Análise exploratória de dados

O objetivo desta nota de aula é apresentar para vocês alguns conceitos básicos que são fundamentais e muito úteis para se realizar uma análise estatística.
A ideia é já começarmos a raciocinar estatisticamente sobre dados, buscando discutir o comportamento das variáveis estudadas tendo em vista gerar conhecimento útil.
Procedimentos de amostragem adequados são essenciais para se realizar uma coleta de dados de forma criteriosa e não tendenciosa. Uma vez coletados os dados, o foco está em como eles serão analisados. Para isso, utilizaremos uma análise quantitativa baseada em medidas numéricas e gráficos básicos que irão resumir o comportamento das variáveis estudadas. Também possibilitarão identificar padrões de comportamento dessas variáveis e valores atípicos.
Fique atento com os elementos do pensamento estatístico e com a clareza sobre a questão estatística motivadora da análise para que sua análise exploratória seja realmente proveitosa, e as respostas ajudem de alguma forma no processo de tomada de decisão e de geração de conhecimento.
Na sequência, serão abordadas as medidas de posição, medidas de dispersão, gráficos básicos, medidas de correlação e associação entre variáveis, formas de apresentação dessas medidas, e também o uso de softwares estatísticos.

O objetivo da* análise exploratória de dados** (AED) ***é a exploração irrestrita dos dados, na busca por padrões interessantes. As conclusões se aplicam somente aos indivíduos e circunstâncias para os quais dispomos de dados. As conclusões são informais, baseadas no que interpretamos dos dados.
Para tal, utilizam-se métodos numéricos e gráficos para descrever as variáveis num conjunto de dados e as relações entre elas. Busca-se detectar padrões, resumir a informação contida nos dados e apresentar os resultados de modo conveniente.

Requisitos importantes:

- Uma visão clara de como os dados foram coletados (população e amostra);
- A identificação dos tipos de variáveis (qualitativas ou quantitativas)
- Uma coerência da análise com os objetivos do trabalho;
- A representação gráfica das variáveis quantitativas e qualitativas;
- A análise de medidas de posição, dispersão, gráficos e de possíveis relações entre as variáveis;
- A identificação de anomalias e situações atípicas

A AED Serve de embasamento para a **inferência estatística**, que por sua vez objetiva responder questões específicas, apresentadas antes dos dados serem obtidos e estudar uma população por meio de evidências decorrentes de uma avaliação minuciosa dos dados amostrais.

Descrição de dados por meio de medidas numéricas

Para formalizar os conceitos abaixo, vamos supor que $x_1 x_2, x_3 \ldots, x_n$ são *n* valores obtidos da observação de uma variável quantitativa *X*.

Medidas de posição: representantes do conjunto de dados

- **Média ($\bar{x}$):** é a média aritmética, ou seja, é a soma das observações dividida pelo número total das mesmas.
- **Mediana (M_d):** é o ponto do meio de uma distribuição, considerando a série de observações ordenada de forma crescente. O número tal que metade das observações são menor do que ele e metade maior. Se o número de observações n for ímpar, a mediana M_d será a observação central na lista ordenada. Se o número n de observações for par, a mediana M_d será a média das duas observações centrais na lista ordenada.
- Você sempre pode localizar a mediana na lista ordenada das observações contando até (n + 1)/2 observações a partir do menor valor da lista.
- **Moda (M_o):** é o dado mais frequente observado em um conjunto de dados. Em uma distribuição, pode haver mais de uma moda.

Medidas de dispersão: sumarizam a variabilidade do conjunto de dados

- **Variância** (s^2): é uma média dos quadrados dos desvios das observações a partir de sua média. $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- **Desvio médio** (dm): é uma média dos desvios das observações em relação à média em valor absoluto. $dm = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$
- **Desvio padrão** (s): é a raiz quadrada da variância
- **Amplitude** ou intervalo: diferença entre os valores máximo ($x_{(n)}$) e mínimo ($x_{(1)}$) do conjunto de dados. $a = x_{(n)} - x_{(1)}$
- **Coeficiente de variação** (*cv*): é a razão entre o desvio padrão (*s*) e a média $\bar{x}$. É uma medida de dispersão relativa. $cv = {}^{S}/_{\bar{x}}$

Quartis:

- Quartis: delimitam a metade central do conjunto de dados. O primeiro quartil (Q_1) é o ponto central entre o mínimo e a mediana. O terceiro quartil (Q_3) é o ponto central entre a mediana e o máximo.
- Intervalo interquartílico: é a diferença entre o terceiro e o primeiro quartis, ou seja, $IQ = Q_3 - Q_1$. Resume a distribuição focando na metade central dos dados.

Resumo de cinco números:

- O resumo de cinco números oferece uma descrição razoavelmente completa de centro e dispersão e serve para construção de um gráfico, o boxplot. Os cinco números são: Mínimo; Q_1, M_d, Q_3, Máximo.

Representação gráfica de variáveis quantitativas

Objetivo: conhecer a forma da distribuição dos dados da variável analisada. A representação gráfica auxilia a interpretar melhor as medidas de posição e dispersão associadas à variável analisada, a simetria e os valores atípicos que se afastam do corpo da distribuição.

- Ramo e folhas: diagrama construído para dar ideia da forma da distribuição
- Dotplot (gráfico de dispersão unidimensional): visão gráfica da frequência com que os valores se repetem
- Histograma: gráfico que auxilia na descrição da distribuição dos dados de uma variável quantitativa e que é baseado numa tabela de frequências. É o principal gráfico para estudar a forma da distribuição de uma variável quantitativa.
- Boxplot (desenho esquemático ou gráfico de caixa): representação gráfica do esquema de cinco números.

Representação gráfica de variáveis qualitativas

- Gráfico de colunas
- Gráfico de barras
- Gráfico de setores