**1.1 Definições**

* **Dados:** consistem em informações provenientes de observações, contagens, medições ou respostas
* **Conjunto de Dados:**  são as informações obtidas através de alguma pesquisa
* **Estatística:** ciência que trata de coleta, organização, análise e interpretação dos dados para a tomada de decisões
* **Sobre conjunto de dados**
  + **População:** é a coleção de todos os resultados, respostas, medições ou contagens que são de interesse
  + **Amostra:** é um subconjunto ou parte de uma população
* **Parâmetro:** é a descrição numérica de uma característica *populacional*
* **Estatística:** é a descrição numérica de uma característica *amostral*
* **Sobre Ramos da Estatística**
  + **Estatística Descritiva:** é o ramo da estatística que envolve a organização, o resumo e a representação dos dados
  + **Estatística inferencial:** é o ramo de estatística que envolve o uso de uma amostra para chegar a uma conclusão sobre uma população. Uma ferramenta básica no estudo da estatística inferencial é a probabilidade.
* **EXEMPLO:**
  + Realizaram uma pesquisa com 350 donos(as) de mercadinhos para saber se “Exibir doces no caixa aumenta sua venda”. 118 donos(as) responderam que “sim”.
    - **População:** todos os proprietários de mercadinhos do Brasil
    - **Amostra:** os 350 proprietários de mercadinhos
    - **Conjunto de dados:** As respostas sim (118 = 33,7%) e não (232 - 66,3%)

**1.2 Classificação dos dados**

* **Tipos de dados:**
  + **Dados qualitativos:** atributos, rótulos ou entradas não numéricas (às vezes pode ser um dado numérico que é classificado como qualitativo)
  + **Dados quantitativos:** medidas numéricas ou contagens
* **Níveis de mensuração:**
  + **Nominal:** são apenas qualitativos. Dados nesse nível são categorizados usando-se nomes, rótulos ou qualidades. Não é possível realizar cálculos matemáticos nesse nível
  + **Ordinal:** são qualitativos ou quantitativos. Dados nesse nível podem ser postos em ordem ou classificados, mas as diferenças entre as entradas de dados não têm sentido matemático
  + **Intervalar:** podem ser ordenados e é possível calcular diferenças que tenham sentido matemático entre as entradas de dados. No nível intervalar, um registro zero simplesmente representa uma posição em uma escala; a entrada não é um zero natural.
  + **Razão:** são similares aos dados no nível intervalar, com a propriedade adicional de que, nesse nível, um registro zero é um zero natural. Uma razão de dois valores pode ser formada de modo que um dado possa ser expresso significativamente como múltiplo de outro.

**1.3 Coleta de dados e planejamento de experimentos**

* **Planejando um estudo estatístico**
  + Identifique as variáveis de interesse (o foco) e a população do estudo
  + Desenvolva um plano detalhado para a coleta de dados. Se usar uma amostra, certifique-se de que a amostra é representativa da população
  + Colete os dados
  + Descreva os dados usando técnicas de estatística descritiva
  + Interprete os dados e torne as decisões sobre a população usando estatística inferencial
  + Identifique quaisquer erros possíveis
* **Tipos de estudo:**
  + **Observacional:** o pesquisador não influencia as respostas. Uma pesquisa de votos é uma pesquisa observacional;
  + **Experimental:** há a influência das respostas pelo pesquisador. Um exemplo é a pesquisa de medicamentos, onde tem o grupo de tratamento (os que recebem o medicamento) e o grupo de controle (os que recebem placebo);
* **Tipos de coletas de dados**
  + **Simulação:** uso de um modelo matemático ou físico para reproduzir as condições de uma situação ou processo. Um modelo de uma explosão de bomba nuclear, modelo de impacto de um boneco ou modelo de como um avião poderia voar são exemplos de simulação (ou até um jogo, dependendo do quão “real” ele seja
  + **Pesquisa:** Investigação de uma ou mais características de uma população. Pesquisas de gostos, em quem cada um vai votar, qual é a cor de pele da população de um determinado lugar, qual é o tipo de consumo e afins podem ser classificados como “pesquisa”.
* **Elementos-chave de experimentos bem realizados:**
  + **Controle:** conseguir controlar o máximo possível as variáveis do estudo. Algumas coisas que podem prejudicar os estudos nesta fase:
    - **Variável de confusão:** ocorre quando um pesquisador não pode distinguir um ou mais fatores que causaram os efeitos provocados sobre a variável em estudo, gerando confusão
      * Você faz uma campanha para diminuir o consumo de água no condomínio e na mesma semana há o racionamento de água no bairro. Não dá para atribuir a diminuição do consumo de água SOMENTE com a campanha de conscientização, pois podemos confundir os resultados da campanha com o racionamento de água
    - **Efeito placebo:** quando há resposta ao tratamento para quem está recebendo o placebo
      * Um medicamento de emagrecimento é dado N pessoas, sendo que metade recebe o medicamento e metade recebe placebo. Pode ser que um ou vários indivíduos comecem a emagrecer e o pesquisador tenha um viés POR SABER QUE AQUELE SER ESTÁ RECEBENDO O PLACEBO.
      * Para resolver este problema é conduzido um experimento **duplo-cego,** onde nem o pesquisador, nem o pesquisado, sabem quem está recebendo placebo ou não. Os pesquisadores só ficam sabendo ao final do experimento.
    - **Efeito Hawthorne:** ocorre em um experimento quando os indivíduos mudam seu comportamento simplesmente porque sabem que estão participando de um experimento
  + **Aleatorização:** é o processo de se designar indivíduos aleatoriamente para diferentes grupos de tratamento
    - **Planejamento completamente aleatorizado:** os indivíduos são designados para diferentes grupos de tratamento (controle ou tratamento) por meio de seleção aleatória;
    - **Planejamento em blocos aleatorizados:** o pesquisador separa os indivíduos com características similares em blocos e, então, dentro de cada bloco, designa-os aleatoriamente para os grupos.
      * Este tipo de planejamento ajuda a diminuir a variável de confusão no experimento
    - **Planejamento de pares combinados:** os indivíduos são colocados em pares de acordo com sua similaridade.Um indivíduo em cada par é selecionado aleatoriamente para receber um tratamento enquanto o outro indivíduo recebe um tratamento diferente.
  + **Replicação:** é a repetição de um experimento sob condições iguais ou semelhantes
    - Em machine learning usa-se o random com um seed (uma variável com um determinado valor), assim a variável “aleatória” sempre voltará o mesmo valor.
  + **Técnicas de amostragem:**
    - **Censo:** contagem ou medição de toda a população
    - **Amostragem:** contagem ou medição de parte da população
  + **Erros de amostragem**
    - **Amostragem enviesada:** quando a amostragem não é representativa da população, seja esse viés proposital ou não.
      * **Exemplo:** o Lula fazer uma enquete no perfil dele do Twitter para quem as pessoas votariam para presidente; uma pesquisa de opinião em um lugar onde as pessoas tendem a uma determinada resposta
  + **Tipos de amostragens aleatórias:**
    - **Amostragem aleatória simples:** aquela na qual cada amostra possível de mesmo tamanho tem a mesma chance de ser selecionada.
      * **Exemplo:** para contar todas as pessoas que moram na cidade de Cafelândia/SP, poderíamos designar um número para cada casa da cidade, usar um programa de números aleatórios, gerar uma amostra de números, selecionar as casas de tais números, contar a quantidade de pessoas que moram dentro delas e estimar o resto