Estatística Aplicada às Ciências Sociais

Conceitos

- **Estatística:** é a ciência que tem por objetivo **planejar, coletar, tabular, analisar e interpretar informações** e delas extrair conclusões que permitam a **tomada de decisões** acertadas mediante incertezas.

- **Áreas:** Estatística **Descritiva** e Estatística **Inferencial** ou Indutiva

Conceitos

- **População:** é o conjunto de elementos (valores, pessoas, medidas etc.) que tem pelos menos uma característica em comum.
  - Alunos de 5 a 12 anos da rede pública do município de Gurupi-TO (para verificação de parasitas intestinais)
  - Idosos integrantes da Unati - Universadade Aberta à Terceira Idade (importância da relação médico – paciente, percepção sobre a atuação do médico)
  - Calendula officinalis L. (ASTERACEA). Influência do processo extrativo nas características físicas e químicas dos extratos.

- **Amostra:** é um subconjunto de elementos extraídos de uma população.

Conceitos

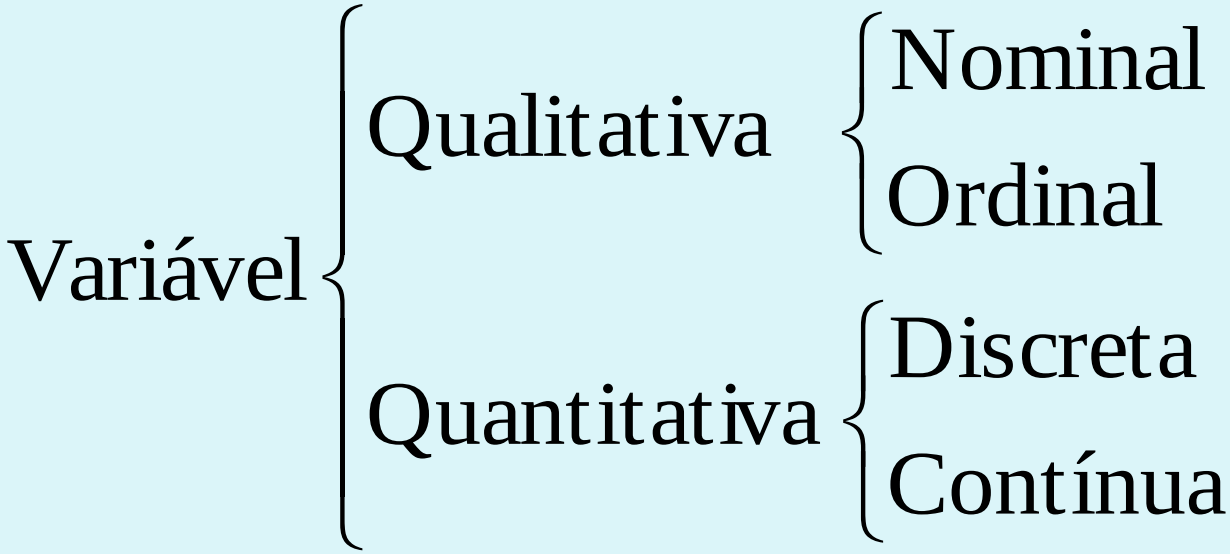
- **Parâmetro:** é uma medida numérica que descreve uma característica de uma população.
- **Estatística:** é uma medida numérica que descreve uma característica da amostra.
- **Dados primários:** dados coletados pelo próprio pesquisador e sua equipe.
- **Dados secundários:** não foram obtidos pelo pesquisador e sua equipe.

Conceitos

- **Censo:** é uma coleção de dados relativos a todos os elementos de uma população.

- **Variável:** é a característica de interesse que é medida em cada elemento da amostra ou população, podendo ter resultados numéricos ou não. Seus valores variam de elemento a elemento.

Variáveis - Classificação



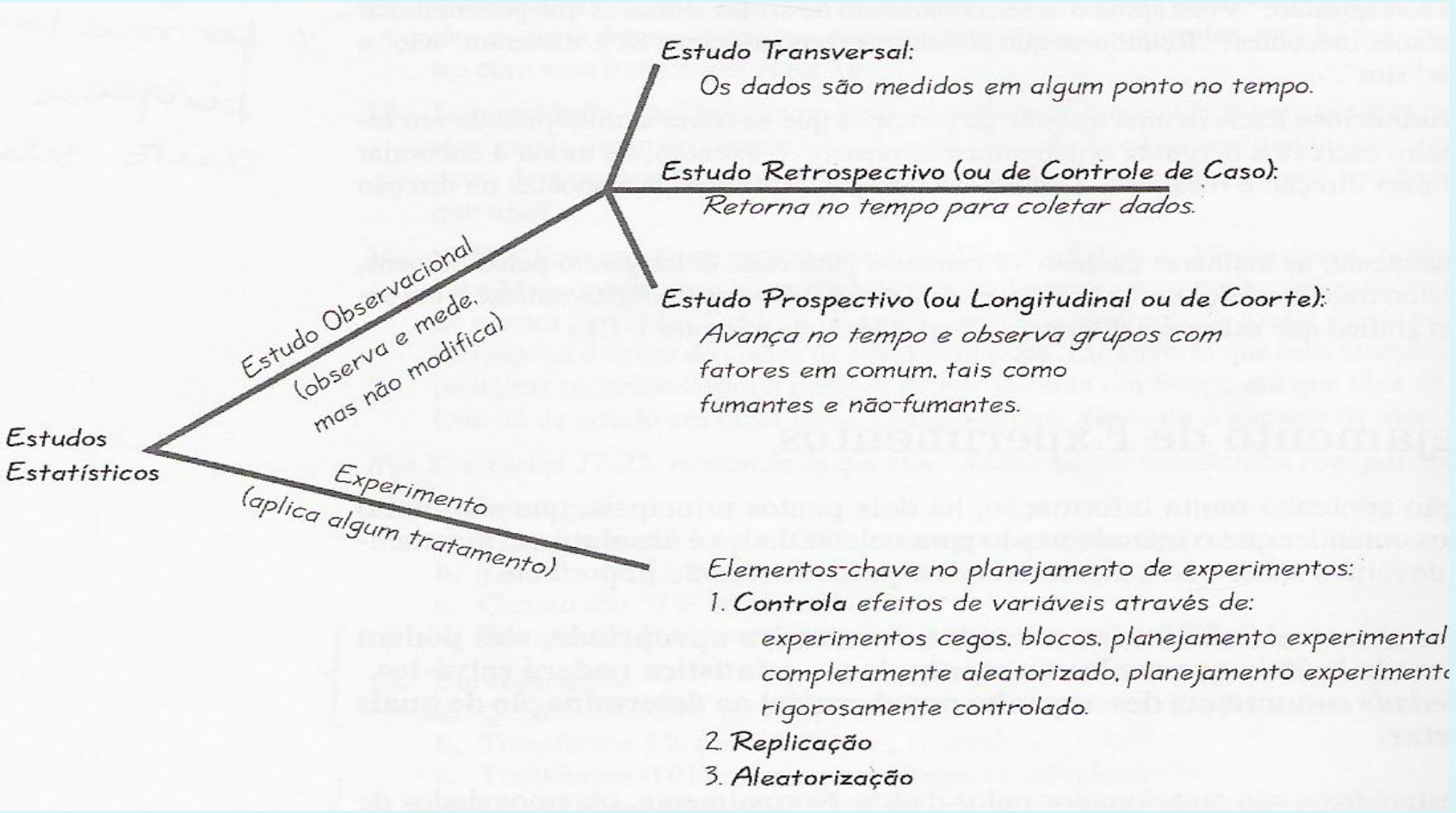
Tipos de estudo

- **Estudo observacional:** verificamos e medimos características específicas, mas não tentamos manipular ou modificar os elementos a serem estudados.
- **Estudo transversal:** dados são observados, medidos e coletados em um ponto no tempo.
- **Estudo retrospectivo ou de caso controle:** os dados são coletados do passado, voltando-se no tempo.
- **Estudo prospectivo ou longitudinal ou de coorte:** os dados são coletados no decorrer do tempo, de grupos (coortes) que compartilham fatores comuns.

Experimentos

- **Controlando os efeitos das variáveis**
  - **Experimentos cegos:** o sujeito não sabe se está recebendo o tratamento ou o placebo.
  - **Planejamento experimental completamente aleatorizado:** os sujeitos são colocados nos tratamentos através de um processo de seleção aleatória.
  - **Planejamento rigorosamente controlado:** sujeitos são escolhidos cuidadosamente de modo que em cada bloco sejam similares.

Tipos de estudos



Levantamento de dados

- **Problemas usuais - Representatividade**
  - Fator associado à forma de amostragem.
  - Na seleção da amostra procura-se reproduzir as características observáveis da população - uso do critério de **proporcionalidade**.
  - Em caso de desconhecimento da composição da população deve-se utilizar algum critério de **aleatoriedade** (sorteio).
  - **Amostra tendenciosa – conclusões sem consistência.**

Levantamento de dados

- **Problemas usuais – Fidedignidade**
  - Relacionada à precisão ou qualidade dos dados.
  - Motivos da falta de precisão:
    - Falhas nos instrumentos de aferição;
    - Problemas nos questionários empregados na obtenção dos dados;
    - Falha humana.

Levantamento de dados

- **A importância da coleta de dados**
  - Cuidado na hora de coletar informações;
    - Não adianta uma metodologia perfeita e um bom planejamento se na hora da coleta dos dados houver alguma influência do entrevistador perante o entrevistado;
    - As pessoas que são contratadas para fazer as entrevistas devem passar por um bom treinamento.

Amostragem

- Se os dados amostrais não forem coletados de maneira apropriada, eles podem ser de tal modo inúteis que nenhuma manipulação estatística poderá salvá-los.
- A aleatoriedade comumente desempenha papel crucial na determinação de quais dados coletar.

Amostragem

- **Vantagens do levantamento por amostragem:** custo menor, menor tempo e objetivos mais amplos.
- **Situações para trabalho com amostras:** população muito grande, dificuldade de acesso, grande número de variáveis.
- **Tipos**
  - Aleatória
  - Estratificada
  - Sistemática
  - Conglomerados
  - Conveniência

Distribuições de Frequências

- Relacionam categorias ou classes de valores, juntamente com contagens (ou frequência) do número de valores que se enquadram em cada categoria.

- **Exemplo: VARIÁVEL QUALITATIVA**

Indígenas por etno-região de origem, Manaus, 2007

Etno-Região	n	%
Juruá, Jutaí, Purus, Javari	51	7,35
Marau-Andirá	148	21,33
Rio Negro	315	45,39
Solimões	129	18,59
Tapajós-Madeira	38	5,48
Outras regiões	13	1,87
Total	694	100,00

Tabelas

- **Tabela de distribuição de frequência**

Considere o seguinte conjunto de dados:

21, 21, 21, 22, 22, 23, 23, 24, 25, 25, 25, 25, 26, 26, 26, 28, 30.

Construa uma distribuição com todas as frequências.

Solução:

Tabelas

X	f_i	f_{ac}	fr	far
21	3	3	3/17	3/17
22	2	5	2/17	5/17
23	2	7	2/17	7/17
24	1	8	1/17	8/17
25	4	12	4/17	12/17
26	3	15	3/17	15/17
28	1	16	1/17	16/17
30	1	17	1/17	17/17
$\sum$	17		1	

Tabelas

- Para a construção de tabelas de frequências para variáveis contínuas, os dados devem ser agrupados em intervalos de classes.

- Para a construção das classes algumas definições são necessárias:

Tabelas

- **Amplitude Total ou "Range" (R):** É a diferença entre o maior e o menor valor observado.

  Ex.: R = 30 - 21 = 9.

Tabelas

- **Intervalos de Classe:** Conjunto de observações apresentadas na forma contínua, sem superposição de intervalos, de tal modo que cada valor do conjunto de observação possa ser alocado em um, e apenas um, dos intervalos.

Tabelas

O número **k** de intervalos para cada conjunto de observações com **n** valores pode ser calculado como:

$$\mathbf{k = 1 + 3{,}322(\log_{10} n)}$$ (fórmula de Sturges)

Ex.: para um conjunto com 50 observações obtemos $\log_{10}(50) \approx \mathbf{1{,}699}$;

$$\mathbf{k = 1 + 3{,}322 \times 1{,}699 \approx 6{,}6 \approx 7 \text{ intervalos}}$$

O tamanho **w** de cada intervalo é obtido pela divisão do valor da diferença entre o maior e o menor valor, **R**, pelo número de intervalos **k**:

$$\mathbf{w = R/k}$$

Tabelas

- **Etapas para a construção de tabelas de frequência para dados agrupados:**

1) Encontrar o menor e o maior valor (mínimo e máximo) do conjunto de dados.
2) Calcular o número de classes que englobem todos os dados sem haver superposição dos intervalos.

Tabelas

3) Contar o número de elementos que pertencem a cada classe.

4) Determinar a frequência relativa de cada classe.

Tabelas

Exemplo:

O conjunto de dados abaixo representa as idades de mulheres responsáveis pelos domicílios. Construa intervalos de classes para o mesmo.

19 19 20 21 23 23 23 23 24 24 25 25 26 26 26 27 27 27 29 29 29 29 30 31 31 31 33 33 33 34 37 37 37 37 40 40 40 40 43 43 44 44 47 48 48 48 51 52 52 53

Tabelas

Solução:
se utilizar a fórmula de Sturges
R = 53 – 19 = 34 e n = 50
Então:
K = 1 + 3,322 x 1,699 ≈ 7 intervalos
W = 34/7 ≈ 5 idades em cada

Intervalo de classe	Freqüência
19 |------- 24	8
24 |------- 29	10
29 |------- 34	11
34 |------- 39	5
39 |------- 44	6
44 |------- 49	6
49 |------- 54	4

Tabelas

Ou construir intervalos empiricamente:

Intervalo de classe	Freqüência
10 |------- 20	2
20 |------- 30	20
30 |------- 40	12
40 |------- 50	12
50 |------- 60	4

Tabelas

- Os extremos dos intervalos são conhecidos como **limites de classes**.

- Procedendo-se desse modo, ao resumir os dados referentes a uma variável contínua perde-se informações.

Representação tabular

- **Apresentação de tabelas**
  - A tabela deve ser simple, clara e objetiva. Grandes volumes de dados devem ser divididos em várias tabelas.
  - A tabela deve ser auto-explicativa.
  - Nenhuma casa da tabela deve ficar em branco, apresentando sempre um número ou um símbolo.
  - As tabelas, excluídos os títulos, serão delimitadas, no alto e em baixo, por traços horizontais grossos, preferencialmente.

Representação tabular

- **Apresentação de tabelas**
  - Recomenda-se não delimitar as tabelas à direita e à esquerda, por traços verticais.
  - Será facultativo o emprego de traços verticais para a separação de colunas no corpo da tabela.
  - Deve-se manter a uniformidade quanto ao número de casas decimais.
  - Os totais e subtotais devem ser destacados.

Gráficos

- Os gráficos são representações pictóricas dos dados.

- Tem por finalidade dar uma ideia, a mais imediata possível, dos resultados obtidos, permitindo chegar-se a conclusões sobre a evolução do fenômeno ou sobre como se relacionam os valores da série.

Gráficos

- A escolha do gráfico mais apropriado ficará a critério do analista.

- Contudo, os elementos simplicidade, clareza e veracidade devem ser considerados quando da elaboração de um gráfico.

Gráficos

- **Gráficos para variáveis qualitativas**

Dentre os gráficos para representar variáveis qualitativas temos o gráfico de barras e em setores (gráfico de pizza).

Gráficos

- **Gráfico de barras:** consiste em construir retângulos ou barras, em que uma das dimensões é proporcional à magnitude a ser representada (f_i).

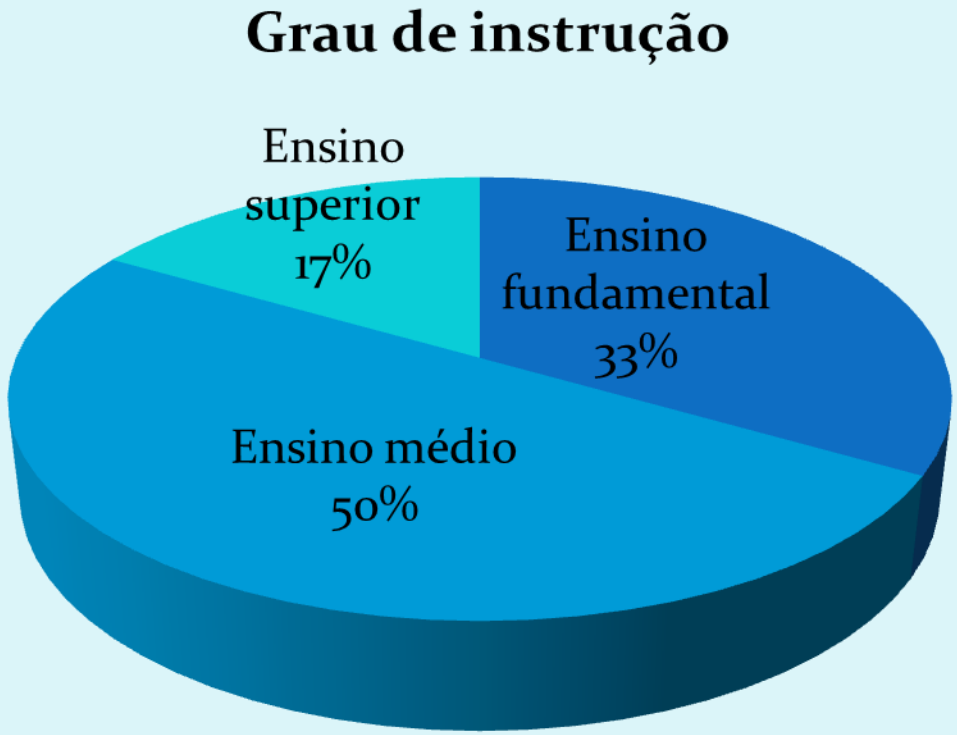
Estas barras são dispostas paralelamente umas às outras, horizontal ou verticalmente.

Gráfico

- **Gráfico de composição em setores**: Destina-se a representar a composição, usualmente em porcentagem, de partes de um todo.

Consiste num círculo de raio arbitrário, representando o todo, dividido e setores, que corresponde as partes de maneira proporcional.

Gráficos


Grau de instrução
Ensino superior 17%
Ensino fundamental 33%
Ensino médio 50%

Gráficos

- **Gráfico para variáveis quantitativas:**

Os tipos de gráficos geralmente são utilizados nesse caso: Gráfico de dispersão, Histograma, polígono de frequência e gráfico de linhas.

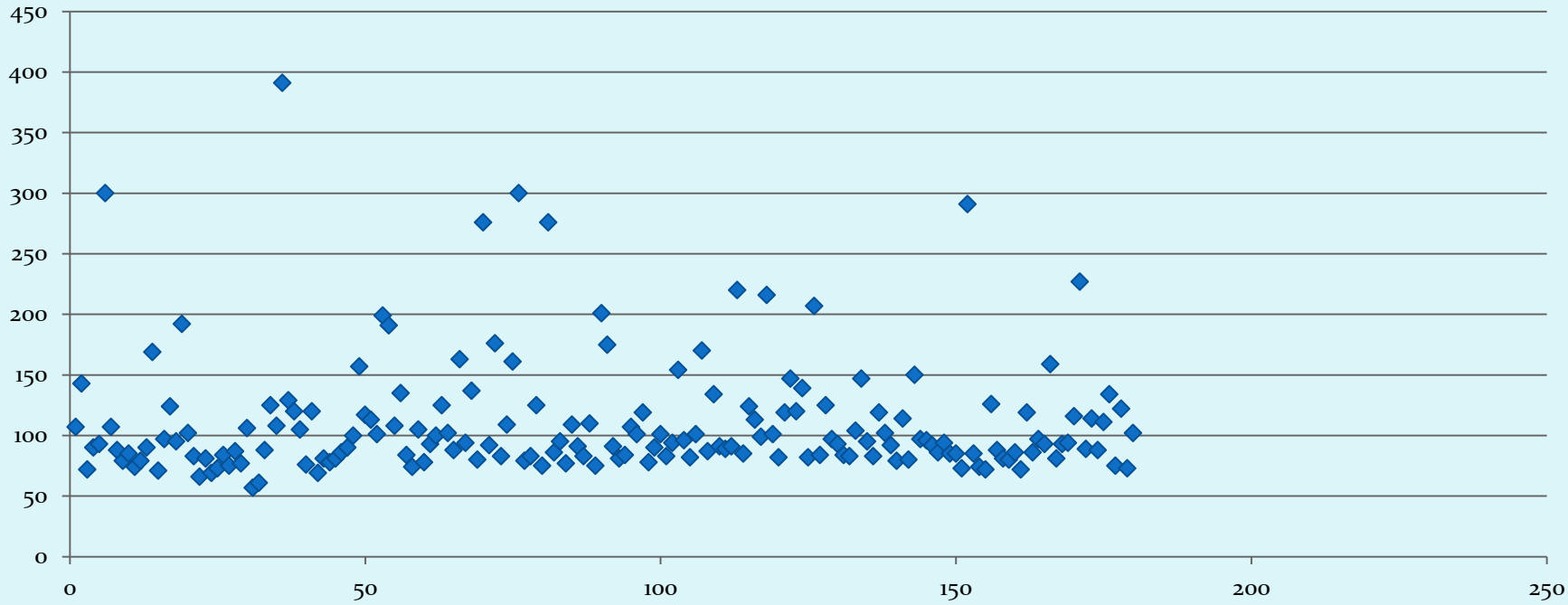
Gráficos

- **Gráfico de dispersão:**

  Os valores são representados por pontos ao longo da reta.

Exemplo: Taxa de glicemia dos idosos que procuram atendimento no Centro de Atenção Integrada da Melhor Idade – CAIMI.

Gráficos



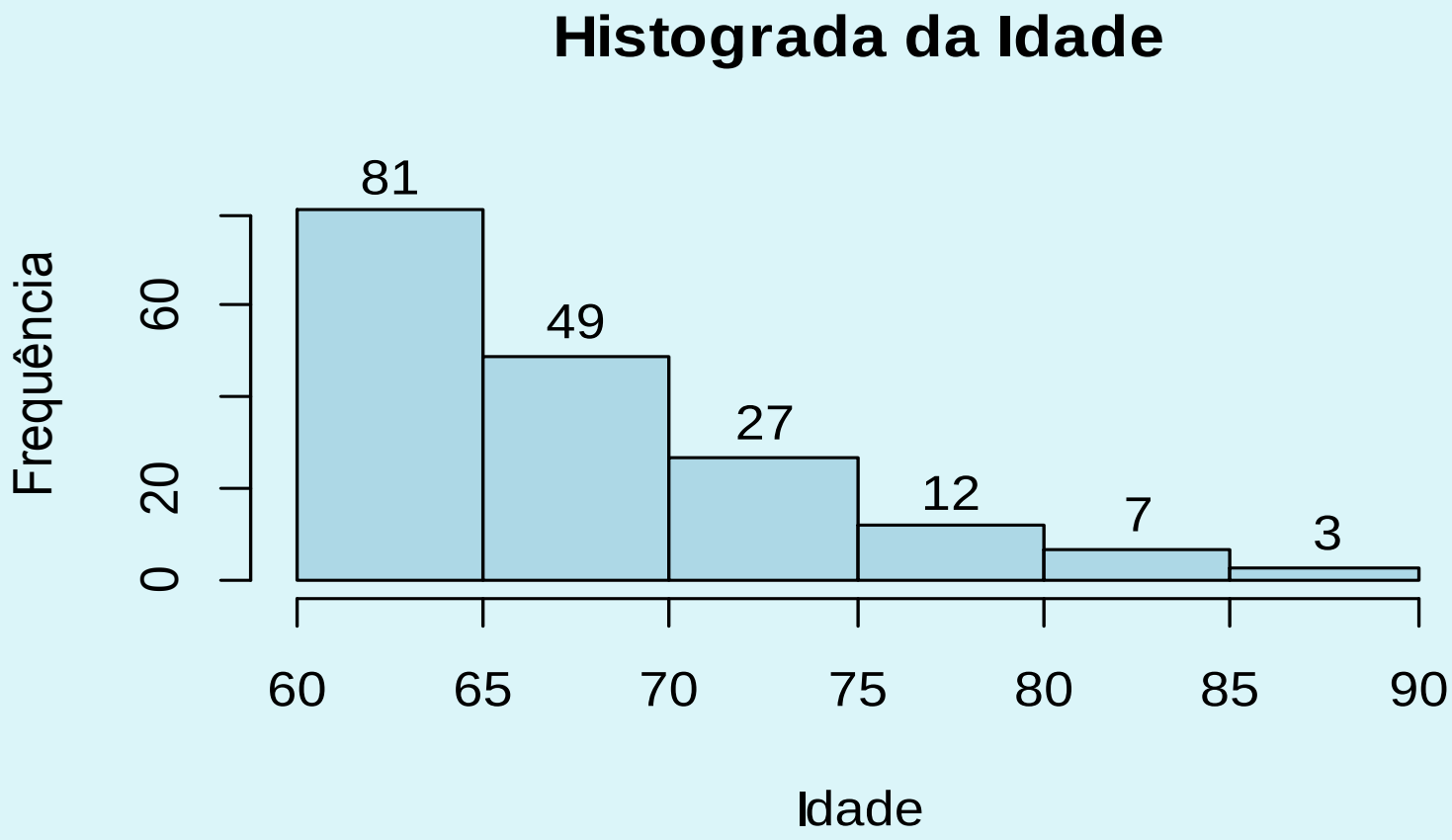
Gráficos

- **Histograma:**

É um gráfico de barras contíguas, com bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva frequência.

Exemplo: Idade dos idosos que procuram atendimento no Centro de Atenção Integrada da Melhor Idade – CAIMI.

Gráficos


Histograda da Idade
Frequência
0
20
60
81
49
27
12
7
3
60
65
70
75
80
85
90
Idade

Gráficos

- **Polígono de frequência**: É um gráfico em linha, onde as frequências são marcadas sobre perpendiculares ao eixo horizontal, levantadas pelos pontos médios dos intervalos de classe. Para conseguir um polígono, ligamos os extremos da linha obtida aos pontos médios da classe anterior à primeira e da posterior à última, da distribuição.

Gráficos

- **Gráfico de linhas:** É indicado para dados coletados ao longo do tempo, ou de medidas repetidas.

- Através desse gráfico é possível constatar algum tipo de tendência e identificar alguns eventos inusitados, como por exemplo, o surto de uma determinada doença.

Distribuições de Frequências

- **Exercício: VARIÁVEL QUANTITATIVA**

- Distribuição de frequência para dados agrupados ou tabulados em classes.

Idade dos sociólogos em anos

36	39	40	40	40
42	43	44	44	45
45	45	47	49	49
50	50	51	52	53
55	57	58	59	59

Medidas de tendência central

Valor do ponto em torno do qual os dados se distribuem

Medidas de Locação ou de Tendência Central

Média Amostral	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1+X_2+\cdots+X_n}{n}$

Exemplo: Considere os pesos de 10 recém-nascidos (em Kg)

3,3	3,1	2,8	2,7	2,9	3,1	3,2	3,0	3,5	3,4
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{3,3 + 3,1 + \cdots + 3,4}{10} = \frac{31}{10} = 3,1$$

Medidas de tendência central

Mediana Amostral ($\widetilde{X}$)

Considere agora a amostra ordenada,

$$X_{(1)}, X_{(2)}, \ldots, X_{(n)}$$

onde $X_{(1)}$ é o menor elemento e $X_{(n)}$ é o maior elemento da amostra.

A mediana amostral é definida como segue:

$$\text{Se } n \text{ é ímpar}, \quad \widetilde{X} = X_{\left(\frac{n+1}{2}\right)}$$

$$\text{Se } n \text{ é par}, \quad \widetilde{X} = \frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2}$$

Medidas de tendência central

Neste caso $n = 10$ e

$$\widetilde{X} = \frac{X_{\left(\frac{10}{2}\right)} + X_{\left(\frac{10}{2}+1\right)}}{2} = \frac{X_{(5)} + X_{(6)}}{2} = \frac{3,1 + 3,1}{2} = 3,1$$

Se a amostra tivesse tamanho $n = 11$ elementos, a mediana seria

$$\widetilde{X} = X_{\left(\frac{n+1}{2}\right)} = X_{\left(\frac{11+1}{2}\right)} = X_{\left(\frac{12}{2}\right)} = X_{(6)}$$

Medidas de tendência central

A *mediana amostral* tem uma vantagem sobre a *média amostral*: a mediana amostral é menos influenciada por observações extremas do que a média amostral.

Por exemplo, suponha que os dados na amostra são

$$1 \quad 2 \quad 3 \quad 4 \quad 6 \quad 7 \quad 8$$

Neste caso

$$\overline{X} = \frac{1+2+3+4+6+7+8}{7} = \frac{31}{7} = 4,43$$

e a mediana é $\widetilde{X} = X_{(\frac{n+1}{2})} = X_{(\frac{7+1}{2})} = X_{(4)} = 4.$

Se os dados agora são

$$1 \quad 2 \quad 3 \quad 4 \quad 2519 \quad 7 \quad 8$$

A *média amostral* é $\overline{X} = 363,43$ mas, a *mediana amostral* continua sendo $\widetilde{X} = X_{(4)} = 4.$

Medidas de tendência central

É claro que este exemplo é radical mas, ilustra bem o fato da *mediana* ser mais "robusta" ao encontrar observações discrepantes do resto da amostra.

Moda (M_0)

A *moda amostral* é simplesmente a observação mais freqüente na amostra.

Se os dados são: 1, 4, 8, 12, 5, 4, 4, 7, a moda é $M_0 = 4$, o valor que ocorreu mais vezes.

Medidas de tendência central

- Média aritmética: Cálculo da média de dados em Tabela de Distribuição de frequência

Classe	Ponto Médio	Frequência
1,5l— 2,0	1,75	3
2,0l— 2,5	2,25	16
2,5l— 3,0	2,75	31
3,0l— 3,5	3,25	34
3,5l— 4,0	3,75	11
4,0 l— 4,5	4,25	4
4,5l— 5,0	4,75	1

n=100

Média ($\overline{X}$): ponto médio de cada classe x respectiva frequência dividido pelo n

$$\overline{X} = \frac{1,75x3 + 2,25x16 + \ldots + 4,25x4 + 4,75x1}{100} = \frac{300}{100} = 3$$

Medidas de Variabilidade

Variância Amostral (S^2)

É a medida mais comum de dispersão. A *variância amostral* é definida como segue

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

A *variância populacional* será denotado por σ^2.

Nota: A *variância* (*amostral* ou *populacional*) é sempre maior ou igual a zero e é uma medida quadrática. Assim, se os dados estão em metros (m), a *variância* é expressa em metros quadrados (m^2). Isso dificulta a interpretação da variância amostral. Para contornar este problema, trabalhamos com o desvio padrão amostral, definido a seguir.

Medidas de Variabilidade

Desvio Padrão Amostral (S)

O *desvio padrão amostral* é definido como sendo a raiz quadrada positiva da *variância amostral*

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

Medidas de Variabilidade

- Medida de dispersão: indicadores do grau de variabilidade dos indivíduos em torno das medidas de tendência central
- Variância: Medir os desvios em relação a média
- Não há média dos desvios pois sua soma é igual a zero

Ex.: 0,4,6,8,7

$\overline{X}$ **(média) :** $\frac{0+4+6+8+7}{5} = \frac{25}{5} = 5$

$X - \overline{X}$ **(desvio em relação a média)**

0 – 5 = - 5

4 – 5 = -1 **A soma dos desvios é igual a zero**

6 – 5 = 1

8 – 5 = 3 **(-5 + -1)+1+3+2= - 6 + 6 = 0**

7 – 5 = 2

Medidas de Variabilidade

- Variância: Soma dos quadrados dos desvios

Dados X	Desvios $(X - \overline{X})$	Quadrado dos desvios $(X - \overline{X})^2$
0	- 5	25
4	- 1	1
6	1	1
8	3	9
7	2	4
$\overline{x} = 5$	$\Sigma (x - \overline{x}) = 0$	$\Sigma (x - \overline{x})^2 = 40$

A soma do quadrado dos desvios não é usada como medida de dispersão, porque o seu valor cresce com o nº de dados

Medidas de Variabilidade

- Variância

Então, para medir a dispersão dos dados em relação à média, usa-se a variância (S^2) que leva em consideração o ***n***

$$S^2 = \frac{\text{soma dos quadrados dos desvios}}{n - 1}$$

Para os dados: 0, 4, 6, 8 e 7 a $S^2 = \frac{40}{5-1} = \frac{40}{4} = 10$

Medidas de Variabilidade

Desvio Padrão

Raiz quadrada da variância, sendo representava por S; tem a mesma unidade de medida dos dados

Ex.: 0,4,6,8,7. S^2 (variância) = 10
s (desvio padrão): $\sqrt{10} = 3,16$

Coeficiente de variância (CV)

Razão entre o desvio padrão e a média x 100

$$CV = \frac{6}{X} \times 100$$

Ex.: Grupo I: 3,1,5 anos (x = 3 anos; $s^2 = 4$; s=2) : CV = 66,7%
Grupo II: 55,57,53 anos (x = 55 anos; $s^2 = 4$; s = 2) : CV = 3,64%

Vejam à dispersão dos dados em ambos os grupos é a mesma, mas os CV são diferentes (no grupo I a dispersão relativa é ALTA)