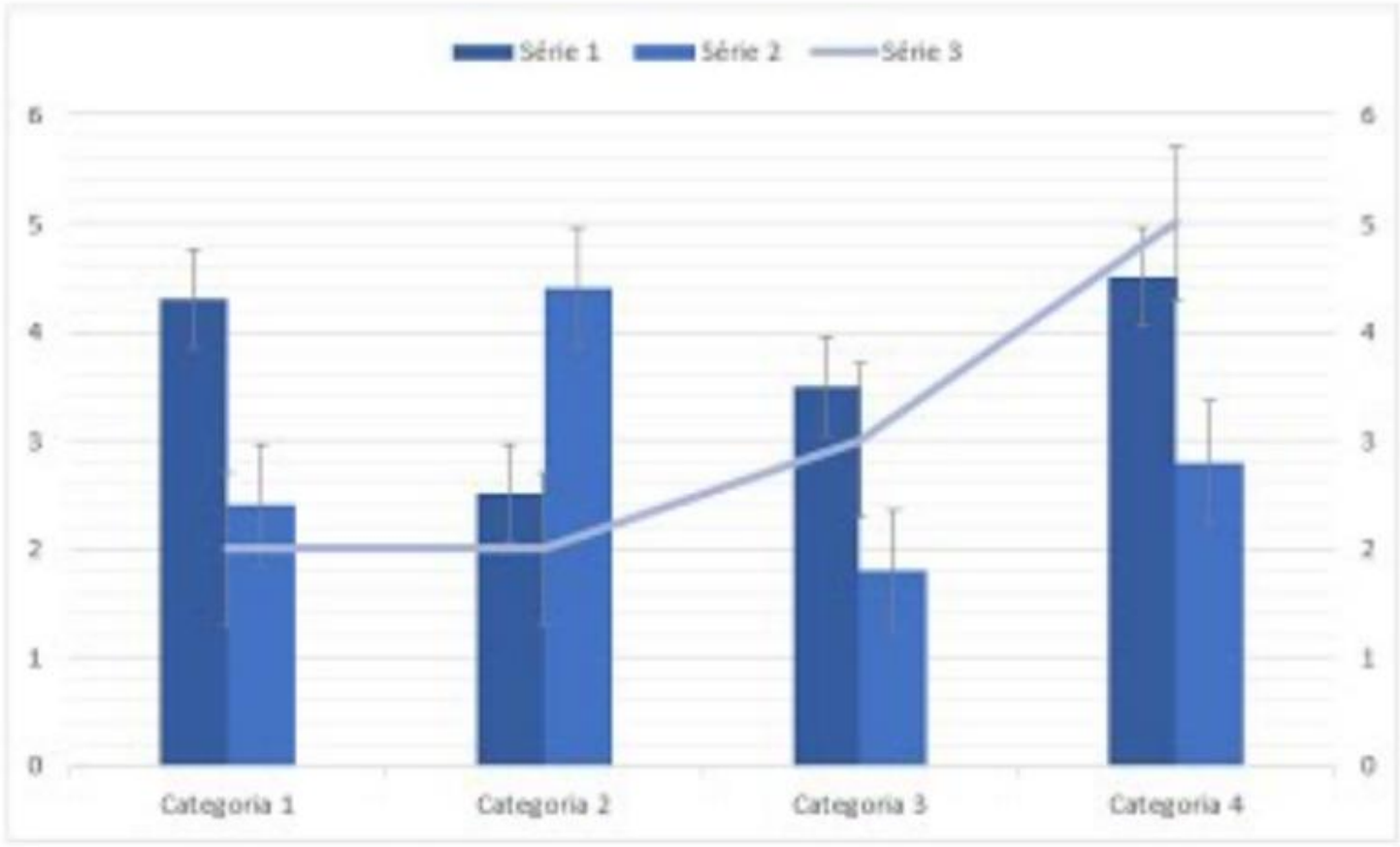
ESTATÍSTICA BÁSICA

Fonte: http://www.portalaction.com.br/estatistica-basica.

Estatística Básica

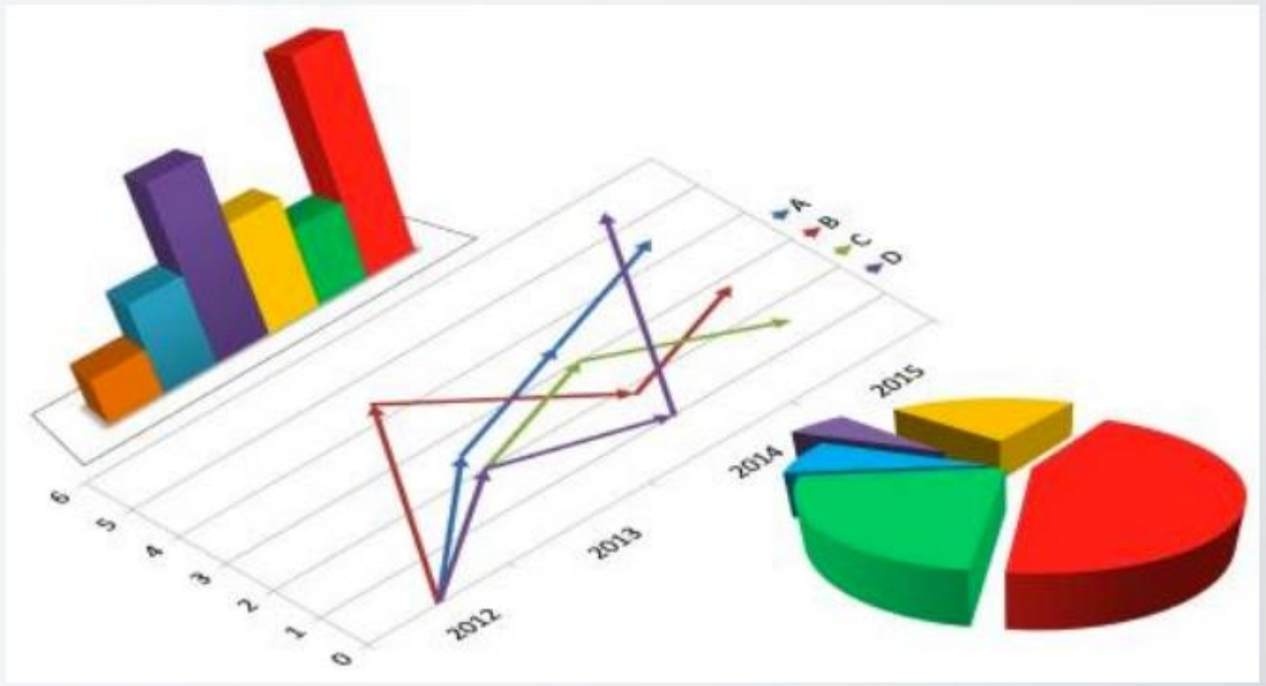


NOÇÕES BÁSICAS DE ESTATÍSTICA

Utilizando o R como Ferramenta

INTRODUÇÃO

- A Estatística (ou ciência Estatística) é um conjunto de técnicas e métodos de pesquisa que entre outros tópicos envolve o **planejamento** do experimento a ser realizado, a **coleta qualificada** dos dados, a **inferência** , o **processamento**, a **análise** e a **disseminação** das informações.



INTRODUÇAO

- Na estatística trabalhamos com dados, nos quais podem ser obtidos por meio de uma **amostra** da **população** em estudo.

  ✓ ***População***

  → conjunto de elementos que tem pelo menos uma característica em comum. Esta característica deve delimitar corretamente quais são os elementos da população.

  ✓ ***Amostra*** :

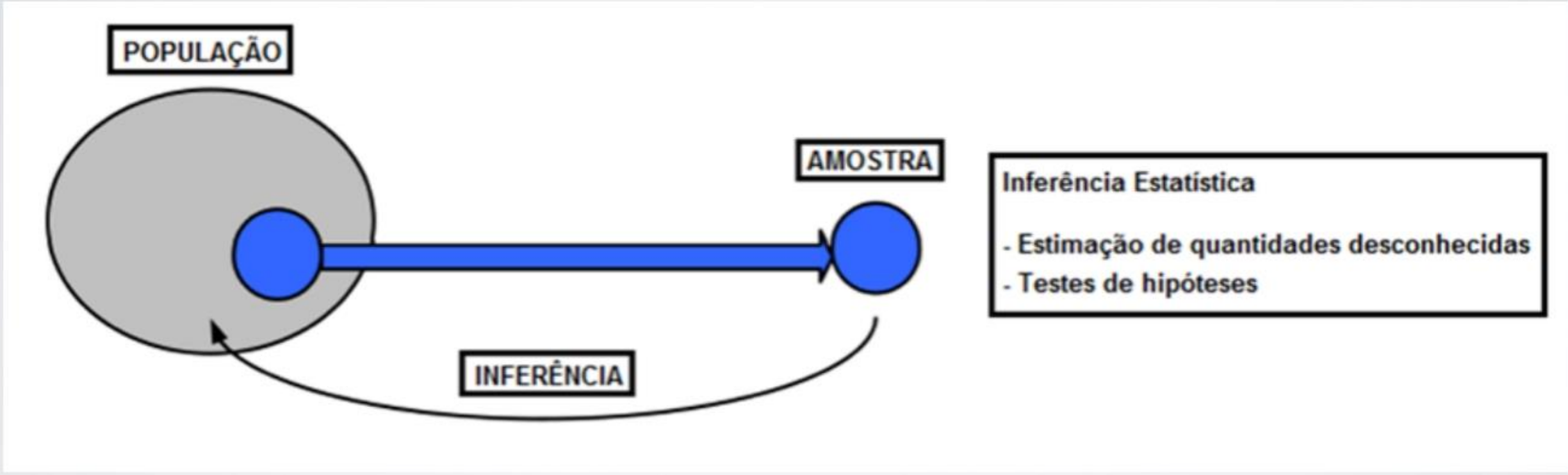
  → subconjunto de elementos de uma população, que são representativos para estudar a característica de interesse da população.

  → A seleção dos elementos que irão compor a amostra pode ser feita de várias maneiras e irá depender do conhecimento que se tem da população e da quantidade de recursos disponíveis.

✓ ***Sempre que resumimos um conjunto de dados, perdemos informação sobre o mesmo, pois condensamos as observações originais.***

✓ ***Entretanto, esta perda de informação é pequena se comparada ao ganho que se tem com a clareza da interpretação proporcionada*** .

CONCEITOS BÁSICOS



Inferência estatística é um ramo da Estatística cujo objetivo é fazer afirmações a partir de um conjunto valores representativo (amostra) sobre um universo (população), assume-se que a população é muito m que o conjunto de dados observados, a amostra.

ESTATÍSTICA-FUNDAMENTOS

- A Estatística utiliza a **variabilidade** presente nos dados para obter para obter informação sobre o **comportamento** de processos e produtos.

- *A variabilidade está presente em todo lugar .*
    - *A posição de uma motocicleta estacionada em uma garagem não é a mesma ao longo dos dias.*
    - *A posição da moto apresenta uma variação.*
    - *Avaliar as variações e obter informações através delas .*

ESTATÍSTICA-FUNDAMENTO S

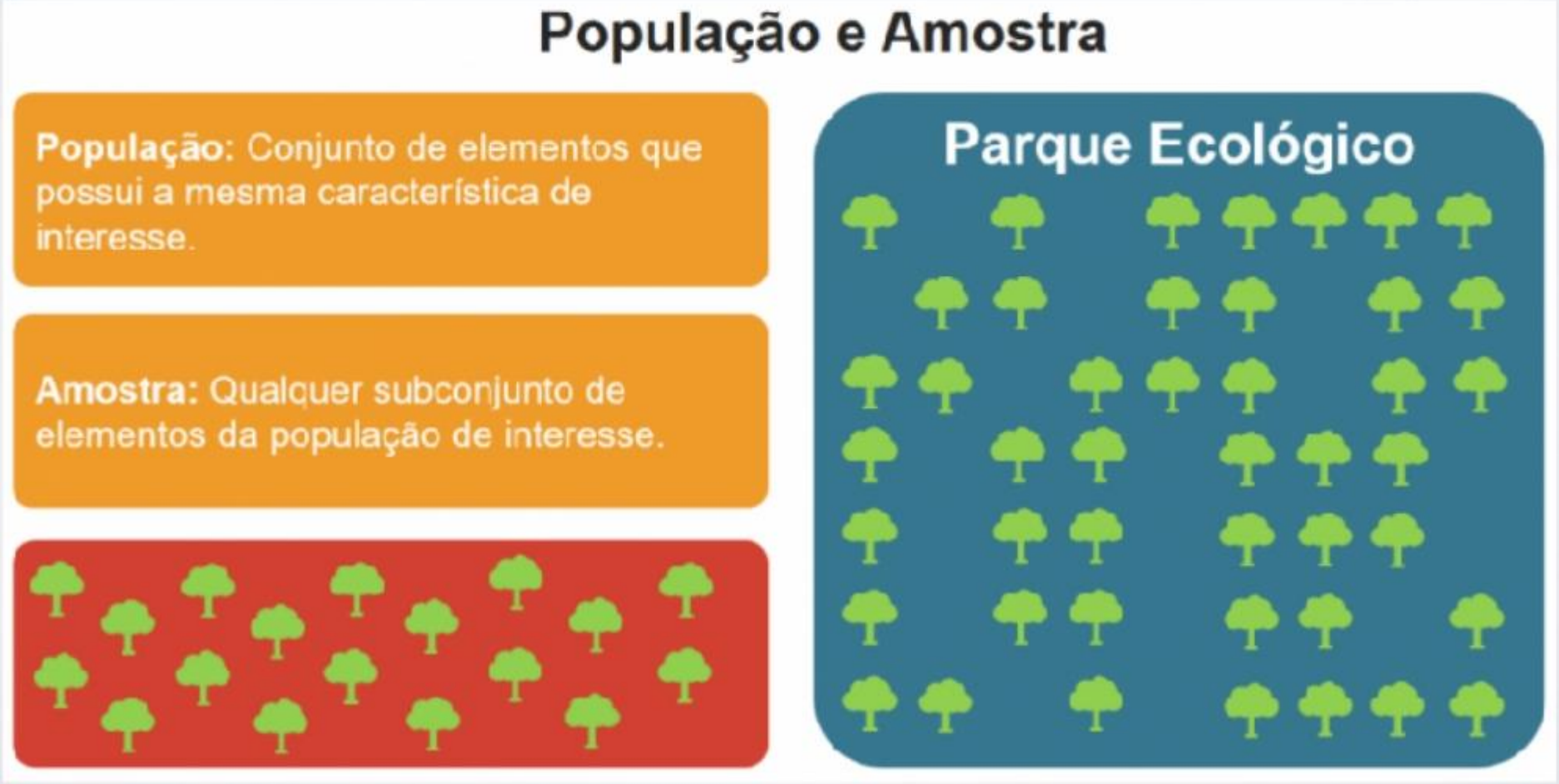
- A aplicação de técnicas estatísticas envolve várias etapas.
- **Importante** : definir o problema que estamos tratando!
- Na sequência, planejar e executar a coleta dos dados.
- A partir dos dados:
    - *aplicar técnicas estatísticas*
    - *extrair informação sobre o problema e sua solução.*

COLETA DE DADOS

- A qualidade da solução do problema está diretamente relacionada com a qualidade dos dados obtidos.
- Não se deve coletar dados sem que antes se tenha definido claramente o problema ou situação a ser enfrentada, bem como os objetivos com relação aos mesmos!
- Devem ser utilizados métodos adequados para coleta de dados de acordo com o problema estudado.

POPULAÇÃO E AMOSTRA

- **População**
  - Agregado de elementos (finitos ou não) do qual desejamos obter informações sobre algumas de suas características
- **Amostra**
  - Uma parcela de uma população que pode conter informações sobre esta população



COLETA DE DADOS
CLASSIFICAÇÃO DOS DADOS OU VARIÁVEIS

- Uma variável é uma característica específica da população
  - taxa selic, idade, sexo ou preferência partidária.
- Cada variável consiste de um conjunto de categorias que descrevem a natureza e o tipo de variação associados com esta característica.
- Algumas variáveis podem ter inúmeras categorias de resposta, dependendo do objetivo e foco do problema em questão.

ESCALA QUALITATIVA

- A escala qualitativa está relacionada com a **descrição** da característica.
- A variável *não pode ser medida* mas *pode ser observada*.
  - ✓ tipos de defeitos em um televisor no final da linha de montagem (pequeno, médio, grande)
  - ✓ nível educacional (ensino fundamental, ensino médio, graduação e pós-graduação)
  - ✓ sexo, cor dos olhos, fumante/não fumante, doente/sadio, sim/não

ESCALA QUANTITATIVA

- A escala quantitativa está relacionada com características que podem ser medidas ou contadas.
- São caracteristicas numéricas.
  - ✓ massa de um comprimido
  - ✓ preço de um ativo no mercado financeiro
  - ✓ número de defeitos em um carro
  - ✓ quantidade de úlcera por pressão em um paciente de um hospital

ESCALAS QUALI/QUANTI

- *As distinções são menos rígidas do que a descrição insinua…*
- *Uma variável originalmente quantitativa pode ser coletada de forma qualitativa.*
    - a variável idade, **medida em anos completos** , é *quantitativa* (contínua);
    - se for informada apenas a **faixa etária** (0 a 5 anos, 6 a 10 anos, etc...), é *qualitativa* (ordinal).
- *Nem sempre uma variável representada por números é quantitativa.*
    - número do telefone de uma pessoa
    - número da casa
    - número de sua identidade.
- Peso dos lutadores de boxe
    - quantitativa (contínua) se trabalhamos com o valor obtido na balança (89,5 kg)
    - qualitativa (ordinal) se o classificarmos nas categorias do boxe
        - peso-pena
        - peso-leve
        - peso-pesado
- Sexo do indivíduo
    - Registrado na planilha de dados como 1 se macho e 2 se fêmea
    - ***Isto não significa que a variável sexo passou a ser quantitativa!***

EXEMPLO

		Mídia Social no Ensino Superior: Dicionário das Váriáveis de Pesquisa		
Seção do Survey	**Nome da Variável**	**Descrição da Variável**	**Tipo de variável e nível de medição**	**Valores e Rótulos**
I. Dados Censitários				
	genero	Gênero do respondente	Tipo: qualitativa Nível de medição: nominal	0. Prefiro não declarar 1.Masculino 2.Feminino
	idade	Faixa de idade	Tipo: qualitativa Nível de medição: ordinal	1.Entre 16 e 20 anos 2.Entre 21e 25 anos 3.Entre 26 e 30 anos 4.Entre 30 e 35 anos 5.Entre 36 e 40 anos 6.Acima de 40 anos
	profal	Característica do respondente	Tipo: qualitativa Nível de medição: nominal	1.Professor 2.Aluno 3.Professor e aluno 4.Prefiro não declarar
	Outro	Outra característica não atendida pela variável profal	Tipo: qualitativa	Texto
	trabalha	Situação trabalhista atual	Tipo: qualitativa Nível de medição: nominal	0.Desempregado 1.Jornada parcial 2.Jorna da integral 3.Estagiário 4.Trabalha por conta própria
	idadefilho	Se tem filhos, qual a faixa etária dos mesmos	Tipo: qualitativa. Nível de medição: nominal	0.Sem filhos 1.Entre 0 e 6 anos 2.Entre 7 e 15 anos 3.Entre 16 e 20 anos 4.Acima de 20 anos 5.1 e 2 (0 a 6/7 a 15)
VI. Avaliação de recursos				
	envinfo	Envio de informações da escola para os pais	Tipo: qualitativa. Nível de medição: ordinal.	1.Muito pobre 2.Pobre 3.Indiferente 4.Bom 5.Excelente

PLANEJAMENTO DE COLETA DE DADOS

- Para estudar adequadamente uma população através de uma amostra , deve-se planejar a coleta de dados.
- Com este objetivo, formulam-se algumas perguntas, tais como:
    - Com que frequência ocorre(m) o(s) problema(s)?
    - Quais são as causas potenciais do problema?
- Um bom planejamento para coleta de dados deve considerar:
    - Qual a pergunta a ser respondida?
    - Como comunicar a resposta obtida?
    - Qual ferramenta de análise pretendemos usar e como utilizar os resultados?
    - Qual tipo de dado é necessário para utilizar as ferramentas desejadas e responder a pergunta?
    - Como coletar esses dados com o mínimo de esforço e erro?
    - Onde acessar estes dados?
    - Quem pode nos fornecer os dados?
    - Qual o período em que os dados serão coletados?

PLANEJAMENTO DE COLETA DE DADOS

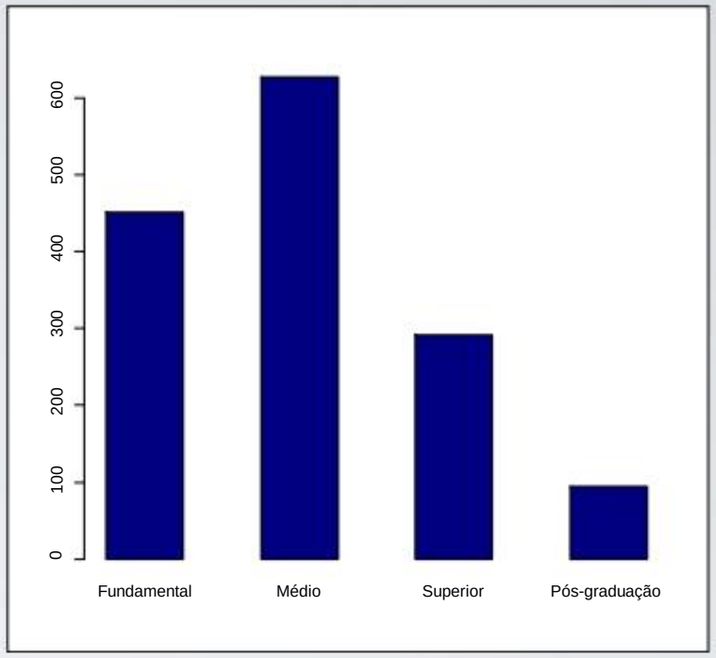
- Tendo as respostas para estas perguntas, deve-se:

    - Construir uma metodologia para nos certificar de que todas as informações estão definidas;
    - Coletar os dados de forma consistente e honesta;
    - Certificar-se de que existe tempo suficiente para a coleta de dados;
    - Definir quais informações adicionais serão necessárias para estudos futuros, referências ou reconhecimento.

PLANEJAMENTO DE COLETA DE DADOS

EXPOSIÇÃO DOS DADOS

Antes da exposição dos dados coletados é necessário que se faça um trabalho de revisão e correção nos dados coletados na tentativa de eliminar possíveis enganos na elaboração do relatório.
Inicialmente, os dados podem ser classificados como "qualitativos" ou "quantitativos".



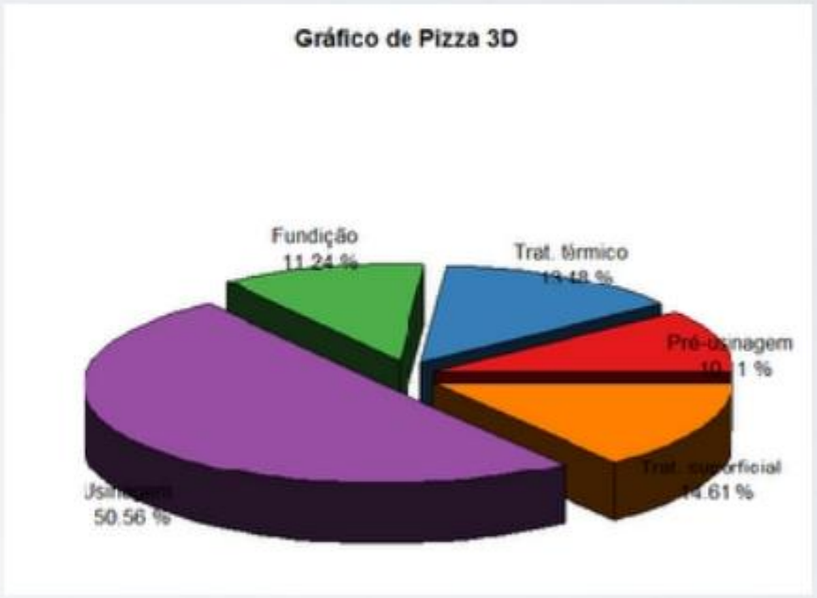
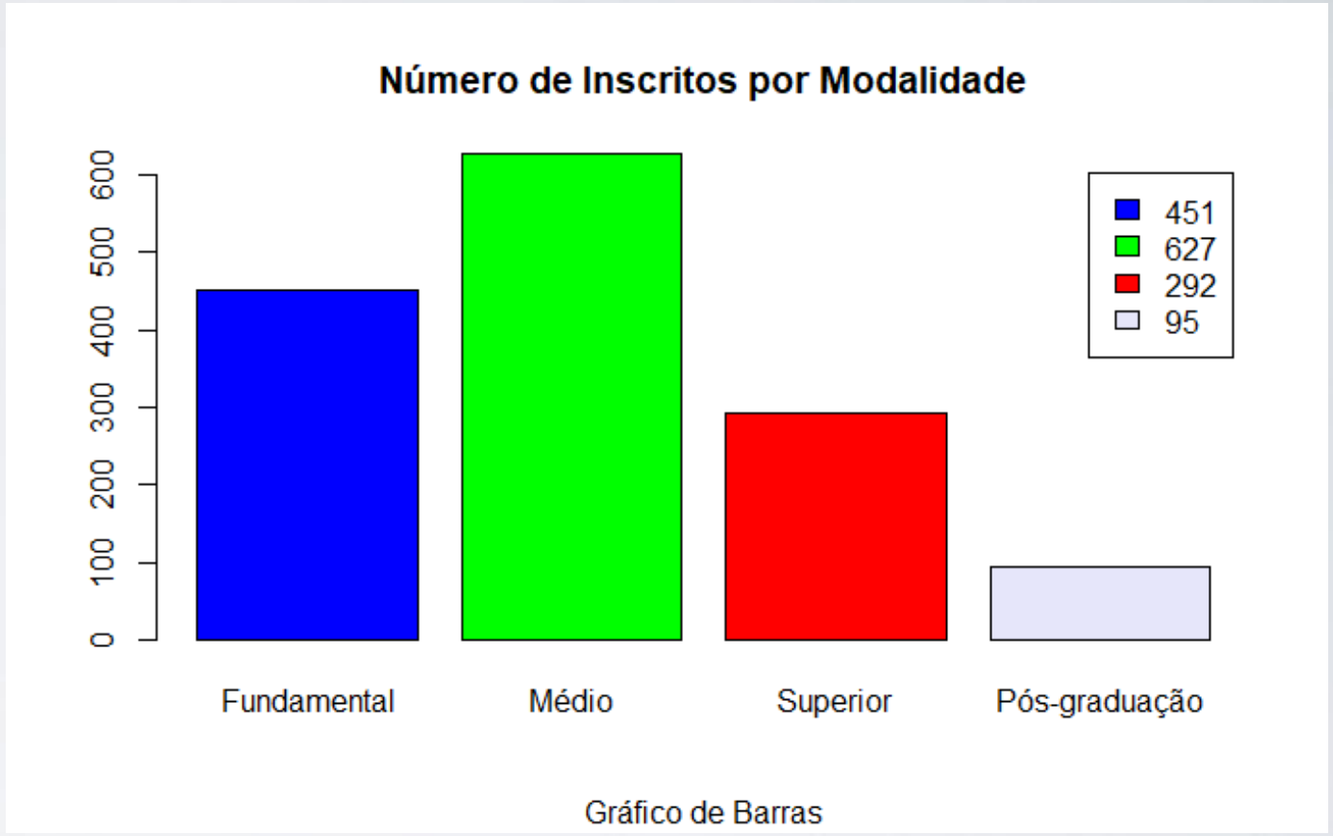


GRÁFICO DE BARRAS

- Apresenta dados categorizados em barras retangulares correspondentes a cada categoria.
- É proporcional ao número de observações na respectiva categoria.
- Utilizado para realizar comparações entre as categorias de uma variável qualitativa ou quantitativa discreta.
- Este gráfico pode ser utilizado na vertical ou horizontal.

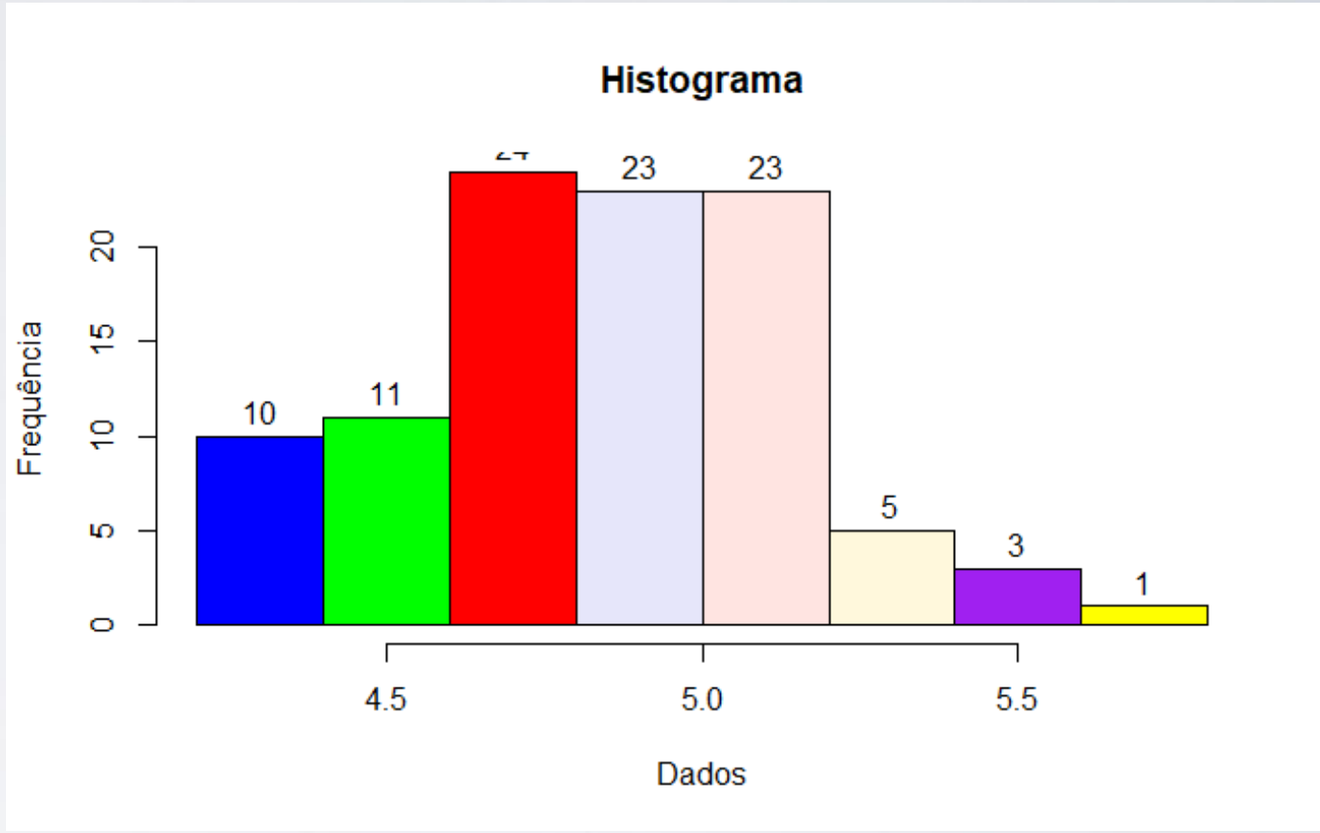



```
library(xlsx)
eb <- read.xlsx("estatistica_basica.xlsx", sheetName = "Plan1")

head(eb)
#       escolaridade inscritos
#1    Fundamental         451
#2          Médio         627
#3       Superior         292
#4 Pós-graduação          95
#
barplot(eb$inscritos, names.arg = eb$escolaridade, col = c("blue",
"green", "red", "lavender"),
        legend.text = eb$inscritos, sub = "Gráfico de Barras",  main =
"Número de Inscritos por Modalidade")
```

HISTOGRAMA

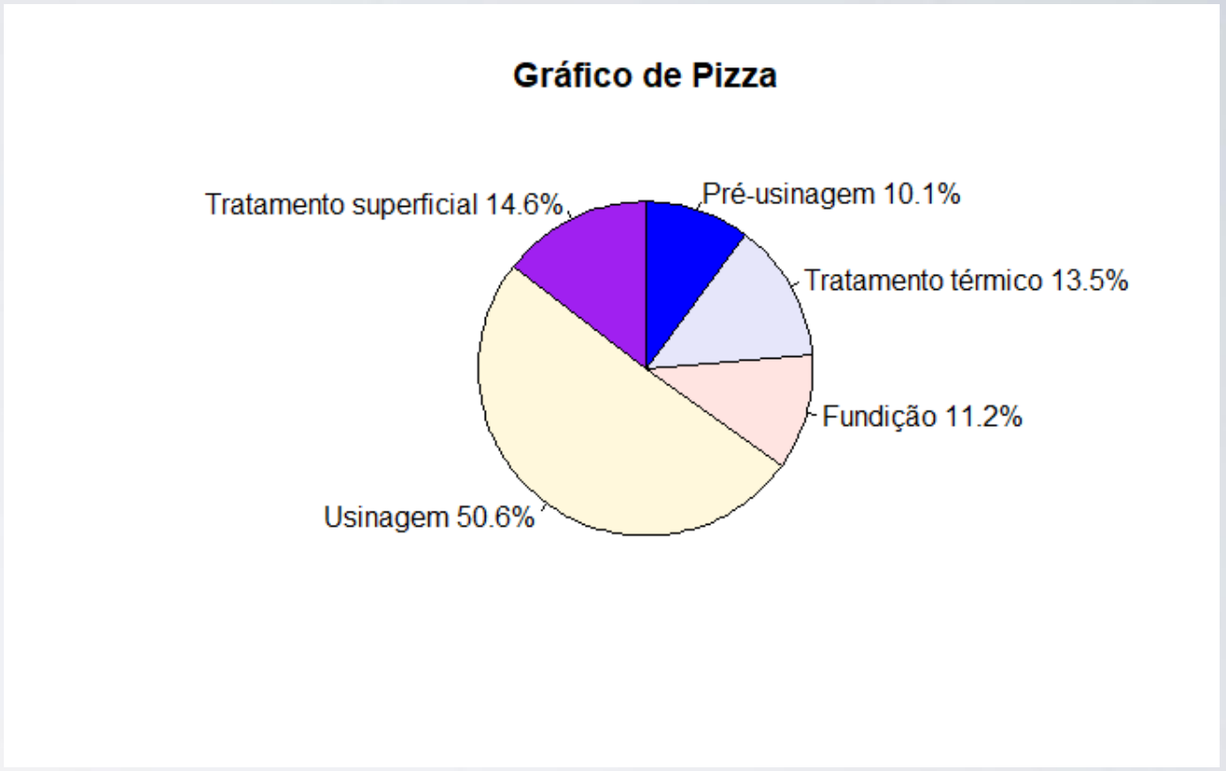
- Representação gráfica da distribuição de frequências de um conjunto de dados quantitativos contínuos.
- Pode ser um gráfico por valores absolutos ou frequência relativa ou densidade.
- Assim, ao construir o histograma, cada retângulo deverá ter área proporcional à frequência relativa ou absoluta correspondente.
- No caso em que os intervalos são de tamanhos (amplitudes) iguais, as alturas dos retângulos serão iguais às frequências relativas (ou iguais às frequências absolutas) dos intervalos correspondentes.




```
eh <- read.xlsx("estatistica_basica.xlsx", sheetName = "Plan2")
head(eh)
#  eixos
#1   4.8
#2   4.9
#3   5.1
#4   4.8
#5   5.1
#6   4.9
hist(eh$eixos,  main = "Histograma", labels = TRUE,
     col = c("blue", "green", "red", "lavender", "mistyrose",
             "cornsilk", "purple", "yellow"),
     ylab = "Frequência",
     xlab = "Dados")
```

GRÁFICO DE PIZZA

- Também conhecido como gráfico de setores ou gráfico circular.
- Diagrama circular onde os valores de cada categoria são proporcionais às respectivas frequências.
- Este gráfico pode vir acompanhado de porcentagens.
- Utilizado para dados qualitativos nominais.
- Para construir um gráfico tipo pizza é necessário determinar o ângulo dos setores circulares correspondentes à contribuição percentual de cada valor no total.

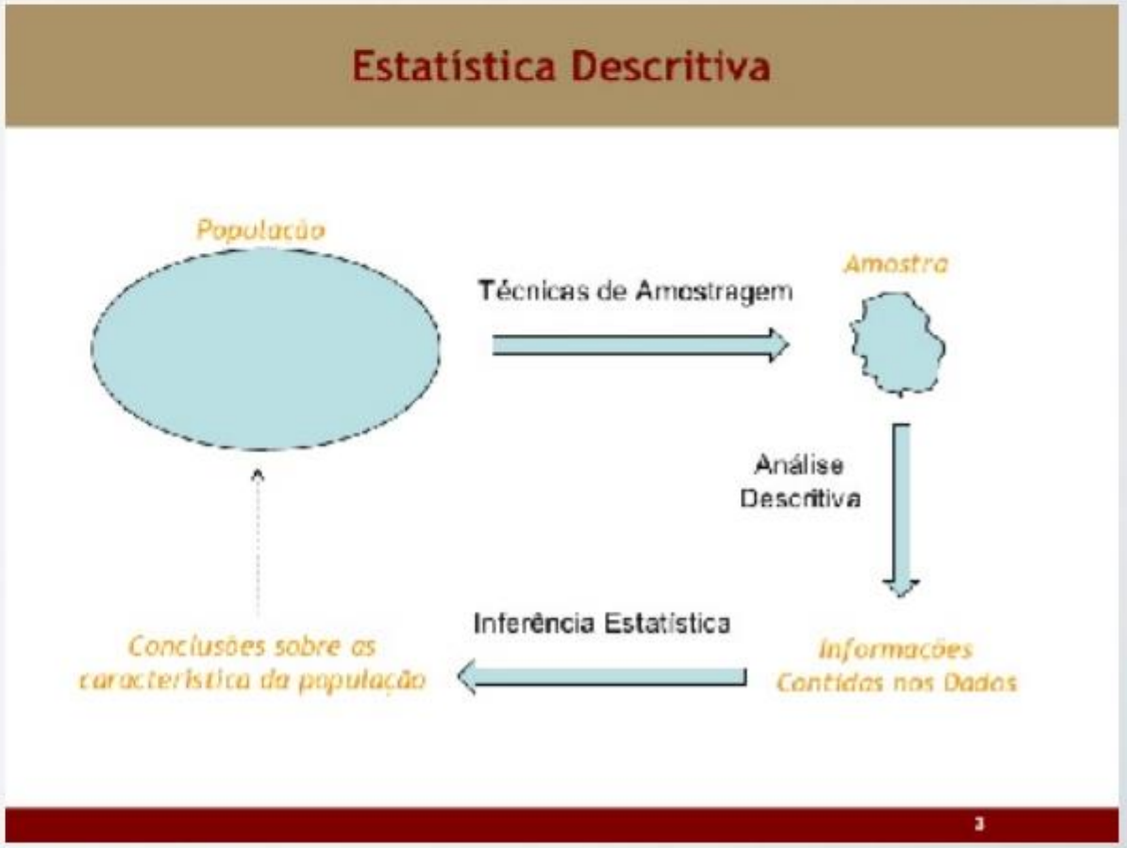



```
# Gráfico Pizza
ep <- read.xlsx("estatistica_basica.xlsx", sheetName = "Plan3")
head(ep)
#                          ccusto defeitos
#1            Pré-usinagem                9
#2 Tratamento térmico          12
#3                    Fundição          10
#4                   Usinagem          45
#5 Tratamento superficial       13
#
# Calculando a porcentagem
#Rótulos para a figura
lbls <- ep$ccusto
pct <- round(ep$defeitos/sum(ep$defeitos)*100, digits=1)
lbls <- paste(lbls, pct) # acrescentar os percentuais aos rótulos
lbls <- paste(lbls,"%",sep="") # acrescentar o símbolo "%" aos rótulos
labels
lbls
pie(ep$defeitos, labels = lbls, edges = 200, radius = 0.8,
    clockwise = TRUE, angle = 45, main = "Gráfico de Pizza",
    col = c("blue", "lavender", "mistyrose", "cornsilk", "purple"),
    lty = NULL)
```

ESTATÍSTICAS DESCRITIVAS

- Números que resumem e descrevem o conjuntos de dados.
- Apenas "descrevem" os dados, não representam generalizações da amostra para a populaçã
- Medidas básicas de análise descritiva:
  - *medidas de posição*
  - *medidas de dispersão*
  - *quartis*
  - *coeficiente de assimetria*
  - *coeficiente de curtose*

- *A técnica utilizada para estender conclusões da amostra para a população é a **inferência** !*



MEDIDAS DE POSIÇÃO - MÉDIA

- Estatísticas que representam uma série de dados orientando-nos quanto à posição da distribuição em relação ao eixo horizontal do gráfico da curva de frequência.
- x: valor de cada indivíduo da amostra.
- $\bar{x}$: média amostral.
- **n** : tamanho amostral.
- **Média populacional**
  - Calculada somando-se todos os valores da população e dividindo o resultado pelo total de elementos da população.
  - Numa população de N elementos, a média populacional é dada por $\mu = \frac{x_1 + \ldots . x_N}{N}$
- **Média Amostral , Média Aritmética ou Média**
  - Calculada somando-se os valores das observações da amostra e dividindo-se o resultado pelo número de valores.
  - Assim, a média amostral é dada por $\bar{x} = \frac{x_1 + \ldots . x_N}{N}$

$$\overline{x} = \frac{4,5 + 4,6 + 4,5 + 4,4 + 4,5}{5} = 4,5$$

```
xma = (4.5+4.6+4.5+4.4+4.5) / 5
xma

mean(4.5,4.6,4.5,4.4,4.5)
```

```
# Média
ma <- read.xlsx("estatistica_basica.xlsx", sheetName =
"Plan4")
head(ma)
#  dados
#1    4.5
#2    4.6
#3    4.5
#4    4.4
#5    4.5
xma = mean(ma$dado)
xma
#[1] 4.5
```

MEDIDAS DE POSIÇÃO - MEDIANA

- Paracalcular a mediana devemos, em primeiro lugar, ordenar os dados do menor para o maior valor.
- Seo número de observações for ímpar, a mediana será a observação central.
- Seo número de observações for par, a mediana será a média aritmética das duas observações centrais.
- Notação: $\tilde{X}$

Consideremos os seguintes dados correspondentes aos comprimentos de 8 rolos de fio de aço: 65, 72, 70, 72, 60, 67, 69, 68

Ordenando os valores temos: 60, 65, 67, 68, 69, 70, 72, 72.

Como o número de observações é par, a mediana é dada pela média dos dois valores centrais que são 68 e 69, isto é $\tilde{X} = \frac{68 + 69}{2} = 68,5$

Classificados:

```
l istamd=sort(md$dados, decreasing=F)
listamd
#[1] 60 65 67 68 69 70 72 72
xbarra = (68 + 69) / 2
xbarra
#[1] 68.5
```

```
# Mediana
md <- read.xlsx("estatistica_basica.xlsx",
sheetName = "Plan5")
head(md)
#  dados
#1     65
#2     72
#3     70
#4     72
#5     60
#6     67
mediana = median(md$dados)
mediana
#[1] 68.5
```

MEDIDAS DE POSIÇÃO - MODA

- É o valor que apresenta a maior frequência.
- R não possui uma função embutida padrão para calcular a moda.
- Então, criamos uma função de usuário para calcular o modo de um conjunto de dados em R.
- Esta função leva o vetor como entrada e dá o valor do modo como saída.
- Não é frequentemente utilizada, mas é muito útil para *dados categóricos e discretos* .

```
# Moda
moda <- read.xlsx("estatistica_basica.xlsx", sheetName = "Plan5")
head(moda)
# Create the function.
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
##[1] 60 65 67 68 69 70 72 72

# Calculate the mode using the user function.
result <- getmode(md$dados)
print(result)
#[1] 72
```

MEDIDAS DE DISPERSÃO

- Dispersão é sinônimo de *variação* ou *variabilidade* .
- Paramedir dispersão, duas medidas são usadas: *amplitude* e *desvio padrão* .
- Amplitude
  - A amplitude é definida como sendo a diferença entre o maior e o menor valor do conjunto de dados.
  - A amplitude é denotada por *R* .
  - Considerando o conjunto de dados ordenado $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n-1)} \leq X_{(n)}$
  - A amplitude R dos dados é dada por: $R = X_{(n)} - X_{(1)}$
- Como o valor máximo do conjunto é 72 e o valor mínimo é 60, temos que a amplitude é: ***R = 72 - 60 = 12*** .

```
# Amplitude
ampl <- read.xlsx("estatistica_basica.xlsx", sheetName = "Plan5")
head(ampl)
range(ampl)
#[1] 60 72
diff( range(ampl) )
#[1] 12
```

Para definirmos desvio padrão é necessário definir variância. A notação mais comumente usada é:

s^2 - variância amostral.

σ^2 - variância populacional.

s - desvio padrão amostral.

σ - desvio padrão populacional.

MEDIDAS DE DISPERSÃO

Variância populacional

A variância de uma população $\{x_1,...,x_N\}$ de N elementos é a medida de dispersão definida como a média do quadrado dos desvios dos elementos em relação à média populacional μ. Ou seja, a variância populacional é dada por:

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

Variância amostral

A variância de uma amostra $\{x_1,...,x_n\}$ de n elementos é definida como a soma ao quadrado dos desvios dos elementos em relação à sua média $\overline{x}$ dividido por (n-1). Ou seja, a variância amostral é dada por:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n - 1}$$

Ao utilizarmos a média amostral como estimador de m para calcularmos a variância amostral, perdemos 1 grau de liberdade em relação à variância populacional.

Desvio padrão populacional

O desvio padrão populacional de um conjunto de dados é igual à raiz quadrada da variância populacional. Desta forma, o desvio padrão populacional é dado por:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}}$$

Desvio padrão amostral

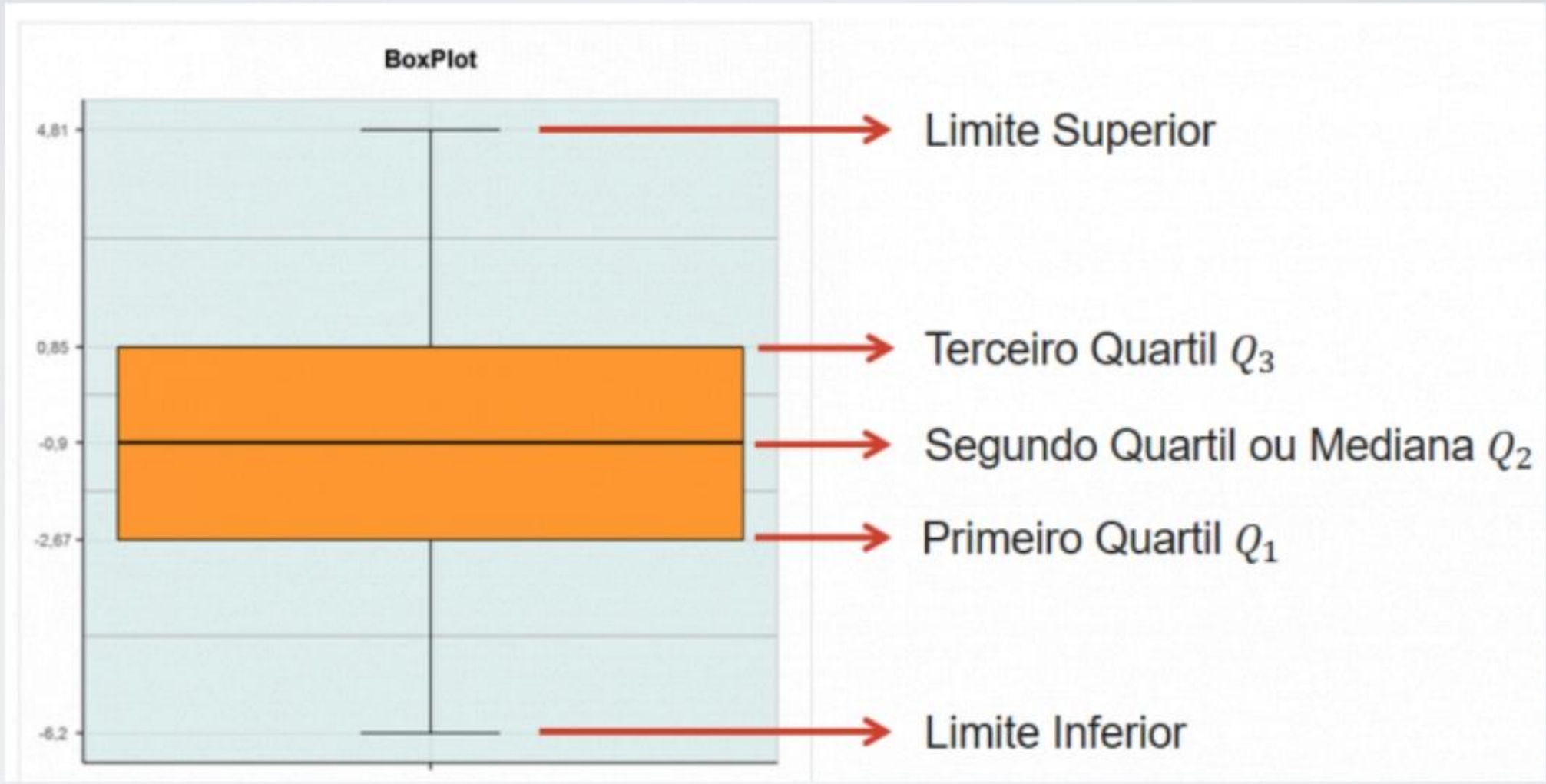
O desvio padrão amostral de um conjunto de dados é igual à raiz quadrada da variância amostral. Desta forma, o desvio padrão amostral é dado por:

$$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n - 1}}$$

QUARTIS

- Valores dados a partir do conjunto de observações ordenado em ordem crescente, que dividem a distribuição em quatro partes iguais.
    - O primeiro quartil, Q1, é o número que deixa 25% das observações abaixo e 75% acima.
    - O terceiro quartil, Q3, deixa 75% das observações abaixo e 25% acima.
    - Q2 é a mediana, deixa 50% das observações abaixo e 50% das observações acima.
- Em uma análise das estatísticas descritivas da amostra é fundamental para resumirmos algumas informações sobre a população.
- Estas informações são utilizadas para tomada de decisão e formação de modelos estatísticos paramétricos.
- Definiremos como:
    - Mínimo: menor elemento da amostra;
    - Máximo: maior elemento da amostra;
- Seja n o número total de elementos da amostra e calcule j(n+1)/4, para j=1,2 e 3. Desta forma Qj será um elemento entre Xk e Xk+1, onde k é o maior inteiro menor ou igual a j(n+1)/4 e será calculado da seguinte forma: $Q_j = X_k + \left(\frac{j(n+1)}{4} - k\right)(X_{k+1} - X_k).$
- Podemos observar que quando $ k $ é um valor inteiro, o quantil será o próprio Xk, isto é, $Q_j = X_k, \text{onde } k = \frac{j(n+1)}{4}, j = 1, 2, 3.$

QUARTIL


BoxPlot
4,81
0,85
-0,9
-2,67
-6,2
Limite Superior
Terceiro Quartil Q_3
Segundo Quartil ou Mediana Q_2
Primeiro Quartil Q_1
Limite Inferior

QUARTIL

```
# Quartis

quartil <- read.xlsx("estatistica_basica.xlsx", sheetName = "Plan6")

listaq <- sort(quartil$dados)

min(listaq)

#[1] 60

max(listaq)

#[1] 77

q1 = (length(listaq) + 1) / 4

#3

listaq[q1]

#[1] 66

## Logo, 25% das observações etão abaixo de 66 e 75% das observações estão acima de 66.

q3 <- 3 * (length(listaq) + 1) / 4

#[1] 9

listaq[q3]

#[1] 71

##Portanto, 75% das observações estão abaixo de 71 e 25% das observações estão acima de 71.

quantile(listaq)

# 0%  25%  50%  75% 100%

#60.0 66.5 69.0 70.5 77.0

boxplot(quantile(listaq) , pch=15, main="Quartiz" , col = "lightblue", pars = list(boxwex = 1))
```