Bioestatística

para Biologia, Enfermagem, Farmácia e Nutrição

Códigos: MAD231 / MAD125 / MAD 237 / MAD120

Oferecido pelo:

DEPARTAMENTO DE MÉTODOS ESTATÍSTICOS

INSTITUTO DE MATEMÁTICA

CENTRO DE CIÊNCIAS MATEMÁTICAS E DA NATUREZA

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO

Bibliografia

- **Livro-texto**: Freire, S., *Estatistica Básica*, 2020.
- Pagano, M. e Gauvreau, K., *Princípios de Bioestatística*. 2ª Ed., Pioneira Thomson Learning, 2004.
- Soares, J. F. e Siqueira, A. L., *Introdução à Estatística Médica*. 2ª Ed, COOPMED, 2002.
- Bussab, W. O. e Morettin, A. M., *Estatística Básica*. Saraiva, 2009.
- Triola, M. F., *Introdução à Estatística*. 10ª Ed., LTC, 2008.
- Apostila: Santos, R., *Material Introdutório sobre R*.

Ementa

- Parte I - Análise Exploratória de Dados:
  - Tipos de variáveis.
  - Resumo tabular de dados.
  - Exibição gráfica de dados.
  - Resumo de dados por meio de estatísticas sumário.
  - Introdução a técnicas de amostragem.

Ementa

- Parte II - Probabilidade:
  - Fundamentos de Probabilidade.
  - Variáveis aleatórias discretas.
  - Variáveis aleatórias contínuas.
  - Vetores aleatórios.

Ementa

- Parte III - Inferência Estatística:
  - Estimadores pontuais.
  - Estimadores Intervalares.
  - Testes de Hipótese.
  - Comparação de Médias.
  - Analise de Variância.
  - Teste de Independência.
  - Regressão Linear
  - Outros tópicos

Análise Exploratória de Dados

Tópicos

1. Análise exploratória de dados (AED)
2. Tabelas
3. Gráficos
4. Estatísticas sumárias: medidas de posição
5. Estatísticas sumárias: quantis
6. Estatísticas sumárias: medidas de dispersão

Análise Exploratória de Dados

- Extrair informação geral da massa numérica de dados original.
- Vizualização da distribuição dos dados - **gráficos**.
- **Resumos numéricos** da massa de dados (medidas resumo: média, variância, correlação, etc.).

Objetivo: entender como os dados se comportam. AED é conjunto de métodos dedutivos (descrição somente do conjunto de dados observado). Entretanto, os elementos levantados na AED podem ser utilizados, posteriormente, para se fazer inferência (processo indutivo) sobre universos maiores que o conjunto de dados observado.

Tipos de Variáveis

- Variável é cada característica registrada sobre uma unidade de observação (ou seja, sobre um indivíduo ou elemento qualquer).
- São exemplos de variáveis: sexo, idade, grau de instrução, peso, altura, IMC, nível de colesterol, nível de glicose, etc.
- Importante determinar o tipo de cada variável a ser analisada: métodos mais adequados dependem do tipo da variável.

Tipos de Variáveis: Categóricas

- **Variáveis Categóricas** são aquelas em que se registra um atributo qualitativo do elemento observado.
- Assume-se que a variável categórica é definida sobre k níveis (categorias) e, ainda, que:
  - as k categorias são mutuamente excludentes, disjuntas ou mutuamente exclusivas (não há interseção entre categorias - cada indivíduo classificado em uma única categoria);
  - as k são exaustivas (a união das categorias dá conta de todas as possibilidades de classificação)

Variáveis Categóricas: Exemplos

- Sexo (Feminino/ Masculino) é variável categórica com $k = 2$ níveis.
- Grau de instrução é variável categórica que pode apresentar diferente quantidade de níveis, por exemplo:
  - ($k = 7$): sem formação, fundamental incompleto, fundamental completo, médio incompleto, médio completo, superior incompleto, superior completo;
  - níveis dicotomizados ($k = 2$), por exemplo: tem ou não nível superior completo.

Variáveis Categóricas - Nominais e Ordinais

- Variáveis Categóricas dividem-se entre os seguintes tipos:
  - **Categóricas Ordinais**: quando há alguma ordenação possível entre as categorias. Exemplo: grau de instrução, classe social;
  - **Categóricas Nominais**: quando não há qualquer ordenação possível entre as categorias. Exemplo: sexo, cor, região em que reside.

Variáveis Categóricas - Codificação

- É usual que variáveis categóricas sejam codificadas numericamente em bases de dados. Por exemplo:
  - uma variável dicotômica, como Sexo, pode ter codificação: (F-0 , M-1) ou (F-1, M-2);
  - uma variável com diversas categorias pode ser codificada numericamente, por exemplo: Região do Brasil em que mora (Norte - 1, Nordeste -2, Centro-Oeste - 3, Sudeste - 4, Sul - 5).
- A codificação numérica facilita a digitação dos dados.
- Os números usados na codificação são meros rótulos e não faz sentido realizar operações com esses rótulos. Por exemplo, não faz sentido calcular a região **média** observada.

Variáveis Quantitativas

- Variáveis quantitativas têm valores numéricos que se originam de processos de contagem ou mensuração.
- Dividem-se em:
  - **Quantitativas discretas**: originadas, tipicamente, de processos de contagem. São exemplos: número de quartos em uma residência, número de leucócitos a cada 1μL;
  - **Quantitativas contínuas**: originadas, tipicamente, por processo de mensuração. São exemplos: nível pressórico, peso, altura.
- É lícito realizar operações aritméticas com variáveis quantitativas, por exemplo calcular o peso médio entre os indivíduos observados.
- Podem ser categorizadas. Exemplo: nível pressórico (baixo, normal, alto).

Análise Exploratória de Dados

- A seguir são apresentados instrumentos para exploração dos diferentes tipos de variáveis, por meio de:
  - tabelas;
  - gráficos; e
  - estatísticas sumárias.

Exemplo: Partos Prematuros

Tabela 1: Histórico de partos prematuros em 189 gestantes.

| | | |

Exemplo: Partos Prematuros

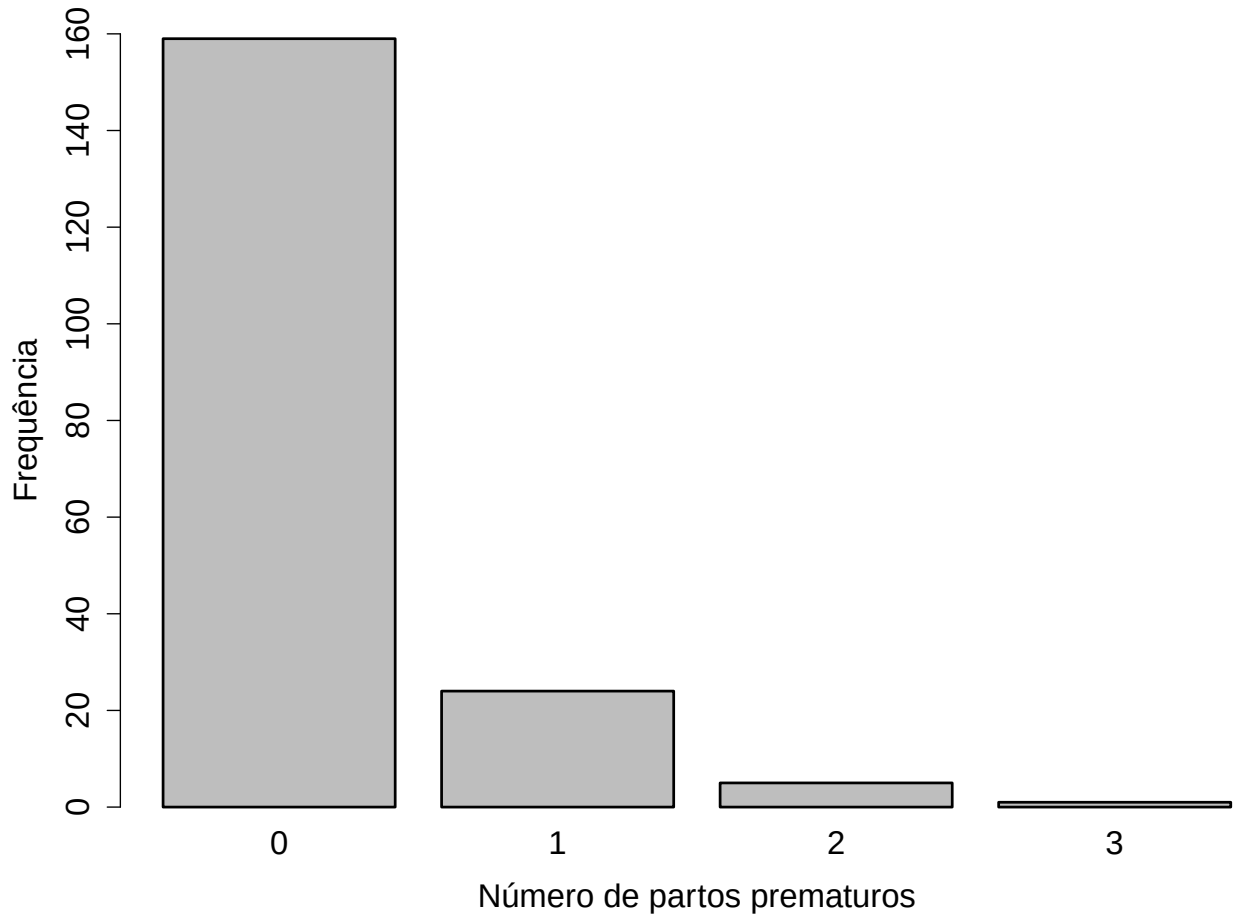


Figura 1: Frequência (absoluta) do número de partos prematuros.

Tabelas de Frequências

- Resumos **numéricos** da massa de dados.
- A construção das tabelas varia de acordo com o tipo de variável:
  - Variável Qualitativa (Ordinal ou Nominal);
  - Variável Quantitativa Discreta; ou
  - Variável Quantitativa Contínua.

Tabela de Frequência para Variáveis Qualitativas e Quantitativas Discretas

- Contagem de quantas observações pertencem à cada valor possível da variável.
- Pode conter frequências relativas.

Tabela 2: Histórico de partos prematuros de 189 gestantes.

	0	1	2	3	Total
Frequência absoluta	159	24	5	1	189
Frequência relativa (%)	84,1	12,7	2,5	0,5	100

Tabelas de Frequência para Variáveis Quantitativas Discretas

Exibir a frequência de cada valor observado da variável.

- 20 domicílios de classe A no Rio de Janeiro.
- Variável: número de filhos por família.
- Observações: 0, 1, 0, 1, 2, 1, 0, 0, 1, 0, 1, 4, 2, 1, 3, 1, 2, 1, 1, 1.

Tabelas de Frequência para Variáveis Quantitativas Discretas

Tabela 3: Frequência (absoluta) e relativa do número de filhos por família.

Nº de filhos	Frequência	Frequência Relativa
0	5	5/20 = 0,25 = 25%
1	10	10/20 = 0,50 = 50%
2	3	3/20 = 0,15 = 15%
3	1	1/20 = 0,05 = 5%
4	1	1/20 = 0,05 = 5%
Total	20	100%

Tabelas de Frequência para Variáveis Quantitativas Contínuas

- Valores em um intervalo da reta;
- Dividir em intervalos menores (classes); e
- Contar quantas observações caem em cada classe.

Perguntas:

- Como escolher as classes?
- Quantas classes? (Evitar extremos)

Observação: aplicável a variáveis discretas se as contagens são elevadas (número de pontos dos jogadores na NBA em um determinado ano).

Exemplo: Idades de Árvores em uma Floresta

Tabela 4: Idades (anos) de 30 árvores em uma floresta.

67	71	78	82	85	86	87	87	92	95
97	99	99	100	101	102	103	103	104	105
107	108	109	112	114	116	118	122	124	125

- Variável: idade da árvore (anos).
- Amostra de 30 árvores ($n = 30$).
- Número de classes: $\sqrt{n} \simeq 5,477$ (sugestão).

Passos para Elaboração da Tabela

Passo 1: Máximo e mínimo (67, 125).

Passo 2: Número de classes (k): 5 ou 6, por exemplo. Digamos 5.

Passo 3: Calcule a **amplitude** (A) da amostra (max-min): $125 - 67 = 58$.

Passo 4: Calcule a amplitude das classes (h): $h = \frac{A}{k}$.

$$h = \frac{58}{5} = 11,6 \approx 12.$$

Passo 5: Calcule os limites das classes. Comece pelo valor mínimo e vá somando h:

$$[67, 79)\ [79, 91)\ [91, 103)\ [103, 115)\ [115, 127)$$

Passo 6: Conte quantas observações pertencem à cada classe:

$$x \in [a, b) \Leftrightarrow a \leqslant x < b$$

Tabela 5: Frequência (absoluta) e relativa das idades das árvores.

Nº de filhos	Frequência	Frequência Relativa (%)
$67 \vdash 79$	3	$3/30 \times 100 = 10{,}00$
$79 \vdash 91$	5	$5/30 \times 100 = 16{,}67$
$91 \vdash 103$	8	$8/30 \times 100 = 26{,}66$
$103 \vdash 115$	9	$9/30 \times 100 = 30{,}00$
$115 \vdash 127$	5	$5/30 \times 100 = 16{,}67$
Total	30	100%

Representações Gráficas

- Aprender a interpretar gráficos.
- Diferentes tipos $\Rightarrow$ diferentes funcionalidades.
- Todo gráfico deve "contar uma história".
- Intencionalidade e senso crítico.
- Ferramenta exploratória que pode ser usada em outras etapas da pesquisa estatística (análise dos dados, divulgação dos resultados ou até mesmo testes formais!)

Gráfico Setorial para Variáveis Qualitativas

Gráfico setorial 2D $\Rightarrow$ Bem aceito.
Gráfico setorial 3D $\Rightarrow$ **Heresia**.

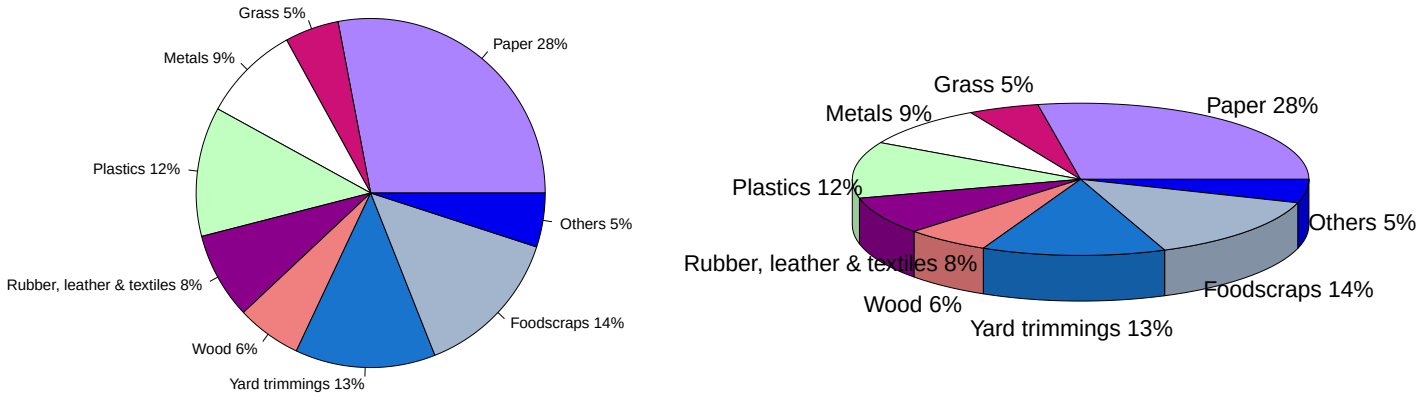


Figura 2: *Municipal Solid Generation* 2010.

Gráfico de Barras

- Representação gráfica da tabela de frequências.
- Aplicável para variáveis quantitativas ou qualitativas. As observações devem estar divididas em classes.

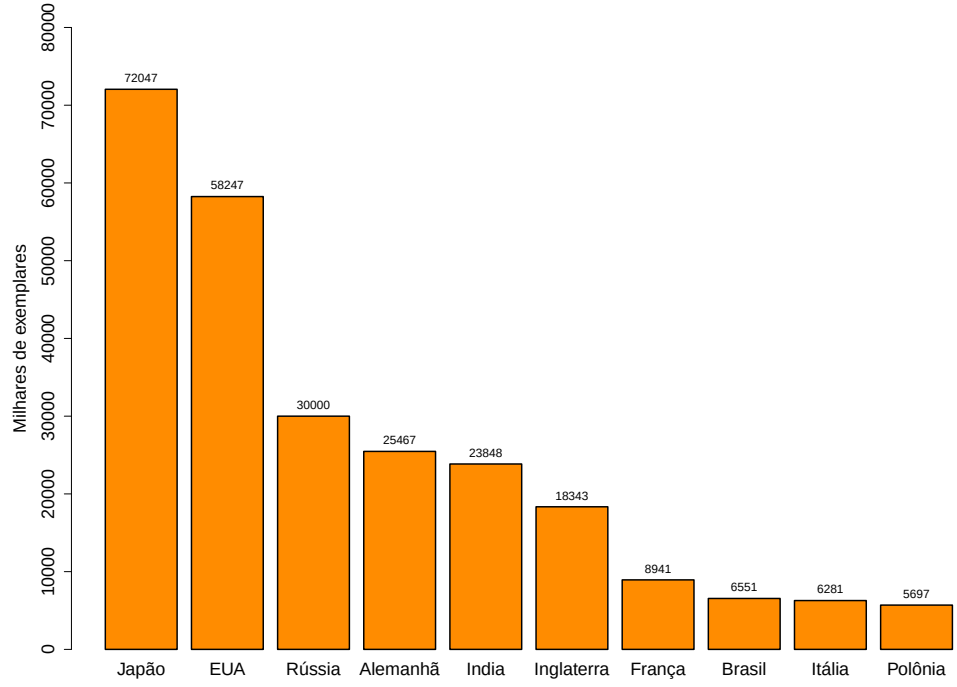


Figura 3: Número de cópias de jornais que circulam diariamente.

Gráfico de Barras

Também podem vir na horizontal.

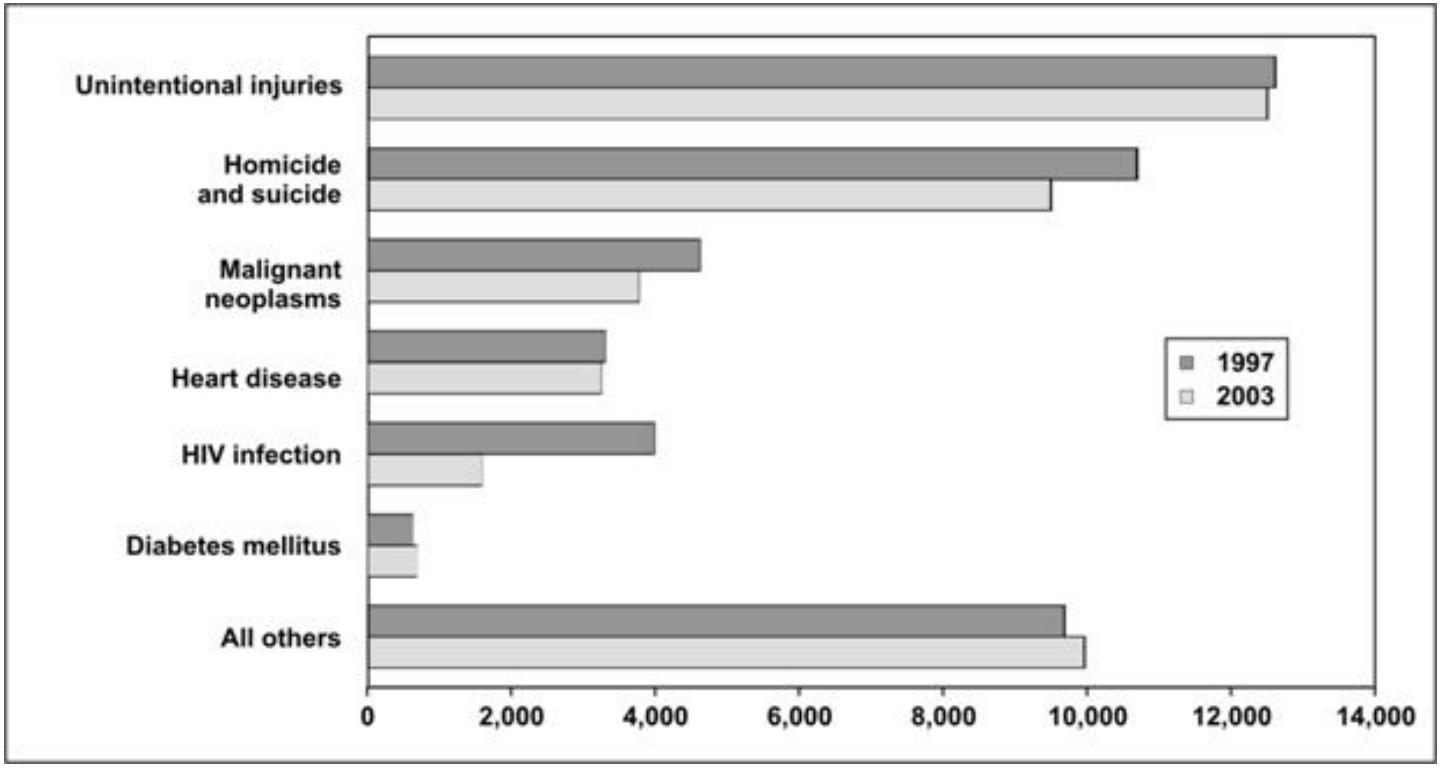


Figura 4: *Number of Deaths by Cause Among 25-34 Year Olds* - EUA.

Fonte: https://archive.cdc.gov/www_cdc_gov/csels/dsepd/ss1978/lesson4/section4.html

Gráfico em Linha para Variáveis Quantitativas

Esse tipo de gráfico pretende mostrar a evolução temporal de uma variável quantitativa (discreta ou contínua).

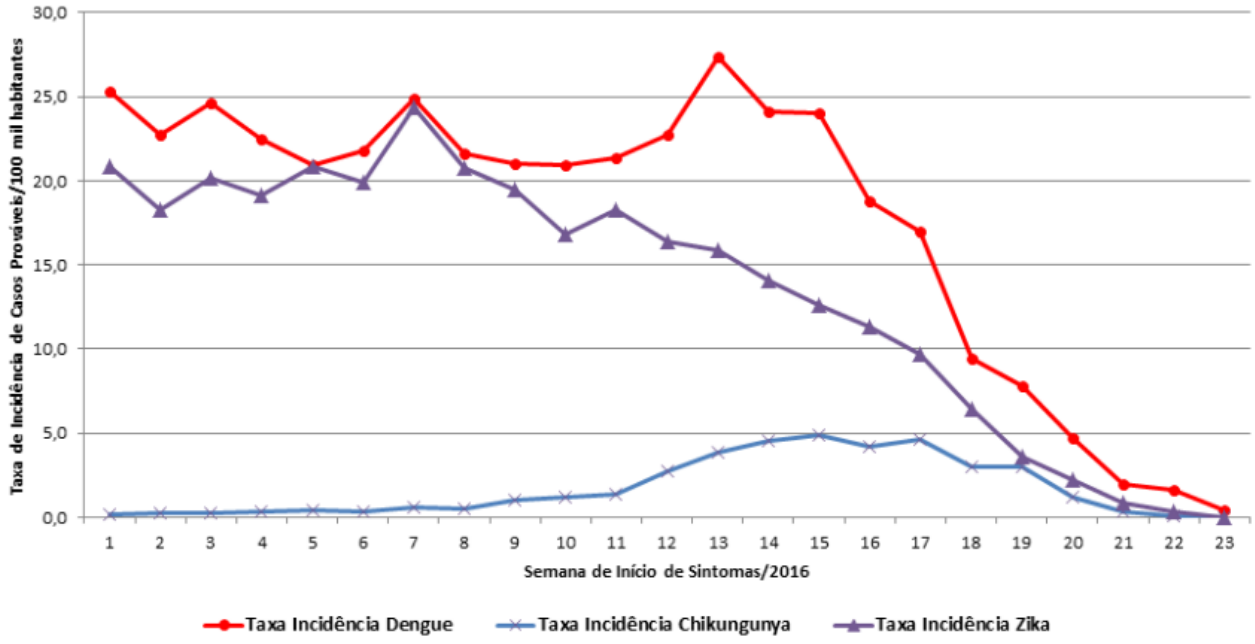


Figura 5: Taxas de incidências de casos prováveis de Dengue, Chikungunya e Zika, por semana epidemiológica de início de sintomas, Estado RJ, ano 2016. Fonte: Boletim Epidemiológico 005/2016, SES-RJ, 20/jun/2016.

Gráfico com Múltiplas Linhas

Comparações entre tratamentos diferentes.

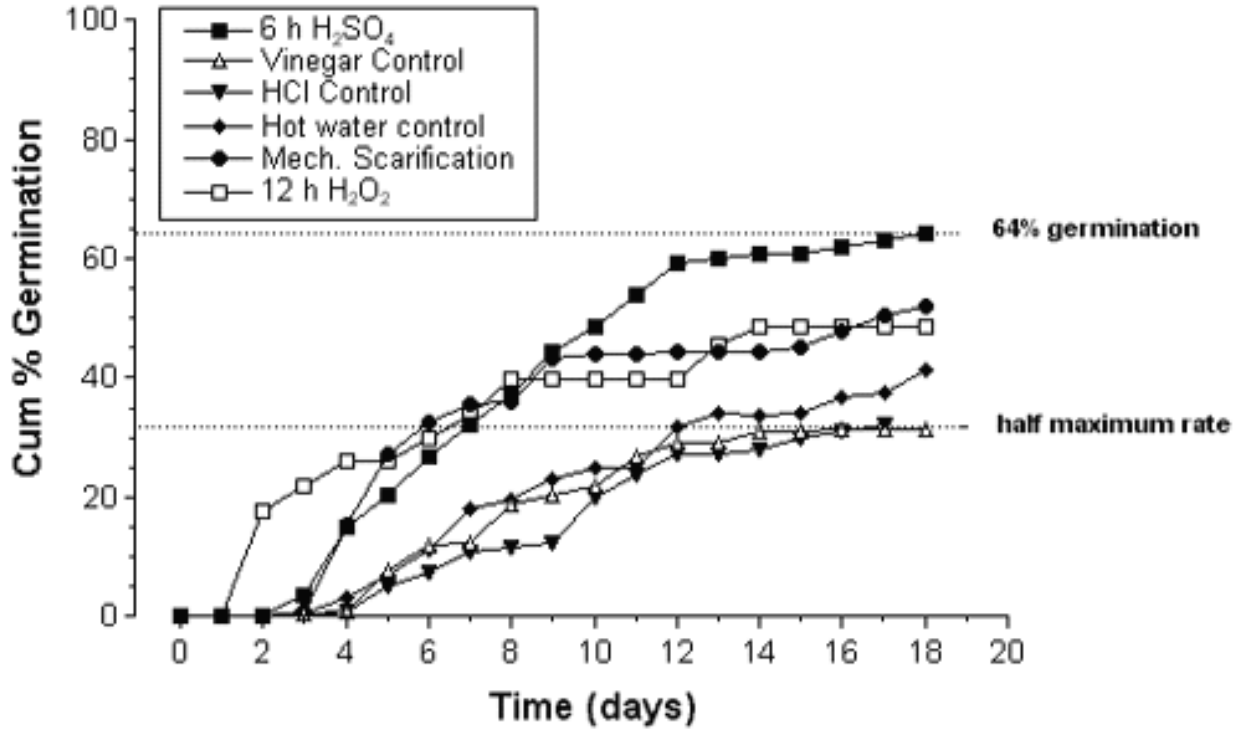


Figura 6: *Cumulative germination of gourd seeds following various pregermination treatment.*

Histograma

- Mimetizar um modelo de probabilidade contínuo.
- Se calculada em proporções, representa uma distribuição de probabilidade advinda de um modelo teórico subjacente (desconhecido).

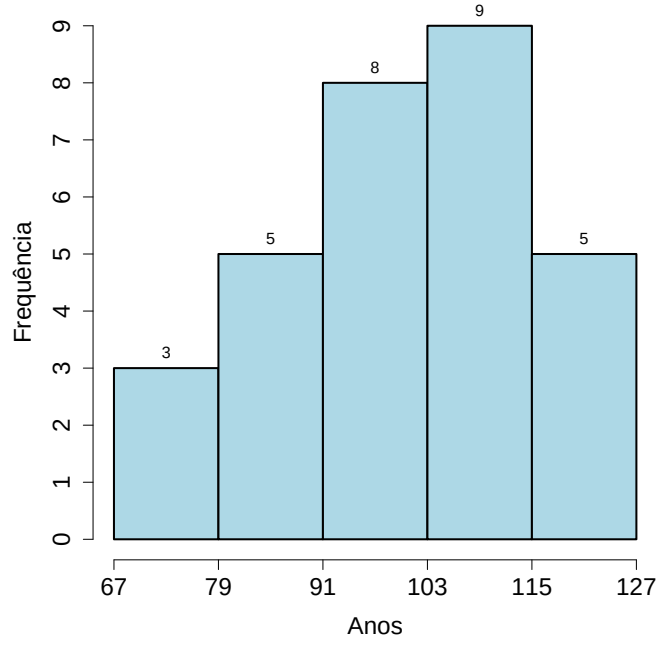


Figura 7: Idades (anos) de árvores em uma floresta.

Múltiplos Histogramas

Comparações entre as distribuições empíricas de duas populações.

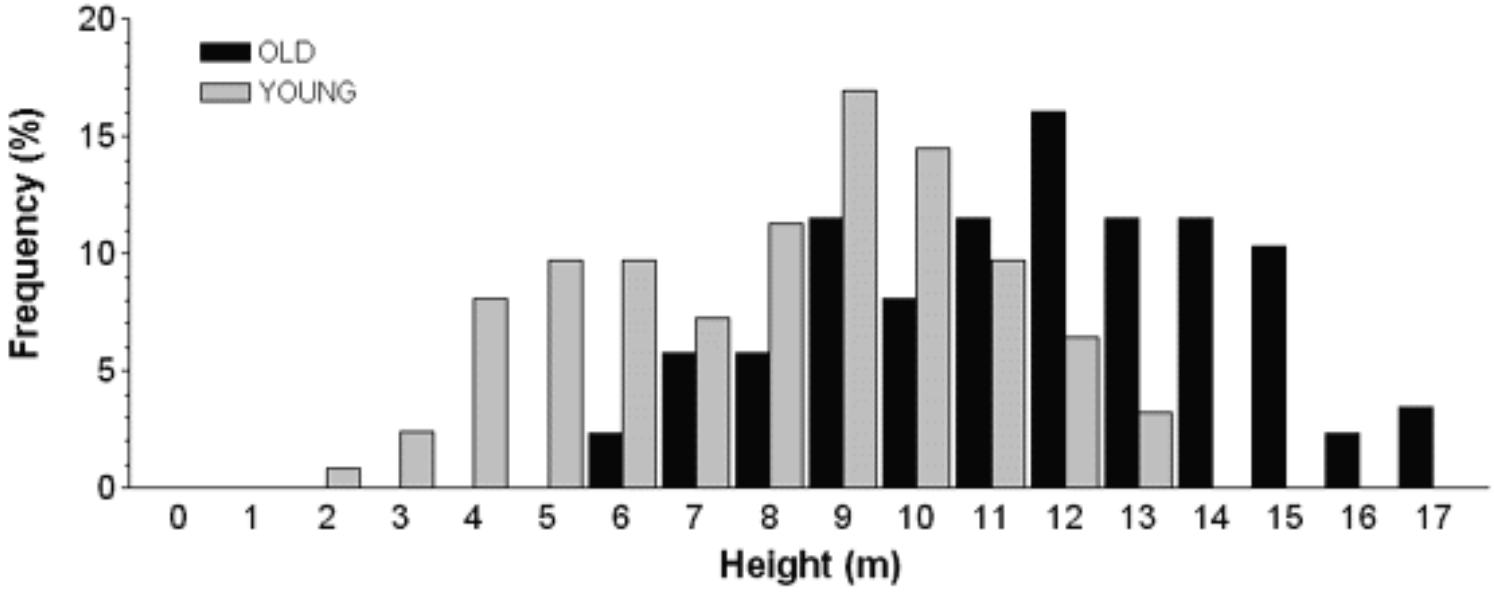


Figura 8: *Height distribution in two recruitment cohort of eastern white pine (Pinus strobus) near the eastern shore of Lake Auburn, Maine, in Januaray 2011.* $n = 88$ *OLD and* 123 *YOUNG trees.*

Diagrama de ramo e folhas

- Gráfico de barras sem perda de informação numérica.
- “As dezenas são como os ramos e as unidades são como as folhas”.
- Os ramos podem ser unidades e as folhas as casas decimais, etc.

```
 6 | 7
 7 | 18
 8 | 25677
 9 | 25799
10 | 0123345789
11 | 2468
12 | 245
```

Figura 9: Idades (anos) de árvores em uma floresta.

Boxplot

Amostra de recém nascidos abaixo do peso:
709 1021 1135 1330 1474 1588 1588 1701 1729 1790 1818 1885
1893 1899 1928 1928 1928 1936 1970 2055 2055 2082 2084 2084
2100 2125 2126 2187 2187 2211 2225 2240 2240 2282 2296 2296
2301 2325 2353 2353 2367 2381 2381 2381 2395 2410 2410 2414
2424 2438 2442 2450 2466 2466 2466 2495 2495 2495 2495

- Quartil 1 (frequência relativa 25%): Q_1 - valor que deixa 1/4 das observações à esquerda;
- Quartil 2, ou mediana (frequência relativa 50%): Q_2 - valor que deixa 2/4 das observações à esquerda;
- Quartil 3 (frequência relativa 75%): Q_3 - valor que deixa 3/4 das observações à esquerda; e
- Amplitude interquartílica: $Q_3 - Q_1$ - medida de dispersão.

Relação Histograma versus *Boxplot*

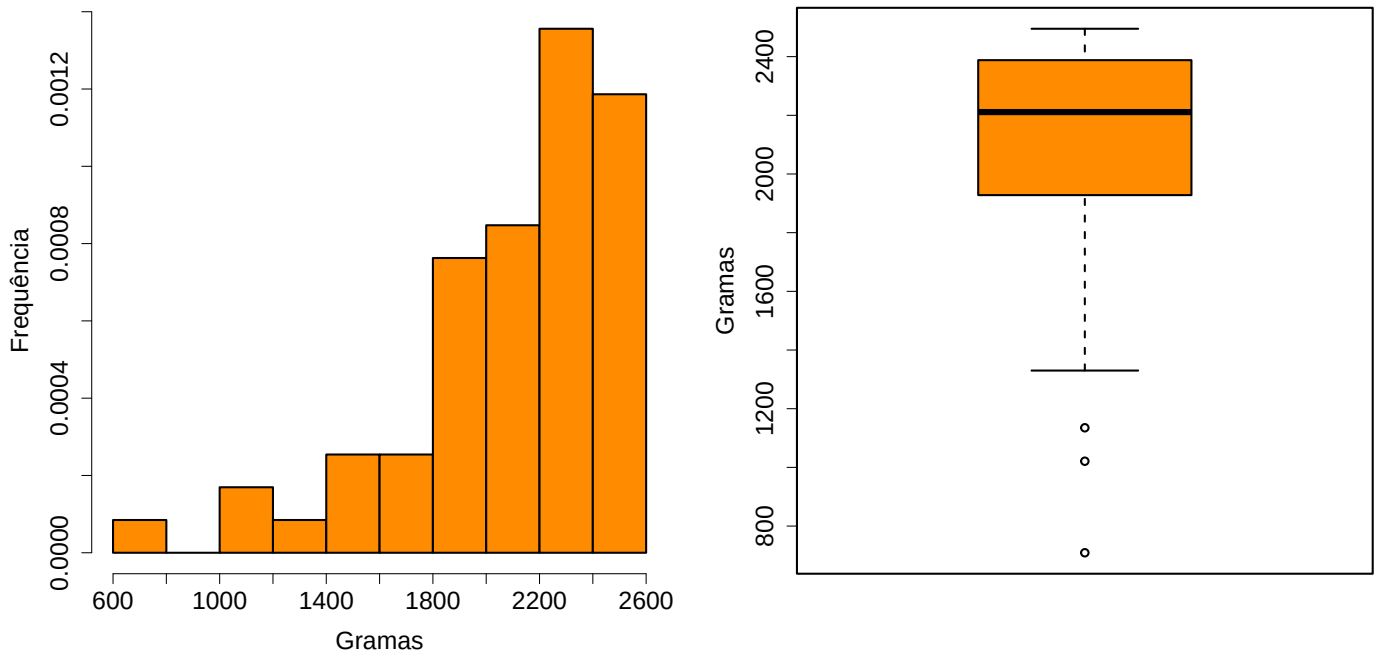


Figura 10: Recém nascidos abaixo do peso.

Construção do *Boxplot*

1. Encontre os 3 quartis.
2. Classifica-se como *outliers à direita*, pontos que se encontram a mais de $1,5\times$ (amplitude interquartílica) unidades à direita do 3º quartil.
3. Classifica-se como *outliers* à esquerda, pontos que se encontram a mais de $1,5\times$ (amplitude interquartílica) unidades à esquerda do 1º quartil.
4. A linha se estende até os valores mais extremos do conjunto de dados que não sejam *outliers*.

Múltiplos *Boxplots*

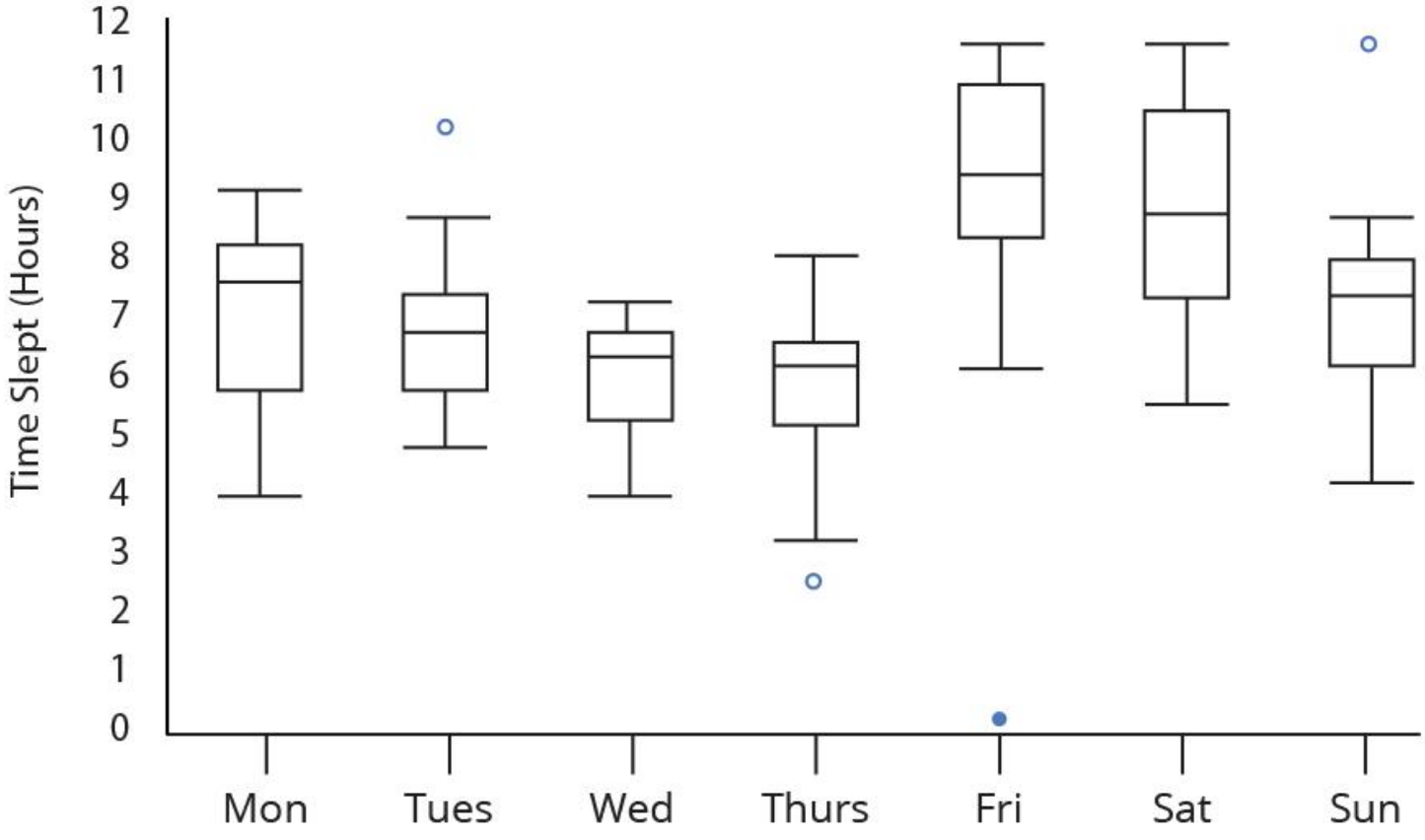


Figura 11: Número de horas dormidas por dia da semana.

Diagrama de Dispersão

- Pares ordenados no plano cartesiano.
- Eixo *x*: variável 1. Eixo *y*: variável 2.
- **Objetivo**: investigar relações entre as variáveis 1 e 2.

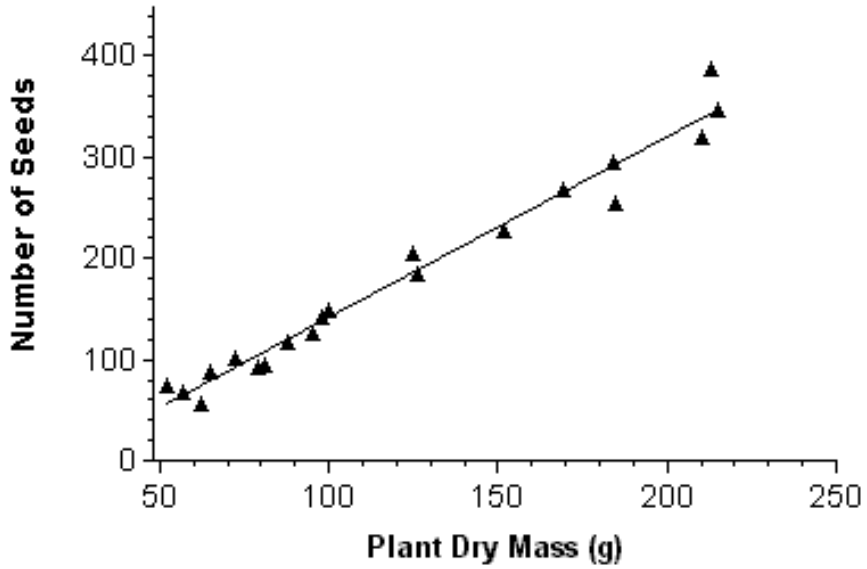


Figura 12: *Seed production as a function of plant biomass in waterlilies (Nuphar luteum) harvested from Great Works Pound in Nothern Maine in August, 2011.*

Correlações Espúrias

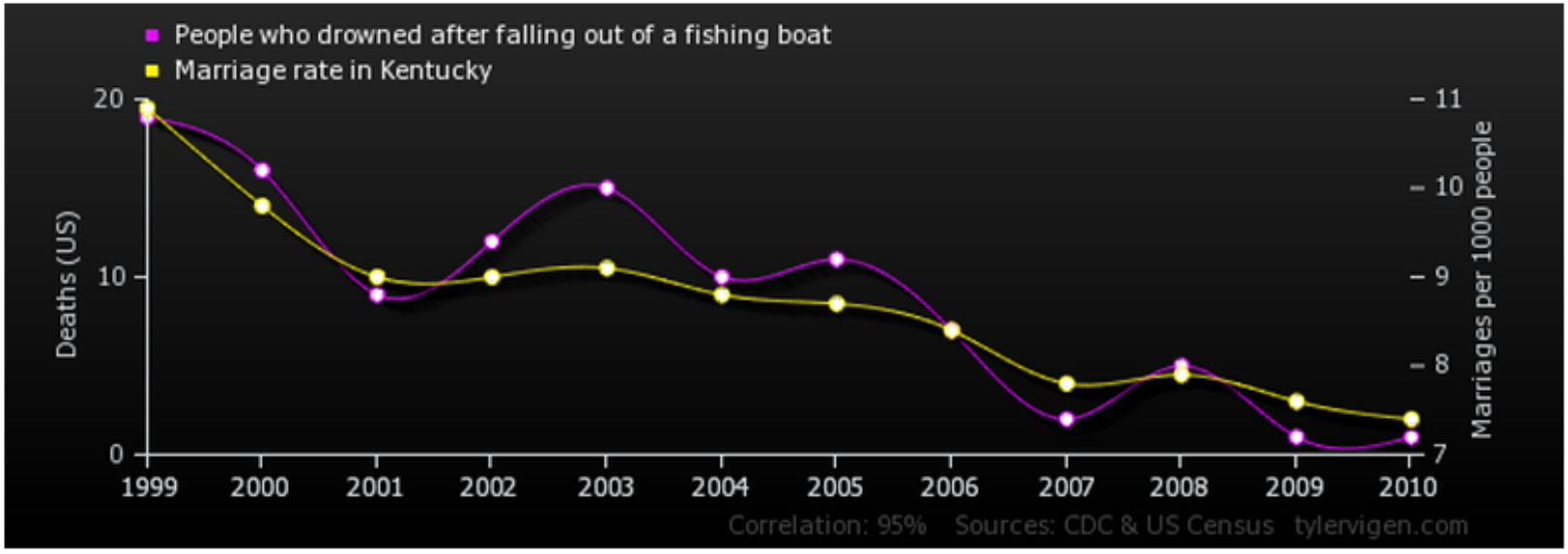


Figura 13: Número de afogamentos após cair de barco de pesca e taxa de casamentos em Kentucky, por ano.

Gráficos Diversos

- Comparações entre populações diferentes.
- O crescimento parece ser estatísticamente diferente para as duas espécies.

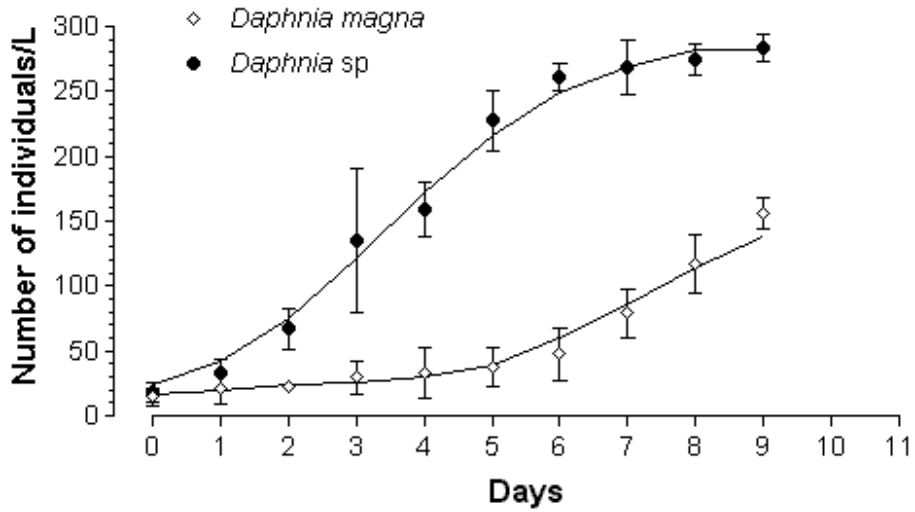


Figura 14: *Mean population density (±1 standard deviation) of two species of Daphnia following artificial eutrophication of a small farm pond by application of organic fertilizer. Six replicate 1L water samples were drwan from 50cm depth at 1100hr each day.*

Daphnia magna

Estatísticas Sumárias

- **Propósito**: resumir características relevantes sobre o comportamento de uma variável.
- Trataremos dos seguintes grupos de medidas:
  - Medidas de Posição (ou Tendência Central);
  - Medidas de Dispersão; e
  - Quantis.

Estatísticas Sumárias: Medidas de Posição

- Medidas de posição (ou de tendência central) objetivam informar em torno de qual ponto da reta real um conjunto de observações se distribui.
- São medidas de posição:
  - média aritmética (ou, simplesmente, média);
  - mediana; e
  - moda.

Medidas de Posição - Média Aritmética

- Sejam $x_1, x_2, \ldots, x_n$ os n valores observados para uma variável X. Por exemplo, os níveis de glicose de n indivíduos.
- A média aritmética de X é dada por

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Exemplo

Considere uma variável X que contém $n = 15$ valores.

10	9	8	12	11	7	10	8,5
9,5	6	14	13	11	12	9	

Daí, podemos calcular a média por

$$\bar{x} = \frac{10 + 9 + 8 + \cdots + 12 + 9}{15} = \frac{150}{15} = 10.$$

Exemplo

Suponha que o primeiro valor seja igual a 100.

100	9	8	12	11	7	10	8,5
9,5	6	14	13	11	12	9	

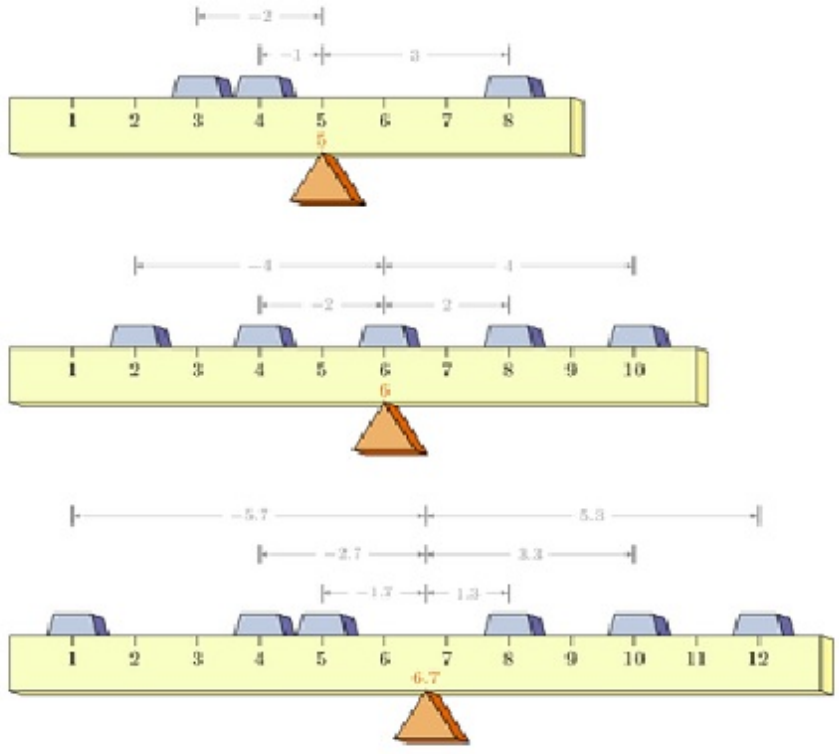
Daí, podemos calcular a nova média por

$$\bar{x} = \frac{100 + 9 + 8 + \cdots + 12 + 9}{15} = \frac{240}{15} = 16.$$

A média pode ser bastante influenciada por valores extremos.

Média Aritmética - Interpretação

A média aritmética pode ser intrepretada como o centro de massa do conjunto de dados:



Medidas de Posição - Mediana

Sejam $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ os n valores **ordenados** da variável X.

A mediana $\widetilde{x}$ é um ponto tal que o número de observações abaixo e acima de $\widetilde{x}$ é o mesmo.

- Se n é ímpar: $\widetilde{x}$ é o valor central no conjunto de dados ordenados:

$$\widetilde{x} = x_{\left(\frac{n+1}{2}\right)}$$

- Se n é par: tome $\widetilde{x}$ como a média entre os dois valores centrais no conjuno de dados ordenados:

$$\widetilde{x} = \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right].$$

Exemplo (continuação)

Para o cálculo da mediana, precisamos ordenar os dados.

6	7	8	8,5	9	9	9,5	10
10	11	11	12	12	13	14	

Note que $n = 15$ e $\dfrac{n+1}{2} = 8$.

Daí, podemos calcular a mediana por

$$\widetilde{x} = x_{(8)} = 10.$$

Ao contrário da média, a mediana não sofre influência de valores extremos. Seu valor permaneceria inalterado se substituíssemos a primeira observação por 10 vezes o valor da média observada.

Medidas de Posição - Moda

A moda é o valor da variável de interesse que é observado com frequência máxima.

Pode haver mais de um valor mais frequente: dados bimodais (duas modas) ou multimodais (três ou mais modas).

Um conjunto de observações sobre uma variável também pode ser amodal (sem moda).

Exemplo

Considere idades de pessoas que fazem parte de um grupo de pais que ensinam os seus filhos bem novos a nadar.

1	1	1	2	2	2	2	3	3
31	31	32	32	32	32	33	33	33

Daí, podemos calcular a **média** por

$$\bar{x} = \frac{1+1+1+\cdots+33+33}{18} = \frac{306}{18} = 17,$$

e a **mediana** por

$$\tilde{x} = \frac{x_{(9)} + x_{(10)}}{2} = \frac{3+31}{2} = 17.$$

Exemplo

A média e a mediana não refletem o valor típico das idades (grupos de pais e grupo de filhos).

Os passos para se obter a moda são:

1. Encontrar todos os valores distintos da variável;
2. Obter a frequência de cada valor distinto; e
3. Selecione o valor (ou valores) com maior freqência para obter a moda.

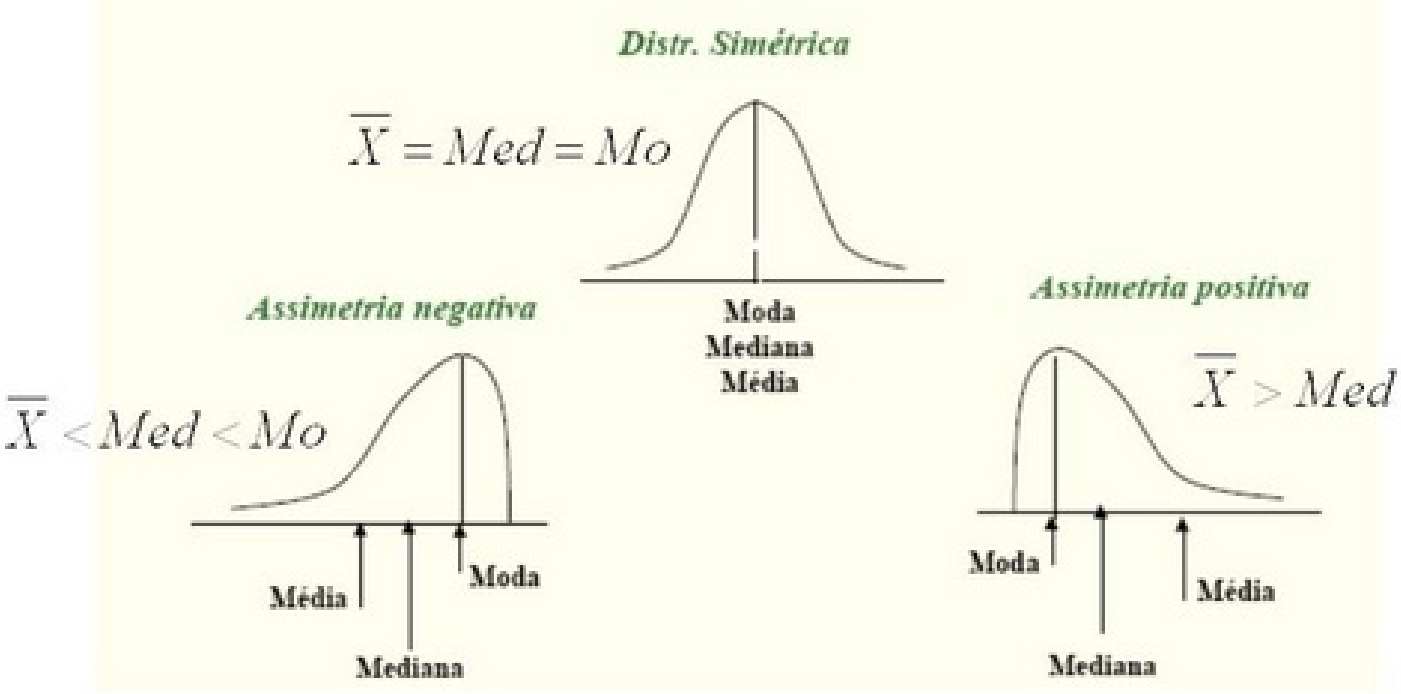
Valor	1	2	3	31	32	33
Frequência	3	4	2	2	4	3

Os dois valores com a maior frequência são 2 e 32.

Esta variável é **bimodal**.

Ordenação entre Medidas de Posição

Ao ordenar os valores calculados de média, mediana e moda, podemos ter ideia do formato da distribuição dos dados:



Observação: Para dados empíricos, dificilmente média, mediana e moda serão idênticas, mas se o formato da distribuição for simétrico, seus valores devem ser próximos.

Quantis

- Quantis são pontos que dividem o conjunto de dados em partes com igual número de observações. São quantis:
  - Quartis: 3 pontos que dividem o conjunto de dados ordenados em 4 partes com igual número de observações;
  - Decis: 9 pontos que dividem o conjunto de dados ordenados em 10 partes com igual número de observações; e
  - Percentis: 99 pontos que dividem o conjunto de dados ordenados em 100 partes com igual número de observações.
- Por exemplo, o primeiro quartil deixa 25% das observações abaixo de seu valor; o quinto percentil deixa 5% das observações abaixo de seu valor.

Estatísticas Sumárias: Medidas de Dispersão

- Medidas de dispersão objetivam informar quão disperso (alternativamente, quão concentrado) é um conjunto de dados.
- São medidas de dispersão:
  - amplitude amostral;
  - amplitude interquartílica;
  - variância amostral;
  - desvio-padrão amostral; e
  - coeficiente de variação amostral.

Medidas de Dispersão: Amplitude Amostral

- A medida mais básica possível de variabilidade em um conjunto de dados é a chamada amplitude amostral, dada pela diferença entre os valores máximo e mínimo observados.
- Algumas questões, ao usar a amplitude amostral:
  - toda a informação entre os extremos é desprezada: pontos entre os extremos são concentrados ou dispersos?
  - os pontos extremos (máximo e mínimo) destoam dos demais (são *outliers*)?

Exemplos (continuação)

Os dados ordenados provenientes do primeiro exemplo:

6	7	8	8,5	9	9	9,5	10
10	11	11	12	12	13	14	

Logo, a amplitude é dada por: $A_1 = 14 - 6 = 8$.

Os dados ordenados provenientes do segundo exemplo:

1	1	1	2	2	2	2	3	3
31	31	32	32	32	32	33	33	33

Logo, a amplitude é dada por: $A_2 = 33 - 1 = 32$.

Medidas de Dispersão: Amplitude Interquartílica

Lembando:

- Quartil 1 (frequência relativa 25%): Q_1 - valor que deixa 1/4 das observações à esquerda;
- Quartil 2, ou mediana (frequência relativa 50%): Q_2 - valor que deixa 2/4 das observações à esquerda; e
- Quartil 3 (frequência relativa 75%): Q_3 - valor que deixa 3/4 das observações à esquerda.
- Amplitude interquartílica: $AI = Q_3 - Q_1$ - medida de dispersão.
- AI avalia a dispersão na porção central dos dados (entre as 50% observações centrais).

Uso da AI para Avaliação de Pontos Discrepantes

- São consideradas discrepantes observações cujos valores são:
  - superiores a $Q_3 + 1,5 \times AI$;
  - inferiores a $Q_1 - 1,5 \times AI$.
- São consideradas extremamente discrepantes observações cujos valores são:
  - superiores a $Q_3 + 3 \times AI$;
  - inferiores a $Q_1 - 3 \times AI$.

Medidas de Dispersão: Variância Amostral

- A variância amostral é dada por:

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

- É uma medida da distância "quadrática média" das observações $x_1, x_2, \ldots, x_n$ de uma variável X, em relação à sua média.
- Quanto maior a distância quadrática média dos observados em relação à sua média, maior a dispersão na variável X.
- Problema: difícil avaliar se dispersão "grande" ou "pequena", pois a variância tem unidade igual ao quadrado da unidade de X. Se X é registrada em kg, por exemplo, a variância s^2 tem unidade kg^2.

Medidas de Dispersão: Desvio-padrão Amostral

- O desvio-padrão amostral é dado pela raiz quadrada positiva da variância amostral:

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

- Tem unidade igual à dos dados observados.
- Quanto maior a distância quadrática média dos observados em relação à sua média, maior a dispersão na variável X.
- Questão: $s = 10kg$ é desvio-padrão "grande" ou "pequeno"? Depende: se $s = 10kg$ se originou das medidas de peso de filhotes de elefantes, provavelmente se trata de elefantinhos com aspectos bastante similares. Já $10kg$ de desvio-padrão entre os pesos de crianças pode indicar pesos bastante heterogêneos.

Coeficiente de Variação

- O coeficiente de variação de uma variável X é dado por:

$$CV = \frac{s}{\bar{x}}$$

- As unidades de s e $\bar{x}$ são idênticas, se cancelam e, portanto, o CV é medida adimensional de dispersão.
- Avalia-se s relativizado pela magnitude média de X.
- Como é adimensional, pode ser usado para comparar dispersão de variáveis com magnitudes e até unidades distintas.

Exemplo

Considere idades de pessoas que fazem parte de um grupo de pais que ensinam os seus filhos bem novos a nadar.

1	1	1	2	2	2	2	3	3
31	31	32	32	32	32	33	33	33

Considere também os dados resumidos pela frequência:

Valor	1	2	3	31	32	33
Frequência	3	4	2	2	4	3

Observação: será apresentado na aula (no quadro) uma forma sistemática de fazer as contas utilizando um formato de tabela. Será apresentado também fórmulas alternativas dos cálculos da média e da variância utilizando frequências absolutas e relativas.

Amostragem

Tópicos

1. População e amostra
2. Introdução às técnicas de amostragem
   - Amostra aleatória simples
   - Amostra aleatória estratificada
   - Amostra aleatória por conglomerados
   - Amostragem sistemática
   - Amostragem por conveniência
3. Estudos transversais e longitudinais
   - Ensaio controlado aleatorizado
   - Série de casos
   - Estudos de coortes
   - Estudos de caso e controle
4. Meta análises

Conceitos Fundamentais da Estatística

Problema geral da Estatística: queremos tirar conclusões sobre toda uma população, mas não conseguimos investigar todos os seus indivíduos.

- Selecionamos uma amostra de indivíduos que represente razoavelmente bem a população.
- Com os resultados amostra, inferimos resultados para a população.

Definições

- **População**: conjunto de todos os elementos de interesse em um estudo (pessoas, objetos, cidades, etc.).
- **Amostra**: subconjunto escolhido dentre os elementos da população.
- **Parâmetro**: quantidade numérica que representa uma característica da população.
- **Estimador**: variável obtida em função da amostra para estimar um parâmetro da população.
- **Inferência**: área que trata de tirar conlusões sobre a população tendo como base uma amostra.

Exemplo: População e Amostra

- Por exemplo, em um estudo sobre as características de pacientes com urolitíase, a **população** seria o conjunto dos pacientes com diagnóstico de urolitíase em um determinado território.
- Uma **amostra** poderia ser os pacientes acompanhados no ambulatório de nefropediatria do Hospital Federal dos Servidores do Estado do Rio de Janeiro (HFSE), no período de janeiro de 2012 a dezembro de 2014, e com o diagnóstico de urolitíase.

Tipos de amostragem probabilística

Há diversos métodos de amostragem probabilísticas, dentre as quais abordaremos:

1. Amostragem aleatória simples;
2. Amostragem aleatória estratificada;
3. Amostragem por conglomerados; e
4. Amostragem sistemática.

O conteúdo desta aula se baseia fortemente no livro texto:

Freire, S. M. (2012), *Bioestatística Básica*.

Amostra Aleatória Simples

Na amostra aleatória simples, cada observação tem a mesma probabilidade de ser extraída da população:

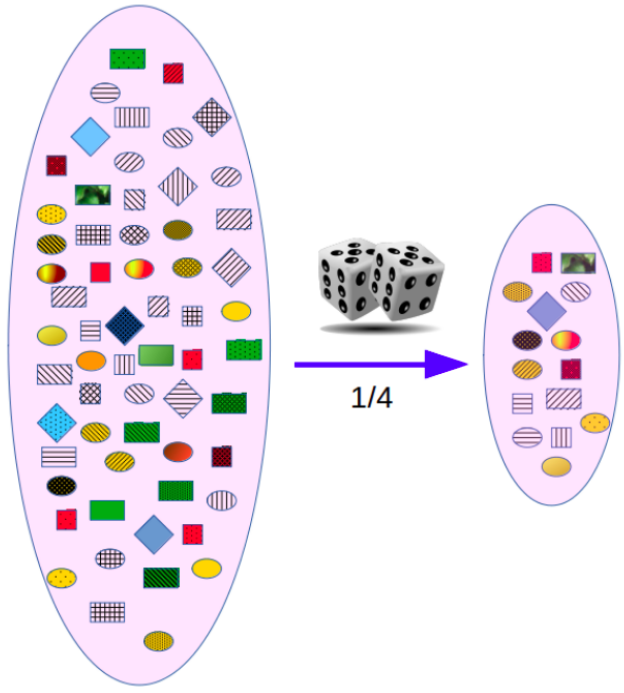


Figura 15: Amostra aleatória simples onde são extraídos 1/4 das unidades populacionais para compor a amostra.

Exemplo: Amostra Aleatória Simples

O estudo “O perfil socioeconômico e a percepção ambiental dos pescadores da Lagoa do Apodi, Rio Grande do Norte, Brasil” (Pinto, Nobre e Mariano, 2020) realiza uma amostragem aleatória simples de indivíduos que fazem uso da lagoa do Apodi para pesca.

Na colônia de pescadores de Apodi, existem 405 pescadores cadastrados na Confederação Nacional de Pescadores e Aquicultores(CNPA), dos quais 240 atuam diretamente na lagoa. Desses 240 pescadores, 52 foram selecionados aleatoriamente para participarem do estudo.

Amostra Aleatória Estratificada

- A população (heterogênea) é dividida em estratos ou subgrupos homogêneos.
- A amostra extraída tem, em cada estrato, uma proporção de indivíduos igual ou semelhante à correspondente proporção na população.

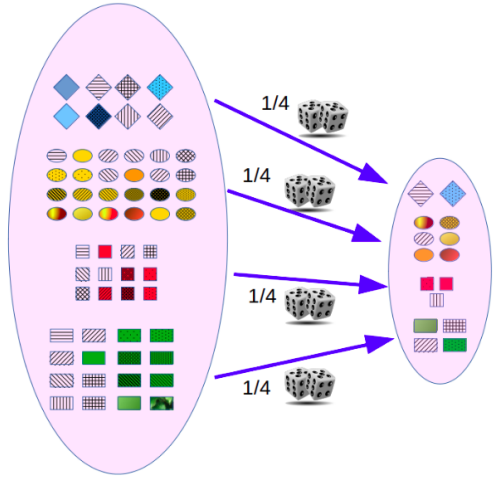


Figura 16: Amostra aleatória estratificada onde são extraídos 1/4 das unidades populacionais em cada estrato para compor a amostra

Exemplo: Amostra Aleatória Estratificada

Estudo multicêntrico “Hipertensão Arterial e Diabetes Mellitus entre trabalhadores da saúde: associação com hábitos de vida e estressores ocupacionais” (Novaes, Araújo e Sousa, 2020).

Processo de amostragem: a partir de listas nominais de todos os trabalhadores em atividade nos serviços de saúde da atenção básica e da média complexidade, fornecidas pelas Secretarias de Saúde dos municípios estudados, estratificou-se a amostra em três níveis: área geográfica, nível de assistência (atenção básica e média complexidade) e grupo ocupacional.

A composição da amostra foi definida com base na participação percentual de cada grupo por nível de estratificação estabelecido, seguindo-se o sorteio dos trabalhadores para compor a amostra (o sorteio foi feito com base em listagem de números aleatórios).

Amostra Aleatória por Conglomerados

- A seleção aleatória é realizada em grupos previamente existentes, como, por exemplo, escolas, unidades de saúde, bairros, etc.
- A população é dividida em grupos (conglomerados) e uma amostra aleatória simples dos grupos é selecionada.
- Em seguida, os elementos em cada conglomerado selecionado são amostrados.
- Se todos os elementos de cada conglomerado for selecionado, então esta amostragem é chamada de amostragem por conglomerados de um estágio. Se uma amostra aleatória é extraída de cada conglomerado selecionado, então esta amostragem é chamada de amostragem por conglomerados de dois estágios.

Amostra Aleatória por Conglomerados

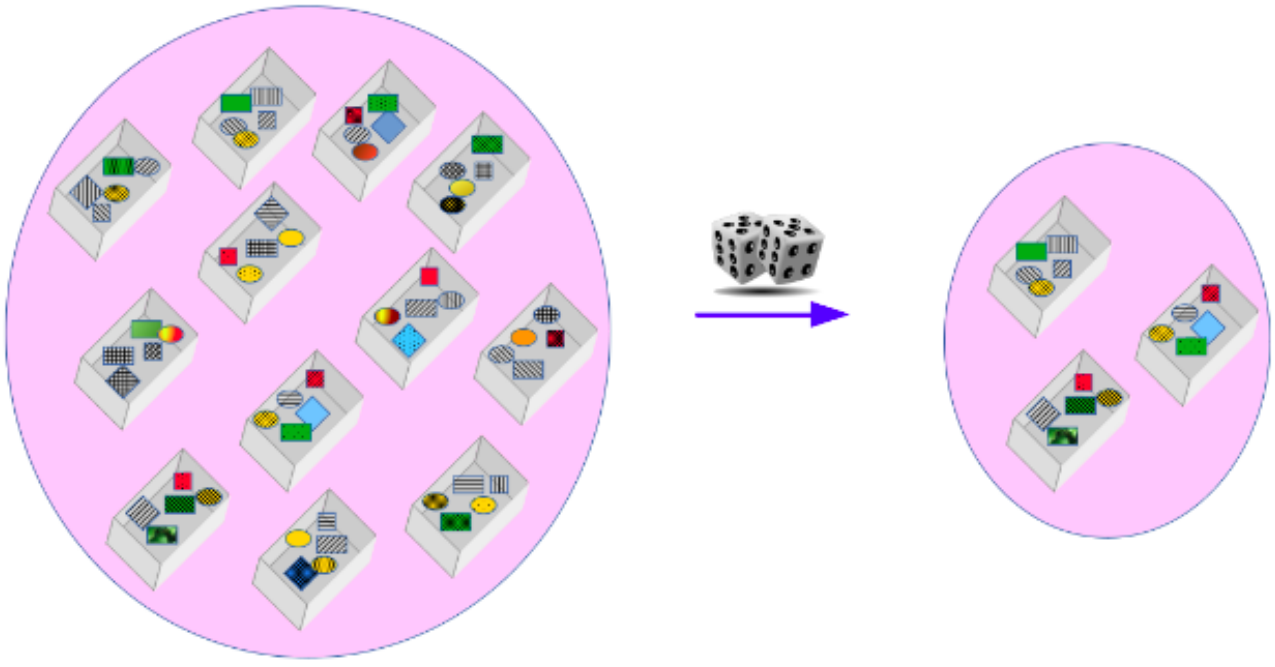


Figura 17: Amostra aleatória por conglomerados de 1 estágio. Os grupos são selecionados aleatoriamente e todos os elementos de cada grupo são selecionados.

Exemplo: Amostra Aleatória por Conglomerados

No estudo "Associação da depressão com as características sociodemográficas, qualidade do sono e hábitos de vida em idosos do Nordeste brasileiro: estudo seccional de base populacional" (Lopes et al., 2015), a população-alvo constou de aproximadamente 40 mil idosos, no ano de 2010, residentes na zona urbana da cidade de Campina Grande, PB.

O tamanho da amostra de idosos foi estimada em 205 idosos. Este estudo realizou uma amostragem por conglomerados, na qual o primeiro estágio consistiu na seleção aleatória de unidades básicas de saúde (UBS) em quatro distritos sanitários urbanos.

No segundo estágio, foi realizada uma amostragem sistemática (abordada a seguir) das residências nas ruas das unidades básicas de saúde selecionadas.

Amostra Sistemática

- A amostragem é feita de forma sistemática.
- Por exemplo, criamos uma lista ordenada dos objetos e sorteamos um número de 1 a 4 para indicar o primeiro item a ser selecionado. A partir deste item, selecionamos os demais pulando de 4 em 4 elementos na lista.

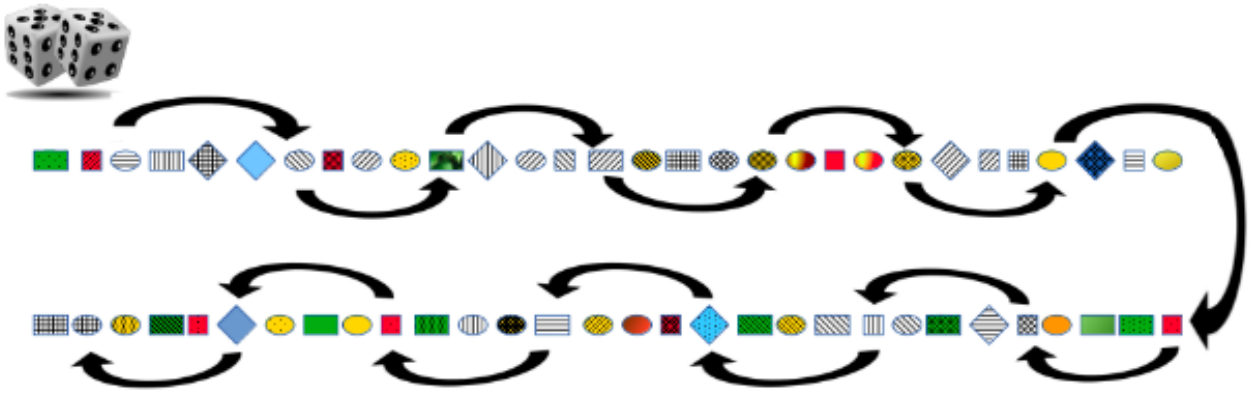


Figura 18: Amostra aleatória sistemática.

Exemplo: Amostra Sistemática

A amostragem sistemática deve ser realizada com cuidado, pois pode gerar uma amostra não representativa da população (ver figura).

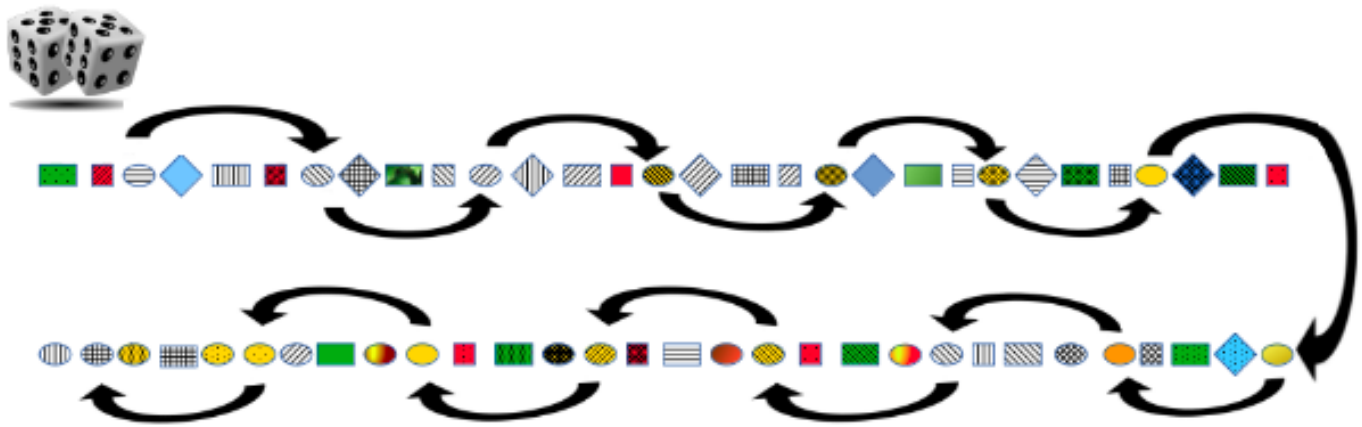


Figura 19: Uma amostra aleatória sistemática não representativa da população.

Amostragem por Conveniência

- Consiste em formar uma amostra da população a partir de itens que estejam mais facilmente disponíveis.
- Muito comum na pesquisa clínico-epidemiológica.
- Em estudos na área clínica, frequentemente as amostras são obtidas simplesmente identificando um número de pacientes que atendem aos critérios para inclusão em um estudo.
- A amostragem por conveniência é um tipo de amostragem não-probabilística.

Amostragem por Conveniência

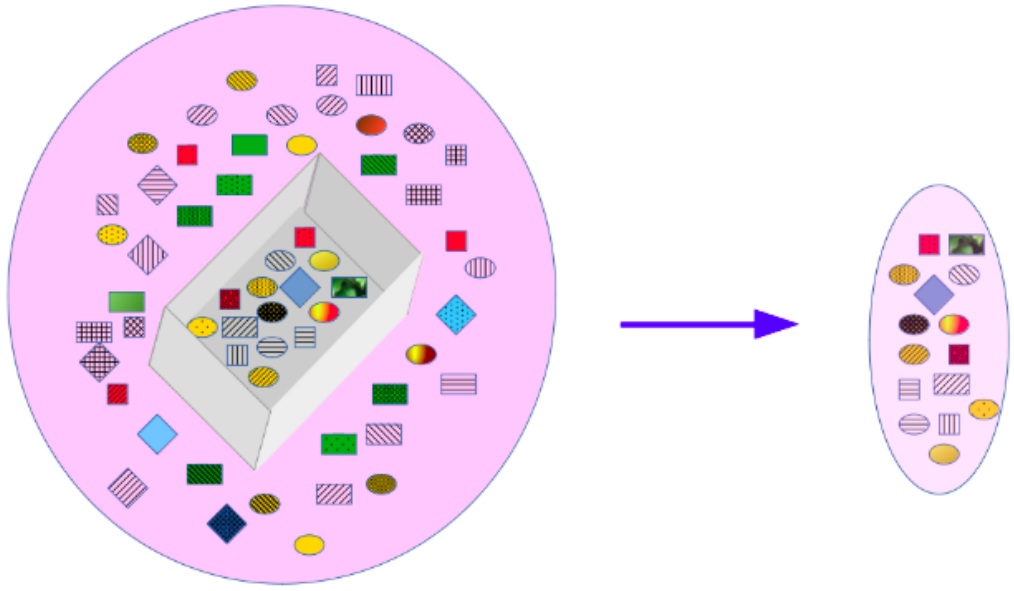


Figura 20: Amostra aleatória por conveniência.

Exemplo: Amostragem por Conveniência

No estudo "Perfil clínico-epidemiológico de 106 pacientes pediátricos portadores de urolitíase no Rio de Janeiro" (Barata e Valete, 2018), os pacientes que constituíram a amostra foram selecionados por meio de consulta ao setor de Estatística e Arquivo do Hospital Federal dos Servidores do Estado do Rio de Janeiro (HFSE).

Identificou-se todos os pacientes acompanhados no ambulatório de nefropediatria do hospital, no período de janeiro de 2012 a dezembro de 2014, e selecionados os prontuários com o diagnóstico de urolitíase.

Exemplo: Amostragem por Conveniência

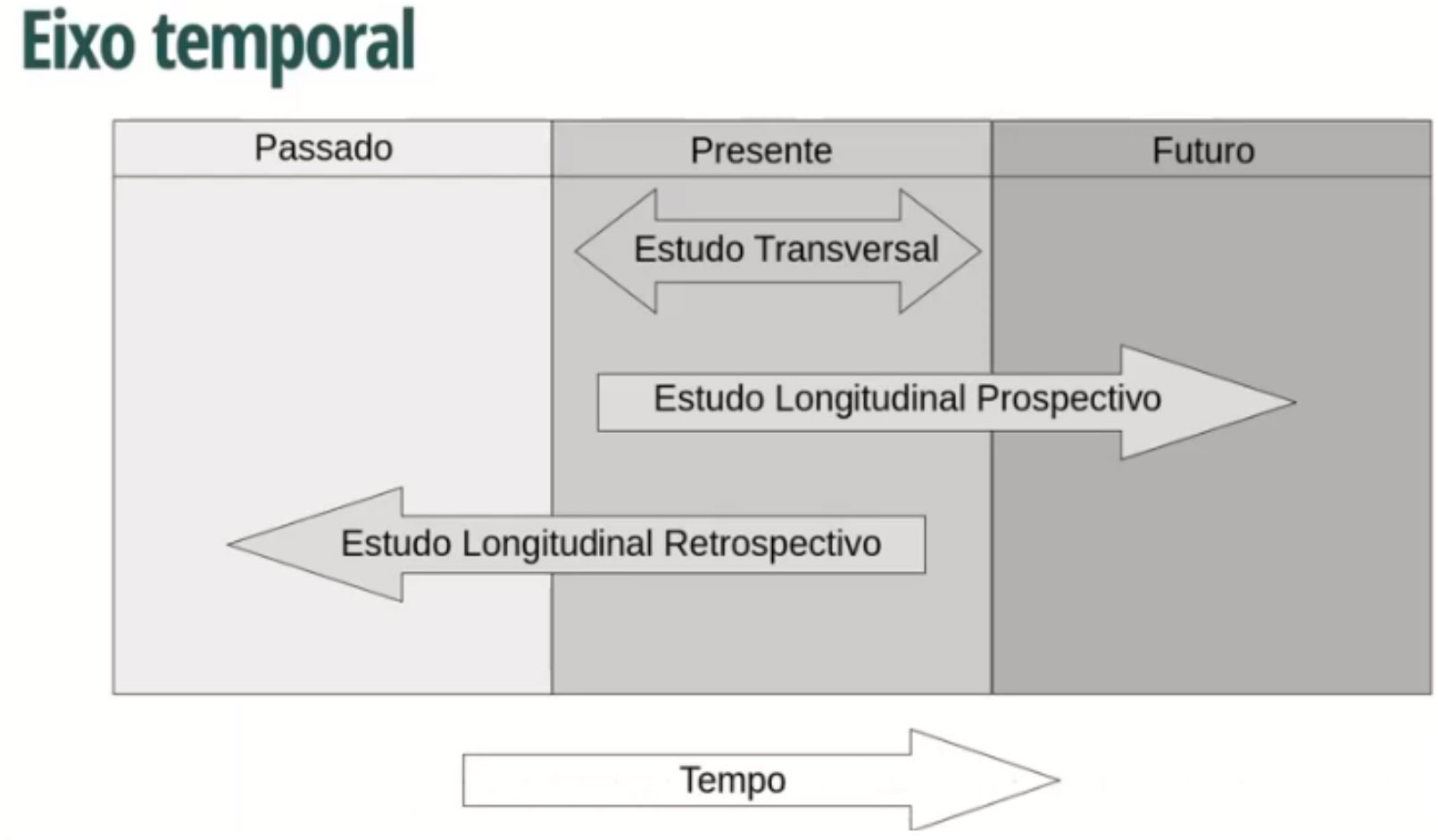
Critérios de inclusão na amostra:

1. Idade entre 1 mês e 18 anos;
2. Confirmação do diagnóstico clínico por pelo menos um exame radiológico, podendo ser a radiografia simples de abdome, a ultrassonografia abdominal ou de aparelho urinário (que pode identificar cálculos $\leqslant 5$ mm) e a tomografia computadorizada helicoidal de abdome sem contraste; e
3. Ser assistido no ambulatório de nefropediatria do HFSE no período citado anteriormente.

Delineamento de Estudos Clínicos e Epidemiológicos

- **Estudo Longitudinal**: as variáveis são mensuradas em mais de um instante do tempo, ou seja, quando as unidades de observação (frequentemente pacientes) são acompanhados ao longo do tempo e diversas variáveis são coletadas para verificar o estado clínico em cada instante.
- **Estudo Transversal**: as variáveis são mensuradas em um único instante do tempo para toda a amostra.
- **Estudo Prospectivo**: variáveis serão coletadas a partir do início do estudo em diante.
- **Estudo Retrospectivo**: as variáveis analisadas já foram coletadas a priori, utilizando-se de diversas fontes de dados como, por exemplo, prontuários, bancos de dados, etc. em instantes anteriores à realização do estudo.

Delineamento de Estudos Clínicos e Epidemiológicos



Ensaio Controlado Aleatorizado

- Comparação entre dois grupos ou populações (tratamento e controle).
- Grupo controle: geralmente recebe um placebo, ou um tratamento padrão já consolidado na literatura.
- Grupo de tratamento: tratamento proposto pelo pesquisador, a ser comparado com o grupo controle.
- A alocação das unidades amostrais nos grupos é feita de forma aleatória.

Exemplo: Ensaio Controlado Aleatorizado

"Repouso de três horas não aumenta complicações após cateterismo cardíaco diagnóstico com introdutor arterial 6F: ensaio clínico randomizado" (Rocha et al., 2009).

Objetivo: avaliar a segurança na redução do tempo de repouso no leito, de seis para três horas, após cateterismo cardíaco diagnóstico com introdutor arterial 6F.

Duas intervenções (tratamentos):

1. Deambulação após três horas após a remoção da bainha; e
2. Deambulação após seis horas após a remoção da bainha.

Os desfechos clínicos são possíveis complicações vasculares:

1. Hematoma no local da punção arterial;
2. Sangramento;
3. Correção cirúrgica da complicação vascular; e
4. Reação vasovagal.

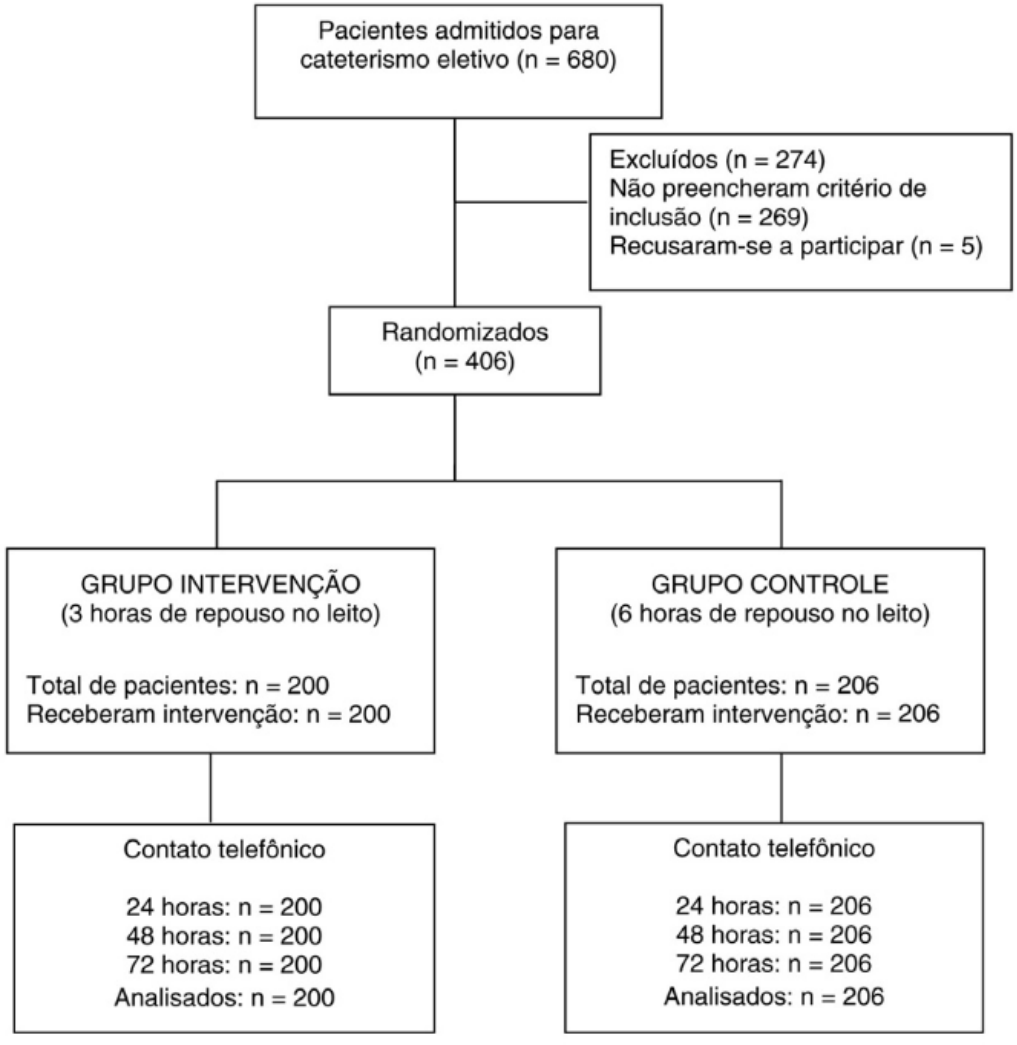


Figura 21: Definição dos grupos e aleatorização.

Características dos Grupos de Tratamento e Controle

- Os dois grupos (tratamento e controle) devem apresentar características similares para não viesar o estudo.
- As variáveis contínuas foram expressas como média $\pm$ desvio padrão e analisadas pelo teste t-Student. As variáveis categóricas foram expressas em frequências e percentuais e comparadas com o teste qui-quadrado de Pearson ou exato de Fisher. $p < 0,05$ foi considerado significante.

Tabela 6: Características demográficas e clínicas.

Características	Todos (n = 406)	Grupo intervenção (n = 200)	Grupo controle (n = 206)	P
Idade, anos	64 ± 9,4	63,5 ± 9,5	64,5 ± 9,3	0,32
Sexo feminino, n (%)	192 (47,3)	101 (50,5)	91 (44,2)	0,20
Peso, kg	75 ± 13,7	75,2 ± 13,8	74,8 ± 13,6	0,76
Altura, cm	1,67 ± 0,08	1,67 ± 0,09	1,67 ± 0,08	0,39
Índice de massa corporal, kg/m²	27 ± 4,3	27,1 ± 4,4	27 ± 4,1	0,92
Diabetes, n (%)	69 (17)	35 (17,5)	34 (16,5)	0,78
Insulina, n (%)	7 (1,72)	4 (2)	3 (1,5)	0,27
Hipertensão arterial sistêmica, n (%)	245 (60,3)	114 (57)	131 (63,6)	0,17
Doença periférica vascular, n (%)	35 (8,6)	20 (10)	15 (7,3)	0,32
Medicações em uso, n (%)				
Ácido acetilsalicílico	226 (55,7)	108 (54)	118 (57,3)	0,50
Clopidogrel	49 (12,1)	24 (12)	25 (12,1)	0,96
Ticlopidina	5 (1,2)	2 (1)	3 (1,5)	0,31
PAS, mmHg	127,6 ± 15,2	126,5 ± 15,6	128,7 ± 14,8	0,14
PAD, mmHg	78,8 ± 8,7	77,9 ± 8,6	79,8 ± 8,6	0,06

n = número de pacientes; PAD = pressão arterial diastólica; PAS = pressão arterial sistólica.

Tabela 7: Complicações pós-cateterismo cardíaco.

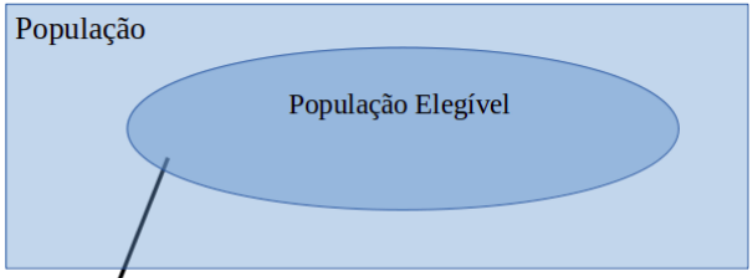
Complicações	Todos (n = 406)	Grupo intervenção (n = 200)	Grupo controle (n = 206)	P
Hematoma, n (%)	7 (1,7)	3 (1,5)	4 (1,9)	0,28
Sangramento, n (%)	5 (1,2)	1 (0,5)	4 (1,9)	0,16
Reação vasovagal, n (%)	11 (2,7)	4 (2)	7 (3,4)	0,38

n = número de pacientes.

Conclusão: pacientes submetidos a cateterismo cardíaco diagnóstico, utilizando introdutores arteriais 6F, podem deambular com segurança após três horas de repouso no leito. A redução do tempo de repouso traz maior conforto ao paciente, diminui o tempo de internação e contribui para aperfeiçoar os recursos existentes em relação à demanda crescente de exames.

Série de Casos

- Não há um grupo controle.
- Um conjunto de indivíduos é submetido a um tratamento e acompanhado por um período de tempo para se avaliar os efeitos do tratamento.
- Pela ausência do grupo controle, as evidências de uma série de casos são fracas, já que não é possível comparar os efeitos do tratamento com um outro tratamento alternativo ou placebo.

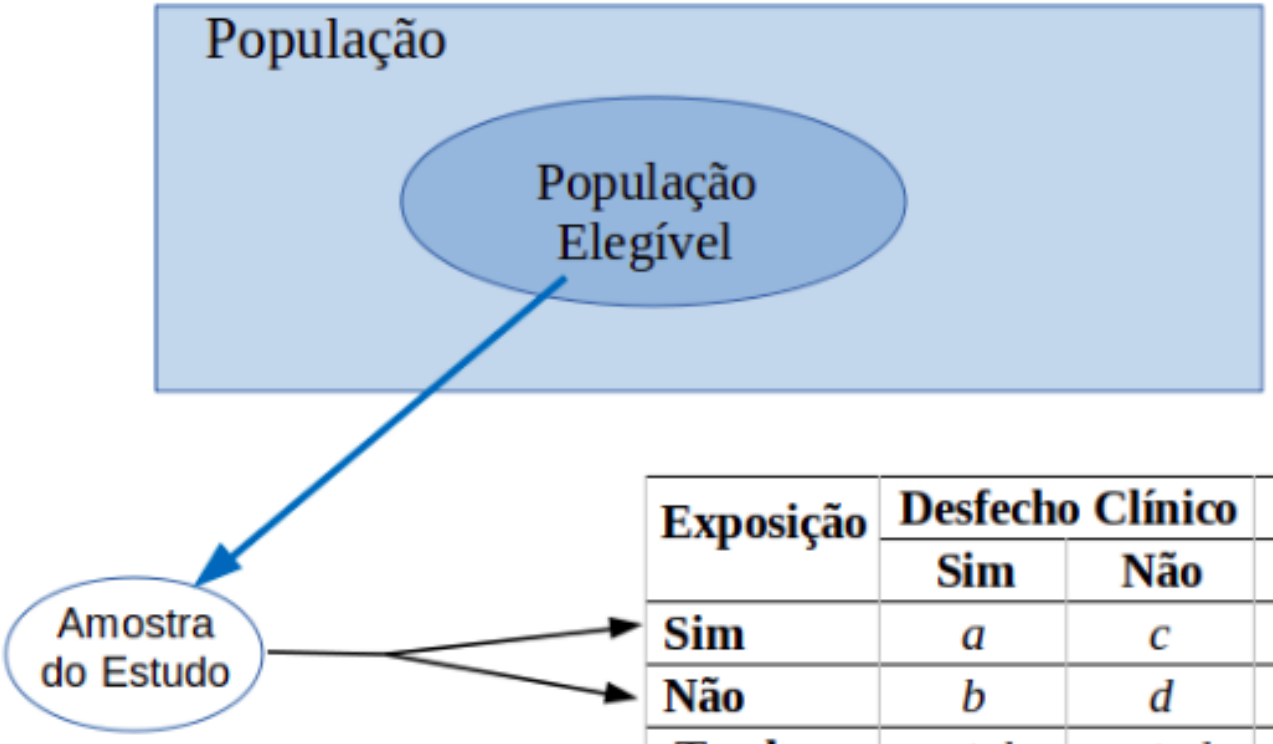


Exposição	Desfecho Clínico		
	Sim	Não	Total
Sim	*a*	*c*	*a* + *c*

Estudos de Coortes

- Quando se deseja estudar a associação entre possíveis fatores etiológicos (álcool, fumo, radiação, etc.) e desfechos clínicos (cirrose, câncer de pulmão, leucemia, etc.), não é ético expor deliberadamente pessoas a esses agentes.
- Nesses casos, deve-se recorrer a estudos observacionais, onde os fatores são estudados se comparando populações que são expostas ao fator (por opção, tipo de ocupação, acidente, etc.) com populações que não foram expostas.
- A alocação de unidades no grupo controle ou tratamento não é feita pelos pesquisadores.
- Amostra-se as unidades com respeito ao fator de exposição e, então, observa-se a ocorrência (ou não) do desfecho.

Estudos de Coortes



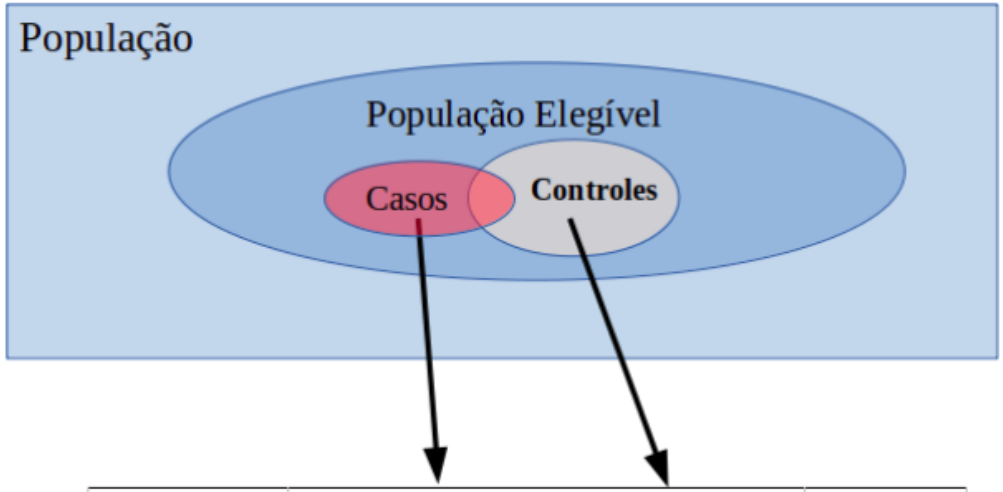
Exposição	Desfecho Clínico		
	Sim	Não	Total
Sim	*a*	*c*	*a* + *c*
Não	*b*	*d*	*b* + *d*
Total	*a* + *b*	*c* + *d*	

Estudos de Coortes

Exposição	Desfecho Clínico		
	Sim	Não	Total
Sim	*a*	*c*	*a* + *c*
Não	*b*	*d*	*b* + *d*
Total	*a* + *b*	*c* + *d*	

Estudo de Caso e Controle

- Amostra-se as unidades **com respeito ao desfecho** e depois observa-se se o indivíduo está exposto (ou não) ao fator de exposição.



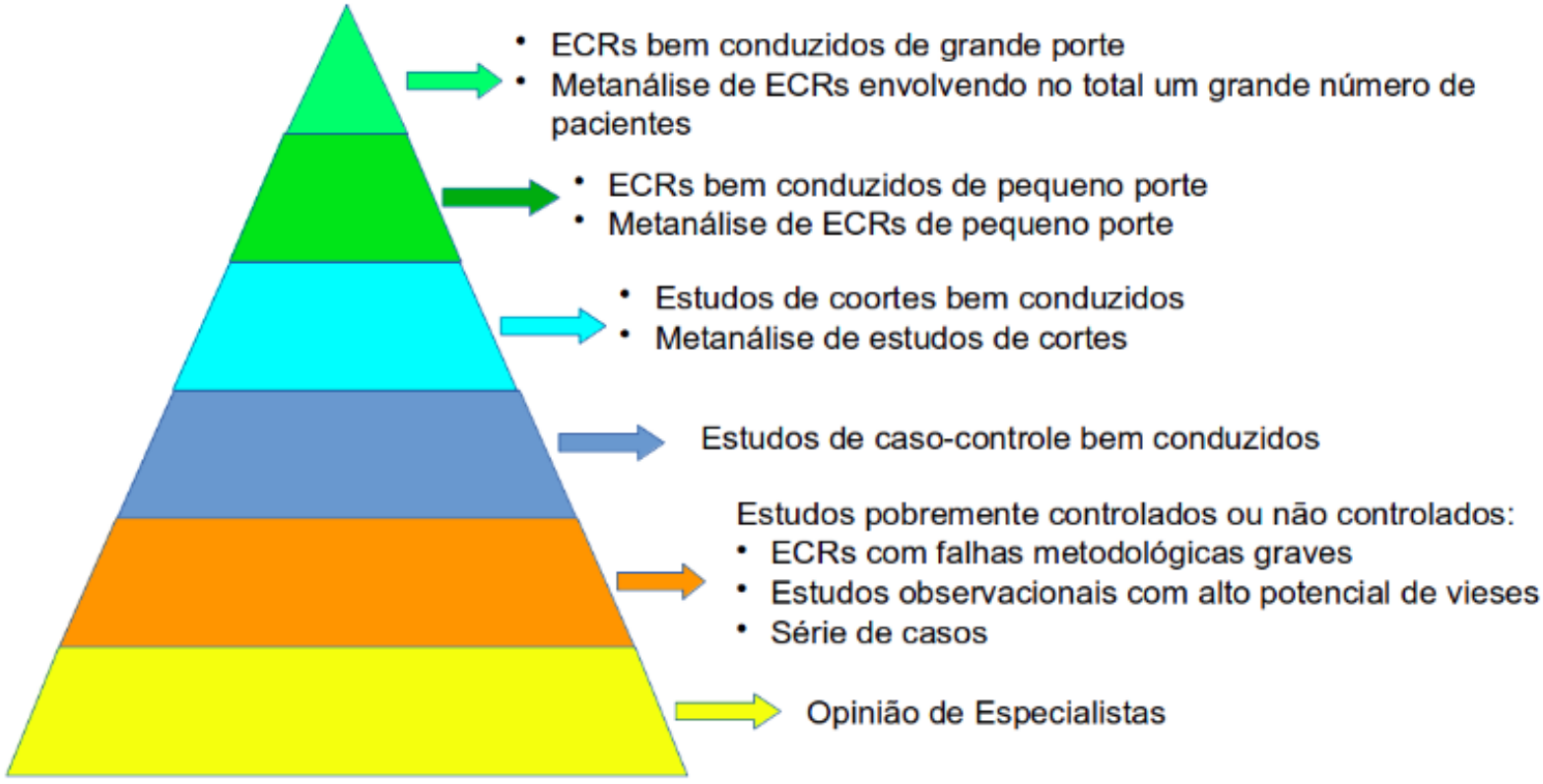
Exposição	Desfecho Clínico		
	Casos (com desfecho)	Controles (c/s desfecho)	Total
Sim	*a*	*c*	*a* + *c*
Não	*b*	*d*	*b* + *d*
Total	*a* + *b*	*c* + *d*	

Exemplo: Estudo de Caso e Controle

- "Estudo caso-controle sobre exposição precoce ao leite de vaca e ocorrência de Diabetes Mellitus tipo 1 em Campina Grande, Paraíba" (Medeiros et al., 2003).
- 64 bebês com diabetes mellitus amostrados, dos quais 54 foram expostos ao leite de vaca nos 4 meses anteriores.
- 64 bebês sem diabetes mellitus amostrados, dos quais 41 foram expostos ao leite de vaca nos 4 meses anteriores.

	Diabetes		
Exposição	Sim	Não	Total
Aleitamento materno exclusivo $\geq$ 4 meses	*10*	*23*	***33***
Exposição precoce ao leite de vaca $<$ 4 meses	*54*	*41*	***95***
Total	***64***	***64***	

Meta Análises

- Em geral, um único estudo não fornece a evidência definitiva sobre um determinado tema.
- Frequentemente vários estudos são realizados sobre um mesmo assunto em locais ou instantes de tempo diferentes.
- “Non-Invasive Ventilation in Patients with Heart Failure: A Systematic Review and Meta-Analysis.” (Bittencourt et al, 2017).

Evidência Científica de Acordo com o Tipo de Estudo