Métodos Quantitativos e Qualitativos (Análise Estatística) - Parte 1 - Estatística Descritiva

Valderio A. Reisen e Bartolomeu Zamprogno

Departmento de Estatística - UFES

JANEIRO 2018

Valdério Anselmo Reisen, PhD

Departamento de Estatística e PPGEA - UFES, CentraleSupeléc-Paris

Bartolomeu Zamprogno, Dr (PPGEA)

Departamento de Estatística - UFES

Estatística

- A Estatística é uma ciência multidisciplinar, parte da Matemática Aplicada, que fornece métodos para coleta, organização, descrição, análise e interpretação de dados e para a utilização dos mesmos na tomada de decisões. Seu objetivo é o estudo da variabilidade, da incerteza, para a tomada de decisões frente a essa incerteza.
- Como a variabilidade e a incerteza estão presentes em todas as áreas do conhecimento, a estatística é uma ciência crucial para resolver uma série de problemas com uso de metodologias científicas apropriadas a partir da análise dos dados coletados. Em suma: **Estatística transforma dado em informação**.

Estatística: áreas

- **Amostragem e planejamento de experimentos:** tratam dos métodos científicos de amostragem e do planejamento de experimentos;
- **Estatística descritiva:** etapa inicial da análise utilizada para descrever, organizar(tabelas e gráficos) e resumir os dados.
- **Teoria das Probabilidades:** ramo da matemática que trata do estudo da incerteza, ou das medidas numéricas da plausibilidade da ocorrência de eventos. Fornece a base matemática para a inferência estatística, ou seja, a tomada de decisão em situações de incerteza;
- **Inferência:** fornece um conjunto de técnicas que permite tomar decisões (estimativas e testar hipóteses) sobre a *população* com base nas observações da *amostra*.

Estatística: Etapas

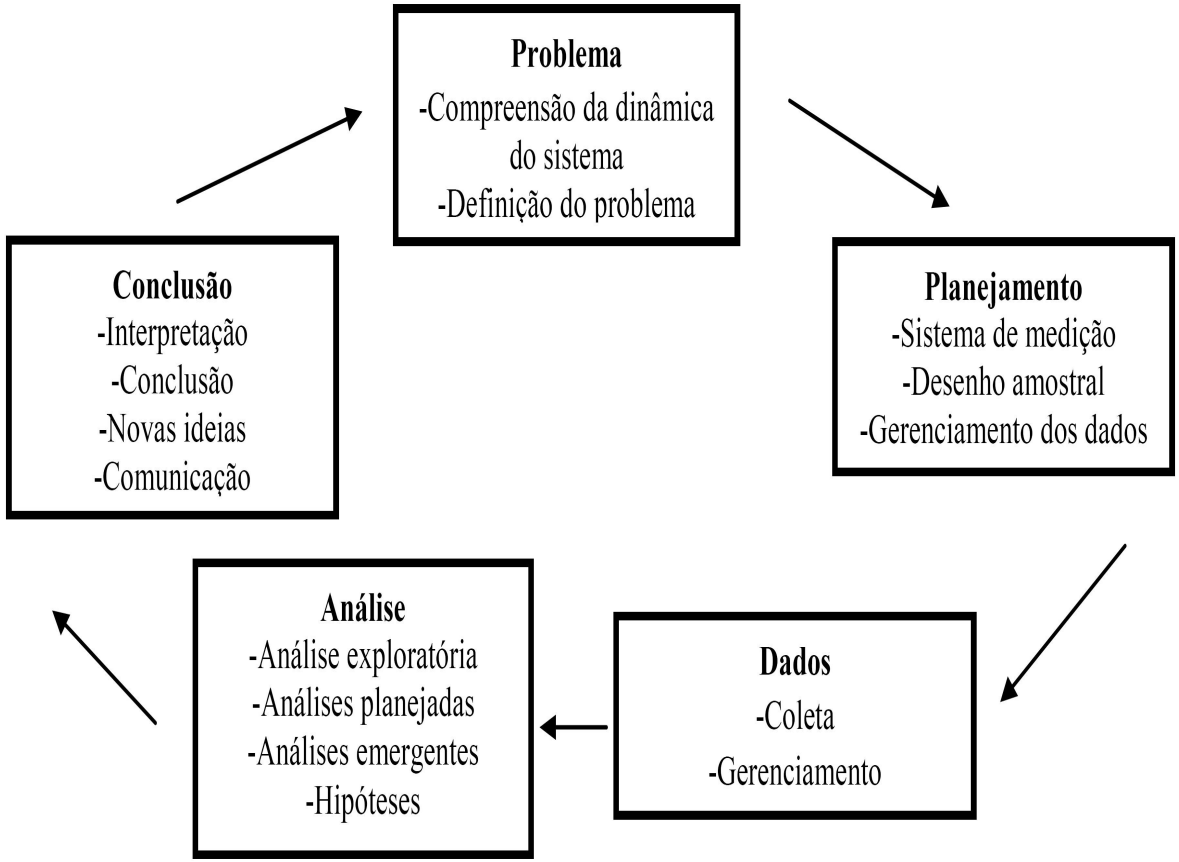
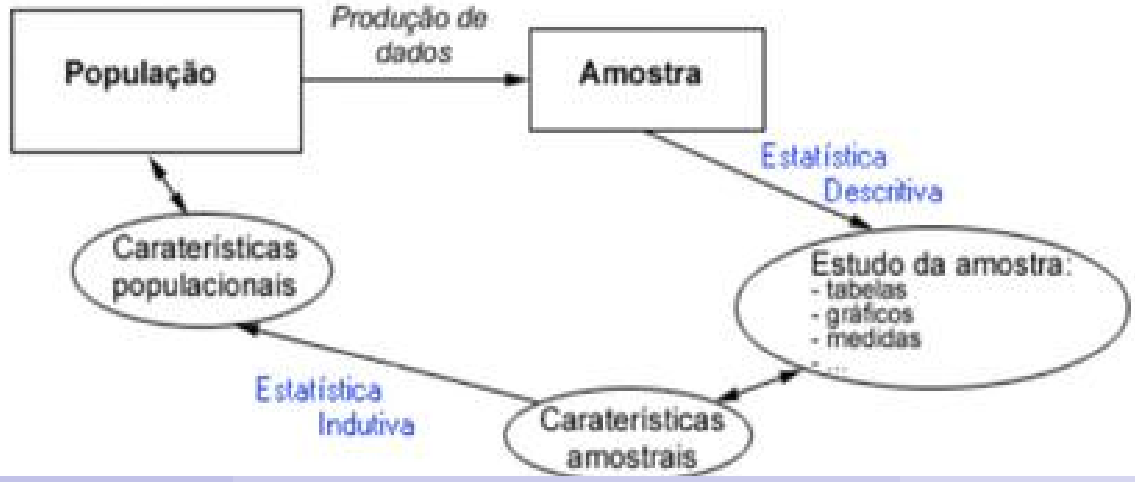


Figura: Etapas da análise estatística

Análise da População: Censo ou amostra?

Defina-se **população** ou universo estatístico ao conjunto de todos os elementos que têm pelo menos uma característica comum.
Defina-se **amostra** a um subconjunto finito da população. Razões para estudar uma amostra: econômicas; comodidade; tempo entre outros.

Esquematicamente, temos:



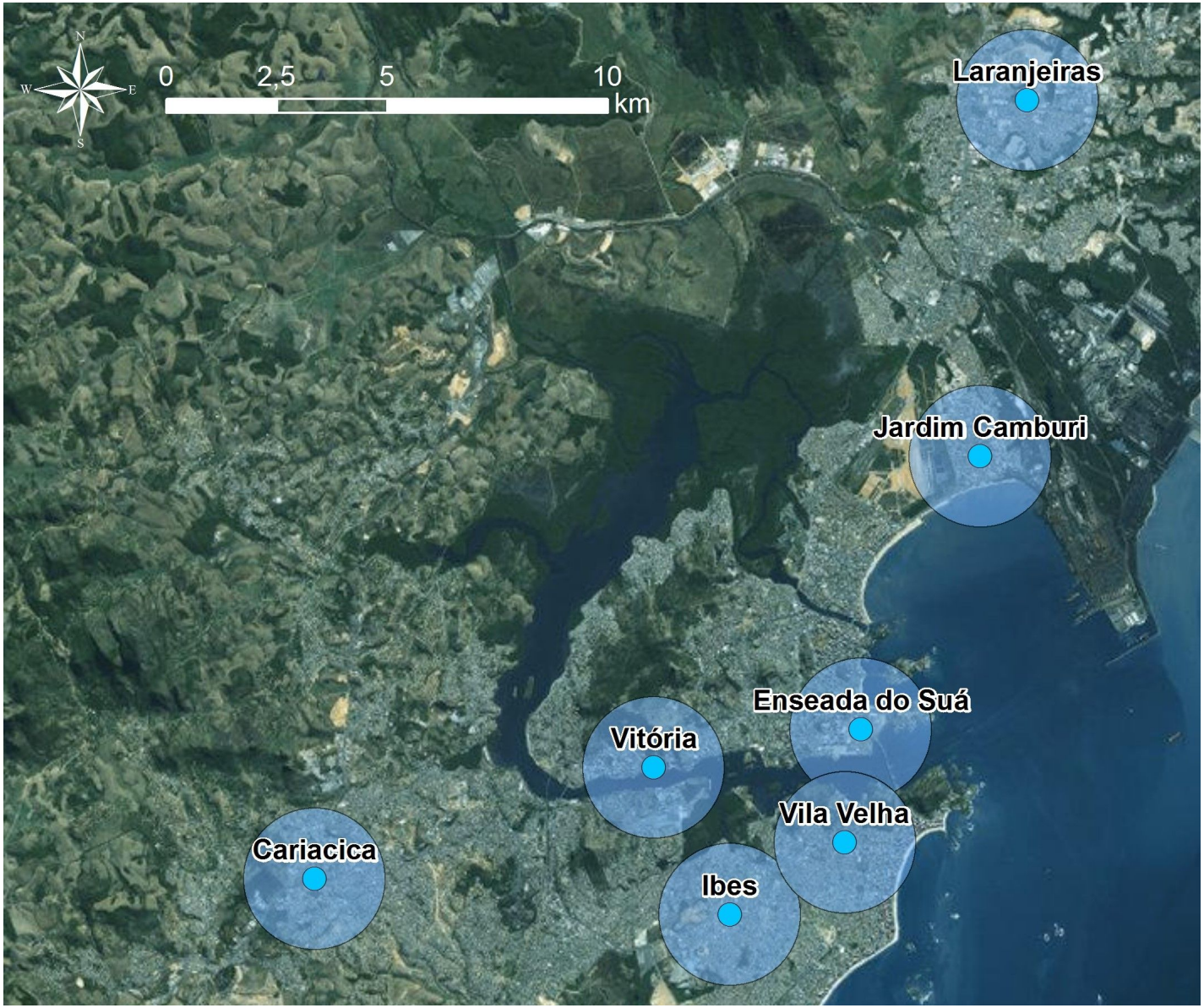
Exemplos do dia a dia de amostra:

- Experimentar uma sopa tomando uma pequena quantidade em uma colher;
- Retirada de sangue, fezes ou urina para análise;
- Ao abrir o chuveiro tomar percepção se a água está fria ou quente;
- Antes de entrar em uma piscina, colocar o pé na água.

Chamar a atenção que em todos os casos uma pequena parcela foi analisada para concluir para o todo.

Técnicas de amostragem

- Existem técnicas científicas para a seleção coleta de amostras. Entre essas técnicas, as mais conhecidas são a amostragem aleatória simples, a amostragem sistemática, a amostragem estratificada e a de conglomerados. Essas são probabilísticas. **Um exemplo para aplicação destas técnicas se dá no estudo do incomodo da poeira, Melo (2015)**. Neste estudo a variável principal é a percepção do incômodo causado pela poluição do ar, e pela presença de poeira. Para medir o incômodo algumas muitas perguntas foram elaboradas, por exemplo: "O sr. (a) se sente incomodado com a poluição do ar?"ou ainda "Durante esse último mês, o sr.(a) se sentiu incomodado com a presença de poeira em sua residência?". Delimitação da área de amostragem, ver p. 61 de Melo (2015).
- Há também as técnicas não probabilísticas, como por exemplo, amostragem por cota (assemelha à amostragem estratificada) e julgamento (aqui julga que determinado elemento amostral deve ser parte da amostra, **como um local de monitoramento que se entende que não pode ficar de fora da amostra**).


N
W
E
S
0
2,5
5
10
km
Laranjeiras
Jardim Camburi
Enseada do Suá
Vitória
Vila Velha
Cariacica
Ibes

- Para observações meteorológicas, análise de água e concentração do poluente há técnicas específicas para obtenção de uma amostra. A escolha de métodos é normalmente ditada pelo ambiente a ser monitorado, o parâmetro de interesse, e os requisitos de qualidade de dados.
  - Tipicamente deve-se selecionar um método científico apropriado, aprovado por uma agência reguladora. Por exemplo, métodos para analisar se a água é potável requerem técnicas de laboratório específicas. No caso de análise do total de chumbo dissolvido na água potável, nos EUA o método EPA 239.2 deve ser usado. Este método requer o uso de graphite furnace atomic absorption spectroscopy. Adicionalmente o método provê procedimentos de laboratório detalhados e requisitos de controle de qualidade para usar com as amostras de água.
  - Diversos métodos analíticos estão disponíveis para a análise de amostras de ar, água, solos, resíduos, plantas e animais.

Conjunto de dados

Consiste em um número de mensurações acerca de um (ou mais) fenômeno ou característica de interesse. Tal característica de interesse é denominada (*variável*).

- Uma característica de interesse pode ser o incomodo decorrente da poluição, a concentração de poluição de determinado poluentes, os tipos de poluentes, as regiões afetadas pela poluição, a quantidade de crianças internadas.
- Como a característica de interesse apresenta algumas respostas distintas, dizemos que essa característica é uma variável.
- As variáveis podem ser classificadas como qualitativas (nominais ou ordinais) e quantitativas (contínuas ou discretas).
- O padrão de variabilidade de uma variável aleatória é denominado de *distribuição*.

Variáveis: definições nas áreas

- Definição geral de variável: Do latim variabilis, uma variável é aquilo que varia ou pode variar. Trata-se de algo instável, inconstante e sujeito a alterações. Por outras palavras, uma variável é um símbolo que representa um elemento não especificado de um determinado conjunto. Esse conjunto é denominado conjunto universal da variável ou universo da variável, e cada elemento do conjunto é um valor da variável.
- Uma variável é um elemento de uma fórmula, de uma proposição ou de um algoritmo, podendo ser substituído ou podendo adquirir um valor qualquer dentro do seu universo. Os valores de uma variável podem ser definidos dentro de um intervalo ou estar limitados por condições de pertença.

- No âmbito da programação (informática), as variáveis são estruturas de dados que podem mudar de conteúdo ao longo da execução de um programa. Essas estruturas correspondem a uma área armazenada na memória principal do computador.
- Na astronomia, as estrelas **variáveis** são aquelas que sofrem variações significativas de luminosidade. Aqui a palavra variável é uma qualidade da estrela.

Definição de variável na estatística:

É a característica de interesse que é medida em cada elemento da amostra ou da população. Como o nome diz, seus valores variam de elemento para elemento. As variáveis podem ter valores numéricos ou não numéricos e podem ser classificadas seguinte forma: **Qualitativas** e **Quantitativas**.

Variáveis Quantitativas:

São as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser contínuas ou discretas.

- Variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são os resultados de contagens. Exemplos: número de filhos, número de cigarros fumados por dia, número de pessoas que sentem incomodo à poluição em determinada região.
- Variáveis contínuas, características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido. Usualmente, devem ser medidas por meio de algum instrumento. Exemplos: peso (balança), altura (régua), tempo (relógio), concentração do poluente, precipitação de chuva.

Variáveis Qualitativas (ou categóricas):

São as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

- Variáveis nominais: não existe ordenação dentre as categorias. Exemplos: sexo, fumante/não fumante, doente/sadio.
- Variáveis ordinais: existe uma ordenação entre as categorias. Exemplos: escolaridade (1^o, 2^o, 3^o graus), estágio da doença (inicial, intermediário, terminal), padrão da qualidade do ar: primário e secundário, classificação do IQA, veja `http://www.qualidade.iema.es.gov.br/scripts/sea0513.asp`.

Observações:

- Os valores de uma variável podem depender dos valores de outra variável, neste caso a chamamos de variável dependente. Quando uma determinada variável não sofre influência dos valores de outra variável, dizemos que a primeira variável é independente da outra variável. Atenção, uma variável qualitativa pode exercer influência em uma variável do tipo qualitativa ou quantitativa, e o mesmo vale para a variável quantitativa.
- Uma variável originalmente quantitativa pode ser coletada de forma qualitativa. Por exemplo, a variável idade, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos, etc...), é qualitativa (ordinal). Outro exemplo é o peso dos lutadores de boxe, uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado, etc.).

Observações:

- Outro ponto importante é que nem sempre uma variável representada por números é quantitativa. O número do telefone de uma pessoa, o número da casa, o número de sua identidade.
- Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa!

- Uma variável pode assumir qualquer resultado dentre um conjunto de possíveis resultados. Dessa forma, seu valor é imprevisível e aleatório, e portanto utilizamos o termo *variável aleatória*.
- Definição de variável aleatória (utilizada a partir da probabilidade em diante): Dado um experimento aleatório, descrito pelo espaço de probabilidades (Ω, E, P), uma função numérica $X : \Omega \to \Re$ será dita uma variável aleatória (do experimento). De forma simples, é uma função que associa os elementos do espaço amostral ao números reais.
- As variáveis aleatórias são classificadas como discretas, contínuas ou mistas (combinação de discreta e contínua).
- Exemplo: E: lançamento de um dado. X = número obtido na face superior é uma variável aleatória. Mais precisamente, $X : \Omega = \{1, 2, ..., 6\} \to \Re$ tal que $X(\omega) = \omega$ é uma função numérica do experimento, e logo é uma variável aleatória.

Alguns exemplos de variáveis aleatórias:

Discretas

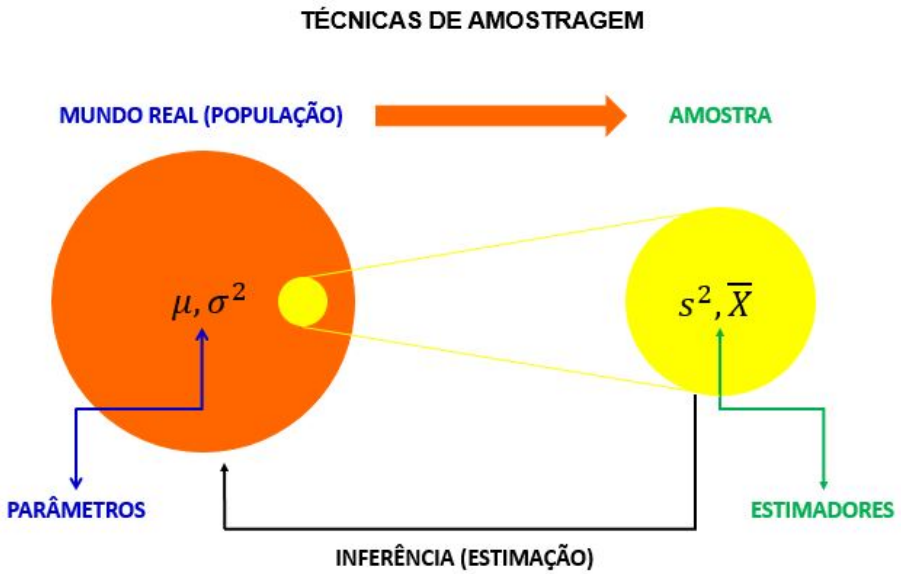
1. Características de um indivíduo sexo: 0 - feminino e 1 - masculino;
2. Incomodo à algum tipo de poluição: 0 - não e 1 - sim;
3. Número de crianças internadas decorrentes de poluição.

Contínuas

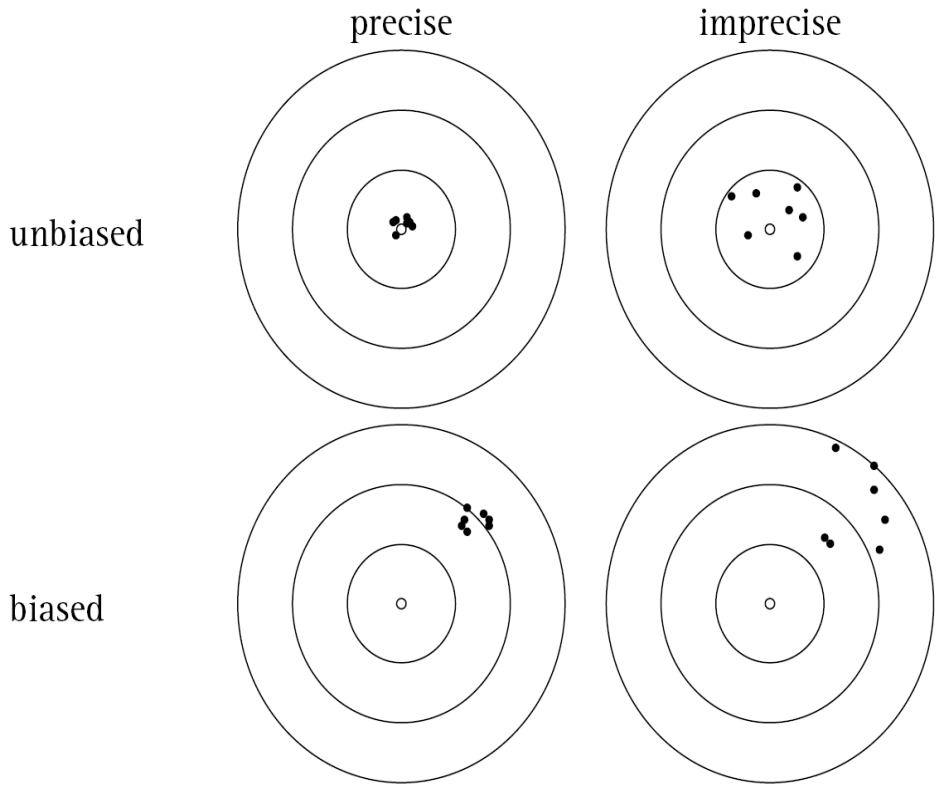
4. Concentração do poluente;
5. Nível de precipitação;
6. Características de um indivíduo: altura, peso, idade;

Obs: Voltaremos a falar de variáveis aleatórias no tópico de noções de probabilidade.

- O conceito de variável aplica-se tanto para uma população quanto para uma amostra oriunda da população. Ou seja, uma variável é quantitativa ou qualitativa independente da totalidade dos elementos que estão sendo considerados.
- Um detalhe deve ser observado, quando uma característica é observada na população, denominamos essa característica de parâmetro. Quando é observada na amostra, a tratamos como estimativa e para estimar necessita-se de um "bom" estimador. Dessa forma, o valor obtido por um estimador é uma estimativa.



Precisão do estimador. Para um estimador ter boas propriedades ele deve ser não viesado e consistente.



Para estimar uma média, um total, uma variância é necessário um bom estimador. Para estimar a concentração de um poluente em uma determinada região um modelo deve ser adotado e para isso as propriedades do modelo devem ser avaliadas.

Uma vez coletadas as observações, o que fazemos com elas?

Primeira etapa: resumo dos dados utilizando Estatística Descritiva. Através da estatística descritiva teremos um cenário geral dos dados coletados, ou seja, do fenômeno estudado.

Representação gráfica:

- Conjunto de técnicas para visualizar de forma simples (tabelas e gráficos) e resumir as informações (medidas de tendência, de dispersão, assimetria e curtose) contidas nos dados.
- Pode fornecer informações sobre a variabilidade dos dados, sobre a forma da sua distribuição e até mesmo sobre a presença de observações atípicas.

Observação

Deve-se considerar a tabela ou o gráfico de acordo com o tipo de variável que está sendo avaliada. Há tipo de tabelas e gráfico para variáveis qualitativas e para variáveis quantitativas.

- Os gráficos de barras, de pizza, de colunas, de pareto devem ser utilizados para variáveis qualitativas.
- Já os ramo e folhas, o histogramas, o box-plot, de linhas (série temporal) são gráficos para variáveis quantitativas.

As variáveis quantitativas discretas podem ser organizadas em *tabelas de frequências* e representadas usando-se *gráficos de barras*.

Organização dos dados em tabelas de frequência: uma tabela de frequência relaciona categorias (ou classes) de valores e contagens (frequências) do número de valores que se enquadram em cada uma das categorias ou classes.

Exemplo: Construção da tabela de frequências.
Conjunto de dados: número de ocorrências de enchentes por ano de 1939 a 1972 na estação de medição de Calamazza no Rio Magra, entre Pisa e Genoa (noroeste da Itália).

Observações: 2 5 7 3 2 1 7 4 6 5 3 4 7 2 5 4 4 4 3 1 2 4 3 4 8 4 2 3 7 3 5 4 2 3

Tabela

O comando table no R resume as informações de uma variável, faz a contagem dos distintos valores.

Tabela: Tabela de frequências: número de ocorrências de enchentes por ano de 1939 a 1972 na estação de medição de Calamazza no Rio Magra, entre Pisa e Genoa (noroeste da Itália).

Nro de enchentes por ano (X_i)	Nro de ocorrências Frequência absoluta (f_i)	Freq. relativa (f_{ri})	Freq. relativa % ($f_{ri} * 100\%$)
0	0	0.00	0
1	2	0.06	5.88
2	6	0.18	17.65
3	7	0.21	20.59
4	9	0.26	26.47
5	4	0.12	11.76
6	1	0.03	2.94
7	4	0.12	11.76
8	1	0.03	2.94
9	0	0.00	0
Total	34	1	100

f_i: frequência absoluta da categoria i (número de observações que pertencem à categoria i);
$f_{ri} = \frac{f_i}{n}$: frequência relativa da categoria i;
$f_{ri}\% = \frac{f_i}{n} * 100\%$: frequência relativa percentual da categoria i;
n: total de observações.

No R faça:

> dados=c(2, 5, 7, 3, 2, 1, 7, 4, 6, 5, 3, 4, 7, 2, 5, 4, 4, 4, 3, 1, 2, 4, 3, 4, 8, 4, 2, 3, 7, 3, 5, 4, 2, 3)

> table(dados)

dados

1 2 3 4 5 6 7 8 2 6 7 9 4 1 4 1

No Excel de forma simples utilize a tabela dinâmica com o uso da função contagem em configurações de campo de valor.

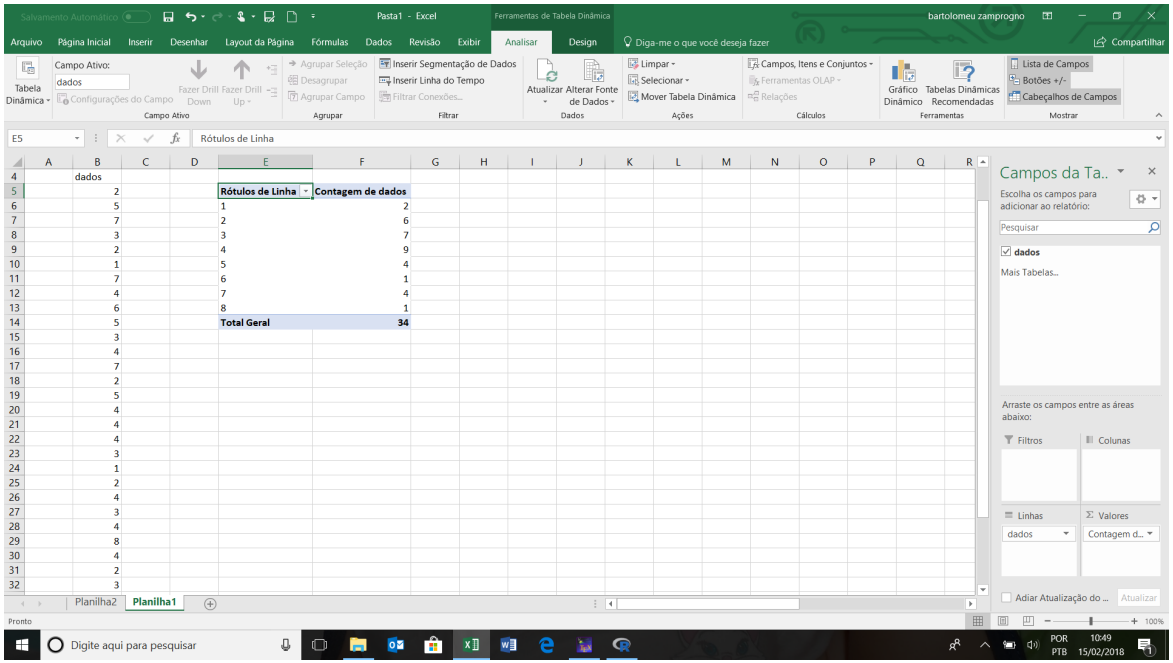
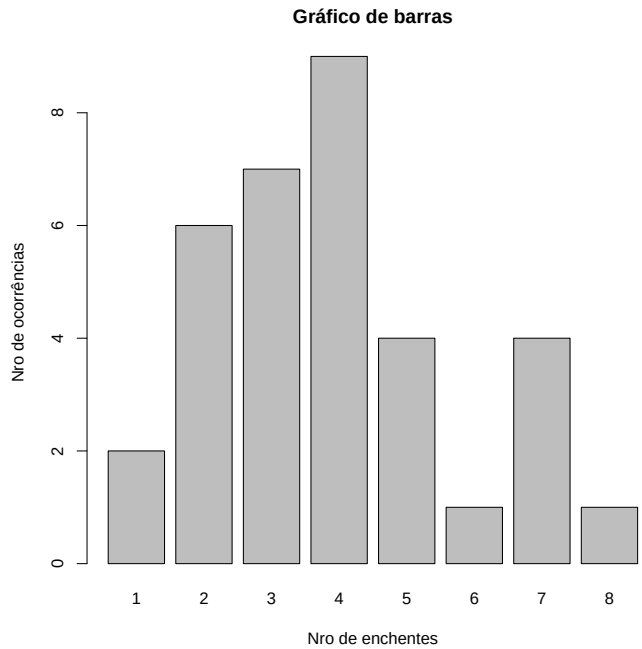
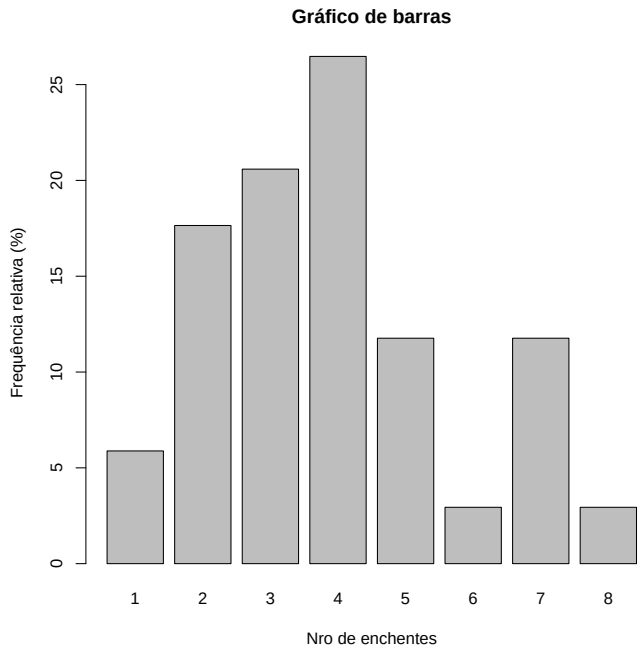


Gráfico de Barras

Para construir o gráfico no R utilize o comando barplot. No Excel vá no menu Inserir e escolha gráfico de barras.



(a)

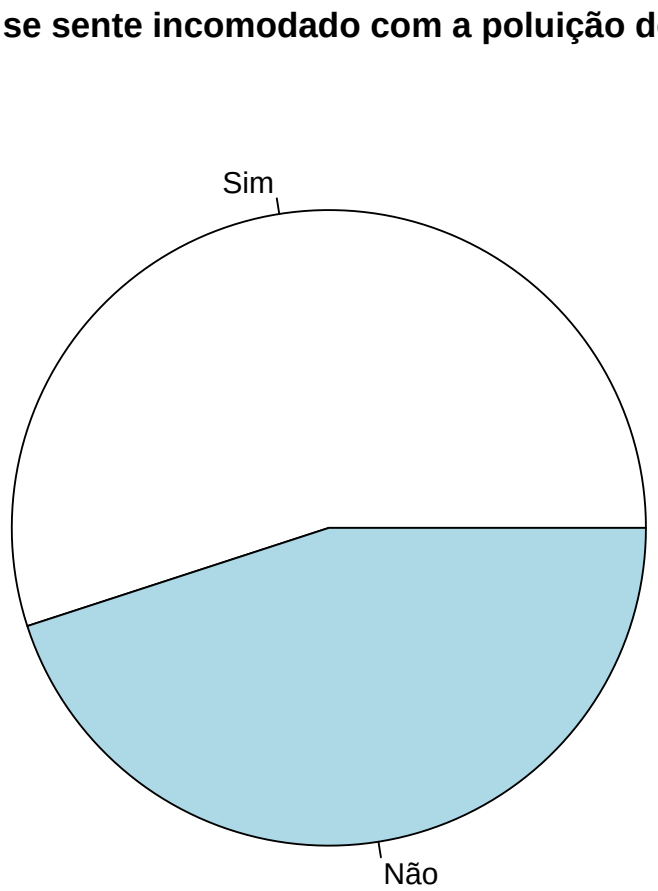


(b)

Figura: Gráficos de barras para os dados de número de ocorrências de enchentes por ano de 1939 a 1972 na estação de medição de Calamazza no Rio Magra, entre Pisa e Genoa (noroeste da Itália). (a) Freq. absoluta, (b) Freq. relativa.

Gráfico de pizza ou setor

No R para construir um gráfico de pizza utilize o comando pie. No Excel vá no menu Inserir e escolha gráfico de pizza.



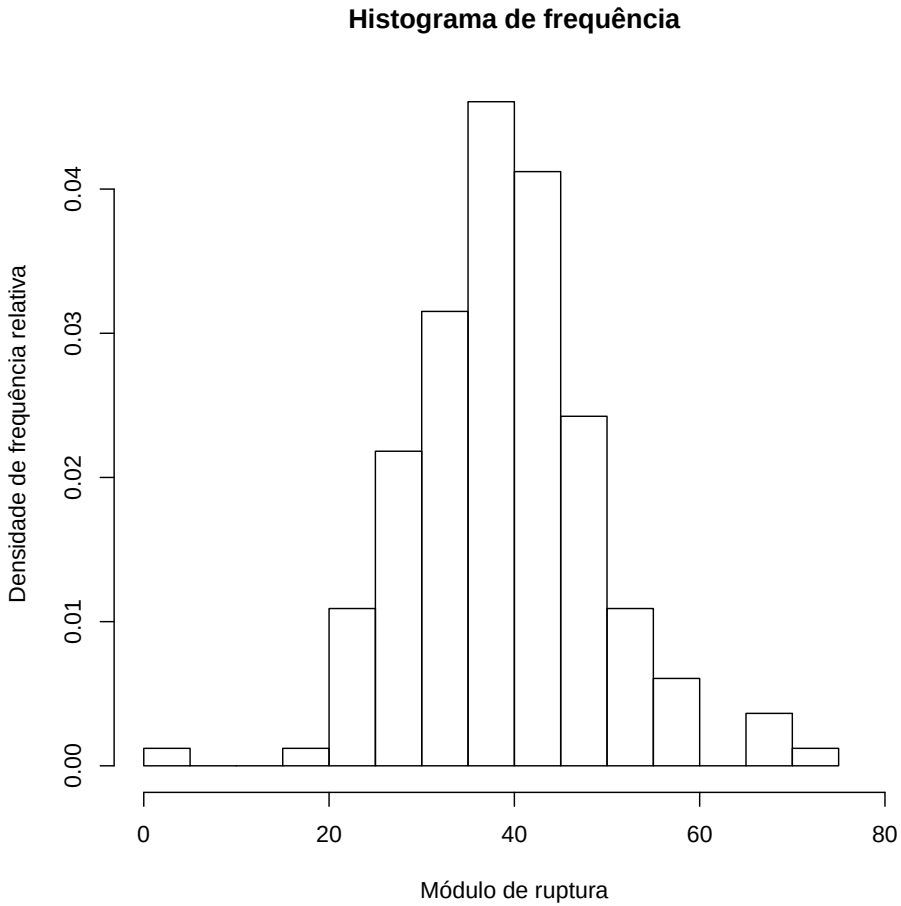
Histograma

- O histograma é um gráfico de barras adjacentes de um conjunto de dados previamente resumido em uma tabela de frequências.
- Indicado para variáveis quantitativas contínuas com no mínimo 25 observações, mas também pode ser utilizado para var. quanti. discretas com muitos valores distintos.

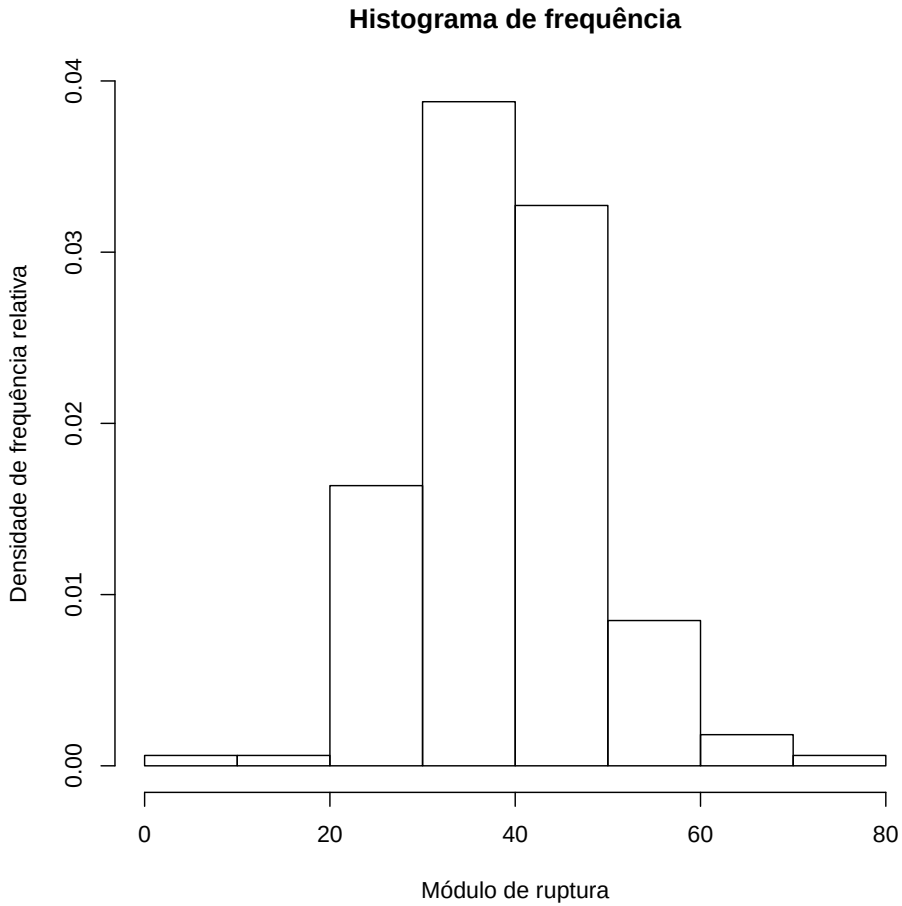
Para construir tabelas de frequências para variáveis contínuas, construímos faixas ou classes de valores e contamos o número de ocorrências em cada faixa de valores.

Usualmente, os intervalos são construídos de tal forma que sejam adjacentes e de mesmo tamanho.

No R pode-se construir o histograma utilizando o comando hist. Há outras formas de construção do histograma no R. No Excel pode construir de forma manual, um processo trabalhoso, ou via tabela dinâmica, com a opção agrupar.



(a)

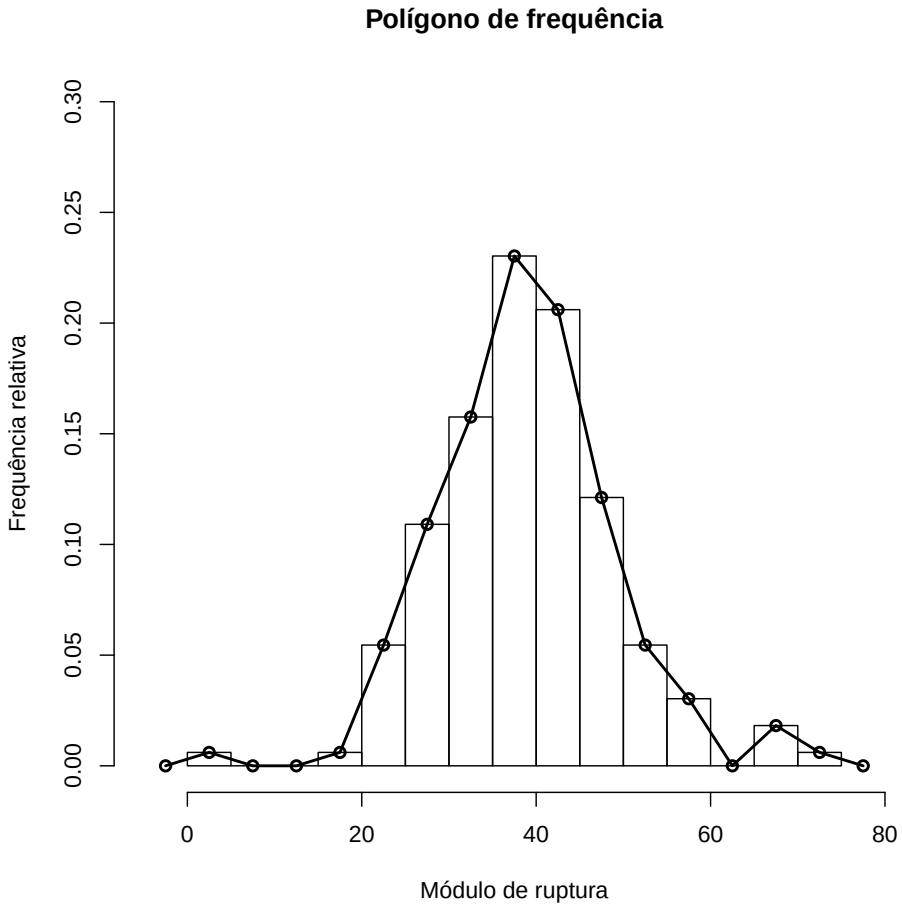


(b)

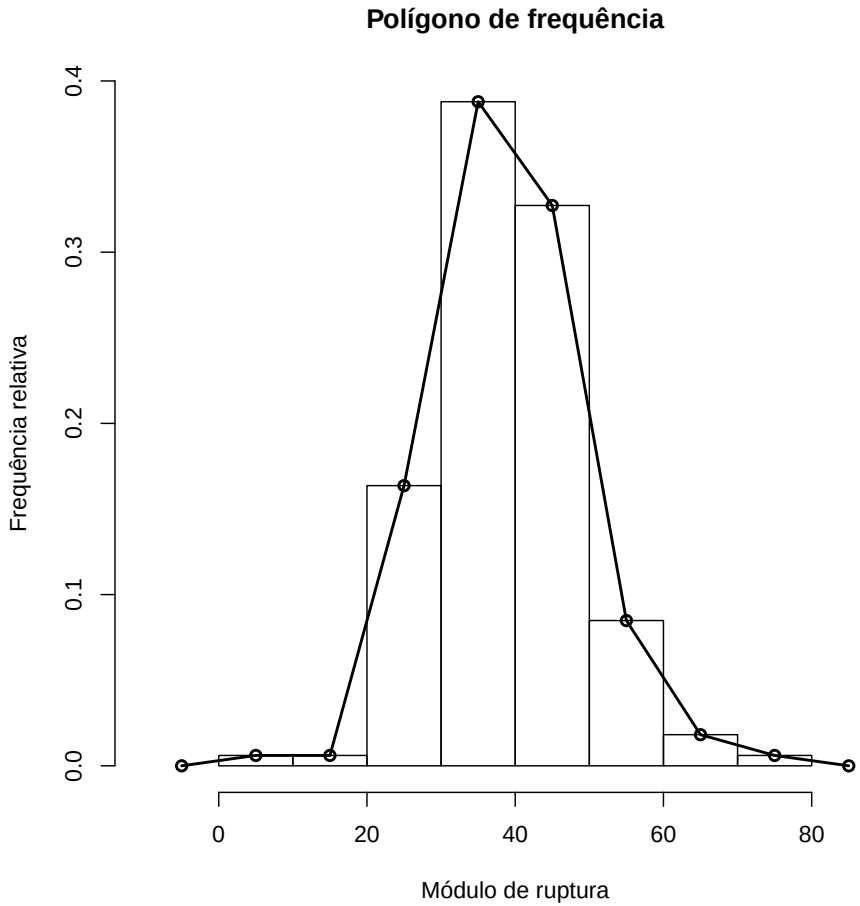
Figura: Histogramas de densidades de frequência relativas para os dados de ruptura de madeira: (a) $n_c = 15$, $r = 75$ e $h = 5$; (b) $n_c = 8$, $r = 80$ e $h = 10$.

Polígono de frequência

- Ferramenta de representação gráfica dos dados que possibilita determinar a forma da distribuição da variável em estudo.
- É construído ligando-se os pontos médios das barras do histograma de frequências e incluindo-se uma classe antes da menor classe e uma classe após a maior classe (desta forma o gráfico tocará o eixo horizontal).



(a)



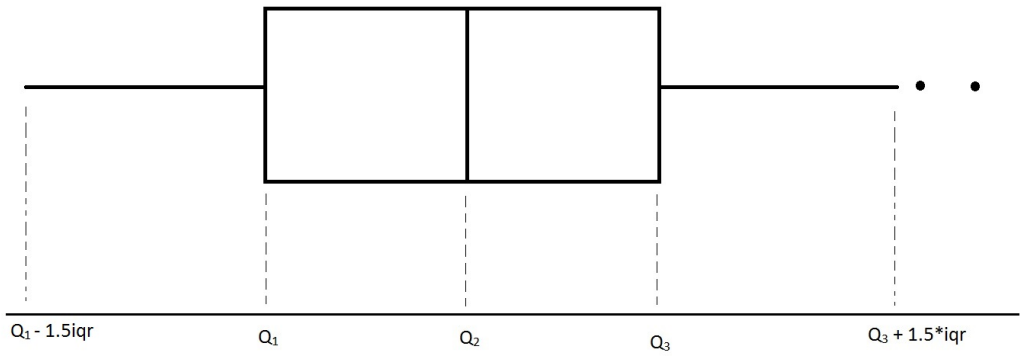
(b)

Figura: Histogramas de frequência relativas para os dados de ruptura de madeira: (a) $n_c = 15$, $r = 75$ e $h = 5$; (b) $n_c = 8$, $r = 80$ e $h = 10$;.

Box-plot

- Representação gráfica dos dados por meio de um retângulo construído com os quartis, isto é, (valor mínimo), Q_1, Q_2, Q_3 e (valor máximo). Os valores mínimos e máximos podem ser substituídos por meio do $P_{5\%}$ e $P_{95\%}$,respc., ou outros extremos.
- Fornece uma representação da distribuição dos dados e da variabilidade dos mesmos.
- Pode identificar a presença de observações atípicas (outliers).

Esquema de construção de um box-plot:



Figura

Construção de um box-plot.

Conjunto de dados: dados extraídos da "1974 Motor Trend US magazine". Observações de consumo de combustível para $n = 32$ modelos de veículos fabricados em 1973-74 (em milhas por galão).

Dados disponíveis no `R`: *data(mtcars)*.

Observações: 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4

BoxPlot no R

Utilize o comando boxplot. Abaixo segue a sequência manual da construção.

Ordenando as observações em ordem crescente, temos:
10.4 10.4 13.3 14.3 14.7 15.0 15.2 **15.2 15.5** 15.8 16.4 17.3 17.8 18.1 18.7
19.2 19.2
19.7 21.0 21.0 21.4 21.4 21.5 **22.8 22.8** 24.4 26.0 27.3 30.4 30.4 32.4 33.9
Calculando as medidas:
(determinando a "caixa")
Q_2 (mediana): $\frac{19.2+19.2}{2} = 19.2$
Q_1: $\frac{15.2+15.5}{2} = 15.35$
Q_3: $\frac{22.8+22.8}{2} = 22.8$
(extensão das linhas a partir da caixa)
$iqr = Q_3 - Q_1 = 22.8 - 15.35 = 7.45$, $1.5iqr = 1.5 * 7.45 = 11.175$
$Q_1 - 1.5iqr = 15.35 - 11.175 = 4.175$, $Q_2 + 1.5iqr = 22.8 + 11.175 = 33.975$
Alternativa: usar os valores mínimo e máximo: $MIN = 10.4$ e $MAX = 33.9$.

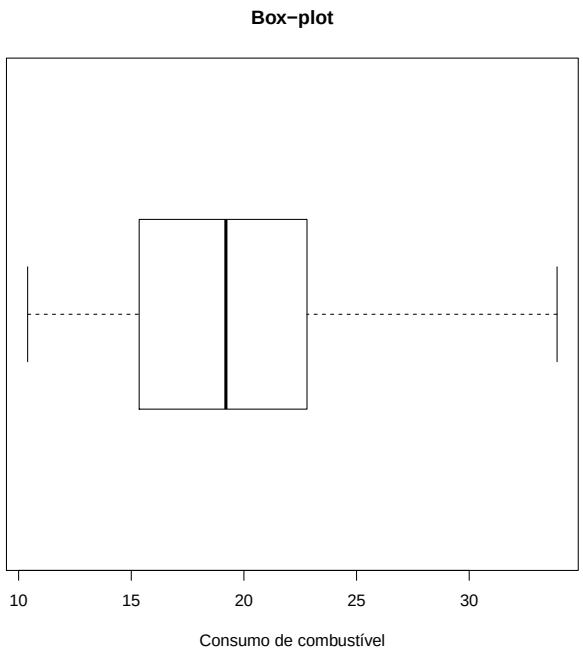


Figura: Box-plot para os dados de consumo de combustível.

Introdução: Medidas Descritivas

- Um sumário numérico dos dados é um número usado para descrever uma característica específica do conjunto de dados.
- Fornecem uma metodologia para extrair e resumir informações relevantes presentes nos conjuntos de dados (com precisão).
- São importantes para fazer inferência sobre a população a partir dos dados observados.
- Por exemplo, qual medida poderia ser utilizada para descrever o centro da distribuição de uma variável?
- Veremos que as medidas descritivas representam características importantes e podem ser reconhecidos nos histogramas e polígonos de frequência.
- As medidas descritivas podem ser classificadas em tendência, dispersão, assimetria e curtose.

Medidas de tendência:

1. Muitas vezes estamos interessados em valores que possam representar todos os resultados de uma variável observada em um experimento.
2. Geralmente, ocorre que os dados observados em experimentos tendem a concentrar-se em torno de um valor específico da variável. Uma medida, conhecida como medida de tendência central, pode ser tomada como representativa dos dados.

As principais medidas de posição são: média, mediana e moda. As medidas de tendência central dão o centro do histograma e do polígono de frequência.

Quartis, decis e percentis também são medidas de tendência e são conhecidas como de separatrizes.

Para dados brutos ou tabelado não agrupados em classes:

- Média: Sejam $x_1, x_2, \ldots, x_n$ n valores observados da variável aleatória quantitativa X. Então, a média amostral, denotada por $\bar{x}$, é definida por

$$\bar{x} = \frac{1}{n} \sum_{n}^{i=1} x_i$$

- Mediana: A mediana é o valor da variável que ocupa a posição central de um conjunto de n observações ordenadas. Para um conjunto de n observações, a posição da mediana é dada pelo valor que ocupa a $\frac{n+1}{2}$ posição. Notação: Md.
- Moda: A moda é o valor da variável que ocorre com maior frequência. Notação: Mo.

Exemplo: Média, mediana e moda.

Conjunto de dados: dados extraídos da "1974 Motor Trend US magazine". Observações de consumo de combustível para $n = 32$ modelos de veículos fabricados em 1973-74 (em milhas por galão). Dados disponíveis no `R`: *data(mtcars)*.

Observações: 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7 15.0 21.4

Ordenando as observações em ordem crescente, temos: 10.4 10.4 13.3 14.3 14.7 15.0 15.2 15.2 15.5 15.8 16.4 17.3 17.8 18.1 18.7 **19.2 19.2** 19.7 21.0 21.0 21.4 21.4 21.5 22.8 22.8 24.4 26.0 27.3 30.4 30.4 32.4 33.9

Média: $\bar{x} = 20.09062$.
Mediana: $\text{Md} = \frac{19.2+19.2}{2} = 19.2$.
Moda: note que o valor 19.2 e o valor 21.0 ocorrem 2 vezes.

Quando deve-se fazer uso da média ou da mediana ou da moda?

Na maioria das situações, não necessitamos calcular as três medidas de tendência central.
Normalmente precisamos de apenas uma das medidas para caracterizar o centro da série. Surge, então, a questão: qual medida deve ser utilizada?

- **A medida ideal em cada caso é aquela que melhor representa a maioria dos dados da série**.
- Se os valores são iguais, qualquer uma representa, mas no geral os valores divergem.
- No caso de forte concentração de dados na área central, utilize a média porque ela tem melhores propriedades estatísticas.
- No caso de forte assimetria para esquerda ou direita, utilize a mediana.
- A moda deve ser a opção como medida de tendência central apenas em séries que apresentam um elemento típico, isto é, um valor cuja frequência é muito superior a frequência dos outros elementos da série.

Medidas de dispersão:

Objetivo: encontrar um valor que resuma a variabilidade de um conjunto de dados.

- Para descrever de modo satisfatório uma variável de interesse, não é suficiente observar apenas uma medida de posição.
- Podemos facilmente encontrar variáveis que apresentam o mesmo valor para uma medida de locação (média, por exemplo), porém com dados apresentando comportamentos completamente diferentes.
- Esses diferentes comportamentos são consequência de dados com diferentes graus de dispersão.
- As medidas de dispersão indicam quanto os valores observados estão "dispersos" e também o quanto os dados "variam".

As principais medidas de dispersão são a variância e desvio padrão amostral. Há várias outras medidas de dispersão como a amplitude, a distância interquartil e o coeficiente de variação.

Exemplo: Considere três conjuntos de dados:
Grupo 1: $x_1 = \{3, 4, 5, 6, 7\}$
Grupo 2: $x_2 = \{1, 3, 5, 7, 9\}$
Grupo 3: $x_3 = \{5, 5, 5, 5, 5\}$.
Note que $\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = 5$ e $\text{Md}_{x_1} = \text{Md}_{x_2} = \text{Md}_{x_3} = 5$.

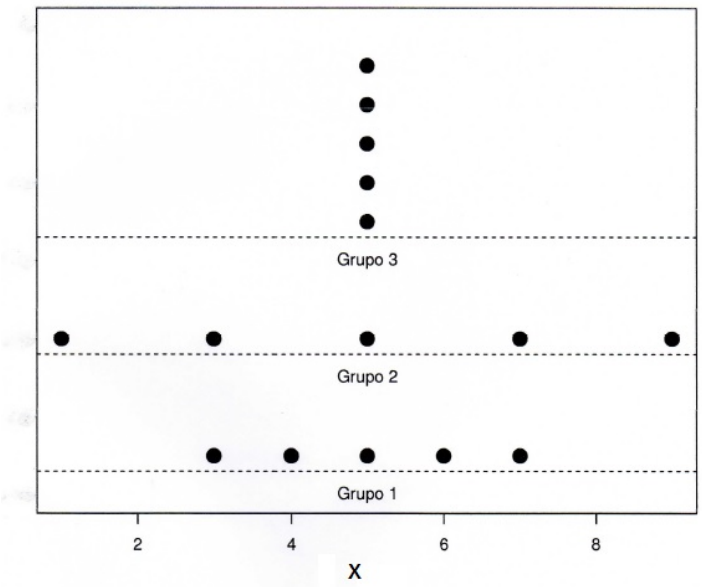


Figura: Diagrama de pontos.

Para dados brutos ou tabelado não agrupados em classes:

Variância: Sejam $x_1, x_2, \ldots, x_n$ n valores observados da variável aleatória quantitativa X. Então, a variância amostral, denotada por s^2, é definida por

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

- O que diz essa fórmula? Por que divisão por $n - 1$?
- Atenção: a variância amostral será uma medida de dimensão igual ao quadrado da dimensão dos dados observados.

Para dados brutos ou tabelado não agrupados em classes:

- Desvio padrão: O desvio padrão amostral é definido por

$$s = \sqrt{s^2}.$$

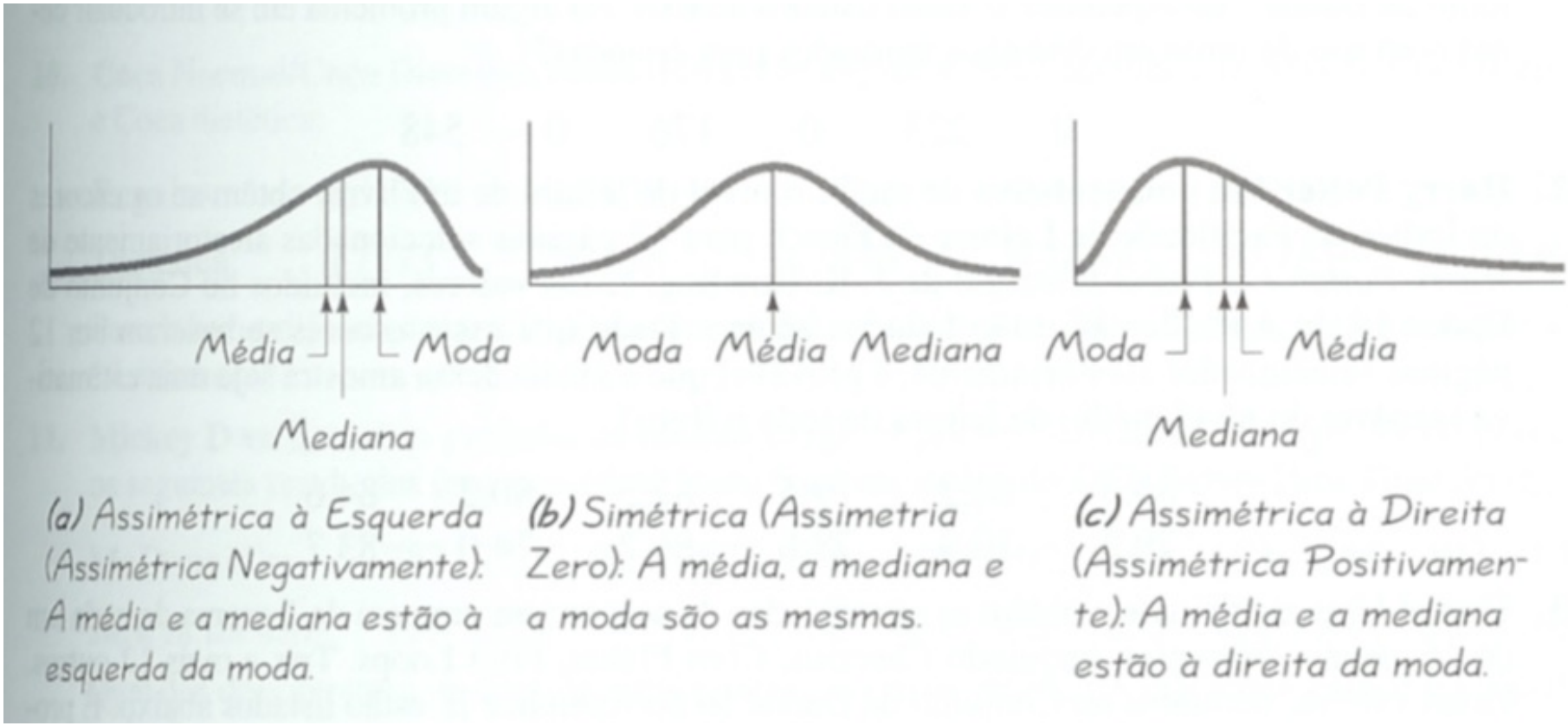
- Coeficiente de variação: $\upsilon = \frac{s}{\bar{x}} * 100\%$. O coeficiente de variação é uma medida adimensional e fornece um grau de variação dos dados, sendo útil para comparar conjuntos de dados (o conjunto com maior υ é mais disperso). Vantagem do coeficiente de variação em relação ao desvio padrão, no caso de duas populações com médias muito diferentes podem ter suas disperções comparadas através do coeficiente de variação.

Medidas de Assimetria:

- Uma comparação da média, mediana e moda, pode revelar uma informação sobre o comportamento dos dados, denominada "assimetria".
- Uma variável é dita ter comportamento (ou distribuição) assimétrico quando os seus valores estão mais concentrados em um dos seus extremos (valores altos ou baixos).
- Há algumas fórmulas para cálculo da assimetria. Destaca aqui o segundo coeficiente de assimetria de Pearson (A_p), dado por

$$A_p = \frac{3(\bar{x} - Md)}{s}.$$

Teoricamente, o segundo coeficiente de assimetria de Pearson pode variar entre -3 e +3. Na prática, porém, raramente ultrapas sará os limites de -1 e +1.

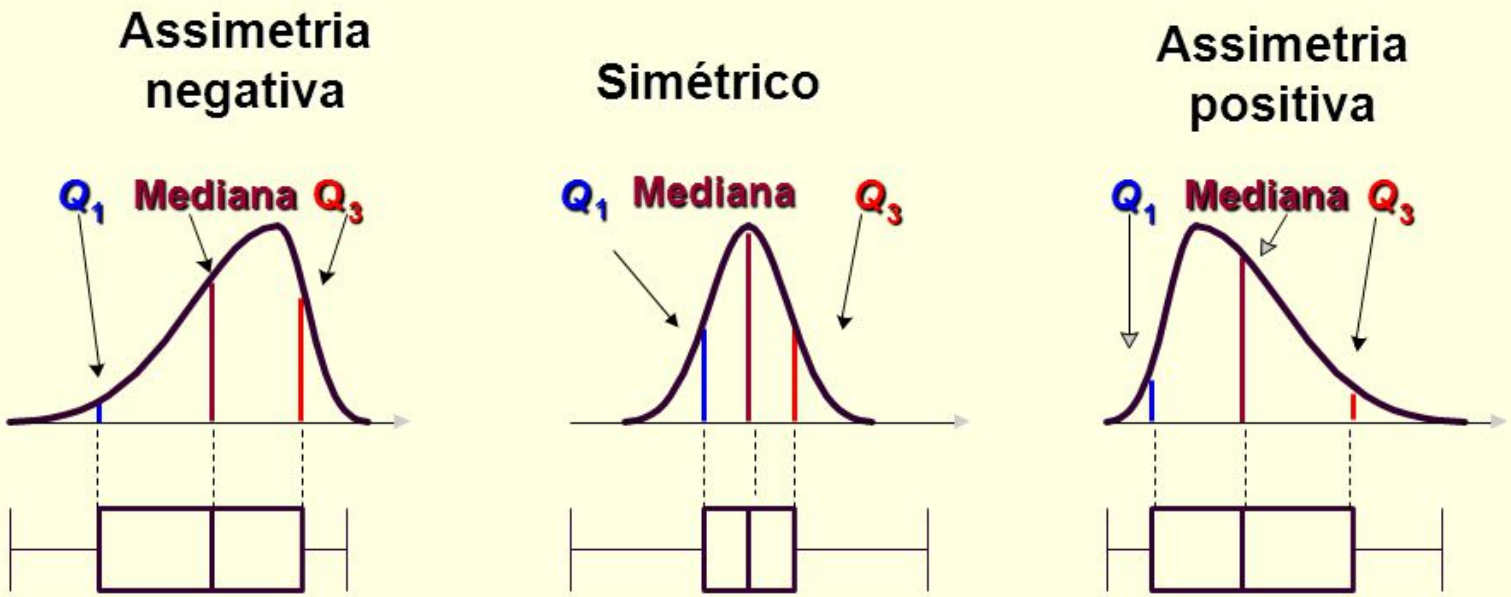


(a) Assimétrica à Esquerda (Assimétrica Negativamente): A média e a mediana estão à esquerda da moda.

(b) Simétrica (Assimetria Zero): A média, a mediana e a moda são as mesmas.

(c) Assimétrica à Direita (Assimétrica Positivamente): A média e a mediana estão à direita da moda.

Forma & Box Plot

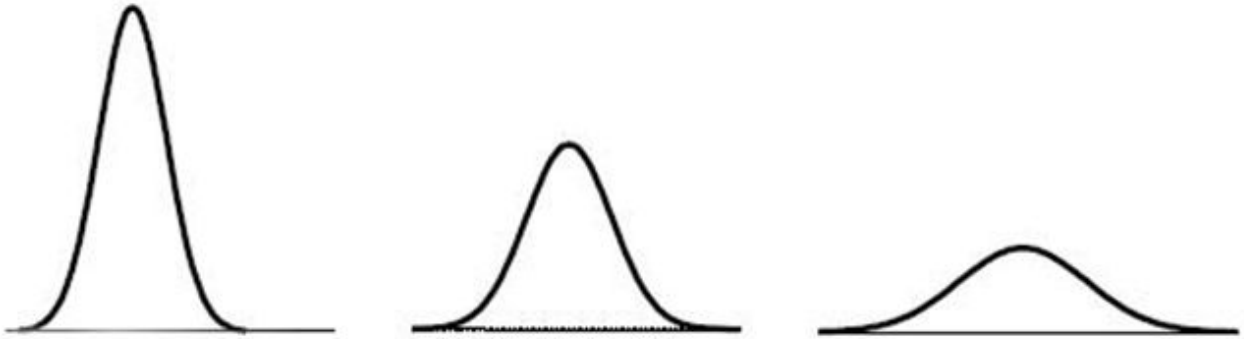


Medidas de Curtose:

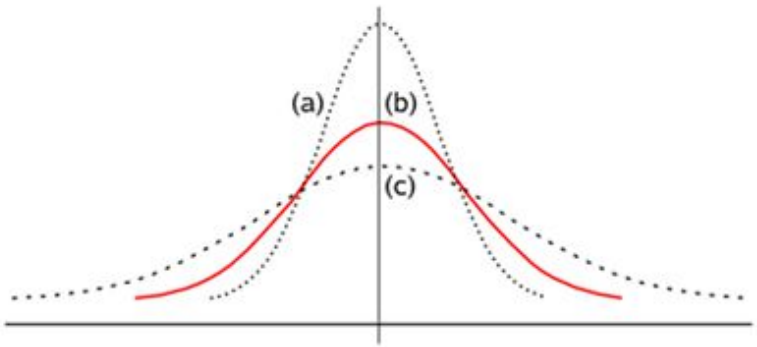
Denomina - se curtose ao grau de achatamento de uma distribuição de freqüências, geralmente unimodal, medido em relação ao de uma distribuição normal (de Gauss) que é tomada como padrão. O que as medidas de curtose buscam indicar realmente é o grau de concentração de valores da distribuição em torno do centro desta distribuição.
Dizemos que uma distribuição de freqüências é:

- Mesocúrtica: quando apresenta uma medida de curtose igual à da distribuição normal.
- Platicúrtica: quando apresenta uma medida de curtose menor que a da distribuição normal.
- Leptocúrtica: quando apresenta uma medida de curtose maior que a da distribuição normal.

Curtose e o coeficiente



Leptocúrtica | Mesocúrtica | Platicúrtica



Coeficiente percentílico de curtose

$$k = \frac{(Q_3 - Q_1)}{2.(P_{90} - P_{10})}$$

k = 0,263 - curva mesocúrtica
k < 0,263 - curva leptocúrtica
k > 0,263 - curva platicúrtica

Variáveis Bidimensionais

Nos problemas anteriores, consideramos apenas a representação gráfica e medidas resumo para conjunto de dados com apenas uma variável.
Na prática, os conjuntos de dados são compostos de duas ou mais variáveis.
Nestes casos, os dados costumam estar dispostos da segunte forma:

Tabela: Tabela de dados com p variáveis.

	Variáveis			
Observação	X_1	X_2	$\ldots$	X_p
1	x_{11}	x_{12}	$\ldots$	x_{1p}
2	x_{21}	x_{22}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{11}	$\ldots$	x_{np}

Tabela: Tabela de dados com 2 variáveis.

	Variáveis	
Observação	X_1	X_2
1	x_{11}	x_{12}
2	x_{21}	x_{22}
$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	x_{11}

Nas análises bidimensionais é possível verificar a relação entre duas variáveis qualitativas, ou duas quantitativas ou entre uma qualitativa com uma quantitativa. Para mais detalhes veja o cap. 4 do livro do Morettin e Bussab (2010).

Exemplos gerais:

- Conjunto de dados 1: medição de temperatura no período de n meses.
  X_1: temperatura na cidade de Vitória-ES **versus** X_2: temperatura na cidade de Campinas-SP.
- Conjunto de dados 2: tempo de serviço e número de clientes de funcionários de uma empresa.
  X_1: tempo de serviço **versus** X_2: número de clientes.
- Conjunto de dados 3: produção de peças em uma fábrica.
  X_1: número de peças produzidas **versus** X_2: número de peças defeituosas.

Exemplos na poluição do ar:

- Avaliação do número de internações causada por um determinado poluente;
- Concentração de um poluente versus uma variável meteorológica (vento, preciptação);
- Incomodo à poluição versus taxa de deposição das partículas.

Objetivo: Encontrar relações e possíveis associações entre as variáveis.
Gráfico de dispersão:

- ferramenta de análise gráfica que permite representar a associação entre **duas variáveis quantitativas**.
- mostra a relação entre duas variáveis quantitativas (medidas sobre os mesmos "indivíduos"). Os valores de uma variável aparecem no eixo horizontal e os da outra no eixo vertical. Cada indivíduo aparece como o ponto do gráfico definido pelos valores de ambas as variáveis para aquele indivíduo.

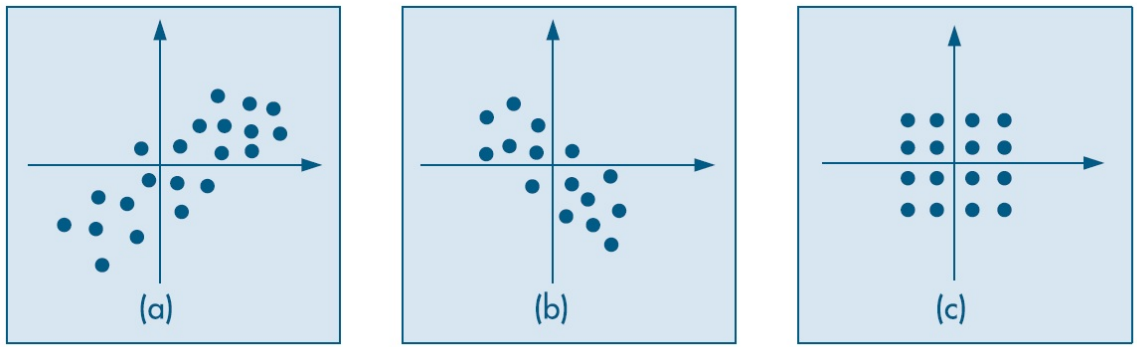


Figura: Exemplos de tipos de associação entre variáveis.

Exemplo: Construção do gráfico de dispersão.
Conjunto de dados: Tempo de serviço (em anos) e número de clientes de funcionários de uma empresa de seguros (ver Tabela 4.12 do livro do Moretim).

Tabela: Tabela de dados: tempo de serviço (em anos) e número de clientes de funcionários de uma empresa de seguros.

Funcionário	Tempo de serviço (X)	Número de clientes (Y)
1	2	48
2	3	50
3	4	56
4	5	52
5	4	43
6	6	60
7	7	62
8	8	58
9	8	64
10	10	72

No R utilize plot(x,y) e no Excel utilize o diagrama de dispersão.

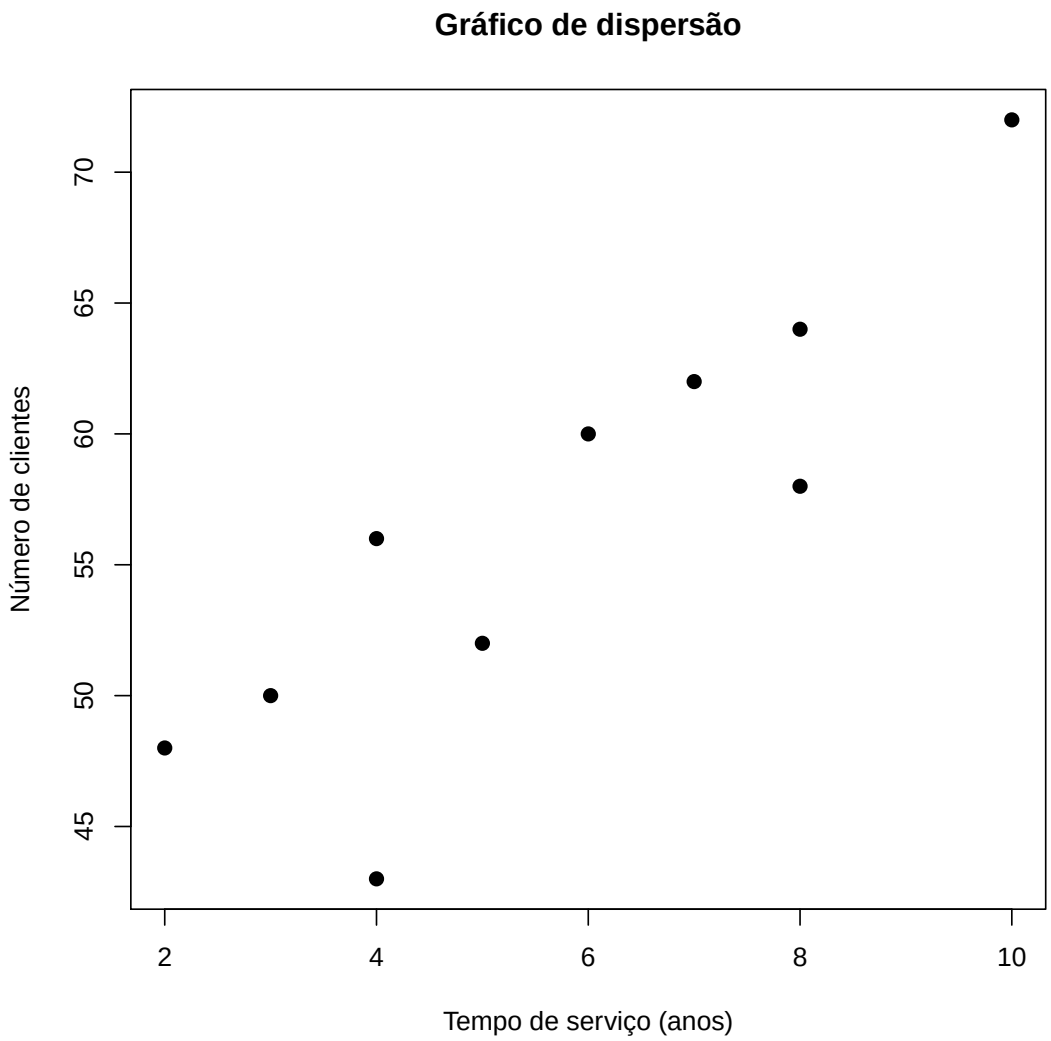


Figura: Gráfico de dispersão para as variáveis tempo de serviço (em anos) (X) e número de clientes (Y) de funcionários de uma empresa de seguros.

Exemplos possíveis de aplicação

Exemplos de dados possíveis para utilizar o Excel e software R para análise das variáveis e construção de gráficos e tabelas, e uso das medidas descritivas.

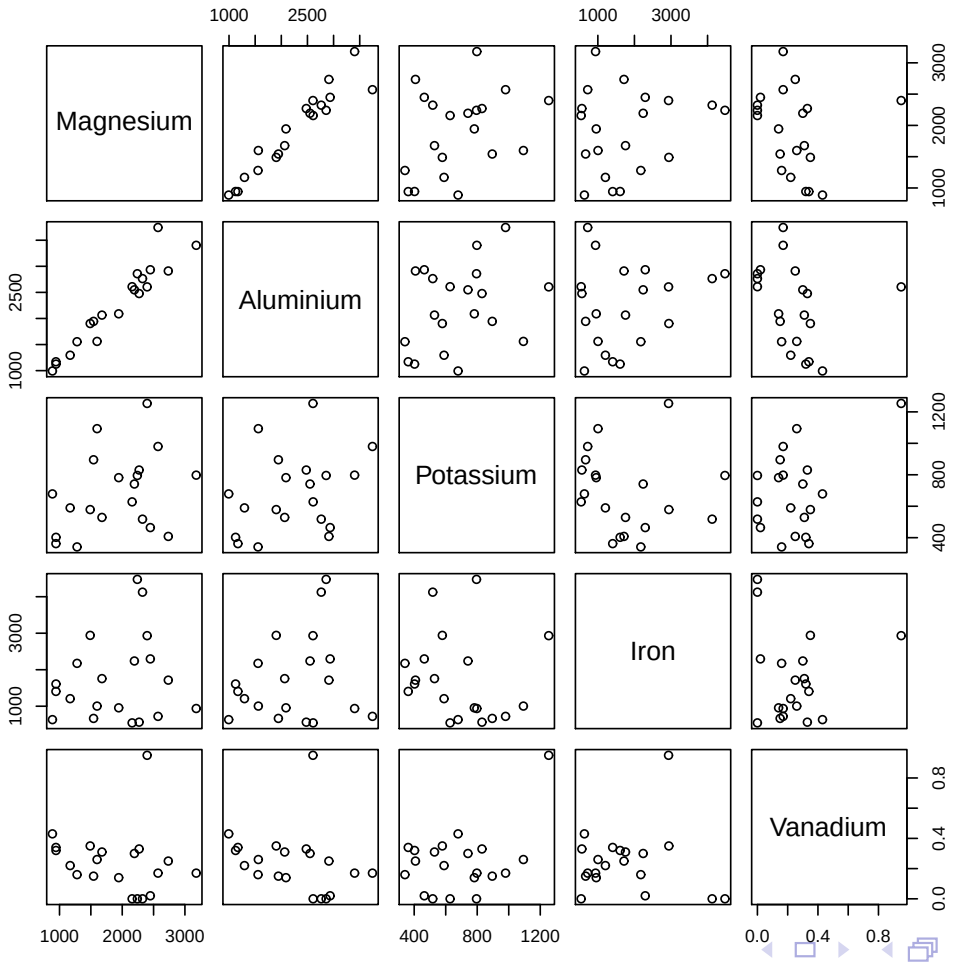
- Composição química das partículas totais em suspensão (PTS) no ar de Vitória - ES, estação de monitoramento da qualidade do ar localizada na Enseada do Suá. Esse ponto de monitoramento sofre influência de intenso tráfego de veículos, de inúmeras obras de construção cível e de diversos pontos comerciais e habitações.
- Concentração de PM_{10} nas estações de monitoramento.
- Dados do incomodo, tese de Melo(2015).

Abaixo segue exemplo no R com dados da PTS.

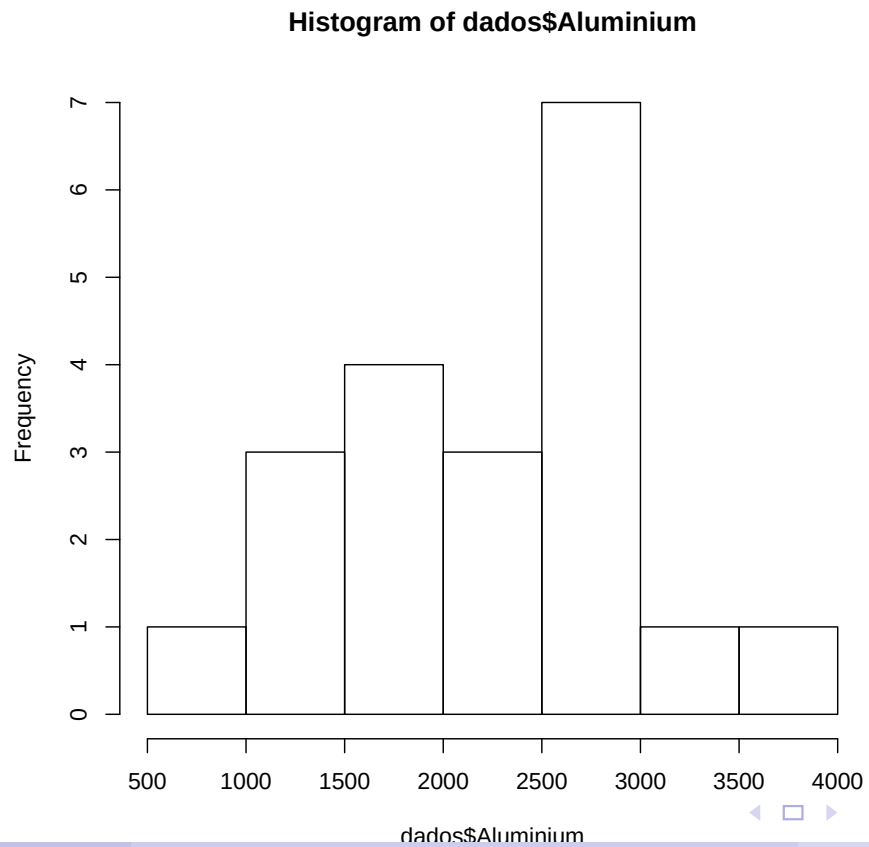
Tabela: Um parte dos dados do PTS, estação Enseada do Suá

Magnesium	Aluminium	Potassium	Iron	Vanadium
942.39	1169.99	362.15	1405.53	0.34
943.56	1125.26	402.34	1608.91	0.32
887.02	993.35	678.43	632.16	0.43
1169.38	1295.15	589.45	1206.42	0.22
1943.92	2088.97	781.61	955.63	0.14
3180.21	3402.11	797.61	936.82	0.17
1488.81	1902.89	578.63	2941.05	0.35
1676.74	2064.45	528.86	1756.21	0.31
1542.77	1944.23	895.43	665.42	0.15
2269.67	2477.54	830.23	562.80	0.33
2570.94	3743.67	979.91	722.07	0.17
2397.51	2605.54	1253.62	2933.09	0.95
1598.65	1560.58	1093.16	1003.79	0.26
2157.69	2608.90	627.72	543.72	0.00
2448.39	2933.73	464.11	2295.79	0.02
1279.70	1554.54	341.20	2174.49	0.16
2323.49	2764.02	517.79	4125.14	0.00
2241.20	2857.59	795.51	4475.25	0.00
2194.07	2548.98	741.17	2239.09	0.30
2734.93	2910.26	408.03	1714.48	0.25

> #ler os dados via memória
> dados=read.table("clipboard",h=T)
> #constroi um diagrama de dispersão dos elementos químicos do poluente PTS
> plot(dados)



> #obtem algumas medidas descritivas do elemento químico Aluminium
> summary(dados$Aluminium)
Min. 1st Qu. Median Mean 3rd Qu. Max.
993.4 1559.0 2283.0 2228.0 2787.0 3744.0
> #faz o histograma dos dos dados do Alumininum
> hist(dados$Aluminium)



Trecho da pergunta "Na opinião do(a) Sr(a) qual é a principal fonte de poluição do ar na sua região? "do questionário aplicado por Melo(2015).

QuestaoC1	Rótulo
2	industrial
2	industrial
2	industrial
2	industrial
2	industrial
2	industrial
4	constrcivil
6	pedreiras
6	pedreiras
2	industrial
4	constrcivil
2	industrial
2	industrial
4	constrcivil
2	industrial
1	veicular
2	industrial
1	veicular
1	veicular
2	industrial

```
> #ler os dados via memória
> dados2=read.table("clipboard",h=T)
> #constroi uma tabela para a questão C1
> table(dados2)
dados2
1 2 4 6
3 12 3 2
```

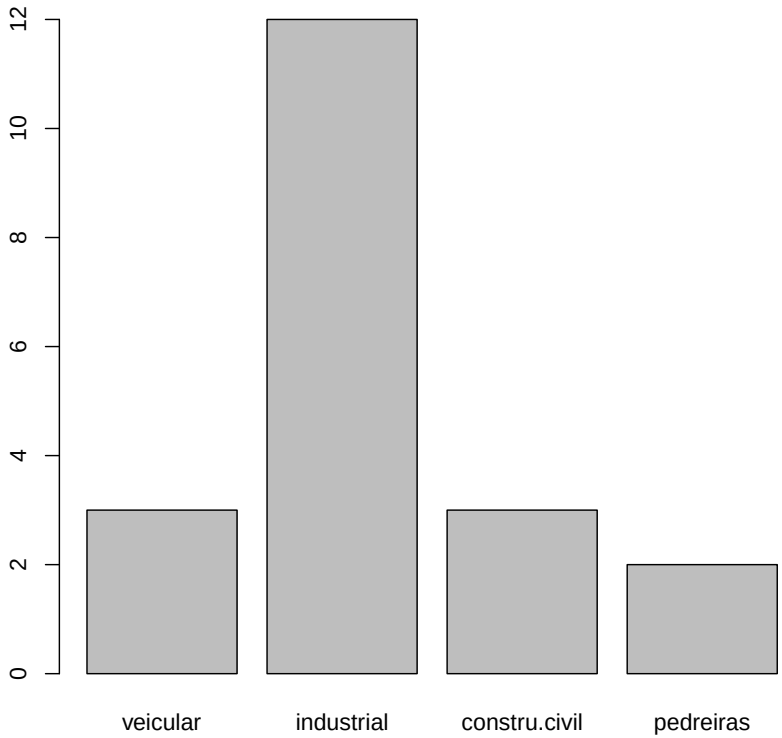
Aqui observa que o código 1 aparece 3 vezes, o código 2 aparece 12 vezes, ...

Para colocar o rótulo da variável, faça:

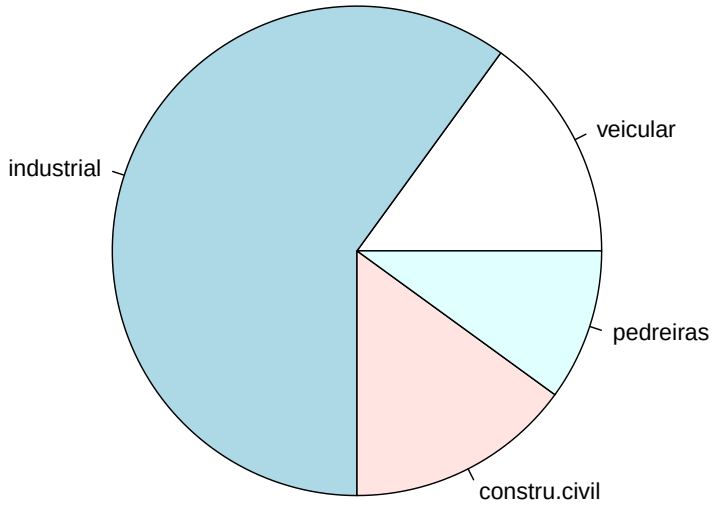
```
> questaoC1=table(dados2)
> names(questaoC1)<-
c("veicular","industrial","constru.civil","pedreiras")
> questaoC1
veicular industrial constru.civil pedreiras
3         12         3              2
```

```
> #Em percentual
> prop.table(questaoC1)
veicular industrial constru.civil pedreiras
0.15        0.60        0.15        0.10
> #gráfico de barras
> barplot(questaoC1)
> #gráfico de pizza
> pie(questaoC1)
```

Os gráficos foram colocados abaixo.



(a) Gráfico de barras



(b) Gráfico de pizza

Figura: Na opinião do(a) Sr(a) qual é a principal fonte de poluição do ar na sua região?

1. MORETTIN, P. A. e BUSSAB, W. O. (2010). Estatística Básica. 6a ed. São Paulo: Saraiva.
2. MARTINS, Gilberto de Andrade e DOMINGUES, Osmar. (2011) Estatística Geral e Aplicada. 4a ed. São Paulo: Atlas.
3. Reinsen, V. A. e Silva, A. N. (2011). O uso da linguagem R para cálculos de Estatística Básica. Edufes.
4. R: linguagem de programação usada para análise estatística e gráficos.
5. Download: https://www.r-project.org/
6. Editores para R:
   - R-Studio: https://www.rstudio.com/
   - Tinn-R: https://sourceforge.net/projects/tinn-r/
7. Melo, M. M. (2015). Tese: "Correlação entre percepção do incômodo e exposição ao material particulado presente na atmosfera e sedimentado". PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA AMBIENTAL (UFES).