusão. Por fim, é importante notar que cada uma tem suas vantagens: o p-valor te informa qual a significância máxima para não rejeitar H0 (que é o próprio p-valor, quando p-valor = significância). Já região crítica te informa os valores limites para que H0 não seja rejeitado, fato que permite o pesquisador testar hipóteses de várias amostras do mesmo

tamanho, caso ele tenha que fazer isso, sem ter que calcular o p-valor de cada uma dessas amostras.

Vale ressaltar que existem também os testes unilaterais. Por exemplo, você poderia testar a hipótese de que opositores são a maioria (p = proporção de opositores):

H0: $p = 0.5$

H1: $p > 0.5$

Significância = 0.05

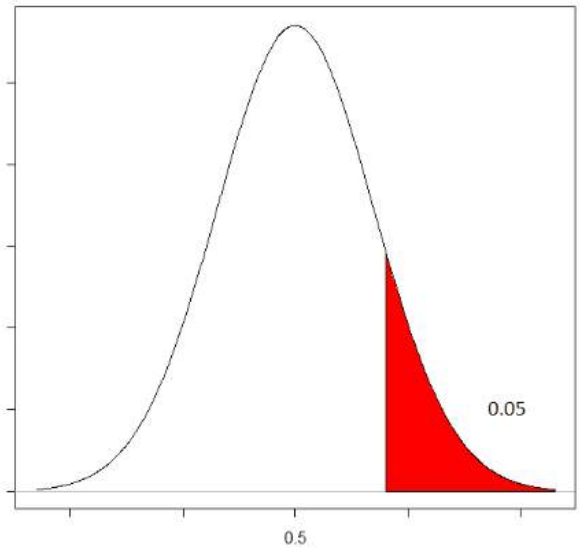


Figura 5.15: Região crítica à direita

Ou você poderia testar a hipótese de que opositores são a minoria (p = proporção de opositores):

H0: $p = 0.5$

H1: $p < 0.5$

Significância = 0.05

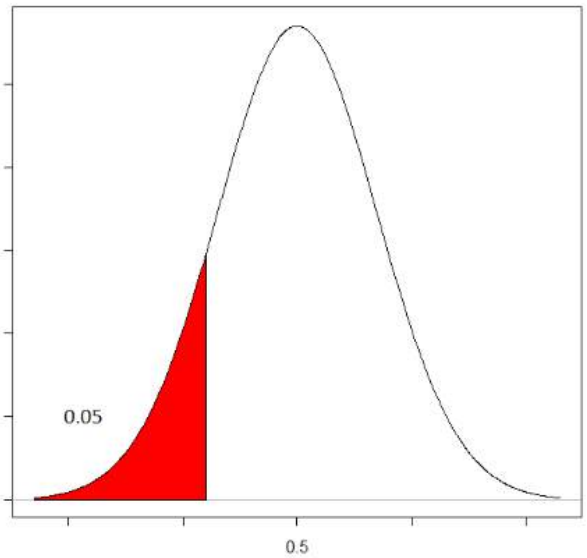


Figura 5.16: Região crítica à esquerda

O procedimento é o mesmo do teste bilateral, supomos que H0 seja verdade (neste caso, ainda supomos que $p = 0.5$) mas a região crítica não será dividida em duas partes, será apenas uma área acima da média contendo todo o valor da significância, caso H1: $p > 0.5$ ou apenas uma área abaixo da média contendo todo o valor da significância, caso H1: $p < 0.5$. Além disso, o p-valor não será 2.P($Z > z$) ou 2.P($Z < z$), e sim P($Z > z$) ou P($Z < z$).

5.2.2 Pela Distribuição Binomial

Vimos na seção anterior como se faz um teste de proporção usando a distribuição normal. Geralmente fazemos esse teste por meio da distribuição normal quando a amostra que estamos trabalhando possui tamanho superior a 30. Caso contrário, usamos a distribuição binomial, a qual vamos apresentar nesta seção.

Suponha que agora estamos trabalhando com uma amostra pequena, de tamanho 10, por exemplo. Suponha que, como no exemplo anterior, queremos testar hipóteses sobre a proporção de opositores e, nessa amostra de tamanho 10, encontramos 7 opositores. Queremos verificar se é plausível que a verdadeira proporção (proporção populacional) seja igual a 0.5 (50%). Então as hipóteses são:

$$\text{H0: } p = 0.5$$
$$\text{H1: } p \neq 0.5$$

Lembremos que a distribuição binomial é uma distribuição de probabilidade que nos informa o quão provável é encontrar a soma de variáveis aleatórias Bernoulli (acontecer ou não acontecer). Em nosso caso, a distribuição binomial nos informa o quão provável é encontrar 0,1,2,3,4,5,6,7,8,9 ou 10 opositores em nossa amostra de 10 pessoas. Como em todo teste de hipóteses supomos que H0 seja verdadeira substituímos, na fórmula da binomial, o p por 0.5 (como está escrito em H0). Feito isto, podemos calcular as probabilidades de ocorrência de qualquer número de opositores em nossa amostra.

Como encontramos 7 opositores em nossa amostra de tamanho 10, a probabilidade disso ou algo mais discrepante ocorrer supondo que a proporção populacional de opositores é igual a 0.5 (50%) é:

- $P(X = 7) = \frac{10!}{7!(10-7)!}0,5^7(1-0,5)^{10-7} = 0.1171$
- $P(X = 8) = \frac{10!}{7!(10-8)!}0,5^8(1-0,5)^{10-8} = 0.0439$
- $P(X = 9) = \frac{10!}{9!(10-9)!}0,5^9(1-0,5)^{10-9} = 0.0097$
- $P(X = 10) = \frac{10!}{10!(10-10)!}0,5^10(1-0,5)^{10-10} = 0.0009$
- $p - valor = 2.(P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)) = 2.(0.1716) = 0.3432$
- Lembre-se que $0! = 1$, $1! = 1$, $2! = 2x1$, $3! = 3x2x1$, ... lembre-se também que qualquer número elevado à zero é igual a um.

O p-valor é igual à duas vezes $(P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$, já que é um teste bilateral: $0.1716x2 = 0.3432$. Se adotarmos uma significância padrão de 0.05 (5%) não rejeitamos H0.

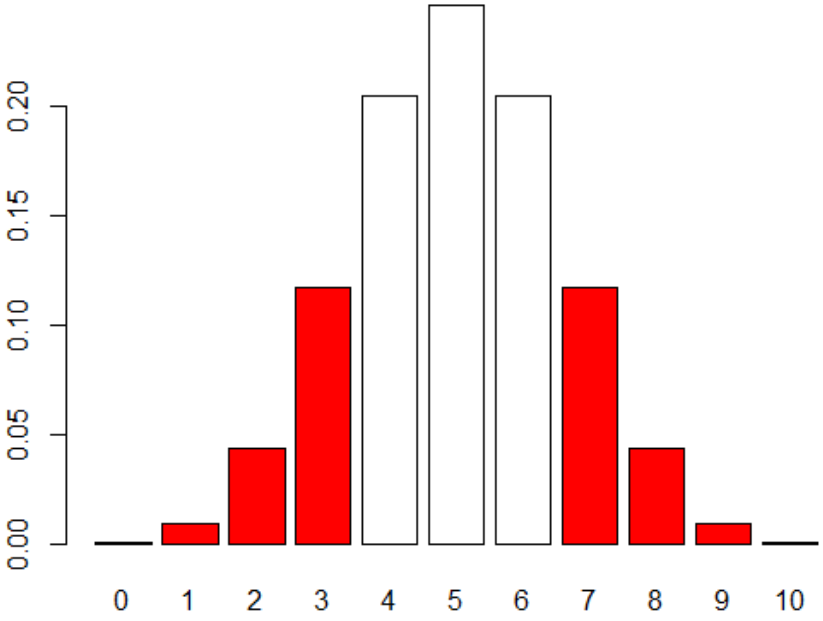


Figura 5.17: Região crítica do teste bilateral na distribuição binomial

Se o teste fosse unilateral não precisaríamos multiplicar $(P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10)$ por 2.

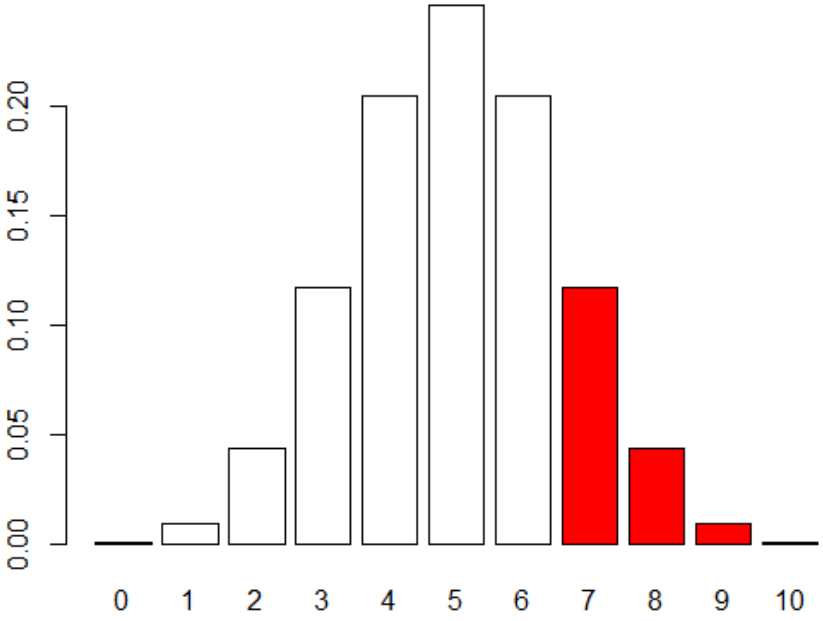


Figura 5.18: Região crítica do teste unilateral à direita na distribuição binomial

E se tivéssemos testando a hipótese de que os opositores são a minoria (teste unilateral à esquerda), calcularíamos a probabilidade de encontrar uma observação menor do que a que encontramos. Por exemplo, se encontrássemos $X = 3$ neste caso, teríamos que calcular: $(P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0)$

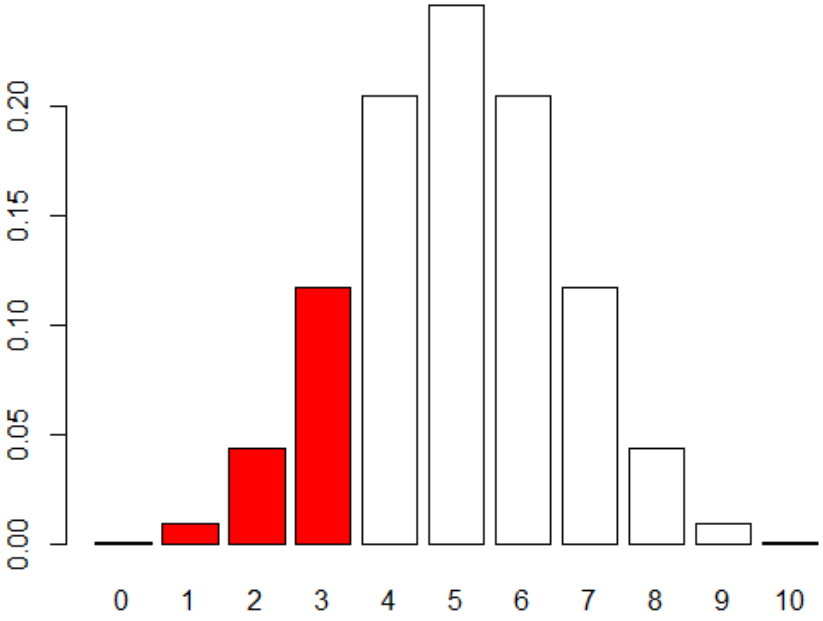


Figura 5.19: Região crítica do teste unilateral à esquerda na distribuição binomial

Por fim, é interessante saber, caso queira utilizar, que quando temos uma amostra grande também podemos usar o teste binomial. Neste caso podemos utilizar a correção de continuidade para aproximar a distribuição binomial de uma normal. Assim, aplicado esta correção, a distribuição de $\hat{p}$ terá média igual a np e variância igual a np(1-p).

A realização de tais testes no Rcommander é feita de forma semelhante a dos testes anteriores, onde utilizamos uma variável categórica(Classe fator) na análise. Novamente, utilizaremos o banco de dados do estresse para exemplificar o teste. Neste caso,suponhamos que o banco conta apenas com uma amostra da turma, e desejamos testar se metade dos alunos moram com os pais, ou seja, se a proporção de alunos que moram com os pais é igual a 0.5. Neste caso, nossas hipóteses são:

$$H_0 : p = 0.5$$
$$H_1 : p \neq 0.5$$

Onde p indica a proporção de alunos que moram com o pais. No Rcommander, com o banco de dados ativo, basta ir no menu superior:

Estatísticas > Frequências/Proporções > Teste de frequência/proporção(1 amostra). Na janela aberta há duas abas. Na aba dados, devemos selecionar a variável referente ao nosso teste, lembre-se que só aparecem as variáveis da classe fator, portanto, se a variável tiver codificada numericamente, ela deverá ser convertida (em sessões anteriores, explicamos como converter)

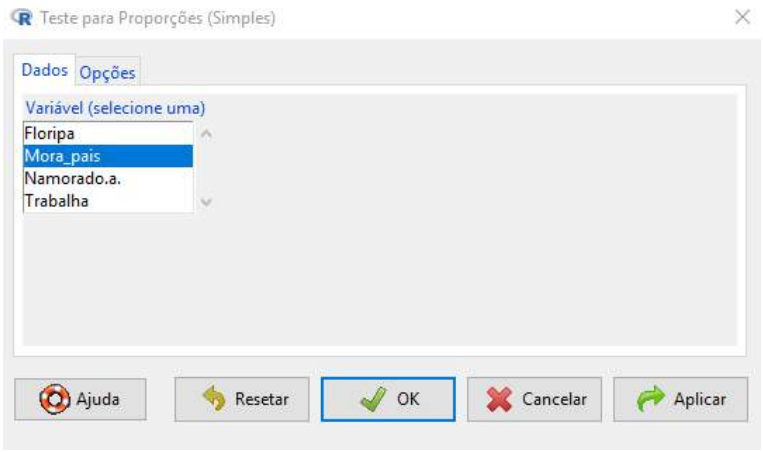


Figura 5.20: Janela de dados

Na aba opções encontramos as especificações do nosso teste, como temos um teste bilateral, especificamos a hipótese alternativa $\neq p0$, e especificamos o valor da hipótese nula, no caso do nosso exemplo 0.5. O nível de confiança (0.95), e escolhemos o tipo de teste. Para este exemplo, o teste utilizado será o binomial exato:

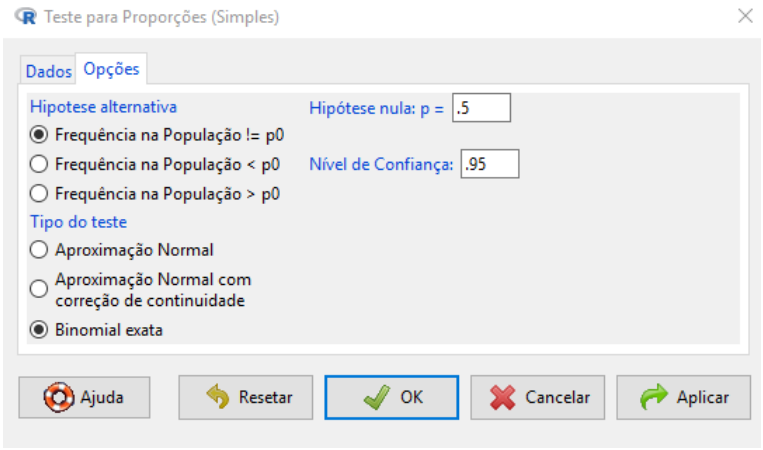


Figura 5.21: Janela de opções

Tendo escolhido, as especificações desejadas, basta clicar em "ok", e observar os resultados obtidos na janela output.

```
Frequency counts (test is for first level):
Mora_pais
Sim Não
 44  51

        Exact binomial test

data:  rbind(.Table)
number of successes = 44, number of trials = 95, p-value = 0.5384
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.3602248 0.5684504
sample estimates:
probability of success
             0.4631579
```

Figura 5.22: Resultado do teste

Os resultados são exibidos de forma bem detalhada. Ele mostra a frequência das observações, especifica a quantidade de sucessos(interesse do teste), o número total de observações (number of trials) , o p-valor, especifica

qual a hipótese alternativa e o intervalo de confiança.

A conclusão do teste pode ser feita tanto através do p-valor quanto através do intervalo de confiança. O intervalo de confiança, vai de 0.3601 até 0.5684. Como o valor testado na hipótese nula (0.5) está presente no intervalo, concluímos que não há evidências para rejeitar a hipótese de que a proporção é igual a 0.5. Olhando para o p-valor e o comparando com o nível de significância, vemos que ele é maior ($0.5384 > 0.05$), obtendo a mesma conclusão que o intervalo de confiança nos forneceu.

5.3 Teste Qui-quadrado

Quando pensamos em testar se existe dependência entre duas variáveis categóricas, é comum pensarmos em usar o teste Qui-quadrado. Por ser um teste tão utilizado, ele não poderia faltar no grupo de testes que o Rcommander faz.

Podemos realizar o teste o primeiro passo é criar nossa tabela de contingência, que é uma tabela com uma variável representada nas linhas e outra nas colunas, e no interior da tabela, os valores representam a frequência em que os pares de categorias acontecem simultaneamente.

Para criar uma tabela de contingência no Rcommander devemos ir no menu superior > estatísticas > tabelas de contingência.

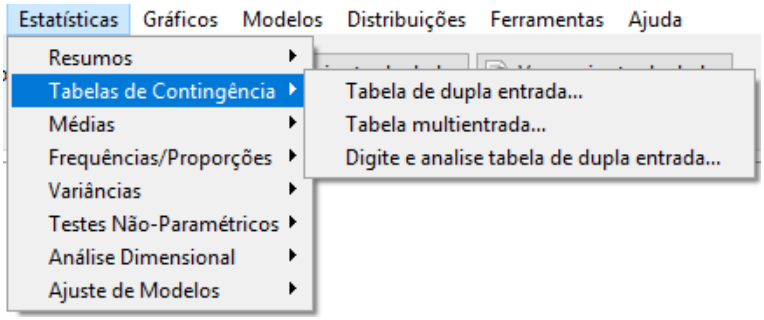


Figura 5.23: Opções para criar tabelas de contingência

Tendo feito isso, temos 3 opções, as opções "tabela de dupla entrada.."e "Tabela multientrada"servem para criar tabelas a partir de um banco já existente, e ultima opção serve para digitar a tabela manualmente. Aqui, continuaremos utilizando o banco de dados do Estresse, mas se digitamos as tabelas manualmente, o processo para definir o teste é o mesmo, portanto bastará digitar a tabela e replicar os passos aqui mostrados.

Antes de aplicar o teste precisamos definir as hipóteses de interesse. Vamos testar aqui, se a turma a qual o aluno pertence tem relação com a naturalidade dele(Floripa ou outras cidades), portanto as hipóteses a serem testadas são:

H_0: Turma e naturalidade são independentes

H_1: Turma e naturalidade são Não são independentes

Como estaremos fazendo o teste com duas variáveis que estão em um banco carregado, selecionaremos a primeira opção : "Tabela de dupla entrada". Será então aberta uma janela com duas abas, a aba dados é onde selecionaremos

as variáveis que farão parte da nossa tabela de contingência. No nosso caso, escolheremos as variáveis Turma e Floripa (que é a variável que indica se a pessoa é natural de Florianópolis ou não).

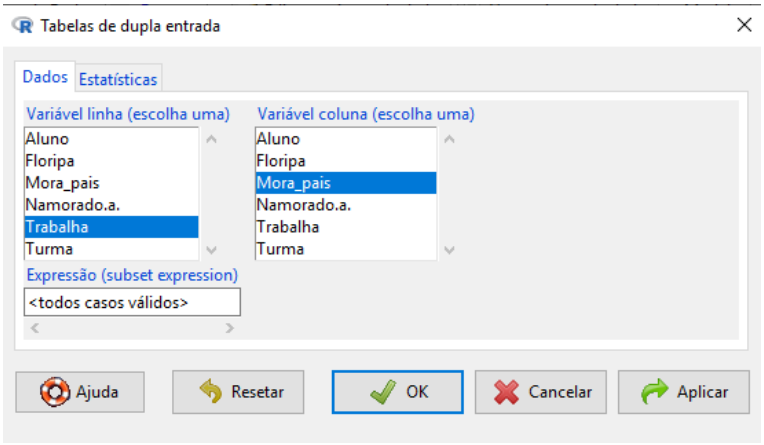


Figura 5.24: Aba para criação da tabela de contingência

Caso fossemos entrar com os dados, bastaria selecionar a terceira opção e digitar manualmente a tabela.

Tendo definido as variáveis para criação da tabela de contingência, iremos utilizar a aba "Estatísticas para definir o teste a ser feito e as medidas a serem exibidas. Podemos optar por exibir as proporções calculadas sob as linhas, colunas ou total. O cálculo desses percentuais não influencia no teste, portanto, não os exibiremos para obter uma visualização mais limpa. Logo em seguida, temos as opções de testar hipóteses, para realizar o teste, basta marcar a caixa relativa ao teste de independência Qui-quadrado, no entanto é sempre importante também selecionar a opção "Apresentar frequências esperadas".

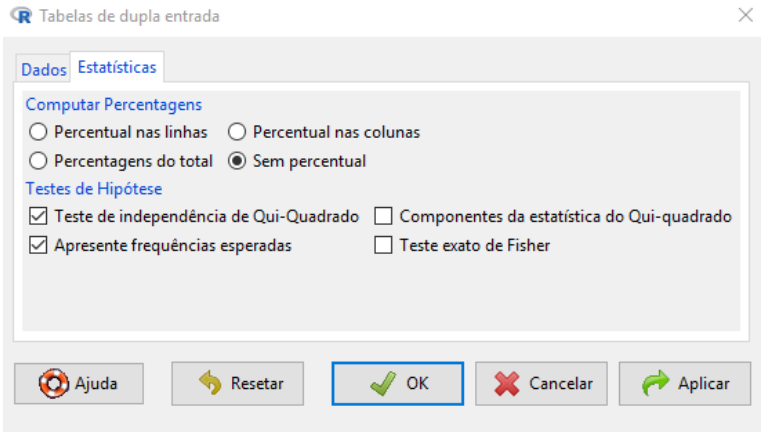


Figura 5.25: Aba para Estatísticas do teste

O teste é feito com base na suposição de que a hipótese nula seja verdeira, portanto, se as variáveis foram independentes, esperamos que os valores observados em cada caselas sejam dados por: $E_{ij} = \frac{n_{i.} * n_{.j}}{n}$, e a estatística de teste será dada por: $\chi^2_{obs} = \sum_{i=1}^{r} \sum_{i=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$ que tem distribuição χ^2 com ((nº de linhas)-1)*((nº de colunas)-1) graus de liberdade, onde O_{ij}, indica o valor observado no encontro da linha i com a coluna j e E_{ij} indica o valor esperado para este mesmo encontro de linha e coluna. Se as duas variáveis de fato forem independentes, os valores observados serão bem próximos dos valores esperados, e a estatística de teste será um valor baixo, para mensurar o quão baixo, calculamos a probabilidade de observarmos o valor observado ou um valor maior que este de acordo com

a distribuição da estatística de teste, esta probabilidade é conhecida como p-valor, e se ele for maior que o nível de significância adotado, não rejeitamos a hipótese de independência.

Sabendo isso, basta observar o output do nosso teste.

```
Frequency table:
        Floripa
Turma     Natural Outra_Cidade
  2007_2      11            17
  2008_1      10            22
  2008_2      12            23

        Pearson's Chi-squared test

data:  .Table
X-squared = 0.43033, df = 2, p-value = 0.8064


Expected counts:
        Floripa
Turma       Natural Outra_Cidade
  2007_2  9.726316     18.27368
  2008_1 11.115789     20.88421
  2008_2 12.157895     22.84211
```

Figura 5.26: Output do teste

O primeiro item do output é a nossa tabela de contingência criada com base nos nossos dados, é dela que retiramos os valores observados. Em seguida temos o resultado do nosso teste: X-squared representa a estatística de teste, df são os graus de liberdade e p-value é o p-valor. No nosso exemplo, como ele é maior que 0.05(nível de significância adotado) nós concluímos que não há evidências contrárias a hipótese nula.

O ultimo item do nosso output são os valores esperados, é importante exibi-los pois se houverem valores esperados iguais a 0, ou muitos valores esperados menores que 5, o teste qui-quadrado deixa de ser a melhor opção para testar a hipótese. Nos casos em que tais condições não forem satisfeitas, é preferível usar o teste exato de Fisher.

Para fazer o teste exato de Fisher, basta seguir os mesmos passos de criar uma tabela de contingência, e na aba estatísticas selecionar a opção referente a tal teste. o método de Fisher é computacionalmente mais pesado, portanto quando as tabelas de contingência são muito grandes, o computador não é capaz de realizar tais testes.

Por fim, vale ressaltar que tanto o teste Qui-quadrado quanto o teste exato de Fisher servem para avaliar se as variáveis são ou não independentes, caso as variáveis sejam dependentes é possível calcular algumas medidas adequadas para identificar "onde"se encontra a dependência.

5.4 Teste de correlação

É extremamente importante entender como é a relação de duas variáveis. Há várias formas de avaliar isso, os coeficientes de correlação servem para medir essa associação, e testes para avaliar se essa associação é significativa. Há vários coeficientes que medem a correlação, no Rcommander podemos usar 3 deles, o de Pearson(r), o de Spearman(ρ) e o de Kendall(τ). Estes 3 coeficientes são valores entre -1 e 1, onde valores negativos indicam correlação negativa,

e valores positivos indicam correlação positiva. Quanto mais próximo de 0, mais fraca é a correlação, portanto, testamos de as correlações são significativamente diferentes de 0.

- **Pearson(r):** O coeficiente de correlação de Pearson mede o grau de correlação LINEAR entre duas variáveis quantitativas. Sendo calculado da seguinte forma: $r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 * \sum(y_i-\bar{y})^2}}$
- **Spearman(ρ):** É uma medida de correlação não-paramétrica que não requer a suposição que a relação entre as variáveis é linear. Pode ser usado tanto para variáveis numéricas quanto para variáveis ordinais. A ideia é ordenar cada variável, para obter se posto, depois olhar os pares e comparar de acordo com o posto.(O posto não altera os pares). Sabendo isso, o cálculo é feito da seguinte forma: $\rho = 1 - \frac{6*\sum(postodex_i - postodey_i)^2}{(n^3-n)}$
- **Kendall(τ):** É uma medida de associação para variáveis ordinais. Seu calculo é feito da seguinte forma: $\tau = \frac{(\text{n° de pares concordantes}) * (\text{n° de pares discordantes})}{n(n-1)/2}$, se houver empates(valores iguais em uma mesma variável) é feito um ajuste nesta fórmula.

Para calcular estas correlações e testar se elas são significativamente diferentes de 0, é preciso estar com elas no formato numérico, ou seja, caso seja ordinal, devemos codificá-la. Usando o banco de dados do Estresse, testaremos se o número de créditos é significativamente correlacionado com a educação, para isso, basta ir no menu superior > Estatísticas > Resumos > Testes de correlação ..

Na janela aberta, basta escolher as variáveis (com a tecla ctrl conseguimos escolher a segunda), o tipo de coeficiente a ser calculado e a hipótese alternativa.

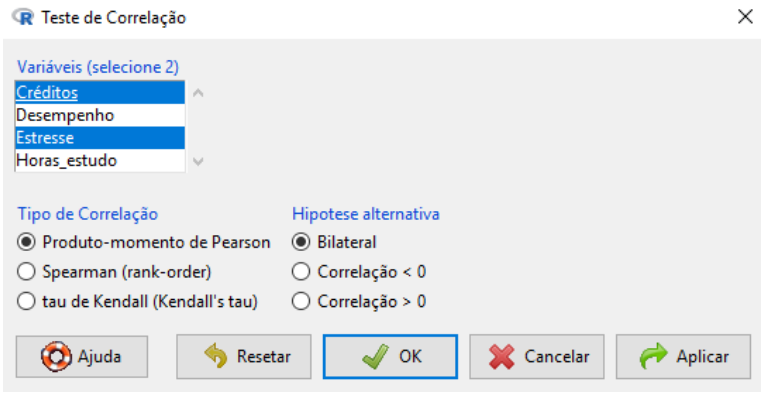


Figura 5.27: opções

Ao clicar em ok, teremos o valor aparece no output, junto o p-valor e o intervalo de confiança;

```
        Pearson's product-moment correlation

data:  Créditos and Estresse
t = -0.56671, df = 92, p-value = 0.5723
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2585080  0.1453739
sample estimates:
        cor
-0.05898032
```

Figura 5.28: Output do teste

Podemos observar que o coeficiente de correlação de Pearson entre créditos e estresse é igual a -0.05, e que ele não é significativamente diferente de 0, pois seu p-valor é igual 0.5723

Capítulo 6

Modelos Lineares

As análises estatísticas vão além de simples testes de hipótese, há vários outros métodos e modelos que resolvem diferentes problemas, e se adéquam a cada estrutura de dados. É comum presenciarmos situações em que desejamos explicar a distribuição de uma variável com base em outra, ou outras. Os modelos de regressão servem para atingir este objetivo. Ou seja, temos uma variável que queremos explicar a qual chamamos de variável dependente, e outras variáveis que queremos usar para entender o comportamento da variável dependente, estas outra variáveis são chamadas de variáveis independentes.

Existem inúmeros tipos de regressão para diferentes tipos de dados e distribuições. Aqui, trataremos apenas do ajuste modelos lineares.

6.1 Regressão Linear simples

O modelo de regressão linear simples usa apenas uma variável independente para explicar a variável dependente. Esse modelo é explicado através de uma reta onde Y é variável dependente e X a variável independente.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

No modelo descrito, β_0 determina o intercepto da reta, ou seja, o valor onde x=0, β_1 indica a inclinação da reta e ϵ indica os resíduos do modelo. Os resíduos são normalmente distribuídos, com média 0 e variância constante.

Nos modelos de regressão linear simples, o método utilizado para estimar a reta de regressão é o método dos mínimos quadrados, que consiste em buscar a reta que minimiza o quadrado da diferença entre os valores reais e os valores ajustados pela reta.

O modelo ajustado será dado por:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Onde o "chapéu"indica os valores estimados. Ou seja cada $\hat{y_i}$ é o valor estimado de y na observação i. E nossos resíduos ϵ_i são dados pelos valores reais de Y menos os valores estimados em cada observação.

Para mostrar como fazer uma regressão linear simples, utilizaremos o banco de dados Duncan, que está no pacote "carData", que é automaticamente carregado quando utilizamos o Rcommander. Este banco de dados é composto de 45 observações de 45 ocupações nos EUA em 1950.

- **type** É uma variável categórica que indica o tipo da ocupação: prof = profissional e gerencial, wc = colarinho branco e wb= colarinho azul
- **income:** Indica a porcentagem de ocupantes no censo que ganharam mais de $ 3.500 ou mais por ano (cerca de $36000 em dólares americanos em 2017)
- **education** Porcentagem de ocupantes ocupacionais em 1950 que eram formados no ensino médio (que, diríamos que é aproximadamente equivalente a um doutorado em 2017)
- **prestige** Porcentagem de entrevistados em uma pesquisa social que classificaram a ocupação como "boa" ou melhor em prestígio

em seções anteriores deste manual, explicamos como carregar um banco de dados de um pacote. Relembrando, basta ir no menu superior > dados > conjunto de dados em pacote > ler dados de pacote "attachado"... ; Na janela aberta dar um clique duplo no pacote carData e em seguida buscar pelo banco Duncan.

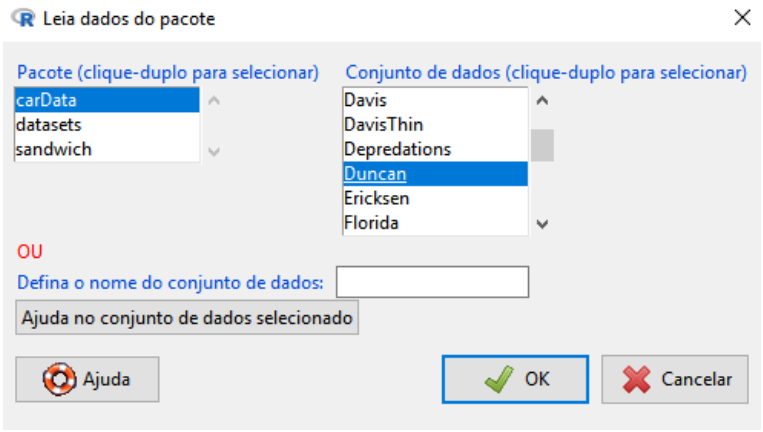


Figura 6.1: Seleção do banco Duncan

Tendo estes dados, vamos tentar entender se a porcentagem de pessoas com classificação boa ou melhor em prestigio, pode ser explicada pela porcentagem de pessoas que eram formadas no ensino médio. Ou seja, a variável prestige será a variável dependente(ou variável resposta) e education a variável independente. Antes de partir para o ajuste de modelo, é recomendado fazer gráficos e olhar a correlação entre as variáveis, pois o procedimento a seguir é válido, se a relação entre as duas variáveis for linear.

Para ajustar o modelo, devemos ir no menu superior e selecionar estatísticas > Ajustes de modelo > Regressão Linear...

Na janela aberta, são exibidos os nomes das variáveis para que possamos escolher a variável resposta e as variáveis explicativas(independentes), como estamos em um modelo simples, escolheremos apenas uma variável explicativa. Portanto, selecionaremos a variável prestige no campo da variável resposta e a variável education no campo das variáveis explicativas.

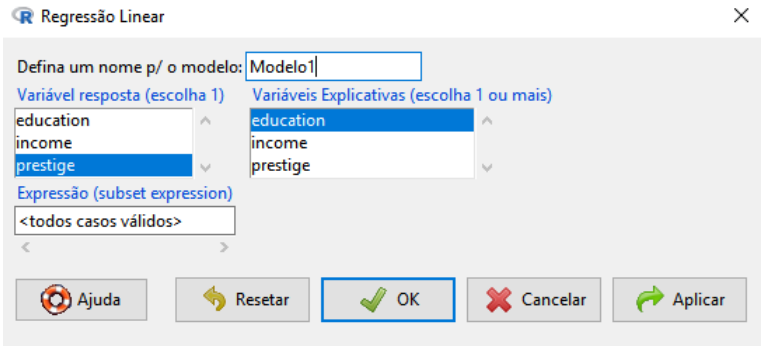


Figura 6.2: Janela de opções para ajustar o modelo

Após clicar em ok, o modelo será ajustado, e na janela output poderemos observar o resumo do nosso modelo. Estes resumos são as informações exibidas na maioria dos softwares que realizam tais análises. Observe o resultado a seguir, e seus resumos serão explicados em seguida.

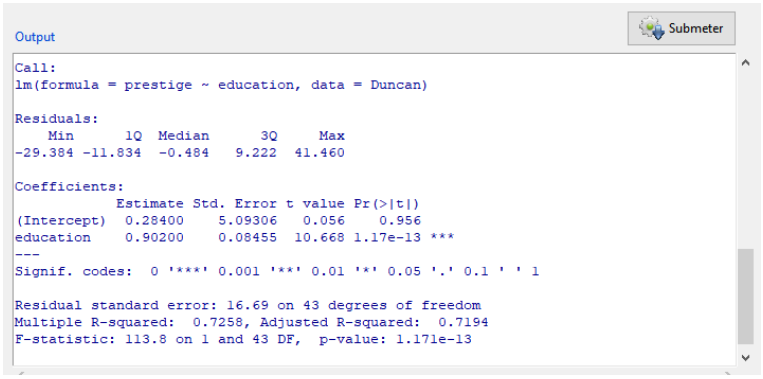


Figura 6.3: Janela de opções para ajustar o modelo

A sessão Call indica justamente a chamada do R para a regressão. Se estivermos usando o software R sem usar o rcommander, a função para ajustar o modelo é ajustada como mostrado nessa seção.

Na sessão residuals temos informações sobre a distribuição dos resíduos. tais como, o mínimo, o 1º quartil, a mediana, o 3º quartil e o máximo. Lembrando que os resíduos são a diferença entre o valor ajustado pela regressão e o valor real observado.

Na sessão Coefficients, podemos encontrar os valores estimados e informações relacionadas a ele. Estimate indica os valores estimados dos coeficientes, Std. Error indica o erro padrão associado ao coeficientes, t value e Pr(>|t|) indicam respectivamente os valores de teste e os p-valores para testar se os coeficientes são significativamente diferentes de 0.

Na linha Signif. codes: temos a indicação do significado dos asteriscos mostrados ao lado de cada p-valor. '***', indica que o p-valor está entre 0 e 0.001, '**' indica que o p-valor está entre 0.001 e 0.01, '*' indica que o p-valor está entre 0.01 e 0.05, '.' indica que o p-valor está entre 0.05 e 0.1, e se não houver nenhum simbolo a frente do p-valor, é por que o p-valor é maior que 0.1.

Na última sessão temos o erro padrão dos resíduos e seus graus de liberdade, temos também o coeficiente de determinação e também a estatística F, junto com seus graus de liberdade e p-valor para testar se todos os coeficientes são iguais a 0.

Outras informações do modelo podem ser extraídas indo no menu superior e na aba modelo. Convidamos o leitor a explorá-la, afinal o uso é bem intuitivo. Algumas partes dessa aba serão exploradas posteriormente.

Outra parte de extremo interesse é enxergar graficamente a reta ajustada. No modelo linear simples é possível fazer isso através das opções padrões de gráfico. Basta ir no menu superior>gráficos> diagrama de dispersão. Em seguida, na aba dados da janela aberta, selecione as variáveis x e y de acordo com o que escolhemos em nosso modelo.(y= prestige e x = education)

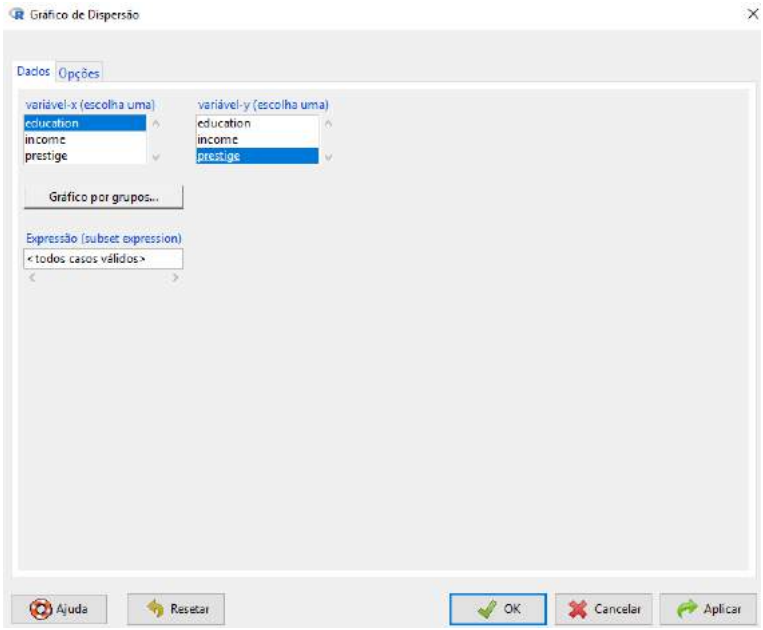


Figura 6.4: Aba de escolha de dados

Em seguida clique na aca de opções, dentro dela há varias opções para personalizar nosso gráfico, a opção de interesse para visualizar a reta de regressão é a opção "linha de quadrados mínimos". Portanto selecionaremos tal opção e depois clicamos em "ok"para gerar o gráfico

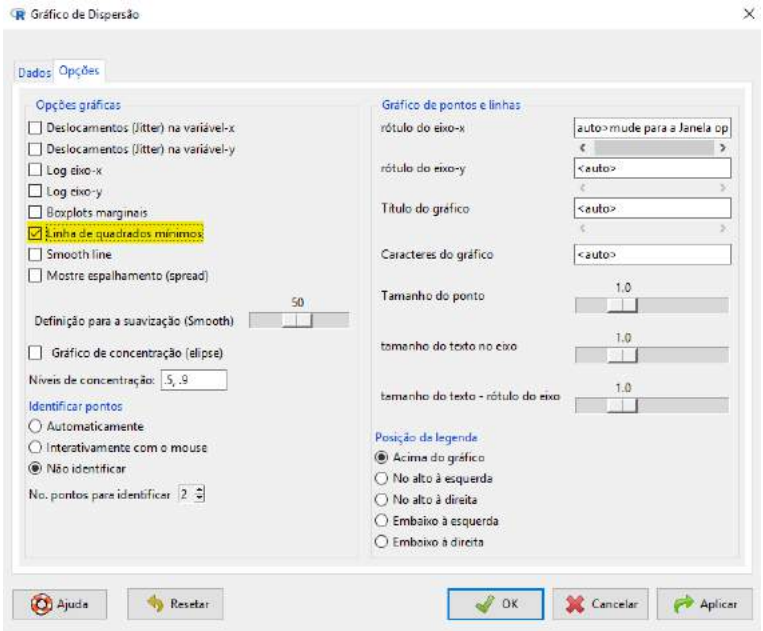


Figura 6.5: Janela de opções do gráfico

Tendo feito isso, o resultado obtido é um diagrama de dispersão com a reta de regressão traçada.

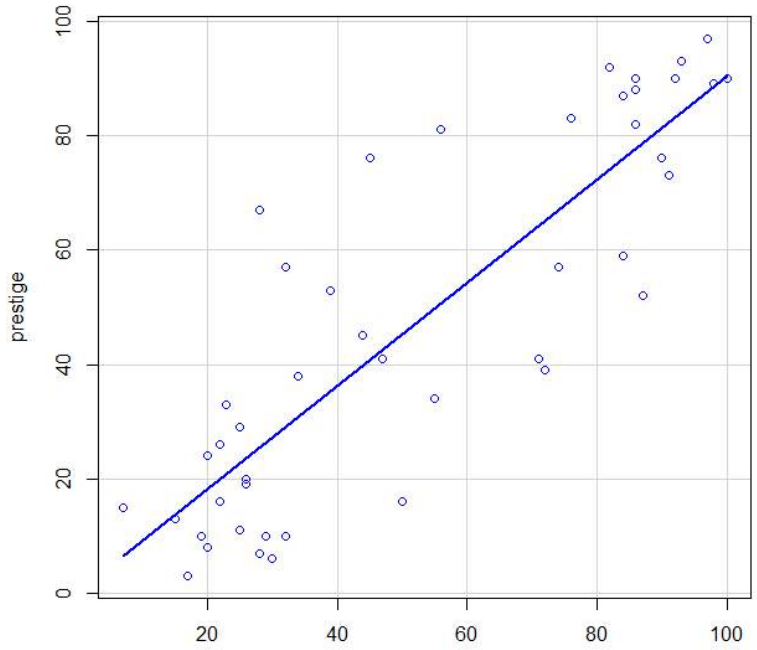


Figura 6.6: Gráfico com a reta de regressão

6.2 Regressão Linear múltipla

O modelo de regressão linear múltipla usa mais de uma variável independente para explicar a variável resposta. O modelo é da forma:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i$$

Onde k é o número de variáveis explicativas. β_0 indica o intercepto da reta, e os outras β's são os coeficientes associados as outras variáveis. Os resíduos são normalmente distribuídos, com média 0 e variância constante.

Novamente, o método utilizado para estimar a reta será o método de mínimos quadrados, e o modelo ajustado será dado por:

$$\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$$

Lembre-se que o ˆ encima, indica que são valores estimados. Portanto, cada $\hat{y_i}$ é o valor estimado de y na observação i. Os nosso resíduos ϵ_i serão dados pela subtração dos valores reais pelos valores estimados.

Continuaremos utilizando o banco de dados Duncan, e agora usaremos as porcentagens de pessoas que tinham o ensino médio e a porcentagem de pessoas que recebiam mais de R$ 3500 para explicar a porcentagem de pessoas com classificação boa ou melhor. Ou seja, nossa variável dependente ou resposta, continuará sendo a variável prestige, e teremos duas variáveis explicativas(ou independentes), education e income.

Lembramos novamente que neste caso utilizamos variáveis que tem relação linear com a variável resposta, caso a relação não seja linear, podemos aplicar transformações para linearizar esta relação(trataremos disso adiante).

Portanto lembre-se de fazer uma boa análise descritiva para certificar-se de que é razoável dizer que a relação é linear(O diagrama de dispersão é uma ótima forma de enxergar tal relação)

Para fazer o ajuste no rcommander, devemos ir no menu superior e depois em estatísticas > Ajustes de modelo > Regressão Linear...

Na janela aberta selecionamos a variável prestige como variável resposta, e as outras variáveis como variáveis explicativas. Para selecionar mais de uma variável na aba explicativa basta segurar o botão direito do mouse e arrastar a seleção ou pressionar a tecla ctrl e clicar em cada variável desejada.

Sempre dê um nome para o modelo que seja fácil de identificar. Muitas vezes precisamos ajustar vários modelos até chegar no ideal, se dermos nomes confusos ao nosso modelo, pode ser que fiquemos perdidos.

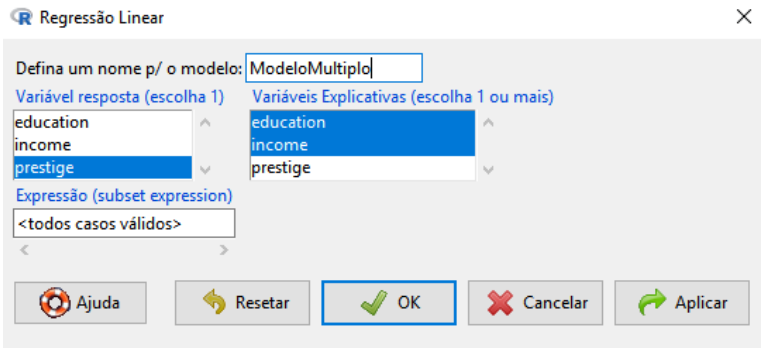


Figura 6.7: Opções para ajustar o modelo

Após clicarmos em OK, o modelo será ajustado, e seu resumo será exibido na parte do output. Os resumos exibidos são os mesmos que para a regressão linear simples, tornado desnecessária a explicação de cada parte novamente.

```
Call:
lm(formula = prestige ~ education + income, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-29.538  -6.417   0.655   6.605  34.641

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420      0.163
education    0.54583    0.09825   5.555 0.00000173 ***
income       0.59873    0.11967   5.003 0.00001053 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared:  0.8282, Adjusted R-squared:   0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Figura 6.8: Resultado do modelo

Capítulo 7

Simulações

7.1 Números aleatórios

Um número aleatório é um número que é escolhido sem obedecer nenhuma sequência ou padrão, por exemplo, ao jogarmos um dado não viciado, o valor a ser sorteado é aleatório. Imagine que colocamos os valores de 1 a 100, em uma caixa, e sorteamos um deles, novamente estamos lidando com valores aleatórios. È como usarmos softwares para obtenção de valores aleatórios.No entanto os computadores geram valores pseudo-aleatórios que, que nada mais são do que valores que tem algumas propriedades dos valores aleatórios.

Nas simulações os valores pseudo aleatórios são amplamente usados. A possibilidade de especificar algumas características e com base nestas gerar valores de forma "aleatória"é muito útil. Imagine que desejamos brincar de cara e coroa, sem usar uma moeda, apenas com o computador. Podemos especificar que queremos dois valores (0 e 1), onde um especifica cara, e outro coroa, tendo os dois valores probabilidade igual de ser sorteada. Quando o software usado sortear o valor 0 ou 1, temos o resultado do jogo cara e coroa. Da mesma maneira, usamos os valores pseudo aleatórios para fazer simulações provenientes de distribuições de probabilidade.

Uma vantagem do uso dos valores pseudo-aleatórios, é o fato de que tal estrutura permite a reprodutibilidade dos valores através de uma "semente". Com uma mesma semente, geramos uma mesma sequência de números pseudo-aleatórios. Ou seja, se duas pessoas usarem a mesma semente, elas obterão os mesmo valores pseudo-aleatórios. No Rcommander, definimos a semente indo no menu superior e selecionando: Distribuições > Definir semente geradora de números aleatórios. Onde será aberta uma janela onde é possível selecionar uma semente entre 1 e 100000.

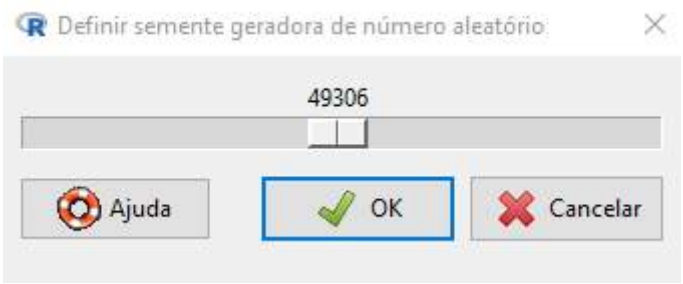


Figura 7.1: Janela para definir a semente para amostragem

Caso essa semente não seja fixada, a cada vez que formos gerar valores, ela é mudada, preservando as características de aleatoriedade.

7.2 Simulações de distribuições de Probabilidade

Para exemplificar o processo de geração de valores provenientes de distribuições de probabilidade usando o Rcommander, vamos gerar uma amostra pseudo-aleatória da distribuição exponencial com taxa = 10. Para isso devemos buscar no menu superior a opção Distribuições, e ir selecionando de acordo com a distribuição de probabilidade que desejamos. No nosso caso: Distribuições> Distribuições Contínuas > Distribuição Exponencial > Amostragem da distribuição exponencial. Onde será aberta uma janela para que possamos escolher nosso valores desejados. Para cada tipo de distribuição aparecerá as opções para especificar seus parâmetros. No caso da distribuição exponencial, o único parâmetro a ser especificado é a taxa. Em seguida, devemos especificar também o tamanho da amostra e o número de observações. Neste caso queremos apenas uma observação de uma amostra de tamanho 1000. No caso em que escolhemos várias amostras, podemos optar por obter também a média, soma e desvio padrão amostral. Como neste caso, usaremos apenas uma amostra, não é necessário adicionar os valores. Podemos também, definir o nome do conjunto de dados, para que seja mais fácil identificá-lo (embora ele fique ativo assim que definimos a amostra):

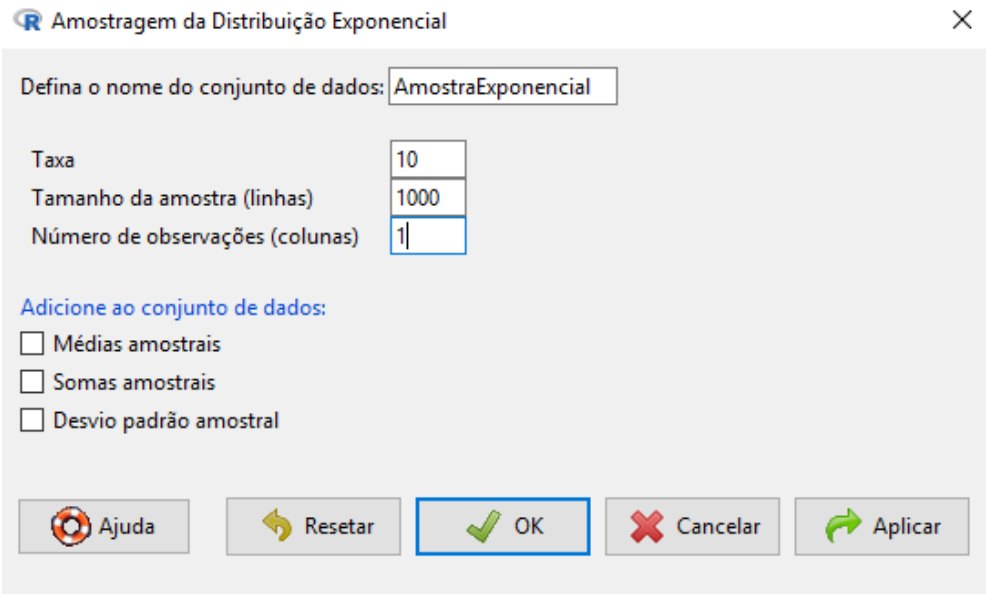


Figura 7.2: Janela de opções para criar a amostra

Para verificar se a amostragem está realmente de acordo com a distribuição, podemos fazer o histograma do valor obtido e comparar com a curva de densidade da distribuição exponencial, que também será obtida no Rcommander. O método para criação do histograma já foi explicado em sessões anteriores (Basta ir em gráficos>histograma e selecionar a variável obs). Para obter a densidade da distribuição exponencial, devemos ir em: Distribuições> Distribuições Contínuas > Distribuição Exponencial > Gráfico da distribuição exponencial.

Na janela aberta, especificamos as características nosso gráfico. Para a comparação desejada, precisamos apenas

de especificar a taxa, selecionar, gráfico da função densidade e x-values:

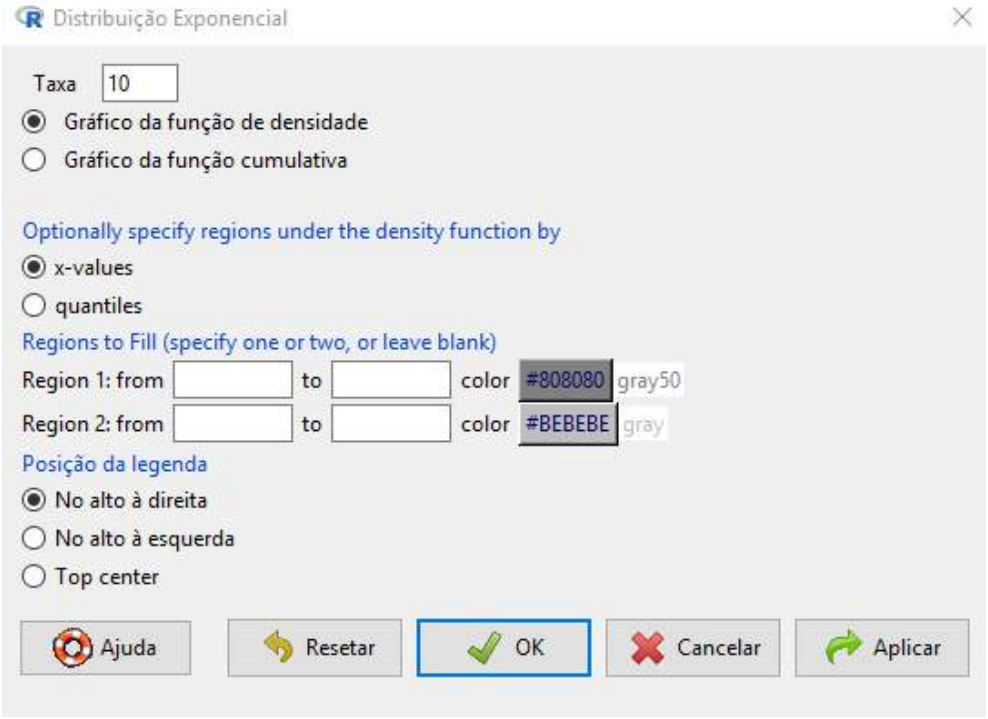


Figura 7.3: Janela de opções para gerar a densidade

O gráfico da densidade original, e o histograma da amostra gerada ficaram da seguinte forma:

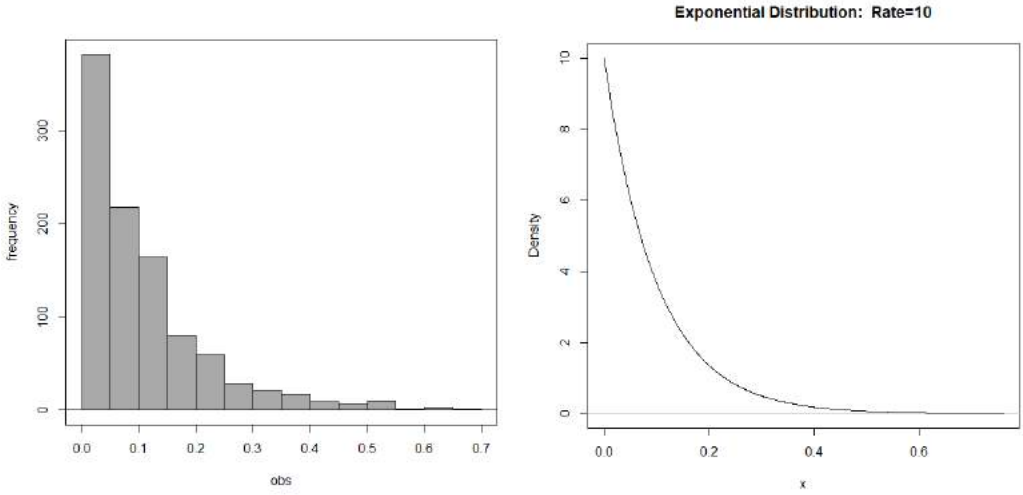


Figura 7.4: Histograma da amostra e densidade teórica

Através da comparação é possível observar a semelhança entre o histograma e a curva de densidade, isso serve para enxergarmos que os valores que geraram o histograma podem ser provenientes de uma distribuição exponencial. Um fato importante a ser considerado é o tamanho da amostra. Neste caso, utilizamos uma amostra de tamanho 1000, que é um tamanho razoável para observar a semelhança. Quanto maior a amostra , mais claro fica a semelhança com a distribuição, pois amostras pequenas podem não ter observações suficientes para uma semelhança visível da distribuição. O que não quer dizer que os valor deixam de ser provenientes da distribuição especificada. Imagine que você simulou uma amostra de tamanho 5. O histograma não será semelhante a densidade, por ainda não há

observações suficientes, mas os 5 valores continuam sendo observações da distribuição exponencial.

7.3 A lei dos grandes números

A lei dos grandes números é uma das principais leis assintóticas(Grandes amostras) da estatística. A ideia básica é que a medida que o tamanho da amostra vai aumentando, o valor da média amostral vai se aproximando da média da distribuição teórica. Esta é uma ideia bem intuitiva, que fica mais clara ainda quando aplicamos em algum exemplo:

Imagine que estamos lançando uma moeda honesta e observando o resultado obtido. Neste caso, tratamos de uma distribuição Bernoulli(1= cara, 0= coroa), com média p=0.5 . Ao lançarmos a moeda algumas vezes, e observamos a média dos lançamentos, poderemos obter qualquer valor entre 0 e 1. Por se tratar de algo aleatório, não há nada que garanta a média da amostra. Imagine que em 3 lançamento, a moeda caiu cara nos 3, neste caso a média é 1 : (1 + 1+ 1)/3. Quando vamos aumentando o tamanho da amostra, é esperado que os dois valores apareçam, e mais, por terem igual probabilidade de ocorrer em cada lançamento, esperamos que eles apareçam em proporções parecidas (fazendo com que a média fique próxima de 0.5)

Um outro exemplo poderia ser o lançamento de um dado honesto. imagine que estamos interessados em observar a face do dado igual a 2. (1 se a face é 2, 0 caso contrario) A probabilidade de em apenas um lançamento, a face ser 2 é 1/6 = 0.16666(Probabilidade teórica), sendo esta a esperança da distribuição Bernoulli especificada. Imagine que vamos lançar o dados repetidas vezes e observar a proporção de faces iguais a 2(no caso de valores 0 e 1, a proporção equivale á média). Quanto mais vezes lançamos o dados, a proporção de faces igual a dois de aproximará de 1/6.

No Rcommander, a simulação da lei dos grandes números é um pouco trabalhosa e manual. Uma das formas mais fáceis é exemplificada a seguir: Usaremos como exemplo a distribuição normal, com média 10 e desvio padrão 3. De acordo com a lei dos grandes números, a medida que o tamanho da amostra for aumentando, a média amostral vai se aproximando da média populacional que é 10.

A ideia aqui consiste em criar um banco de dados onde armazenaremos os tamanhos de amostra e suas respectivas médias. E em seguida gerar cada amostra, observar a média e armazená-la manualmente ao banco de dados criado.

Já explicamos anteriormente, neste manual como criar um banco de dados, portanto aqui citaremos sem entrar em detalhes. Basta ir no menu superior, selecionar Dados e em seguida novo conjunto de dados, na janela aberta escrever um nome para o banco(no nosso caso, usei Lei_G_Numeros), e clicar em ok. Na nova janela Aberta colocar a quantidade de linhas e colunas desejadas. No nosso caso, utilizaremos 20 tamanho diferentes de amostra, portanto, serão 20 linhas(Uma para cada amostra) e duas colunas(Uma contendo o tamanho da amostra e a outra contendo a média amostral). Na coluna dos tamanho amostrais basta colocar em cada linha os tamanhos que usaremos. Neste exemplo usarei amostras de tamanho 1 ao 15 e mais 5 amostras de tamanho 20,25,30,35 e 40. Neste passo, podemos deixar a coluna da média com dados faltantes, pois iremos inseri-los manualmente nos próximos passos.

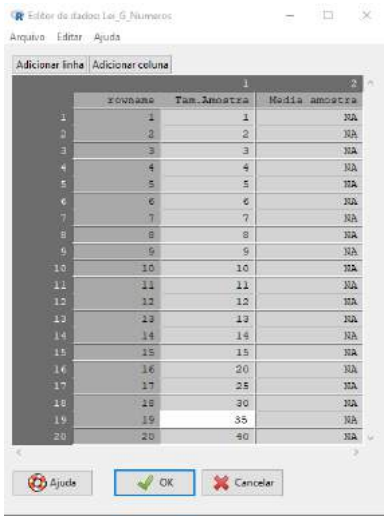


Figura 7.5: Prévia do banco de dados

Depois de clicar em ok, o banco estará criado. E poderá ser visualizado e editado. O próximo passo é gerar amostra para os tamanhos desejados. Infelizmente o Rcommander não nos dá meios para gerar amostras de tamanhos diferentes de uma vez só, portanto precisaremos criar uma por uma e ir atualizando o banco (A minha recomendação é anotar num papel e depois passar para o banco, mas você pode atualizar o banco a cada amostra gerada).

Para gerar a primeira amostra de tamanho 1, iremos em Distribuições> Distribuições Contínuas > Distribuição Normal > Amostragem da distribuição Normal. Na janela aberta basta definir a média e o desvio padrão desejados e o tamanho da amostra. O numero de observações(colunas) será sempre 1, neste caso.

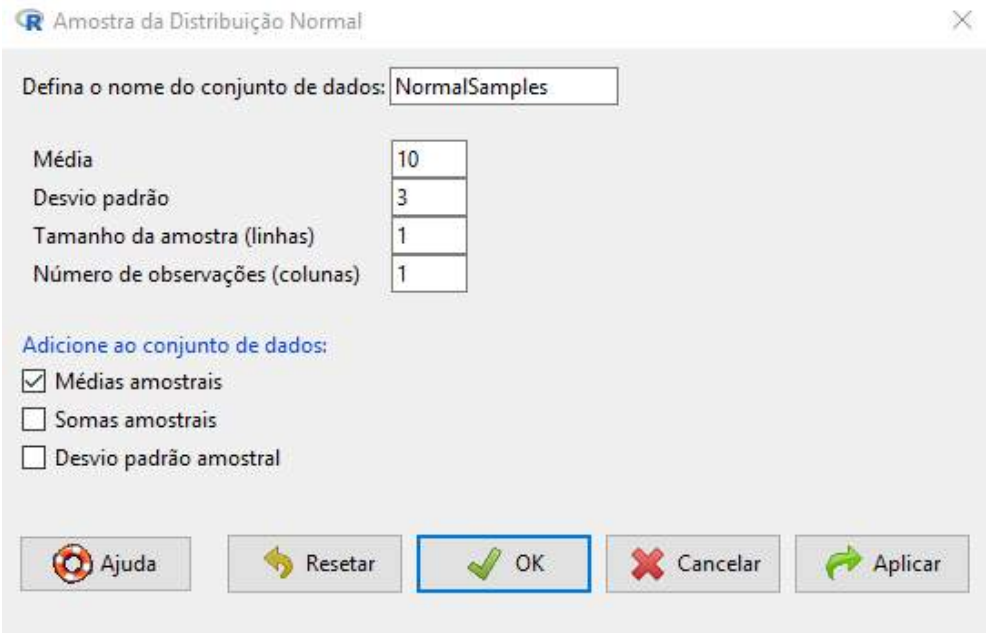


Figura 7.6: Opções para gerar amostra de tamanho 1

Como a amostra é apenas 1, basta visualizar a amostra e anotar este valor como sua média, portanto, após observar tal valor, vá em conjunto de dados, selecione o banco criado para a lei dos grandes números e deixe-o ativo. Em seguida va em editar conjunto de dados, e na linha f=referente a amostra de tamanho um, coloque o valor observado na coluna da média.

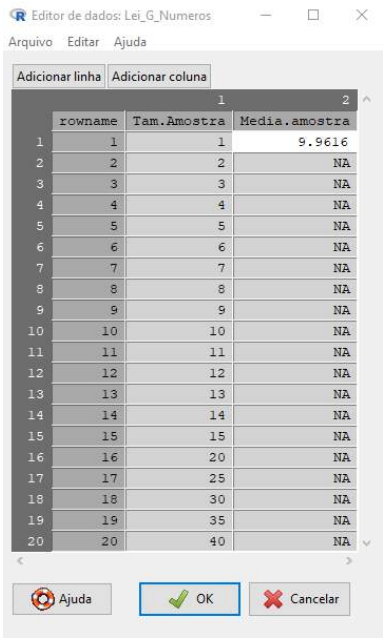


Figura 7.7: Opções para gerar amostra de tamanho 1

Salve a alteração e passe a gerar o segundo valor, que e uma amostra de tamanho 2. Novamente iremos em Distribuições> Distribuições Contínuas > Distribuição Normal > Amostragem da distribuição Normal. E neste caso, selecionaremos o tamanho da amostra 2, e usaremo a mesma média e desvio padrão estabelecidos(10 e 3)

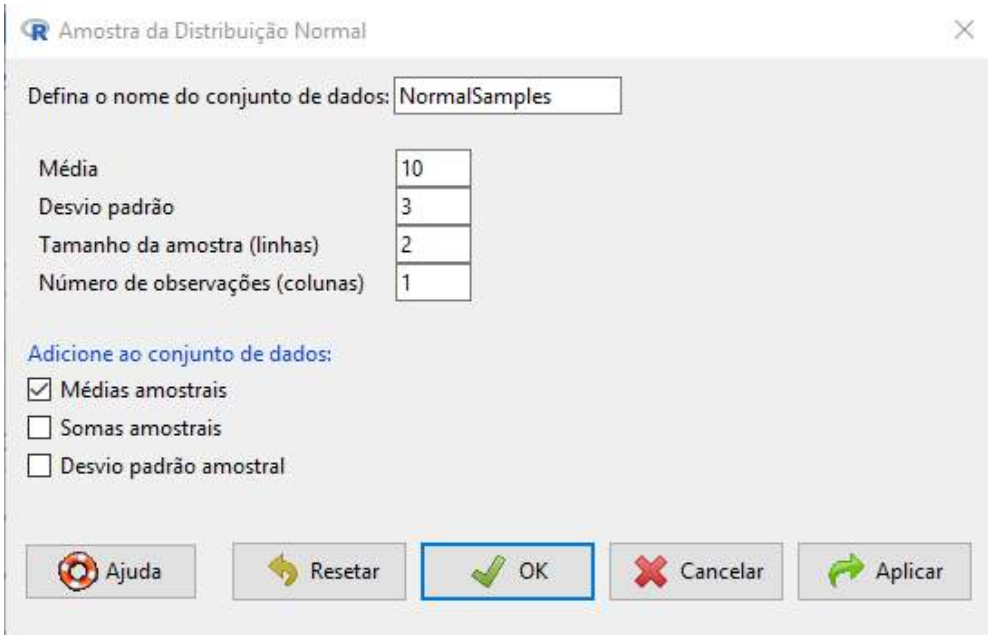


Figura 7.8: Opções para gerar amostra de tamanho 2

Como agora geramos uma amostra de tamanho 2, precisamos obter sua média. Para isso, vamos no Menu Superior, em Estatísticas > Resumos> Resumos numéricos. Como haverá apenas uma variável basta clicar em ok. e na janela output serão exibidos os resumos numéricos, onde deveremos observar nosso resumo de interesse, que é a média.

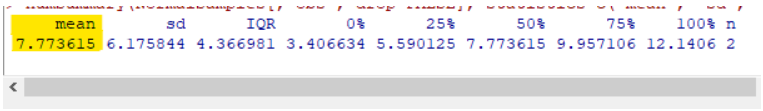


Figura 7.9: Resultado na janela Output

Pegamos a média observada e novamente a registramos no banco de dados. Novamente, basta ir em conjunto de dados, selecionar o banco de dados que criamos para a lei dos grandes números e atualiza-lo com a média da amostra de tamanho 2.

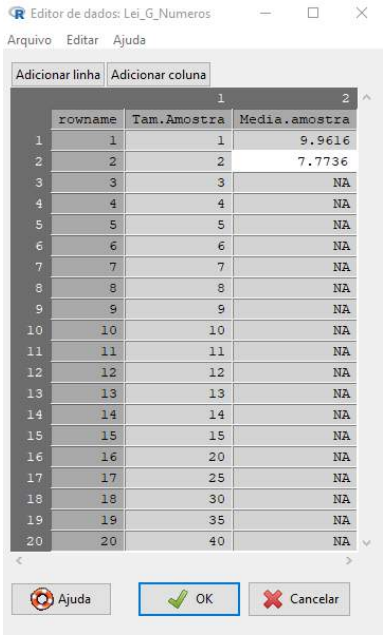


Figura 7.10: Banco atualizado para a amostra de tamanho 2

Basta repetir estes passos para todos os tamanhos de amostra desejado(Gera a amostra, obtém a média e atualiza o Banco). A ultima amostra é a de tamanho 40:

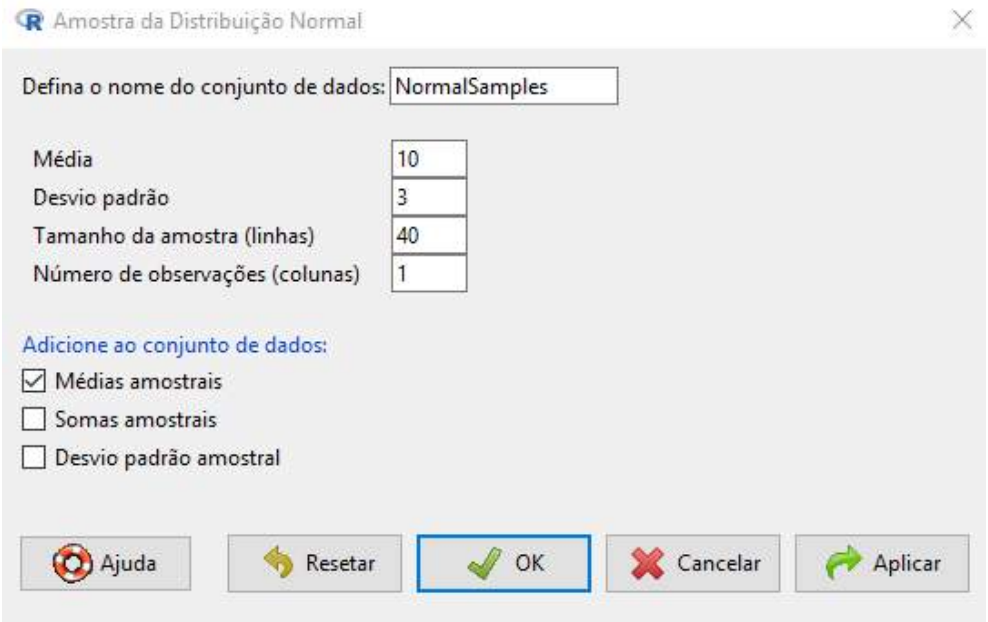


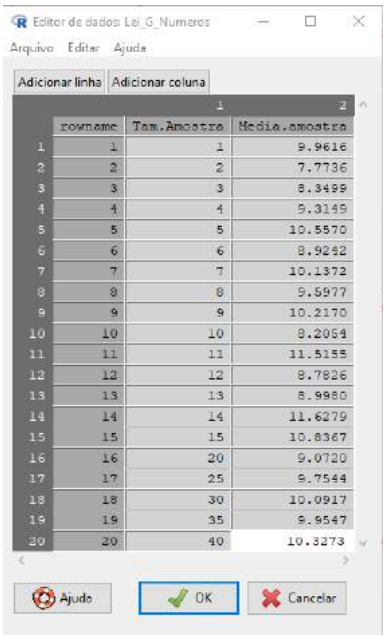
Figura 7.11: Opções para gerar a amostra de tamanho 40

Observamos a média amostral

mean	sd	IQR	0%	25%	50%	75%	100%	n
10.32733	2.976777	3.200977	4.140818	8.613664	10.14866	11.81464	17.80871	40

Figura 7.12: Resumo numérico da amostra de tamanho 40

E obteremos o banco final.



	rowname	Tam.Amostra	Media.amostra
1	1	1	9.9616
2	2	2	7.7736
3	3	3	8.3499
4	4	4	9.3149
5	5	5	10.5570
6	6	6	8.9242
7	7	7	10.1372
8	8	8	9.5977
9	9	9	10.2170
10	10	10	8.2054
11	11	11	11.5155
12	12	12	8.7826
13	13	13	8.9980
14	14	14	11.6279
15	15	15	10.8367
16	16	20	9.0720
17	17	25	9.7544
18	18	30	10.0917
19	19	35	9.9547
20	20	40	10.3273

Figura 7.13: Banco de dados final

Por fim , tendo o banco final vamos fazer uma gráfico de dispersão para observar o comportamento da média amostral de acordo com o tamanho da amostra, onde colocaremos no eixo x o tamanho amostral e no eixo Y a média amostral. Portanto basta ir em : Gráficos> Diagrama de dispersão. Na janela aberta selecionamos as variáveis a serem inseridas em cada eixo

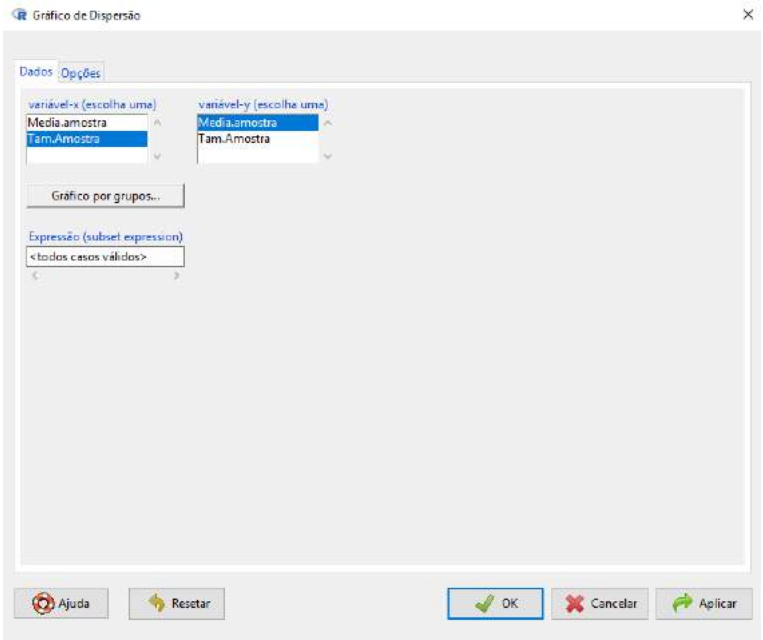


Figura 7.14: Opções para gerar o gráfico

No gráfico gerado vamos lembrar que a média real é 10, portanto, tracei uma linha(Fora do Rcommander) na média igual a 10

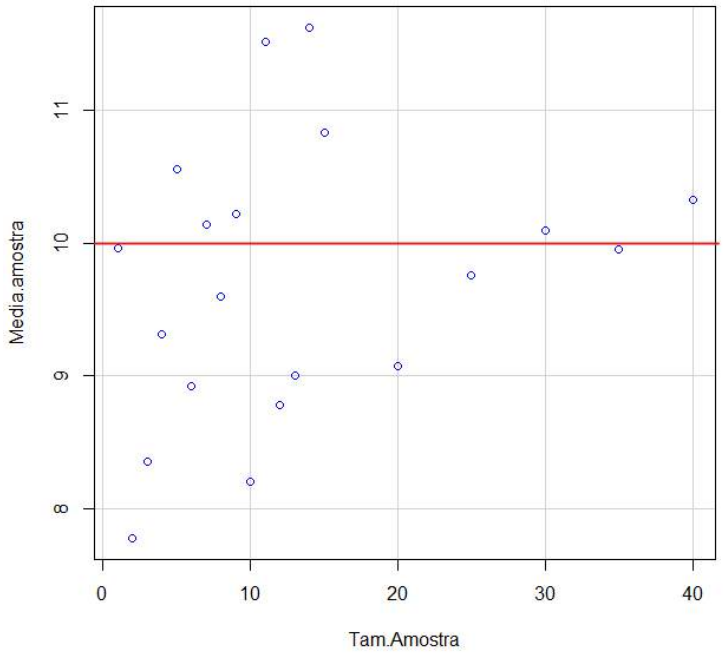


Figura 7.15: Gráfico gerado

Um gráfico de linhas também pode ser útil para enxergar a tendência, a única diferença do gráfico de linhas para o de pontos, é que no gráfico de linhas há uma linha ligando os pontos, Para criar o gráfico de linhas basta ir em : Gráficos> Gráfico de linha. E na janela aberta, selecionar qual variável será adicionada no eixo X e qual será adiciona no eixo Y.

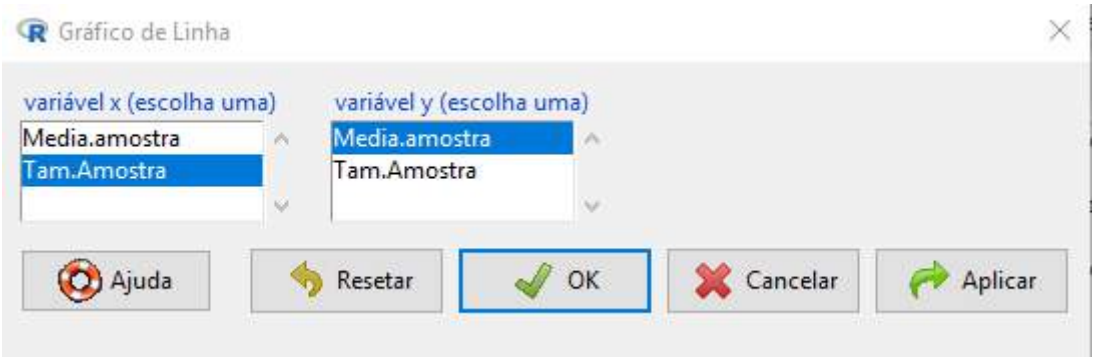


Figura 7.16: Opções para gerar o gráfico de linhas

Apos clicar em ok, o gráfico será gerado, novamente, traçamos uma linha(Fora do Rcommander) no valor da média populacional, para que a comparação fique mais visual:

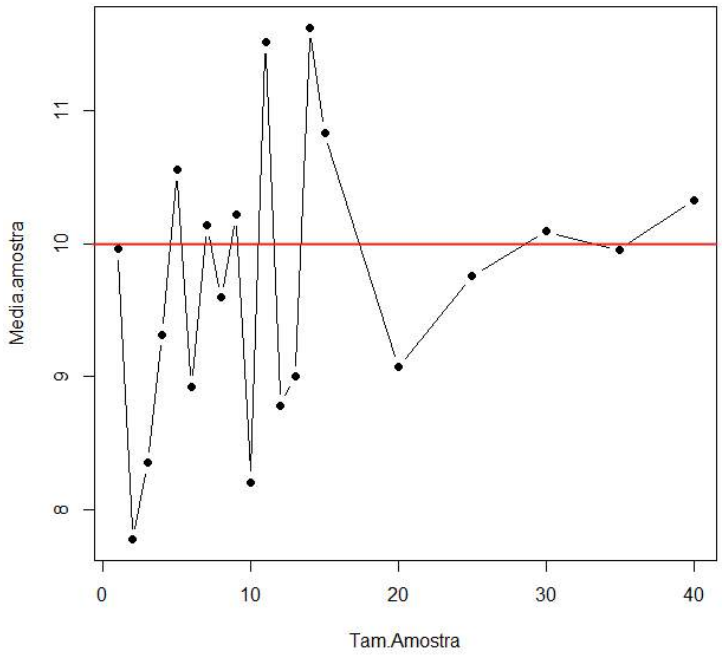


Figura 7.17: Gráfico de linhas

Através dos gráficos, conseguimos observar que nas amostras menores a média é mais dispersa, e a medida que vamos aumentando o tamanho amostral, as médias começam a flutuar bem mais próximo da média populacional. Que é exatamente o que a lei dos grandes números nos diz.

O ideal aqui seria gerar vários tamanhos de amostra, e ir observando a média de todos, no enatando, a limitação do Rcommander, fazendo este processo ser tão manual nos faz reduzir a quantidade de observações, mas note que mesmo com a penas 20 tamanhos de amostras observados já é possível ver a tendência. Sempre que possível utilize a maior quantidade de tamanhos de amostas possíveis nesse tipo de simulação.

Vale lembrar também que ao simular isso em seu computador, os resultados numéricos serão diferentes, pois estamos gerando amostras pseudo-aleatórias. Mas como o intuito é exemplificar a lei dos grandes números, o comportamento esperado será o mesmo, independente da distribuição e da semente usada, a média amostral irá tender a média populacional a medida que os tamanhos de amostra forem aumentando!

7.4 TCL-Teorema Central do Limite

Primeiramente, vamos entender como a ciência opera suas investigações no que diz respeito à coleta de dados. Quando investigamos alguma realidade muitas vezes não temos acesso à essa realidade por completo. Por exemplo, se quisermos saber qual é a média da altura dos estudantes da UFMG é totalmente infactível coletar essa medida de todos os alunos. Nesse sentido, coletamos uma amostra, ou seja, um subconjunto da população. O tamanho da amostra é arbitrário, ou seja, o pesquisador é quem escolhe o tamanho dela. Mas você concorda comigo que a média de uma amostra não necessariamente é a média que procuramos, ou seja, a média populacional? Concorda também que se decidirmos coletar uma amostra de tamanho n (n é um número qualquer) temos várias possibilidades

de diferentes amostras de tamanho n?

Para ser ainda mais claro, imagine que você está conduzindo essa pesquisa da altura média dos estudantes da UFMG. Daí, você e sua equipe decidem escolher alunos, de maneira aleatória (que dê chances de todos participarem), para comporem uma amostra de tamanho igual a 90, por exemplo (n = 90). Nesse sentido, concorda que você tem muitas opções de amostras de tamanho igual a 90 e, com isso, muitas opções de médias amostrais, já que a população de estudantes é bem maior que 90? Pois então é de seu interesse, nesse caso, descobrir qual é a distribuição de probabilidade das médias dessas amostras de tamanho igual a 90 assim como sua média e variância, para que se obtenha uma boa precisão na estimativa da verdadeira média de altura dos alunos da UFMG.

Na prática, não coletamos todas as médias de todas as amostras possíveis de tamanho n para construir a distribuição de probabilidade dessas médias amostrais. Coletamos apenas uma amostra de tamanho igual a n. É aqui que entra o Teorema Central do Limite, pois ele nos informa que quanto maior o n, ou seja, quanto maior for a nossa amostra, mais próximo está a distribuição de probabilidade das médias de amostras de tamanho n de uma distribuição normal com média igual a μ (média populacional – isto é, a média das médias de amostras de tamanho n é igual à média populacional) e variância igual a $\frac{\sigma^2}{n}$ (isto é, a variância das médias de amostras de tamanho n é igual à variância populacional dividido pelo tamanho da amostra).

Por fim, provamos que isso ocorre por meio de simulações computacionais, utilizando o Rcmdr.

Primeiramente, vamos simular diferentes amostras de uma distribuição qualquer. Vamos usar a distribuição exponencial, como exemplo. Começamos gerando 30 amostras de tamanho igual a 10. Não é necessário e nem possível, neste caso, gerar todas as amostras possíveis de tamanho igual a 10. Para fazer isto no Rcmdr basta seguir o seguinte passo:

Nas opções que aparecem na parte de cima da tela, clicar em: **Distribuições » Distribuições contínuas » Distribuição Exponencial » Amostragem da distribuição exponencial**

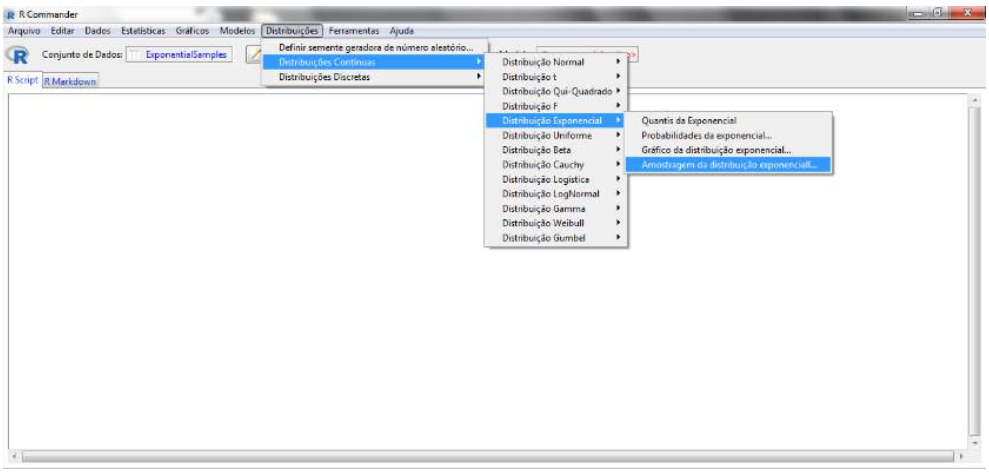


Figura 7.18: Amostragem da distribuição exponencial

Em seguida geramos as 30 amostras de tamanho 10 e colocamos o valor 10 como a taxa da nossa distribuição exponencial. Sendo assim, na primeira opção "Taxa"digitamos o valor 10, na segunda opção "Tamanho da amostra (linhas)"digitamos 30, pois essa opção nos pergunta, na verdade, a quantidade de amostras que queremos e, por fim, a última opção "Número de observações", que é o tamanho das nossas amostras, digitamos 10. Abaixo, selecione

apenas "Médias amostrais", pois o histograma que iremos construir se refere à apenas essa medida. Por fim, clique em "Ok".

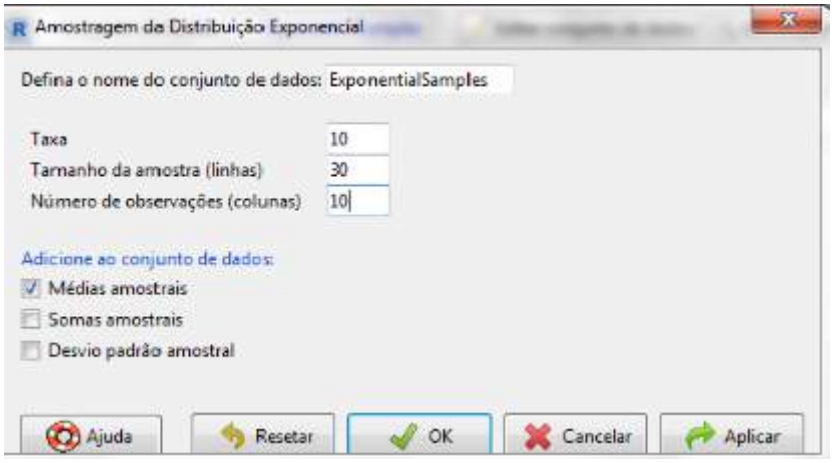


Figura 7.19: Gerando 30 amostras de tamanho igual a 10

Se quiser visualizar os dados, basta clicar em "Ver conjunto de dados"na parte superior da tela. Se não conseguir visualizar dessa forma, clique na opção ao lado:"Editar conjunto de dados". Você verá na tabela dos dados que existe, no final dela, uma variável denominada "mean", que significa "média". Esta última coluna contém a média de cada uma das 30 amostras de tamanho 10. É com essas médias que se deve construir o histograma. Para fazer isto, você clica em "Gráficos"na parte superior da tela e depois em "Histograma".

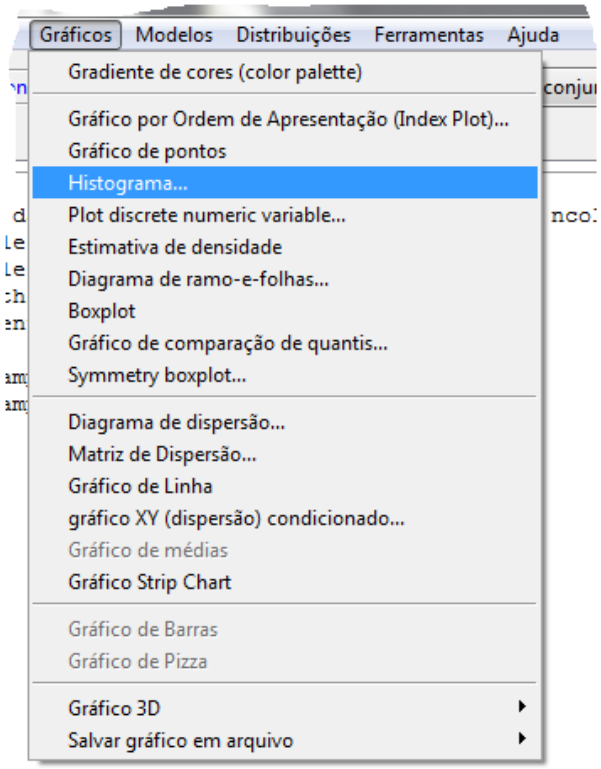


Figura 7.20: Construindo o histograma

Aparecerá uma janela e você deve selecionar "mean", pois é a nossa variável de interesse o qual desejamos plotar o histograma. Depois de selecionar essa variável clique em "Ok". Por fim, aparecerá o histograma plotado da média

das 30 amostras de tamanho 10.

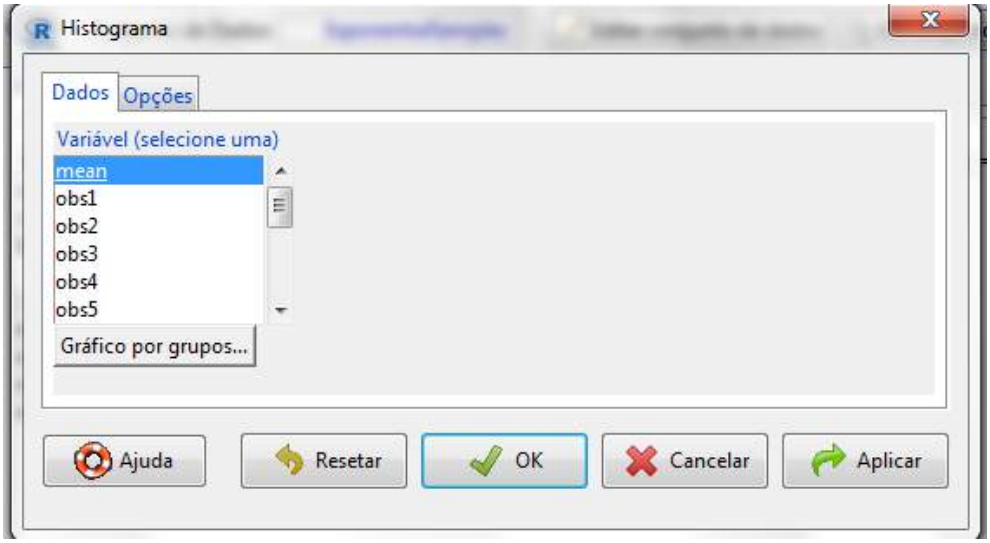


Figura 7.21: Construindo o histograma

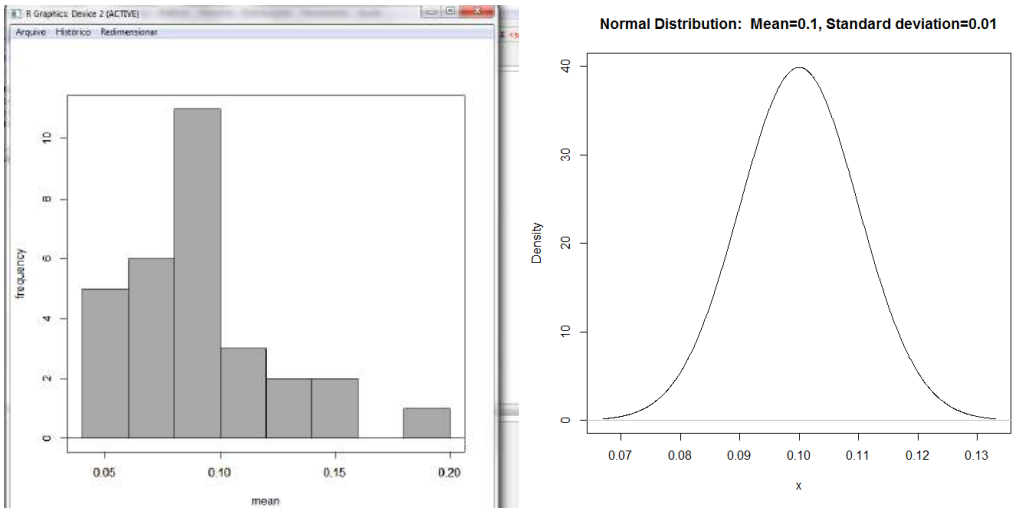


Figura 7.22: Construindo o histograma: n = 10

Feche o gráfico e faça exatamente o mesmo procedimento descrito acima para construir o histograma das médias de outras 30 amostras, mas agora estas devem ter tamanho igual a 30. Você deverá substituir o número 10 por 30 em "Número de observações (colunas)"e deverá definir um outro nome pro conjunto de dados nesta mesma janela em "Defina o nome do conjunto de dados". Para este último, você pode apenas acrescentar um número na frente, por exemplo: "ExponentialSamples2"ou "ExponentialSamples3", etc. Clique em "Ok"e gere o novo histograma.

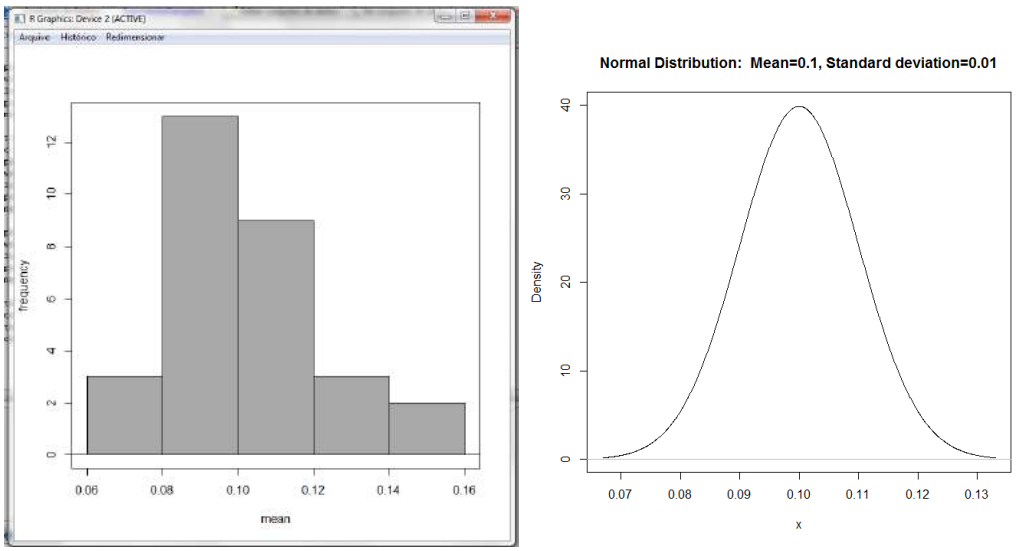


Figura 7.23: Construindo o histograma: n = 30

Repare que à medida que você aumenta o tamanho da amostra a distribuição das médias amostrais, descrita pelo histograma, começa a se parecer cada vez mais com uma distribuição normal com média igual a 0.10, que é a média populacional de qualquer população que segue distribuição exponencial com taxa igual a 10 (definimos esta taxa neste exemplo), pois sua média será $\frac{1}{\lambda} = \frac{1}{10} = 0.10$.

Repare como fica o histograma das médias amostrais quando o tamanho das amostras é igual a 100:

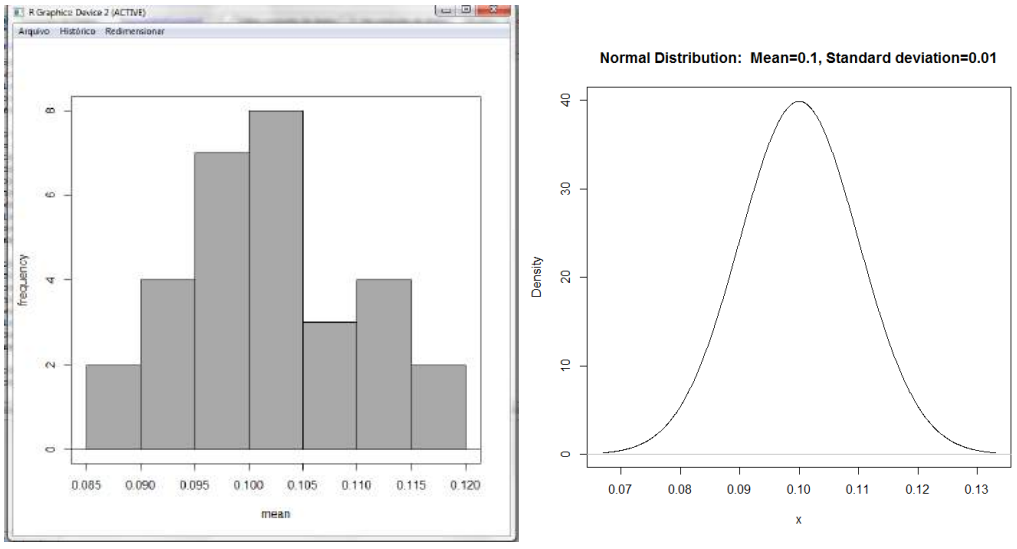


Figura 7.24: Construindo o histograma: n = 100

Quando aumentamos para 150:

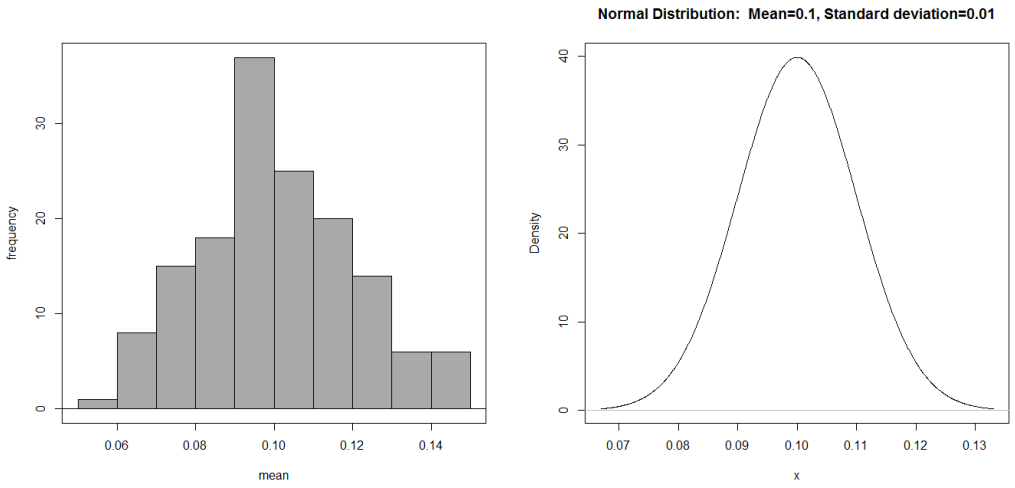


Figura 7.25: Construindo o histograma: n = 150

7.5 QQplot - Comparação de quantis

O gráfico quantil-quantil ou qqplot é utilizado para verificar se um conjunto de dados segue uma determinada distribuição. A ideia básica consiste em calcular o valor teoricamente esperado para cada ponto considerando que a distribuição suposta seja verdadeira. Se os dados de fato seguirem tal distribuição, ao traçarmos os valores observados em um eixo e os esperados em outro, obteremos uma linha reta. Ou seja, se os pontos formarem uma linha reta, ou algo próximo disso, teremos que os dados seguem á distribuição suposta.

O gráfico quantil-quantil mais utilizado é o da distribuição normal, vamos explicar aqui como é feita a criação deste:

- Calcula-se a média e o desvio padrão amostral dos dados;
- Colocamos as obervações da amostra em ordem crescente;
- Para cada valor da amostra (x_i) calculamos uma; aproximação para a proporção de dados menores que x_i,essa aproximação é calculada por f_i tal que:

$$f_i = \begin{cases} \frac{i-(3/8)}{n+(1/4)}, \text{ se } n \leq 10 \\ \frac{i-(1/2)}{n}, \text{ se } n > 10 \end{cases}$$

- Para cada valor f_i, calculamos o quantil da distribuição normal padrão que deixa uma área f_i abaixo dele, ou seja $q(f_i)$;
- Por fim fazemos um diagrama de dispersão com os valores amostrais no eixo y, e $q(f_i)$ no eixo x.

Para exemplificar a criação destes gráficos no Rcommander continuaremos usando o banco de dados do estresse, e iremos tentar compara o nível de estresse(variável estresse do banco) das pessoas com a distribuição normal. Para isso, devemos estar com o conjunto de dados ativo, e depois ir no **menu superior > Gráficos > Gráfico de comparação de quantis..**

Após isso será aberta uma janela com duas abas. Na aba "Dados"escolhemos a variável que desejamos comparar.

Figura 7.26: Aba dados, para criação do gráfico

Após selecionar a variável, iremos para a aba "Opções". É nesta aba que decidiremos com qual distribuição desejamos comparar. Há 4 distribuições já pré-definidas para que possamos escolher, Na distribuição normal, não é preciso definir nenhum parâmetro, na outras é preciso especificar os parâmetros.

Nesta aba também podemos definir o nome dos nossos eixos, e definir se serão mostrados os valores ou só os pontos, podemos deixar que o Rcommander defina automaticamente, que nenhum ponto seja mostrado com o valor, ou que o valor apareça ao clicarmos encima dele.

Para fim de exemplo, vamos apenas selecionar a distribuição normal e deixar o Rcommander identificar os pontos automaticamente.

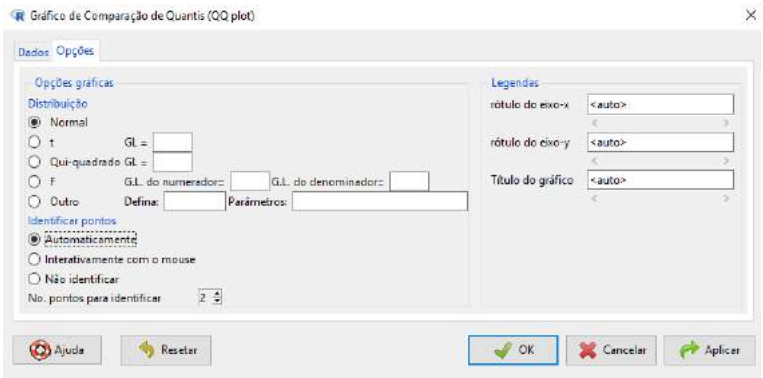


Figura 7.27: Especificações do gráfico

O resultado obtido após clicarmos em "ok"é o gráfico que exibe os pontos e um intervalo de confiança.

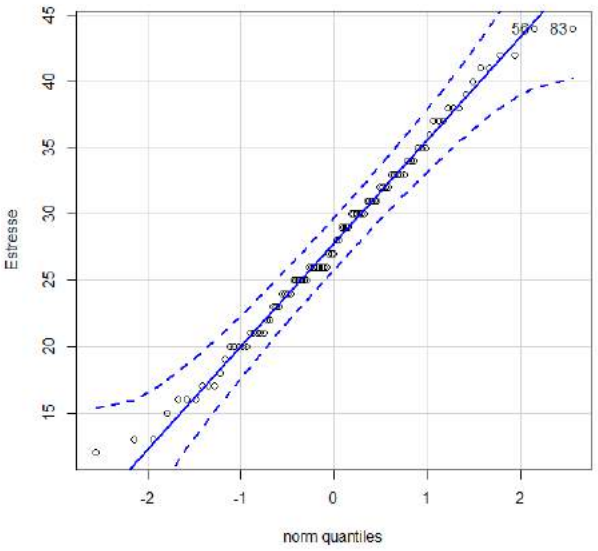


Figura 7.28: Gráfico obtido

Através do gráfico obtido, seria razoável pensar na distribuição normal para esta variável. Mas por se tratar de um recurso visual, sua interpretação é subjetiva, portanto ainda é recomendável fazer um teste de normalidade para tomar decisões.

Para testar as outras distribuições o procedimento é o mesmo, basta selecionar a variável e depois especificar a distribuição desejada.

[*]

Referências

[1] Pedro Barbetta. *Estatıstica Aplicada às Ciências Sociais*. 7ª edição. Editora da UFSC, 2008.

[2] J Fox. "The R commander: a basic-statistics graphical". Em: *User Interface to* 2005 (2005), p. 14.

[3] John Fox et al. "Package 'Rcmdr'". Em: (2020).

[4] João Ismael D. Pinheiro. *Probabilidade e Estatıstica. Quantificando a incerteza*. Ed. por Elsevier. 2012.

[5] Mario F. Triola. *Introdução à Estatıstica*. Ed. por LTC. 11ª ed. 2013.