

# Capítulo 2

## Princípios gerais da Estatística

### 2.1. Histórico

Desde a antiguidade a Estatística está presente na vida do homem. As necessidades que exigiam o conhecimento numérico surgiram anos antes de Cristo, por exemplo, com o recenseamento das populações. Os registros egípcios de presos de guerra em 5000 a.C. foram os primeiros dados estatísticos disponíveis. Outro momento importante na história da estatística ocorreu em 3000 a.C., quando os egípcios registraram a falta de mão-de-obra relacionada à construção de pirâmides.

Em 2238 a.C., foi realizado o primeiro recenseamento na China com fins agrícolas e comerciais. Em 600 a.C., todos os egípcios tinham que declarar suas profissões e fontes de rendas ao governo de suas províncias, com a penalidade de morte para quem não o fizesse.

No Brasil, a formação da estatística está ligada a alguns fatos importantes como:

- a) 1º censo geral da população brasileira em 1872;
- b) Criação do Instituto Brasileiro de Geografia e Estatística (IBGE) em 1936;
- c) Início do ensino de Estatística nas escolas: Escola Nacional de Ciências Estatísticas (ENCE – Rio de Janeiro) e Escola de Estatística da Bahia, em 1953;
- d) Inclusão da Estatística no Ensino Fundamental e Médio em 1997.

Hoje em dia a estatística é uma ferramenta indispensável para todo profissional que analisa informações diariamente para tomar decisões, pois seu campo de aplicação se estende a várias áreas do conhecimento humano. Porém, devido à facilidade de alguns aplicativos estatísticos, muitos pesquisadores se sentem aptos a fazerem análises estatísticas sem um conhecimento mais profundo de conceitos e teorias, o que muitas vezes gera interpretações equivocadas ou até errôneas.

A literatura médica está repleta de artigos nos quais técnicas estatísticas são usadas e para avaliar criticamente as evidências numéricas apresentadas, deve-se ter algum conhecimento de estatística.

### 2.2. Objetivos da estatística

A estatística é um conjunto de técnicas que permite coletar, organizar, descrever, analisar e interpretar dados, numéricos ou não, provenientes de estudos ou experimentos realizados em qualquer área do conhecimento.

Antes da coleta dos dados, deve ser feito um planejamento estatístico com a finalidade de garantir que as hipóteses de interesse do pesquisador possam ser testadas. O planejamento estatístico é necessário, pois o que se procura são dados confiáveis que confirmem ou não as hipóteses formuladas pelo pesquisador, relacionadas aos objetivos específicos da pesquisa.

A verificação das hipóteses se realiza com a aplicação de métodos de análise estatística. A aplicação desses métodos depende da maneira como os dados foram coletados. O planejamento

do experimento e a análise estatística dos seus resultados estão associados e devem ser aplicados em uma seqüência lógica na pesquisa científica, nas diversas áreas do conhecimento.

O planejamento de uma pesquisa consiste de diversas etapas, que dependem de um bom entendimento entre o pesquisador e o estatístico planejador.

Os seguintes passos fazem parte de um planejamento estatístico:

- 1) Definição da coleta dos dados: como coletar amostras representativas da população de interesse e cálculo do tamanho da amostra;
- 2) Definição das variáveis a serem analisadas;
- 3) Definição das medidas descritivas, tabelas e gráficos;
- 4) Definição da análise inferencial: diferença entre grupos; diferença entre tempos; associação entre variáveis (correlações); relações temporais (sobrevida); acurácia de um teste diagnóstico; modelos de previsão.

### 2.2.1 Coleta dos dados

Para a coleta dos dados é necessário definir qual é a população de interesse e quais informações (variáveis) serão obtidas de cada um dos elementos.

População é o grande conjunto de dados que contém a(s) característica(s) de interesse em um determinado estudo. O termo população refere-se não somente a uma coleção de indivíduos, mas também ao alvo (objeto) sobre o qual reside nosso interesse.

Amostra é qualquer subconjunto extraído da população, em geral com tamanho sensivelmente menor, isto é, são alguns elementos representativos da população.

A seleção dos elementos que irão compor a amostra pode ser feita de várias maneiras e dependerá do conhecimento que se tem da população e da quantidade de recursos disponíveis. Nesse passo, tenta-se fornecer um subconjunto de elementos o mais parecido possível com a população que lhe dá origem.

Existem dois métodos de amostragem:

a) *Probabilísticos*, onde cada elemento da população possui a mesma probabilidade de ser selecionado. Os métodos de amostragem probabilísticos envolvem as técnicas de amostragem: casual ou aleatória simples, estratificada, sistemática e por conglomerados. Na *amostragem casual* (aleatória simples), os elementos da amostra são selecionados ao acaso e todos têm a mesma chance de serem escolhidos.

Na *amostragem estratificada*, a população de interesse apresenta alguma característica (estratificação) conhecida a priori, que será utilizada para dividi-la em estratos e de cada estrato será selecionada uma amostra aleatória.

Na *amostragem sistemática*, os elementos são escolhidos na mesma ordem, por exemplo, a cada 10 ou a cada 20 ou a cada  $k$  elementos.

Na *amostragem por conglomerados*, a área populacional é dividida em seções que são selecionadas aleatoriamente e todos os elementos das seções selecionadas fazem parte da amostra.

b) *Não probabilísticos*, onde há escolha deliberada dos elementos da amostra. Esse método não permite a generalização dos resultados da amostra para a população e envolve técnicas de amostragem de conveniência (acidental ou intencional), onde os elementos selecionados são de fácil acesso.

Uma vez selecionada a amostra, as características de interesse (variáveis) são obtidas para confirmar ou não as hipóteses formuladas pelo pesquisador.

*Variáveis* são características observadas, medidas ou contadas nos elementos da amostra, como por exemplo, sexo, estado civil, peso, altura, etc, e que podem variar, isto é, assumir um valor diferente para cada elemento da amostra. Uma variável pode ser quantitativa (discreta ou contínua) ou qualitativa (nominal ou ordinal).

*Variável quantitativa* é uma variável numérica. *Variável quantitativa discreta* pode ser vista como resultante de contagens, assumindo assim, em geral, valores inteiros. Exemplos: número de filhos, número de dias de internação, etc. *Variável quantitativa contínua* assume valores em intervalos dos números reais e, geralmente, é proveniente de uma mensuração. Exemplos: peso, altura, etc.

Um subtipo especial de variável quantitativa contínua, e de muita utilidade em oncologia, é a denominada *variável data* ou tempo até ocorrência de um evento de interesse. A expressão gráfica desta variável é apresentada, nos artigos científicos, sob a forma de uma curva de sobrevida. Há duas maneiras de se coletar dados para a caracterização desta variável: pode-se calcular diretamente, para cada indivíduo, o intervalo de tempo de interesse e expressá-lo por um valor numérico contínuo (por exemplo, 33 meses), ou pode-se assinalar as datas de início e final do intervalo de interesse, deixando que o software estatístico escolhido processe esta informação da forma solicitada. Esta é a forma preferida, hoje em dia, devido às facilidades de computação.

*Variável qualitativa* é uma variável não numérica cujos resultados possíveis representam atributos e/ou qualidades. *Variável qualitativa ordinal* segue certa ordenação, indicando intensidades crescentes de realização. Exemplo: Tamanho (pequeno, médio, grande), classe social (baixa, média, alta), nível de instrução (ensino fundamental, ensino médio, ensino superior), etc. Variável qualitativa nominal não apresenta nenhuma ordenação nas possíveis realizações. Exemplo: sexo, estado civil, fumante (sim, não), etc.

Uma vez coletados os dados de todas as variáveis envolvidas num estudo, o próximo passo é organizar esses dados para apresentá-los.

### **2.2.2. Organização dos dados**

Geralmente, os dados das variáveis são coletados através de entrevistas realizadas com os pacientes ou de prontuários dos pacientes arquivados em hospitais. As informações dos pacientes (sexo, idade, escolaridade, resultados de exames laboratoriais, etc.) ficam em fichas, as quais devem ser armazenadas para que possam ser analisadas.

O armazenamento dos dados deve ser feito através de um banco de dados (tabela com os dados brutos, isto é, de cada unidade de observação), que poderá ser lido ou exportado em diferentes softwares (estatísticos ou não), como Excel, SAS, SPSS, etc. No banco em Excel, a primeira linha da planilha deve conter os nomes das variáveis. Cada coluna da planilha está associada a uma característica (variável). Cada linha da planilha está associada a uma unidade de observação (por exemplo, paciente). Não se deve colocar a informação de duas variáveis na mesma coluna (por exemplo, tabagismo: fuma 2 maços de cigarro por dia). Nesse caso, uma coluna deve conter informação sobre tabagismo (sim ou não) e, no caso de sim, outra coluna deve conter a informação número de cigarros diários ou semanais. Deve-se, também, padronizar as unidades de medida. Segue abaixo um exemplo de ficha clínica e de banco de dados.

**Sexo:**  Mas.       Fem.  
**Idade:** 35 anos  
**Escolaridade:**  Ensino Fundamental       Ensino Médio       Ensino Superior  
**Peso:** 87 kg  
**Altura:** 1,78 metros  
**Pressão arterial:** 15 / 12 mmHg  
**Hipertensão arterial:**  Não     Sim      Tempo de diagnóstico (ano) \_\_\_\_\_

**FIGURA 2.1 - Exemplo de ficha para um paciente**

Para facilitar o armazenamento das informações de um estudo, na maioria das vezes é necessário codificar os níveis das variáveis categorizadas, pois texto livre está associado a alguns inconvenientes, tais como:

- a) Dificuldade de leitura de diferentes caligrafias;
- b) Dificuldade de armazenamento em meio eletrônico;
- c) Falta de uniformidade (melhorou um pouco; melhorou, mas ainda sente dor).

A codificação deve contemplar os seguintes itens:

- I) Todas as possíveis categorias devem ser incluídas.
- II) Nos casos envolvendo muitas categorias, algumas podem ser agrupadas dependendo de sua utilização no estudo.
- III) Em muitas situações uma categoria rotulada “Outros” pode ser incluída. Em outras situações categorias rotuladas: “Não aplicável” ou “Não informado” também podem ser incluídas.
- IV) Cada unidade de investigação deve ser classificada em uma única categoria. Situações em que se deseja classificar algumas unidades de investigação em mais do que uma categoria podem ser contornadas redefinindo a variável, criando uma categoria rotulada “Duas ou mais” ou gerando uma nova variável dicotômica (com níveis Sim e Não) para cada nível da variável original.
- V) Sempre que possível, as categorias devem ter rótulos (M = Masculino, F = Feminino ou numéricos 1 = Masculino, 2 = Feminino).
- VI) Os códigos devem obedecer à ordenação natural das categorias (1 = ruim, 2 = razoável, 3 = bom, 4 = excelente).

Uma vez que os dados foram tabulados, o próximo passo é realizar uma análise descritiva para conhecer as informações dos mesmos, o que será estudado no próximo capítulo.

No entanto, é importante ressaltar que a análise exploratória dos dados constitui uma fase preliminar na análise, destinada a obter informações que indiquem possíveis modelos a serem utilizados numa fase final, que seria chamada inferência estatística (ou análise confirmatória de dados)

### **2.2.3 Análise confirmatória e Interpretação dos dados**

Nesta etapa, o maior interesse do pesquisador é tirar conclusões que o auxiliem na resolução do problema. É possível arriscar algumas generalizações, as quais envolverão algum grau de incerteza, pois o pesquisador pode não estar seguro de que o observado naquela amostra reflita o comportamento da população. Para atingir esse objetivo, o pesquisador utiliza os métodos de inferência estatística.

Com a análise e interpretação dos dados, o pesquisador tem subsídios para a tomada de decisões diante de incertezas e para as justificativas científicas dessas decisões.

## 2.3 Tipos de estatística

A Estatística subdivide-se em duas áreas: estatística descritiva e estatística inferencial.

A estatística descritiva é, em geral, utilizada na etapa inicial da análise, objetivando tirar conclusões de modo informal e direto a respeito de características de interesse. Em outras palavras, pode ser definida como um conjunto de técnicas destinadas a descrever e resumir os dados para tirar conclusões a respeito de características de interesse sem pretender conclusões de caráter extensivo.

A estatística inferencial ou inferência estatística é o estudo de técnicas que possibilitam a extrapolação das informações e conclusões obtidas a partir da amostra para a população.

O uso de informações da amostra para concluir sobre o todo faz parte da atividade diária da maioria das pessoas. Basta observar como uma cozinheira verifica se o prato que ela está preparando tem ou não a quantidade adequada de sal. Essa decisão é baseada em procedimento amostral.

A inferência estatística é a conceituação formal desse princípio intuitivo do dia-a-dia, para que possa ser utilizado cientificamente em situações mais complexas.

Um tópico muito utilizado na área médica para tomada de decisões é o Teste de Hipóteses que será estudado no capítulo 4.

## Referências

1. Bussab WO, Morettin PA. Estatística Básica. 5ed. São Paulo :Saraiva, 2003.
2. Fonseca JS, Martins GA. Curso de Estatística. 3ed. São Paulo:Atlas, 1982.
3. Pagano M, Gauvreau K. Princípios de Bioestatística. São Paulo: Pioneira Thomson Learning, 2004.
4. Triola MF. Introdução à Estatística. 10ed. Rio de Janeiro :LTC, 2008.