

Conceitos Estatísticos Elementares

Estatística

Estatística é o estudo das inferências sobre uma população, usando amostras [1].

População e Amostra

A população é o conjunto de indivíduos (*i.e.*, objetos estatísticos - não necessariamente pessoas), definido a partir de critérios fixos, com propriedades bem definidas.

Ao contrário, a amostra é um sub-conjunto populacional, selecionado através de um mecanismo aleatório (que permite ao cientista inferir sobre a população e quantificar incertezas ao redor dos resultados).

Dados

Dado é a unidade de informação tipicamente resultante de uma medida (numérico) ou contagem (categórico).

Variáveis

Variáveis servem como um reservatório para dados.

- Numéricas e categóricas

Variáveis podem ser numéricas e categóricas. As variáveis **numéricas ou contínuas** aceitam qualquer valor dentro de um intervalo finito ou infinito (e.g., altura, peso, temperatura, glicose no sangue, etc). Variáveis **categóricas ou discretas** aceitam dois ou mais valores (categorias). As categorias ordinais podem ter uma ordenação intrínseca (e.g., baixo, médio, alto), enquanto as nominais, não (e.g., gênero).

- Dependentes e Independentes

Variáveis dependentes (**Y**) são também chamadas de **variáveis de resultado** ou de resposta. São normalmente, objeto de predição.

As variáveis independentes (**X**) são também chamadas de **variáveis preditoras ou covariadas**. São normalmente obtidas a partir do espaço amostral e usadas como entrada para se obter o resultado (Y) a partir de um dado modelo paramétrico [2].

Parâmetros e Modelos

Modelos paramétricos são frequentemente usados para fazer inferências em um problema estatístico (referente a uma dada população). Assim, do ponto de vista conceitual, os parâmetros de um modelo se relacionam fortemente com o conceito de população, sendo definido como qualquer descritor dos elementos de uma população [3].

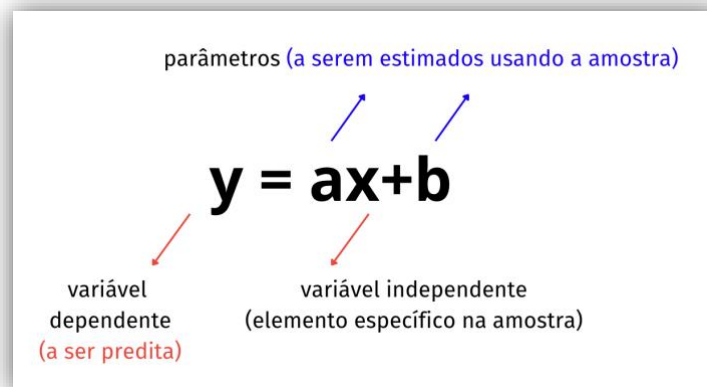


Figura 1. Equação da reta

Estimativa

Estimativa é o processo pelo qual a amostra é usada para aprender sobre a população (e.g., inferindo os **parâmetros** de um modelo) - uma vez que a média amostral é uma estimativa natural da média populacional e a mediana amostral é uma estimativa natural da mediana populacional. Pode-se dizer que se trata de uma **inferência a posteriori**, já que foi feita após o aprendizado da amostra.

Predição

Predição é uma **estimativa a priori**. Isto é, antes da ocorrência do evento. Trata-se do uso do modelo ajustado (construído) para **calcular novos valores** randômicos (e que não pertencem ao conjunto amostral existente).

Distribuição

Cada **variável aleatória** tem uma distribuição que descreve a **probabilidade** assumir qualquer valor permitido na amostra.

Variáveis categóricas podem assumir qualquer número finito de valores **discretos**, sendo a distribuição dada pelo finito conjunto de probabilidades correspondente a esses possíveis valores, com a **probabilidades somando 1**.

Variáveis contínuas possuem várias (de fato, um **número infinito** de) possíveis valores. Consequentemente, a **probabilidade de cada valor é muito pequena**. A princípio, pode-se imaginar a distribuição de uma variável contínua como um histograma extremamente detalhado em que as larguras dos intervalos se tornam muito pequenas.

Assim, do histograma é possível obter uma curva chamada **Função de Densidade de Probabilidade (FDP)** da variável aleatória e a área em baixo da FDP **sobre qualquer intervalo** representa a probabilidade do valor aleatório terá um valor dentro daquele intervalo.

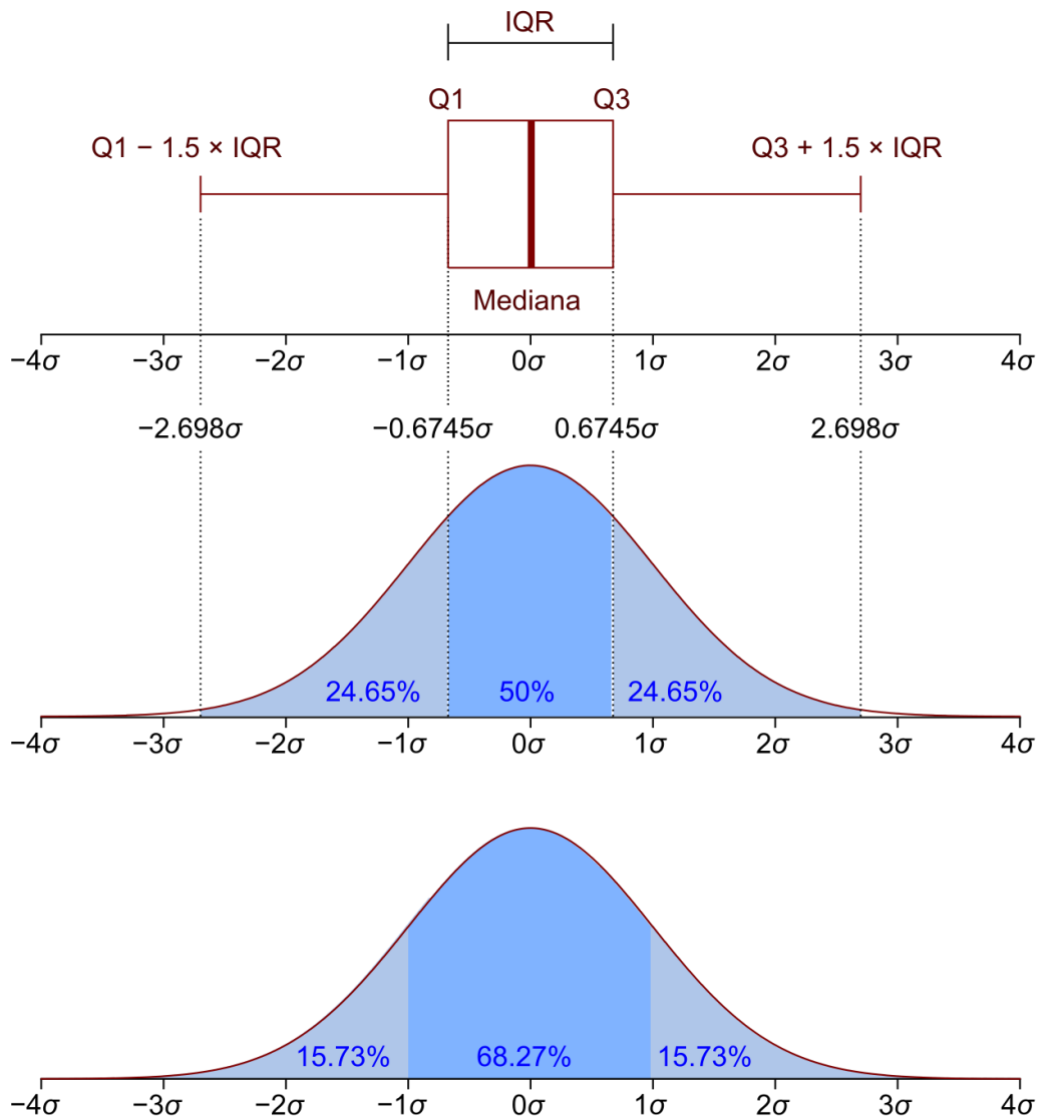


Figura 2. Função de Densidade de Probabilidade de uma curva normal e diagrama de caixa.

Bibliografia

- [1] R. Etzioni, M. Mandel e R. Gulati, Statistics for Health Data Science - An Organic Approach, Pittsburg, PA - USA: Springer, 2020.
- [2] S. Sayad, An Introduction to Data Science, Toronto, Canadá: Saed Sayad, 2010.
- [3] R. Blair e R. Taylor, Bioestatística para Ciências da Saúde, Sao Paulo, SP: Pearson, 2013.