Estatística

Prof. Ismael Bastos

UFRJ

Introdução à Estatística e Informações da Disciplina

O que é Estatística

Entendemos a Estatística como 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.

Áreas da Estatística

Podemos dividir a Estatística em três áreas:

- Estatística Descritiva
- Probabilidade
- Inferência Estatística

Áreas da Estatística

- **Estatística Descritiva** (*Conheça seus dados*) : Etapa inicial em qualquer análise de dados. É um conjunto de métodos estatísticos que tem por objetivo a organização, descrição e o resumo de dados.

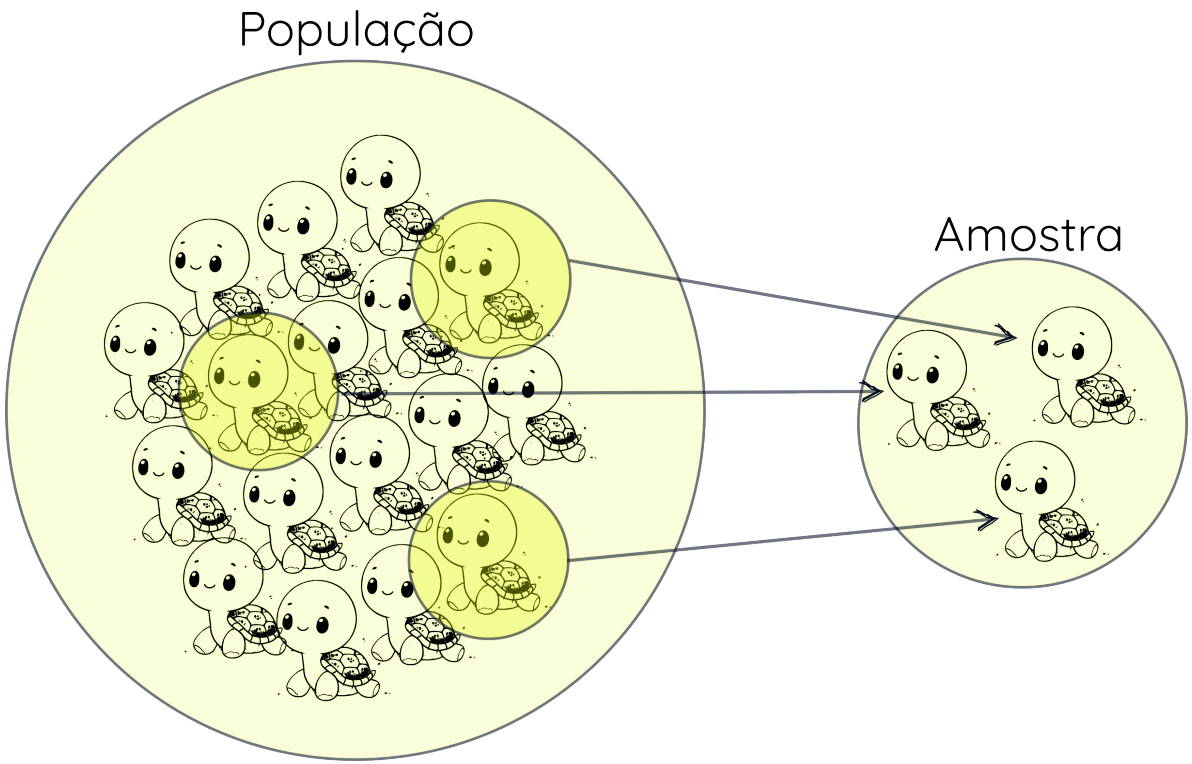
Áreas da Estatística

- **Estatística Descritiva** (*Conheça seus dados*) : Etapa inicial em qualquer análise de dados. É um conjunto de métodos estatísticos que tem por objetivo a organização, descrição e o resumo de dados.
- **Probabilidade** (*Qual a incerteza associada aos dados?*) : Auxilia na modelagem de fenômenos aleatórios, ou seja, aqueles em que há incerteza.

Áreas da Estatística

- **Estatística Descritiva** (*Conheça seus dados*) : Etapa inicial em qualquer análise de dados. É um conjunto de métodos estatísticos que tem por objetivo a organização, descrição e o resumo de dados.
- **Probabilidade** (*Qual a incerteza associada aos dados?*) : Auxilia na modelagem de fenômenos aleatórios, ou seja, aqueles em que há incerteza.
- **Inferência Estatística** (*Quais conclusões podemos tirar a partir destes dados?*): É um método estatístico que possibilita tirar conclusões sobre uma população a partir de informações contidas na amostra.

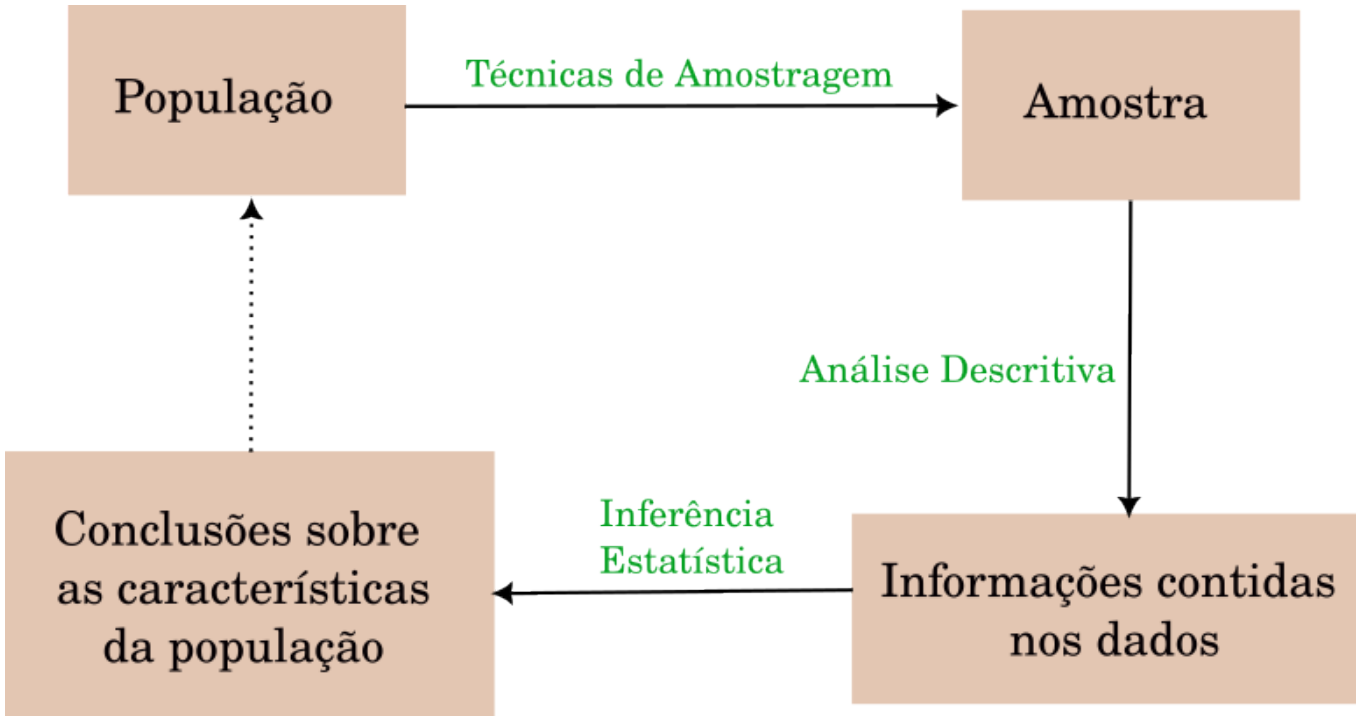
População x Amostra



População x Amostra

Por que precisamos selecionar uma amostra?

Áreas da Estatística



Informações da Disciplina

- Trabalho 1 valendo 10 pontos
- Trabalho 2 valendo 10 pontos
- Prova 1 valendo 10 pontos
- Prova 2 valendo 10 pontos

Antiga:

$$\text{Nota Final} = 0,2 \cdot T_1 + 0,4 \cdot P_1 + 0,4 \cdot P_2$$

Nova:

$$\text{Nota Final} = 0,4 \cdot T_1 + 0,2 \cdot T_2 + 0,4 \cdot P_2$$

Funcionamento da disciplina

- Aulas expositivas com conteúdo em slide.
- Resolução de exemplos e exercícios no quadro *.
- Uso da linguagem de programação R.
- Disponibilização de informações da disciplina através do site da disciplina e do SIGA.
- Listas de exercício
- Alguns códigos estarão presentes nos slides, mas alguns outros serão disponibilizados separadamente.

*: A resolução de exemplos e exercícios não será disponibilizada no site.

Linguagem de programação R

- Por que usar a linguagem R para estatística?
- Instalação do R e R Studio.
- Posit Cloud.

Perguntas iniciais

- Como garantir que uma amostra representa bem a população?
    - Qual metodologia devo utilizar para selecionar minha amostra? **Exemplo**: Quero conduzir uma pesquisa eleitoral na cidade de São Paulo, como devo escolher as pessoas que vou entrevistar?

Perguntas iniciais

- Como garantir que uma amostra representa bem a população?
  - Qual metodologia devo utilizar para selecionar minha amostra?
    Exemplo: Quero conduzir uma pesquisa eleitoral na cidade de São Paulo, como devo escolher as pessoas que vou entrevistar?
  - Qual tamanho de amostra selecionar?
    Exemplo: Sei que o tamanho da população de São Paulo é de aproximadamente $11,45$ milhões de habitantes, quantas pessoas devo escolher?

Amostragem

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.
3. Colocamos esses papeis em uma urna bem misturados.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.
3. Colocamos esses papeis em uma urna bem misturados.
4. Tiramos os n papeis correspondentes à amostra.

Amostragem aleatória simples

Amostragem aleatória simples: Consideremos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória simples é feita da seguinte forma:

1. Numeramos os itens da população com número de 1 até N.
2. Escrevemos cada um desses números em uma pedaço de papel.
3. Colocamos esses papeis em uma urna bem misturados.
4. Tiramos os n papeis correspondentes à amostra.

Pergunta: Essa retirada é com ou sem reposição?

Amostragem aleatória simples - Exemplo

O estudo O perfil socioeconômico e a percepção ambiental dos pescadores da Lagoa de Apodi, Rio Grande do Norte, Brasil propõe uma amostragem aleatória simples como metodologia de amostragem.

Amostragem aleatória simples - Exemplo no R

Nesse exemplo do R estamos assumindo que temos uma população de 10 indivíduos cujos nomes estão armazenadas no vetor *populacao* e desejamos obter uma amostra de tamanho 3 utilizando a amostragem aleatória simples.

Amostragem aleatória simples no R

```
n = 3
populacao = c("Tom", "Lia", "Ema", "Max", "Ana",
↪  "Lua", "Mia", "Isa", "Ilo", "Gal")
amostra = sample(populacao, n)
```

Amostragem aleatória simples - Exemplo no R

Nesse exemplo do R estamos assumindo que temos uma população de 10 indivíduos cujos nomes estão armazenadas no vetor *populacao* e desejamos obter uma amostra de tamanho 3 utilizando a amostragem aleatória simples.

Amostragem aleatória simples no R

```
n = 3
populacao = c("Tom", "Lia", "Ema", "Max", "Ana",
↪  "Lua", "Mia", "Isa", "Ilo", "Gal")
amostra = sample(populacao, n)
```

A função *sample* gera uma amostra aleatória simples de tamanho n, que nesse caso é 3.

Amostragem aleatória simples - Exemplo no R

Nesse exemplo do R estamos assumindo que temos uma população de 10 indivíduos cujos nomes estão armazenadas no vetor *populacao* e desejamos obter uma amostra de tamanho 3 utilizando a amostragem aleatória simples.

Amostragem aleatória simples no R

```
n = 3
populacao = c("Tom", "Lia", "Ema", "Max", "Ana",
↪  "Lua", "Mia", "Isa", "Ilo", "Gal")
amostra = sample(populacao, n)
```

A função *sample* gera uma amostra aleatória simples de tamanho *n*, que nesse caso é 3.

Observação: A função *sample* por padrão gera uma amostra selecionada sem reposição, para gerar uma amostra com reposição, basta adicionar o argumento *replace* = *TRUE*.

Amostragem aleatória estratificada

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).

Amostragem aleatória estratificada

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:

Amostragem aleatória estratificada

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada estrato é proporcional ao tamanho do estrato.
   - **Igualitária**: Selecionamos o mesmo número de indivíduos em cada estrato, não levando em conta o tamanho.

Amostragem aleatória estratificada

Amostragem aleatória estratificada: Consideremos que temos uma população (heterogênea) de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Determinamos o tamanho da amostra que iremos retirar de cada estrato, podendo ser feito de duas principais formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada estrato é proporcional ao tamanho do estrato.
   - **Igualitária**: Selecionamos o mesmo número de indivíduos em cada estrato, não levando em conta o tamanho.
3. Realizamos a amostragem aleatória simples em cada estrato.

Amostragem aleatória estratificada - Exemplo

O estudo Hipertensão Arterial e Diabetes Mellitus entre trabalhadores da saúde: associação com hábitos de vida e estressores ocupacionais propõe uma amostragem aleatória estratificada como metodologia de amostragem.

Amostragem aleatória estratificada - Exemplo no R

Amostragem aleatória estratificada (Proporcional) no R

```
n = 3
populacao = c("Tom", "Lia", "Ema", "Max", "Ana",
 ↪  "Lua", "Mia", "Isa", "Ilo", "Gal")
populacao_M = c("Tom", "Max", "Ilo")
populacao_F = c("Lia", "Ema", "Ana", "Lua", "Mia",
 ↪  "Isa", "Gal")
amostra_1 = sample(populacao_M, round(0.3 * n))
amostra_2 = sample(populacao_F, round(0.7 * n))
amostra = c(amostra_1, amostra_2)
```

Amostragem aleatória estratificada

Amostragem aleatória estratificada em duas etapas:
Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada em duas etapas é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).

Amostragem aleatória estratificada

Amostragem aleatória estratificada em duas etapas:
Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada em duas etapas é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Selecionamos aleatoriamente subconjuntos desses estratos.

Amostragem aleatória estratificada

Amostragem aleatória estratificada em duas etapas:
Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada em duas etapas é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Selecionamos aleatoriamente subconjuntos desses estratos.
3. Realizamos amostragem aleatória simples dentro de cada grupo.

Amostragem aleatória estratificada

Amostragem aleatória estratificada em duas etapas:
Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória estratificada em duas etapas é feita da seguinte forma:

1. Dividimos a população em estratos (com base em Idade, Renda, Sexo, Escolaridade ou outra variável).
2. Selecionamos aleatoriamente subconjuntos desses estratos.
3. Realizamos amostragem aleatória simples dentro de cada grupo.

Atenção: O número de etapas não está necessariamente relacionado ao número de variáveis.

Amostragem aleatória por conglomerados

Amostragem aleatória por conglomerados: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).

Amostragem aleatória por conglomerados

Amostragem aleatória por conglomerados: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).
2. Selecionamos aleatoriamente um subconjunto desses conglomerados.

Amostragem aleatória por conglomerados

Amostragem aleatória por conglomerados: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória por conglomerados é feita da seguinte forma:

1. Dividimos a população em conglomerados (como, por exemplo, escolas, unidades de saúde, bairros, etc.).
2. Selecionamos aleatoriamente um subconjunto desses conglomerados.
3. Todos os indivíduos dentro dos conglomerados selecionados são estudados

Amostragem aleatória por conglomerados

Como garantir que ao final teremos os n indivíduos que desejávamos?

Amostragem aleatória por conglomerados

Como garantir que ao final teremos os n indivíduos que desejávamos?

Amostragem aleatória por conglomerados em duas etapas:

1. Realizamos a amostragem aleatória por conglomerados conforme descrita no slide anterior.
2. Definimos o tamanho da amostra retirada de cada conglomerado, podendo ser feito de duas formas:
   - **Proporcional**: A quantidade de indivíduos amostrados em cada conglomerado é proporcional ao tamanho do estrato.
   - **Igualitário**: Selecionamos o mesmo número de indivíduos em cada conglomerado, não levando em conta o tamanho.
3. Realizamos a amostragem aleatória simples em cada conglomerado.

Amostragem aleatória por conglomerados X Amostragem aleatória estratificada

Importante: Não confundir amostragem aleatória por conglomerados com amostragem aleatória estratificada.

Amostragem aleatória por conglomerados	**Amostragem aleatória estratificada**
Grupos (Conglomerados) heterogêneos em seu interior	Grupos (Estratos) homogêneos em seu interior
Grupos (Conglomerados) homogêneos entre si	Grupos (Estratos) heterogêneos entre si

Amostragem aleatória sistemática

Amostragem aleatória sistemática: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória sistemática é feita da seguinte forma:

1. Determinamos o intervalo de amostragem $k = \frac{N}{n}$.

Amostragem aleatória sistemática

Amostragem aleatória sistemática: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória sistemática é feita da seguinte forma:

1. Determinamos o intervalo de amostragem $k = \frac{N}{n}$.
2. Selecionamos aleatoriamente um número entre 1 e k como ponto de partida.

Amostragem aleatória sistemática

Amostragem aleatória sistemática: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória sistemática é feita da seguinte forma:

1. Determinamos o intervalo de amostragem $k = \frac{N}{n}$.
2. Selecionamos aleatoriamente um número entre 1 e k como ponto de partida.
3. Selecionamos os elementos da amostra a partir do ponto de partida, seguindo o intervalo k.

Amostragem aleatória sistemática

Amostragem aleatória sistemática: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória sistemática é feita da seguinte forma:

1. Determinamos o intervalo de amostragem $k = \frac{N}{n}$.
2. Selecionamos aleatoriamente um número entre 1 e k como ponto de partida.
3. Selecionamos os elementos da amostra a partir do ponto de partida, seguindo o intervalo k.

Perguntas:

1. No passo 1, precisamos de fato definir k dessa forma? Em qual situação que poderíamos defini-lo de outra forma?

Amostragem aleatória sistemática

Amostragem aleatória sistemática: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória sistemática é feita da seguinte forma:

1. Determinamos o intervalo de amostragem $k = \frac{N}{n}$.
2. Selecionamos aleatoriamente um número entre 1 e k como ponto de partida.
3. Selecionamos os elementos da amostra a partir do ponto de partida, seguindo o intervalo k.

Perguntas:

1. No passo 1, precisamos de fato definir k dessa forma? Em qual situação que poderíamos defini-lo de outra forma?
2. Qual o principal possível problema dessa abordagem?

Amostragem aleatória sistemática

Amostragem aleatória sistemática: Consideremos que temos uma população de tamanho N e desejamos uma amostra de tamanho n. A amostragem aleatória sistemática é feita da seguinte forma:

1. Determinamos o intervalo de amostragem $k = \frac{N}{n}$.
2. Selecionamos aleatoriamente um número entre 1 e k como ponto de partida.
3. Selecionamos os elementos da amostra a partir do ponto de partida, seguindo o intervalo k.

Perguntas:

1. No passo 1, precisamos de fato definir k dessa forma? Em qual situação que poderíamos defini-lo de outra forma?
2. Qual o principal possível problema dessa abordagem?

Amostragem aleatória por conglomerados e sistemática - Exemplo

O 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. propõe uma amostragem aleatória por conglomerados seguido pela amostragem aleatória sistemática como metodologia de amostragem.

Amostragem aleatória sistemática - Exemplo no R

Amostragem aleatória sistemática no R

```
populacao = c("Tom", "Lia", "Ema", "Max", "Ana",
↪  "Lua", "Mia", "Isa", "Ilo", "Gal")
N = length(populacao)
n = 5
k = N/n
amostra = c()
valor_inicial = sample(seq(1,k), 1)
for(i in seq(valor_inicial,N, k)){
  amostra = c(amostra, populacao[i])
}
```

Amostragem aleatória sistemática - Exemplo no R

O código so slide anterior realiza a amostragem aleatória sistemática em um cenário em que se deseja obter uma amostra de tamanho 5.

Outros tipos de amostragem

Considere, agora, que um aluno esteja interessado em avaliar a opinião dos alunos da UFRJ sobre o serviço de transporte entre os diversos campi, oferecido pela administração da universidade. Como ele não tem **condições** nem **tempo** de selecionar uma amostra de todos os alunos a UFRJ, decide entrevistar seus colegas de turma.

Outros tipos de amostragem

Considere, agora, que um aluno esteja interessado em avaliar a opinião dos alunos da UFRJ sobre o serviço de transporte entre os diversos campi, oferecido pela administração da universidade. Como ele não tem **condições** nem **tempo** de selecionar uma amostra de todos os alunos a UFRJ, decide entrevistar seus colegas de turma.

- A decisão desse aluno é razoável? Ou seja, essa amostra é representativa?
- Qual a diferença desse método para os que estudamos anteriormente?

Outros tipos de amostragem

Considere, agora, que um aluno esteja interessado em avaliar a opinião dos alunos da UFRJ sobre o serviço de transporte entre os diversos campi, oferecido pela administração da universidade. Como ele não tem **condições** nem **tempo** de selecionar uma amostra de todos os alunos a UFRJ, decide entrevistar seus colegas de turma.

- A decisão desse aluno é razoável? Ou seja, essa amostra é representativa?
- Qual a diferença desse método para os que estudamos anteriormente?

Esse tipo de amostragem é chamado de **amostragem por conveniência** e é um tipo de amostragem não-probabilística.

Outros tipos de amostragem

Os métodos de amostragem não-probabilísticos são ruins e não devem ser utilizados???

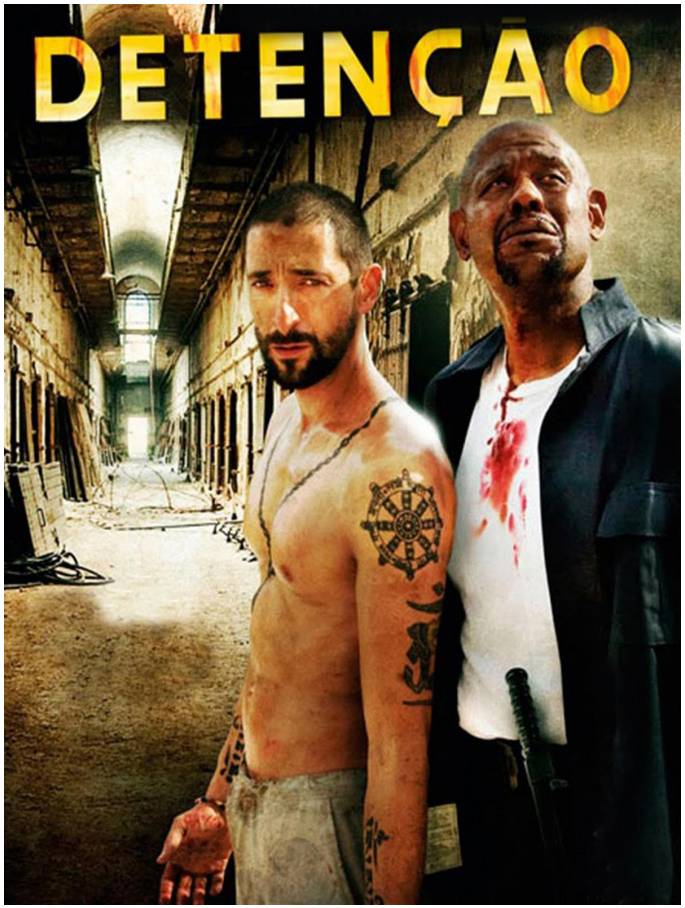
Problemas dos métodos de amostragem probabilísticos discutidos

- Como vou ter acesso a lista de todas as pessoas da população?

Problemas dos métodos de amostragem probabilísticos discutidos

- Como vou ter acesso a lista de todas as pessoas da população?
- Será que a informação sobre a proporção de elementos em cada estrato é confiável?

O problema do viés - Experimento de aprisionamento de Stanford (1971)



O problema do viés - Experimento de aprisionamento de Stanford (1971)

- Participantes de uma comunidade local foram convidados através de um anúncio de jornal que recrutava estudantes do sexo masculino para participar de um "**estudo psicológico da vida em uma prisão**".

O problema do viés - Experimento de aprisionamento de Stanford (1971)

- Participantes de uma comunidade local foram convidados através de um anúncio de jornal que recrutava estudantes do sexo masculino para participar de um "**estudo psicológico da vida em uma prisão**".
- Caso o estudante fosse aprovado, receberia \$15 por dia (equivalente a aproximadamente \$120 atualmente).

O problema do viés - Experimento de aprisionamento de Stanford (1971)

- Participantes de uma comunidade local foram convidados através de um anúncio de jornal que recrutava estudantes do sexo masculino para participar de um "**estudo psicológico da vida em uma prisão**".
- Caso o estudante fosse aprovado, receberia \$15 por dia (equivalente a aproximadamente \$120 atualmente).
- Os participantes foram submetidos a testes psicológicos e alocados aleatoriamente na posição de prisioneiro ou carcereiro.
- **Conclusão do Experimento**: Pessoas "boas" podem ter um "mal" comportamento quando colocadas em posições de poder ou sob condições de estresse.

Exemplos no nosso cotidiano

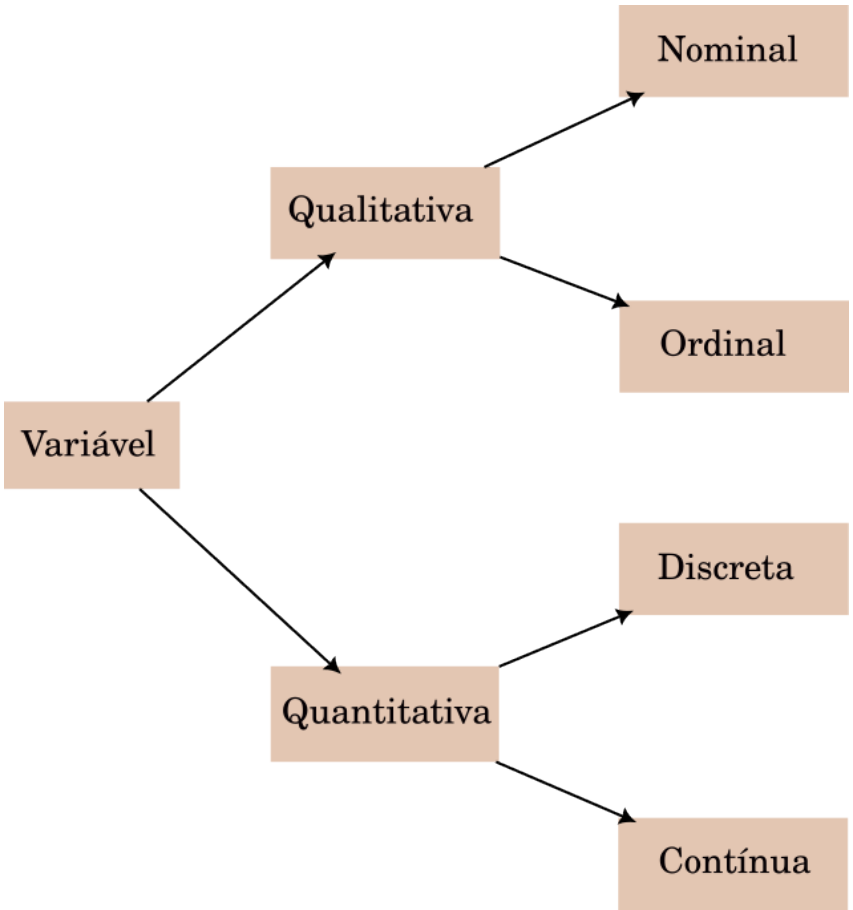
- Pesquisa eleitoral sobre intenções de voto para 2026 - UOL.
- Pesquisa eleitoral sobre intenções de voto para 2026 - SP - CNN.
- Pesquisa eleitoral sobre intenções de voto para 2026 - SP - Paraná Pesquisas.
- Pharmacokinetics and pharmacodynamics of cisatracurium in young and elderly adult patients.
- The Women's Health Initiative Randomized Trials and Clinical PracticeA Review

Análise Exploratória de Dados

Tipos de variáveis

- **Variável:** Característica analisada (Nome da coluna)
  - **Quantitativa:** Assume valores numéricos
    - Discreta: Assume valores discretos, valores inteiros.
    - Contínua: Assume valores no intervalo dos números reais.
  - **Qualitativa:** Assume valores que representam atributos e/ou qualidades
    - **Ordinal**: Valores assumem ordem, indicando intensidade crescentes de realização
    - **Nominal**: Valores não assumem ordem

Tipos de variáveis



Tipos de variáveis

Exemplo inicial

Est. Civil	Instrução	Filhos	Salário	Idade
Solteiro	Fundamental	0	4,00	26
Casado	Médio	1	4,56	32
Casado	Superior	3	6,20	34
Solteiro	Médio	2	5,23	21
Casado	Superior	2	3,23	24
Solteiro	Médio	1	7,90	56
Casado	Fundamental	4	6,45	67
Casado	Fundamental	0	4,56	34
Solteiro	Médio	1	6,78	56
Solteiro	Superior	0	3,56	34

Tipos de variáveis

Atenção: Ao analisar as variáveis estamos olhando para a informação trazida e não para o valor em si.

Tipos de variáveis

Atenção: Ao analisar as variáveis estamos olhando para a informação trazida e não para o valor em si.

Est. Civil	Instrução	Filhos	Salário	Idade
1	1	0	4,00	26
2	2	1	4,56	32
2	3	3	6,20	34
1	2	2	5,23	21
2	3	2	3,23	24
1	2	1	7,90	56
2	1	4	6,45	67
2	1	0	4,56	34
1	2	1	6,78	56
1	3	0	3,56	34

Tipos de variáveis

Atenção: Ao analisar as variáveis estamos olhando para a informação trazida e não para o valor em si.

Est. Civil	Instrução	Filhos	Salário	Idade
1	1	0	4,00	26
2	2	1	4,56	32
2	3	3	6,20	34
1	2	2	5,23	21
2	3	2	3,23	24
1	2	1	7,90	56
2	1	4	6,45	67
2	1	0	4,56	34
1	2	1	6,78	56
1	3	0	3,56	34

Nesse caso a tabela acima é equivalente a vista no slide anterior e a interpretação das variáveis permanece a mesma.

Tipos de variáveis

Vamos para um exemplo real. Abrir base de dados do curso.

Tabela de frequência

Uma tabela de frequência relaciona categorias ou classes de valores juntamente com as frequências do número de valores que se enquadram em cada categoria ou classe.
Denotaremos por:

- n_i: Frequência absoluta
- f_i: Frequência relativa

onde, $f_i = \frac{n_i}{n}$

Tabela de frequência

Exemplo 1

Construa a tabela de frequência para a variável formação utilizando os dados da tabela abaixo

idade	experiência(anos)	sexo	cidade	formação	salário (R$)
20	2	1	Rio de Janeiro	Farmácia	6000
25	5	0	São Paulo	Engenharia Química	5000
35	7	1	Rio de Janeiro	Farmácia	10000
40	15	1	Rio de Janeiro	Engenharia Química	6000
45	25	1	Rio de Janeiro	Farmácia	7500
22	1	0	São Paulo	Farmácia	1740
34	7	0	Rio de Janeiro	Farmácia	7000
56	25	1	São Paulo	Engenharia Química	6500
19	0	1	Rio de Janeiro	Engenharia Química	1990
29	9	0	São Paulo	Engenharia Química	8000

Tabela de frequência

Formação	n_i	f_i
Farmácia	5	0,5
Engenharia Química	5	0,5
total	n=10	1

Tabela de frequência - Exemplo no R

No exemplo a seguir é criado o vetor *populacao* contendo 10 indivíduos, sendo armazenado o sexo (M - Masculino, F - Feminino).

Tabela de frequência no R (Frequência Absoluta)

```
populacao = c("F", "M", "M", "F", "M", "F", "M", "F",
↪  "F", "F")
freq_absoluta = table(populacao)
print(freq_absoluta)
```

Tabela de frequência - Exemplo no R

No exemplo a seguir é criado o vetor *populacao* contendo 10 indivíduos, sendo armazenado o sexo (M - Masculino, F - Feminino).

Tabela de frequência no R (Frequência Absoluta)

```
populacao = c("F", "M", "M", "F", "M", "F", "M", "F",
↪  "F", "F")
freq_absoluta = table(populacao)
print(freq_absoluta)
```

Ao executar o código percebemos que a função *table* retorna apenas uma tabela contendo a frequência absoluta.

Tabela de frequência - Exemplo no R (Frequência Relativa)

Uma forma de mostrar a frequência relativa é apresentada no código abaixo.

Tabela de frequência no R (Frequência Relativa)

```
populacao = c("F", "M", "M", "F", "M", "F", "M", "F",
↪  "F", "F")
freq_absoluta = table(populacao)
freq_relativa = prop.table(freq_absoluta)
print(freq_relativa)
```

Tabela de frequência - Exemplo no R (Frequência Relativa)

Uma forma de mostrar a frequência relativa é apresentada no código abaixo.

Tabela de frequência no R (Frequência Relativa)

```
populacao = c("F", "M", "M", "F", "M", "F", "M", "F",
↪  "F", "F")
freq_absoluta = table(populacao)
freq_relativa = prop.table(freq_absoluta)
print(freq_relativa)
```

Atenção: Perceba que no código acima, a função que gera a tabela de frequência relativa (*prop.table*) recebe como argumento a própria tabela de frequência absoluta.

Tabela de frequência - Frequência Acumulada

Uma coluna que pode ser últil dentro de uma tabela de frequncia é a coluna referente à fequência acumulada.
Denotaremos a frequência acumulada por f_{ac}, sendo obtida pela soma acumulada da coluna de frequencia relativa (f_i).

frametitleFrequencia acumulada - Exemplo no R Para obter a frequência acumulada, basta usar a função *cumsum* do R

Tabela de frequência no R (Frequência Acumulada)

```
populacao = c("F", "M", "M", "F", "M", "F", "M", "F",
↪  "F", "F")
freq_absoluta = table(populacao)
freq_relativa = prop.table(freq_absoluta)
freq_acumulada = cumsum(freq_relativa)
```

Tabela de frequência - Exemplo no R

- Como gero uma tabela contendo as duas informações?
- Como eu gero uma tabela de frequência a partir de dados reais?

Pausa - Aprendendo sobre DataFrames no R

Nome dos arquivos de código:

- Como criar e trabalhar com DataFrames em R: **criacao_dataframe.R**
- Como criar uma tabela de frequência completa no R: **tabela_freq.R**.
- Como ler bases de dados no R: **leitura_base_dados.R**

Pausa - Aprendendo sobre DataFrames no R

Nome dos arquivos de código:

- Como criar e trabalhar com DataFrames em R: **criacao_dataframe.R**
- Como criar uma tabela de frequência completa no R: **tabela_freq.R**.
- Como ler bases de dados no R: **leitura_base_dados.R**

Para aprender um pouco mais, recomendo a leitura da página 30 do livro **Uma introdução à programação com o R**.

Exemplo 2

Construa uma tabela de frequência para a variável Ocupacao da base de dados adotada no curso.

Tabela de frequência com divisão em classes

Exemplo 3

Construa uma tabela de frequência para a variável Idade da base de dados adotada no curso.

Tabela de frequência com divisão em classes

Exemplo 3

Construa uma tabela de frequência para a variável Idade da base de dados adotada no curso.

Será que usar uma tabela de frequência é a melhor forma de apresentar esses dados?

Frequência com divisão em classes

A tabela de frequência com divisão em classes é semelhante à tabela de frequência, mas com a diferença que os dados são agrupados em intervalos numéricos.

Frequência com divisão em classes

A tabela de frequência com divisão em classes é semelhante à tabela de frequência, mas com a diferença que os dados são agrupados em intervalos numéricos.

- Geralmente a divisão em classes é bastante útil para estudar variáveis quantitativas contínuas, pois geralmente os valores não se repetem.
- Tomemos como exemplo a variável salário da tabela vista anteriormente, nela apenas o valor 6000 se repete, todos os demais são únicos. Se contruísimmos uma tabela de frequência comum, a maior parte dos valores teria frequência absoluta igual a 1.

Frequência com divisão em classes

Exemplo 7: Em uma turma do curso de Engenharia foi feito o registro da idade de cada um dos estudantes dessa turma.

22	23	44	33	20	27	19	54	37	22
25	29	40	23	20	30	24	39	28	21

Construa uma tabela de frequência partindo da menor idade para a variável idade. Defina a amplitude como sendo igual a 5 e inclua apenas o limite inferior dos intervalos.

Frequência com divisão em classes

Exemplo 7: Em uma turma do curso de Engenharia foi feito o registro da idade de cada um dos estudantes dessa turma.

22	23	44	33	20	27	19	54	37	22
25	29	40	23	20	30	24	39	28	21

Construa uma tabela de frequência partindo da menor idade para a variável idade. Defina a amplitude como sendo igual a 5 e inclua apenas o limite inferior dos intervalos.

Tabela de frequência com divisão em classes

Idade	n_i	f_i	f_{ac}
[19, 24)	8	0,4	0,4
[24, 29)	4	0,2	0,6
[29, 34)	3	0,15	0,75
[34, 39)	1	0,05	0,8
[39, 44)	2	0,1	0,9
[44, 49)	1	0,05	0,95
[49, 54)	0	0	0,95
[54, 59)	1	0,05	1
total	n=20	1	

Tabela de frequência com divisão em classes

- Como determinar a **amplitude dos intervalos**?

Tabela de frequência com divisão em classes

- Como determinar a **amplitude dos intervalos**?
- Todos os intervalos devem possuir a mesma amplitude?

Tabela de frequência com divisão em classes

- Como determinar a **amplitude dos intervalos**?
- Todos os intervalos devem possuir a mesma amplitude?
- Devemos usar intervalos abertos, semi-abertos ou fechados?

Tabela de frequência com divisão em classes - Exemplo no R

No exemplo a seguir, iremos construir uma tabela de frequência com divisão em classes para a variável idade, definindo a amplitude como sendo igual a 5 e incluindo apenas o limite inferior dos intervalos.

Tabela de frequência com divisão em classes no R (Amplitudes Iguais)

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
idades = dados$Idade
tabela_freq = table(cut(idades, seq(min(idades),
↪  max(idades), 5)))
```

Tabela de frequência com divisão em classes - Exemplo no R

No exemplo do slide anterior, utilizamos a função *cut* passando como segundo argumento uma sequência. Podemos também passar como argumento o número de intervalos, dessa forma a própria função define a amplitude.

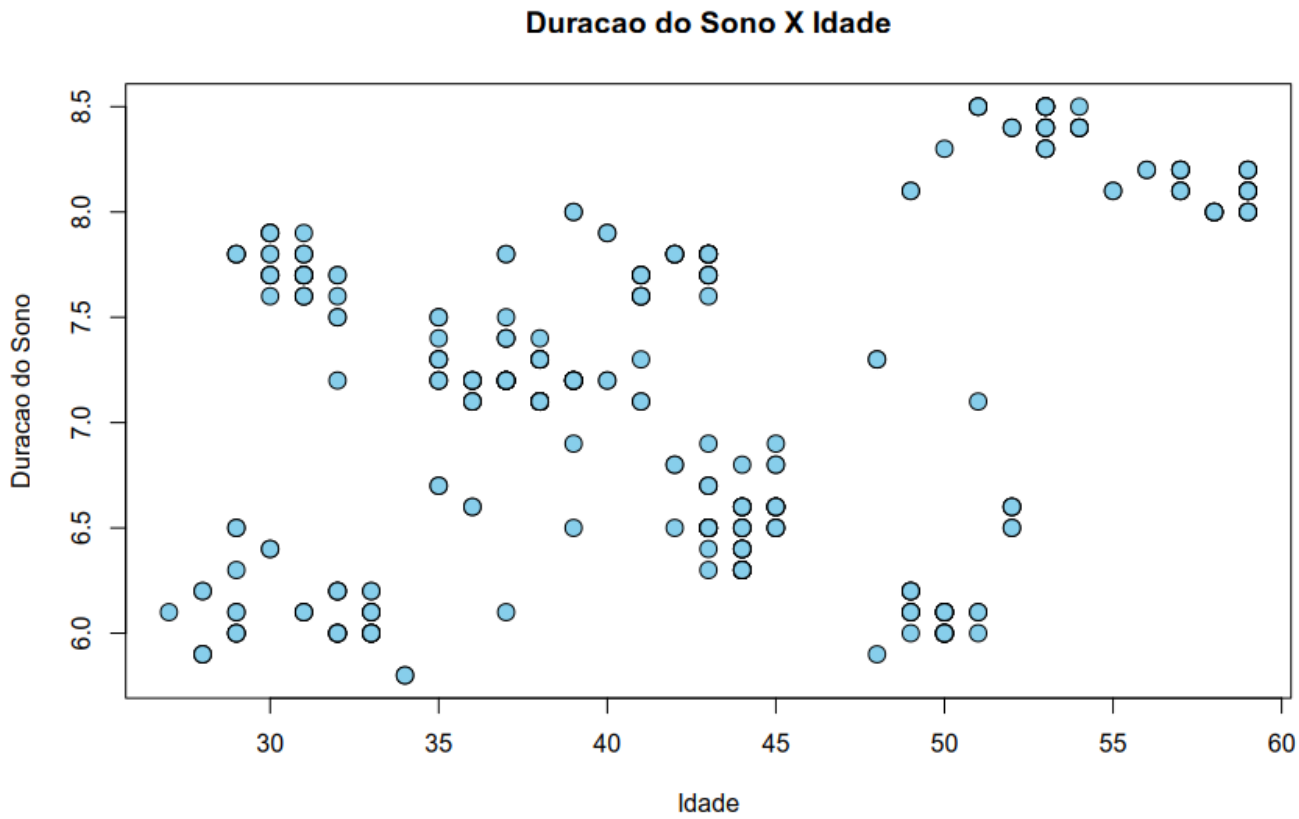
Tabela de frequência com divisão em classes no R (Amplitudes Iguais)

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
idades = dados$Idade
tabela_freq = table(cut(idades, 6))
```

Note que ao rodar o código acima temos 6 intervalos, não necessariamente de mesma amplitude.

Gráficos - Gráfico de dispersão

O gráfico de dispersão permite representar dados de duas ou três variáveis. Consiste em basicamente em dispor cada ponto no planno cartesiano.



Interpretação - Gráfico de dispersão

Em geral, a interpretação de um gráfico de dispersão se concentra em avaliar o comportamento de uma variável em relação à outra. Na Figura do slide anterior, podemos observar que **parece** que ao aumentar a idade, também há o aumento da duração do sono.

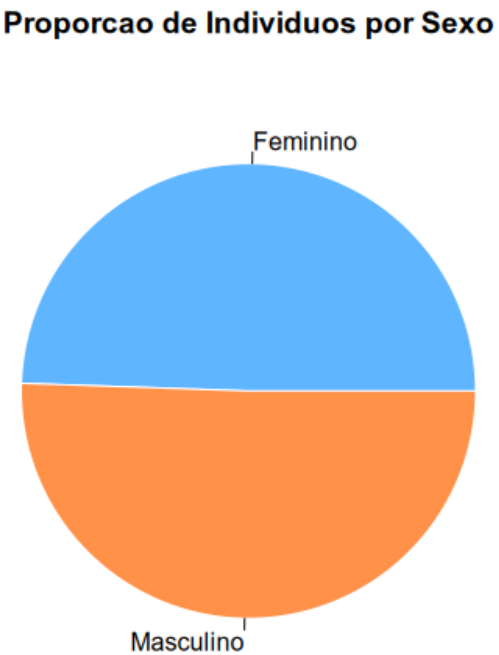
Gráficos - Gráfico de dispersão - Exemplo no R

Gráfico de dispersão no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
duracao_sono = dados$`Duracao do Sono`
idade = dados$Idade
plot(x=idade, y=duracao_sono,
     main="Duracao do Sono X Idade",
     xlab ="Idade", ylab = "Duracao do Sono",
     bg = "#87ceeb", # Cor dos pontos
     cex = 1.5, # Tamanho dos pontos
     pch=21) # Tipo do ponto
```

Gráficos - Gráfico de setores (pizza)

Consiste em dividir um círculo (pizza) em diferentes setores (fatias), cada um representando a proporção do elemento analisado em relação ao conjunto de estudo.



Atenção: Olhando para o gráfico do slide anterior, qual sexo parece ter maior proporção?

Atenção: Olhando para o gráfico do slide anterior, qual sexo parece ter maior proporção?
Devido a esse problema, recomenda-se evitar o uso de gráficos de pizza. Sendo recomendado:

- Colocar as porcentagens por extenso caso existam poucas categorias.
- Usar outro gráfico, como por exempo o gráfico de barras.

Gráficos - Gráfico de setores (pizza) - Exemplo no R

Gráfico de pizza no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
sexo = dados$Sexo
freq_sexo = table(sexo)
pie(freq_sexo,
    border="white", # Coloca bordas brancas
    col=c("#60B5FF", "#FF9149"), # Cor de cada fatia
    main = "Proporcao de Individuos por Sexo")
```

Gráficos - Gráfico de barras

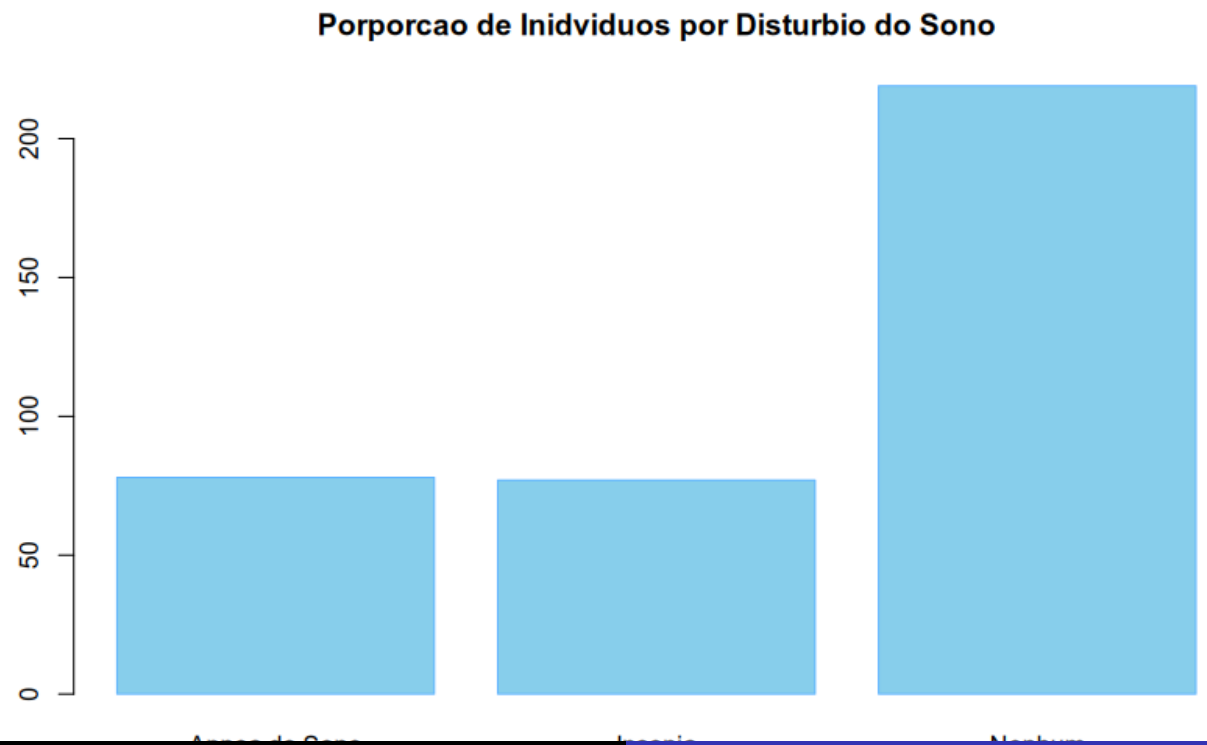
- Utiliza o plano cartesiano com os valores da variável no eixo das abcissas e as frequências ou porcentagens no eixo das ordenadas.

Gráficos - Gráfico de barras

- Utiliza o plano cartesiano com os valores da variável no eixo das abcissas e as frequências ou porcentagens no eixo das ordenadas.
- Utilizado, geralmente, para representar visualmente uma tabela de frequência.

Gráficos - Gráfico de barras

- Utiliza o plano cartesiano com os valores da variável no eixo das abcissas e as frequências ou porcentagens no eixo das ordenadas.
- Utilizado, geralmente, para representar visualmente uma tabela de frequência.



Interpretação - Gráfico de barras

- No gráfico de barras, estamos interessados em entender a altura da barra referente a cada categoria analisada.

Interpretação - Gráfico de barras

- No gráfico de barras, estamos interessados em entender a altura da barra referente a cada categoria analisada.
- Na Figura do slide anterior, podemos perceber que **parece** que temos a mesma proporção de indivíduos com apnea do sono e insonia, sendo o fato de não ter nenhum disturbio a Característica maoritariamente presente no conjunto de dados.

Gráficos - Gráfico de barras - Exemplo no R

Gráfico de barras no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
disturbio_sono = dados$`Disturbio do Sono`
freq_disturbio = table(disturbio_sono)
barplot(freq_disturbio,
        main="Porporcao de Inidviduos por Disturbio
        ↪  do Sono",
        border="#60B5FF",# Cor do contorno das barras
        col="#87ceeb") # Cor das barras
```

Gráficos - Histograma

- O histograma é bastante semelhante ao gráfico de barras.

Gráficos - Histograma

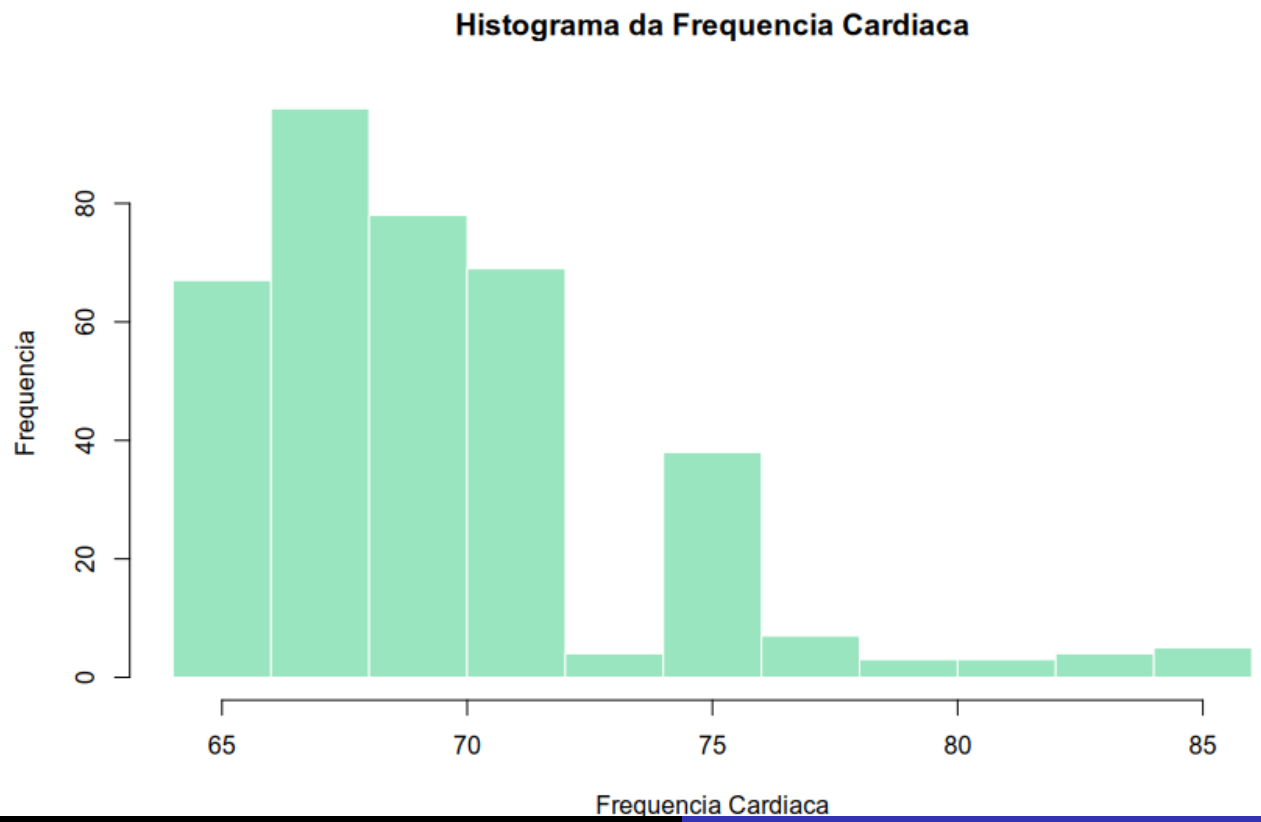
- O histograma é bastante semelhante ao gráfico de barras.
- Utilizado, em geral, como uma uma representação visual da uma tabela agrupada em classes.

Gráficos - Histograma

- O histograma é bastante semelhante ao gráfico de barras.
- Utilizado, em geral, como uma uma representação visual da uma tabela agrupada em classes.
- Evidencia a distribuição de uma variável quantitativa.

Gráficos - Histograma

- O histograma é bastante semelhante ao gráfico de barras.
- Utilizado, em geral, como uma uma representação visual da uma tabela agrupada em classes.
- Evidencia a distribuição de uma variável quantitativa.



Interpretação - Histograma

- Ao olhar para um histograma estamos interessados em analisar a distribuição dos valores, tentando perceber a concentração dos valores para a variável analisada.

Interpretação - Histograma

- Ao olhar para um histograma estamos interessados em analisar a distribuição dos valores, tentando perceber a concentração dos valores para a variável analisada.
- Uma outra possibilidade é tentar identificar algum padrão, podendo relacionar com distribuições estatísticas conhecidas.
- Na Figura do slide anterior, poodemos perceber que há uma maior concentração de frequências cardiacas antes do valor 72 batimentos por minuto.

Gráficos - Histograma - Exemplo no R

Histograma no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
freq_cardiaca = dados$`Frequencia Cardiaca`
hist(freq_cardiaca,
     main="Histograma da Frequencia Cardiaca",
     xlab = "Frequencia Cardiaca",
     ylab = "Frequencia",
     breaks=8, # Numero de barras (Nem sempre)
     col="#33CC8080", # Cor das barras
     border=FALSE) # Elimina as bordas
```

Gráficos - Histograma

- O argumento breaks define o número de barras, entretanto, para esse gráfico, todas as barras devem ter o mesmo tamanho.

Gráficos - Histograma

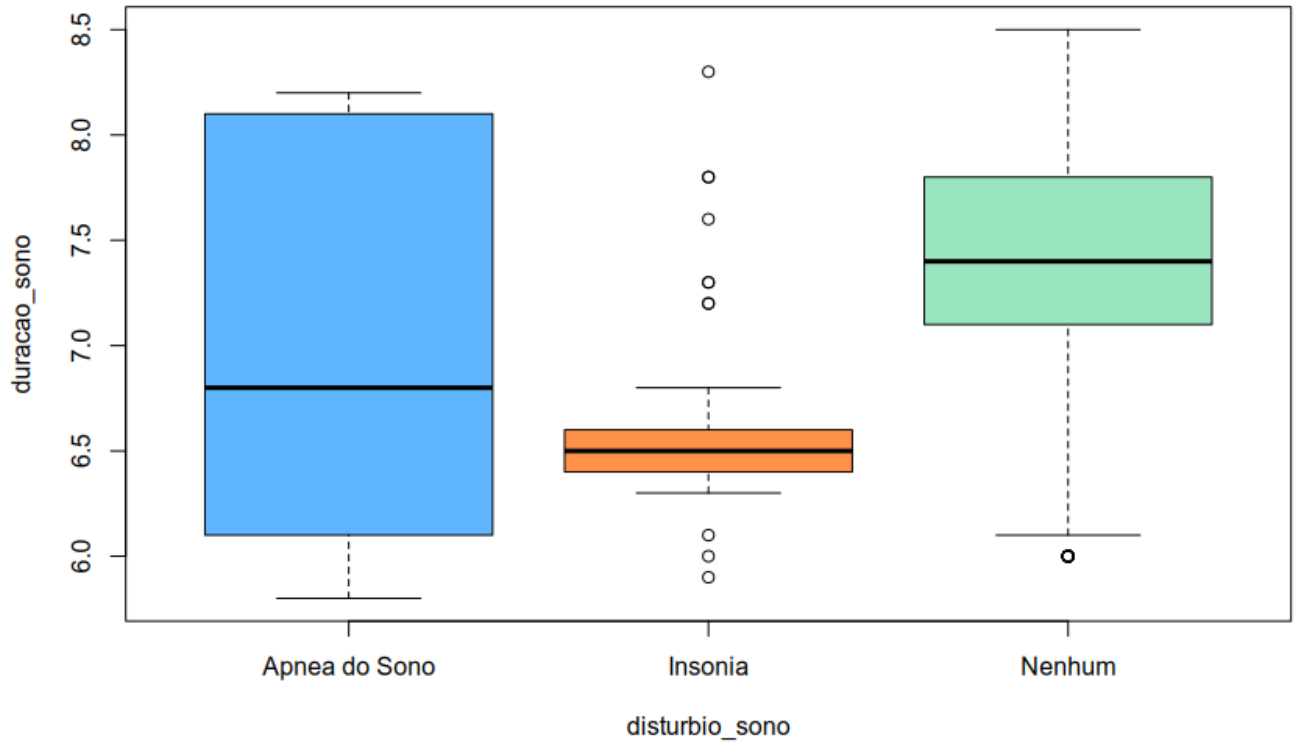
- O argumento breaks define o número de barras, entretanto, para esse gráfico, todas as barras devem ter o mesmo tamanho.
- Também é possível passar uma tabela de frequencia em classes dentro da função.

Gráficos - Boxplot

De forma inicial, o gráfico Boxplot apresenta a distribuição dos dados, mas, diferentemente do histograma, apresenta alguns elementos específicos, os quais serão vistos posteriormente.

Gráficos - Boxplot

De forma inicial, o gráfico Boxplot apresenta a distribuição dos dados, mas, diferentemente do histograma, apresenta alguns elementos específicos, os quais serão vistos posteriormente.



Gráficos - Extras

Gráfico de uma função matemática

```
modulo = function(x){
    return(abs(x))
}
curve(modulo,
    -2, 2, # Intervalo do eixo x
    ylim=c(0, 3), # Intervalo do eixo y
    col="#60B5FF", # Cor do grafico
    ann=FALSE) # Remove nomes nos eixos
```

Gráficos - Extras

Adicionar linhas verticais e horizontais

```
modulo = function(x){
    return(abs(x))
}
curve(modulo, -2, 2, ylim=c(0, 3),
      col="#60B5FF",
      ann=FALSE)
abline(h=0) # Adiciona uma linha horizontal y=0
abline(v=0) # Adiciona uma linha vertical x=0
```

Gráficos - Extras

Vários Gráficos em uma mesma Imagem

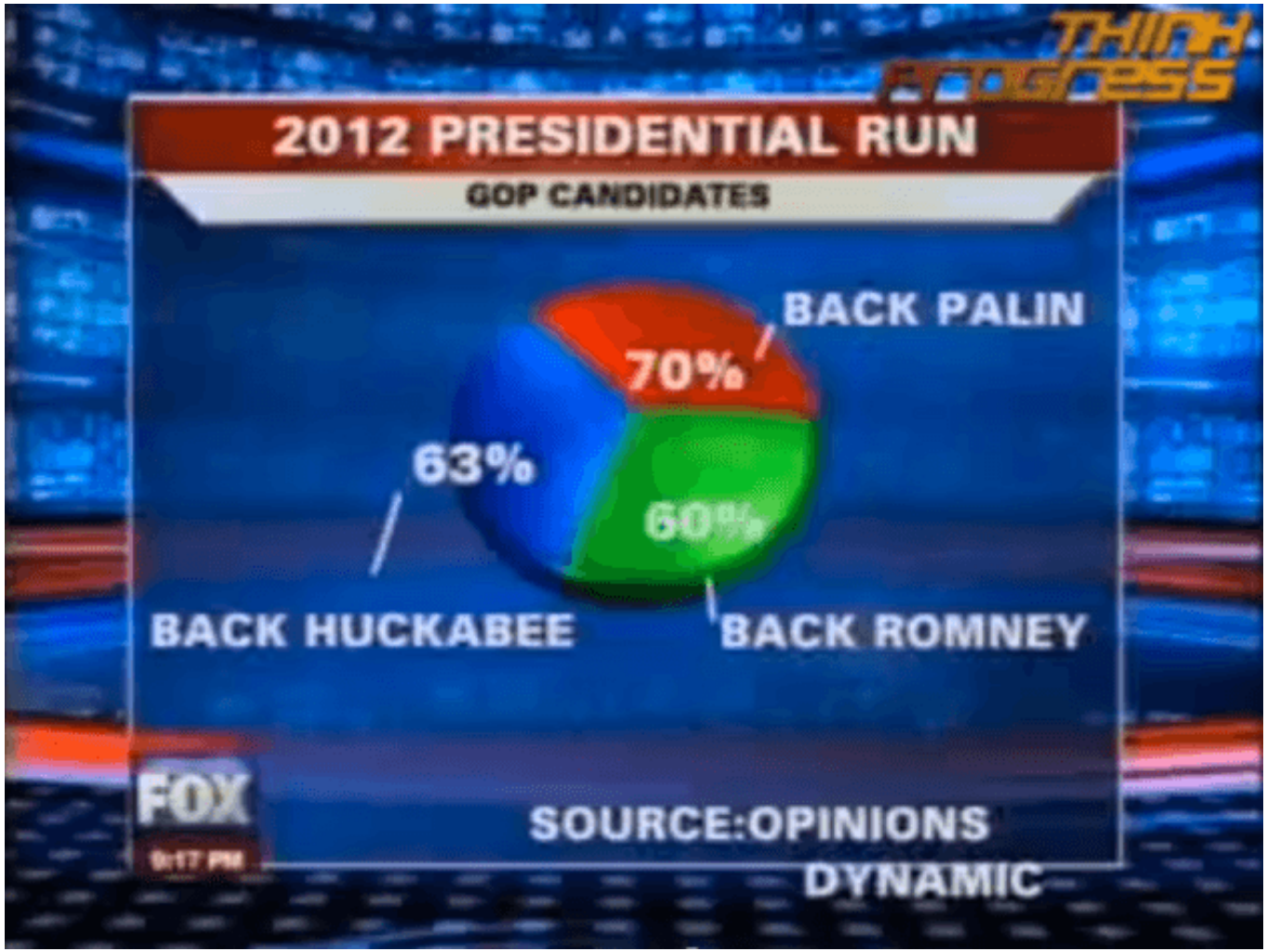
```
modulo = function(x){
    return(abs(x))
}
quadratica = function(x){
    return(x^2)
}
curve(modulo, -2, 2, col="#60B5FF",
      ann=FALSE, ylim=c(0, 3))
par(new=TRUE)
curve(quadratica, -2, 2, col="#FF9149",
      axes=FALSE, ann=FALSE, ylim=c(0, 3))
```

Gráficos - Extras

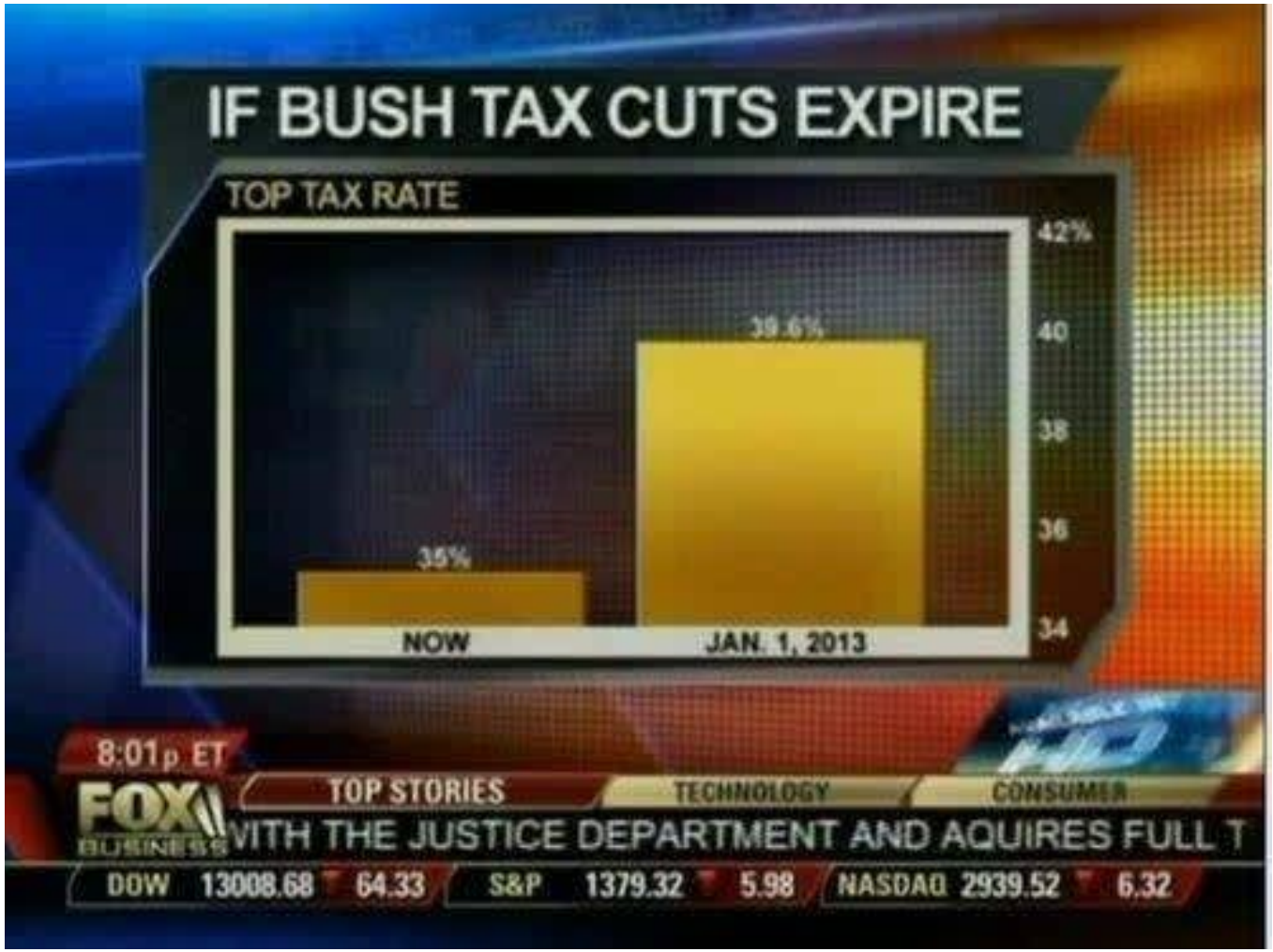
Adicionando pontos no plano cartesiano

```
modulo = function(x){
    return(abs(x))
}
quadratica = function(x){
    return(x^2)
}
curve(modulo, -2, 2, col="#60B5FF",
      ann=FALSE, ylim=c(0, 3))
par(new=TRUE)
curve(quadratica, -2, 2, col="#FF9149",
      axes=FALSE, ann=FALSE, ylim=c(0, 3))
points(-1, 1, col='red', pch=19) # Adiciona um ponto
↪  vermelho no ponto (-1,1)
points(1, 1, col='red', pch=19)  # Adiciona um ponto
```

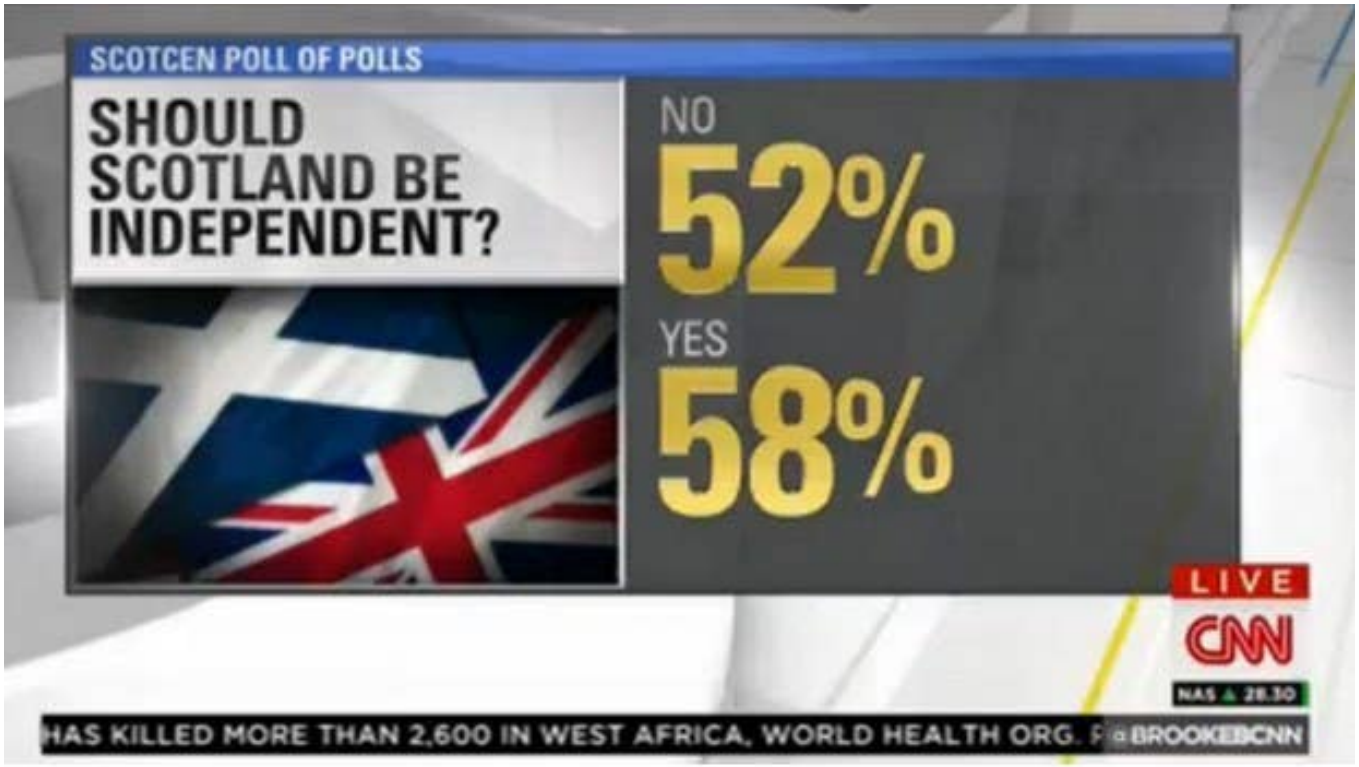
Gráficos - Exemplos a não serem seguidos



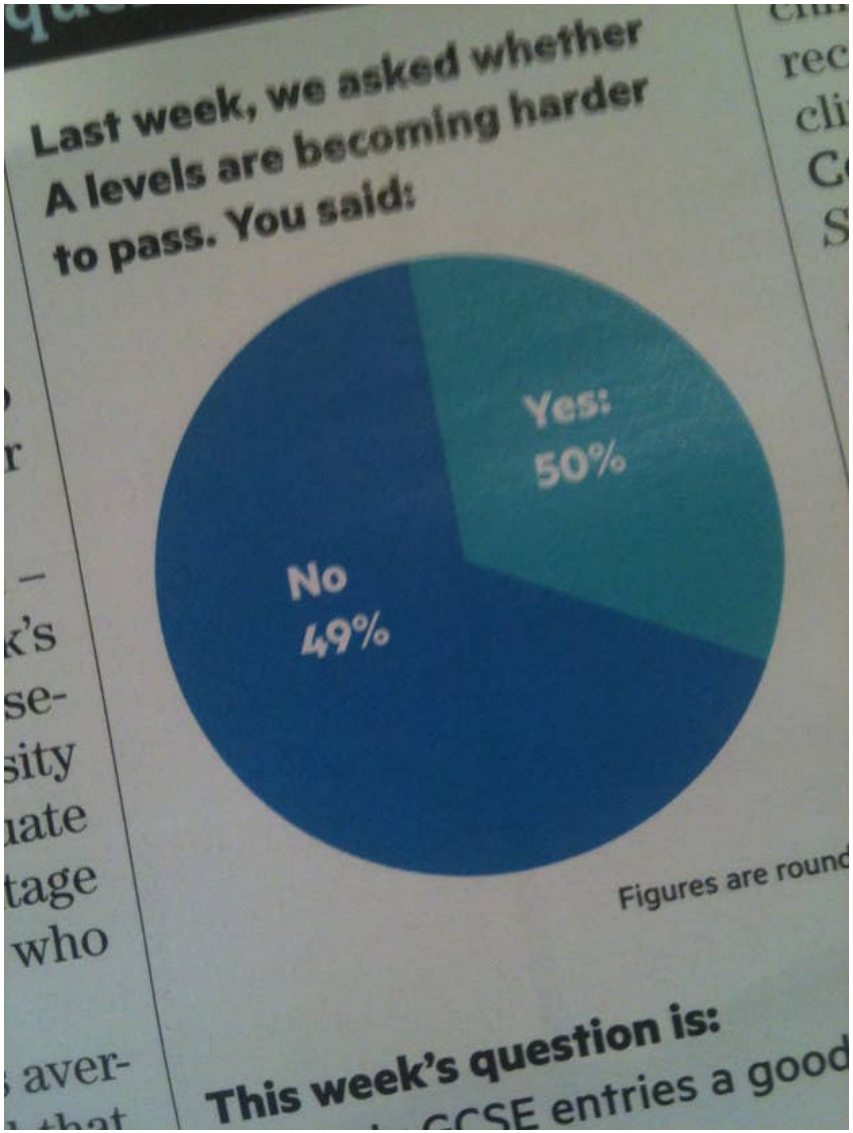
Gráficos - Exemplos a não serem seguidos



Gráficos - Exemplos a não serem seguidos



Gráficos - Exemplos a não serem seguidos



Gráficos - Exemplos a não serem seguidos

Gráficos - Exemplos a não serem seguidos

Medidas de Posição

Podemos ter o interesse em resumir nosso conjunto de dados, apresentando um ou mais valores que de certa forma representem todo o conjunto de dados.
Para determinar esses valores, utiliza-se as **medidas de posição central**: Média, Mediana e Moda.

Média Aritmética

Dado um conjunto de n observações $x_1, x_2, \ldots, x_n$ a média aritmética é definida como:

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

Média Aritmética

Exemplo 4

Suponha que tenhamos coletado a altura de 10 *alunos de uma turma, as alturas obtidas foram as seguintes:*

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Logo,

$$\bar{x} = \frac{1,90 + 1,67 + 1,87 + 1,55 + 1,76 + 1,87 + 1,95 + 1,66 + 1,75 + 1,60}{10} = 1,76$$

Média Aritmética - Exemplo no R

Para calcular a média de um conjunto de dados no R, utilizamos a função *mean*.

Calculando a Média Aritmética no R

```
altura =  c(1.90, 1.67, 1.87, 1.55, 1.76, 1.87, 1.95,
↪  1.66, 1.75, 1.60)
media_altura = mean(altura)
```

Média Aritmética

Exemplo 5

Obtenha a média da Idade dos indivíduos do conjunto de dados adotado na disciplina.

Moda

A moda de um conjunto de dados, representada por x^*, é o valor que mais se repete, ou seja, o valor mais frequente.

Moda

A moda de um conjunto de dados, representada por x^*, é o valor que mais se repete, ou seja, o valor mais frequente.
Atenção : Um conjunto de dados pode ter mais de uma moda.

Mediana

Seja $x_1, x_2, \ldots, x_n$ um conjunto de n observações, e seja $x_{(i)}, i = 1, \ldots, n$ o conjunto das observações ordenadas, de modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Então, a mediana Q_2 é definida como o valor tal que 50% das observações são menores e 50% são maiores que ela.

Mediana

Seja $x_1, x_2, \ldots, x_n$ um conjunto de n observações, e seja $x_{(i)}, i = 1, \ldots, n$ o conjunto das observações ordenadas, de modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Então, a mediana Q_2 é definida como o valor tal que 50% das observações são menores e 50% são maiores que ela.

Exemplo 6

Suponhamos que o peso de 7 crianças de uma determinada turma escolar tenha sido coletada, os pesos coletados são exibidos a seguir:

$$28, 34, 45, 39, 42, 33, 50$$

Obtenha a mediana desse conjunto de dados

Mediana

Primeiramente vamos ordenar os dados:

$$28, 33, 34, 39, 42, 45, 50$$

Mediana

Primeiramente vamos ordenar os dados:

$$28, 33, 34, 39, 42, 45, 50$$

Logo, a mediana será o valor 39

Mediana

Exemplo 7

Retomemos o exemplo das alturas dos alunos, temos os seguintes dados:

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Para calcular a mediana, primeiramente vamos ordenar os dados:

$$1,55; 1,60; 1,66; 1,67; 1,75; 1,76; 1,87; 1,87; 1,90; 1,95$$

Mediana

Exemplo 7

Retomemos o exemplo das alturas dos alunos, temos os seguintes dados:

$$1,90; 1,67; 1,87; 1,55; 1,76; 1,87; 1,95; 1,66; 1,75; 1,60$$

Para calcular a mediana, primeiramente vamos ordenar os dados:

$$1,55; 1,60; 1,66; 1,67; 1,75; 1,76; 1,87; 1,87; 1,90; 1,95$$

Temos 10 elementos, logo, não conseguimos encontrar o valor que divide exatamente ao meio o conjunto de dados, nesse caso, a mediana é dada pela média dos elementos que ocupam a posição $\frac{n}{2}$ e $\frac{n}{2}+1$.

$$Q_2 = \frac{1,75 + 1,76}{2} = 1,755$$

Mediana

Portanto,

$$Q_2 = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se n é impar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{se n é par} \end{cases}$$

Médiana - Exemplo no R

Para obter a mediana de um conjunto de dados no R, basta usar a função *median*

Calculando a Mediana no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
idades = dados$Idade
mediana_idade = median(idades)
```

Média X Mediana

- Imagine que você precisa atravessar um lago e existe uma placa que diz: Altura média do lago: $1,5$ metros. Você o atravessaria?

Média X Mediana

- Imagine que você precisa atravessar um lago e existe uma placa que diz: Altura média do lago: $1,5$ metros. Você o atravessaria?
- Dependendo do conjunto de dados, nem sempre a média é uma boa medido da resumo. A média é fortemente influenciada por valores aberrantes (outliers).

Média X Mediana

- Imagine que você precisa atravessar um lago e existe uma placa que diz: Altura média do lago: $1,5$ metros. Você o atravessaria?
- Dependendo do conjunto de dados, nem sempre a média é uma boa medido da resumo. A média é fortemente influenciada por valores aberrantes (outliers).

Exemplo 8

Os seguintes dados referentes aos salários (em R$) de cinco funcionários de uma firma:

$$136; 210; 350; 360; 2500$$

Calcule o salario médio.

Média X Mediana

No caso do slide anterior, o salário médio é igual a $R\$647,20$. No entanto, esse valor não representa, de forma adequada, os salários mais baixos e os salários mais altos, isso porque o mais alto é muito diferente dos demais.

Média X Mediana

- Por outro lado, a mediana desse conjunto de dados é o valor 350, que indica que metade dos valores é menor que 350 e a outra metade maior.
- Ainda assim, não conseguimos capturar o quanto menor ou maior são.

Observações

- Quando a média é próxima a mediana dizemos que isso é um indicativo de que a amostra não apresenta pontos aberrantes (possíveis outliers).
- A média é fortemente influenciada pela presença de pontos aberrantes, enquanto a mediana não é.

Média Ponderada

A média aritmética ponderada de números $x_1, x_2, \ldots, x_n$ com pesos $w_1, w_2, \ldots, w_n$ é definida como:

$$\bar{x}_w = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_n x_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Quartis - Elementos fundamentais

- O primeiro quartil, que indicaremos por Q_1, deixa 25% das observações abaixo e 75% acima dele.

Quartis - Elementos fundamentais

- O primeiro quartil, que indicaremos por Q_1, deixa 25% das observações abaixo e 75% acima dele.
- A mediana é o segundo quartil

Quartis - Elementos fundamentais

- O primeiro quartil, que indicaremos por Q_1, deixa 25% das observações abaixo e 75% acima dele.
- A mediana é o segundo quartil
- O terceiro quartil, Q_3, deixa 75% das observações abaixo e 25% acima dele.

Quartis - Elementos adicionais

- Amplitude Interquartil (AIQ) $= Q_3 - Q_1$

Quartis - Elementos adicionais

- Amplitude Interquartil (AIQ) $= Q_3 - Q_1$
- Limite Superior $= Q_3 + 1,5 \cdot AI$

Quartis - Elementos adicionais

- Amplitude Interquartil (AIQ) $= Q_3 - Q_1$
- Limite Superior $= Q_3 + 1,5 \cdot AI$
- Limite Inferior $= Q_1 - 1,5 \cdot AI$

Quartis - Cálculo dos quartis

$$Q_1 = x_{\left(\frac{1}{4}(n+1)\right)}$$

Quartis - Cálculo dos quartis

$$Q_1 = x_{\left(\frac{1}{4}(n+1)\right)}$$

$$Q_3 = x_{\left(\frac{3}{4}(n+1)\right)}$$

Importante: Se o valor obtido não for um inteiro, fazer uma média do elemento que ocupa a parte inteira e seu sucessor.

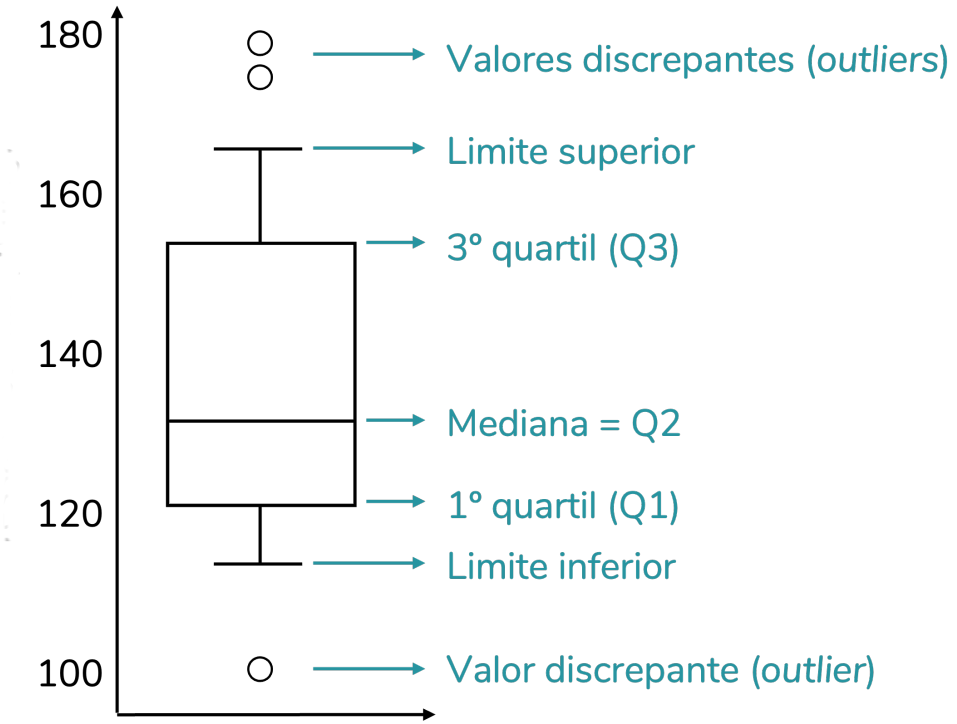
Quartis - Exemplo no R

Para obter os quartis de um conjunto de dados no R, utilizamos a função *quantile*.

Variância no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
duracao_sono = dados$`Duracao do Sono`
q1 = quantile(duracao_sono, 0.25) # Primeiro quartil
q3 = quantile(duracao_sono, 0.75) # Terceiro quartil
```

Revisitando o Boxplot



Boxplot - Exemplo no R

Gráfico Boxplot no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
disturbio_sono = dados$`Disturbio do Sono`
duracao_sono = dados$`Duracao do Sono`
boxplot(duracao_sono ~ disturbio_sono,
        col=c("#60B5FF", "#FF9149", "#33CC8080"))
```

Medidas de dispersão

Muitas vezes estamos interessados em entender o comportamento dos dados e as medidas de posição não são suficientes:

Exemplo 9

Calcule a média dos seguintes conjuntos de dados:

$$salarios_1 = 2.000,00; 3.500,00; 4.900,00; 3.100,00$$

$$salarios_2 = 3.375,00; .3375,00; 3.375,00; 3.375,00$$

Medidas de dispersão

Exemplo 10

Calcule a mediana dos seguintes conjuntos de dados:

$$salarios_1 = 2.000,00; 3.500,00; 4.900,00; 3.100,00; 4.100,00$$

$$salarios_2 = 2.000,00; 3.500,00; 70.900,00; 3.100,00; 32.000,00$$

Em ambos os casos anteriores, percebemos que a Média e Mediana não são suficientes para descrever o comportamento dos dados. Devido a isso, trabalharemos agora com medidas de dispersão, sendo elas: Amplitude, Devio médio absoluto, Variância, Desvio Padrão e Coeficiente de Variação.

Amplitude

A amplitude de um conjunto de dados é a distância entre o maior valor e o menor valor.

Amplitude

A amplitude de um conjunto de dados é a distância entre o maior valor e o menor valor.

$$\Delta_{total} = V_{max} - V_{min}$$

Amplitude

A amplitude de um conjunto de dados é a distância entre o maior valor e o menor valor.

$$\Delta_{total} = V_{max} - V_{min}$$

A amplitude nos dá uma ideia do intervalo de possíveis valores que a variável analisada pode assumir.

Amplitude - Exemplo no R

Para calcular a amplitude de um conjunto de dados no R, utilizamos a função *var*.

Amplitude no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
duracao_sono = dados$`Duracao do Sono`
amplitude_duracao_sono = max(duracao_sono) -
↪  min(duracao_sono)
```

Desvio médio absoluto

O desvio médio absoluto de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por:

$$DMA(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Desvio médio absoluto

O desvio médio absoluto de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por:

$$DMA(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- O desvio médio tem como objetivo verificar quanto os valores, em média, estão se distanciando da média.

Desvio médio absoluto

O desvio médio absoluto de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por:

$$DMA(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- O desvio médio tem como objetivo verificar quanto os valores, em média, estão se distanciando da média.
- Note que o desvio médio absoluto é sempre um valor positivo.

Desvio médio absoluto

O desvio médio absoluto de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por:

$$DMA(X) = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- O desvio médio tem como objetivo verificar quanto os valores, em média, estão se distanciando da média.
- Note que o desvio médio absoluto é sempre um valor positivo.

Exemplo 11

Calcule o desvio médio para os salários dos exemplos anteriores.

Variância

A variância σ^2 de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

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

Variância

A variância σ^2 de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

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

- Perceba que a variância é bastante semelhante ao desvio médio absoluto, no entanto, penaliza grandes desvios.

Variância

A variância σ^2 de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

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

- Perceba que a variância é bastante semelhante ao desvio médio absoluto, no entanto, penaliza grandes desvios.
- Observe que a variância é sempre um número **positivo**.

Variância

A variância σ^2 de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

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

- Perceba que a variância é bastante semelhante ao desvio médio absoluto, no entanto, penaliza grandes desvios.
- Observe que a variância é sempre um número **positivo**.
- Note que a variância **não** fornece um valor na mesma unidade de medida dos dados.

Variância

A variância σ^2 de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

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

- Perceba que a variância é bastante semelhante ao desvio médio absoluto, no entanto, penaliza grandes desvios.
- Observe que a variância é sempre um número **positivo**.
- Note que a variância **não** fornece um valor na mesma unidade de medida dos dados.
- A variância não tem interpretação prática direta, podendo ser utilizada como comparação.

Variância

Exemplo 12

Calcule a variância para os salários dos exemplos anteriores.

Variância - Exemplo no R

Para calcular a variância de um conjunto de dados no R, utilizamos a função *var*.

Variância no R

```
require("readxl")
dados = read_excel(file.choose(), sheet=1)
duracao_sono = dados$`Duracao do Sono`
var_duracao_sono = var(duracao_sono)
```

Desvio padrão

O desvio padrão de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido como a raiz quadrada da variância:

Desvio padrão

O desvio padrão de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido como a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2}$$

Desvio padrão

O desvio padrão de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido como a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2}$$

- Diferentemente da variância, o desvio padrão possui interpretação prática, pois ele volta o valor para a unidade de medida dos dados.

Desvio padrão

O desvio padrão de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido como a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2}$$

- Diferentemente da variância, o desvio padrão possui interpretação prática, pois ele volta o valor para a unidade de medida dos dados.
- Interpretamos o desvio padrão como quanto, em média, os valores do conjunto de dados estão se ditanciando da média.

Desvio padrão

O desvio padrão de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido como a raiz quadrada da variância:

$$\sigma = \sqrt{\sigma^2}$$

- Diferentemente da variância, o desvio padrão possui interpretação prática, pois ele volta o valor para a unidade de medida dos dados.
- Interpretamos o desvio padrão como quanto, em média, os valores do conjunto de dados estão se ditanciando da média.
- Note que, apesar de interpretável, é difícil dizer se temos um desvio padrão alto ou baixo, mas, com o auxilio da amplitude, podemos tirar boas conclusões.

Coeficiente de variação

Dado um conjunto de observações $x_1, x_2, \ldots, x_n$; o coeficiente de variação (CV) é definido como a razão entre o desvio-padrão dos dados e sua média, ou seja,

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

- Em geral, multiplica-se o valor do CV por 100 para ter um valor percentual.

Coeficiente de variação

Dado um conjunto de observações $x_1, x_2, \ldots, x_n$; o coeficiente de variação (CV) é definido como a razão entre o desvio-padrão dos dados e sua média, ou seja,

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

- Em geral, multiplica-se o valor do CV por 100 para ter um valor percentual.
- Quanto menor for o valor do CV, mais homogêneo é um conjunto de dados.

Coeficiente de variação

Podemos adotar o seguinte critério:

- $CV < 0,10 \implies$ variabilidade baixa
- $0,10 \leq CV < 0,20 \implies$ variabilidade intermediária
- $0,20 \leq CV < 0,30 \implies$ variabilidade alta
- $CV \geq 0,30 \implies$ variabilidade muito alta

Observação: Só podemos calcular o CV quando a média amostral for diferente de zero.

Escores padronizados

Suponhamos que tenha sido feito um estudo visando entender o desempenho dos alunos do curso de Engenharia Química no curso de Cálculo em comparação com o curso de Estatística.

Escores padronizados

Suponhamos que tenha sido feito um estudo visando entender o desempenho dos alunos do curso de Engenharia Química no curso de Cálculo em comparação com o curso de Estatística.

As notas de 9 alunos da turma foram coletadas e são apresentadas abaixo:

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Olhando para as notas, será que tirar 6 em Estatística tem o mesmo "peso" que tirar 6 em Cálculo?

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular a média das notas de cada um dos cursos:

$$\bar{x}_E = \frac{6+4+5+7+8+5+5+5+7}{9} = \frac{52}{9} = 5,78$$

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular a média das notas de cada um dos cursos:

$$\bar{x}_E = \frac{6+4+5+7+8+5+5+5+7}{9} = \frac{52}{9} = 5,78$$

$$\bar{x}_C = \frac{6+8+9+10+7+7+8+9+3}{9} = \frac{67}{9} = 7,44$$

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular a média das notas de cada um dos cursos:

$$\bar{x}_E = \frac{6+4+5+7+8+5+5+5+7}{9} = \frac{52}{9} = 5,78$$

$$\bar{x}_C = \frac{6+8+9+10+7+7+8+9+3}{9} = \frac{67}{9} = 7,44$$

- Podemos perceber que, o aluno 1, por exemplo, ficou acima da média em Estatística, mas abaixo da média em Cálculo.

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular a média das notas de cada um dos cursos:

$$\bar{x}_E = \frac{6+4+5+7+8+5+5+5+7}{9} = \frac{52}{9} = 5,78$$

$$\bar{x}_C = \frac{6+8+9+10+7+7+8+9+3}{9} = \frac{67}{9} = 7,44$$

- Podemos perceber que, o aluno 1, por exemplo, ficou acima da média em Estatística, mas abaixo da média em Cálculo.
- Outra forma de verificar essa afirmação é por meio do **desvio**.

Escores Padronizados - Desvio

O desvio de uma observação x_i em torna da média é definido como:

$$d_i = x_i - \bar{x}$$

Escores Padronizados - Desvio

O desvio de uma observação x_i em torna da média é definido como:

$$d_i = x_i - \bar{x}$$

Calculando os desvios da nota do primeiro aluno, temos:

$$d_{E_1} = 6 - 5,78 = 0,22$$

$$d_{C_1} = 6 - 7,44 = -1,44$$

Escores Padronizados - Desvio

O desvio de uma observação x_i em torna da média é definido como:

$$d_i = x_i - \bar{x}$$

Calculando os desvios da nota do primeiro aluno, temos:

$$d_{E_1} = 6 - 5,78 = 0,22$$

$$d_{C_1} = 6 - 7,44 = -1,44$$

- Dessa forma, temos que o aluno 1 ficou, aproximadamente, $0,22$ pontos acima da média em Cálculo e 1,44 abaixo da média em Estatística.

Escores Padronizados - Desvio

O desvio de uma observação x_i em torna da média é definido como:

$$d_i = x_i - \bar{x}$$

Calculando os desvios da nota do primeiro aluno, temos:

$$d_{E_1} = 6 - 5,78 = 0,22$$

$$d_{C_1} = 6 - 7,44 = -1,44$$

- Dessa forma, temos que o aluno 1 ficou, aproximadamente, $0,22$ pontos acima da média em Cálculo e 1,44 abaixo da média em Estatística.
- Ainda asssim, não temos como comparar se o desempenho do aluno foi melhor em Estatística ou em Cálculo, pois, novamente, as médias são diferentes.

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular o desvio padrão das notas de cada um dos cursos.

$$\sigma_E^2 = 1,51 \implies \sigma_E = 1,23$$

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular o desvio padrão das notas de cada um dos cursos.

$$\sigma_E^2 = 1,51 \implies \sigma_E = 1,23$$

$$\sigma_C^2 = 3,80 \implies \sigma_C = 1,95$$

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular o desvio padrão das notas de cada um dos cursos.

$$\sigma_E^2 = 1,51 \implies \sigma_E = 1,23$$

$$\sigma_C^2 = 3,80 \implies \sigma_C = 1,95$$

- Qual disciplina apresenta maior variabilidade das notas?

Escores Padronizados

Aluno	1	2	3	4	5	6	7	8	9
Estatística	6	4	5	7	8	5	5	5	7
Cálculo	6	8	9	10	7	7	8	9	3

Vamos calcular o desvio padrão das notas de cada um dos cursos.

$$\sigma_E^2 = 1,51 \implies \sigma_E = 1,23$$

$$\sigma_C^2 = 3,80 \implies \sigma_C = 1,95$$

- Qual disciplina apresenta maior variabilidade das notas?
- Faça o cálculo da variância manualmente e verifique os resultados apresentados nesse slide.

Escores Padronizados

O escore padronizado de uma observação x_i é definido como:

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

Escores Padronizados

O escore padronizado de uma observação x_i é definido como:

$$z_i = \frac{x_i - \bar{x}}{\sigma_x}$$

- Ao dividirmos pelo desvio padrão, a escala passa a ser definida em termos do desvio padrão e cada escore padronizado nos informa que a observação está acima ou abaixo da média por um determinado número de desvios-padrão. Dessa forma, removemos o efeito das médias e o fato das variabilidades serem diferentes.
- Importante: A média dos escores padronizados é, independente do conjunto de dados, sempre igual a 0 e a variância igual a 1.

Escores Padronizados

Vamos analisar agora as notas de Estatística e Cálculo em termos dos escores padronizados:

Aluno		1	2	3	4	5	6	7	8	9
Estatística	Nota	6	4	5	7	8	5	5	5	7
	Escore	0,18	-1,45	-0,63	1,00	1,81	-0,63	-0,63	-0,63	1,00
Cálculo	Nota	6	8	9	10	7	7	8	9	3
	Escore	-0,74	0,29	0,80	1,13	-0,23	-0,20	0,29	0,80	-3,28

- Podemos perceber que a nota 6 está, aproximadamente, $0,18$ desvios-padrão acima da média em Estatística e aproximadamente $0,74$ desvios-padrão abaixo da média das notas de Cálculo.
- Perceba que tirar a nota 10 em Cálculo é menos "surpreendente" que tirar 8 em Estatística.

Escores Padronizados - Teorema de Chebyshev

Para qualquer distribuição de dados, pelo menos $(1 - 1/z^2)$ dos dados estão dentro de z desvios-padrão da média, onde z é qualquer valor maior que 1. Ou seja, pelo menos $(1\text{-}1/z^2)$*dosdadosestãonointervalo*

$$\bar{x} - z\sigma; \bar{x} + z\sigma$$

.

Escores Padronizados - Teorema de Chebyshev

Exemplos:

- Para $z = 2$ temos que $1 - 1/z^2 = 3/4 = 75\%$ dos dados estão dentro de dois desvios-padrão da média. Ou, equivalentemente, 75% dos escores padronizados estão no intervalo $(-2, 2)$.

Escores Padronizados - Teorema de Chebyshev

Exemplos:

- Para $z = 2$ temos que $1 - 1/z^2 = 3/4 = 75\%$ dos dados estão dentro de dois desvios-padrão da média. Ou, equivalentemente, 75% dos escores padronizados estão no intervalo $(-2, 2)$.
- Para $z = 3$ que $1 - 1/z^2 = 8/9 = 89\%$ dos dados estão dentro de três desvios-padrão da média. Ou, equivalentemente, 89% dos escores padronizados estão no intervalo $(-3, 3)$.

Escores Padronizados - Teorema de Chebyshev

Exemplos:

- Para $z = 2$ temos que $1 - 1/z^2 = 3/4 = 75\%$ dos dados estão dentro de dois desvios-padrão da média. Ou, equivalentemente, 75% dos escores padronizados estão no intervalo $(-2, 2)$.
- Para $z = 3$ que $1 - 1/z^2 = 8/9 = 89\%$ dos dados estão dentro de três desvios-padrão da média. Ou, equivalentemente, 89% dos escores padronizados estão no intervalo $(-3, 3)$.
- Para $z = 4$ que $1 - 1/z^2 = 15/16 = 93,75\%$ dos dados estão dentro de quatro desvios-padrão da média. Ou, equivalentemente, 93,75% dos escores padronizados estão no intervalo $(-4, 4)$.

Covariância e Correlação

- Até agora, vimos como analisar medidas referentes a uma única variável. Entretanto, na prática, podemos nos deparar com situaçãos nas quais se faz interessante estudar a relação entre duas ou mais variáveis.

Covariância e Correlação

- Até agora, vimos como analisar medidas referentes a uma única variável. Entretanto, na prática, podemos nos deparar com situaçãos nas quais se faz interessante estudar a relação entre duas ou mais variáveis.
- Nesse sentido, surge a Covariância e a Correlação, que são medidas usadas para analisar a relação entre duas ou mais variáveis.

Covariância

A covariância entre duas variáveis X e Y é definida por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

Covariância

A covariância entre duas variáveis X e Y é definida por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- A covariância mede o grau de associação **linear** entre as variáveis.

Covariância

A covariância entre duas variáveis X e Y é definida por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- A covariância mede o grau de associação **linear** entre as variáveis.
- A unidade de medida é dada pelo produto das unidades de medida das variáveis X e Y.

Covariância

A covariância entre duas variáveis X e Y é definida por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- A covariância mede o grau de associação **linear** entre as variáveis.
- A unidade de medida é dada pelo produto das unidades de medida das variáveis X e Y.
- Importante: A covariância depende da escala dos dados, o que faz com que seja difícil estabelecer comparações.

Covariância

A covariância entre duas variáveis X e Y é definida por:

$$Cov(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

- A covariância mede o grau de associação **linear** entre as variáveis.
- A unidade de medida é dada pelo produto das unidades de medida das variáveis X e Y.
- Importante: A covariância depende da escala dos dados, o que faz com que seja difícil estabelecer comparações.
- Seus valores podem variar de $-\infty$ a ∞

Covariância - Exemplo no R

Para calcular a covariância entre duas variáveis no R, utilizamos a função *cov*.

Covariância no R

```
dados = read.csv(file.choose())
duracao_sono = dados$`Duracao do Sono`
idade = dados$Idade
cov_dsono_idade = cov(duracao_sono, idade)
```

Impacto da escala dos dados no cálculo da covariância

Exemplo no R, arquivo: escala_covariancia.R

Coeficiente de correlação

O coeficiente de correlação entre duas variáveis X e Y é definido por:

$$Corr(X, Y) = \rho(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_X} \right) \left(\frac{y_i - \bar{y}}{\sigma_Y} \right)$$

Coeficiente de correlação

O coeficiente de correlação entre duas variáveis X e Y é definido por:

$$Corr(X, Y) = \rho(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_X} \right) \left(\frac{y_i - \bar{y}}{\sigma_Y} \right)$$

- O coeficiente de correlação também mede o grau de associação **linear** entre as variáveis.

Coeficiente de correlação

O coeficiente de correlação entre duas variáveis X e Y é definido por:

$$Corr(X, Y) = \rho(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_X} \right) \left(\frac{y_i - \bar{y}}{\sigma_Y} \right)$$

- O coeficiente de correlação também mede o grau de associação **linear** entre as variáveis.
- O coeficiente de correlação **não** depende da escala dos dados, permitindo comparações.

Coeficiente de correlação

O coeficiente de correlação entre duas variáveis X e Y é definido por:

$$Corr(X, Y) = \rho(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_X} \right) \left(\frac{y_i - \bar{y}}{\sigma_Y} \right)$$

- O coeficiente de correlação também mede o grau de associação **linear** entre as variáveis.
- O coeficiente de correlação **não** depende da escala dos dados, permitindo comparações.
- Seus valores podem variar entre -1 e 1.

Classificação do coeficiente de correlação

Valor de ρ (+ ou −)	**Interpretação**
0,00 a 0,19	Correlação bem fraca
0,20 a 0,39	Correlação fraca
0,40 a 0,69	Correlação moderada
0,70 a 0,89	Correlação forte
0,90 a 1,00	Correlação muito forte

Correlação - Exemplo no R

Para calcular o coeficiente de correlação entre duas variáveis no R, utilizamos a função *cor*.

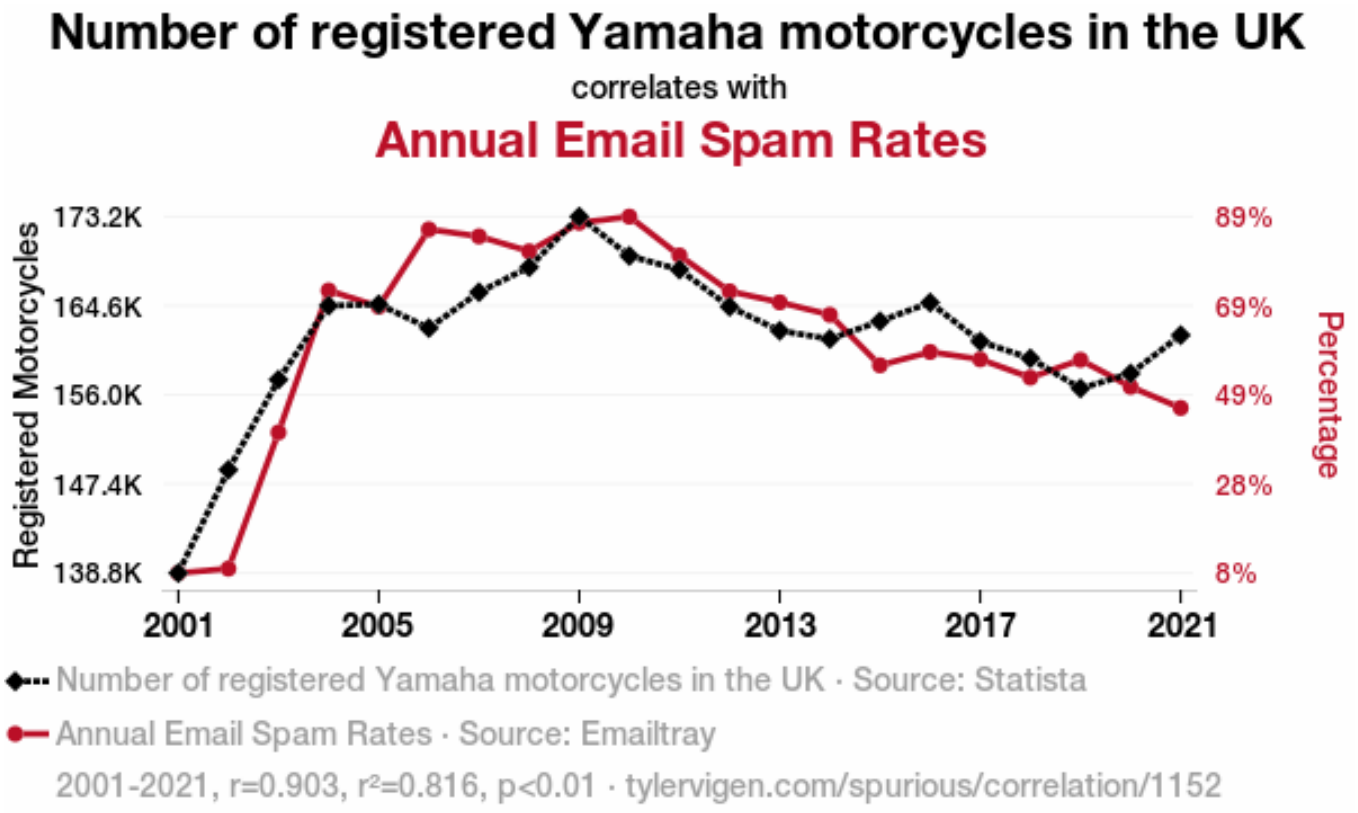
Correlação no R

```
dados = read.csv(file.choose())
duracao_sono = dados$`Duracao do Sono`
idade = dados$idade
cor_dsono_idade = cor(duracao_sono, idade)
```

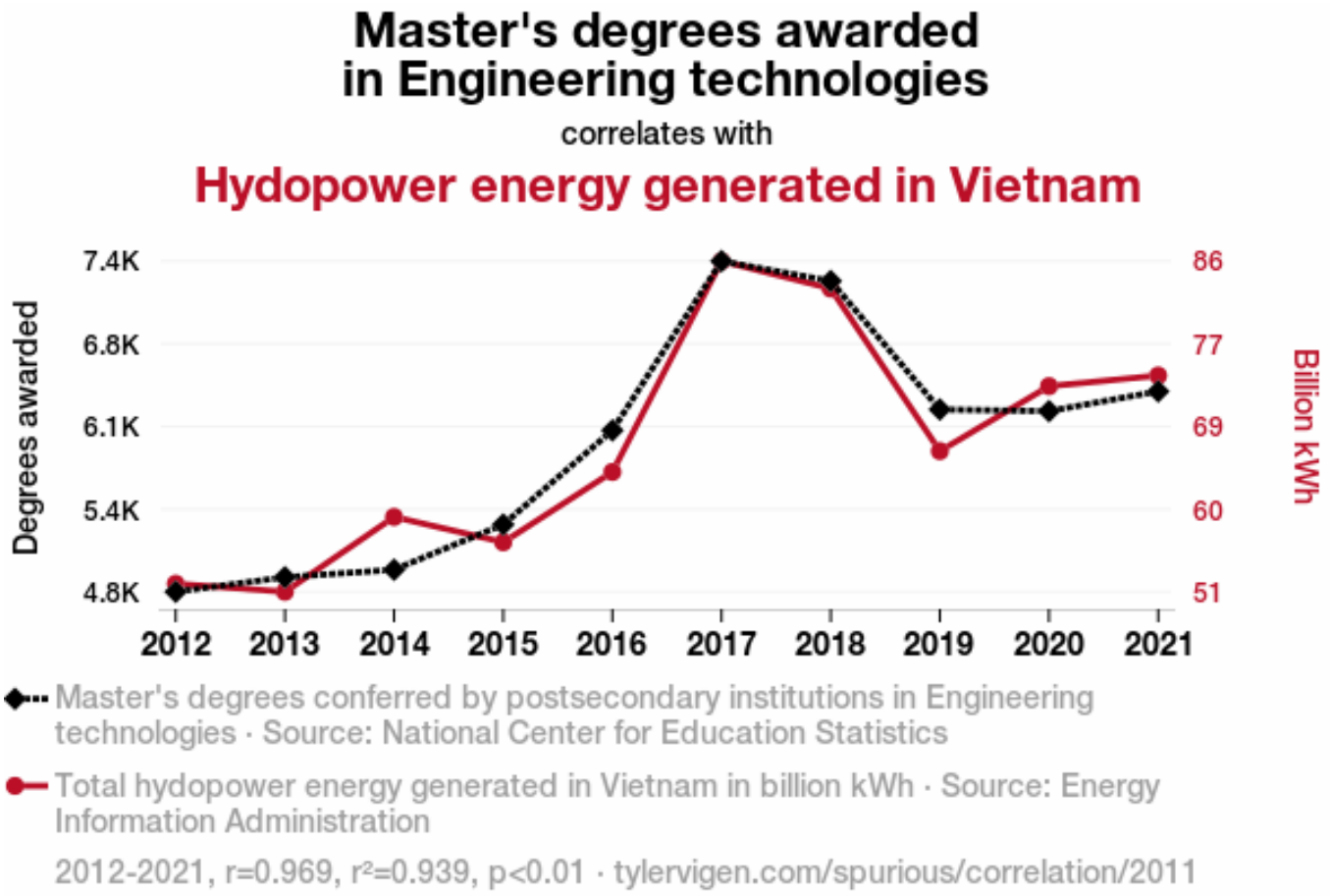
Correlação **Linear** (Apenas Linear)

Exemplo no R, arquivo: correlacao_linear.R

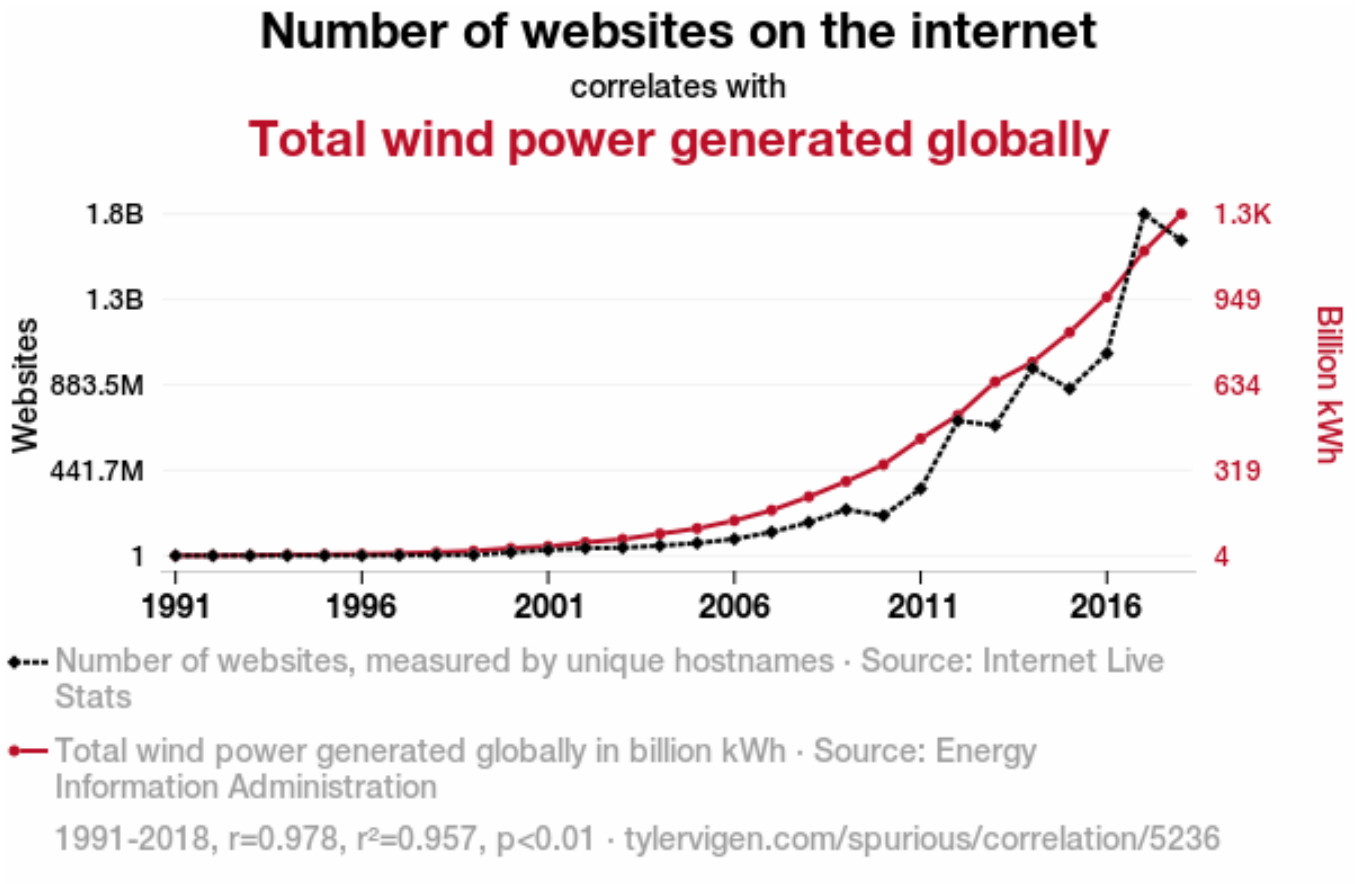
Correlação não implica causalidade



Correlação não implica causalidade



Correlação não implica causalidade



Fim!

Chegamos ao fim da primeira parte da matéria de estatística!

Fim!

Chegamos ao fim da primeira parte da matéria de estatística!

Antes de passarmos para a Probabilidade, vamos falar sobre os seguintes pontos:

- Como usar o Quarto.
- Lacunas do R (salvar DataFrame, filtrar DataFrame, valores únicos e obter dimensões do dataframe)
- Explicação do trabalho.