DESCRIÇÃO DE UMA VARIÁVEL NUMÉRICA

[Documento Metodics] [1]

[1] Revisado por Mafalda, Janeiro 2023

A descrição estatística univariada consiste na caraterização da distribuição dos valores registrados por uma única variável. Esta caracterização inclui a consideração de 4 aspetos: a tendência central dos dados; a dispersão dos dados em relação ao centro; a forma geral da distribuição; e a presença de outliers. A continuação examinamos as principais ferramentas numéricas e visuais estatísticas para a análise desses 4 aspetos com uma variável numérica.

1. Antes da Análise: pergunta e dados

Pergunta de investigação

A duração em funções dos ministros do governo é uma dimensão analítica chave na literatura sobre formação de governos. Ministros que duram mais, estão associados a sistemas políticos mais estáveis e eficientes. Porém, a investigação tem focado em países grandes ou democracias consolidadas, não havendo muitos estudos sobre países pequenos com democracias recentes. Interessa-nos estudar quanto duram os ministros nestas democracias.

> *Quanto duram os ministros de governo nas democracias novas de países pequenos?*

Dados

Para responder à pergunta de investigação formulada, utilizamos a base tutorial_1. Esta é uma base fictícia que tem como unidade de análise ministros de governo de uma democracia nova e pequena, com 16 observações e X variáveis. Interessa-nos a variável 'duração', que registra o número de dias que o ministro ficou no governo.

```
load ("BaseTutorial_1.RData")
B <- BaseTutorial_1
```

Com fins didáticos, criamos mais quatro variáveis para duas amostras diferentes:

```
dura_a2 <- rnorm(16, mean = 154, sd = 5)

dura_a3 <- rnorm(16, mean = 154, sd = 50)

dura_a4 <- rnorm(16, mean = 154, sd = 30)
dura_a4[dura_a4==max(dura_a4)] <- 180        # com outlier

dura_a5 <- rnorm(16, mean = 154, sd = 30)
dura_a5[c(3,5,9)] <- NA                      # com NAs
```

Controlo dos dados

Vimos que as variáveis numéricas registram variações de valores numéricos e podem ser discretas (números inteiros), continuas (números fraccionais, por ex., 1.4), de rácio (quando o zero representa ausência real, por exemplo, na quantidade de laranjas, ter zero representa "não ter nenhuma laranja").

Esta falta de precisão por vezes pode ser fonte erro na realização de cálculos ou da obtenção de resultados válidos. Assim, o controlo da "classe" da variável deve estar sempre entre os primeiros passos do controlo de uma variável numérica (em realidade, de qualquer variável).

```
class(B$duração)
class(dura_a2)
class(dura_a3)
```

```
[1] "numeric"
[1] "numeric"
[1] "numeric"
```

As 3 variáveis são numéricas.

Também para qualquer variável, é importante assegurar se os dados variam entre os valores previstos, se existem valores em falta (NA), e se estes estão identificados com um número específico. As tarefas de limpeza e preparação dos dados são particularmente importantes, na medida em que as estatísticas descritivas e respetivas visualizações podem encobrir incongruências.

Controlo da presença de valores em falta:

```
table (is.na(B$duração))
```

```
FALSE
   16
```

O output dá-nos uma tabela de frequências dos valores que são missings (TRUE) e os que não são missings (FALSE). Para a variável duração, não existem valores em falta.

Controlo da presencia de valores em falta:

```
table ((B$duração))
```

```
 80 110 120 140 150 190 200 210 230 240
  3   2   1   2   1   2   2   1   1   1
```

2. Dados 'crus' (desagregados)

Toda a análise exploratória deve ter em consideração o estado 'natural' dos dados, a sua situação desagregada e anterior a serem tratados pela estatística.

Um simple '**print**' da variável devolve a distribuição dos seus valores, tal como aparecem (foram introduzidos) na base de dados.

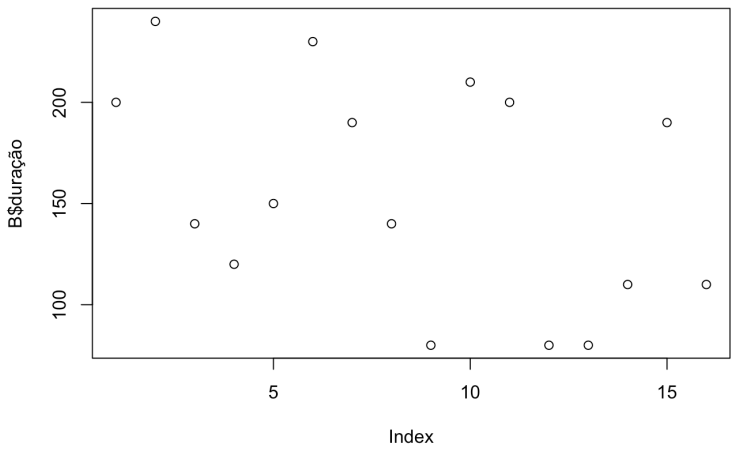
```
print (B$duração)
```

```
[1]200 240 140 120 150 230 190 140  80 210 200  80  80 110 190 110
```

A variável observa 16 indivíduos (N=16), onde o primeiro indivíduo durou no governo 200 dias, o quinto durou 150 e o último 110 dias.

Certamente esta modalidade é pouco prática com variáveis extensas, típicas das bases estatísticas, mas é uma ferramenta que deve estar sempre presente pelas suas utilidades básicas e aplicabilidade a qualquer objeto R. A alternativa mais efetiva para a exploração de uma variável, sem importar o seu número de observações, é a representação visual de um **gráfico de dispersão** (*plot, scaterplot*)

```
plot(B$duração)
```



O indivíduo 1 durou 200 dias, que o indivíduo 5 150 dias, e que o indivíduo 10 mais de 200 dias.

O eixo vertical indica o intervalo entre os valores mínimos e máximos observados. O eixo horizontal indica a sequência das observações, que correspondem às linhas de base e, pelo tanto, aos indivíduos observados.

A informação fornecida pelos plots costuma ser de interpretação aproximativa mas dá uma ideia da distribuição geral dos valores registados pela variável e da sua extensão, da existência de outliers, e também de possíveis erros de codificação (por exemplo, se neste caso tivéssemos encontrado alguma observação abaixo de 0)

Porém, uma das tarefas essenciais da estatística é produzir medições, agregações informativas dos dados como veremos de seguida.

3. Análise

Rasgo 1: Medidas de Tendência Central

Média: o valor médio

As medidas de tendência central orientam-se para detetar o centro das distribuições, um ponto médio. A primeira é seguramente a mais intuitiva destas medidas: a média aritmética, que consiste na soma dos valores registrados, divididos pela número total de observações:

```
(200+240+140+120+150+230+190+140+80+210+200+80+80+110+190+110)/16
```

```
[1] 154.375
```

Neste primeiro caso, foi calculado o valor da média manualmente, que é 154 dias.

Pode ser calculada de várias formas,

```
sum (B$duração)/length (B$duração)
```

```
[1] 154.375
```

Outra forma de calcular a média é dividir a soma de todos os valores da variável pelo comprimento da variável (ou seja, o número de observações).

Mas a mais prática é a mais simples:

```
mean (B$duração)
```

```
[1] 154.375
```

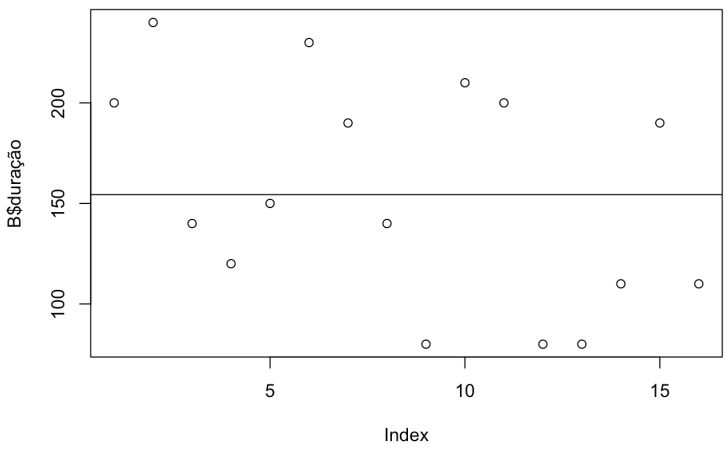
A média da duração no governo dos indivíduos da nossa amostra é de pouco mais de 154 dias.

Com um único número, sem importar o número de observações, a média fornece uma representação numérica do centro da distribuição.

E a informação obtida pode ser incorporada no gráfico de dispersão.

Scaterplot

```
plot (B$duração)
abline (h= mean(B$duração))
```



Mediana: a posição média

Um segundo centro da distribuição é o registrado pela mediana, que indica o valor que divide as observações em dois grupos, deixando 50% dos indivíduos de cada lado desse valor (i. e., 50% dos indivíduos tem um valor acima da mediana e 50% dos indivíduos tem um valor abaixo da mediana). Este é um valor posicional, que precisa de ter em conta a ordem sequencial dos valores observados.

Para calcular a mediana manualmente: primeiro devem ordenar-se os valores. Logo,

a) Se extensão da variável (N) é impar, a mediana será o valor posicionado no centro

4, 1,7,5,12 (N=5)
1,4**, 5,**7,12
mediana = 5

b) Se extensão da variável (N) é par, a mediana será a media dos dois valores centrais

4, 1,7,5,12,20 (N=6)
1,4**, 5,7,**12,20.
mediana = 6

Tendo a nossa variável uma extensão de 16 observações, a sua mediana encontra-se entre a posição 8 e 9.

```
sort (B$duração)
```

```
80  80  80 110 110 120 140 140 150 190 190 200 200 210 230 240
```

Todos os valores da variável duração, em ordem crescente.

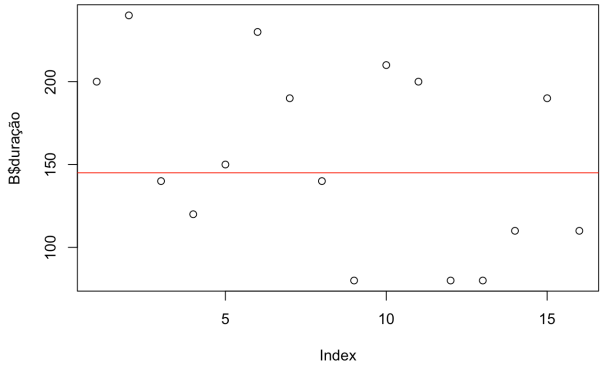
```
median (B$duração)
```

```
[1] 145
```

A mediana da duração dos indivíduos da nossa amostra é de 145 dias.

E a informação obtida pode ser incorporada no gráfico de dispersão:

```
plot (B$duração)
abline (h=mean(B$duração))
abline (h= median(B$duração), col="red")
```



Quando a distribuição dos valores de uma variável está repartida de modo mais ou menos simétricos acima e abaixo da média, os valores da mediana e da média serão similares (50% das observações estará por cima/debaixo da média). Este tipo de configurações é nomeada de distribuição normal.

A utilização da mediana torna-se útil em duas situações: quando a distribuição dos valores não não é simétrico[2]; e/ou quando existem outliers (a média é muito sensível a valores extremos).

Moda

Uma terceira medida de tendência central é a moda, que identifica o valor com a maior frequência. O seguinte output apresenta uma tabela de frequência dos valores da variável duração.

```
table (B$duração)
```

```
  80 110 120 140 150 190 200 210 230 240
  3   2   1   2   1   2   2   1   1   1
```

A moda é o valor com maior frequência.

```
# install.packages("lsr")
library("lsr")
modeOf(B$duração)
```

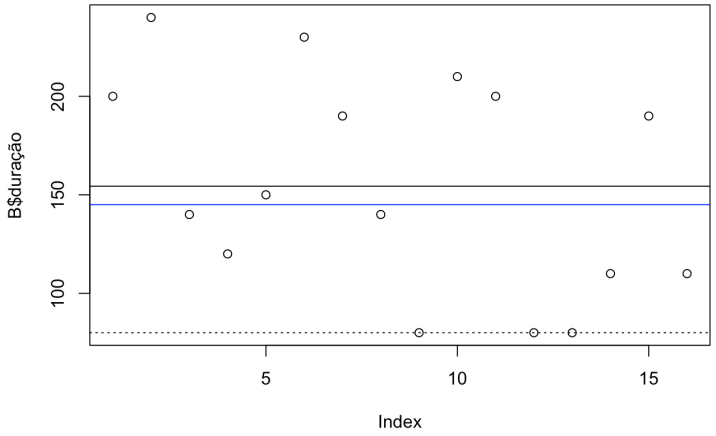
```
[1] 80
```

A moda da duração dos indivíduos da nossa amostra é de 80 dias

[2]Quanto mais alta/baixa a mediana é face à média, menos simétrica é a distribuição, ou seja, menos equilibrada é a distribuição de casos. Em alguns testes estatísticos, em que se pretende estimar os parâmetros da população com base nas estatísticas da amostra, a normalidade da distribuição dos dados é um requisito.

O seguinte gráfico de dispersão inclui as três medidas de tendência central.

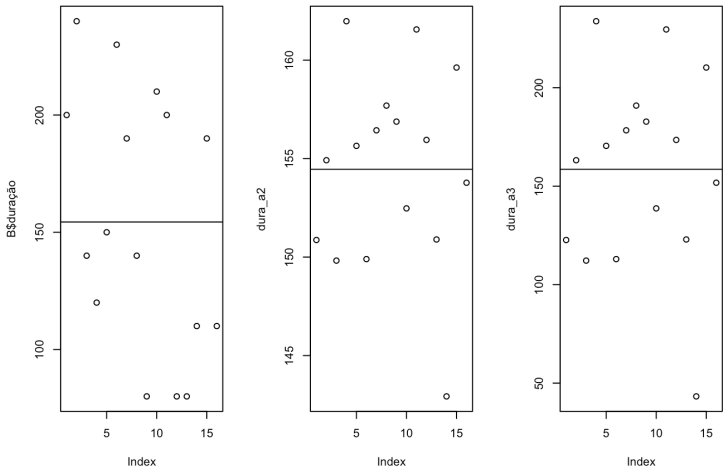
```
plot (B$duração)
abline (h=mean(B$duração))
abline (h=median(B$duração),col="blue")
abline (h=modeOf(B$duração),lty = 3)
```



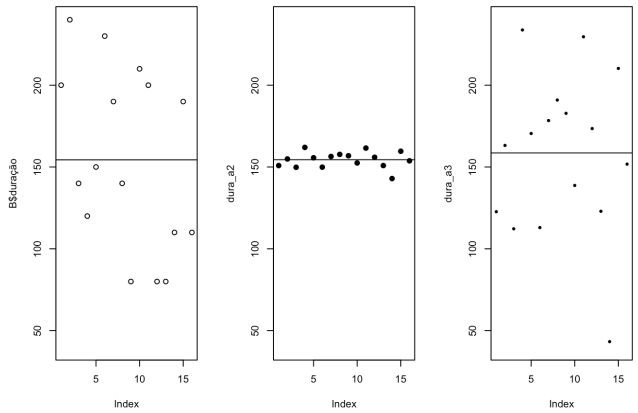
Gráficos múltiplos

Uma imagem pode valer mais de 1000 palavras, mas nunca é inocente. Os seguintes 2 conjuntos de gráficos mostram os mesmos dados.

```
par (mfrow =c(1,3))

plot(B$duração)
abline (h= mean (B$duração))

plot(dura_a2)
abline (h= mean (dura_a2))

plot(dura_a3)
abline (h= mean (dura_a3))
```



```
par (mfrow =c(1,3), pch=1)
plot(B$duração, ylim=c(40,240))
abline (h= mean (B$duração))

plot(dura_a2, ylim=c(40,240) , pch=19)
abline (h= mean (dura_a2))

plot(dura_a3, ylim=c(40,240) , pch=19,  cex = 0.5)
abline (h= mean (dura_a3))
```



No primeiro grupo, o intervalo do eixo y não está especificado, construindo-se em função de distribuição empírica da cada amostra, e tornando a comparação entre amostras difícil. No segundo grupo, este problema está corrigido, especificando-se um intervalo similar (40-240) para as três amostras.

Rasgo 2: Medidas de Dispersão

Se a tendência central é um primeiro aspeto que carateriza uma variável numérica, um segundo são os modos em que os valores se afastam em direção aos extremos da distribuição. Esta variabilidade é captada pelas medidas de dispersão. As três mais importantes são as seguintes.

Intervalo entre mínimo e máximo (range)

O intervalo é o modo mais intuitivo de capturar a dispersão de uma distribuição, e consiste na distância entre o valor mínimo e o valor máximo dos valores observados.

```
min (B$duração)
max (B$duração)
range(B$duração)
```

```
[1] 80
[1] 240
[1] 80 240
```

A duração dos ministros no governo na amostra 1, vai de um mínimo de 80 dias até uma duração máximo de 240 dias, sendo o intervalo total da dispersão de 160 dias.

Intervalo entre quartis (inter-quartile range)

Os quartis dividem as observações em quatro intervalos iguais. Assim,
- o primeiro quartil indica o valor abaixo do qual se encontram 25% dos valores observados,
- o segundo quartil (Q2) indica 50% dos valores (ou a mediana);
- o terceiro (Q3) indica 75%,
- e finalmente o quarto (Q4) indica 100%.

Estatisticamente, podemos analisar uma distribuição olhando para os valores que separam os dados em partes iguais. Estas porções dos conjuntos dos dados (as nossas linhas, os nossos sujeitos), são chamados genericamente de quartis (remete para quatro partes iguais). Os quartis dividem então os dados ordenados em 4 partes iguais. Por exemplo, para a variável duração:

80 80 80 110 [-] 110 120 140 140 [-] 150 190 190 200 [-] 200 210 230 240

Quartil 1º:

```
quantile (B$duração, .25)     # 25th quartiles = 11
```

```
25%
110
```

25% das observações apresenta uma duração no governo abaixo de 110 dias.

Quartil 2º:

```
quantile (B$duração, .50)
```

```
 50%
145
```

50% das observações apresenta uma duração no governo abaixo de 145 dias.

Quartil 3º :

```
quantile (B$duração, .75)      # 75th quartiles = 16.5
```

```
75%
200
```

75% das observações apresenta uma duração no governo abaixo de 200 dias.

O **interquartile range (IQR)** é a diferença entre o quartil 3 e o quartil 1 e indica o intervalo onde se encontra 50% das observações no centro da distribuição.

```
IQR (B$duração) # amplitude interquartil (q3-q1)
```

```
[1] 90
```

A amplitude interquartil é 90, ou seja, a distância entre o percentil 25 e o percentil 75 é 90.

O IQR utiliza-se para detectar outliers. Considera-se outlier qualquer observação que se localize para além do seguinte valor:

$$<Q1-1,5IQR \text{ ou } >Q3 +1,5IQR$$

Outras segmentações muito usadas são os decis (dividem os dados em 10 partes iguais), ou os percentis (dividem os dados em 100 partes, equivale ao valor percentual).

Variância, Desvio-padrão, Erro padrão [MODELO NULO]

O **desvio padrão** (junto com as suas variantes) é a representação estrela da dispersão, sendo com a média as duas ferramentas centrais na execução das técnicas da estatística. Se bem que esta medição apresenta um grau de sofisticação maior do que as estatísticas examinadas até agora, torna-se muito intuitiva após compreendidos os passos da sua computação. Vamos reconstruir esses passos, calculando a desvio padrão para a variável duração.

Começamos representando numa matriz o número de indivíduos, a variável duração e a média de toda amostra.

```
par (mfrow =c(1,1))

id <- 1:16
duração <- B$duração
S <- data.frame(id, duração)

S$media <- mean (S$duração)
S
```

id <int>	duração <dbl>	media <dbl>
1	200	154.375
2	240	154.375
3	140	154.375
4	120	154.375
5	150	154.375
6	230	154.375
7	190	154.375
8	140	154.375
9	80	154.375
10	210	154.375

A média é considerada o modelo estatístico nulo (Judd et al., 2017), e o modo mais básico para predizer o comportamento de uma variável. O nosso modelo nulo prediz que os indivíduos duram em média no seu cargo 154 dias. Este modelo funcionará sempre melhor para uns do que para outros. Ou seja, a predição será mais acertada para os indivíduos que estejam mais perto dessa média, e terá mais dispersão ou 'erro' quando mais longe dela fique. Por exemplo, para o indivíduo 1, o modelo tem o erro de quase 46 dias (200-154), no entanto que para o indivíduo 2 o erro é de 86 dias (200-154)

Cálculo do erro em relação à média (*error mean*) do nosso modelo nulo:

```
S$mediaerror <- S$duração - S$media    # error mean
S
```

id <int>	duração <dbl>	media <dbl>	mediaerror <dbl>
1	200	154.375	45.625
2	240	154.375	85.625
3	140	154.375	-14.375
4	120	154.375	-34.375
5	150	154.375	-4.375
6	230	154.375	75.625
7	190	154.375	35.625
8	140	154.375	-14.375
9	80	154.375	-74.375
10	210	154.375	55.625

Um passo básico na construção do desvio padrão é calcular a quantidade total de erro do modelo nulo. Mas a simples soma dos erros dá como resultado 0. Porquê? Porque os valores positivos (acima de média)anulam com os valores negativos (abaixo da média)

```
sum (S$mediaerror) # soma de erros (sum of error mean)
```

```
[1] 0
```

Se só somarmos os valores dos erros obtemos zero.

A solução proposta pela estatística é considerar o quadrado dos erros. Deste modo conseguem-se duas coisas:

(i) evitar a anulação entre valores positivos e negativos, e

(ii) dar mais peso aos valores mais afastados da média fazendo que, por exemplo, um modelo com duas observações a 2 pontos da média se considere como tendo menos erro do que um modelo com uma observação a 4 pontos de distância.

2^2 + 2^2= 8

4^2 = 16

Cálculo do quadrado dos erros:

```
S$mediaerror2 <- S$mediaerror^2 # soma de quadrado dos erros
S
```

A soma do quadrado dos erros fornece uma primeira medição da dispersão do erro da variável:

id <int>	duração <dbl>	media <dbl>	mediaerror <dbl>	mediaerror2 <dbl>
1	200	154.375	45.625	2081.64062
2	240	154.375	85.625	7331.64062
3	140	154.375	-14.375	206.64062
4	120	154.375	-34.375	1181.64062
5	150	154.375	-4.375	19.14062
6	230	154.375	75.625	5719.14062
7	190	154.375	35.625	1269.14062
8	140	154.375	-14.375	206.64062
9	80	154.375	-74.375	5531.64062
10	210	154.375	55.625	3094.14062

```
sum (S$mediaerror2) # variância
```

```
[1] 44993.75
```

O quadrado da soma das médias é 4493.75

Vimos que para o cálculo da média, primeiro se soma o valor das observações, e depois se divide pelo número de indivíduos (ou tamanho da amostra). Um passo similar faz-se para chegar à ***variância***: a soma total do quadrado dos erros é dividida pelo tamanho da amostra, com uma pequena correção matemática[3]:

$$\sigma^2 = (\Sigma (x-\mu)^2) / N-1$$

Podemos calcular a variância manualmente:

```
varianciaDuração <- sum (S$mediaerror2) /
(length(S$duração)-1)
varianciaDuração
```

```
[1] 2999.583
```

Ou podemos simplesmente fazê-lo utilizando um função:

```
var(S$duração)
```

```
[1] 2999.583
```

[3] Esta correção tem a ver com os graus de liberdade. Em estatística não podemos fazer análises infinitas aos dados, por cada análise que fazemos perdemos graus de liberdade. Por isso não dividimos por N, mas N-k, em que o k.

Agora, dizer que a variância da variável duração é 2999 não nos diz muito porque, depois de aplicar o cálculo quadrático (o quadrado dos erros em lugar dos erros), perdemos a escala original dos dados (duração em dias). Este problema resolve-se aplicando a raiz quadrada à variância, o que dá como resultado o ***desvio padrão*** *(standard deviation)* da variável.

Podemos calcular o desvio padrão quase manualmente:

```
sqrt  (varianciaDuração)
```

```
[1] 54.76845
```

Ou mais simplesmente usando função:

```
sd(S$duração)
```

```
[1] 54.76845
```

A média da duração no cargo é de 154 dias, com um desvio padrão de quase 55 dias.

O desvio padrão resulta da raiz quadrada da variância, basicamente é o valor da variância apresentado na mesma escala dos dados. Esta estatística é utilizada para construir os famosos ***margem de erro*** e ***intervalos de confiança***, a partir dos ***erros padrão*** *(standard error)*, estando também na base de muitas estimações e testes estatísticos que serão explorados no módulo dedicado à inferência.

```
sd(S$duração) /  sqrt (length(S$duração))# erro padrão (standard error)
```

```
[1] 13.69211
```

Aspeto 3: Outliers

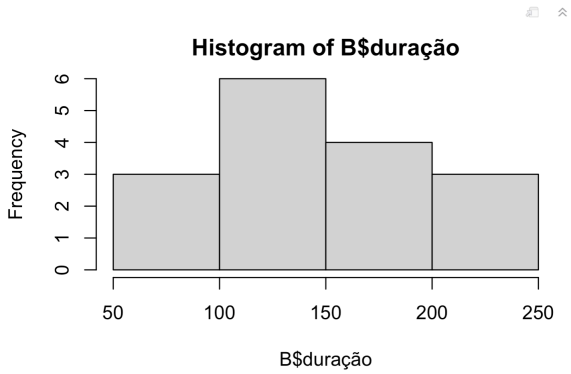
Os outliers são observações isoladas que se afastam exageradamente do centro de uma distribuição. A sua presença costuma alterar os resultados estatísticos e por isso devem ser identificados. Por exemplo, na distribuição de riqueza no mundo existem vários outliers, sobretudo as pessoas mais ricas do mundo. Se estivéssemos a fazer um estudo para perceber quais os rendimentos da população portuguesa e escolhermos 30 pessoas, nas quais esteja incluído o Cristiano Ronaldo, quando fizermos a média, vamos ter um valor muito mais elevado do que a real média de rendimentos da população portuguesa (o que nos lembra também da importância de uma boa amostragem!).
São fáceis de observar com ferramentas como os histogramas e as caixas de bigodes (ver mais abaixo).

Aspeto 4: Forma da Distribuição e medidas integrais

As medições pontuais fornecidas por estatísticas como a média e o desvio padrão, junto com a deteção de outliers, são complementadas com a análise da forma geral que adota a distribuição da variável. O histograma é um ótimo recurso para a sua análise.

Histogramas

Os histogramas são gráficos de barras que representam como se distribuem as observações por grupos de valores, indicando as frequências absolutas dos indivíduos no eixo Y, e os valores ordenados de modo crescente e agrupados nas barras no eixo X. [4]



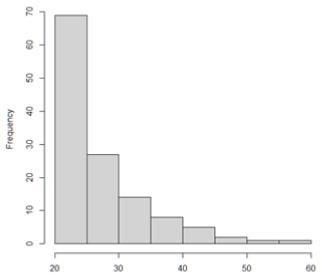
Para a variável duração, o histograma indica que

- *3 observações registram um valor entre 50 e menos de 100,*
- *6 observações registram um valor entre 100 e menos de 150,*
- *4 observações registram um valor entre 150 e menos de 200,*
- *3 observações registram um valor entre 200 e menos de 250,*

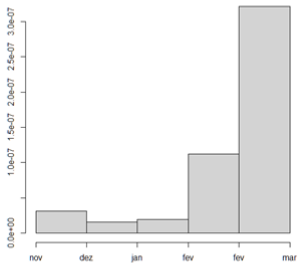
Modas e simetria

Os histogramas são um ótimo instrumento para ter uma primeira visualização da forma geral de distribuição. Vejam-se as seguintes situações

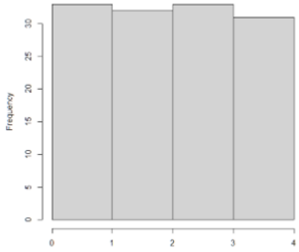
Distribuição assimétrica à direita



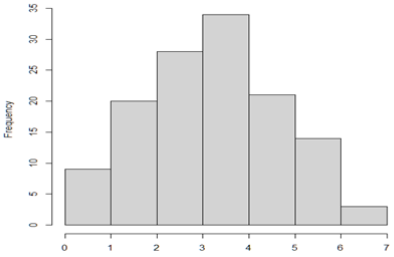
Distribuição assimétrica à esquerda



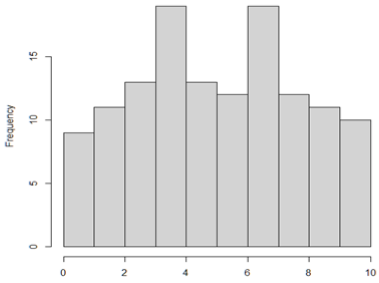
Distribuição simétrica, uniforme



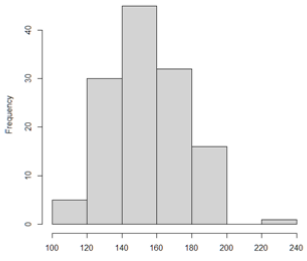
Distribuição unimodal (só tem uma moda)



Distribuição simétrica bimodal



Distribuição com outliers



[4] As pré-definições destes gráficos permitem uma análise rápida dos dados, mas é possível manipular cada um dos seus parâmetros básicos (ver abaixo "mais histogramas").

([5])

Basicamente, a forma da distribuição é dada pelo contorno superior do histograma. É a silhueta que se formaria lançando um esparguete por cima dele e deixando que fique sobre as barras (metáfora da OpenIntro). A forma de uma distribuição compõe-se basicamente de dois atributos:

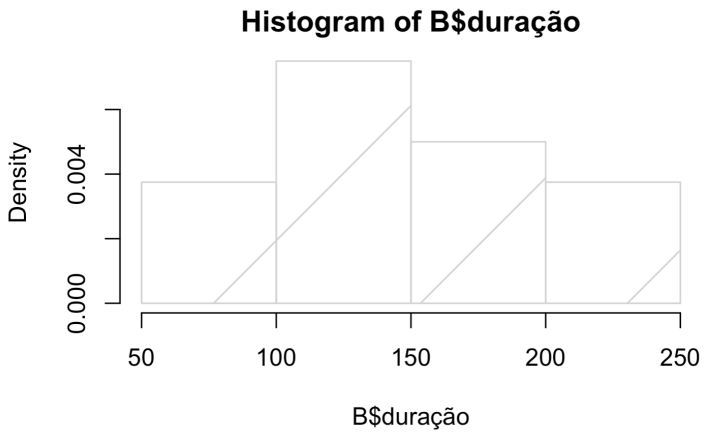
1) O número de picos ou modas da distribuição, sendo as mais comuns:
   a) a unimodal (um pico, figura X);
   b) a bimodal (figura X).
2) A simetria da dispersão do centro para os extremos, sendo as mais comuns:
   a) simétrica uniforme;
   b) simétrica em forma de sino;
   c) assimétrica à direita;
   d) assimétrica à esquerda.

Densidade

A densidade da distribuição refere-se à quantidade de observações registradas em cada valor.
No histograma, as barras mais altas representam onde os dados são relativamente mais frequentes.A área total coberta é 1.

O mesmo histograma com percentagens em vez de frequências ilustra melhor este aspeto:

```{r}
hist (B$duração, freq = FALSE, density = TRUE)
```



[5] Figuras Mafalda.

Uma tabela complementa a análise, fornecendo os valores precisos

```{r}
x <- table (B$duração)
y <- prop.table(table (B$duração))*100
```

```{r}
x
```

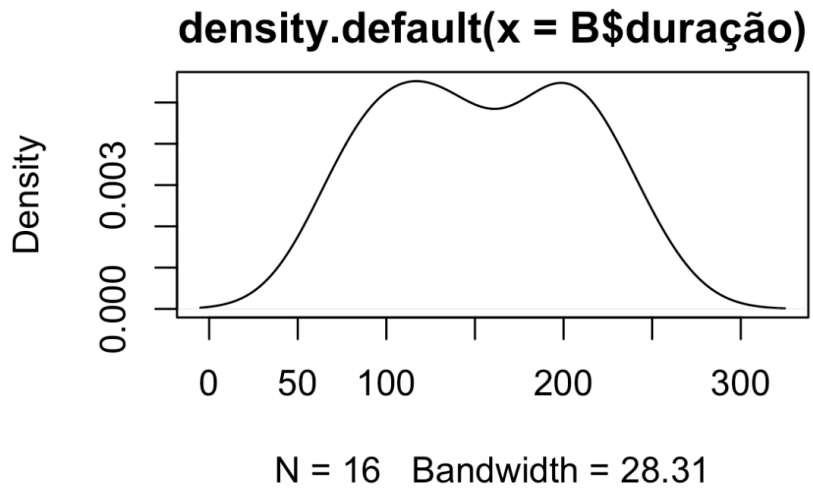
```
 80 110 120 140 150 190 200 210 230 240
  3   2   1   2   1   2   2   1   1   1
```

```{r}
y
```

```
 80    110    120    140    150    190    200    210    230    240
18.75 12.50  6.25 12.50  6.25 12.50 12.50  6.25  6.25  6.25
```

O (Kernel Density Plots)...[6]

```
plot(d) # plots the results
polygon(d, col="yellow", border="red")
```



[6] https://www.statmethods.net/graphs/density.html

Medidas integrais

O histograma fornece uma representação integral de uma distribuição porque, para além da **forma**, permite também visualizar o **centro** (intuir a média, mediana); a **dispersão** (entre mínimos e máximos, intuir os quartis); e os **outliers**. Outras medidas integrais de uso frequente são os sumários, a caixa de bigodes, e a tabela de frequências

Sumário (summary)

A função *summary* dá-nos simultaneamente informação sobre a tendência central (mediana e média), a forma (q1, q2, q3) e a dispersão (mínimos e máximos) duma distribuição.

```
summary (B$duração)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   80.0   110.0   145.0   154.4   200.0   240.0 
```

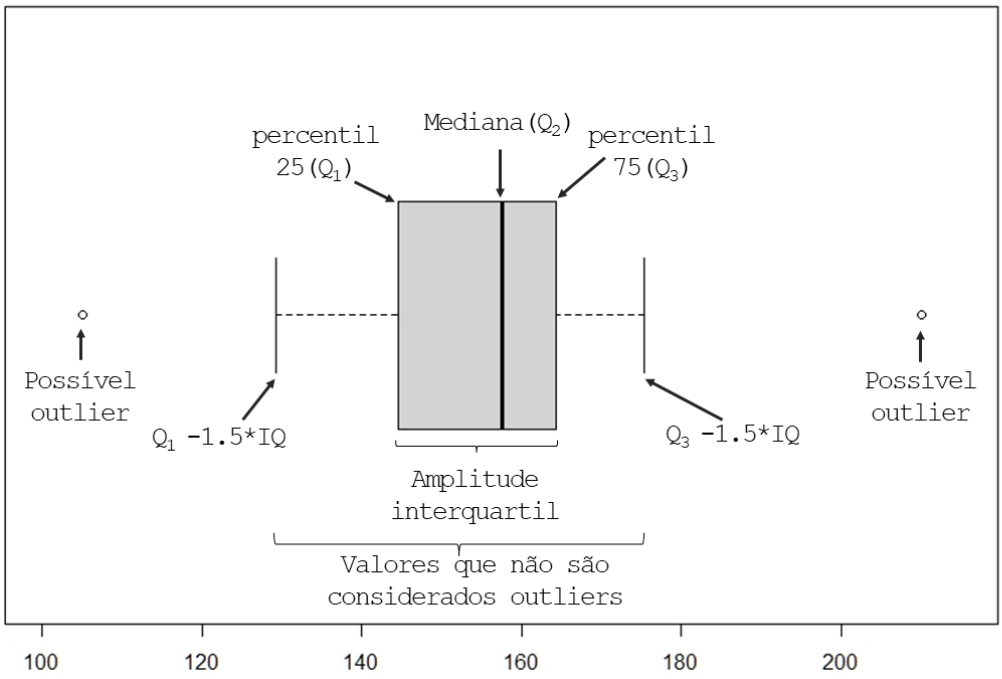
A média superior à mediana indica uma certa assimetria para a direita, mas a relativa equidistância entre o Q1 e Q3 em relação à média(44.4 e 45.6), dá ideia de simetria. Do mesmo modo, a relativa equidistância entre o valor mínimo e máximo da média (o valor mínimo está a 74.4 dias da média e o valor máximo está a 85.6).

Caixa de Bigodes (boxplot)

A caixa de bigodes pode ser considerada a representação visual do sumário, exibindo o valor mínimo, percentil 25, mediana, percentil 75, valor máximo e possíveis valores outliers.
Como vimos, um percentil é o valor abaixo do qual uma certa percentagem de dados cai. Por exemplo, se 25% das observações têm valores menores que 200, então 200 é o percentil 25 dos dados. O percentil 50 indica o valor que separa o 50% dos valores inferiores dos 50% superiores a esse valor.

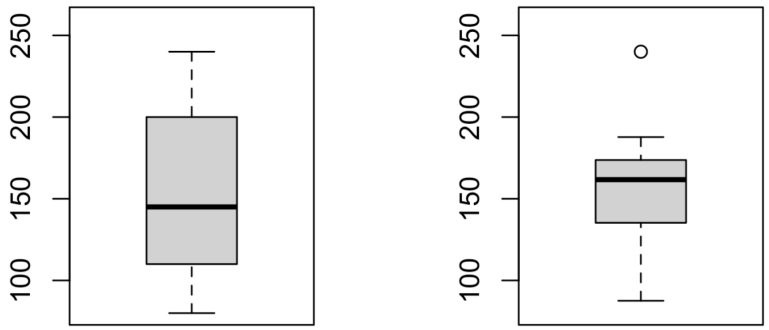
Na caixa de bigodes:

- o retângulo do gráfico representa "intervalo interquartil" (ou o percentil 75º menos o 25º).
- A linha próxima do meio do retângulo representa a mediana (ou valor médio do conjunto de dados).
- Os bigodes em ambos os lados do IQR representam os quartis mais baixos e mais altos dos dados.
- As extremidades dos bigodes representam o máximo e o mínimo dos dados, e
- os pontos individuais além dos bigodes representam valores discrepantes no conjunto de dados.



(7)

```
boxplot (B$duração)
# boxplot (B$duração, range=0)
par (mfrow =c(1,2))
boxplot (B$duração, ylim = c(80, 260))
boxplot (dura_a4, ylim = c(80, 260))
```



[7] Figuras Mafalda.

```
summary(B$duração)
```

```
   80.0   110.0   145.0   154.4   200.0   240.0
```

```
summary(dura_a4)
```

```
  Min.    1st Qu.  Median    Mean    3rd Qu.  Max.
  142.9   150.9    155.3     155.6   157.1    180.0
  87.56   135.32   161.70    157.93  172.49   240.00
```

*Aqui estamos a comparar duas variáveis com uma média semelhante, mas distribuições diferentes. Na primeira linha temos uma variável que varia entre 142.9 e 180, a amplitude interquartil é 6.2 e por isso os outliers são os valores superiores a 166.4 (157.1+1.5*6.2) e seriam os valores inferiores a 141.6 (150.9-1.5*6.2), mas como o mínimo é 142.9, não existem outliers no limite inferior, só no limite superior. Na segunda variável, os valores variam entre 87.56 e 240. O IQR é 37.17. E também só temos outliers no limite superior entre 228.245 (172.49+1.5*37,17) e 240.*
A dispersão é maior na segunda variável.
Na primeira variável a média e a mediana praticamente sobrepõem-se, enquanto que na segunda a média e a mediana estão mais afastadas, indicando uma distribuição não simétrica.

Os ***outliers***, representados pelos círculos afastados, podem ser identificados com o seguinte comando:

```
boxplot.stats(B$duração)$out
boxplot.stats(dura_a4)$out
```

Por defeito, são considerados como outliers todos os valores inferiores ao valor do:

Q1 - 1.5xIQR

ou superiores ao valor:

Q3 + 1.5xIQR.

Q1= 110 Q3= 200 IQR= 90 (200-110) Q1 - 1.5*IQR = 110-1.5*90= -55 Q3 + 1.5*IQR = 200 +1.5*90= 335	Q1= 135.32 Q3= 172.49 IQR= 37.17 (172.49-135.32) Q1 - 1.5*IQR = 135.32-1.5*37.17= 79.565 Q3 + 1.5*IQR = 172.49 +1.5*37.17= 228.245

Limite inferior para identificar outliers

Variável duração:

```
quantile(B$duração, .25) - 1.5* IQR(B$duração)
```

Variável dura_a4:

```
quantile(dura_a4, .25) - 1.5* IQR(dura_a4)
```

Limite superior para identificar outliers

Variável duração:

```
quantile(B$duração, .75) + 1.5* IQR(B$duração)
```

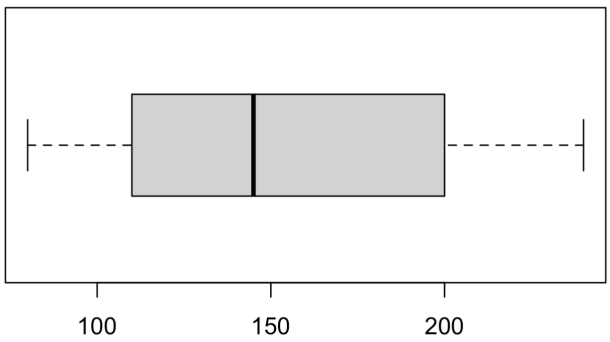
Variável dura_a4:

```
quantile(dura_a4, .75) + 1.5* IQR(dura_a4)
```

Caixa de bigodes + histogramas

Histogramas e caixas de bigodes (horizontais) podem ler-se em simultâneo:

```
par (mfrow =c(2,1))
boxplot (B$duração, horizontal=TRUE)
hist(B$duração)
```



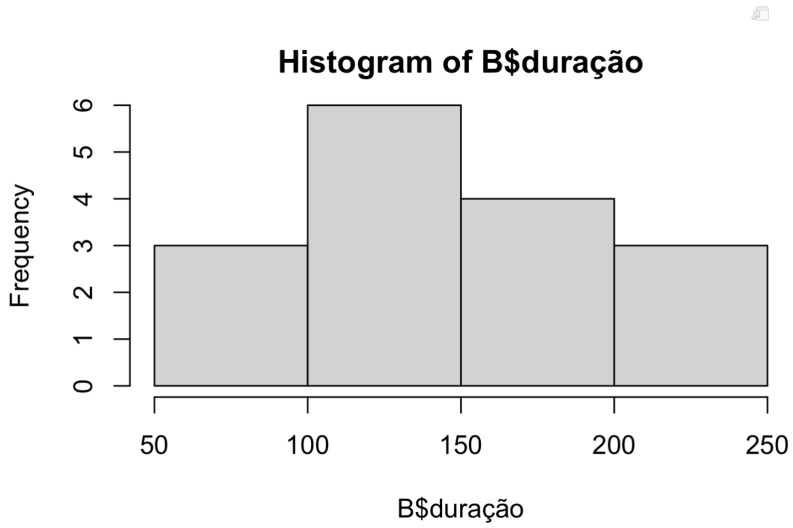


Tabela de frequências

Uma table de frequências inclui:
- Frequência Absoluta
- Frequência Relativa (proporção)
- Percentagem
- Frequência cumulativa (frequências absolutas)
- Cumulative f (frequências relativas)

Proporções :

```
t1 <-table (B$duração)

t2 <- cbind(Freq=t1,
                Cumul=cumsum(t1),
                Relative=prop.table(t1),
                Cum.Rel.=cumsum(prop.table(t1) ) )
t2
```

Percentagens:

```
t2per <- cbind(Freq=t1,
               Cumul=cumsum(t1),
               Relative=prop.table(t1)*100,
               Cum.Rel.=cumsum(prop.table(t1)
               *100))  ;
t2per
```

	Freq	Cumul	Relative	Cum.Rel.
80	3	3	0.1875	0.1875
110	2	5	0.1250	0.3125
120	1	6	0.0625	0.3750
140	2	8	0.1250	0.5000
150	1	9	0.0625	0.5625
190	2	11	0.1250	0.6875
200	2	13	0.1250	0.8125
210	1	14	0.0625	0.8750
230	1	15	0.0625	0.9375
240	1	16	0.0625	1.0000

	Freq	Cumul	Relative	Cum.Rel.
80	3	3	18.75	18.75
110	2	5	12.50	31.25
120	1	6	6.25	37.50
140	2	8	12.50	50.00
150	1	9	6.25	56.25
190	2	11	12.50	68.75
200	2	13	12.50	81.25
210	1	14	6.25	87.50
230	1	15	6.25	93.75
240	1	16	6.25	100.00

4. Exercícios

Secção PANORAMA

1) Identificar e escolher 5 artigos vinculados a um tópico do seu interesse, que utilizem análise estatística.
2) Para cada um dos 5 artigos escolhidos, identificar que tipo de análise aplicam (descritivo, explicativo) e que técnicas e ferramentas (coeficientes, regressões, caixa de bigodes, etc.).
3) Para 2 dos 5 artigos escolhidos,
   a) identificar e descrever a base de dados utilizada (unidade de observação, número de observações, variáveis) e
   b) reconstruir o codebook incluindo pelo menos 5 variáveis.

Secção UMA QUANTITATIVA

4) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável numérica. Transcrever essa descrição.

Secção UMA QUALITATIVA

5) Para 1 dos 5 artigos escolhidos, identificar a descrição de uma variável qualitativa. Transcrever essa descrição.

Secção DUAS QUANTITATIVAS

1) Para 1 dos 5 artigos escolhidos, identifique a variável independente e a variável dependente. Classifique-as (e.g., categórica, numérica).
2) Agora escolha um artigo com uma variável independente numérica e uma variável dependente numérica. Calcule o coeficiente de correlação entre as duas variáveis.
3) Por fim, calcule um modelo de regressão linear e interprete os resultados.

5. Referências

Bryman et al. (2021). Bryman's Social Research Methods, Oxford University Press.

Diez et al (2021). OpenIntro Statistics. https://stats.libretexts.org/@go/page/270

Judd, C. M., McClelland, G. H., & Ryan, C. S. (2017). Data analysis: A model comparison approach to regression, ANOVA, and beyond. Routledge.

6. Apêndice (R)

O R tem muitos tipos de vetores numéricos (ver a tabela) e nem sempre é tão certinho na sua utilização. Por exemplo, pode ler como numéricas a variáveis dicotómicas categóricas que utilizem 0 e 1 ou pode ler como variáveis de texto se algum participante escrever um número por extenso

Tabela X: Diferentes tipos de vetores (variáveis) numéricos no R

integer	1, 2, 3, 4
double	1.1 , 1.2 , 1.3, 1.4
numeric	

Gráficos com Pacote Base

Histogramas

As pré-definições dos histogramas permitem uma análise rápida dos dados, mas é possível editar cada dos seus elementos.

```
hist(B$duração)
```

- bin [50-100) = 3 (0, 1, 1)
- bin [100-150)= 6 (,,,)
- bin [150-200)= 4 (...)
- bin [200-250)= 3 (...)

```
hist(B$duração, right=F)   # valor da direita no incluido (o 15 do bin 3)
```

```
hist(B$duração, 8)
```

```
hist(B$duração, 15)
```

Histogramas múltiplas

sem correção dos eixos

```
par (mfrow =c(3,1))
```

```
hist(B$duração)
hist(dura_a2)
hist(dura_a3)
```

com correção do eixo Y

```
par (mfrow =c(3,1))       #

hist(B$duração, ylim= c(0,8))

hist(dura_a2,     ylim= c(0,8))

hist(dura_a3,     ylim= c(0,8))

```

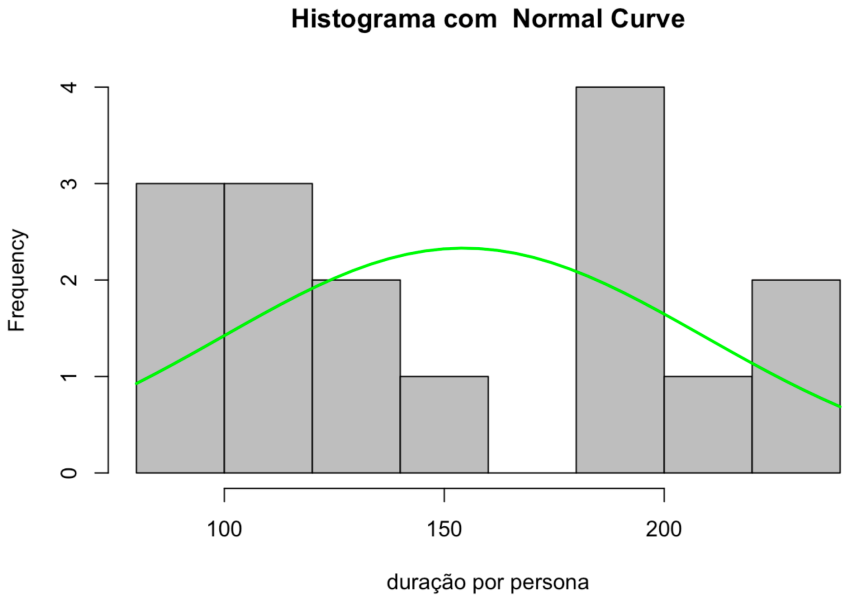
com correção dos eixos Y e X

```
par (mfrow =c(3,1))       #

hist(B$duração, ylim= c(0,10), xlim=c(0,50))
      abline (v=median(dura_a2), col="red")

hist(dura_a2, ylim= c(0,10), xlim=c(0,50))
     abline (v=median(dura_a2), col="red")

hist(dura_a3, ylim= c(0,10), xlim=c(0,50))
     abline (v=median(dura_a3), col="red")

```

Histograma com curva de normalidade

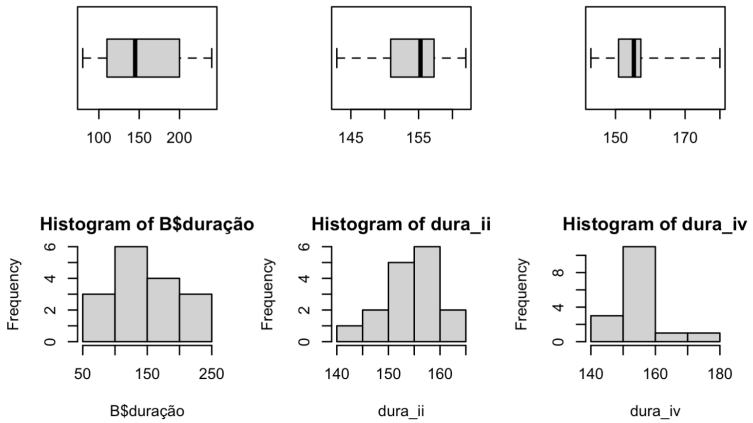
```
h<-hist(B$duração, breaks=10, col="grey", xlab="duração por
persona",
   main="Histograma com  Normal Curve")

xfit<-seq(min(B$duração),max(B$duração),length=40)

yfit<-dnorm(xfit,mean=mean(B$duração),sd=sd(B$duração))

yfit <- yfit*diff(h$mids[1:2])*length(B$duração)

lines(xfit, yfit, col="green", lwd=2)
```

"Histograms can be a poor method for determining the shape of a distribution because it is so strongly affected by the number of bins used."



Histograma e caixa de bigodes múltiplas

```r
par (mfrow =c(2,3))
boxplot (B$duração, range=0, horizontal=TRUE)
boxplot (dura_ii, range=0, horizontal=TRUE)
boxplot (dura_a4, range=0, horizontal=TRUE)
hist(B$duração)
hist(dura_ii)
hist(dura_a4)
```



Stemplots

Os stemplots oferecem uma representação numérica da frequência dos valores observados, semelhante a um histograma vertical.

```
stem(B$duração)
  The decimal point is 2 digit(s) to the right of the |
  0 | 888          # ler: 80, 80, 80
  1 | 11244        # ler: 110, 110, 112, 140, 140
  1 | 599          # ler: 150, 190, 190
  2 | 00134        # ler: 200, 200, 210, 230, 240
```

O resultado indica que,

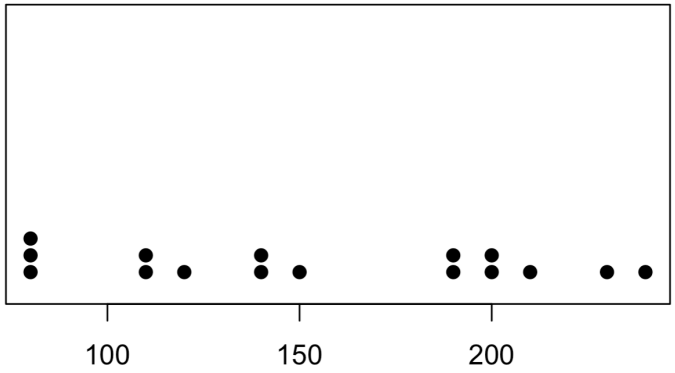
- Três observações registram valores de 80
- 2 observações registram um valor de 110
- 1 observação registra um valor de 112,
- etc.

Doplots

Os [**dotplots**] oferecerem uma representação visual similar aos stemplots, mas invertida, amostrando a frequência absoluta de observações (eixo X vertical) por valores registrados ordenados de menor a maior (eixo Y horizontal). O resulta abaixo para a nossa variável indica que

- 3 observações registram o valor mínimo (80),
- 2 observações registram uma duração algo superior a 100 dias (as duas observações de 110,
- 1 observação registra uma valor de 112;
- etcétera.

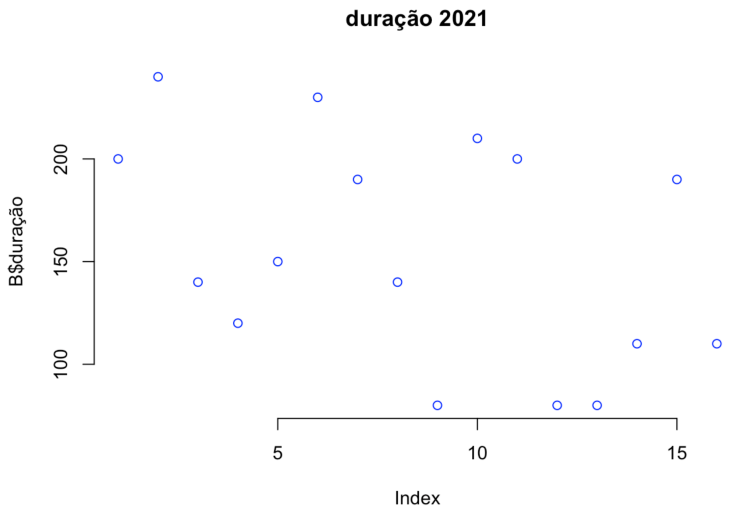
- [grafico: agregar todos los valores de Y?)



Estas ferramentas nos da uma imagem geral da distribuição dos valores, util para N pequenos mas difícil de utilizar com amostras largas. Para estes casos e mais recomendável a utilização dos histogramas.

Plots

```
plot (B$duração,
      frame=FALSE,
      col = "blue",
       main = "duração 2021")
```



Plots com etiquetas

[Labels_to_Points_in_Scatterplot](https://rpubs.com/RatherBit/188960)

```
plot(B$duração ~B$lealdade, col="lightblue", pch=19,
cex=2,data=cars)
text(B$duração ~B$lealdade, labels=dist,data=cars, cex=0.9,
font=2, pos=4)
```