AnÃ¡lise exploratÃ³ria de dados

# IntroduÃ§Ã£o

Nesta sesÃ£o vamos ver alguns (mas nÃ£o todos!) mÃ©todos para fazer uma anÃ¡lise exploratÃ³ria descritiva de um conjunto de dados.

Uma boa forma de iniciar uma anÃ¡lise descritiva adequada Ã© verificar os tipos de de variÃ¡veis disponÃ­veis.

## VariÃ¡veis

Quando fazemos uma **amostragem**, coletamos nÃ£o apenas a informaÃ§Ã£o sobre a caracterÃ­stica de interesse, mas *diversas outras informaÃ§Ãµes* que auxiliarÃ£o no entendimento desta caracterÃ­stica.

Cada uma das caracterÃ­sticas da populaÃ§Ã£o amostrada, como peso, altura, sexo ou idade, Ã© denominada de uma **variÃ¡vel**.

As *variÃ¡veis* podem assumir diferentes valores, que basicamente podem ser separados em

* **Quantitativos** ou numÃ©ricos.
* **Qualitativos** ou nÃ£o numÃ©ricos, ou *categÃ³ricos*.

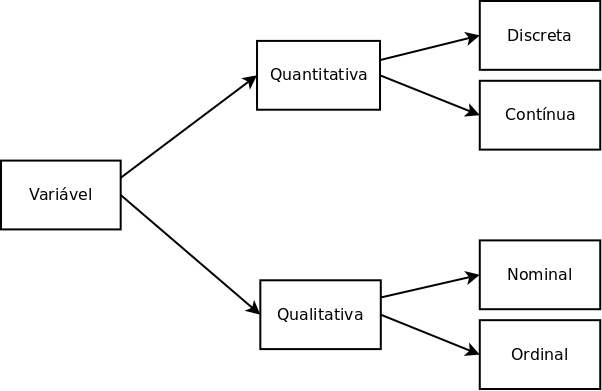
## ClassificaÃ§Ã£o de variÃ¡veis

As *variÃ¡veis quantitativas* ou numÃ©ricas podem ser

* **Discretas**: assumem apenas valores inteiros. Ex.: nÃºmero de irmÃ£os, nÃºmero de passageiros
* **ContÃ­nuas**: assume qualquer valor no intervalo dos nÃºmeros reais. Ex.: peso, altura

As *variÃ¡veis qualitativas* ou categÃ³ricas podem ser

* **Nominais**: quando as categorias nÃ£o possuem uma ordem natural. Ex.: nomes, cores, sexo.
* **Ordinais**: quando as categorias podem ser ordenadas. Ex.: tamanho (pequeno, mÃ©dio, grande), classe social (baixa, mÃ©dia, alta), grau de instruÃ§Ã£o (bÃ¡sico, mÃ©dio, graduaÃ§Ã£o, pÃ³s-graduaÃ§Ã£o)



# ApresentaÃ§Ã£o e organizaÃ§Ã£o de dados

A **organizaÃ§Ã£o** dos dados coletados Ã© fundamental para que nÃ£o hajam erros de processamento e perda de informaÃ§Ãµes. Deve ser feito em um programa apropriado. Exemplo: planilhas eletrÃ´nicas e bancos de dados.

A **apresentaÃ§Ã£o** dos dados depende do tipo de variÃ¡vel e daquilo que se quer mostrar.

**Tabelas** e **grÃ¡ficos** podem mostrar a mesma informaÃ§Ã£o, mas alguns sÃ£o mais apropriados dependendo do objetivo.

Existe um nÃºmero considerÃ¡vel de **pacotes estatÃ­sticos**, alguns especÃ­ficos para algumas Ã¡reas e outros mais gerais. Qualquer que seja o programa a ser utilizado, existem trÃªs etapas que envolvem seu uso:

1. Entrada de dados.
2. ExecuÃ§Ã£o da anÃ¡lise estatÃ­stica.
3. InterpretaÃ§Ã£o de resultados.

A entrada de dados deve assumir certas convenÃ§Ãµes:

* Os dados devem estar no formato de **matriz**
* Cada linha da matriz corresponde a uma **unidade experimental**
  + Elemento da populaÃ§Ã£o ou amostra no qual observamos as variÃ¡veis
* Cada coluna da matriz corresponde a uma **variÃ¡vel**

# Dados brutos

Quando fazemos uma coleta de dados, e armazenamos de forma correta, temos em mÃ£os o que se chama de **dados brutos**, pois consiste das observaÃ§Ãµes “puras”, sem nenhum tipo de processamento ou resumo.

Uma anÃ¡lise de dados sempre deve comeÃ§ar com uma planilha de dados brutos, pois serÃ¡ a partir deles que iremos resumir e visualizar as informaÃ§Ãµes de interesse.

## O conjunto de dados milsa

O livro “EstatÃ­stica BÃ¡sica” de W. O. Bussab e P. A. Morettin traz no segundo capÃ­tulo um conjunto de dados hipotÃ©tico de atributos de 36 funcionÃ¡rios da companhia “Milsa”. Os dados estÃ£o reproduzidos na tabela abaixo.

## Importando os dados para o R  
milsa <- read.csv("dados/milsa.csv")

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Funcionario | Est.civil | Inst | Filhos | Salario | Anos | Meses | Regiao |
| 1 | solteiro | 1o Grau | NA | 4.00 | 26 | 3 | interior |
| 2 | casado | 1o Grau | 1 | 4.56 | 32 | 10 | capital |
| 3 | casado | 1o Grau | 2 | 5.25 | 36 | 5 | capital |
| 4 | solteiro | 2o Grau | NA | 5.73 | 20 | 10 | outro |
| 5 | solteiro | 1o Grau | NA | 6.26 | 40 | 7 | outro |
| 6 | casado | 1o Grau | 0 | 6.66 | 28 | 0 | interior |
| 7 | solteiro | 1o Grau | NA | 6.86 | 41 | 0 | interior |
| 8 | solteiro | 1o Grau | NA | 7.39 | 43 | 4 | capital |
| 9 | casado | 2o Grau | 1 | 7.59 | 34 | 10 | capital |
| 10 | solteiro | 2o Grau | NA | 7.44 | 23 | 6 | outro |
| 11 | casado | 2o Grau | 2 | 8.12 | 33 | 6 | interior |
| 12 | solteiro | 1o Grau | NA | 8.46 | 27 | 11 | capital |
| 13 | solteiro | 2o Grau | NA | 8.74 | 37 | 5 | outro |
| 14 | casado | 1o Grau | 3 | 8.95 | 44 | 2 | outro |
| 15 | casado | 2o Grau | 0 | 9.13 | 30 | 5 | interior |
| 16 | solteiro | 2o Grau | NA | 9.35 | 38 | 8 | outro |
| 17 | casado | 2o Grau | 1 | 9.77 | 31 | 7 | capital |
| 18 | casado | 1o Grau | 2 | 9.80 | 39 | 7 | outro |
| 19 | solteiro | Superior | NA | 10.53 | 25 | 8 | interior |
| 20 | solteiro | 2o Grau | NA | 10.76 | 37 | 4 | interior |
| 21 | casado | 2o Grau | 1 | 11.06 | 30 | 9 | outro |
| 22 | solteiro | 2o Grau | NA | 11.59 | 34 | 2 | capital |
| 23 | solteiro | 1o Grau | NA | 12.00 | 41 | 0 | outro |
| 24 | casado | Superior | 0 | 12.79 | 26 | 1 | outro |
| 25 | casado | 2o Grau | 2 | 13.23 | 32 | 5 | interior |
| 26 | casado | 2o Grau | 2 | 13.60 | 35 | 0 | outro |
| 27 | solteiro | 1o Grau | NA | 13.85 | 46 | 7 | outro |
| 28 | casado | 2o Grau | 0 | 14.69 | 29 | 8 | interior |
| 29 | casado | 2o Grau | 5 | 14.71 | 40 | 6 | interior |
| 30 | casado | 2o Grau | 2 | 15.99 | 35 | 10 | capital |
| 31 | solteiro | Superior | NA | 16.22 | 31 | 5 | outro |
| 32 | casado | 2o Grau | 1 | 16.61 | 36 | 4 | interior |
| 33 | casado | Superior | 3 | 17.26 | 43 | 7 | capital |
| 34 | solteiro | Superior | NA | 18.75 | 33 | 7 | capital |
| 35 | casado | 2o Grau | 2 | 19.40 | 48 | 11 | capital |
| 36 | casado | Superior | 3 | 23.30 | 42 | 2 | interior |

Estes dados estÃ£o disponÃ­veis em um arquivo csv no endereÃ§o <http://www.leg.ufpr.br/~fernandomayer/dados/milsa.csv>.

ObservaÃ§Ã£o

Um arquivo csv Ã© um arquivo de texto (assim como um txt), mas as colunas de um conjunto de dados sÃ£o separados por vÃ­rgula (*csv* = *comma separated values*). Arquivos de texto como esse podem ser abertos em qualquer editor de texto, e tambÃ©m em planilhas eletrÃ´nicas.

O nosso objetivo agora Ã© entÃ£o fazer uma anÃ¡lise descritiva destes dados, com a intenÃ§Ã£o de se identificar padrÃµes e relaÃ§Ãµes (se houverem).

Portanto, o primeiro passo Ã© classificar todas as variÃ¡veis desse conjunto de dados:

|  |  |
| --- | --- |
| VariÃ¡vel | ClassificaÃ§Ã£o |
| Funcionario | Quantitativa discreta |
| Est.civil | Qualitativa nominal |
| Inst | Qualitativa ordinal |
| Filhos | Quantitativa discreta |
| Salario | Quantitativa contÃ­nua |
| Anos | Quantitativa discreta |
| Meses | Quantitativa discreta |
| Regiao | Qualitativa nominal |

Naturalmente as variÃ¡veis Anos e Meses, por se tratarem de tempo, sÃ£o contÃ­nuas. No entanto, da maneira como foram coletadas, nÃ³s podemos classificÃ¡-las como discretas, sem muita perda de informaÃ§Ã£o. Esse processo Ã© denominado de **discretizaÃ§Ã£o** de uma variÃ¡vel contÃ­nua, e pode ser utilzado em pelo menos dois casos:

1. Quando nÃ£o hÃ¡ perda de informaÃ§Ã£o ao discretizar a variÃ¡vel (como nesse caso), ou quando o detalhamento nÃ£o Ã© fundamental para o entendimento da caracterÃ­stica.
2. Quando uma variÃ¡vel contÃ­nua Ã© medida por um equipamento que nÃ£o Ã© capaz de mensurar com uma determinada precisÃ£o. Por exemplo, se o peso de uma pessoa for medido por uma balanÃ§a que sÃ³ mede em escala de 1 em 1 kg, o peso seria considerado como variÃ¡vel discreta mesmo que saibamos que de fato Ã© contÃ­nua.

Estes sÃ£o dados no “estilo planilha”, com variÃ¡veis de diferentes tipos: categÃ³ricas e numÃ©ricas (qualitativas e quantitativas). Portanto segue as normas que estabelecemos anteriormente:

* Cada linha corresponde a uma unidade amostral, nesse caso Ã  uma pessoa
* Cada coluna corresponde a uma variÃ¡vel diferente

NOTE algumas coisas importantes:

* Cada variÃ¡vel possui um tipo de resposta, e estas respostas seguem sempre o mesmo padrÃ£o. Por exemplo, na variÃ¡vel Est.civil as possÃ­veis respostas sÃ£o: solteiro e casado. Se por acaso houvesse o termo Casado, poderia haver confusÃ£o na maioria dos pacotes estatÃ­sticos, e este termo poderia ser (erroneamente) uma terceira categoria de resposta.
* Quando a informaÃ§Ã£o de uma (ou mais) variÃ¡veis nÃ£o estiver disponÃ­vel (por exemplo: falta de resposta, rasura na marcaÃ§Ã£o da resposta, ou simplesmente nÃ£o foi possÃ­vel coletar aquela informaÃ§Ã£o para aquele indivÃ­duo), esta informaÃ§Ã£o deve ser marcada (ou indicada) com algum sÃ­mbolo especial. Em planilhas eletrÃ´nicas, normalmente a cÃ©lula fica em branco, mas na maioria dos pacotes estatÃ­sticos Ã© necessÃ¡rio preencher estas cÃ©lulas com um sÃ­mbolo (que varia entre os pacotes). Aqui, por exemplo, usamos o sÃ­mbolo NA (de Not Available) para indicar estas observaÃ§Ãµes.

Importante!

NÃ£o preencha casos perdidos com 0 (zero), pois o zero nÃ£o significa falta de informaÃ§Ã£o, e sim uma quantidade que foi medida, mas foi verificada como nula.

Agora que os dados estÃ£o prontos podemos comeÃ§ar a anÃ¡lise descritiva. A seguir mostramos como fazer anÃ¡lises descritivas uni e bi-variadas. Sugiro ainda que vocÃª use algum programa ou pacote estatÃ­stico para reproduzir os resultados mostrados aqui.

# AnÃ¡lise univariada

A anÃ¡lise univariada consiste basicamente em, para cada uma das variÃ¡veis individualmente:

* Classificar a variÃ¡vel quanto a seu tipo: qualitativa (nominal ou ordinal) ou quantitativa (discreta ou contÃ­nua)
* Obter tabelas, grÃ¡ficos e/ou medidas que resumam a variÃ¡vel

A partir destes resultados pode-se montar um resumo geral dos dados. Quando se estuda uma variÃ¡vel, o maior intersse do pesquisador Ã© conhecer o *comportamento* dessa variÃ¡vel, analisando a ocorrÃªncia de suas possÃ­veis realizaÃ§Ãµes. Nesse sentido, as **distribuiÃ§Ãµes de frequÃªncia** serÃ£o o principal recurso que utilizaremos para resumir uma Ãºnica variÃ¡vel.

A seguir vamos mostrar como obter tabelas e grÃ¡ficos simples. Para isto vamos selecionar uma variÃ¡vel de cada tipo para que o leitor possa, por analogia, obter resultados para as demais.

## VariÃ¡vel Qualitativa Nominal

A variÃ¡vel Est.civil Ã© uma qualitativa nominal. Desta forma podemos obter:

1. Uma tabela de frequÃªncias (absolutas e/ou relativas)
2. Um grÃ¡fico de barras ou de setores
3. A “moda”, *i.e.* o valor que ocorre com maior frequÃªncia

Por ser uma variÃ¡vel qualitativa, para obter a distribuiÃ§Ã£o de frequÃªncia desta variÃ¡vel, basta contarmos quantas vezes ocorre cada categoria (ou nÃ­vel), e organizar em uma tabela.

civil.tb <- table(milsa$Est.civil)  
cbind("f" = civil.tb)

f  
casado 20  
solteiro 16

Esta simples contagem Ã© chamada de **frequÃªncia absoluta**. Podemos tambÃ©m incluir nessa tabela a soma total de observaÃ§Ãµes de todas as categorias.

cbind("f" = addmargins(civil.tb))

f  
casado 20  
solteiro 16  
Sum 36

Com essa informaÃ§Ã£o adicional, podemos agora calcular a **frequÃªncia relativa**, ou seja, a frequÃªncia absoluta de cada categoria, dividida pelo total.

cbind("fr" = prop.table(civil.tb))

fr  
casado 0.5555556  
solteiro 0.4444444

Note que, nesse caso, a soma das categorias deve somar 1 (ou 100%)

cbind("fr" = addmargins(prop.table(civil.tb)))

fr  
casado 0.5555556  
solteiro 0.4444444  
Sum 1.0000000

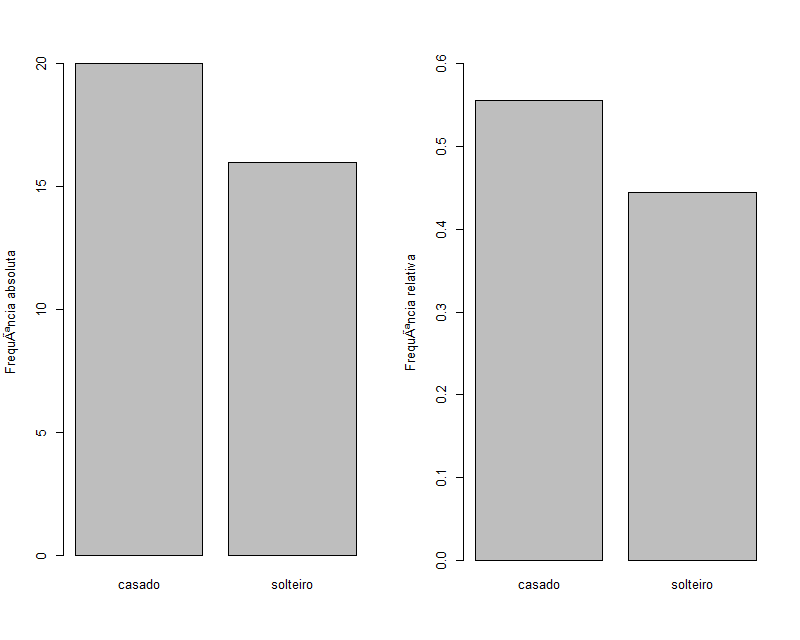
Com isso, podemos definir alguns tipos de frequÃªncia.

Tipos de frequÃªncia

* FrequÃªncia absoluta (): nÃºmero total de elementos em cada classe
* FrequÃªncia relativa (): razÃ£o entre cada valor da frequÃªncia absoluta e o total de observaÃ§Ãµes
* FrequÃªncia percentual (): frequÃªncia relativa em porcentagem

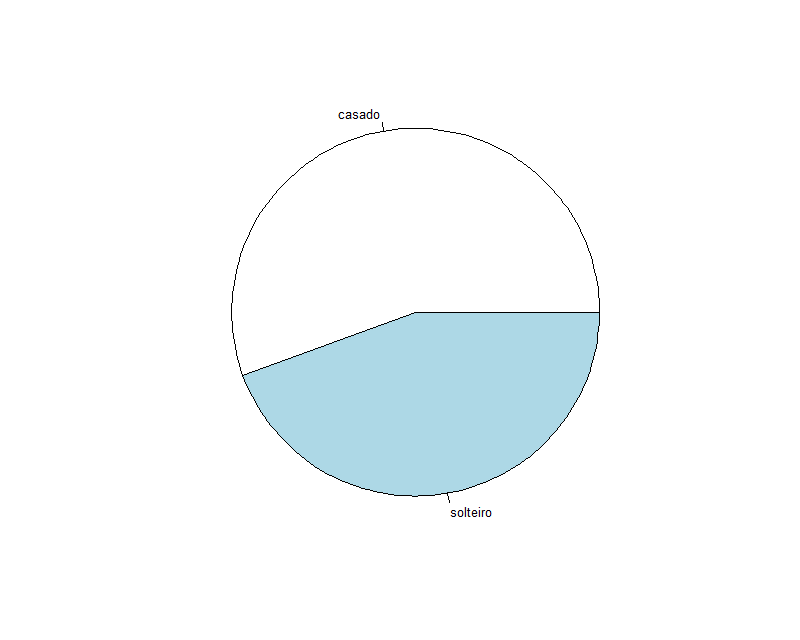
Os grÃ¡ficos de barras e de setores sÃ£o adequados para representar esta variÃ¡vel. O **grÃ¡fico de barras** Ã© formado pelas categorias no eixo X, e pela frequÃªncia no eixo Y. A frequÃªncia utilizada pode ser tanto a absoluta quanto a relativa, conforme for o caso.

par(mfrow = c(1, 2))  
barplot(civil.tb, ylab = "FrequÃªncia absoluta")  
barplot(prop.table(civil.tb), ylab = "FrequÃªncia relativa",  
 ylim = c(0, .6))  
par(mfrow = c(1,1))



O **grÃ¡fico de setores** (ou de pizza, ou torta, ou diagrama circular) tambÃ©m pode ser utilizado, mas apresenta uma maior limitaÃ§Ã£o. Independente da frequÃªncia utilizada, cada setor terÃ¡ a mesma Ã¡rea. AlÃ©m disso, quando existem muitas categorias, e/ou as categorias possuem frequÃªncias semelhantes, a diferenciaÃ§Ã£o dos setores Ã© dificultada.

pie(civil.tb)



A **moda** de qualquer variÃ¡vel Ã© definida como o valor mais frequente encontrado na amostra. No caso de variÃ¡veis qualitativas, a moda Ã© a categoria que apresenta maior frequÃªncia. Nesse exemplo, a moda seria entÃ£o

names(civil.tb)[which.max(civil.tb)]

[1] "casado"

## VariÃ¡vel Qualitativa Ordinal

Para exemplificar como obter anÃ¡lises para uma variÃ¡vel qualitativa ordinal vamos selecionar a variÃ¡vel Inst, que verificou o grau de instruÃ§Ã£o dos funcionÃ¡rios.

As tabelas de frequÃªncias sÃ£o obtidas de forma semelhante Ã  mostrada anteriormente. A frequÃªncia absoluta Ã© a contagem do nÃºmero de vezes que cada categoria foi observada. Note que aqui, a ordem tem importÃ¢ncia, portanto, a tabela tambÃ©m deve seguir a ordem natural das categorias. Abaixo, mostramos a tabela de frequÃªncia absoluta jÃ¡ com o somatÃ³rio de todas as classes.

inst.tb <- table(milsa$Inst)  
cbind("f" = addmargins(inst.tb))

f  
1o Grau 12  
2o Grau 18  
Superior 6  
Sum 36

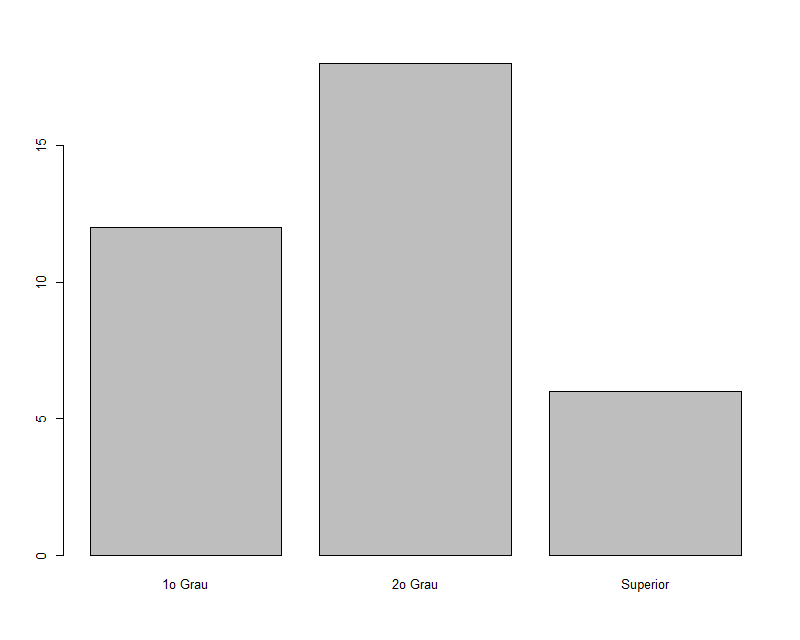
As frequÃªncias relativas tambÃ©m sÃ£o obtidas atravÃ©s da divisÃ£o da frequÃªncia absoluta de cada classe pelo total, ou seja,

cbind("f" = addmargins(inst.tb),  
 "fr" = addmargins(prop.table(inst.tb)))

f fr  
1o Grau 12 0.3333333  
2o Grau 18 0.5000000  
Superior 6 0.1666667  
Sum 36 1.0000000

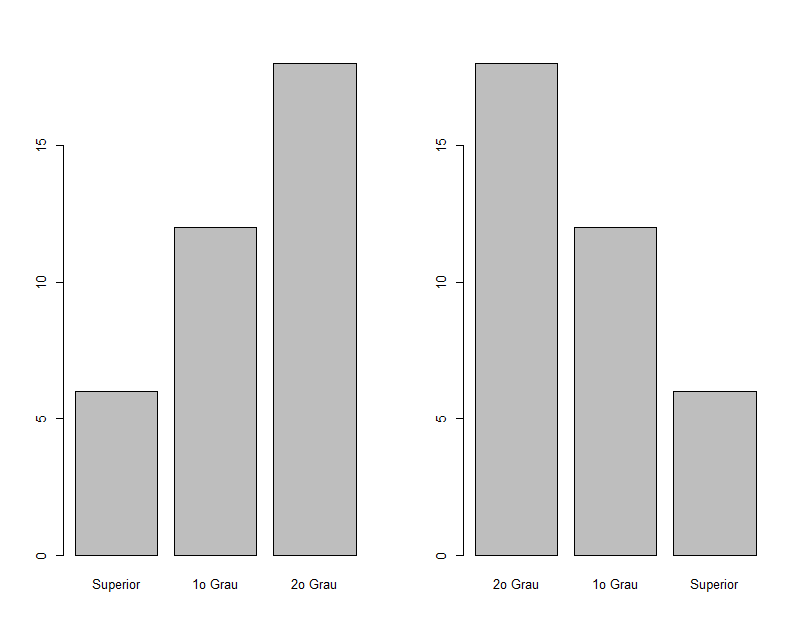
O grÃ¡fico de setores nÃ£o Ã© adequado para este tipo de variÃ¡vel por nÃ£o expressar a ordem dos possÃ­veis valores. Usamos entÃ£o apenas um grÃ¡fico de barras conforme mostrado abaixo

barplot(inst.tb)



Em alguns casos podemos querer mostrar o grÃ¡fico de barras com as barras classificadas da menor para a maior, ou vice-versa, independente da ordem dos nÃ­veis. O importante Ã© sempre deixar claro as categorias de cada barra.

par(mfrow = c(1,2))  
## Menor para maior  
barplot(sort(inst.tb))  
## Maior para menor  
barplot(sort(inst.tb, decreasing = TRUE))  
par(mfrow = c(1,1))



Para uma variÃ¡vel ordinal, a moda tambÃ©m Ã© especificada como a categoria de maior frequÃªncia, ou seja,

names(inst.tb)[which.max(inst.tb)]

[1] "2o Grau"

## VariÃ¡vel quantitativa discreta

Vamos agora usar a variÃ¡vel Filhos (nÃºmero de filhos) para ilustrar algumas anÃ¡lises que podem ser feitas com uma quantitativa discreta.

FrequÃªncias absolutas e relativas sÃ£o obtidas como anteriormente. Nesse caso, assumimos que cada valor numÃ©rico Ã© uma categoria, e construÃ­mos as tabelas de frequÃªncia como se a variÃ¡vel fosse qualitativa ordinal. Note, no entanto, que quando existem poucos valores numÃ©ricos, essa abordagem Ã© viÃ¡vel. Mas contagens podem assumir muitos valores diferentes, e nesses casos, fazer uma tabela de frequÃªncia pode nÃ£o ajudar a resumir aquela variÃ¡vel. Quando esse for o caso, adotaremos a mesma tÃ©cnica que usamos para resumir variÃ¡veis quantitativas contÃ­nuas, como veremos na prÃ³xima sessÃ£o.

Abaixo, temos a frequÃªncia absoluta, com o total de observaÃ§Ãµes

filhos.tb <- table(milsa$Filhos)  
cbind("f" = addmargins(filhos.tb))

f  
0 4  
1 5  
2 7  
3 3  
5 1  
Sum 20

Note que a soma foi 20, ao invÃ©s das 36 observaÃ§Ãµes totais da planilha original. Como vocÃª deve imaginar ao ter inspecionado a tabela, isso ocorre pelo fato de que existem algumas observaÃ§Ãµes *perdidas* para essa variÃ¡vel. Mais especificamente, existem 16 observaÃ§Ãµes faltantes, aquelas marcadas com NA. Se for desejÃ¡vel, pode-se incluir a contagem de observaÃ§Ãµes faltantes na tabela de frequÃªncia.

cbind("f" = addmargins(table(milsa$Filhos, useNA = "always")))

f  
0 4  
1 5  
2 7  
3 3  
5 1  
<NA> 16  
Sum 36

Por ora, vamos usar a tabela de frequÃªncia sem a contagem destes valores perdidos. Nestes casos, Ã© importante ter cuidado com as frequÃªncias relativas, que devem ser calculadas com o total da tabela, e nÃ£o com o total de observaÃ§Ãµes. Portanto, para esse exemplo, as frequÃªncias relativas seriam

cbind("f" = addmargins(filhos.tb),  
 "fr" = addmargins(prop.table(filhos.tb)))

f fr  
0 4 0.20  
1 5 0.25  
2 7 0.35  
3 3 0.15  
5 1 0.05  
Sum 20 1.00

Para variÃ¡veis cujos valores possuem ordenaÃ§Ã£o natural (qualitativas ordinais e quantitativas em geral), faz sentido calcularmos tambÃ©m as **frequÃªncias acumuladas**. A frequÃªncia acumulada atÃ© um certo valor Ã© obtida pela soma das frequÃªncias de todos os valores da variÃ¡vel, *menores ou iguais ao valor considerado*.

filhos.tba <- as.table(cumsum(filhos.tb))  
cbind("f" = addmargins(filhos.tb),  
 "fr" = addmargins(prop.table(filhos.tb)),  
 "F" = c(filhos.tba, NA))

f fr F  
0 4 0.20 4  
1 5 0.25 9  
2 7 0.35 16  
3 3 0.15 19  
5 1 0.05 20  
Sum 20 1.00 NA

TambÃ©m podemos definir a **frequÃªncia acumulada relativa**, que Ã© o resultado da divisÃ£o das frequÃªncias acumuladas pelo total de observaÃ§Ãµes, ou seja

filhos.tba <- as.table(cumsum(filhos.tb))  
cbind("f" = addmargins(filhos.tb),  
 "fr" = addmargins(prop.table(filhos.tb)),  
 "F" = c(filhos.tba, NA),  
 "Fr" = c(filhos.tba/20, NA))

f fr F Fr  
0 4 0.20 4 0.20  
1 5 0.25 9 0.45  
2 7 0.35 16 0.80  
3 3 0.15 19 0.95  
5 1 0.05 20 1.00  
Sum 20 1.00 NA NA

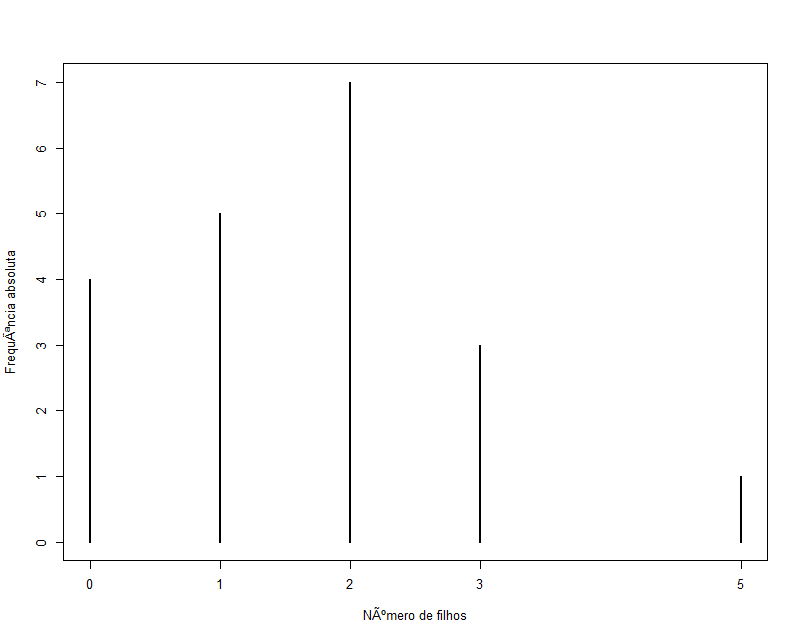
Com isso, definimos tambÃ©m as frequÃªncias acumuladas.

Tipos de frequÃªncia acumulada

* FrequÃªncia acumulada (): soma das frequÃªncias simples de uma classe com a frequÃªncia simples da classe anterior
* FrequÃªncia acumulada relativa (): frequÃªncia acumulada da classe dividida pelo total de observaÃ§Ãµes
* FrequÃªncia acumulada percentual (): frequÃªncia acumulada relativa em porcentagem

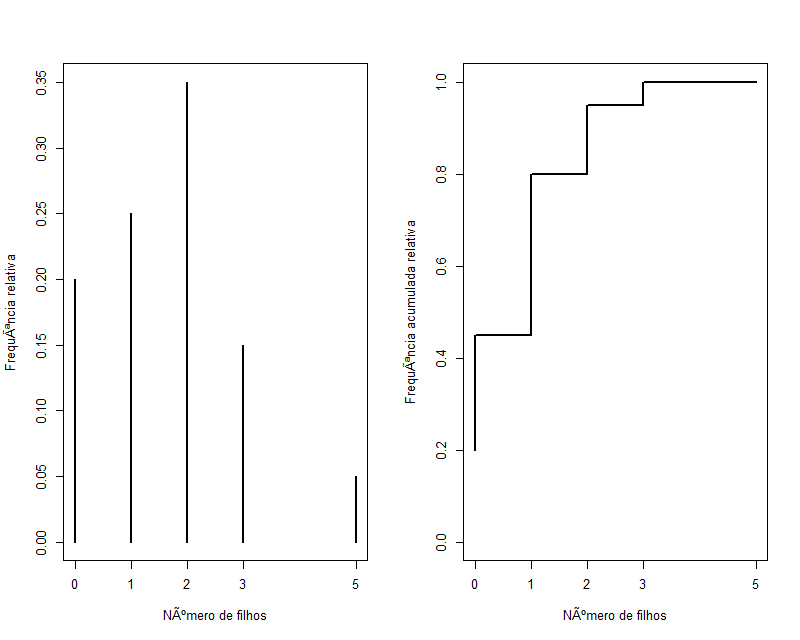
O grÃ¡fico adequado para frequÃªncias absolutas, relativas, ou acumuladas de uma variÃ¡vel discreta Ã© parecido com um grÃ¡fico de barras, mas nesse caso, as frequÃªncias sÃ£o indicadas por linhas.

plot(filhos.tb, xlab = "NÃºmero de filhos",  
 ylab = "FrequÃªncia absoluta")



Outra possibilidade seria fazer grÃ¡ficos de frequÃªncias relativa e de frequÃªncias relativas acumuladas conforme mostrado nas figuras abaixo.

par(mfrow = c(1,2))  
## FrequÃªncia relativa  
plot(prop.table(filhos.tb), xlab = "NÃºmero de filhos",  
 ylab = "FrequÃªncia relativa")  
## FrequÃªncia relativa acumulada  
plot(filhos.tba/20, type = "S", # tipo step (escada)  
 xlab = "NÃºmero de filhos",  
 ylab = "FrequÃªncia acumulada relativa")  
par(mfrow = c(1,1))



## VariÃ¡vel quantitativa contÃ­nua

Para concluir os exemplos para anÃ¡lise univariada vamos considerar a variÃ¡vel quantitativa contÃ­nua Salario.

Para se fazer uma tabela de frequÃªncias de uma variÃ¡vel contÃ­nua, Ã© preciso primeiro agrupar os dados em classes prÃ©-estabelecidas. A escolha dos **intervalos de classe** geralmente Ã© arbitrÃ¡ria, e a familiaridade do pesquisador com os dados Ã© que lhe indicarÃ¡ quantas e quais classes devem ser utilizadas. No entanto, algumas regras simples podem ser Ãºteis para auxiliar na construÃ§Ã£o das classes. Uma forma seria calcular a **amplitude de classe** () atravÃ©s da relaÃ§Ã£o

onde Ã© a **amplitude total** dos dados, e Ã© um nÃºmero *estimado* de intervalos de classes para um conjunto de dados com observaÃ§Ãµes ( pode ser calculado por outras definiÃ§Ãµes, como a regra de Sturges, por exemplo).

Note que, para facilitar o cÃ¡lculo da amplitude de classe e posteriormente para montar a tabela de distribuiÃ§Ã£o de frequÃªncia, podemos **ordenar** o vetor de dados brutos. Por exemplo, os valores ordenados (em ordem crescente) de Salario sÃ£o

sort(milsa$Salario)

[1] 4.00 4.56 5.25 5.73 6.26 6.66 6.86 7.39 7.44 7.59 8.12  
[12] 8.46 8.74 8.95 9.13 9.35 9.77 9.80 10.53 10.76 11.06 11.59  
[23] 12.00 12.79 13.23 13.60 13.85 14.69 14.71 15.99 16.22 16.61 17.26  
[34] 18.75 19.40 23.30

Com isso, naturalmente jÃ¡ visualizamos os valores mÃ¡ximo e mÃ­nimo, e sabemos que a amplitude total Ã©

Como temos 36 observaÃ§Ãµes, o nÃºmero estimado de classes Ã© . Portanto, a amplitude de classe Ã©

O valor de amplitude de classe pode ser arredondado para um nÃºmero inteiro, geralmente para facilitar a interpretaÃ§Ã£o da tabela. Nesse caso, poderÃ­amos arredondar a amplitude para 3 ou para 4 (inteiros mais prÃ³ximos). Como o valor mÃ­nimo Ã© 4 e o mÃ¡ximo estÃ¡ prÃ³ximo de 24, Ã© natural arredondarmos a amplitude para 4. ApÃ³s esse arredondamento, vamos denominar a amplitude de classe definitiva por , apenas para diferenciar do valor calculado em .

Dessa forma, as classes seriam formadas pela sequÃªncia, ou “quebra de classes” definida por

(quebra <- seq(4, 24, 4))

[1] 4 8 12 16 20 24

Portanto, ficamos com um total de 5 classes. A Ãºltima etapa Ã© definir os extremos de classe que serÃ£o abertos (permitem incluir aquele valor exato) ou fechados (nÃ£o incluem aquele valor). Para isso, usaremos a notaÃ§Ã£o a seguir

|  |  |  |  |
| --- | --- | --- | --- |
| Classe | NotaÃ§Ã£o | DenominaÃ§Ã£o | Resultado |
| [a,b) |  | Fechado em a, aberto em b | Inclui a, nÃ£o inclui b |
| (a,b] |  | Aberto em a, fechado em b | NÃ£o inclui a, inclui b |

Nesse exemplo, vamos usar intervalos do tipo fechados Ã  esquerda (abertos Ã  direita), pois o primeiro valor de quebra de classe Ã© o igual ao valor mÃ­nimo do conjunto de dados, e dessa forma garantimos que todos os valores pertenÃ§am a alguma das classes.

Portanto, as classes completas com suas respectivas frequÃªncias absolutas estÃ£o na tabela abaixo.

classes <- cut(milsa$Salario, breaks = quebra, right = FALSE)  
classes.tab <- table(classes)  
cbind("f" = addmargins(classes.tab))

f  
[4,8) 10  
[8,12) 12  
[12,16) 8  
[16,20) 5  
[20,24) 1  
Sum 36

Para completar esta tabela, podemos obter os demais tipos de frequÃªncia, conforme visto anteriormente: (frequÃªncia relativa), (frequÃªncia acumulada), e (frequÃªncia acumulada relativa).

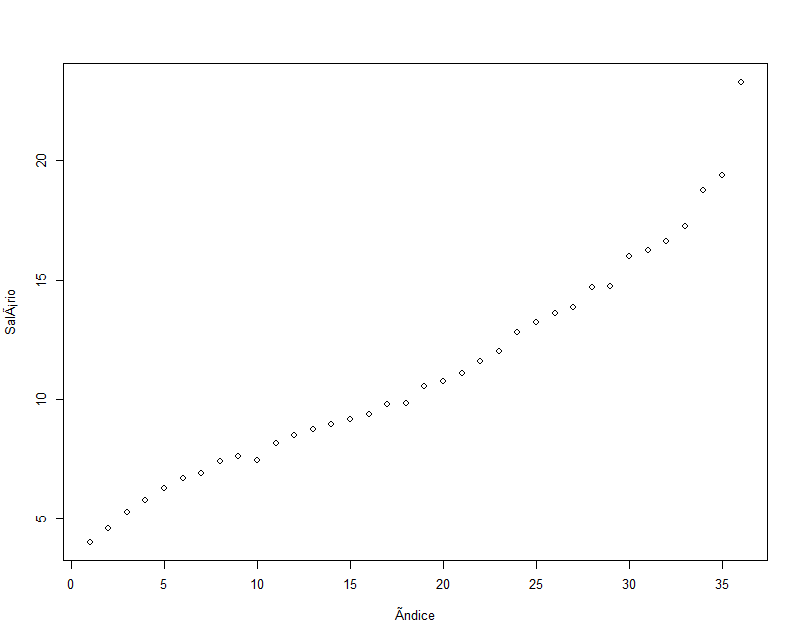
cbind("f" = addmargins(classes.tab),  
 "fr" = addmargins(prop.table(classes.tab)),  
 "F" = c(cumsum(classes.tab), NA),  
 "Fr" = c(cumsum(classes.tab), NA)/36)

f fr F Fr  
[4,8) 10 0.27777778 10 0.2777778  
[8,12) 12 0.33333333 22 0.6111111  
[12,16) 8 0.22222222 30 0.8333333  
[16,20) 5 0.13888889 35 0.9722222  
[20,24) 1 0.02777778 36 1.0000000  
Sum 36 1.00000000 NA NA

Na sequÃªncia vamos mostrar trÃªs possÃ­veis grÃ¡ficos para variÃ¡veis contÃ­nuas: o de dispersÃ£o, o histograma, e o de ramo-e-folhas.

O grÃ¡fico de dispersÃ£o unidimensional consiste da variÃ¡vel de interesse no eixo Y, plotada com seus respectivos Ã­ndices (entrada) da tabela de dados. Nesse exemplo, o grÃ¡fico seria

plot(milsa$Salario, xlab = "Ãndice", ylab = "SalÃ¡rio")



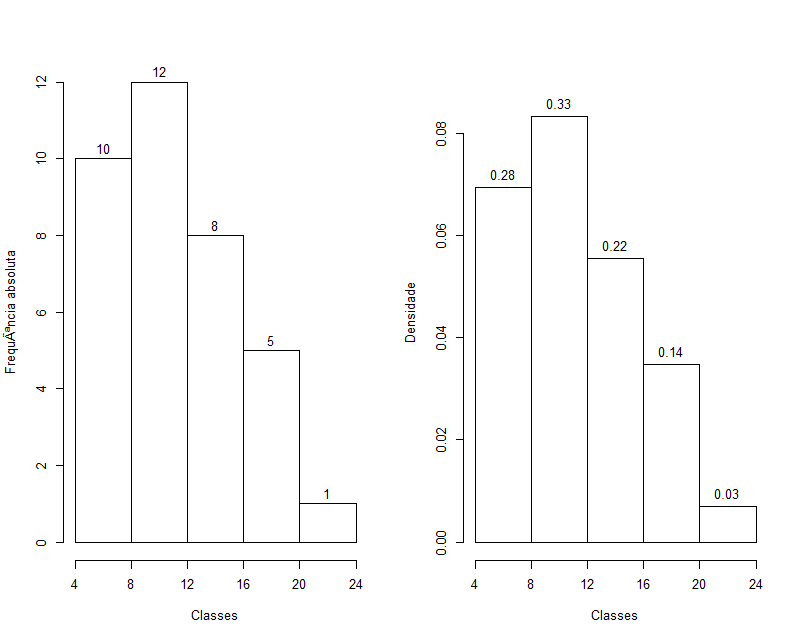
Note que, como os dados nesta tabela estÃ£o ordenados (em ordem crescente) pelos salÃ¡rios, entÃ£o vemos esse padrÃ£o de cresimento. Mas a ordenaÃ§Ã£o nÃ£o Ã© um prÃ©-requisito para fazer esse grÃ¡fico. A interpretaÃ§Ã£o desse grÃ¡fico Ã© limitada pelo fato de que sÃ³ faz sentido dizer alguma coisa sobre os salÃ¡rios se os Ã­ndices (nesse caso, os funcionÃ¡rios) tivessem algum tipo de identificaÃ§Ã£o. Por esse motivo, o histograma Ã© uma forma mais eficiente de resumir variÃ¡veis contÃ­nuas.

O **histograma** Ã© um grÃ¡fico de barras *contÃ­guas*, com as bases proporcionais aos intervalos de classe, e a Ã¡rea de cada retÃ¢ngulo proporcional Ã  respectiva frequÃªncia.

Pode-se usar tanto a frequÃªcia absoluta () quanto a relativa (). Indicamos a amplitude do -Ã©simo intervalo por . Para que a Ã¡rea do retÃ¢ngulo respectivo seja proporcional a , a sua altura deve ser proporcional a , que Ã© chamada **densidade de frequÃªncia**, ou simplesmente **densidade** da -Ã©sima classe. Quando todos os intervalos de classe forem iguais a (como Ã© o caso nesse exemplo), entÃ£o a densidade de frequÃªncia simplifica para . Com essa convenÃ§Ã£o, garantimos que a Ã¡rea total do histograma serÃ¡ igual a um, uma propriedade que serÃ¡ importante quando estudarmos probabilidade.

Abaixo, vemos os histogramas com as frequÃªncias absolutas, e com a densidade de frequÃªncia.

par(mfrow = c(1,2))  
hist(milsa$Salario, breaks = quebra, right = FALSE, main = "",  
 xlab = "Classes", ylab = "FrequÃªncia absoluta",  
 xlim = c(4, 24), labels = TRUE, axes = FALSE)  
axis(1, at = quebra, labels = quebra); axis(2)  
h <- hist(milsa$Salario, breaks = quebra, right = FALSE, main = "",  
 xlab = "Classes", ylab = "Densidade",  
 xlim = c(4, 24), ylim = c(0, .09), freq = FALSE, axes = FALSE)  
axis(1, at = quebra, labels = quebra); axis(2)  
text(h$mids, h$density, labels = round(h$density \* 4, 2), pos = 3)  
par(mfrow = c(1,1))



Note que, para construir o grÃ¡fico com as densidades, precisamos entÃ£o de mais uma coluna na tabela, que seria a densidade .

cbind("f" = addmargins(classes.tab),  
 "fr" = addmargins(prop.table(classes.tab)),  
 "F" = c(cumsum(classes.tab), NA),  
 "Fr" = c(cumsum(classes.tab), NA)/36,  
 "Dens" = c(prop.table(classes.tab)/4, NA))

f fr F Fr Dens  
[4,8) 10 0.27777778 10 0.2777778 0.069444444  
[8,12) 12 0.33333333 22 0.6111111 0.083333333  
[12,16) 8 0.22222222 30 0.8333333 0.055555556  
[16,20) 5 0.13888889 35 0.9722222 0.034722222  
[20,24) 1 0.02777778 36 1.0000000 0.006944444  
Sum 36 1.00000000 NA NA NA

O que devemos observar em um histograma:

* Amplitude de valores
* Forma da distribuiÃ§Ã£o (assimetria):
  + Positiva
  + Negativa
  + SimÃ©trica
* TendÃªncia central
* Valores extremos

Tanto o histograma quanto os grÃ¡ficos de barra, dÃ£o uma ideia da **forma da distribuiÃ§Ã£o** de uma variÃ¡vel. Existem vÃ¡rias medidas para resumir uma variÃ¡vel, como veremos mais adiante, mas a forma da distribuiÃ§Ã£o Ã© tÃ£o importante quanto estas medidas. Por exemplo, saber que a mÃ©dia de salÃ¡rios da empresa “Milsa” Ã© 11.12 Ã© importante, mas saber como os salÃ¡rios se distribuem entre os funcionÃ¡rios da empresa (como vimos pelo histograma) pode ser mais importante.

Outro mÃ©todo para se resumir uma variÃ¡vel com o objetivo de obter uma ideia da forma de sua distribuiÃ§Ã£o, Ã© atravÃ©s do “grÃ¡fico” de **ramo-e-folhas**. Este Ã© um mÃ©todo bem simples, e foi criado na Ã©poca em que fazer grÃ¡ficos no computador era uma tarefa difÃ­cil. Mas ainda hoje Ã© uma forma de visualizaÃ§Ã£o muito efetiva. Uma vantagem sobre o histograma Ã© que nÃ£o perdemos (ou perdemos muito pouca) informaÃ§Ã£o sobre os dados em si.

NÃ£o existe uma regra fixa para construir o ramo-e-folhas, mas a ideia bÃ¡sica Ã© dividir cada observaÃ§Ã£o em duas partes: a primeira (o *ramo*) Ã© colocada no lado esquerdo de uma linha vertical, e a segunda (a *folha*) Ã© colocada Ã  direita. Geralmente, a divisÃ£o natural Ã© separar um nÃºmero contÃ­nuo em sua parte inteira Ã  esquerda, e a parte decimal Ã  direita. Dessa forma o ramo-e-folha para a variÃ¡vel Salario estÃ¡ abaixo.

stem(milsa$Salario, scale = 2)

The decimal point is at the |  
  
 4 | 06  
 5 | 37  
 6 | 379  
 7 | 446  
 8 | 157  
 9 | 01488  
 10 | 58  
 11 | 16  
 12 | 08  
 13 | 269  
 14 | 77  
 15 |   
 16 | 026  
 17 | 3  
 18 | 8  
 19 | 4  
 20 |   
 21 |   
 22 |   
 23 | 3

Note que, por simplificaÃ§Ã£o, os valores foram arredondados para uma casa decimal, assim, por exemplo, o segundo valor de 4,56 foi arredondado para 4,6, e representado na primeira linha do grÃ¡fico, junto com o primeiro valor que Ã© 4,0.

Algumas informaÃ§Ãµes que podemos obter a partir do ramo-e-folhas:

* HÃ¡ um destaque para o valor 23,3.
* Os demais valores se concentram mais fortemente entre 4,0 e 14,7.
* Um valor mais ou menos tÃ­pico para essa variÃ¡vel seria em torno de 9.
* HÃ¡ uma leve assimetria em direÃ§Ã£o aos valores grandes.

Existem diversas variaÃ§Ãµes do grÃ¡fico de ramo-e-folhas, mas nÃ£o entraremos em detalhes.

# ExercÃ­cios

1. Em um questionÃ¡rio aplicado aos estudantes que frequentavam a livraria do campus, foi perguntado como classificariam o serviÃ§o prestado. As respostas foram:

ex <- c("Ã³timo", "bom", "bom", "pÃ©ssimo", "bom", "bom", "Ã³timo",  
 "Ã³timo", "bom", "Ã³timo", "bom", "Ã³timo", "bom", "bom", "Ã³timo",  
 "bom", "pÃ©ssimo", "bom", "pÃ©ssimo", "bom", "pÃ©ssimo", "bom", "bom",  
 "bom", "bom", "Ã³timo", "bom", "pÃ©ssimo", "Ã³timo", "Ã³timo", "bom",  
 "pÃ©ssimo")  
ex

[1] "Ã³timo" "bom" "bom" "pÃ©ssimo" "bom" "bom"   
 [7] "Ã³timo" "Ã³timo" "bom" "Ã³timo" "bom" "Ã³timo"   
[13] "bom" "bom" "Ã³timo" "bom" "pÃ©ssimo" "bom"   
[19] "pÃ©ssimo" "bom" "pÃ©ssimo" "bom" "bom" "bom"   
[25] "bom" "Ã³timo" "bom" "pÃ©ssimo" "Ã³timo" "Ã³timo"   
[31] "bom" "pÃ©ssimo"

1. Como vocÃª classificaria estas respostas?
2. Crie uma tabela e um grÃ¡fico para representar e resumir esta informaÃ§Ã£o.
3. Considere os dados abaixo

ex1 <- c(11,13,10,9,0,16,28,15,14,1,2,17,  
 18,12,18,8,9,22,16,20,34,15,21,13,  
 12,14,13,19,0,2,17,11,18,16,13,6,  
 19,8,8,12,13,21,11,19,1,14,4,16)  
sort(ex1)

[1] 0 0 1 1 2 2 4 6 8 8 8 9 9 10 11 11 11 12 12 12 13 13 13  
[24] 13 13 14 14 14 15 15 16 16 16 16 17 17 18 18 18 19 19 19 20 21 21 22  
[47] 28 34

1. Monte uma tabela construindo as classes conforme os critÃ©rios apresentados.
2. Calcule as frequÃªncias: absoluta, relativa, percentual, acumulada, acumulada relativa, acumulada percentual.
3. FaÃ§a um histograma com a *densidade* das classes.
4. FaÃ§a um diagrama de ramo-e-folhas.