# Dados e amostragem

Conjunto de dados, tipos de variáveis, fontes de dados, amostragem probabilística e não probabilística

Prof. Me. Lineu Alberto Cavazani de Freitas

Departamento de Estatística
Laboratório de Estatística e Geoinformação

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

# Dados

# O que são dados?

- Dados são **conjuntos de valores**.
- Podem ser de diferentes fontes, tais como **estudos** e **experimentos**.
- Podem conter **variáveis** de diferentes tipos.
- Podem surgir em formatos **estruturados** e **não estruturados**.

Figura 1. Extraído de pixabay.com.

# Conjunto de dados

- Em Estatística, em geral, lidamos com **dados estruturados em um formato tabular**.
- Os dados nem sempre começam nessa forma. Muitas vezes a informação deve ser processada e tratada de modo a chegar nesta estrutura.
- O conjunto de dados completo e sem tratamentos é denominado conjunto de **dados brutos**.
- Um conjunto de dados considerado **arrumado** é aquele em que:
  - Cada **coluna** representa uma **variável**.
  - Cada **linha** representa uma **observação**.
  - Cada **célula** representa o **valor** observado.

# Conjunto de dados

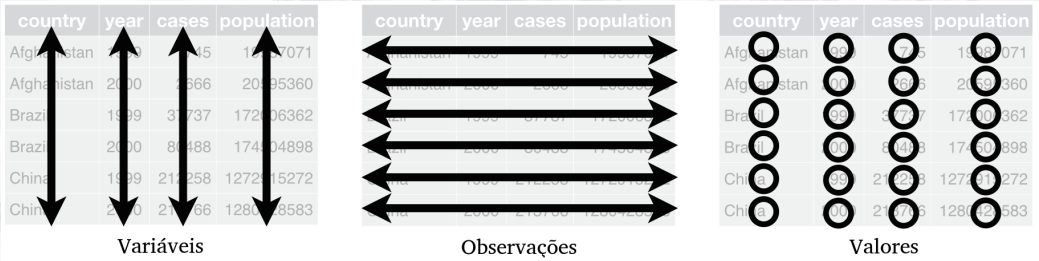


Figura 2. Adaptado de https://r4ds.had.co.nz.

# Conjunto de dados

Tabela 1. Exemplo de conjunto de dados

| ID | Sexo | Escolaridade | Altura | Peso | Irmãos |
|---|---|---|---|---|---|
| 1 | Masculino | Ensino superior | 182 | 80 | 0 |
| 2 | Feminino | Ensino médio | 160 | 46 | 1 |
| 3 | Feminino | Ensino superior | 160 | 55 | 4 |
| 4 | Feminino | Mestrado | 165 | 58 | 3 |
| 5 | Masculino | Ensino médio | 183 | 55 | 1 |

# Fontes de dados

# De onde vêm os dados?

**Alguns exemplos**:

- Estudos de caso.
- Experimentos.
- Pesquisas.
- Registros administrativos.
- Dados em repositórios online.
- Bancos de dados corporativos.
- Sensores.
- Textos, imagens e vídeos.

Figura 3. Extraído de pixabay.com.

# Dados observacionais x dados experimentais

**Dados observacionais**

- **Observação passiva** da realidade.
- Sem modificação das condições.

**Dados experimentais**

- **Intervenção** na realidade.
- Condições controladas.
- Observação dos efeitos das intervenções.

Figura 4. Extraído de pixabay.com.

# Dados observacionais x dados experimentais

- Cada tipo de estudo induz **relações** diferentes entre as observações e **modelos estatísticos** diferentes para modelar a incerteza destas relações.
- Um **conjunto de dados** é um dos subprodutos de um estudo. Ele contém as características principais (variáveis) que se tem interesse em estudar em uma população ou amostra.
- Estas características podem ser **qualitativas** ou **quantitativas** e a partir do conjunto de dados as análises inferenciais são feitas.
- As variáveis são assim chamadas porque seus valores não são constantes e variam segundo regras ou leis naturais que podem ser conhecidas ou desconhecidas.

# Tipos de variáveis

# Tipos de variáveis

- Na prática, podemos coletar variáveis de diferentes tipos e naturezas.
- Antes de de qualquer análise precisamos ser capazes de compreender os tipos de variáveis pois estes tipos conduzirão às análises e métodos estatísticos que poderão ser aplicados.

- Existem dois tipos (básicos) de variáveis:
  - Numéricas (**quantitativas**).
  - Não numéricas (**qualitativas**).


Figura 5. Tipos básicos de variáveis.

# Variáveis quantitativas

- **Variáveis Quantitativas**: assumem valores numéricos.
  - **Discretas**: características mensuráveis que podem assumir apenas um número finito ou infinito **contável** de valores.
  - **Contínuas**: características mensuráveis que assumem valores em uma **escala contínua**, isto é, na reta real.

**Exemplos**

- Altura.
- Peso.
- Idade.
- Percentual de gordura corporal.
- Número de filhos.
- Número de fraturas.
- Número de faltas.
- Número de peças defeituosas em um lote.

# Variáveis qualitativas

- **Variáveis Qualitativas**: são as características definidas por categorias, ou seja, representam uma classificação dos indivíduos e não uma característica numérica.
  - **Nominais**: não existe ordenação nem peso entre as categorias.
  - **Ordinais**: existe uma ordenação entre as categorias.

**Exemplos**

- Estado civil.
- Orientação sexual.
- Turma.
- Posição em que joga em um time.
- Severidade de uma lesão.
- Escolaridade.
- Grau de proficiência em língua inglesa.
- Risco de infarto.

# Cuidados com variáveis

- Existem particularidades na classificação de variáveis devido a situações como:
  - Discretização de variáveis contínuas.
  - Limitações em instrumentos de mensuração.
  - Utilização de quantidades numéricas para representação de variáveis categóricas.
  - Dentre outras.
- Deve-se sempre estar atento a este tipo de situação pois podem levar a implicações nas análises e consequentemente nos resultados.
- Existem outros tipos de variáveis que ocorrem em situações particulares que requerem técnicas específicas de análise.

# Análise de dados

# No que devemos pensar antes de analisar nossos dados?

- O que estamos interessados em avaliar?
- Quais são as variáveis de interesse?
- Quais são as variáveis que queremos avaliar se influenciam a variável de interesse?
- Quais são os métodos disponíveis para análise de variáveis deste tipo?
- Quais os métodos disponíveis que permitem responder nossa pergunta de pesquisa?
- Como coletar os dados?
- Os dados são válidos?

# Métodos de amostragem

# Amostras

- Uma amostra é um **subconjunto da população**.
- Na prática costuma ser inviável trabalhar com a população toda.
- A alternativa então é trabalhar com uma **amostra** e **inferir** os resultados para a população.
- A seleção da amostra pode ser feita de diversas maneiras.

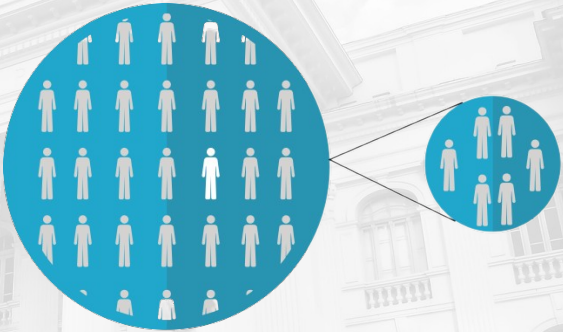

Figura 6. Extraído de pixabay.com.

# Amostras

- Os métodos de amostragem servem para selecionar subconjuntos da população de forma mais **representativa** possível.
- A forma adequada de amostragem conduz a um **menor tamanho amostral** para obtenção de uma **precisão satisfatória**.
- São características desejáveis de uma amostra:
  - Capacidade de generalização.
  - Imparcialidade e representatividade.
  - Capacidade de medir a precisão das estimativas.
- Podemos dividir os métodos em:
  - Amostragem probabilística.
  - Amostragem não probabilística.

# Um caso clássico: a história do Literary Digest

- O **Literary Digest** era uma revista americana de publicação semanal fundada em 1890.
- Em 1936 ocorreu a 38ª **eleição presidencial** dos Estados Unidos.
- Como candidatos haviam nomes como: Franklin Roosevelt, Alf Landon, William Lemke, Norman Thomas, dentre outros.
- **Roosevelt** e **Landon** eram vistos como os favoritos.

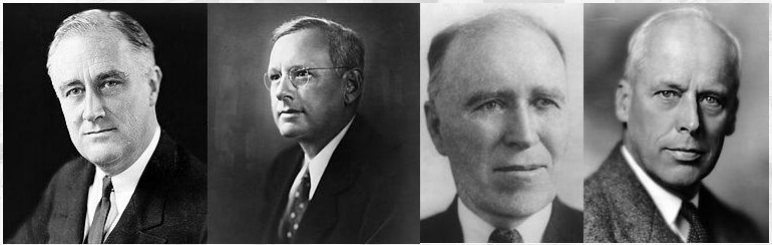

Figura 7. Franklin Roosevelt, Alf Landon, William Lemke e Norman Thomas.

# Um caso clássico: a história do Literary Digest

- No ano da eleição, o Literary Digest conduziu uma pesquisa de intenção de votos com **mais de 10 milhões de respondentes** com base em sua base de assinantes e outras listas de indivíduos.
- Enquanto isso, George Gallup, fundador da Gallup Poll, conduziu pesquisas quinzenais com apenas **2 mil indivíduos**.
- O Literary Digest previu a vitória de Landon, Gallup previu a vitória de Roosevelt. Qual dos dois acertou?

# Um caso clássico: a história do Literary Digest

O resultado da eleição foi:

1. **Franklin D. Roosevelt, 27.752.648 de votos.**
2. Alf Landon, 16.681.862 de votos.
3. William Lemke, 892.378 votos.
4. Norman Thomas, 187.910 votos.
5. Outros, 132.901 votos.

- Gallup acertou, Literary Digest errou.
- O que deu errado na pesquisa do Literary Digest?
  - A resposta é: a **composição da amostra**.

# Um caso clássico: a história do Literary Digest

- O Literary Digest optou por **quantidade**, prestando pouca atenção ao método de seleção.
- A amostra foi de **conveniência** e representava apenas o grupo da população com nível socioeconomico relativamente alto: seus próprios assinantes e pessoas que possuiam luxos da época como telefones.
- Isso gerou um **viés de amostragem**, ou seja, a amostra era diferente, de modos importantes e não aleatórios, da população que deveria representar. Ou simplesmente: **a amostra não era representativa**.
- Por outro lado, a amostra de Gallup era bem mais modesta, contudo o método de seleção gerou uma **amostra representativa da população** em que todas as camadas de votantes estavam presentes.

# Amostragem probabilística

# Amostragem probabilística

- Amostragem probabilística deve ser usada **sempre que possível**.
- O objetivo é dimensionar amostras que sejam capazes de **estimar** as quantidades de interesse com uma certa **precisão** desejada.
- Existem diversos métodos disponíveis.

**Alguns métodos são:**

- Amostragem aleatória simples (com ou sem reposição).
- Amostragem sistemática.
- Amostragem estratificada.
- Amostragem por conglomerados.

# Amostragem aleatória simples (com ou sem reposição)

- Selecionar um conjunto de elementos da população em que **todos os elementos tenham a mesma probabilidade de serem sorteados**.
- Os sorteios de cada elemento são **independentes**.
- Pode ser com ou sem reposição.
  - **Com reposição:** um elemento sorteado pode ser sorteado novamente.
  - **Sem reposição:** um elemento sorteado não pode ser sorteado novamente.
  - A diferença entre com e sem reposição só é relevante se o tamanho da população for pequeno.
- Muitos métodos estatísticos foram desenvolvidos pensando no plano de amostragem aleatória simples.

# Amostragem aleatória simples (com ou sem reposição)

**Exemplo**

- Suponha que uma empresa deseja avaliar se os produtos de determinado lote estão dentro das especificações de qualidade.
- Avaliar todos os produtos é inviável, mas podemos observar uma amostra.
- Retira-se aleatoriamente produtos deste lote para inspeção.

# Amostragem sistemática

- Sorteia **um elemento a cada número de unidades definido** inicialmente (1 a cada 10, 1 a cada 20, etc).
- É necessário definir um **tamanho de intervalo inicial** e selecionar uma **unidade de partida**.
- Todos os elementos dentro do intervalo apresentam a mesma probabilidade de serem sorteados.
- A partir da unidade de partida a próxima sorteada é a da posição correspondente à **inicial mais o tamanho do intervalo** e assim sucessivamente.

# Amostragem sistemática

**Exemplo**

- Suponha que desejamos selecionar uma amostra sistemática de 500 alunos em uma população de 10.000 e que haja um cadastro desses alunos.
- Podemos selecionar 1 a cada 20.
- Primeiro seleciona-se a unidade de partida, um número aleatório entre 1 e 20. Este é o primeiro elemento da amostra. Suponhamos que tenha sido o aluno número 5.
- O próximo elemento é o da posição 5+20, e assim por diante.
- Serão selecionados os alunos 5, 25, 45, 65,...

# Amostragem estratificada

- Usada quando a população de interesse possui algum tipo de **estratificação natural** (por exemplo, cidades possuem bairros).
- Dentro de cada estrato podemos coletar uma amostra.
- A amostra final é composta pela **união das amostras** obtidas em cada estrato.
- O tipo de amostragem dentro de cada estrato pode variar.

# Amostragem estratificada

**Exemplo**

- Suponha que existe interesse em avaliar a proporção de crianças em situações de risco em determinada região.
- Esta região possui 3 bairros: A, B e C.
- Considere que o bairro A possui 60% das crianças, o bairro B possui 30% e o bairro C possui 10%.
- Considerando uma amostra aleatória estratificada de 500 indivíduos, poderiam ser selecionados
  - 300 crianças do bairro A (60% da amostra).
  - 150 do bairro B (30% da amostra).
  - 50 do bairro C (10% da amostra).

# Amostragem por conglomerados

- Parecida com a ideia de amostragem estratificada.
- Os conglomerados são **conjuntos de observações**.
- Inicialmente **sorteamos os conglomerados** (diferente do que acontece na amostragem estratificada).
- A amostra pode ser composta por todos os elementos de todos os conglomerados sorteados.
- Outra alternativa é sortear dentro dos conglomerados amostrados.

# Amostragem por conglomerados

**Exemplo**

- Suponha que o interesse reside em avaliar as notas médias de alunos de uma escola.
- O interesse é obter uma amostra por conglomerados.
- Cada turma é um conglomorado.
- Primeiro sorteiam-se as turmas (conglomerados).
- Dentro de cada conglomerado sorteiam-se as unidades que vão compor a amostra.

# Diferença amostragem estratificada e por conglomerados

- No caso da amostragem estratificada todos os estratos fornecem elementos para a amostra.
- No caso da amostragem por conglomerados não são todos os conglomerados que cedem elementos para a amostra.
    1. Primeiro selecionam-se conglomerados.
    2. Depois selecionam-se as unidades.

# Amostragem não probabilística

# Amostragem não probabilística

- Em muitos casos não é possível fazer uso de métodos de amostragem probabilística.
- Surgem então os métodos de amostragem não probabilística.
- Uma avaliação da “representatividade” dos métodos de amostragem não probabilística não pode ser feita.
- Devemos tomar muito cuidado ao interpretar resultados baseados em métodos de amostragem não probabilísticos.
- Em geral, estas amostras carregam um alto risco de não serem representativas.
- Não há métodos para análise probabilística ou inferencial dos resultados.

# Amostragem não probabilística

Alguns métodos são:

- Amostragem por conveniência.
- Amostragem intencional ou julgamento.
- Amostragem bola de neve.

# Amostragem por conveniência

- Os elementos da amostra não são obtidos por meio de sorteio, mas sim de acordo com sua **disponibilidade**.

**Exemplo**

- Suponha que um pesquisador trabalha com animais criados em cativeiro.
- Não existe qualquer cadastro da população alvo.
- Por isso, o pesquisador avalia os animais disponíveis.

# Amostragem intencional ou julgamento

- Um especialista (expert) no problema **escolhe os elementos** que julga representativos para compor a amostra.

**Exemplo**

- Suponha um problema congênito que só pode ser identificado por um especialista altamente treinado.
- Para isso um conjunto de indivíduos é selecionado e deste conjunto o especialista seleciona para a amostra aqueles em que ele identifica o problema congênito.

# Amostragem bola de neve

- Identifica-se algumas unidades e estas **unidades indicam novas unidades** para compor a amostra.

**Exemplo**

- Suponha que um aluno criou um formulário para obter dados para seu trabalho de conclusão de curso.
- Não existe um cadastro para a população alvo.
- Por isso, o aluno repassa o formulário para indivíduos que ele sabe que fazem parte da população alvo e pede que estes indivíduos indiquem outros possíveis respondentes.

**O que foi visto:**

- Dados.
- Tipos de variáveis.
- Fontes de dados.
- Estudos observacionais e experimentais.
- Amostras.
- Métodos de amostragem.

**Próximos assuntos:**

- Introdução à análise exploratória.
- Análise exploratória univariada para variáveis qualitativas.
- Análise exploratória univariada para variáveis quantitativas.