Aula de revisão: Estatística Descritiva

Estatística e Probabilidades

https://avramaral.github.io/AulasEstProb/

André Victor Ribeiro Amaral

avramaral@gmail.com

O que é Estatística?

Estatística é um conjunto de técnicas que permite, de forma **sistemática**, *organizar*, *descrever*, *analisar* e *interpretar* dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

Conceitos importantes

Ao conjunto de elementos (pessoas, objetos, etc.) que possuem pelo menos uma característica em comum e de interesse do pesquisador, damos o nome de **população**.

Exemplos: população brasileira, produção de uma fábrica em determinado dia, etc.

Porém estudar toda a população pode ser tarefa difícil (ou impossível). Por isso, estamos normalmente interessados em algum subconjunto da população (que represente "o todo" da maneira mais fiel possível), ao qual damos o nome de **amostra**.

Exemplos: habitantes do estado de São Paulo, itens fabricados por uma fábrica no turno da noite, etc.

Resolução de problemas

É possível representar as componentes que acabamos de descrever através do diagrama abaixo.

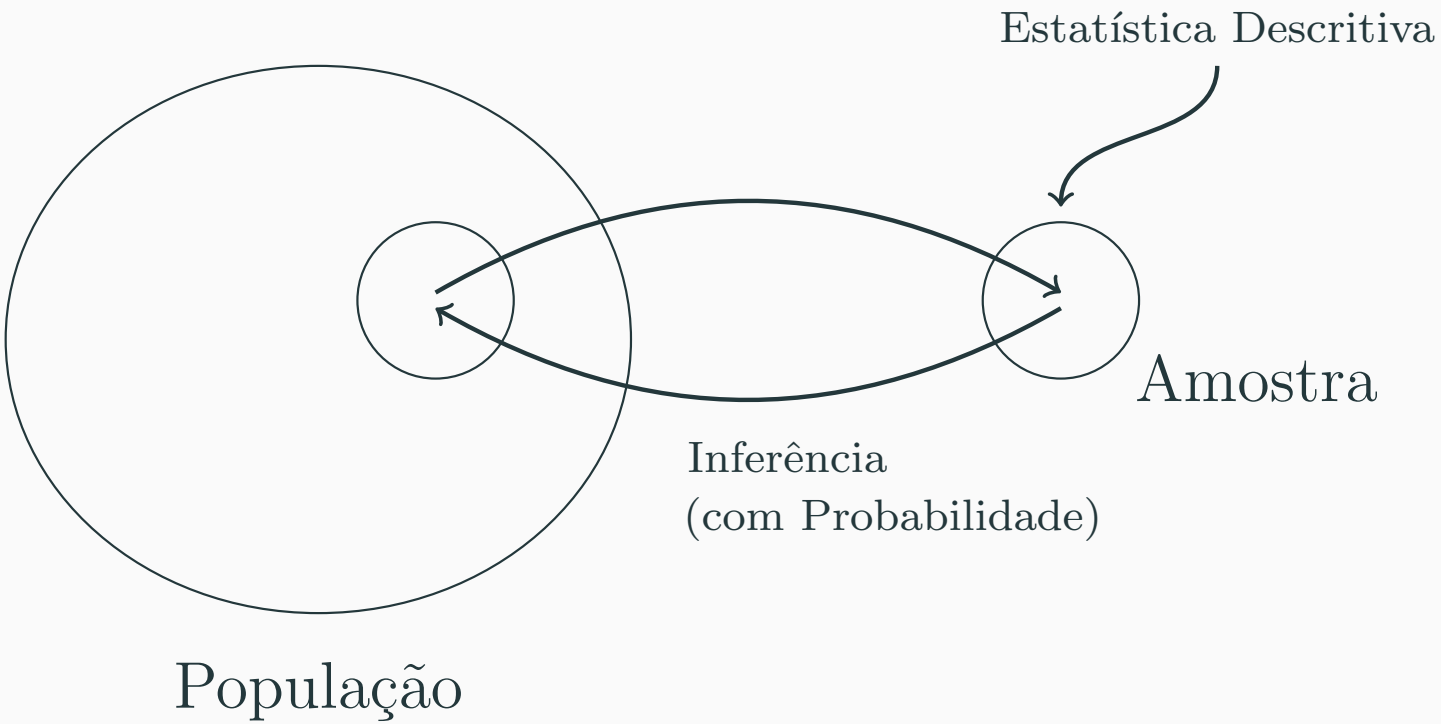


Figura 1: *Simplificação* do processo de resolução de problemas.

Tipos de dados

Em relação às variáveis, podemos classificá-las como:

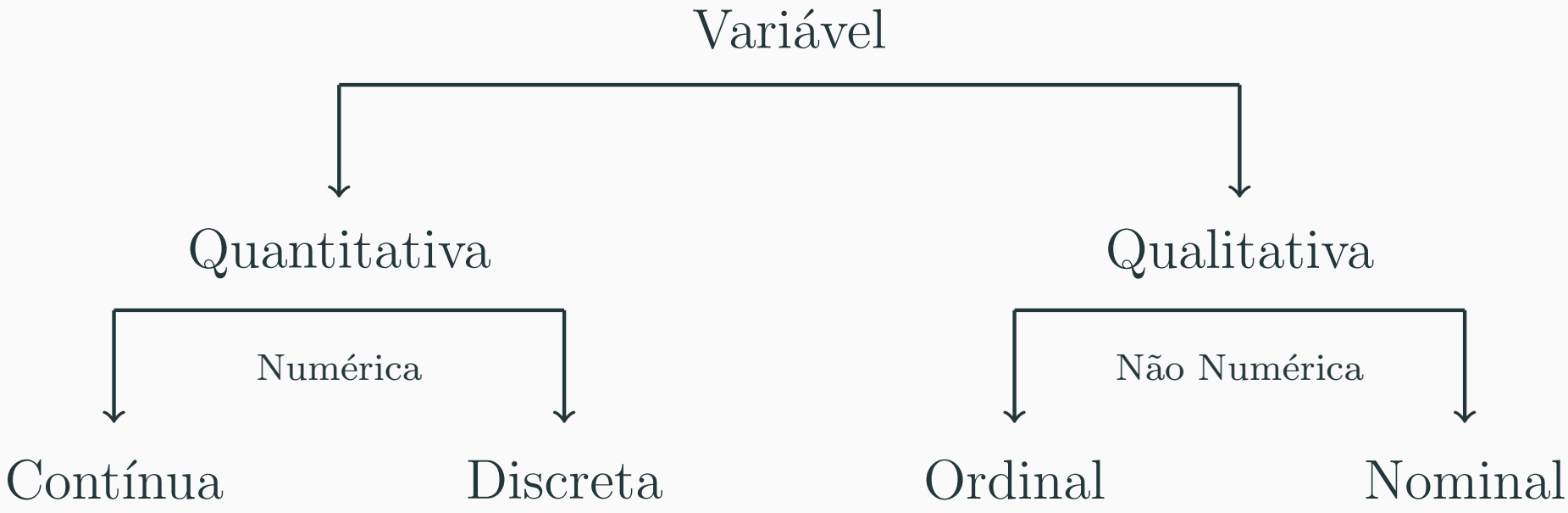


Figura 2: Tipos de dados.

Tipos de dados

1. Quantitativo
   - Discreto: O conjunto resposta é enumerável. Em geral, é resultado de um processo de contagem. *Exemplos*: idade (em anos completos), número de irmãos/irmãs, etc.
   - Contínuo: O conjunto resposta é não-enumerável. Potencialmente um subconjunto dos números reais. *Exemplos*: temperatura, tempo, peso, etc.
2. Qualitativo (Variáveis Categóricas)
   - Nominal: O conjunto resposta não possui ordenação. *Exemplos*: raça, sexo, etc.
   - Ordinal: O conjunto resposta possui uma ordenação natural. *Exemplos*: classe social, escolaridade, etc.

Tabela de frequência

Assuma o seguinte conjunto de idades (em anos), com $n = 36$:

17, 18, 18, 18, 19, 20, 18, 21, 20, 22, 21, 19, 21, 22, 23, 23, 22, 22, 18, 19, 18, 20, 21, 18, 21, 22, 18, 21, 17, 20, 21, 21, 18, 21, 20, 19,

com frequência relativa $f_i = \frac{n_i}{n}$.

Aqui,

- n_i: é a contagem de cada classe de respostas; e
- f_i: é a frequência relativa para cada classe de respostas.

Além disso, podemos definir frequência acumulada como $f_{ac}^{(i)} = \sum_{j=1}^{i} f_j$.

Tabela de frequência

Classe (Idade)	n_i	f_i	$f_{ac}^{(i)}$
17	2	0.06	0.06
18	9	0.25	0.31
19	4	0.11	0.42
20	5	0.14	0.56
21	9	0.25	0.81
22	5	0.14	0.95
23	2	0.05	1.00
total	$n = 36$	1	–

Tabela 1: Tabela de frequência (discreto) para conjunto de idades.

Tabela de frequência

Assuma o seguinte conjunto de pesos (em kg), com $n = 24$:

60.8, 80.1, 70.3, 50.6, 52.0, 71.9, 99.9, 90.0, 79.8, 96.1, 59.4, 50.1, 58.8, 61.9, 70.2, 90.9, 59.9, 65.8, 99.2, 80.6, 70.8, 75.2, 55.1, 73.2.

Classe (Peso)	n_i	f_i	$f_{ac}^{(i)}$
$[50,\ 60)$	7	0.29	0.29
$[60,\ 70)$	3	0.12	0.42
$[70,\ 80)$	7	0.29	0.71
$[80,\ 90)$	2	0.08	0.79
$[90, 100]$	5	0.21	1.00
total	$n = 24$	1	–

Tabela 2: Tabela de frequência (contínuo) para conjunto de pesos.

Tabela de frequência

Nesse caso (dos pesos), perceba que acabamos por definir a amplitude de cada uma das classes de maneira arbitrária; nesse caso, 10 kg.

Via de regra, queremos definir um tamanho de amplitude tal que existam algo entre 5 e 8 classes. Assim:

$$\text{ampl.} = \frac{max - min}{\text{nº de classes}},$$

onde nº de classes $\in \{5, 6, 7, 8\}$.

No nosso exemplo, $\frac{99.9-50.1}{5} = 9.96 \approx 10$.

Tabela de frequência

Também podemos construir tabelas bivariadas. Suponha que, para o conjunto de idades que listamos no Slide 7, os 18 primeiros indivíduos são do sexo feminino e os 18 indivíduos seguintes são do sexo masculino; então:

	Faixa Etária			
Sexo	$[17, 19)$	$[19, 21)$	$[21, 23]$	Total
F	5 (27.8%)	4 (22.2%)	9 (50.0%)	18 (100%)
M	6 (33.3%)	5 (27.8%)	7 (38.9%)	18 (100%)
Total	11 (30.5%)	9 (25.0%)	16 (44.5%)	36 (100%)

Tabela 3: Tabela de frequência bivariada (com soma por linha).

Tabela de frequência

	Faixa Etária			
Sexo	$[17, 19)$	$[19, 21)$	$[21, 23]$	Total
F	5 (27.8%)	4 (22.2%)	9 (50.0%)	18 (100%)
M	6 (33.3%)	5 (27.8%)	7 (38.9%)	18 (100%)
Total	11 (30.5%)	9 (25.0%)	16 (44.5%)	36 (100%)

Comentários:

1. 27.8% das mulheres têm menos que 19 anos, ao passo que 33.3% dos homens pertencem à essa faixa etária;
2. Existe maior porcentagem de indivíduos (independente do sexo) entre 21 e 23 anos (inclusive), etc.

Gráfico

A fim de facilitar a visualização e interpretação de um determinado conjunto de dados, podemos utilizar diferentes tipos de **gráficos** para representá-lo.

Via de regra, é necessário que um gráfico contenha:

- Título (quando o gráfico está inserido em um documento que apresente contexto das informações, o "título" pode ser dispensado);
- Eixo horizontal;
- Eixo Vertical; e
- Legenda.

Gráfico de barra

Gráfico de barra: esse tipo de gráfico se adapta melhor às variáveis discretas ou qualitativas ordinais.

Sobre o Slide 8,

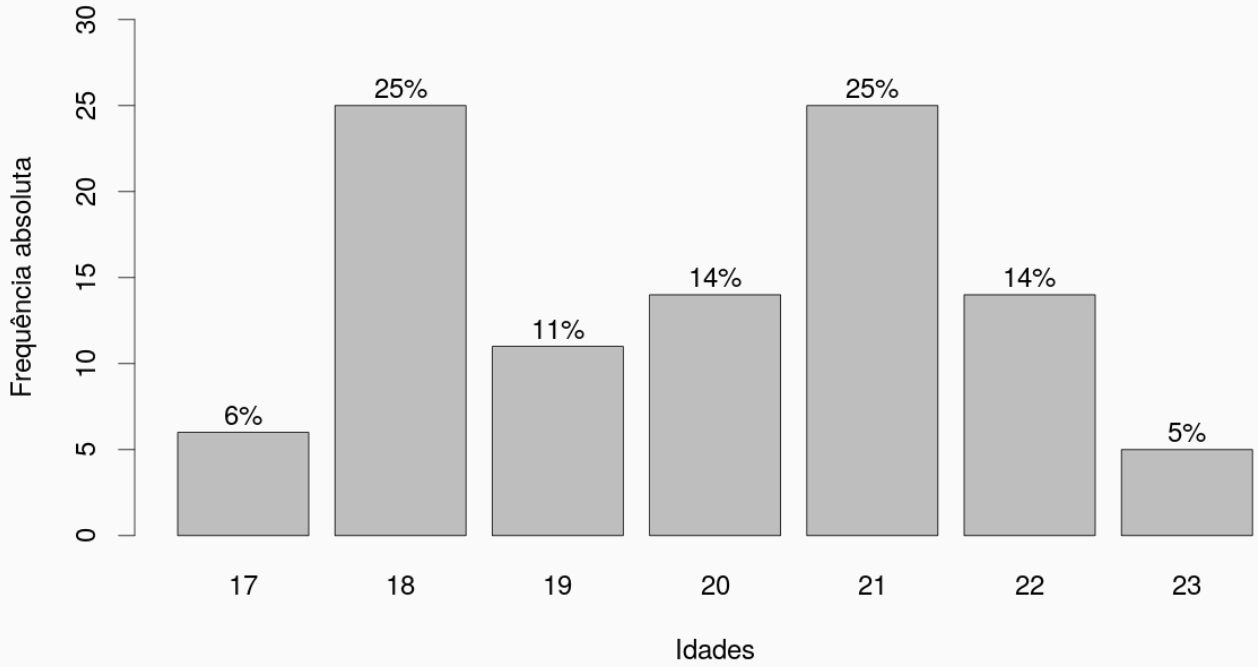


Figura 3: Gráfico de barras para frequência de idades.

Gráfico de linhas

Gráfico de linhas: aplicável, principalmente, para séries temporais. Exibe um a evolução de uma variável ao longo de, por exemplo, o tempo.

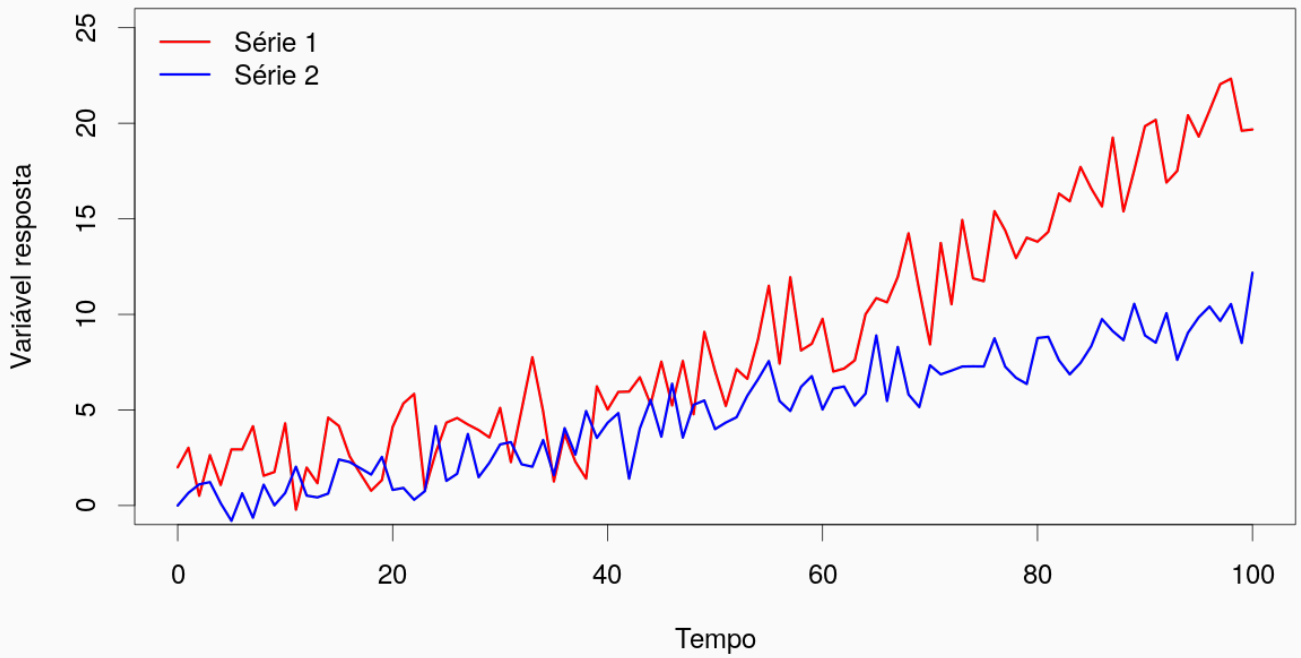


Figura 4: Gráfico de linhas para duas séries temporais arbitrárias.

Histograma

Histograma: utilizado para representar variáveis contínuas. Como exemplo, vamos fazer referência à tabela de frequência que construímos no Slide 9.

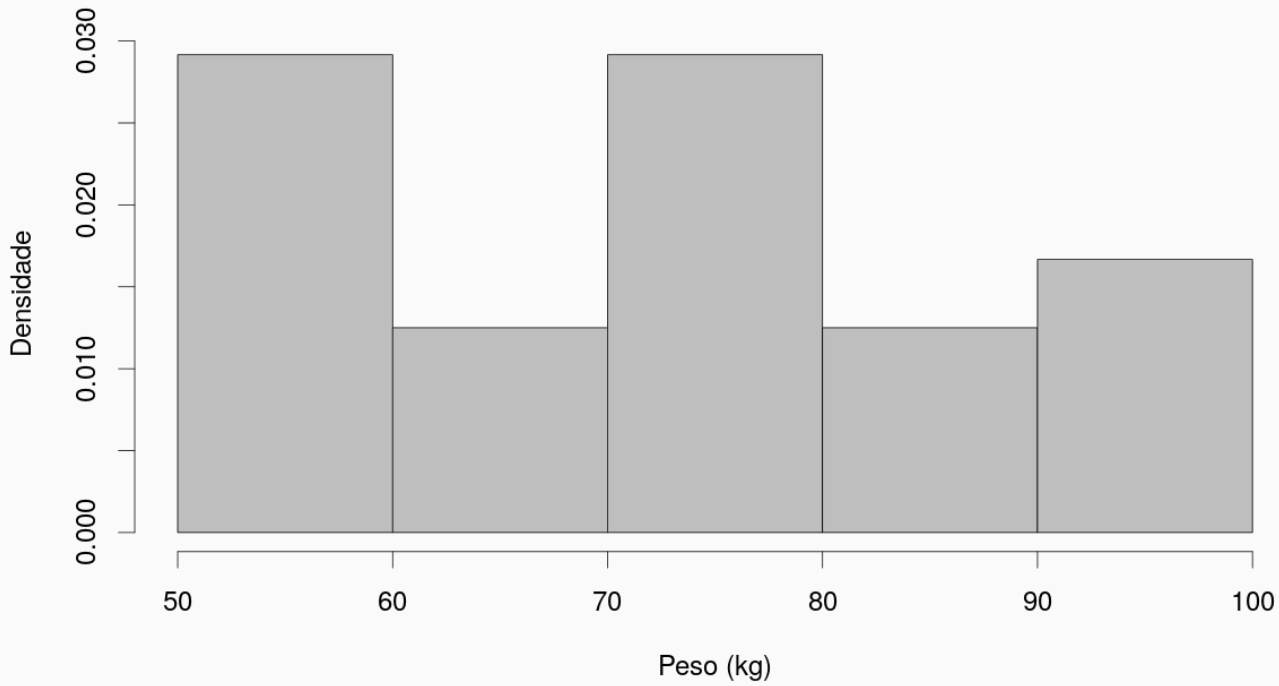


Figura 5: Histograma para os pesos (em kg).

Distribuição de frequência e Medidas Resumo

Ao estudarmos a distribuição de frequência de uma variável quantitativa − seja apenas de um grupo, seja comparando grupos −, devemos verificar basicamente três características:

- Tendência central: o que é mais frequente?
- Variabilidade
- Forma: a distribuição é simétrica ou assimétrica?

Medidas de tendência central

Medidas de tendência central fornecem uma ideia do comportamento central dos dados; ou seja, os valores mais comuns na amostra.

Exemplo: média, moda e mediana.

Usualmente se posicionam nas regiões do gráfico com maior frequência.

Média

Denote as n observações que compõem uma **amostra** por $x_1, \cdots, x_n$; então, a **média amostral** é definida por:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

onde $\bar{x}$ é estimador.

Para uma amostra dada por "$1, 1, 3, 5, 2, 1, 2, 2, 3, 2$", temos

$$\bar{x} = \frac{1+1+3+5+2+1+2+2+3+2}{10} = 2.2$$

A fim de verificar formas de calcular a média amostral a partir de tabelas de frequência, bem como alternativas ao cálculo de $\bar{x}$ (e.g., média aparada, média ponderada, etc.), estudar `Aula 02`.

Moda

A **moda** de uma amostra é aquele valor que ocorre com mais frequência; ou seja, aquele que mais se repete.

Observações:

- Se dois valores ocorrem com a mesma frequência máxima, a variável é bimodal;
- Se mais de dois valores ocorrem com a mesma frequência máxima, a variável é multimodal; e
- Quando nenhum valor se repete, a variável não tem moda.

Exemplos:

1. Amostra$_1$: 1, 1, 1, 2, 2, 2, 2, 3, 3, 5; Aqui, $Mo = 2$.
2. Amostra$_2$: 1, 1, 1, 2, 2, 2, 3, 3, 3, 5; Aqui, $Mo = 1, 2$ e 3.

Mediana

A **mediana**, representada por Md, é o valor que ocupa a posição central dos dados *ordenados*.

Definição: a mediana é **qualquer** valor tal que 50% das observações são menores ou iguais a ele.

Como regra para cálculo da mediana, podemos adotar:

- se n é ímpar, a mediana será o valor que ocupa a posição do meio dentre no conjunto de dados *ordenado*.
- se n é par, a mediana será o ponto médio entre os dois valores que o ocupam as posições centrais no conjunto de dados *ordenado*.

Comparação de medidas de tendência central

Média

- Vantagem: leva em conta todos os valores da amostra e é utilizada em muitos métodos estatísticos.
- Desvantagem: é afetada por valores extremos.

Moda

- Vantagem: não é afetada por valores extremos.
- Desvantagem: não leva em conta todos os valores da amostra; além disso, é raramente utilizada e pode nem existir.

Mediana

- Vantagem: é utilizada com frequência e não é afetada por valores extremos.
- Desvantagem: não leva em conta todos os valores da amostra.

Percentil de ordem $\alpha\%$

Percentil de ordem $\alpha\%$: é definido como qualquer número tal que $\alpha\%$ das observações são menores ou iguais ao valor do percentil.

Observações:

- Os percentis de ordem 25%, 50% e 75% são denotados por primeiro, segundo e terceiro quartil, respectivamente;
- O percentil de ordem 50% é a mediana; e
- Os percentis de ordem 10%, 20%, $\cdots$, 90% também são chamados de decis.

Medida de variabilidade

Medidas **de variabilidade** (ou **de dispersão**) são medidas que tentam quantificar o "espalhamento" dos dados

Nesse sentido, as principais medidas de variabilidade são **variância** e **desvio padrão**. Quanto maior qualquer um desses valores, maior a variação dos dados em torno da média.

A variância e desvio padrão na amostra são definidos por

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

onde s^2 é a **variância** e s é o **desvio padrão** da amostra.

Medida de assimetria

Uma distribuição de dados é assimétrica quando se estende mais para um lado do que para o outro.

Uma das medida para esse tipo de comportamento é chamada de "**Coeficiente de Assimetria de Pearson**", definida por:

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

Aqui, se $A_p \geq 1$ ou $A_p \leq -1$, então os dados podem ser considerados fortemente assimétricos.

Observações atípicas (*Outliers*)

Observações atípicas (ou ***Outliers***) são valores muito altos ou muito baixos em relação ao restante do conjunto de dados.
Podem ser divido em dois tipos:

1. Não genuíno: erro de digitação, erro de medição, etc.
2. Genuíno: não são resultantes de erros e são valores importante ao estudo, podendo fornecer informações valiosas sobre a característica que está sendo estudada.

Em caso de *outliers* não genuínos, estes valores devem ser corrigidos ou excluído do banco de dados. Em contrapartida, para valores atípicos genuínos, devemos deixá-los no banco de dados, tomando cuidado com interpretações futuras.

Escore padronizado

O **escore padronizado** (ou "escore z") representa o número de desvios padrão pelo qual uma observação x_i dista da média (para mais ou para menos); e é calculado como

$$z = \frac{x_i - \bar{x}}{s}.$$

- O escore padronizado permite distinguir entre os valores usuais e valores raros.
- São considerados valores usuais os que possuem escore z entre -2 e 2; e raros os que possuem escore z menor que -2 ou maior que 2.

Boxplot

Boxplot é um gráfico em forma de caixa que utiliza os quartis.

Sejam Q_1, Q_2 e Q_3 o primeiro, segundo e terceiro quartis, respectivamente. Além disso, defina $DI = Q_3 - Q_1$ como "distância interquartílica".

Boxplot

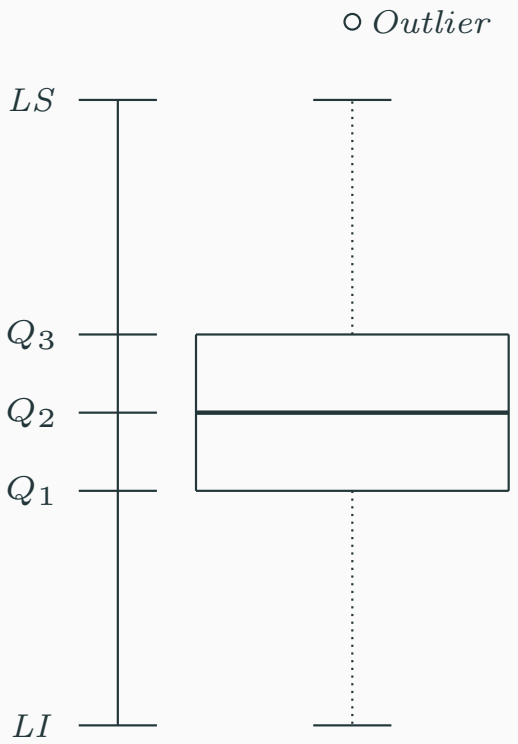


Figura 6: Boxplot.

Definimos, baseado na figura da esquerda:

- $LI = Q_1 - 1.5 \cdot DI$; e
- $LS = Q_3 + 1.5 \cdot DI$.

Assim, dados menores que LI e maiores que LS são marcados com "pontos" e são chamados de *outliers*.

Boxplot

O boxplot pode ser usado, principalmente, para:

- Detecção de valores discrepantes;
- Comparação da tendência central (mediana) de dois ou mais conjuntos de dados; e
- Comparação da variabilidade de dois ou mais conjuntos de dados.

Associação entre variáveis

Um dos principais objetivos de uma distribuição conjunta é descrever a associação que existe entre as variáveis.

Nesse caso, conhecer o grau de dependência entre as variáveis nos ajuda a predizer melhor o resultado de uma delas quando conhecemos a realização da outra.

Exemplos: a relação que existe ao compararmos indivíduos com respeito às seguintes características

- Sexo e peso;
- Idade e peso, etc.

Para cálculo do **coeficiente de contingência** para associação de variáveis categóricas, estudar `Aula 03`.

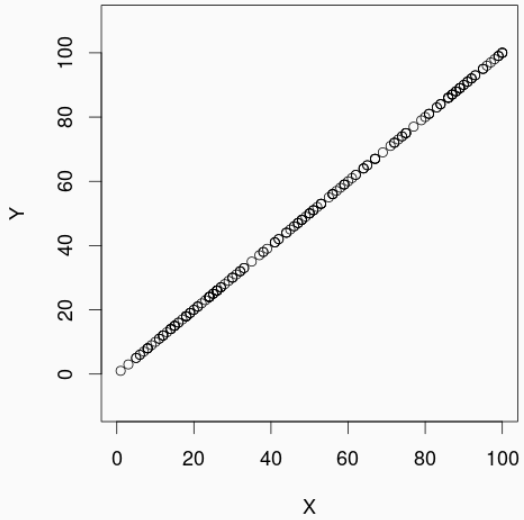
Associação entre variáveis quantitativa

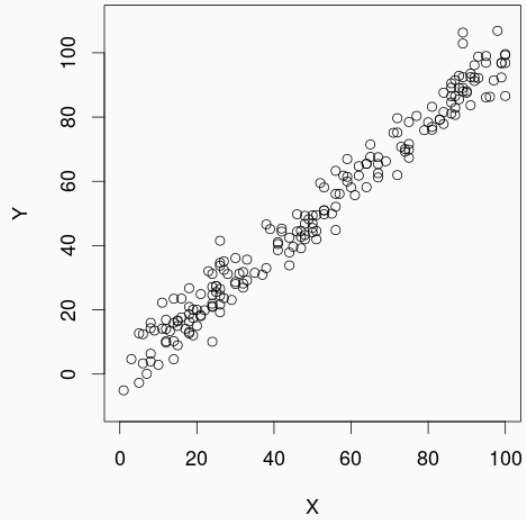
Para analisar a associação que existe entre variáveis quantitativas, a nossa primeira ferramenta será o gráfico de dispersão.

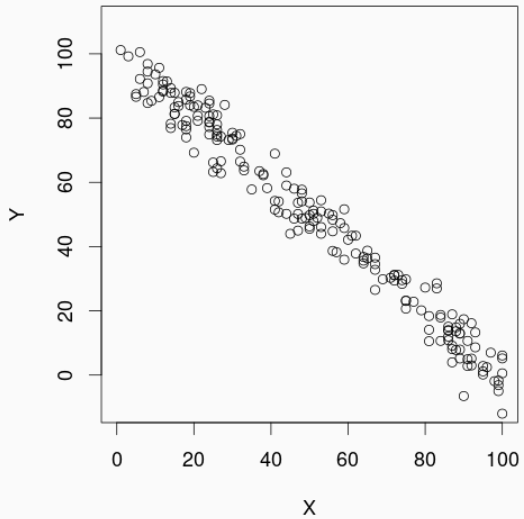
O objetivo desse tipo de gráfico é tentar descobrir se existe relação entre duas variáveis quantitativas (por exemplo, X e Y).

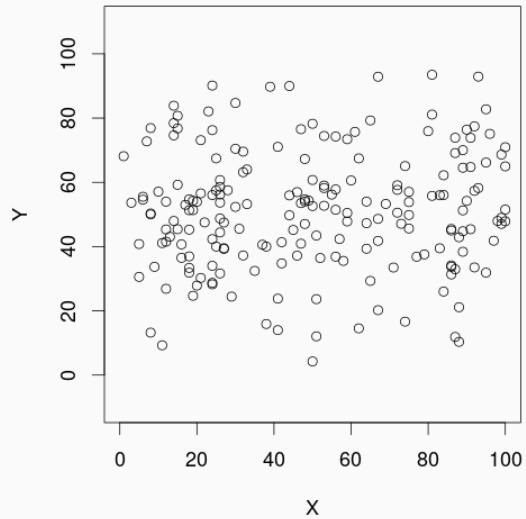
No gráfico de dispersão, cada indivíduo da amostra é representado por um ponto de tal forma que o eixo horizontal represente seu valor para a variável X e o eixo vertical represente o seu valor para a variável Y.

Gráfico de dispersão









Coeficiente de Correlação de Pearson

Mais uma vez, ao invés de nos prendermos, somente, à interpretação visual dos gráficos, vamos quantificar a medida correlação entre duas variáveis X e Y. Para isso, considere o "Coeficiente de Correlação de Pearson".

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \bar{x})^2)(\sum_{i=1}^{n}(y_i - \bar{y})^2)}},$$

onde r é o Coeficiente de Correlação de Pearson **amostral**.

Coeficiente de Correlação de Pearson

Observações importantes

1. $-1 \leq r \leq 1$, de modo que:
   - se $r \approx +1 \rightarrow$ forte correlação **linear** positiva entre X e Y;
   - se $r \approx -1 \rightarrow$ forte correlação **linear** negativa entre X e Y; e
   - se $r \approx \ 0 \rightarrow$ não existe correlação **linear** entre X e Y.
2. Uma possível classificação mais refinada para essa medida é dada por:
   - se $0 \leq |r| < 0.4$, então existe correlação fraca;
   - se $0.4 \leq |r| < 0.7$, então existe correlação moderada;
   - se $0.7 \leq |r| < 1$, então existe correlação forte; e
   - se $|r| = 1$, então existe correlação perfeita.

Correlação e causalidade

Alto coeficiente de correlação **não** se traduz, necessariamente, em causalidade.

Correlação indireta: existe correlação entre duas variáveis, mas isso é justificado por uma terceira variável (que pode ser conhecida ou não).

Exemplo: Número de palavras que uma criança conhece com a altura desse(a) menino(a).

Correlação espúria: $|r|$ é alto, mas não existe relação alguma entre as variáveis X e Y.

Exemplo: Em certa região da Europa registrou-se aumento de avistamento de cegonhas e aumento da taxa de natalidade.