Estatística Básica

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa
Campus UFV - Florestal

Sumário

Vimos que o resumo de dados por meio de tabelas de frequências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados.

Vimos que o resumo de dados por meio de tabelas de frequências e ramo-e-folhas fornece muito mais informações sobre o comportamento de uma variável do que a própria tabela original de dados.

Muitas vezes, queremos resumir ainda mais estes dados, apresentando um ou alguns valores que sejam representativos da série toda. **Quando usamos um só valor, obtemos uma redução drástica dos dados.** Usualmente, emprega-se uma das seguintes medidas de posição (ou localização) central: **média, mediana ou moda.**

A moda é definida como a realização mais frequente do conjunto de valores observados. Por exemplo, considere a variável Z, número de filhos de cada funcionário casado, resumida na Tabela de dados da companhia MB. Vemos que a moda é 2, correspondente à realização com maior frequência, 7. Em alguns casos, pode haver mais de uma moda, ou seja, a distribuição dos valores pode ser bimodal, trimodal etc.

A mediana é a realização que ocupa a posição central da série de observações, quando estão ordenadas em ordem crescente. Assim, se as cinco observações de uma variável forem $3, 4, 7, 8$ e 8, a mediana é o valor 7, correspondendo à terceira observação. Quando o número de observações for par, usa-se como mediana a média aritmética das duas observações centrais. Acrescentando-se o valor 9 à série acima, a mediana será $(7 + 8)/2 = 7,5$.

Finalmente, a média aritmética, conceito familiar ao leitor, é a soma das observações dividida pelo número de observações.

Se $x_1, \cdots, x_n$ são os n valores (distintos ou não) da variável X, a média aritmética, ou simplesmente média, de X pode ser escrita como:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

Agora, se tivermos n observações da variável X, das quais n_1 são iguais a x_1, n_2 são iguais a x_2, n_k iguais a x_k, então a média de X pode ser escrita como:

$$\bar{x} = \frac{n_1 x_1 + \cdots + n_k x_k}{n_1 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i x_i}{n} \tag{2}$$

Consideremos, agora, as observações ordenadas em ordem crescente. Vamos denotar a menor observação por $x_{(1)}$, a segunda por $x_{(2)}$, e assim por diante, obtendo-se

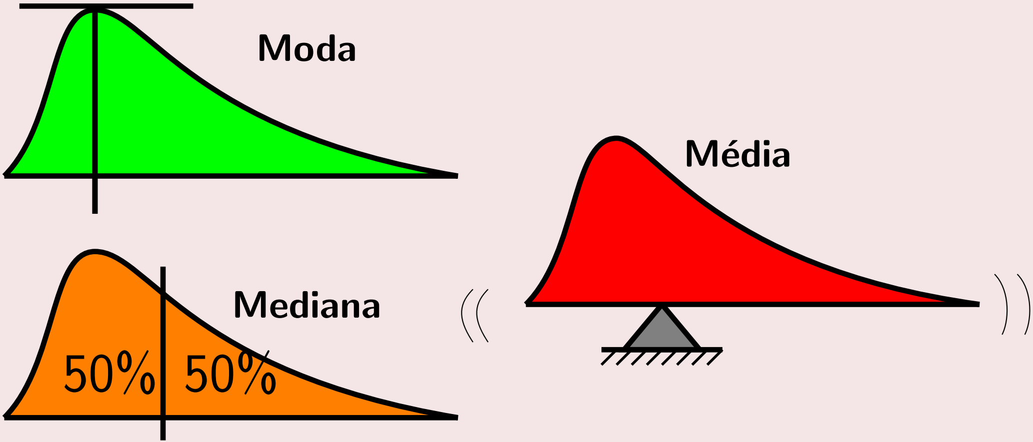
$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}. \tag{3}$$

As observações ordenadas como em (3) são chamadas **estatísticas de ordem**. Com esta notação, a mediana da variável X pode ser definida como:

$$md(X) = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{se} \quad \mathsf{n} \quad \text{é impar,} \\ \\ \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se} \quad \mathsf{n} \quad \text{é par.} \end{cases}$$

A mediana é uma medida mais robusta que a média, quando submetida a mudanças nos valores observados, ou a incorporação de mais observações no conjunto de dados original.

Figura: Visualização Geométrica da moda, média e mediana de uma função densidade de probabilidade arbitrária



Referências