Roteiro de aulas - Estatística descritiva (Análise descritiva de dados)

Disciplina: *Probabilidade e Estatística*

Professor: André Oliveira

23 de setembro de 2025

Estatística descritiva

Em sentido amplo, a Estatística é dividida em duas áreas: **Estatística descritiva** e **Estatística Inferencial. Estatística Descritiva (Estatística dedutiva)** consiste em métodos para se organizar, exibir e descrever dados utilizando tabelas, gráficos e medidas resumidas. Já a **Estatística Inferencial** (Estatística Indutiva) consiste em métodos que utilizam resultados de amostras para auxiliar na tomada de decisão ou na realização de prognósticos sobre uma população de interesse. O diagrama seguinte mostra o contexto em que se situa o estudo da Estatística, aqui subdividido em **Estatística Descritiva** e **Estatística Indutiva** (ou Inferencial).

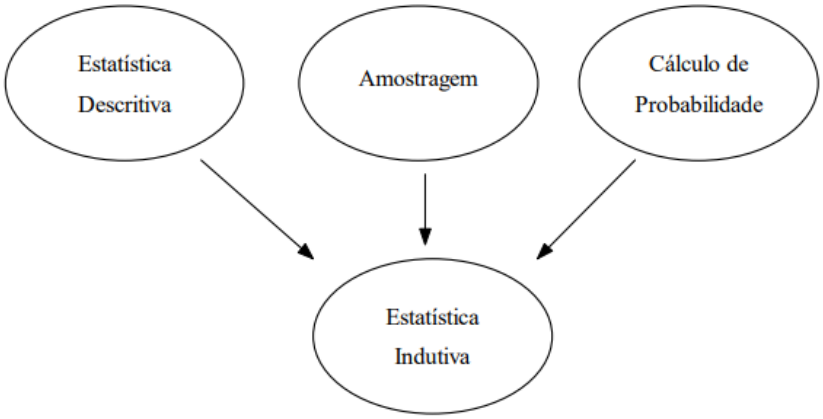

Figura 1: Diagrama de associação entre a Estatística Descritiva e Estatística Inferencial

Probabilidade

É que fornece uma medida da possibilidade de que um determinado resultado venha a ocorrer, atua como uma ligação entre a **Estatística descritiva** e a **Estatística Inferencial** (Estatística Indutiva).

Censo

Uma pesquisa que inclua todos os membros da população é chamada de censo.

Parâmetro

É uma medida usada para caracterizar uma população. Assim sendo para se obter o valor de um parâmetro é necessário coletar a informação a respeito de uma ou mais variáveis em todos os indivíduos dessa população, ou seja, realizar um censo da mesma. Em geral os parâmetros mais estudados são **médias, variâncias, desvio-padrão, mediana, moda, proporção**.

Amostras

A maioria das vezes, as decisões são tomadas com base em **parcelas de populações**. O conjunto composto por um número de elementos selecionados a partir de uma população é chamado de **amostra.** Uma parcela da população selecionada para fins de estudo é conhecida como uma amostra. É um subconjunto não vazio de observações, indivíduos ou objetos de uma população de interesse.

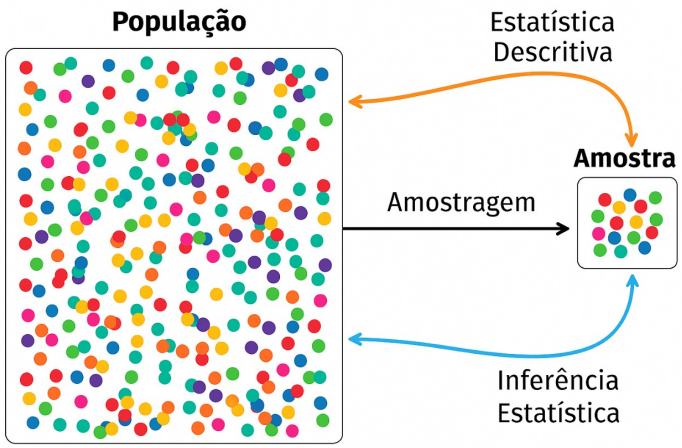

Figura 2: Diagrama sobre a relação entre Amostra, Amostragem, Populações, Estatística Descritiva e Inferência

Estimador

Na grande maioria das situações, não é possível realizar o censo de uma **população**, porque ou a população é muito grande ou é de difícil acesso. Para contornar este problema, o pesquisador pode retirar uma amostra da população e a partir desta amostra caracterizar a população de onde a amostra foi retirada sem nenhum viés. Para alcançar este objetivo se usa **fórmulas estatísticas**, conhecidas como **estimadores** (estatísticas), que apresentem características estatísticas desejáveis, tais como não tendenciosidade, variância mínima, fornecer **estimativas** que se aproximem do valor paramétrico à medida que o tamanho da amostra aumenta ou a medida que a amostra seja representativa da população. O **parâmetro** é sempre um valor constante, pois para a obtenção do mesmo são usados todos os elementos da população. Por outro lado, o **estimador** representa uma variável aleatória, pois os seus valores mudam de **amostra para amostra**. Isto acontece porque os elementos que pertencem a uma amostra geralmente não são os mesmos em outras amostras e asssim é possível estabelecer uma **distribuição de probabilidades**. Para o parâmetro, isto não é possível, pois se assume que ele tem um valor constante. Por isto recomenda-se muito cuidado para usar corretamente a simbologia para o **parâmetro** e para o **estimador**.

Estimativa

Conforme mencionado anteriormente, os estimadores podem assumir valores diferentes em amostras diferentes. Estes diferentes valores que um estimador assume são também conhecidos como **estimativas**.

Amostra Representativa

Uma amostra que representa, o mais próximo possível, as características da população é chamada de amostra representativa. A **estimativa** só será uma estimativa sem viés se a amostra representar bem as caracteristicas da população em estudo.

Uma amostra pode ser aleatória ou não aleatória. Em uma amostra aleatória, cada elemento da população tem uma chance de ser incluído na amostra. Entretanto, em uma amostra não aleatória esse pode não ser o caso. **Amostra Aleatória** Uma amostra extraída de maneira tal que cada elemento da população tenha uma chance de ser selecionado é chamada de amostra aleatória. Caso todas as amostras de um mesmo tamanho, selecionadas de uma determinada população, tenham a mesma probabilidade de vir a ser selecionadas, damos a esse procedimento o nome de amostragem aleatória simples. Essa amostra é conhecida como amostra aleatória simples (AAS).

Variável

Uma variável corresponde a uma **característica sob estudo**, que assume diferentes valores para diferentes elementos. Em contraposição a uma variável, o valor de uma constante é **fixo**. Conhecer a classificação da variável define o tipo de **representação**

gráfica (histograma, *boxplot*, gráfico de barras ou de setores) e o método (modelo) de análise da mesma. É uma característica mensurável de acordo com alguma escala e que pode assumir valores diferente de elemento para elemento. É uma característica a ser observada em cada elemento da amostra. É importante ressaltar que uma variável não são necessariamente numéricos, uma vez que podem dizer respeito a atributos qualitativos observados na população.

Observação ou Medição

O valor de uma variável para um elemento específico da população é chamado de **observação** ou **medição.**

Conjunto de Dados

Um conjunto de dados é uma compilação de observações sobre uma ou mais variáveis.

Estatística descritiva

As medidas descritiva se dividem em **medidas de posição** e **medidas de dispersão.**

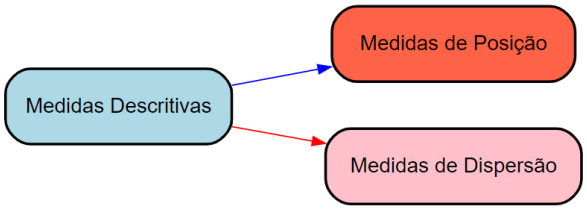

Figura 3: Divisão da Estatística Descritiva

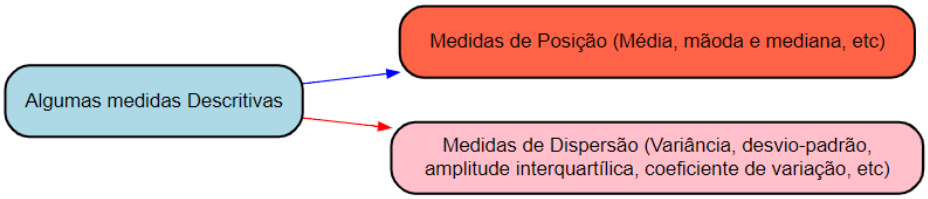

Figura 4: Algumas medidas descritivas

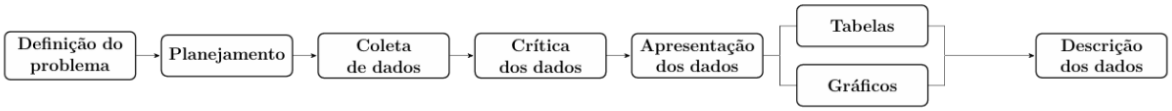

Figura 5: Fluxo da Estatística Descritiva

Medidas de tendência central ou Medida resumo (Medidas de posição).

Média, moda e mediana são as medidas mais comuns na Estatística Descritiva.

Média aritmética simples

Considere uma variável X com observações representadas por $(x_1,\ x_2,\ x_3,\ \cdots, x_n)$. Por atender certos critérios de **estimadores** o melhor estimador da média populacional é a soma dos valores dividido pelo número de observações (n), ou seja, o estimador da média populacional é $\widehat{\mu}_x = \dfrac{\sum_{i=1}^{n} X_i}{n}$. A média de uma população (que é fixa) é representada pela letra grega minúscula, sendo definida por $\mu_x = \dfrac{\sum_{i=1}^{N} X_i}{N}$.

Média aritmética ponderada

Para os dados agrupados em uma tabela de distribuição de frequência a média deve ser obtida ponderando-se o valor médio da classe (ou variável) pela sua respectiva frequência $\widehat{\mu}_x = \dfrac{\sum_{i=1}^{n}(f_i.X_i)}{\sum_{i=1}^{n} f_i}$, em que $\sum_{i=1}^{n} f_i$ é o tamanho da amostra.

Mediana

A mediana é o valor que ocupa a posição **central** dos dados quantitativos ordenados *(Rol)*. Assim 50% dos valores estão acima da mediana e 50% estão abaixo da mediana.

$$M_d = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2} & \text{se } n \text{ é par} \end{cases}$$

Moda

O termo **moda** foi utilizado pela primeira vez em 1895 por Karl Pearson (1857-1936), possivelmente em referência ao seu significado usual. Embora a palavra **moda** possa estar relacionada a desfiles e roupas em geral, em um sentido mais amplo, significa uma ação, uma atitude ou um pensamento que é mais praticado ou frequente. A moda é o valor mais frequente. Um conjunto de valores amostrais pode ser unimodal, bimodal, multimodal ou amodal.

Exemplo

As informações a seguir referem-se ao consumo mensal de dados móveis de Carlos, medido em *Gigabytes (GB)*, ao longo de 24 meses. O plano contratado por Carlos possui uma franquia de 60 GB por mês.

Dados brutos: $[19, 15, 15, 14, 22, 17, 17, 13, 13, 19, 11, 19, 16, 17, 17, 33, 27, 18, 18, 22, 17, 19, 18, 17]$

Dados em *ROL*: $[11, 13, 13, 14, 15, 15, 16, 17, 17, 17, 17, 17, 17, 18, 18, 18, 19, 19, 19, 19, 22, 22, 27, 33]$

Neste caso a **estimativa** da **moda** são 17,00 *Gigabit* (GB), pois foi o que mais frequente ($f = 6$)

A **estimativa** da **média** é $18,04$ *Gigabit* (GB), pois:

$$\widehat{\mu}_x = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{19 + 15 + \cdots + 17}{24} = \frac{433}{24} = 18,04 \quad \textit{Gigabit} \text{ (GB)}$$

A **estimativa** da mediana são 17,00 *Gigabit* (GB), pois neste caso ($n = 24$) que é um núemro par, assim o estimador:

$$M_d = \begin{cases} X_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

Se reduz a $M_d = \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+2}{2}\right)}}{2}$, pois $n = 24$ que é um número par.

Assim $M_d = \dfrac{X_{\left(\frac{24}{2}\right)} + X_{\left(\frac{24+2}{2}\right)}}{2} = \dfrac{X_{(12)} + X_{(13)}}{2} = \dfrac{17 + 17}{2} = 17,00 \quad \textit{Gigabit}$ (GB)

Assim a estimativa da **moda, média, mediana** do consumo são respectivamente 17,00; 18,04 e 17,00 *Gigabit* (GB), neste caso pode-se dizer que o valor médio de consumo está muito distante dos 60 *Gigabit* (GB) contratados. Observando os valores da amostra constata que o maior valor foi de 33 GB, e este fato pode auxiliar ao Carlos descacartar a necessidade de contratar novo serviço com mais de 60 GB em caso de uma oferta da operadora.

Observação

Vale rassaltar a diferença entre as **nomenclaturas/notações** de x_{12} *e* $x_{(12)}$ como já descrito na aula sobre **amostra** e **populações**, e, esta diferença na notação, faz toda diferença para estimar a mediana. Outro fato que deve ser ressaltado é que se diz estimativa, pois a **moda, média, mediana** acima foram obtidas de uma amostra e não de **toda** população.

Medidas de variabilidade (dispersão)

As medidas de posição (média, moda e mediana) para um conjunto de dados ao serem apresentadas é necessário que está seja acompanhada de uma medida de variabilidade

(dispersão) dos dados. Para os métodos estatísticos, a medida de dispersão é de fundamental importância, pois a estatística só existe porque os fenômenos têm variabilidade. Algumas medidas de variabilidade serão apresentadas abaixo.

Amplitude total (AT):

A amplitude total é uma medida de variabilidade e é definida por $AT = X_{max} - X_{min}$ em que X_{max} é o maior valor observado e X_{min} é o menor valor observado.

Variância

A variância é uma medida que serve para medir variabilidade. Em geral usada para comparar grupos (tratamentos) expressos numa mesma medida e com médias aproximadamente iguais. Sendo assim quanto menor a valor da variância, maior é o grau de homogeneidade do conjunto de dados analisados. Os estimadores da variância **populacional** e da variância **amostral** podem ser obtidas respectivamente por:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(X_i - \widehat{\mu}_x)^2}{N} \qquad \text{e} \qquad \widehat{\sigma}_x^2 = \frac{\sum_{i=1}^{n}(X_i - \widehat{\mu}_x)^2}{n-1}$$

Alternativamente a variância **populacional** e **amostral** podem ser obtidas respectivamente por.

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N} \qquad \text{e} \qquad \widehat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$$

A variância apresenta um inconveniente de ordem prática, pois, como ela é expressa em unidades ao quadrado, isto implica na dificuldade de interpretação. Uma medida de variabilidade, calcula com base na variância, é **desvio-padrão**, o qual é expresso na mesma unidade dos dados originais.

Desvio-padrão

Avalia a variabilidade dos dados individuais em torno da média amostral ou populacional. Fundamental para uma análise descritiva. O desvio-padrão nada mais é que a raiz quadrada da variância. O estimador do *desvio-padrão* **populacional** e do desvio-padrão **amostral** são respectivamente $\sigma_x = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu_x)^2}{N}}$ e $\widehat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \mu_x)^2}{n-1}}$

Alternativamente o desvio-padrão **populacional** e **amostral** podem ser obtidas respectivamente por.

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N}} \qquad \text{e} \qquad \widehat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}}$$

Coeficiente de variação populacional (CV)

O coeficiente de variação é uma medida de variabilidade definida em relação a média $CV_x = 100.\left(\frac{\sigma_x}{\mu_x}\right)$.

Coeficiente de variação amostral (CV)

O coeficiente de variação é uma medida de variabilidade definida em relação a média $cv_x = 100.\left(\frac{\widehat{\sigma}_x}{\widehat{\mu}_x}\right)$.

Exemplo

Voltamos ao caso das informações sobre o consumo mensal de dados móveis de Carlos, medido em *Gigabytes (GB)*, ao longo de 24 meses. O plano contratado por Carlos possui uma franquia de 60 GB por mês.

Dados brutos: $[19, 15, 15, 14, 22, 17, 17, 13, 13, 19, 11, 19, 16, 17, 17, 33, 27, 18, 18, 22, 17, 19, 18, 17]$

Para os dados desta amostra:

A amplitude total é $AT = X_{max} - X_{min} = 33 - 11 = 22$ $(Gigabit\ GB)$

$\widehat{\mu}_x = \dfrac{\sum_{i=1}^{n} X_i}{n} = \dfrac{433}{24} = 18,04$ (Gigabit GB).

Assim a variância $\widehat{\sigma}_x^2 = \dfrac{\sum_{i=1}^{N} (X_i - \widehat{\mu}_x)^2}{n-1} = \dfrac{(19-18,04)^2 + (15-18,04)^2 + \cdots + (17-18,04)^2}{24-1}$

$\widehat{\sigma}_x^2 = 21,09$ $(Gigabit\ GB)^2$.

O desvio-padrão $\widehat{\sigma}_x = \sqrt{21,09\ (Gigabit\ GB)^2} = 4{,}59$ (Gigabit GB)

Por fim quem optar por medir variabilidade dos dados pelo coeficiente de variação:

$CV_x = 100.\left(\frac{\sigma_x}{\mu_x}\right) = 100.\left(\frac{4,59}{18,04}\right) = 25,44\%$

Comentários

Para compor relatórios de uma amostra o ideal é que divulgue uma medida de tendência central (posição) associada a uma medida de variação (dispersão).

Medidas descritivas de uma Tabela de frequências absolutas (f_i)

As medidas descritivas podem ser obtidas para dados de uma tabela de frequência.

Exemplo

Paulo assinou um contrato com a empresa $NETVIX$ que oferece planos de internet de 70 MB. Usando o aplicativo $FAST$ e com a finalidade de estimar a velocidade real da entrega prevista no contrato ele mensurou a velocidade após a instalação do modem por 12 dias.

Tabela 1: Dados da frequência da medição emn MB

Megabytes (MB)	fi
56	3
57	1
59	1
61	3
62	2
64	1
65	1

Fonte: Elaboração própria

Assim a estimativa da **média amostral** e da **variancia amostral** da distribuição de **frequências absolutas (fi)** para esta amostra pode ser obtida pela **Tabela 1**.

$$\widehat{\mu}_x = \frac{\sum_{i=1}^{n}(f_i.X_i)}{\sum_{i=1}^{n} f_i} = \frac{3\times 56 + 1\times 57 + \cdots + 1\times 65}{3+1+\cdots+1} = \frac{720,00}{12,00} = 60,00 \text{ MB.}$$

$$\widehat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} f_i.\left(X_i - \hat{\mu}_x\right)^2}{\sum_{i=1}^{n} f_i - 1} = \frac{3\times(56-60)^2 + 1\times(57-60)^2 + \cdots + 1\times(65-60)^2}{12-1} = 10,00 MB^2$$

Assim $\widehat{\sigma}_x^2 = 10\ MB^2$ e $\widehat{\mu}_x = 60\ MB$

No caso de uma tabela de frequências a estimativa da variância também pode ser obtida por $\widehat{\sigma}_x^2 = \dfrac{\sum_{i=1}^{n} f_i.X_i^2 - \dfrac{\left(\sum_{i=1}^{n} f_i.X_i\right)^2}{n}}{n-1}$

$$\sum_{i=1}^{n}(f_i \times X_i^2) = 3 \times 56^2 + 1 \times 57^2 + \cdots + 1 \times 65^2 = 43310,00$$

$$\sum_{i=1}^{n}(f_i \times X_i) = 3 \times 56 + 1 \times 57 + \cdots + 1 \times 65 = 720,00$$

$$\widehat{\sigma}_x^2 = \frac{\sum_{i=1}^{n} f_i.X_i^2 - \dfrac{\left(\sum_{i=1}^{n} f_i.X_i\right)^2}{n}}{n-1} = \frac{43310 - \dfrac{720^2}{12}}{11} = 10,00\ MB^2$$

Comentários

Pode-se dizer que há **evidência** de que a velocidade contratada de 70 MB não está sendo entregue, pois a média amostral foi de 60 MB e a variância amostral de 10 MB^2.