Modelos Lineares
Medidas de tendência central e de variabilidade

Professora Ariane Ferreira

Instituto Politécnico
Campus Regional da UERJ
Nova Friburgo - RJ

Dados Empíricos

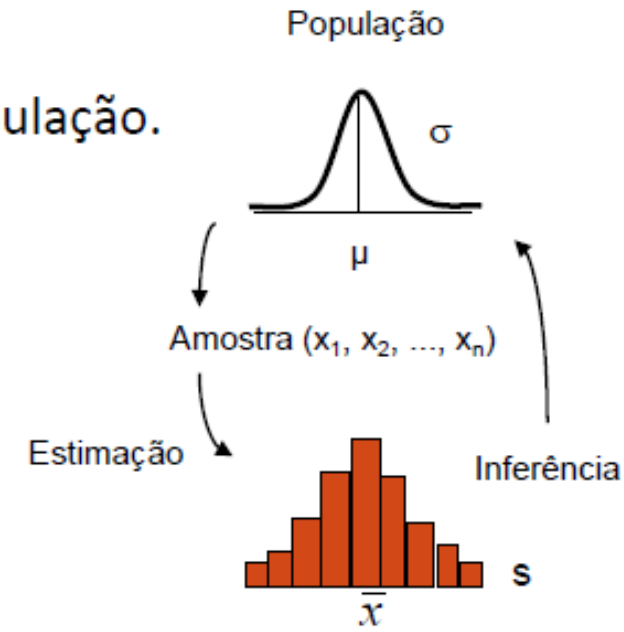
Os dados empíricos coletados de um processo devem formar a base para as decisões e ações.

Uma vez que os dados brutos tenham sido coletados, eles devem ser tabulados e convertidos em “informação” através do uso de métodos estatísticos.

Coleta de dados

População: corresponde ao sistema ou ao todo que se quer descrever. É um conjunto de elementos com características comuns.

- **Censo**: inspeciona todos os elementos de uma população.
  - **Parâmetros**: valor desconhecido associado a uma característica (média = μ, variância = σ^2)
- **Amostra**: é uma parte representativa da população.
  - **Estimador**: função que estima o valor de um Parâmetro baseando-se nas observações (média = $\bar{x}$, variância = s^2)



Amostra Representativa

- Cada individuo da População tem exatamente a mesma probabilidade de ser selecionado na amostra.
- O tamanho n da amostra é suficientemente elevado.
- Ele depende da homogeneidade da população e da precisão desejada.

↑ n ↑ p (detectar pequenas mudanças)

↑ freqüência de amostragem ↑ p (detectar mudanças)

Ideal amostras grandes com alta freqüência → não é viável economicamente

Dicotomia:

↓ n ↑ freqüência mais utilizado na indústria

↑ n ↓ freqüência

Estratificação dos dados

Trabalha-se com dados classificados em agrupamentos:

- camadas ou estratos

- Tempo: os resultados são diferentes de manhã, à tarde ou a noite?
- Local: os resultados são diferentes nas linhas de produção?
- Tipo: os resultados obtidos são diferentes entre os fornecedores?
- Indivíduo: é possível comparar os operadores?

Tipos de dados

- **Contínuos = Variáveis quantitativas**
  - Os valores são um conjunto infinito de $\Re$
  - infinitos valores possíveis entre dois extremos
    - Tempo (1h:35min),
    - Pressão (1.013,105 KPa) ,
    - Dimensão (16,54 mm),
    - Temperatura (23,5℃)
- **Discretos**
  - Os valores são um conjunto infinito de N
  - exemplo: número de filhos
- **Categoricos = Variáveis qualitativas**
  - São conjuntos de valores finitos— numéricos (codigos e não quantidades)
- **Textuais**
  - Letras, textos

Análise Exploratória dos Dados

- **Explorar a distribuição das Variaveis**
- **Verificar a confiabilidade das variaveis**
  - Valores incoerentes ou faltantes
    - Imputação ou supressão
- **Detectar valores extremos (outliers)**
  - Eliminar os valores aberrantes
- **Variaveis Continuas**
  - Detectar a não monotia ou a não-linearidade (para discretisação)
  - Testar a normalidade das variables (sobretudo para amostras de pequenos efetivos), transforma-las se necessario para aumentar a normalidade
  - facultatifvo: testar a **homoscédasticidade** (igualidade das variâncias-covariâncias)

Análise Exploratória dos Dados

- **Variaveis discretas**
  - agrupar certas modalidade muito numerosas ou com efetivos muito pequenos (peso muito grande) valores incoerentes ou faltantes.
- **Criar indicadores pertinentes à partir dos dados brutos**:
  - Conjunto das variaveis « produto Pi comprado (Oui/Non) » permite deduzir o numero de produtos comprados
  - Numero e valor de compras ⇒ valor médio de uma compra
  - datas de compras ⇒ frequência das compras
- **Detectar as ligaçéoes entre as variaveis**
  - entre variaveis explicativas e a variavel a ser explicada (bon)
  - Nas variaveis explicativas entre si (multicolinearidade: ruim para certos métodos lineares)

Valores Faltantes

- **Outras soluções :**
  - Supressão das observações (se elas são pouco numerosas)
  - Trocar a variavel por outra correlacionada, mas sem valores faltantes
  - Tratar os valores faltantes
  - Preencher os valores faltantes com auxilio de fonte externa
- **Imputação estatistica**
  - Usar a moda, a média ou a mediana
  - Usar uma regressão ou arvore de regressão

A imputação não é jamais neutra:

- Os dados não são faltantes ao acaso;
- Deformação das variâncias e correlações

Descrição de uma variavel quantitativa

Estudo de uma variavel numérica **X**

- Seja uma variavel numérica **X** com os valores $x_1,\ldots, x_i,\ldots, x_N$ na população e $x_1,\ldots, x_i,\ldots, x_n$ **na amostra**.
- Ela é resumida pelas estatisticas :
  - de tendência Central (média, médiana) e
  - de dispersão (variância, desvio-padrão, amplitude, quantis, distância entre quantis)
- A dispersão de **X** é visualisada através de
  - Box plot
  - histograma.
- Ajustemento de uma lei de probabilidade conhecida aos dados.

Description d'une variable quantitative

Étude d'une variable numérique X

- Une variable numérique **X** prend des valeurs $x_1,\ldots, x_i,\ldots, x_N$ sur une population et $x_1,\ldots, x_i,\ldots, x_n$ sur un échantillon.
- Elle est résumée par des statistiques :
  - de tendance centrale (moyenne, médiane) et
  - de dispersion (variance, écart-type, étendue, quartiles, distance interquartiles)
- La dispersion de **X** est visualisée par
  - la boîte-à-moustache et
  - l'histogramme.
- Ajustement d'une loi de probabilité connue aux données.

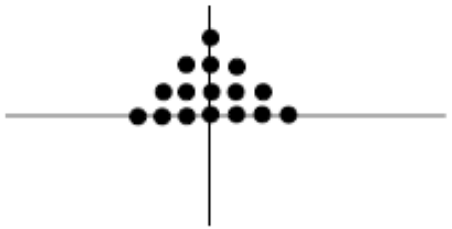
Variavel Quantitativa

1) Medidas de tendência central
2) Medidas de variabilidade
3) Histograma
4) Boxplot
5) Distribuição de probabilidade Normal
6) Gráfico de normalidade

Variavel Quantitativa
Medidas de Tendência Central

A tendência central é uma medida do centro de um conjunto de dados segundo uma regra estabelecida a priori (média aritmética, geométrica, harmônica, ponderada, etc.)

- Média aritmética
- Mediana
- Moda



Variavel Quantitativa

Medidas de Tendência Central

Média aritmética $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- Anotamos a temperatura de uma pessoa de 1 em 1 hora, durante 8 horas. Qual a média da temperatura?
- Valores observados: 37, 37, 38, 39, 37, 39, 39°C.
- O tamanho da amostra é n = 7

$$\bar{x} = \frac{37+37+38+39+37+39+39}{7} = 38°C$$

A média amostral é bom um estimador da média populacional . Quanto maior **n** melhor a estimativa.

Variavel Quantitativa

Medidas de Tendência Central

Mediana

$$\tilde{x} = \begin{cases} x_{((n+1)/2)} & n \quad \text{ímpar} \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & n \quad \text{par} \end{cases}$$

- Ela é não é influenciada pelos dados atípicos
- Deve-se ordenar os dados em ordem crescente
- Qual a mediana da temperatura?
- Valores observados: 37, 37, 38, 39, 37, 39, 39ºC.
- Valores ordenados: 37, 37, 37, 38, 39, 39, 39ºC
- n = 7 é ímpar – mediana valor central $\tilde{x} = 38°C$

Variavel Quantitativa

Medidas de Tendência Central

Moda: observação que ocorre com mais freqüência

- Qual a moda da temperatura?
- Valores observados: 37, 37, 37, 38, 39, 39, 39ºC
- Duas modas: 37 e 39ºC

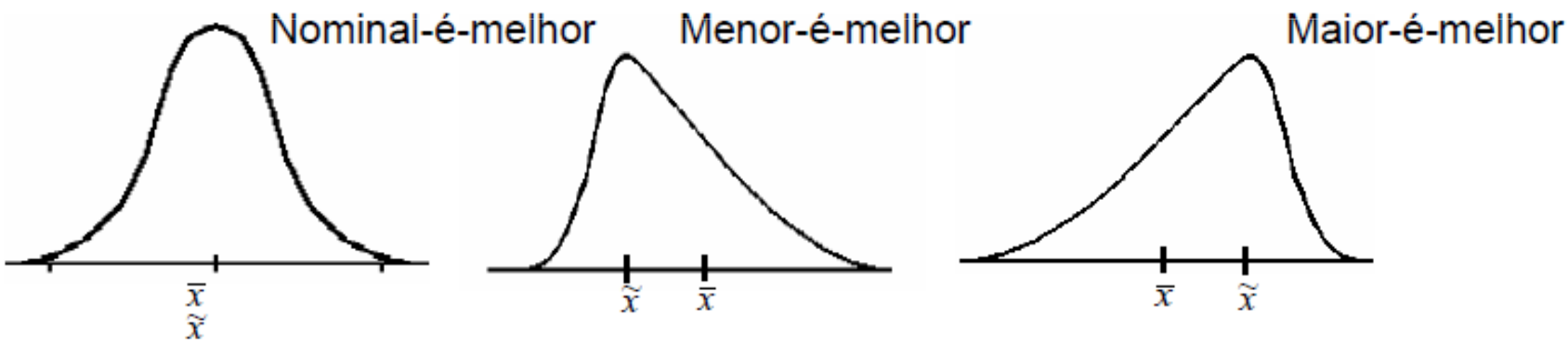
Variavel Quantitativa

Medidas de Tendência Central

Relação entre média e mediana → fornece a forma da dispersão

A	Distribuição simétrica	10 12 14 16 18	$\bar{x} = 14 = \tilde{x} = 14$
B	Distribuição assimétrica à direita	10 12 14 16 23	$\bar{x} = 15 > \tilde{x} = 14$
C	Distribuição assimétrica à esquerda	05 12 14 16 18	$\bar{x} = 13 < \tilde{x} = 14$

Mediana tem maior robustez a dados atípicos do que a média



Simétrica
Forma de Sino

Assimétrica à Direita
Assimetria Positiva

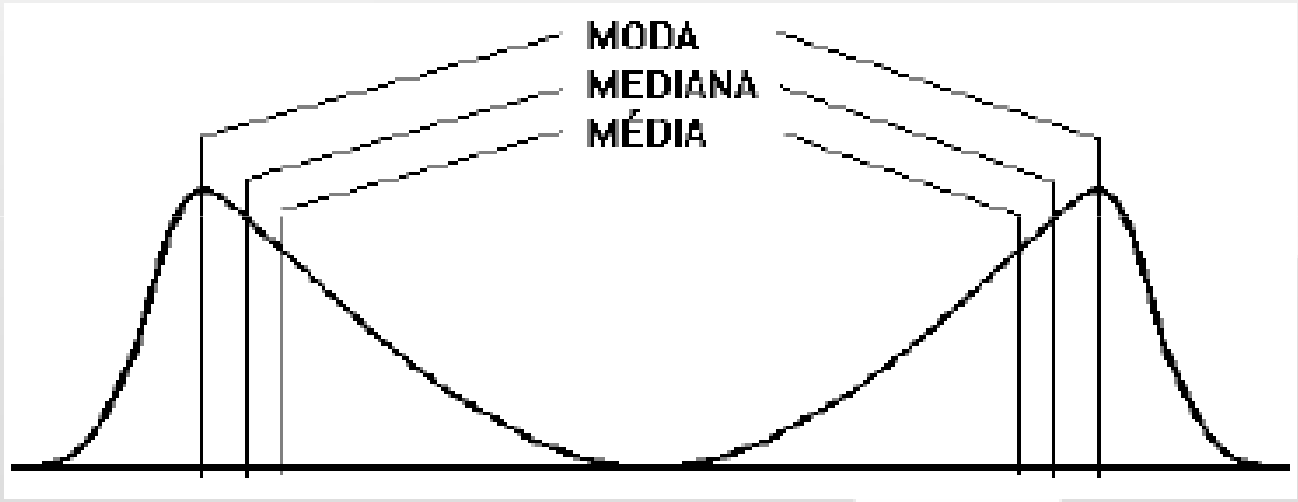
Assimétrica à Esquerda
Assimetria Negativa

Variavel Quantitativa

Medidas de Tendência Central

Para distribuições simétricas a média, a mediana e a moda coincidem aproximadamente.

Para distribuições assimétricas observa-se o seguinte:



Variavel Quantitativa

Medidas de Tendência Central

Média Geométrica (G)

É a raiz de ordem *n* do produto dos valores da amostra:

$$G = \sqrt[n]{X_1 X_2 \ldots . X_n}$$

Exemplo

A média geométrica de 12 14 16 é:

$$G = \sqrt[3]{12 \times 14 \times 16} = 13{,}90$$

Variavel Quantitativa

Medidas de Tendência Central

Média Harmônica (H)

É o inverso da média aritmética dos inversos das observações.

$$H = \frac{1}{\frac{1}{n}\sum\frac{1}{X_i}} = \frac{n}{\sum\frac{1}{X_i}}$$

Exemplo

A média harmônica de 12 14 16 é:

$$H = \frac{3}{\frac{1}{12}+\frac{1}{14}+\frac{1}{16}} = 13{,}81$$

Variavel Quantitativa

Medidas de Tendência Central

Relação entre Média Aritmética, Geométrica e Harmônica:

A média geométrica e a média harmônica são menores, ou no máximo igual, à média aritmética.

A igualdade só ocorre no caso em que todos os valores da amostra são idênticos.

Quanto maior a variabilidade, maior será a diferença entre as médias harmônica e geométrica e a média aritmética.

$$H \leq G \leq \overline{X}$$

Exemplo: Para a amostra *12 14 16* tem-se

$H = 13{,}81 < G = 13{,}90 < \overline{X} = 14{,}00$

Variavel Quantitativa

Medidas de Variabilidade

Amplitude: R = Xmax - Xmin

Exemplo: 8,5 8,7 8,9 10,1 10,5 10,7 11,5 11,9

R = 11,9 - 8,5 = 3,4

- A amplitude é fácil de calcular e fornece uma idéia da magnitude da faixa de variação dos dados.
- Não informa a respeito da dispersão dos valores que caem entre os dois extremos.
- Ela é influenciada pelos dados atípicos
- Quando n < 10 pode resultar em uma medida de variação bastante satisfatória.

Variavel Quantitativa
Medidas de Variabilidade

Quartis

É qualquer um dos três valores que divide o conjunto ordenado de dados em quatro partes iguais, e assim cada parte representa 1/4 da amostra ou população. Ela não é influenciada pelos dados atípicos

- **1º quartil ou quartil inferior (Q1)** = valor aos 25% da amostra ordenada
- **2º quartil ou mediana (Q2)** = valor até ao qual se encontra 50% da amostra ordenada
- **3º quartil ou quartil superior (Q3)** = valor a aos 75% da amostra ordenada

Variavel Quantitativa

Medidas de Variabilidade

Exemplo

Amostra: 36, 40, 7, 41, 15, 39

Amostra ordenada: 7, 15, 36, 39, 40, 41

Q1 = 15

Q2 = (39+36)/2 = 37,5

Q3 = 40

Intervalo inter-quartil: Q3-Q1 (40 - 15 = 25)

Regra para descobrir os quartis

1) use a mediana para dividir os dados ordenados em duas metades, não inclua a mediana nas metades
2) o quartil inferior (ou superior) é a mediana da metade inferior (ou superior).

Variavel Quantitativa

Medidas de Variabilidade

Variância

Quadrado da distância de todos os valores x_i em relação a sua média

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Desvio-padrão

A raiz quadrada da variância (é expresso na unidade original dos dados)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$$

Variavel Quantitativa

Medidas de Variabilidade

Nem sempre se conhece a variância e o desvio padrão populacional. Desta forma, deve-se usar um estimador a partir de uma amostra

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \quad populacional$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \quad populacional$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad amostral$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad amostral$$

A correção de Bessel (eliminando 1 grau de liberdade para n < 30) torna a variância amostral um estimador da variância populacional não-viesado.

Variavel Quantitativa
Medidas de Variabilidade

Exemplo

Amostra: 10 12 14 16 18 (*meses*)

A média é 14 cm, a variância e o desvio-padrão são:

$$s^2 = \frac{(10-14)^2 + (12-14)^2 + (14-14)^2 + (16-14)^2 + (18-14)^2}{5-1} = 9{,}98\ meses^2$$

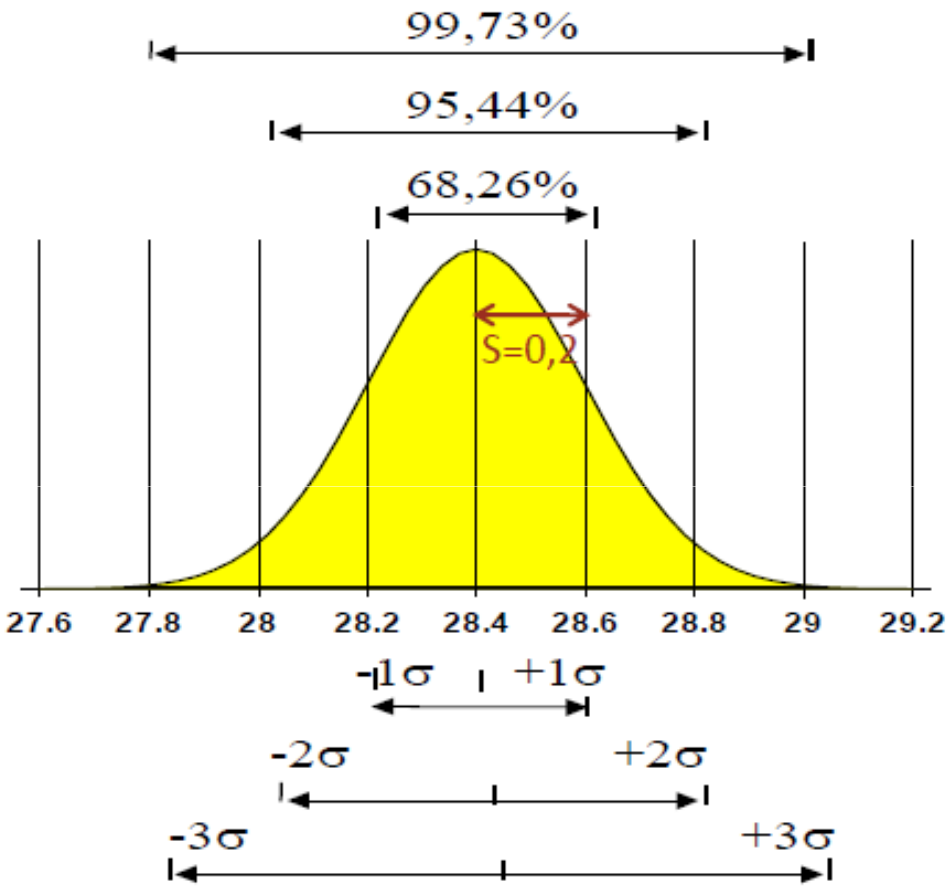
Os desvios de cada valor em relação à média totalizam zero pois a média é o valor central

$$s = \sqrt{9{,}98\ cm^2} = 3{,}16\ meses$$

A média e o desvio padrão possuem a mesma unidade de medida

Variavel Quantitativa
Medidas de Variabilidade

Seja um processo com média 28,4 e desvio-padrão S= 0,2



Variavel Quantitativa

Medidas de Variabilidade

Coeficiente de variação $CV = \frac{s}{\overline{x}} \times 100$

- Um desvio padrão pode ser considerado grande ou pequeno dependendo da ordem de grandeza da média da variável.
- Quanto menor o CV mais homogêneo é o conjunto de dados.
- Medida adimensional, útil para comparar resultados de amostras cujas unidades podem ser diferentes.