Autor: Prof. Me. Walter Garcia Alvares *Versão atualizada 2º sem./2018*

P R O G R A M A   C O M P L E T O   D O   C U R S O

Dados estatísticos. Definições preliminares.

Formas de apresentação dos dados estatísticos.

Representação gráfica dos resultados.

Teoria e tipos de amostragem.

Medidas de tendência central - médias, mediana e moda.

Medidas de dispersão – variância, desvios, desvio padrão e coeficiente de variação

Estudos de probabilidades.

Distribuições de Probabilidades – Normal, Binomial, Poisson e Exponencial.

Estatística - é derivada da palavra latina "status" – estado de coisas. Preliminarmente se resumia somente a coletar e à contagem de dados. Divide-se em: Estatística Descritiva e Estatística Indutiva (ou Estatística Dedutiva ou Inferência Estatística).

O primeiro fato estatístico histórico notável, e que está descrito nas Santas Escrituras, no Evangelho de Jesus Cristo segundo São Lucas – Cap. 02 - versíc. de 1 a 7, que narra: "Aconteceu que naquele tempo César Augusto publicou um decreto, ordenando o recenseamento de toda a terra. Este primeiro recenseamento foi feito quando Quirino era governador da Síria. Todos iam registrar-se na sua cidade natal. Por ser da família e da descendência de Davi, José subiu de Nazaré, na Galiléia à Judéia, à Cidade de Davi, chamada Belém, para registrar-se com sua esposa Maria, que estava grávida. Enquanto estavam em Belém, completaram-se os dias para o parto. E Maria deu à luz o seu filho primogênito Jesus. Ela o enfaixou, e o colocou na manjedoura; pois não havia lugar para eles na hospedaria". Nessa Santa Noite Feliz nasceu o menino Deus, em Belém num presépio, fato que é comemorado por toda a humanidade como o Santo Natal de Jesus.

Atualmente os países coligados a O.N.U., dentre eles o Brasil, por orientação desse órgão das Nações Unidas, todos devem fazer recenseamentos decenais (a cada 10 anos), em anos com final zero, universais (com todos os habitantes). Aconteceu um censo em 2010, deverá ser realizado outro em 2020, pelo órgão I.B.G.E. ligado ao Governo Federal, com o objetivo de colher dados dos habitantes no Brasil, com o objetivo de implementar importantes programas governamentais, tais como: construção de moradias, escolas, hospitais, usinas hidroelétricas, estradas, etc.

Modernamente a Estatística, como ramo da Matemática, está dividida em duas áreas:

Parte I - Estatística Descritiva – é a parte da Estatística que tem por objetivo descrever e analisar cientificamente uma determinada população, sem pretender tirar conclusões em caráter mais genérico sobre a mesma. É usada para esse estudo a Metodologia de Distribuição de Frequências, de acordo com o que iremos estudar na Parte I das Anotações de Estatística.

Parte II - Estatística Indutiva - é a parte da Estatística que, baseando-se em fatos reais já ocorridos anteriormente, e dos resultados obtidos através de uma amostra de uma população, procura deduzir, estimar, induzir ou inferir as Leis de comportamento dessa população, da amostra da qual ela foi retirada. Através desse método científico podemos aceitar ou rejeitar as hipóteses, que poderão surgir sobre as características da população, através de uma amostra representativa (**$n \geq 30$** elementos no mínimo), da qual ela foi retirada. Vamos estudar esses temas na Parte II das Anotações de Estatística.

Parte I - Estatística Descritiva – Definições preliminares - Coleta de dados estatísticos.

– População: é um conjunto universo, formado por todos os elementos. Representação: **N;**

– Amostra: é um subconjunto ou uma parte de uma população. Representação: **n.**

Como podemos determinar quantas pessoas numa população apresentam certa característica? Por exemplo, quantos eleitores apoiam um Presidenciável? Ou então: da população de determinado Estado, quantas pessoas são crianças, quantas vivem em centros urbanos, quantas estão desempregadas?
Uma forma de responder à essas questões, é entrevistar todas as pessoas. Mas este é um processo demorado e muito caro.
Outro método possível consiste em consultar um grupo de pessoas, que constituem uma amostra. Se essa amostra é fiel e representa de fato toda a população, podemos utilizar as características dos seus elementos para estimar as características de toda população.
Coletar dados estatísticos é trabalhoso, demanda tempo e de altos custos, por isso devemos decidir, em 1º lugar, quais os tipos de informações que devem ser pesquisadas. Se por acaso essa escolha for incorreta, as conclusões da pesquisa também serão incorretas, e vice-versa. Vamos usar nesses estudos 2 tipos de variáveis usadas na matemática: as Quantitativas e as Qualitativas.

Variáveis Quantitativas: são aquelas em que os dados podem sempre ser quantificados numericamente. **Exemplo:** Dados sobre idades das pessoas num Município do Estado de S. P.
Existem dois tipos de variáveis quantitativas: as **Discretas** e as **Contínuas.**

Variáveis Quantitativas Discretas – são aquelas que assumem somente valores inteiros (valores pertencentes ao conjunto numérico $\mathbb{Z}^+$ dos nº. inteiros), num conjunto de dados estatísticos de uma população ou de uma amostra.
Exemplos: nº de nascimentos, nº de óbitos, nº de cadeiras, nº ou um conjunto de peças, nº de veículos fabricados num dia, etc.

Variáveis Quantitativas Contínuas – são aquelas que podem assumir quaisquer valores numéricos reais (contidos no conjunto $\mathfrak{R}$ dos números reais), de um conjunto ou subconjunto populacional.
Exemplos: pesos, alturas, idades, comprimentos ou diâmetros de peças, etc.

As variáveis do tipo **qualitativas** raramente são utilizadas pelas amostras estatísticas, pois são difíceis de quantificá-las e enumerá-las de forma crescente ou decrescente.
Exemplos: o belo e a beleza; o amor de mãe; o aroma de um perfume; a cor de uma tinta; etc.
Poder-se-á quantificar essas variáveis dando pontuações observadas pelas pessoas pesquisadas

TIPOS DE AMOSTRAGEM
AMOSTRAGEM ALEATÓRIA SIMPLES (A.A.S.)
É um processo bastante fácil e muito usado. Neste processo todos os elementos da população têm igual probabilidade de serem escolhidos, do início ao fim do processo.
Método – enumerar todos os elementos da população ou de uma amostra; efetuando sucessivos sorteios, com reposição, até completar o tamanho da amostra.
O sorteio pode ser feito de várias formas, p.ex.: usando programas de computador "randômicos"; utilizando-se de urnas, ou ainda utilizando uma "tábua de números aleatórios".
A tábua de números aleatórios apresenta os dígitos de 0 a 9 dispostos aleatoriamente em linhas e colunas, e o pesquisador escolhe, ao acaso, uma linha ou coluna e a sequência que irá seguir.
Exemplo: Numa população com 1.000 indivíduos, que são numerados de 000 a 999, vamos selecionar uma amostra correspondente a 10% dessa população. Para isso escolhemos, ao acaso, uma linha ou coluna e extraímos dela a sequência de 3 dígitos. Repetimos o processo até completar o total da amostra. Se sair um nº repetido, ele será desconsiderado, e o processo continuará sucessivamente. Exemplo de uma tábua (tabela) de nºs. aleatórios:

3045	9213	5431	2549	3674	4512	1354	7895	1465	2539	0214	3502	3145
2365	0123	0547	4578	9876	5423	1245	3265	1953	1234	5789	7985	4512

Exemplo: Escolhemos iniciando-se pela linha 01, extraindo dela os 3 primeiros algarismos da 1ª coluna, depois os 3 últimos. Após iniciar na 2ª coluna e assim por diante, então teríamos: 304, 045, 921, 213, 543, 431, 254,........, até completar o total da amostra.

AMOSTRAGEM SISTEMÁTICA

É uma variação da amostragem aleatória simples, muito utilizada quando a população já está naturalmente ordenada, como p. ex., em listas de matrículas, fichários, arquivos, prontuários, etc.

Ex. do método: **N =** tamanho da população em estudo; **n =** tamanho da amostra a ser pesquisada. Exemplo: Calcular o intervalo da amostragem, através da razão: $a = \frac{N}{n}$ (sendo $\boldsymbol{a}$ igual ao nº. inteiro mais próximo). Sorteia-se um nº entre 01 e "a", formando uma amostra dos números correspondentes ao conjunto: $x;\ x+a;\ \ x+2.a;\ x+3.a;\ldots..;\ x+(n-1).a$

- Sejam: $N = 500$ e $n = 50$, logo: $a = \frac{N}{n} = \frac{500}{50} = 10 \implies a = 10$

- sorteia-se aleatoriamente um nº entre 1 e 10, por exemplo: 5, então ($x = 5$).
- os números considerados para a amostra seriam: 5, 15, 25, 35, 45, ... e assim sucessivamente.

AMOSTRAGEM ESTRATIFICADA

Muitas vezes a população se divide em subpopulações, subconjuntos ou estratos, sendo razoável supor que em cada estrato a variável de interesse na pesquisa apresenta comportamentos diversos. Por outro lado, podemos supor também que o comportamento é homogêneo dentro de cada estrato. Nesse caso é recomendável que cada estrato seja proporcional ao tamanho da população a qual representa.

Num estudo de uma população de adultos, que depende da situação civil (casados, solteiros e outros), é claro que o tamanho desses estratos é diferente.

Exemplo: sabendo-se que na faixa de idades em estudo: 70% são casados, 10% são solteiros e 20% são divorciados. Ao construirmos uma amostra de 800 pessoas, ela deverá ser dividida proporcionalmente ao estrato a que pertence cada grupo civil, e depois somando-os, vamos obter

Casados..........	800 x 70% =	560
Solteiros.........	800 x 10% =	80
Divorciados.....	800 x 20% =	160
Total...............	....	800

Tipos de levantamento de dados estatísticos:

I – **Dados Registrados**: são aqueles obtidos através de cartórios, de sacristias, dos censos populacionais, prefeituras municipais, Governos Estaduais, de Hospitais, de Pronto Socorros, etc.

II – **Dados não registrados**: são aqueles em que os dados foram omitidos, ou perdidos, ou não foram efetuados, tais como nascidos sem registros, mortes de indigentes sem documentos, etc.

III – **Dados não existentes**: são aqueles que somente podem ser obtidos através de pesquisas.

IV – **Dados Brutos** – são aqueles que ainda não foram organizados numericamente, como p. ex.: os dados colhidos recentemente através de uma pesquisa populacional.

V – **Dados em Rol** – são aqueles que já foram organizados numericamente em ordem crescente ou decrescente.

TIPOS DE AMOSTRAGENS PROBABILÍSTICAS:

Distinguiremos 2 tipos de amostragem: **a probabilística e a não-probabilística.**

A amostragem será **probabilística** se todos os elementos da população possuírem probabilidade conhecida de pertencer à sua amostra, e diferente de zero. Caso contrário, a amostragem será não probabilística. Seguindo essa definição, a amostragem probabilística implica num sorteio com regras bem determinadas, cuja realização só será possível se a população for finita e totalmente acessível.

Exemplo: Numa empresa deseja-se escolher 3 diretores entre seus chefes executivos. A escolha é aleatória e não depende do prestígio, da capacidade, dos anos de serviço, etc. Temos daí uma amostragem probabilística.

As técnicas da Estatística pressupõem que as amostras utilizadas sejam probabilísticas, o que muitas vezes não se pode conseguir. No entanto o bom senso irá indicar quando o processo de amostragem, embora não sendo probabilístico, pode ser, para efeitos práticos, considerado como tal. Isso amplia consideravelmente as possibilidades de utilização do método estatístico em geral.

A utilização de uma amostragem probabilística, é a melhor recomendação que se deve fazer no sentido de se garantir a representatividade da amostra. Pois o acaso será o único responsável por eventuais discrepâncias entre a população e a amostra. O que é levado em consideração pelos métodos de análise da Estatística Indutiva.

Apresentamos a seguir alguns modelos de **amostragens probabilísticas**:

Amostragem por Conglomerado -

A população é dividida em diferentes conglomerados (grupos), extraindo-se uma amostra apenas dos conglomerados selecionados, e não de toda a população. O ideal seria que cada conglomerado representasse tanto quanto possível o total da população. Na prática, selecionam-se os conglomerados geograficamente. Escolhem-se aleatoriamente algumas regiões, em seguida algumas sub-regiões e finalmente, alguns lares. Esse processo possibilita ao pesquisador, entrevistar grupos não muito grandes de pessoas.

Amostragem Estratificada -

Se uma população pode ser dividida em subgrupos que consistem, todos eles, em indivíduos bastante semelhantes entre si. Podemos obter uma amostra aleatória de pessoas em cada grupo. Esse processo pode gerar amostras bastante precisas, mas só é viável quando a população pode ser dividida em grupos homogêneos, ou seja: com as mesmas características.

Amostragem Aleatória Simples -

Esta amostragem é a maneira mais fácil para selecionarmos uma amostra probabilística de uma população. Inicia-se introduzindo o conceito de A.A.S. de uma população finita, para a qual temos uma listagem de todas as unidades elementares. Podemos obter uma amostra nessas condições, escrevendo cada elemento num cartão. Misturando todos numa urna e sorteando tantos cartões quantos desejarmos na amostra. Esse procedimento torna-se inviável quando a população é muito grande. Nesse caso, usa-se um processo alternativo, no qual os elementos são numerados, e em seguida sorteados por meio de uma tabela de números aleatórios. Utilizando-se um procedimento aleatório, sorteia-se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de serem selecionados. Repete-se o procedimento até que sejam sorteadas todas as unidades da amostra.

Podemos ter uma A.A.S. com reposição, se for permitido que uma unidade possa ser sorteada mais de uma vez, e sem reposição, se esta unidade sorteada for removida da população.

Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Contudo, a amostragem com reposição conduz a um tratamento teórico mais simples, pois ela implica que tenhamos independência entre as unidades selecionadas. Essa independência facilita o desenvolvimento das propriedades dos estimadores que serão considerados. Se a população for infinita então as retiradas com e sem reposição serão

equivalentes, isto é, se a população for infinita (ou então se for muito grande), o fato de se recolocar o elemento retirado de volta na população, não vai afetar em nada a probabilidade de extração do elemento seguinte.
Se, no entanto, a população for finita (pequena), será necessário fazer uma distinção entre os dois procedimentos, pois na extração com reposição as diversas retiradas serão independentes, mas no processo sem reposição haverá dependência entre as retiradas, isto é, o fato de não recolocar o elemento retirado afeta a probabilidade de o elemento seguinte ser retirado. A amostragem sem reposição é mais eficiente do que a amostragem com reposição, e reduz a variabilidade uma vez que não é possível retirar elementos extremos mais do que uma vez.

Amostragem Sistemática -
Quando os elementos da população se apresentam ordenados, e a retirada dos elementos da amostra é feita periodicamente, temos uma amostragem sistemática. Assim por exemplo: numa linha de produção, podemos, a cada 10 itens produzidos, retirar um para pertencer a uma amostra sistemática da produção diária.

Amostras **não-probabilísticas** são também, muitas vezes, empregadas em trabalhos estatísticos, por simplicidade ou por impossibilidade de se obterem amostras probabilísticas, como seria desejável. No entanto processos não-probabilísticos de amostragem têm também sua importância. Sua utilização, entretanto, deve ser feita com cuidado e criteriosamente.
Apresentamos a seguir, alguns modelos de amostragens não-probabilísticas:

Inacessibilidade a toda população -
Esta situação ocorre com muita frequência na prática. Por exemplo, seja a população que nos interessa constituída de todas as peças produzidas por certa máquina. Ora, mesmo estando a máquina em funcionamento normal, existe uma parte da população que é formada pelas peças que ainda vão ser produzidas. Ou então se nos interessar a população de todos os portadores de febre tifoide, estaremos diante de um caso semelhante. Deve-se notar que, em geral, estudos realizados com base nos elementos da população amostrada terão, na verdade, seu interesse de aplicação voltado para os elementos restantes da população. Este caso de amostragem não-probabilística pode ocorrer também quando, embora se tenha a possibilidade de atingir toda a população, retiramos a amostra de uma parte que seja prontamente acessível. Assim, se fôssemos recolher uma amostra de um monte de minério, poderíamos por simplificação retirar a amostra de uma camada próxima da superfície do monte, pois o acesso as porções interiores seria problemático.

Amostragem a esmo -
É a amostragem em que o amostrador, para simplificar o processo, procura ser aleatório sem, no entanto, realizar propriamente o sorteio usando algum dispositivo aleatório confiável. Por exemplo, se desejarmos retirar uma amostra de 100 parafusos de uma caixa contendo 10.000, evidentemente não faremos uma AAS, pois seria muito trabalhosa, mas retiramos simplesmente a esmo.
Os resultados da amostragem a esmo são, em geral, equivalentes aos da amostragem probabilística se a população é homogênea e se não existe a possibilidade de o amostrador ser inconscientemente influenciado por alguma característica dos elementos da população.

Amostragens intencionais -
Enquadram-se aqui os diversos casos em que o amostrador deliberadamente escolhe certos elementos para pertencer à amostra, por julgar tais elementos bem representativos. O perigo desse tipo de amostragem é grande, pois o amostrador pode facilmente se enganar em seu pré-julgamento.

Amostragem por voluntários -
Ocorre, por exemplo, no caso da aplicação experimental de uma nova droga em pacientes, quando a ética obriga que haja concordância dos escolhido.

TRABALHO PRÁTICO 1 (T.P. 1): *"APRENDIZADO DIRIGIDO DE ESTATÍSTICA"*

1 – O que é População ? Dê exemplos.

2 – O que é Amostra ? Dê exemplos.

3 – O que é Amostragem aleatória simples ? Dê exemplos.

4 - O que é Amostragem estratificada ? Dê exemplos.

5 - O que é Amostragem sistemática ? Dê exemplos.

6 - O que é Amostragem probabilística e não-probabilística ? Dê exemplos.

7 - O que é Amostragem por conglomerados? Dê exemplos.

8 - O que é Amostragem a esmo? Dê exemplos.

9 - O que é Amostragem por voluntários? Dê exemplos.

10 – O que são parâmetros? Dê exemplos.

11 – O que é estimação? Dê exemplos.

12 - Quais as formas de apresentação dos dados estatísticos? Dê exemplos.

2 - Formas de apresentação dos dados estatísticos:

2.1 – Tabular ou por Tabelas: vide exemplo abaixo de uma tabela de distribuição de frequências

ESTADO CIVIL	Nº de Pessoas ($\mathbf{f_i}$) (frequência bruta)	FREQUÊNCIA (relativa) (fr_i)	FREQUÊNCIA fr_i (%) (relativa porcentual)
C A S A D O S	16	0,2963	29,63
S O L T E I R O S	30	0,5556	55,56
DIVORCIADOS	08	0,1481	14,81
T O T A L ($\boldsymbol{\Sigma}$)	54	1,0000	100,00

Fonte: Cartórios de registro civil

2.2 – Gráficos:

2.2.1 - Gráfico de setores (disco de pizza). Exemplo:

1) Casados: 100 % ———— 360º $\alpha \cong 107°$

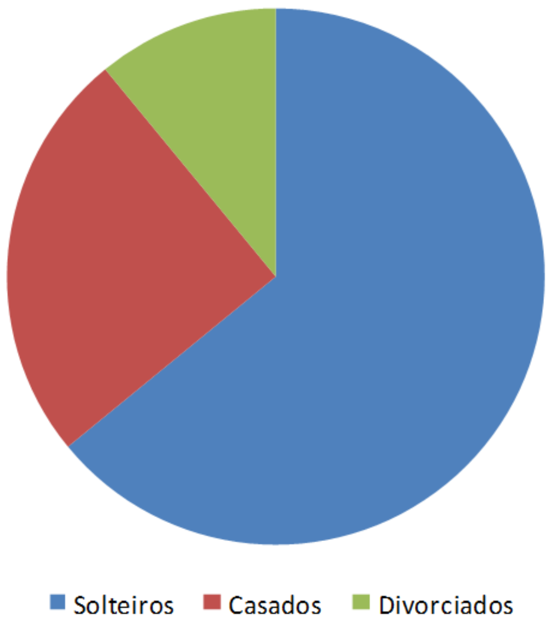
29,63 % ———— α

2) Solteiros: 100 % ———— 360º $\beta \cong 200°$

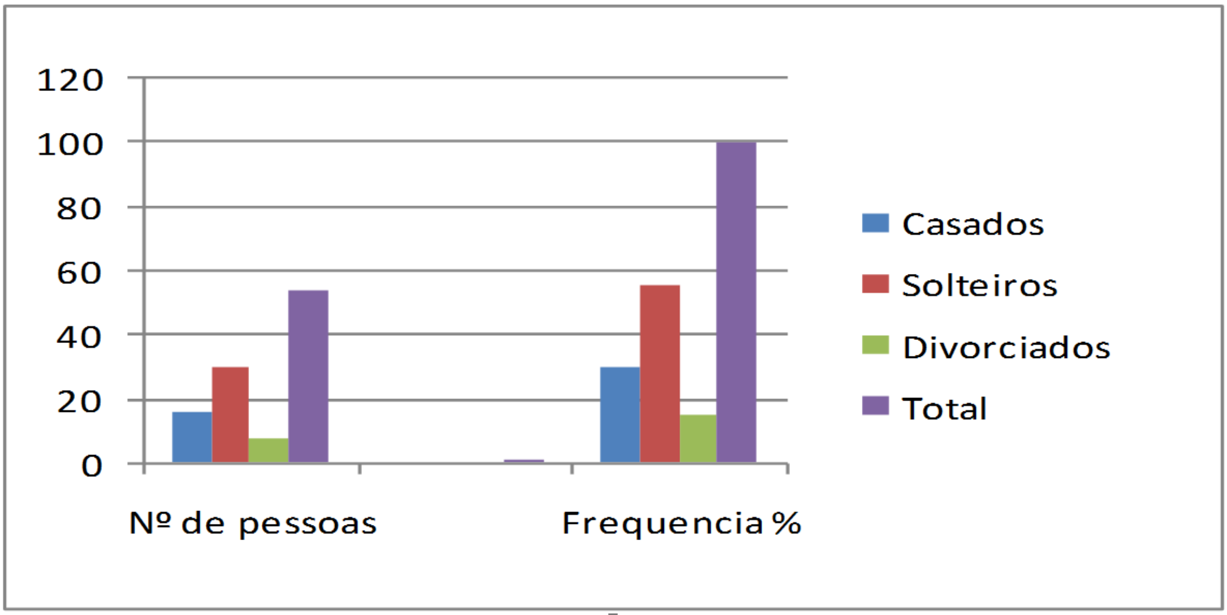
55,56 % ———— β

3) Divorciados: 100 % ———— 360° $\delta \cong 53°$

14,81 % ———— δ



2.2.2 – Gráficos de Barras, de Retângulos ou de Colunas:



O método utilizado pela Estatística Descritiva é o de distribuir frequências de variáveis quantitativas por grupos, faixas, intervalos ou classes de valores, que são intervalos usados na distribuição.

Exemplo: Idade infantil: 0 a 12 anos;

Idade da adolescência (“teen age”): 13 a 19 anos;

Idade adulta: 20 a 50 anos;

Terceira idade: > 50 anos.

O nº de elementos de cada classe ou intervalo é a frequência absoluta (bruta) da classe ($\mathbf{f_i}$).

O agrupamento dos dados por classes de valores fornece a distribuição de frequências das variáveis aleatórias quantitativas (aquelas que não sabemos previamente o seu resultado). A organização dos dados por classes traz vantagens no estudo desses dados, facilitando sua numeração e seus cálculos posteriores, entretanto podem eliminar, por vezes, detalhes originais da pesquisa que foi realizada. Todas as classes devem ter sempre intervalos de classes "***h***" iguais.

Exemplo prático: Dada a coleta, organizada em rol, de 40 pessoas adultas hipertensas, em relação as suas idades. Esse levantamento foi baseado numa pesquisa feita para defesa de uma tese de mestrado na Faculdade de Medicina da U.S.P., em 1.989.

22	27	30	31	34	35	35	35	36	36
36	36	36	40	42	43	44	44	45	46
46	47	48	50	50	53	53	53	56	58
59	59	60	63	63	65	65	65	67	**67**

Vamos considerar as seguintes nomenclaturas:

Número total de dados da amostra: $\boldsymbol{n} = 40$ pessoas;

Limite superior de todas as classes: $\boldsymbol{L_s} = 67$ anos;

Limite inferior de todas as classes: $\boldsymbol{L_i} = 22$ anos;

Amplitude total dos dados da amostra: $\boldsymbol{A_t} = \boldsymbol{L_s} - \boldsymbol{L_i} = 67 - 22 = 45$ anos;

Limite superior de uma determinada classe. Ex.: $\boldsymbol{l_3} = 46$ anos (limite superior da 3ª classe).

Limite inferior de uma determinada classe. Ex.: $\boldsymbol{l_5} = 54$ anos (limite inferior da 5ª classe).

Ponto médio de uma classe = $\frac{(l_s+l_i)}{2}$ **Ex.:** O ponto médio da 4ª classe = $\frac{54+46}{2}$ = 50 anos.

Regra de "Sturges" para determinação do número de intervalos de classes:

$$\boldsymbol{K \cong 1 + 3{,}322 \cdot \log_{10}(n)} \quad \text{ou:} \quad \boldsymbol{K \cong \sqrt{n}}$$

No exemplo dado acima, vamos calcular ***"K"*** para ***n*** = 40, e iremos obter:

$K \cong 1 + 3{,}322 \cdot \log_{10}(40) = 1 + 3{,}322 \cdot 1{,}602 = 6{,}322 \cong 6\ intervalos$ (arredado p/menos)

Formula para determinação da amplitude "*h*" dos intervalos de classes:

$$\boldsymbol{h = \frac{A_t}{K}}$$

No exemplo dado acima, vamos obter:

$$h = \frac{45}{6} = 7{,}5 \cong 8\ anos \text{ (arredado p/mais)}$$

3 - <u>MEDIDAS</u> <u>DE</u> <u>TENDÊNCIA</u> <u>CENTRAL</u> (ou Medidas de **Posição**)

3.1 – <u>Média</u> ($\overline{X}$): é a medida que caracteriza um grupo de pessoas, objetos ou genericamente um conjunto de dados estatísticos em estudo. A média de idades de um grupo de alunos dos 1º. ciclos dos cursos da FATEC/RL é diferente de outro grupo de alunos dos 6º ciclos; a estatura média das pessoas da região Nordeste do Brasil, é menor do que a altura média das pessoas nascidas na Alemanha; o salário mínimo médio na Suíça é maior do que o salário mínimo médio em Moçambique. A média possui a característica de se concentrar no centro (ou no meio) da distribuição de valores, ou seja: entre o menor e o maior valor da distribuição. Por isso é chamada de medida de tendência central. A média pode ser calculada por:

$$\boldsymbol{\overline{X} = \frac{(x_1 + x_2 + x_3 + \ldots + x_i + \ldots + x_n)}{n}}$$

simplificando, e usando símbolos matemáticos, podemos escrever:

$$\boldsymbol{\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}}$$

Quando temos um grupo de **n** dados estatísticos organizados em rol; podemos usar a formula dada abaixo, no caso dos dados sejam agrupados por frequências:

$$\boldsymbol{\overline{X} = \frac{(x_1 \cdot f_1 + x_2 \cdot f_2 + x_3 \cdot f_3 + \ldots + x_i \cdot f_i + \ldots + x_n \cdot f_n)}{\sum f_i}}$$

simplificando, adotando símbolos matempaticoa podemos escrever: $\boldsymbol{\overline{X} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i}}$

No exemplo dado na pág. 8 anterior, a média das 40 pessoas adultas hipertensas é calculada por:

$$\bar{X} = \frac{26 \cdot 2 + 34 \cdot 11 + 42 \cdot 6 + 50 \cdot 9 + 58 \cdot 5 + 66 \cdot 7}{40}$$

$$\bar{X} = \frac{52 + 374 + 252 + 450 + 290 + 462}{40} = \frac{1.880}{40} \; ; \quad \bar{X} = 47 \text{ anos}$$

3.2 - Cálculo da Média pelo Processo Breve ou Abreviado:

Fórmula para cálculo da média $\bar{X}$ pelo processo breve:

$$\bar{X} = x_0 + \frac{(\sum \alpha \cdot f_i) \cdot h}{\sum f_i}$$

onde: $\alpha = 0$,deve ser localizado na classe da moda ou na classe da mediana (a de maior frequência);

x_o = valor de x_i , onde: $\alpha = 0$ e h = amplitude de classes.

Exercício: Dada a tabela de distribuição de frequências por classes de valores de alturas de 30 pessoas, dados na tabela da pág. 11 - seguinte, calcular a média pelo processo breve ou abreviado.

3.3 - Média Geométrica (M_g): é definida como a raiz n-ésima do produto dos n valores amostrais obtidos numa pesquisa. Sua fórmula é:

$$M_g = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \ldots \cdot x_n}$$

Exemplo: Calcular a média geométrica M_g para os valores: 3; 5; 8; 11 e 16.

$$M_g = \sqrt[5]{3 \cdot 5 \cdot 8 \cdot 11 \cdot 16} = \sqrt[5]{21120} = (21120)^{\frac{1}{5}} = (21120)^{0,2} \cong 7{,}327$$

Quando o conjunto de dados possuir um número n muito elevado de valores, para facilitar os cálculos, vamos usar o logaritmo decimal – $\log n$ (ou o logaritmo natural ou neperiano – $\ln n$) dos valores dados. Para isso usaremos solidariamente as seguintes formulas:

$$\log M_g = \frac{\sum (f_i \cdot \log x_i)}{\sum f_i} \qquad \text{e} \qquad M_g = 10^{\log M_g} \; (\text{antilog de } M_g)$$

Exercício: Calcular a média geométrica M_g para os valores de alturas de 30 pessoas, dados na tabela do T.P.2 da pag. 10 – seguinte.

3.4 – Média Harmônica (M_h): é definida como o inverso da média aritmética dos inversos dos valores amostrais obtidos numa pesquisa estatística. Sua fórmula é:

$$M_h = \frac{\sum f_i}{\sum \left(\frac{f_i}{x_i}\right)}$$

Exercício: Calcular a média harmônica M_h, para os valores de alturas (em cm.) de 30 pessoas dados na tabela do T.P. 2 abaixo.

3.5 – Média Quadrática ($\overline{M_q}$): é definida como a raiz quadrada da média aritmética do quadrado dos dados amostrais obtidos numa pesquisa. Sua fórmula é:

$$\overline{M_q} = \sqrt{\frac{\sum x_i^2 \cdot f_i}{\sum f_i}}$$

Exercício: Calcular a média quadrática M_q para os valores de alturas (em cm.) de 30 pessoas dados na tabela do T.P. 2 abaixo.

Propriedade: As médias calculadas anteriormente possuem valores crescentes, como abaixo:

$$\overline{M_q} > \overline{X} > \overline{M_g} > \overline{M_h}$$

Trabalho Prático 2 (**T.P. 2**) – **1)** Calcular todas as médias $\bar{X}$ pelos processos já ensinados, usando a tabela dada abaixo:

i	Alturas (cm.)	Nº de pessoas (f_i)	x_i						
1	100 [----------110	01							
2	110 [----------120	04							
3	120 [----------130	06							
4	130 [----------140	08							
5	140 [----------150	05							
6	150 [----------160	04							
7	160 [----------170	02							
Σ	Total	30							

2) Organize os dados a seguir em uma tabela de distribuição de frequências, contendo o número e o intervalo de classes, a frequência absoluta, a frequência acumulada crescente, a frequência relativa (em %), a frequência acumulada crescente (em %) e a frequência acumulada decrescente.

20,4	22,3	26,2	27,7	23,8	24,1	27,3	25,3	27,1
26,0	25,0	25,1	25,3	24,3	25,4	28,0	25,7	25,8
26,6	26,1	28,3	23,1	26,3	26,5	26,0	26,7	26,8
27,1	24,6	24,3	25,7	25,3	27,9	25,6	26,2	28,7

3.6 – Moda (M_o): é definida como o(s) valor(es) que ocorre(m) com mais frequência(s) num conjunto de dados estatísticos amostrais. A Moda pode não existir, e se existe, pode não ser única. Neste caso denominamos a distribuição de valores de: "unimodal" e de "bimodal".

Exemplos:

I. A Moda do conjunto: *2, 3, 3, 5, 5, 5, 7, 7, 8 e 9* é igual a **5** (pois o dígito **5** ocorre **3** vezes).

II. Os dados: *13, 15, 17, 18, 18, 18, 18, 21, 21, 21, 21, 23, 23, e 25* possuem Modas **18** e **21** (distribuição de valores "bimodal"), pois os dígitos 18 e 21 ocorrem **4** vezes.

III. Os valores: 3, *5, 5, 7, 7, 8, 9, 9, 10, 13, 13, 15, 16, 16, 18* **não** possuem Moda, pois não há um valor que se destaca com maiores frequências ou maior nº de repetições.

Para um conjunto de valores distribuídos por classes, a Moda é calculada através do:

Processo ou Fórmula de "*Czuber*" para cálculo da Moda : $M_o = l_i + \left[\left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) \cdot h\right]$

onde: l_i = limite inferior da classe modal (aquela que possui a maior frequência);
Δ_1 = diferença entre a frequência da classe modal e a da classe imediatamente anterior;
Δ_2 = diferença entre a frequência da classe modal e a da classe imediatamente posterior;
h = amplitude de classes.

No exemplo anterior, dado na pag. 8 dos 40 indivíduos hipertensos, vamos obter a Moda por:

$l_i = 30$ anos; $\Delta_1 = 11 - 2 = 9$ anos; $\Delta_2 = 11 - 6 = 5$ anos; $h = 8$ anos; portanto:

$$\boldsymbol{M_o} = 30 + \left[\left(\frac{9}{9+5}\right) \cdot 8\right] \cong \mathbf{35,14\ anos}$$

3.7 – Mediana ($\boldsymbol{M_e}$): é o valor central **ou** a média dos dois valores centrais de um conjunto de dados organizados em rol.

Exemplos:

I. A distribuição: 1, 3, 4, 7, **9**, 12, 15, 18 e 21, possui Mediana igual a **9** (valor central);

II. Os valores: 19, 21, 24, **26, 28**, 30, 33 e 34, possui Mediana = $\frac{26+28}{2}$ = **27** (média dos 2 valores centrais).

3.7.1 – Cálculo da Posição da Mediana:

a) **se *n* é ímpar:** a mediana ocupa a posição do: $\left(\frac{n+1}{2}\right)^{\underline{o}}$ lugar;

b) **se *n* é par:** a mediana ocupa a posição entre o: $\left(\frac{n}{2}\right)^{\underline{o}}$ e o $\left(\frac{n+2}{2}\right)^{\underline{o}}$ lugares.

Para um conjunto de valores distribuídos por classes, a Mediana pode ser calculada através da formula:

$$\boldsymbol{M_e} = \boldsymbol{l_i} + \left\{\left[\frac{\left(\frac{N}{2}\right) - f_{i_{ac.CANT}}}{f_{i_{C.Me}}}\right] \cdot \boldsymbol{h}\right\}$$

onde: $l_i =$ limite inferior da classe que contem a mediana;

$\frac{N}{2} =$ metade do número de dados da distribuição;

$f_{i_{ac.CANT}} =$ frequência acumulada crescente até a classe anterior a da classe da mediana;

$f_{i_{C.Me}} =$ frequência da classe que contem a mediana; e $h =$ amplitude de classes.

No exemplo dado na pág. 8 dos 40 indivíduos hipertensos, vamos obter a Mediana por:

$l_i = 46$ anos; $\frac{N}{2} = 20$ pessoas; $f_{i_{ac.CANT}} = 19$; $f_{i_{C.Me}} = 9$; $h = 8$ anos, portanto:

$$\boldsymbol{Me} = 46 + \left\{\left[\frac{(20-19)}{9}\right] \cdot 8\right\} \cong \mathbf{46,89\ anos}$$

4 – Medidas de Dispersão (ou de **Espalhamento**):

Apesar da média, moda e mediana fornecerem uma idéia do comportamento das variáveis em estudo, elas podem camuflar ou esconder informações importantes a respeito da distribuição desses dados, ou seja: essas medidas podem não ser suficientes para descrever e/ou discriminar detalhes de diferentes conjuntos de dados. Tomemos por exemplo, o conjunto de idades: 0, 20 e 40 anos; se calcularmos a sua média, vamos facilmente obter 20 anos. Também o conjunto: 20, 20 e 20 anos possui média igual a 20 anos. É evidente que esses 2 conjuntos possuem valores bem distintos, apesar de possuírem a mesma média. No 1º caso a variação (ou dispersão) entre a primeira idade e a terceira é grande, ou seja: 40 anos. No 2º caso não há diferença entre as idades limites do intervalo (entre a 1ª º e a última), ou seja: não há qualquer variação entre as idades.

Em muitas dessas situações existem diferenças em relação ao espalhamento ou dispersão dos dados, isto significa que devemos considerar a maneira como os valores de cada conjunto estão variando (ou dispersos). Para analisar essas situações a Estatística dispõe de medidas para cálculo dessas variações, chamadas de medidas de dispersão ou de variabilidade, sendo as mais usuais: a Amplitude Total; o Desvio; o Desvio Médio; a Variância; o Desvio Padrão; a Amplitude entre quartis, etc.

4.1 – Desvio (ou Afastamento): é a diferença entre cada valor da distribuição x_i e a sua média $\overline{x}$:

$$d_i = x_i - \overline{x}$$

4.2 – Desvio Médio: é a média aritmética dos desvios (ou afastamentos) em módulo (valor absoluto). Sua fórmula é:

$$dm = \frac{\sum [\, x_i - \overline{x} \,] \cdot f_i}{\sum f_i}$$

No ex. dado na pág. 8 anterior dos 40 indivíduos hipertensos, o desvio médio é calculado por:

$$dm = \frac{[26-47]\cdot 2 + [34-47]\cdot 11 + [42-47]\cdot 6 + [50-47]\cdot 9 + [58-47]\cdot 5 + [66-47].7}{40}$$

$$dm = \frac{430}{40} = 10{,}75 \text{ anos}$$

4.3 – Variância: é definida como a média dos quadrados dos desvios, e é representada por: σ^2, para o caso de uma população $(n > 30)$; e por: S^2, para o caso de uma amostra $(n \geq 30)$. Sua formula é:

$$\sigma^2 = S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i}{\sum_{i=1}^{n} f_i}$$

No ex. dado na pág. 8 anterior das 40 pessoas hipertensas, a Variância é calculada por:

$$\sigma^2 = S^2 = \frac{6.104}{40} \cong \mathbf{152{,}6\ anos^2}$$

4.4 – Desvio Padrão: é a raiz quadrada da Variância, ou seja:

$$\sigma = S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot f_i}{\sum_{i=1}^{n} f_i}}$$

portanto:

$$S = \sqrt{\text{Variância}}$$

O desvio padrão é uma medida de variação padronizada pela Estatística, e que mede um espalhamento em relação à média, e possui algumas propriedades importantes que serão abordadas oportunamente, quando for estudada a Distribuição Normal (ou de Gauss) - Parte II das Anotações de Estatística Aplicada, no 2º bimestre deste semestre letivo.

No ex. da pág. 8 – anterior, o Desvio Padrão é calculado por: $\sigma = S = \sqrt{152{,}56} \cong \mathbf{12{,}35\ anos}$

Observe que o desvio padrão assume novamente a unidade original da distribuição de valores.

4.5 – Coeficiente de Variação: dá a medida de variação percentual do desvio padrão em relação a média da distribuição. Sua fórmula é:

$$C.V. = \left(\frac{S}{\overline{x}}\right).\, 100\ \%$$

(atenção a resposta será sempre em porcentagem)

No exemplo da pág. 8 - anterior, o Coeficiente de Variação é calculado por:

$$C.V. = \left(\left({}^{12{,}35}/_{47}\right).\, 100\right)\% \cong \mathbf{26{,}28\ \%}$$

Trabalho Prático 3 (**T.P. 3**)

1) Dada a tabela abaixo dos dados de uma amostra de observação etária de 60 pessoas adultas com idades variando de 55 até 83 anos, pede-se:

55	61	57	77	62	75	63	55	64	60
70	57	61	57	67	62	69	67	68	59
65	72	65	61	68	73	65	62	75	80
76	61	69	76	72	57	75	68	83	55
69	64	66	74	65	76	65	58	65	64
75	60	65	80	66	80	68	55	66	71

a) Organizar a tabela em rol crescente;
b) Construir uma tabela de distribuição de frequências, condensando os valores etários apresentados. Utilizar a "Regra de Sturges" no cálculo de "K" (para determinação do número de intervalos de classes) e de "h" (para determinação da amplitude dos intervalos de classes), citados na fl. 09 das Anotações de Estatística – Parte I;
c) Determine as frequências absolutas de cada intervalo de classe;
d) Determine as frequências relativas porcentuais de cada classe (2 casas decimais);
e) Determine as frequências acumuladas de cada classe;
f) Determine as frequências acumuladas percentuais de cada classe (4 casas decimais);
g) Determine o ponto médio de cada classe;
h) Calcular as médias: aritmética (pelos 2 processos); geométrica; harmônica e quadrática dos dados agrupados em classes (arredondas as respostas para o nº inteiro mais próximo, usando o critério de arredondamento matemático);
i) Calcular o desvio médio, a moda e a mediana dos dados agrupados;
j) Calcular a Variância, o Desvio padrão e o Coeficiente de variabilidade porcentual;
k) Esboçar um Histrograma das Frequências Absolutas em relação as classes.

2) Uma pesquisa com 526 pessoas de ambos os sexos, em relação a pigmentação da pele, revelou que 264 eram de pele branca, 132 possuíam pele parda, 95 eram de pele do tipo caucasiana e 35 eram de pele negra. Organizar esses dados numa tabela onde devem ser apresentadas as frequências relativas com 4 casas decimais, e as frequências percentuais com 2 casas decimais. Esboçar um Gráfico de Setores dos tipos de pele humana dessa pesquisa.

Nesta parte encerra-se e Parte I da matéria a ser lecionada neste 1º bimestre letivo.

A seguir realizaremos em sala de aula o **Trabalho Prático 4**, que resume toda a matéria que foi lecionada nesta "Parte I das Anotações de Estatística Aplicada":

Disciplina: Estatística Aplicada	Curso:	T.P. 4	Apresentação: data da P.1	Turma
Nome: ____________________ Nº. __________				

Trabalho Prático para ser realizado em Sala de Aula: A tabela abaixo apresenta o número de questões certas de uma Prova de Estatística realizada por 84 candidatos a cargos para o Serviço Público Estadual em São Paulo:

32	13	14	20	23	21	22	12	30	20	16	25
21	29	17	25	21	21	15	22	19	22	22	19
28	15	19	27	26	13	23	23	29	26	30	15
36	19	29	12	22	30	21	27	22	33	15	17
27	21	23	32	12	22	22	22	22	23	38	38
19	19	34	24	17	37	31	16	18	21	40	24
20	34	17	40	18	36	29	21	24	31	35	16

Organizar os dados em rol, calcular o valor de ***h*** e ***K*** (usando a Regra de "Sturges"), construir a tabela de distribuição de frequências por classes, e em seguida pede-se calcular:

1) as frequências brutas ou absolutas de cada classe;
2) as frequências relativas porcentuais de cada classe;
3) os valores médios de cada classe (x_i);
4) as frequências acumuladas crescentes;
6) as frequências acumuladas crescentes percentuais;
7) as frequências acumuladas decrescentes;
8) as frequências acumuladas decrescentes percentuais;
9) a média da distribuição pelos 2 processos (aritmético e abreviado);
10) o desvio médio;
11) a média geométrica;
12) a média quadrática;
13) a média harmônica;
14) a moda;
15) a mediana;
16) a variância;
17) o desvio padrão; e
18) o coeficiente de variação porcentual.