Estatística

J. P. S. de Sousa

Contents

Part I

Análise Exploratória

Chapter 1

Introdução

A estatítica é um ramo da matemática aplicada e da metodologia científica que busca extrair, reduzir, analisar e modelar um conjunto de dados que amostram uma população. Para isso, são usadas principalmente as ferramentas teóricas providas pela Teoria das Probabilidades, pois entende-se que os dados originam-se de fenômenos aleatórios, e, uma vez modelados, geralmente dois objetivos visam ser cumpridos: estimar ou predizer o comportamento dos dados no futuro, ou realizar inferência, que consiste em detectar os padrões embutidos. A inferência pode ser feita por duas abordagens: a inferência dedutiva, que aceita determinadas premissas para chegar às conclusões, ou a inferência indutiva, que parte de casos particulares para generalizar.

Ainda sobre a inferência estatística, quando tenta-se modelar um determinado conjunto de dados, tentamos identificar padrões e tendências de comportamento através de modelos estatísticos. Supondo que temos um conjunto de dados D, e indentificamos que conseguimos ajustar a eles um modelo M, teremos então que

$$D = M + R$$

onde R descreve a parte aleatória dos dados que não pode ser capturada pelo modelo. Normalmente, busca-se que R deve conter nenhuma suavidade, pois isso implicaria em padrões que o modelo falhou em capturar, e mais suavização por parte de M será necessária.

Chapter 2

Resumo de Dados

2.1 Tipo de Variáveis

Ao trabalhar com dados, geralmente lidamos com dois grandes grupos de variáveis: as quantitativas, que representam uma medida ou contagem, e as qualitativas, que representam uma qualidade ou categoria de um indivíduo ou objeto. As primeiras ainda se subdividem em contínuas – a exemplo de salário, altura e peso – e discretas, como número de filhos ou idade. Já as qualitativas podem ser nominanais, que representam categorias, ordinais, cujos valores podem ser ordenados, ou dicotômicas, que só podem ter duas realizações: sucesso ou fracasso.

2.2 Distribuição de Frequências

Para variáveis qualitativas, uma forma de resumir seus dados é através de tabelas de frequência, que geralmente constam dois atributos: a frequência absoluta de cada classe, e a proporção de cada classe em relação ao total de observações. Para variáveis quantitativas, visando a geração de tabelas de frequência, é geralmente necessária agrupar os valores em intervalos ou faixas. A construção desses intervalos pode ser feita de diversas formas, a depender do tipo de pesquisa sendo realizado. Contudo, recomenda-se a construção de 5 a 15 grupos de mesma amplitude. Poucos grupos podem acabar omitindo informações importantes, enquanto muitos grupos podem dificultar a análise dos dados.

2.3 Escalas de Medição

Alternativamente às classificações apresentadas, podemos tipificar variáveis através de escalas de medição de seus valores. As classes são similares às classificações apresentadas anteriormente, sendo elas:

- Escala Nominal – para dados nessa escala, só podemos dizer que seus valores são diferentes de outros, sendo usada para categorizar indivíduos. Um exemplo simples é o sexo de uma pessoa: feminino ou masculino. Essa escala não suporta operações matemáticas, e uma medida de centralidade comum para resumir dados nessa escala é a moda.
- Escala Ordinal – nessa escala, os valores podem ser ordenados, e podemos então dizer que, além de diferentes, um valor é maior do que o outro. Essa escala tem sua estrutura preservada por operações que preservem a ordem. Um exemplo de variável ordinal é o nível de escolaridade: fundamental, médio e superior. Medidas de tendência central comuns para dados nessa escala são a moda e a mediana.
- Escala Intervalar - essa escala possui uma origem arbitrária e necessita de uma unidade de medida, sendo um exemplo a temperatura de um ambiente. Podemos afirmar que valores são diferentes, maiores e quão maior em relação a outro, e transformações afim – do tipo $ax + b$ – não alteram a estrutura dessa escala. Podemos utilizar medidas como média, moda e mediana.
- Escala Razão – nessa escala, existe uma origem absoluta, e podemos dizer que um valor é o quão maior do que outro através de razões. Um exemplo de variável nessa escala é o peso de um indivíduo. Transformações do tipo ax preservam a estrutura dessa escala, e podemos utilizar medidas como média, moda e mediana.

Chapter 3

Medidas Resumo

3.1 Medidas de Posição

Além das técnicas de tabelas de frequência ou métodos gŕaficos para resumir informações, há formas de resumir ainda mais os dados. Geralmente, as mais comuns são medidas de centralidade: moda, média e mediana.

A moda é mede qual das realizações de uma variável X é a mais frequente, isto é, aquela que maximiza a função de probabilidade da variável.

$$\text{moda}\,(X) = \arg\max_x P(X \leq x)$$

Para obter a moda amostral – ou empírica – de uma distribuição, basta contarmos o valor que mais frequente na amostra.

A média populacional de uma variável aleatória é o seu valor esperado $E[X]$, e, dada uma amostra, a média amostral que estima a populacional é dada por:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

onde n é o número de amostras obtidas. Por fim, a mediana de uma variável aleatória é o valor x tal que o quantil de probabilidade seja $1/2$, ou seja,

$$P(X \leq x) = 1/2$$

Para estimar a mediana de X, suponha que $\vec{v} = (x_1, \ldots, x_n)$ seja um vetor de realizações de X, e $k = \lfloor n/2 \rfloor$, então a mediana q será dada por:

1. Se n é impar, então faça a mediana igual a x_k.

2. Se n é par, então a mediana é dada pore

$$\frac{x_k + x_{k+1}}{2}$$

Repare que, no caso de n par, deve-se ter cuidado com variáveis não contínuas.

3.2 Medidas de Dispersão

Além da análise de centralidade e de valores típicos, pode ser que deseja-se estudar a variabilidade de um conjunto de dados. Com isso, introduziremos algumas medidas de dispersão. O primeiro deles será o desvio médio, dado como

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

O segundo é a variância, que se assemelha ao desvio-médio, mas com a exceção de que os desvios são elevados ao quadrado para se acentuar os desvios maiores.

$$\mathrm{Var}\,(X) = \sum_{i=1}^{n} \frac{|x_i - \bar{x}|^2}{n-1}$$

A variância gera um valor cuja a unidade está elevada ao quadrado, sendo menos interpretável. Por isso, utiliza-se também o desvio-padrão, dado como

$$\mathrm{std}\,(X) = \sqrt{\sum_{i=1}^{n} \frac{|x_i - \bar{x}|^2}{n-1}}$$

Com base no que foi dito sobre o quadrado dos desvios, a variância e o desvio-padrão são mais sensíveis a outliers, e a amplitude dos dados de modo geral, sendo adequadas para quando a distribuição dos dados é aproximadamente normal. O desvio-médio é uma medida menos sensível.

3.3 Quantis

O quantil q de probabilidade p é um número que satisfaz

$$\mathrm{P}(X \leq q) = p$$

por exemplo: a mediana é o quantil de probabilidade de 50%. Os quantis de probabilidade de 0.25, 0.5 e 0.75 são, respectivamente, chamados de quartis, sendo os mais utilizados em resumos de dados.

A estimação de um quantil, quando a função de distribuição não está disponível, pode ser calculado utilizando-se da função de distribuição empírica da variável. Para isso, computamos a probabilidade empírica acumulada para cada realização da variável, com os valores ordenados em grandeza, e usamos uma interpolação linear para obter o valor do quantil q de probabilidade p. O método está implementado no Algoritmo 1.

Algoritmo 1: Método de Interpolação do Quantil Empírico

Entrada: uma amostra $\vec{x} = (x_1, \ldots, x_n)$ e um $p \in [0, 1]$
Saída: O quantil empírico q

início
  $\vec{s} \leftarrow$ ordene($\vec{x}$);
  para cada i *de 1 até* n **faça**
    $p_i \leftarrow \frac{i-0.5}{n}$;
  fim
  se $\exists i \in [n]\ (p_i = p)$ **então**
    retorna s_i;
  fim
  senão
    Encontre i tal que $p_i < p < p_{i+1}$;
    $f_i \leftarrow \frac{p-p_i}{p_{i+1}-p_i}$;
    $q \leftarrow (1 - f_i)s_i + f_i \cdot s_{i+1}$;
    retorna q;
  fim
fim

3.4 Transformações

São conhecidos diversos procedimentos para se aplicar a dados com distribuição normal ou que sejam aproximadamente simétricos. No entanto, os dados do mundo real possuem, em grande parte, uma forma assimétrica. A fim de possibilitar a aplicação das técnicas para distribuições simétricas, podemos utilizar transformações sobre os valores dos dados, de modo a consertar a sua assimetria. A família de transformações mais comuns para um variável X são

$$T(X, p) = \begin{cases} x^p, & \text{se } p > 0 \\ \log x, & \text{se } p = 0 \\ -x^p, & \text{se } p < 0 \end{cases}$$

onde p é um valor empírico geralmente adotado na sequência

$$-3,\ -2,\ -1,\ -1/2,\ -1/3,\ -1,4,\ 0,\ 1/4,\ 1/3,\ 1/2,\ 1,\ 2,\ 3.$$

Quanto maior for a assimetria à direita da distribuição, menor deverá ser o valor de p a fim de corrigi-la. Para assimetrias à esquerda, valores negativos de $p > 1$ provocam uma compressão de valores pequenos e expandem os valores maiores. Se a distribuição for levemente assimétrica à direita, valores positivos de $p < 1$ provocarão uma correção leve em valores grandes. Transformações como a logaritmíca são adequadas para assimetrias à direita moderadas, especialmente quando o desvio padrão é proporcional à média, ou quando o efeito modelado é multiplicativo invés de aditivo. Já transformações com $p < 0$ são aplicadas em assimetrias à direita severas, em que o sinal negativo nos valores é usado para preservar a sua ordem.

Chapter 4

Análise Bidimensional