Estatística Descritiva com o R

Henrique Alvarenga da Silva

18/01/2018

```
library(tidyverse)
library(gridExtra)
```

Introdução

Objetivos de aprendizagem:

1. Moda, Média, Mediana
2. Amplitude
3. Amplitude interquartil
4. Percentil
5. Variância
6. Desvio Padrão
7. Propriedades do desvio padrão
8. Assimetria

Estatistica Descritiva

Parte 1: Carregamento do dataset *mpg*

Para experimentarmos usar as diversas funções estatísticas do R, vamos primeiro carregar um conjunto de dados que vem junto com o pacote `ggplot2` . Se você já instalou o pacote `tidyverse` então o `ggplot2 já foi instalado. Você pode também instalar o pacote` ggplot2 `isoladamente se desejar. O comando para instalar os pacotes é o` install.pac)`. Lembre-se de que a instalação de um pacote deve ser sempre feita no console e nunca num R script ou num R Notebook.

Para instalar o pacote tidyverse, use o comando abaixo no console: (recomendado)
`install.packages("tidyverse")` .

Caso prefira, instalar o ggplot2 isoladamente, use o comando abaixo no console:
`install.packages("ggplot2")`

Com os pacote `ggplot2` instalado podemos agora usar os datasets que vem junto com esse pacote. para isso é necessário primeiro carregar o pacote `ggplot2 na memória do computador, o que é feito com a função` library(ggplot2) **.** `Atenção: observe que ao usar a função` library() não precisamos usar as aspas.

Com o pacote `ggplot2` estão carregados na memória e podemos usar os datasets que vem junto com esse pacote. Vamos usar o dataset `mpg` para nosso treinamento de estatística descritiva no R. O acrônimo `mpg` significa *Miles Per Gallon* - uma medida de quantas milhas um carro pode viajar se você colocar apenas um galão de gasolina ou diesel em seu tanque. (1 galão equivale a 3.79 litros e uma milha equivale a 1.6km).

Esta medida padronizada serve comparar carros com base na sua eficiência. O conjunto de dados `mpg` que vem junto com o `ggplot2` é um subconjunto dos dados de economia de combustível que a EPA (Enviroment Protection Agency - USA) disponibiliza em http://fueleconomy.gov (http://fueleconomy.gov). O conjunto completo dos dados podem ser obtidos nesse site, no link seguir: http://fueleconomy.gov/feg/download.shtml (http://fueleconomy.gov/feg/download.shtml).

Para facilitar essa aula, vamos usar simplesmente o dataset `mpg` que já vem com o `ggplot2` . Para carregar esse dataset basta usar o comando abaixo:

```
data(mpg)
```

Com a função class( ) podemos verificar que esse dataset é um data frame do R.

```
class(mpg)
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

Em primeiro lugar, vamos visualizar as primeiras linhas desse dataset com a função `head():

```
head(mpg)
```

manufacturer <chr>	**model** <chr>	**displ** <dbl>	**year** <int>	**cyl** <int>	**trans** <chr>	**drv** <chr>	**cty** <int>	**hwy** <int>	**fl** <chr>
audi	a4	1.8	1999	4	auto(l5)	f	18	29	p
audi	a4	1.8	1999	4	manual(m5)	f	21	29	p
audi	a4	2.0	2008	4	manual(m6)	f	20	31	p
audi	a4	2.0	2008	4	auto(av)	f	21	30	p
audi	a4	2.8	1999	6	auto(l5)	f	16	26	p

manufacturer <chr>	model <chr>	displ <dbl>	year <int>	cyl <int>	trans <chr>	drv <chr>	cty <int>	hwy <int>	fl <chr>
audi	a4	2.8	1999	6	manual(m5)	f	18	26	p

6 rows | 1-10 of 11 columns

Com esse comando você pode visualizar a tabela com os dados do dataset `mpg`. Podemos também usar o comando str( ) para visualizarmos a estrutura desse data frame, as variáveis e seus tipos:

```
str(mpg)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    234 obs. of  11 variables:
##  $ manufacturer: chr  "audi" "audi" "audi" "audi" ...
##  $ model       : chr  "a4" "a4" "a4" "a4" ...
##  $ displ       : num  1.8 1.8 2 2 2.8 2.8 3.1 1.8 1.8 2 ...
##  $ year        : int  1999 1999 2008 2008 1999 1999 2008 1999 1999 2008 ...
##  $ cyl         : int  4 4 4 4 6 6 6 4 4 4 ...
##  $ trans       : chr  "auto(l5)" "manual(m5)" "manual(m6)" "auto(av)" ...
##  $ drv         : chr  "f" "f" "f" "f" ...
##  $ cty         : int  18 21 20 21 16 18 18 18 16 20 ...
##  $ hwy         : int  29 29 31 30 26 26 27 26 25 28 ...
##  $ fl          : chr  "p" "p" "p" "p" ...
##  $ class       : chr  "compact" "compact" "compact" "compact" ...
```

Esse dataset possui 243 linhas (observações) com 11 variáveis, conforme a tabela abaixo:

variável	significado
manufacturer	marca
model	modelo
displ	cilindradas
year	ano de fabricação
cyl	número de cilindros
trans	tipo de marcha: automática / manual
drv	tração: f=frontal, r=traseira, 4=4x4
cty	milhas por galão na cidade
hwy	milhas por galão na estrada
fl	tipo de combustível: r=regular, p=premium, d=diesel, e=ethanol, c=CNG (gás)
class	tipo de carro

Parte 2 - Medidas de Tendência Central

Média e Mediana

A medida mais frequentemente investigada num conjunto de dados é seu centro, ou o ponto no qual as observações tendem a se concentrar. Medidas de tendência central são as estatísticas que descrevem um conjunto de dados pela sua posição central.

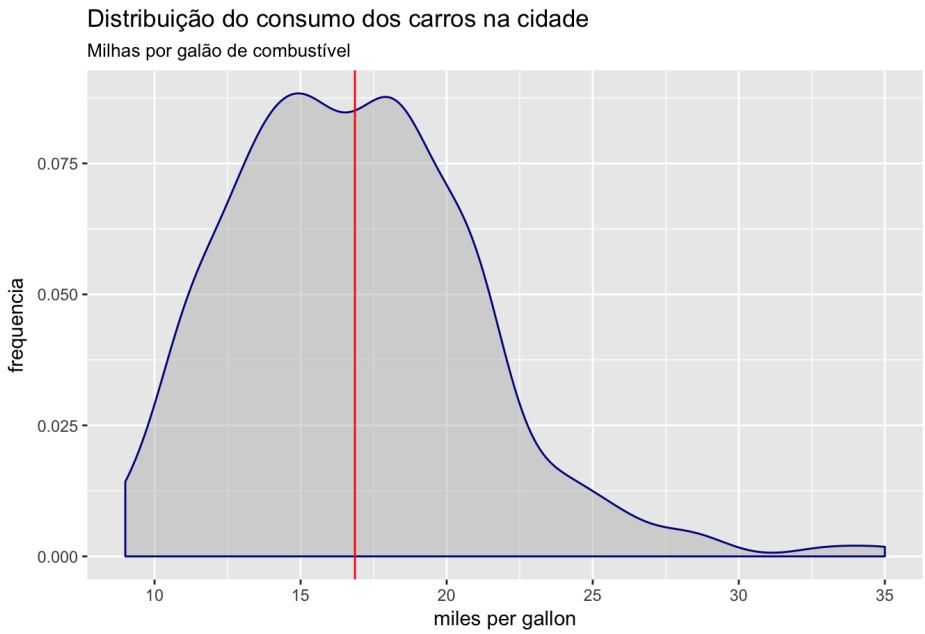
Existem várias medidas (estatísticas) que identificam a posição central de um conjunto de dados. As principais estatísticas que resumem um conjunto de dados pela posição central são a média, a mediana e a moda. A média e a mediana são facilmente calculadas no R através das funções `mean()` e `median()`.

Vamos avaliar a média do consumo dos carros na cidade, lembrando que esses dados estão na variável `cty` do dataset `mpg`. Lembre-se que para acessar uma variável de um data frame usamos o operador `$` da seguinte maneira: `nome.do.data.frame$nome.da.variável`.

```
mean(mpg$cty) # média do consumo na cidade - milhas percoridas por galão
```

```
## [1] 16.85897
```

Vejamos graficamente a posição da média na distribuição dos dados de consumo na cidade:



Fonte: dataset mpg do pacote ggplot2

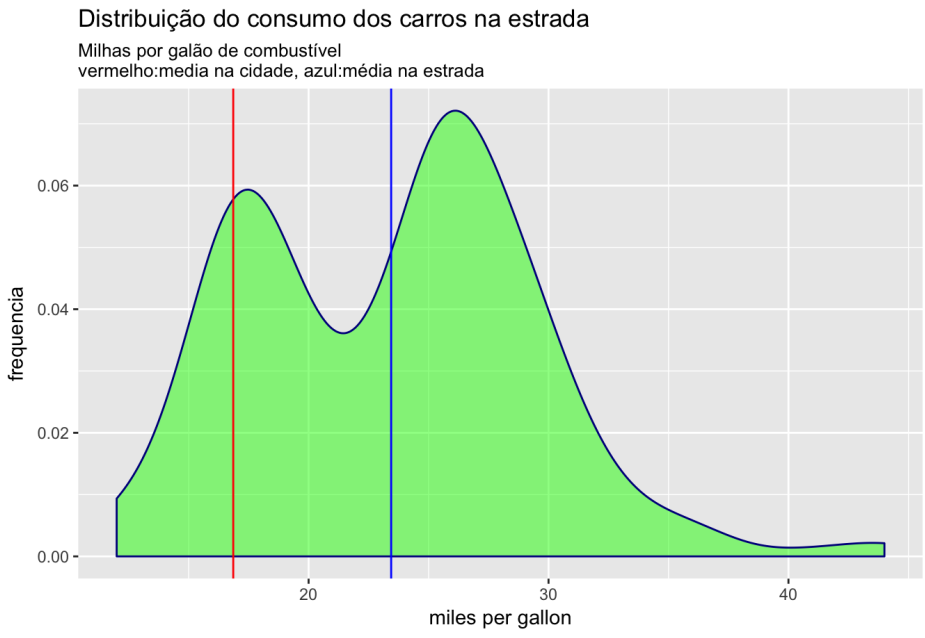
Podemos ver no gráfico que existem carros que percorrem menos de 10 milhas com um galão e alguns poucos carros que percorrem até 35 milhas com um galão. A distribuição dos dados ao redor da média é assimétrica à direita, pois os valores mais extremos estão à direita.

Vamos verificar agora a média do consumo na estrada:

```
mean(mpg$hwy)
```

```
## [1] 23.44017
```

A quantidade de milhas percorridas com um galão na estrada é bem maior que na cidade, como seria esperado. Mas a análise visual da distribuição poderá nos mostrar mais informações:



Fonte: dataset mpg do pacote ggplot2

Podemos ver que a média do consumo na estrada é bem maior, ou seja, na estrada os carros percorrem mais milhas que na cidade. Observe também a diferença na forma da distribuição do consumo na estrada: existem dois picos. Essa distribuição é chamada de *bimodal* devido a isso.

Vamos analisar agora a distribuição das cilindradas dos veículos, calculando inicialmente a média e a mediana. (variável é `displ`).

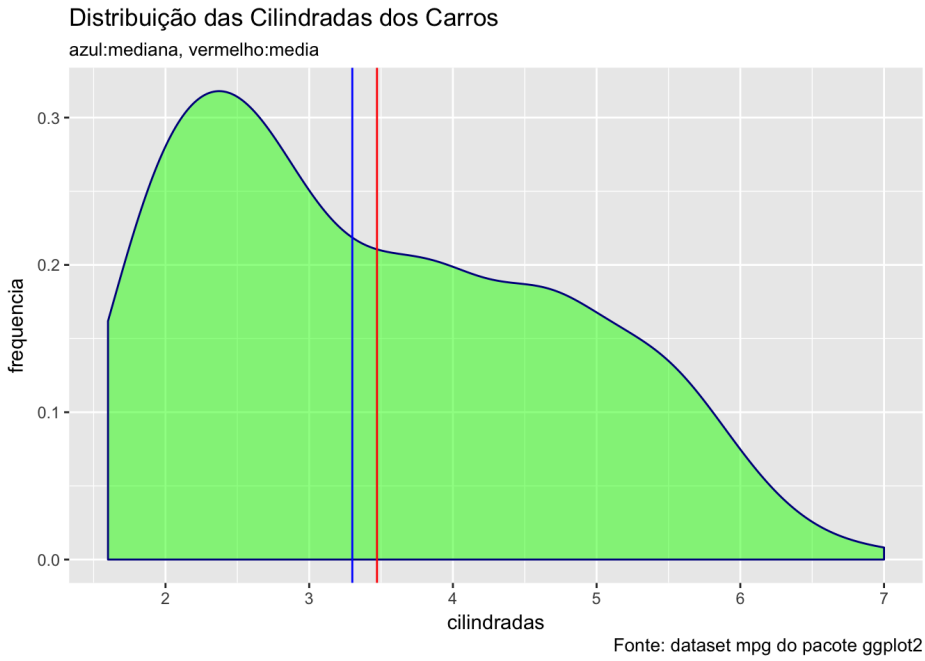
```
mean(mpg$displ)   # media das cilindradas
```

```
## [1] 3.471795
```

```
median(mpg$displ) # mediana das cilindradas
```

```
## [1] 3.3
```

A mediana é um pouco diferente da média, ligeiramente inferior à média. Se a distribuição desse dados fosse simétrica seria esperado que a média e a mediana fossem iguais, o que não é o caso. Vamos analisar visualmente a distribuição das cilindradas.



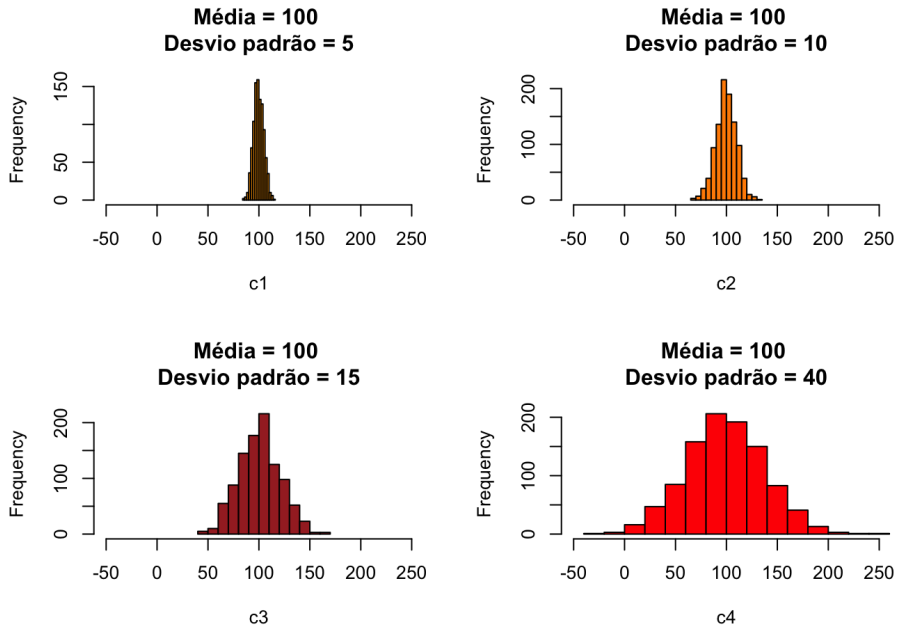
Fonte: dataset mpg do pacote ggplot2

Veja que a distribuição é assimétrica à direita, ou seja, tem valores mais extremos à direita. Esses valores aumentam o valor da *média*. A *mediana*, por outro lado, é menos afetada por valores extremos. Você pode verificar no gráfico acima que a mediana (azul) é menor que a média (vermelho). Analisar a simetria e dispersão de uma dispersão é assunto do próximo tópico.

Parte 3: Medidas de Dispersão

1. amplitude (range)
2. amplitude interquartil (IQR - interquartile range)
3. percentis (quantiles)
4. variância
5. desvio padrão (sd - standard deviation)

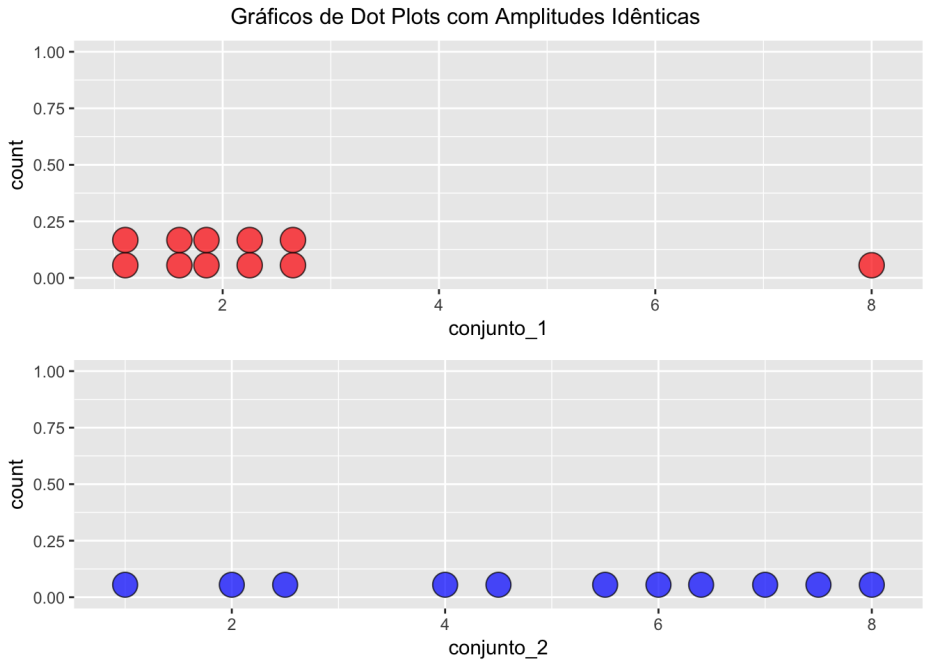
Conhecer o ponto central de uma distribuição não é suficiente para descrever completamente como os dados estão distribuídos. Alguns conjuntos de dados podem ter a mesma média e serem bastante diferentes. Veja nos gráficos abaixo alguns exemplos de dados com a mesma média, mas com grande diferença no que se refere à dispersão dos dados ao redor da média.



A necessidade descrever melhor um conjunto de dados tornou necessária a criação de medidas que descrevessem a dispersão dos dados ao redor da média.

Amplitude

A medida mais simples de dispersão dos dados é amplitude (range). A amplitude é a diferença entre o valor máximo e o valor mínimo. Veja no exemplo abaixo uma ilustração visual do significado dessa medida:



Podemos calcular os limites inferior e superior com a função `range()`. Essa função mostra os valores máximo e mínimo de um conjunto de dados.

```
range(conjunto_1)
```

```
## [1] 1 8
```

```
range(conjunto_2)
```

```
## [1] 1 8
```

A medida da amplitude é simplemente a diferença entre os valores máximo e minimo. Nesse caso a amplitude dos dois conjunto é identica, ou seja, 7 em ambos. Embora os limites sejam idênticos e a amplitude seja a mesma, esses dois conjuntos tem distribuições muito diferentes.

Vamos descobrir quais limites superiores e inferiores da distancia percorrida pelos carros, com um galão de combustível, na cidade, usando a função range:

```
range(mpg$cty)
```

```
## [1]  9 35
```

Podemos ver que existem carros que percorrem apenas 9 milhas com um galão e outros que percorrem 35 milhas com um galão. Ou seja, a amplitude dessa amostra é de 35-9 = 26 milhas.

Entretanto, essa informação é uma medida muito grosseira, pois a amplitude é calculada usando apenas 2 dados do conjunto, os valores máximo e mínimo.

Quartis, Percentil, Qualtil e Amplitude Interquartil

A amplitude interquartil é uma medida de variabilidade, baseada na divisão de um conjunto de dados em quartis. Os quartis dividem um conjunto de dados ordenados em quatro partes iguais. Os valores que separam partes são chamados de primeiro, segundo e terceiro quartis, e são denotados por Q1, Q2 e Q3, respectivamente. O primeiro quartil (Q1) é o valor abaixo do qual estão 25% dos dados, o segundo quartil (Q2) é o valor abaixo do qual estão 50% do dados e o terceiro quartil (Q3) é o valor abaixo do qual estão 75% dos dados.

A divisão dos conjunto de dados em percentuais é o que se denomina de *percentil*. Os percentis mais famosos são justamente o 1º quartil (= percentil 25%), o 2º quartil (= percentil 50% = mediana), e o 3º quartil (= percentil 75%).

Em estatística é comum usar o termo *quantil* para se referir ao percentil. A única diferença é que o quando usamos percentil usamos o número em sua forma de porcentagem (percentil 50%) e quando usamos o termo quantil usamos o número em sua forma decimal (quantil 0.5). O R tem uma função denominada `quantile()` para que você possa saber o quantil (ou percentil) que desejar.

Vamos calcular os quantis mais importantes, que são o primeiro, segundo e terceiro quartis: Q1, Q2 e Q3 no conjunto de dados da cilindrada dos carros no dataset `mpg`.

```
Q1 <- quantile(mpg$displ, probs = 0.25)
Q2 <- quantile(mpg$displ, probs = 0.50)
Q3 <- quantile(mpg$displ, probs = 0.75)

Q1
```

```
## 25%
## 2.4
```

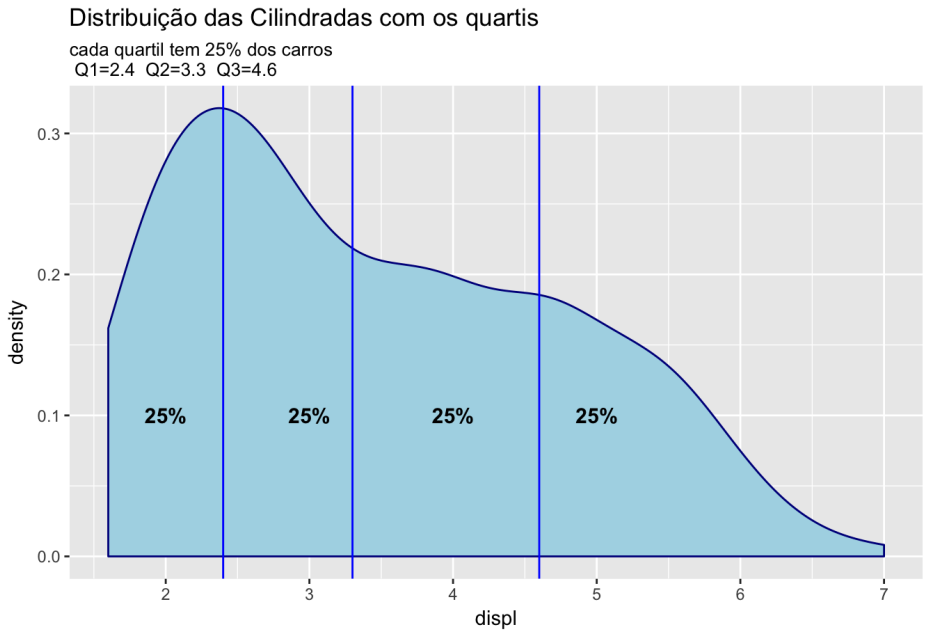
```
Q2
```

```
## 50%
## 3.3
```

```
Q3
```

```
## 75%
## 4.6
```

Podemos ver que o 2º quartil é exatamente a própria mediana. O gráfico abaixo ajuda a visualizar o significado dos quartis.



Fonte: dataset mpg do pacote ggplot2

Na estatística descritiva, a amplitude interquartil (IQR) é uma medida de dispersão estatística, sendo igual à diferença entre o quartil superior (Q3) e quartil inferior (Q1). É também chamada de `midspread` ou `middle 50%`, ou `H-spread`.

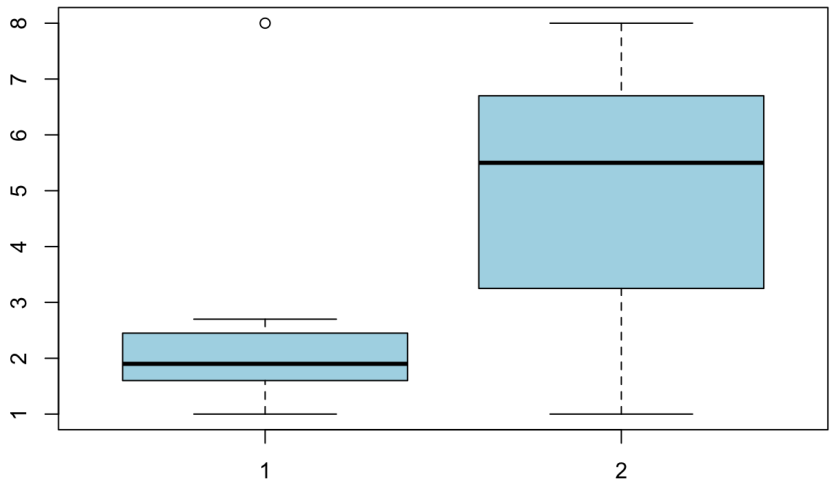
$$IQR = Q3 - Q1.$$

Em outras palavras, a amplitude interquartil (IQR) é `3º quartil` menos o `1º quartil`. Para compreender o que é o IQR observe os seguintes conjuntos de dados:

```
a <- c(1, 1.2, 1.5, 1.7, 1.8, 1.9, 2.2, 2.3, 2.6, 2.7, 8)
b <- c(1, 2, 2.5, 4, 4.5, 5.5, 6, 6.4, 7, 7.5, 8)
```

Esses dados tem a mesma amplitude, mas seu IQR é bastante diferente e pode ser visualizado num gráfico de box plot como abaixo

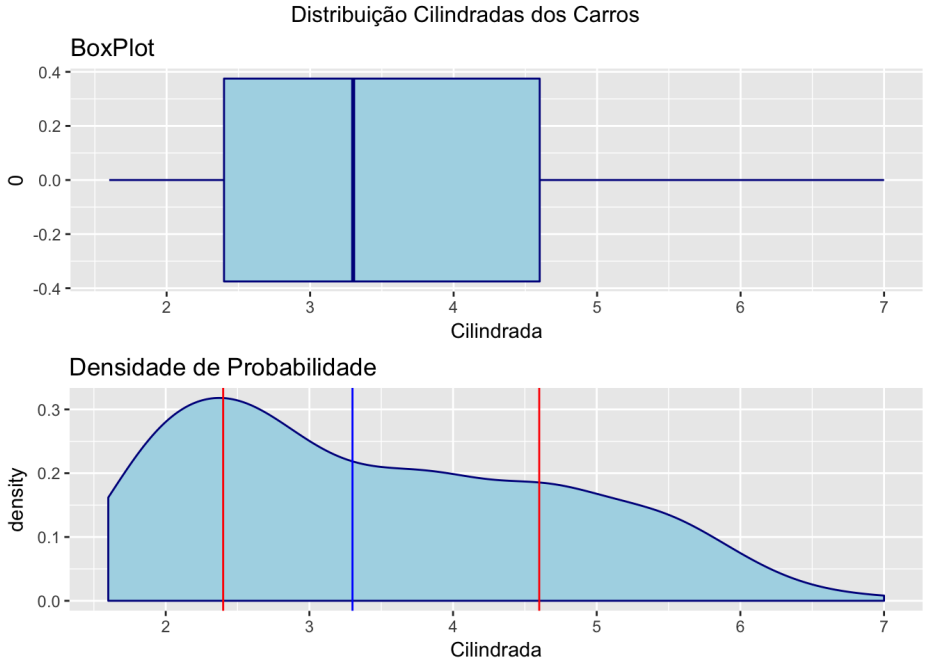
```
boxplot(a,b,
        col="lightblue")
```



Num boxplot o quadrado representa os 50% dos dados centrais do conjunto, ou seja, os limites das caixas de um boxplot são o 1º e o 3º quartis.

A medida IQR (amplitude interquartil) é mais robusta que a amplitude simples, pois desconsidera os 25% de dados superiores e os 25 inferiores.

Agora vamos criar um grafico de densidade e um boxplot com os dados da cilindrada para visualizarmos essas medidas.



O R tem funções para calcular o percentil que você desejar, mas como é um software estatístico, a função é `quantil()` e o não percentil. Para obter o valor correspondente ao quantil desejado, é preciso fornecer como argumento da função o objeto com os dados e o quantil desejado. O resultado é o valor que corresponde àquele quantil.

vamos testar calculando o quantil 0.5, que o mesmo que a mediana, usando os dados das cilindradas:

```
# calculando o quantil 0.5 = percentil 50%, que é o mesmo que mediana
quantile(mpg$displ, 0.5)
```

```
## 50%
## 3.3
```

```
# calculando a mediana, para verificarmos que o resultado é o mesmo
median(mpg$displ)
```

```
## [1] 3.3
```

O significado da mediana=3.3: 50% dos carros nesse dataset tem sua cilindrada acima de 3.3 50% dos carros nesse dataset tem sua cilindrada abaixo de 3.3

Variância e Desvio Padrão

Como vimos, a amplitude interquartil (IQR) é uma medida mais interessante que a mera amplitude, pois descarta os valores extremos, sendo assim menos sensível a outliers, ou seja, é uma medida considerada mais *robusta* que a simples amplitude. entretanto, o IQR também sofre do mesmo problema que a amplitude: é um cálculo que só usa dois dados, deixando muitos dados fora da análise.

Precisamos encontrar uma nova medida que represente melhor a dispersão do conjunto de dados, usando *todos* os dados desse conjunto.

Uma forma de fazer isso é analisar a distância de cada dado em relação à média, somar todas essas distâncias e dividir o valor encontrado pelo número de elementos do conjunto de dados (234 no nosso caso do dataset `mpg`). Entretanto, essa simples soma iria resultar sempre em *zero*, pois esse conjunto de distâncias de cada dado em relação à média iria conter valores positivos e negativos que iriam se cancelar totalmente. Vamos testar isso com o R no conjunto de dados do consumo dos carros na cidade do dataset `mpg`.

```
desvios <- (mpg$cty - mean(mpg$cty))
resultado <- sum(desvios)/234
resultado
```

```
## [1] 1.275333e-15
```

Para visualizarmos esse número da forma usual usamos a função format

```
format(resultado, scientific = FALSE)
```

```
## [1] "0.000000000000001275333"
```

O resultado dessa conta no R não foi exatamente ZERO devido a problemas de arrendondamento. Mas, teoricamente, se não houvessem erros de arredondamente, o valor seria exatamente sempre ZERO. Logo, essa medida é inútil.

Uma possibilidade alternativa é somarmos os módulos das distâncias, para que os valores não se cancelem, podemos fazer isso no R usando a função `abs()`.

```
resultado2 <- sum(abs(desvios))/234
resultado2
```

```
## [1] 3.347359
```

Entretanto, a operação de módulo (ou valor absoluto) é uma operação matemática com muitas limitações. Existe um outro modo melhor de evitarmos valores negativos nas distancias calculadas: podemos usar o quadrado das distâncias. Essa é a medida conhecida como *variância*.

Variância

```
resultado3 <- sum((desvios)^2)/234
resultado3
```

```
## [1] 18.03567
```

É preciso, entretanto, fazer uma observação sobre o denominador desse fórmula. Quando calculamos a variância de uma população usando todos os dados da população o denominador é justamente o tamanho dessa população. Entretanto, na maioria das vezes, trabalhamos apenas com uma amostra da população. Uma das principais funções da estatística é fazer inferências sobre a população usando como referência uma amostra. Nesse caso, quando estamos analisando uma amostra, o denominador da variância precisa de um pequeno ajuste, ao invés de `n`, o deve ser `n-1`. Ou seja, no caso de nossa amostra de 234 carros, usaremos no denominador o valor 233 (234-1).

```
resultado4 <- sum((desvios)^2)/233
resultado4
```

```
## [1] 18.11307
```

Essa é a medida denominada de variância amostral. E o R tem uma função própria para isso:

```
var(mpg$cty)
```

```
## [1] 18.11307
```

A variância encontrada foi de 18.11307 milhas ao quadrado.

Veja que o resultado da operação `var(mpg$cty)` foi idêntico ao calculado acima passo-a-passo. Ou seja, a fórmula da função `var()` utiliza no denominador o valor `n-1`. Para calcular a variância populacional deveríamos usar `n` no denominador, mas o R não tem essa função.

Para pensar: como você pode fazer para obter o valor da variância populacional a partir da função `var()` do R? Podemos encontrar o resultado desejado simplesmente multiplicando o resultado encontrado por $(n-1)/n$

Dica: escreva uma função chamada varPop!

Solução:

```
varPop <- function(x){
  n <- length(x) # calcula o tamanho da amostra com a função lenght()
  var(x)*(n-1)/n # calcula a variância amostral e multiplica por (n-1)/n
}
```

Vamos testar? Se nossa função estiver correta, ambos os cálculos abaixo deve obter o mesmo resultado que obtivemos acima quando calculamos nossa variancia usando 403 no denominador

```
varPop(mpg$cty)
```

```
## [1] 18.03567
```

A variância, calculada com a nova fórmula de variância populacional, é de 18.03567 milhas ao quadrado.

Desvio Padrão

Apesar da importância da variância, essa medida tem um pequeno probleminha: ao elevar ao quadrado, a unidade de medida dos dados também foi elevada ao quadrado (milhas ao quadrado). Portanto, se nossa amostra se refere à altura em metros, a unidade da variância é metros ao quadrado, se a unidade de nossa medida são dias, horas, centímetros, a unidade de medida da variância será $dias^2$, $horas^2$, cm^2. Essas medidas são de difícil interpretação, daí a necessidade de ajustarmos essas medidas.

O modo mais simples de fazer isso é justamente fazendo a *raiz quadrada da variância*. Desse modo, a unidade de medida volta a ser a original. A raiz quadrada da variância foi nomeada de desvio padrão, por ser a medida mais usada (padrão) de desvio dos dados, ou seja, de quanto os dados se desviam da medida central (média).

O R tem uma função preparada para o cálculo do desvio padrão: `sd` . Lembre-se que na língua inglesa desvio padrão se escreve 'standard deviation', daí `sd` . Vamos calcular o desvio padrão do consumo dos carros na cidade:

```
sd(mpg$cty)
```

```
## [1] 4.255946
```

Propriedades do Desvio Padrão

É a medida de dispersão mais comum na estatística, que utiliza em seus cálculos todos os dados. Devido ao fato de uma das etapas do cálculo do desvio padrão envolver a operação de elevar ao quadrado, o desvio padrão será sempre um valor positivo (ou zero). Se todos os valores dos dados forem iguais, o desvio padrão será ZERO. Uma das vantagens do desvio padrão é que a unidade de medida é a mesma dos dados da amostra, ao contrário da variância, na qual a unidade de medida é o quadrado da unidade original.

Ao analisar o desvio padrão, é preciso ter em mente que desvios pequenos indicam que os dados se agrupam em torno da média e desvios grandes indicam que os dados estão dispersos numa grande amplitude. Finalmente, como o cálculo do desvio padrão leva em conta todos os elementos do conjunto de dados, o resultado é que o desvio padrão é sensível aos outliers (valores extremos).

Conclusões finais

Nesse capítulo estudamos as medidas estatísticas de tendencia central e de dispersão. Vimos também que essas medidas numéricas são melhor compreendidas quando associadas a gráficos que ilustrem o que os números indicam. A próxima etapa da análise descritiva é justamente aprender a criar gráficos que ilustrem os dados.