Material Didático

Série

Estatística Básica

Estatística Descritiva

1 2 3 4 5
6 7 8 9
10 11 12
13 14 15
16 17 18
19 20

Enfoque:
Sociais

Prof. Lori Viali, Dr.

SUMÁRIO

ESTATÍSTICA DESCRITIVA

1. 1. GENERALIDADES

1.1. INTRODUÇÃO

Por onde quer que se olhe ou escute uma coleção de números são normalmente enunciados como **estatísticas**. Estes números referem-se aos mais diversos campos de atividades: esportes, economia, finanças, etc. Assim tem-se por exemplo:

* O número de carros vendidos no país aumentou em 30%.
* A taxa de desemprego atinge, hoje, 7,5%.
* As ações da Telebrás subiram R$ 1,5, hoje.
* Resultados do Carnaval no trânsito: 145 mortos, 2430 feridos.

Um número é denominado uma **estatística** (singular). No fechamento da bolsa as ações da Vale foram cotadas a R$ 45.50. As vendas de uma empresa no mês constituem uma estatística. Já uma coleção de números ou fatos é denominado de **estatísticas** (plural). Por exemplo, As vendas da empresa Picuínhas totalizaram: 2,5 milhões em janeiro, 2,7 em fevereiro e 3.1 em março. No entanto o termo Estatística tem um sentido muito mais amplo, do que apenas números ou coleção de números. A **Estatística** pode ser definida como:

A ciência de coletar, organizar, apresentar, analisar e interpretar dados numéricos com o objetivo de tomar melhores decisões.

Assim como advogados possuem "regras de evidência" e contabilistas possuem "práticas comumente aceitas", pessoas que tratam com dados numéricos seguem alguns procedimentos padrões. Alguns destes métodos serão vistos nesta disciplina e outros em uma segunda disciplina. Não esquecendo que mesmo duas disciplinas de Estatística não esgotam o assunto, ou seja, elas dão apenas uma idéia dos procedimentos e técnicas existentes para se lidar com dados numéricos.

1.2. DIVISÃO DA ESTATÍSTICA

A Estatística que lida com a organização, resumo e apresentação de dados numéricos é denominada de Estatística Descritiva**.** Assim pode-se definir a **Estatística Descritiva** como sendo:

Os procedimentos usados para organizar, resumir e apresentar dados numéricos.

Conjuntos de dados desorganizados são de pouco ou nenhum valor. Para que os dados se transformem em informação é necessário organizá-los, resumi-los e apresentá-los. O resumo de

conjuntos de dados é feito através das medidas e a organização e apresentação através das distribuições de freqüências e dos gráficos ou diagramas.

Estatística Indutiva. Muitas vezes, apesar dos recursos computacionais e da boa vontade não é possível estudar todo um conjunto de dados de interesse. Neste caso estuda-se uma parte do conjunto. O principal motivo para se trabalhar com uma parte do conjunto ao invés do conjunto inteiro é o custo.

O conjunto de todos os elementos que se deseja estudar é denominado de *população*. Note-se que o termo população é usado num sentido amplo e não significa, em geral, conjunto de pessoas. Pode-se definir uma **população** como sendo:

Uma coleção de todos os possíveis elementos, objetos ou medidas de interesse.

Assim, são exemplos de populações:

1. O conjunto das rendas de todos os habitantes de Porto Alegre;
2. O conjunto de todas as notas dos alunos de Estatística;
3. O conjunto das alturas de todos os alunos da Universidade; etc.

Um levantamento efetuado sobre toda uma população é dito de levantamento censitário ou simplesmente censo.

Fazer levantamentos, estudos, pesquisas, sobre toda uma população (censo) é, em geral, muito difícil. Isto se deve à vários fatores. O principal é o custo. Um censo custa muito caro e demanda um tempo considerável para ser realizado. Assim, normalmente, se trabalha com partes da população denominadas de *amostras*. Uma **amostra** pode ser caracterizada como:

Uma porção ou parte de uma população de interesse.

Utilizar amostras para se ter conhecimento sobre populações é realizado intensamente na Agricultura, Política, Negócios, Marketing, Governo, etc., como se pode ver pêlos seguintes exemplos:

* Antes da eleição diversos órgãos de pesquisa e imprensa ouvem um conjunto selecionado de eleitores para ter uma idéia do desempenho dos vários candidatos nas futuras eleições.

* Uma empresa metal-mecânica toma uma amostra do produto fabricado em intervalos de tempo especificados para verificar se o processo está sob controle e evitar a fabricação de itens defeituosos.

* O IBGE faz levantamentos periódicos sobre emprego, desemprego, inflação, etc.

* Redes de rádio e tv se utilizam constantemente dos índices de popularidade dos programas para fixar valores da propaganda ou então modificar ou eliminar programas com audiência insatisfatória.

* Biólogos marcam pássaros, peixes, etc. para tentar prever e estudar seus hábitos.

O processo de escolha de uma amostra da população é denominado de amostragem.

Riscos da amostragem. O processo de amostragem envolve riscos, pois toma-se decisões sobre toda a população com base em apenas uma parte dela. A **teoria da probabilidade** pode ser utilizada para fornecer uma idéia do risco envolvido, ou seja, do erro que se comete ao utilizar uma amostra ao invés de toda a população, desde que, é claro, a amostra seja selecionada através de critérios probabilísticos, isto é, ao acaso.

Baseado nos conceitos anteriores pode-se definir **Estatística Indutiva** ou **Inferencial** como sendo:

A coleção de métodos e técnicas utilizados para se estudar uma população baseados em amostras probabilísticas desta mesma população.

1.3. MENSURAÇÃO

1.3.1. INTRODUÇÃO

O processo de selecionar o modelo matemático ou estatístico a ser utilizado com uma dada técnica de pesquisa ou procedimento operacional envolve algumas decisões importantes. A tomada de decisão do modelo matemático ou estatístico a ser aplicado costuma ser precedida pela mensuração do fenômeno envolvido. E uma primeira dificuldade surge na necessidade de se definir o que é mensuração. Se o termo se referir somente aqueles tipos de medidas comumente utilizados em ciências tais como a física (por exemplo: medidas de comprimento, massa ou tempo) não haverá muitos problemas na escolha do sistema matemático. Mas se o conceito de medida for ampla o suficiente para incluir certos procedimentos de categorização normalmente utilizados em Ciências Sociais, então o problema torna-se mais complexo. Pode-se distinguir entre diversos níveis de mensuração e para cada um existem diferentes modelos estatísticos apropriados.

1.3.2. FORMAS DE MENSURAÇÃO

Existem quatro formas de mensuração ou tipos ou níveis de medidas ou ainda escalas que são conhecidas como: *nominal, ordinal, intervalar e razão*.

Nível nominal. A operação básica e mais simples em qualquer ciência é a de classificação. Na classificação tenta-se separar conjuntos de elementos com respeito a certas categorias, tomando decisões sobre quais elementos são mais parecidos e quais são diferentes. O objetivo é colocar os elementos em categorias tão homogêneas quanto possível quando comparados com as diferenças existentes entre as categorias.

Os termos nível nominal de medida ou escala nominal são utilizadas para se referir a àqueles dados que só podem ser classificados em categorias. Se bem que no sentido estrito não existe na realidade uma medida ou escala envolvida. Existe apenas uma contagem. Variáveis que podem ser colocadas nesta categoria são, por exemplo, a classificação das pessoas quanto à religião, sexo, estado civil, etc. Não existe uma ordem particular entre as categorias ou grupos e além disso duas categorias quaisquer são mutuamente excludentes, isto é, uma pessoa não pode ser ao mesmo tempo católico e protestante. Além disso as categorias são exaustivas, significando que um membro da população deve aparecer em uma e somente uma das categorias. Observe a tabela um abaixo.

Tabela 01 - Exemplo de variável nominal

Estado civil	Número de pessoas
Casado	340
Solteiro	250
Viúvo	40
Divorciado	50
Total	**700**

Deve-se ser salientado que as classes ou categorias podem ser rotuladas com números, mas isto não significa que se pode efetuar operações aritméticas com estes números. Neste caso a função dos números á a mesma dos nomes, apenas identificar uma categoria.

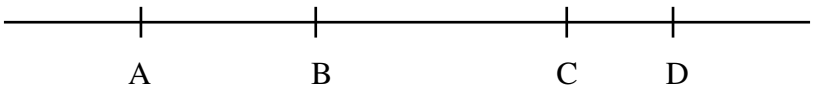
Nível ordinal. O nível ordinal é o tipo nominal em que se pode ordenar as categorias. A única diferença entre os dois níveis é a relação de ordem que se pode estabelecer entre as categorias. No entanto, não é possível afirmar o quanto uma categoria é maior do que a anterior, isto é, não se pode afirmar o quanto uma categoria possui da característica. A avaliação através de conceitos é feita por uma escala ordinal. Veja um exemplo na tabela dois abaixo.

Tabela 02 - Exemplo de variável em escala ordinal

Conceitos	Número de alunos
A	4
B	6
C	15
D	3
E	2
Total	**30**

Não se pode afirmar neste caso que quem tirou A teve um número de acertos duas vezes maior que quem tirou C. A única coisa que se sabe é que quem tem A acertou mais questões do quem tem B e este de quem tem C e assim por diante.

As famílias podem ser classificadas de acordo com seu estatus sócio econômico em : alta, média alta, média, média baixa, baixa. Não é possível entretanto afirmar que a diferença entre a alta e a média alta seja a mesma que entre a média e a média baixa. Esta escala além de possuir a propriedade simétrica da escala nominal é também *anti-simétrica* no sentido de que a relação que existe entre A e B pode não existir entre B e A. Por exemplo, a relação *maior que* é anti-simétrica, no sentido que se A > B então pode não ser verdade que B > A. A relação transitiva continua verdadeira, isto é: se A > B e B > C então A > C.



Pode-se dizer que a distância $\overline{AD} = \overline{AB} + \overline{BC} + \overline{CD}$, mas não se pode comparar as distâncias $\overline{AB}$ e $\overline{CD}$, em outras palavras, quando se traduz relações de ordem em operações matemáticas não se pode, em geral, utilizar as operações usuais de adição, subtração, multiplicação e divisão.

Nível intervalar. No sentido estrito da palavra o termo mensuração pode ser utilizado para se referir a situações em que se pode, não somente ordenar objetos com respeito ao grau de que eles possuem certa característica, mas também indicar a exata distância entre eles. Se isto for possível, será obtido o que se denomina de uma escala de intervalo.

A escala de medida intervalar é uma escala nominal em que a distância entre as categorias, ao contrário da ordinal, é sempre a mesma. Ou seja, ela possui todas as características da escala ordinal mais o fator de que a distância entre as diversas categorias (ou valores) é sempre constante. As escalas de medir temperaturas como a Fahrenheit e a Celsius são exemplos de escalas de intervalo. No entanto, não se pode afirmar que uma temperatura de 40 graus é duas vezes mais quente que uma de 20 graus, embora se possa dizer que a diferença entre 20 graus e 40 graus é a mesma que entre 75 graus e 95 graus. Isto porque este tipo de escala não possui um zero absoluto. Ou seja, o valor zero na escala é apenas um ponto de referência e não significa a ausência de calor. Escores padronizados são também exemplos deste tipo de nível de medida.

Torna-se evidente que uma escala de intervalo requer o estabelecimento de algum tipo de unidade física a qual todos concordem, isto é, um padrão e que seja replicável, isto é, possa ser aplicada muitas vezes e fornecendo sempre os mesmos resultados. Comprimento é medido em termos de cm ou metros, tempo em segundos, temperatura em centígrados ou Fahrenheit, renda em dólar ou

reais. Por outro lado não existem tais unidades para inteligência, autoritarismo, ou prestígio que seja unânime entre todos os cientistas sociais e que possa ser assumida constante de uma situação para outra.

Nível de razão. Este é o mais alto nível de medida. É caracterizado por apresentar todas as características da escala intervalar mais um zero absoluto. Aqui o zero pode ser entendido como a ausência da característica e as comparações de valor (razão) tem sentido. Um exemplo de variável deste tipo é o peso. Um valor igual a zero significa ausência de peso e um valor de 20 kg é duas vezes mais pesado que um de 10 kg.

2. RESUMO DE PEQUENOS CONJUNTOS DE DADOS

2.1. INTRODUÇÃO

Para se analisar um conjunto de valores é necessário primeiramente, para fins de notação, distinguir se este conjunto é resultado de um censo ou de uma amostragem.

A Estatística Descritiva pode ser estudada considerando os conjuntos de valores analisados como sendo amostras ou então populações. Como o caso mais comum é a obtenção de amostras a notação apresentada será feita considerando os valores como resultados de amostragens. No entanto, convém ficar atento, com a bibliografia, pois dependendo do autor a orientação pode ser outra. A diferença, considerada do ponto de vista da descrição dos dados, é apenas notacional. Assim o tamanho de uma população (quando finita) é representado, normalmente por "N", enquanto que o tamanho de amostra é representado por "n". Afora algumas exceções os valores calculados na amostra são representados por letras latinas enquanto que os correspondentes na população o são pelas mesmas letras só que gregas.

Para facilitar o estudo da Estatística Descritiva os conjuntos de valores serão considerados como pequenos e grandes. Assim se um conjunto tiver 30 ou menos valores a análise será feita sem o agrupamento. Caso o conjunto tenha mais do que 30 valores então primeiramente será feito o agrupamento de acordo com o tipo de variável considerada. O valor 30 é apenas um ponto de referência escolhido arbitrariamente e dependendo da situação pode-se considerar o agrupamento com mais ou menos valores envolvidos.

Um conjunto de dados, de qualquer tamanho, pode ser resumido de acordo com os seguintes medidas:

1. **Medidas de tendência central ou posição**
2. **Medidas de dispersão ou variabilidade.**
3. **Medidas de assimetria.**
4. **Medidas de achatamento ou curtose.**

2.2. MEDIDAS DE POSIÇÃO OU TENDÊNCIA CENTRAL

Um conjunto de valores (amostra) será representada por: $\mathbf{x_1, x_2, ..., x_n}$, onde "**n**" é o número de elementos do conjunto, isto é, o tamanho da amostra.

2.2.1. AS MÉDIAS

(a) A média aritmética

A média aritmética do conjunto $x_1, x_2, ..., x_n$ é representada por $\bar{x}$ e calculada por:

$$\bar{x} = (x_1 + x_2 + ... + x_n) / n = \sum \frac{x_i}{n}$$

(b) A média geométrica

A média geométrica dos valores positivos: $x_1, x_2, ..., x_n$, é representada por m_g e calculada por:

$$m_g = \sqrt[n]{x_1 x_2 \cdots x_n}$$

(c) A média harmônica

A média harmônica dos valores positivos $x_1, x_2, ..., x_n$ é representada por m_h e calculada por:

$$m_h = \frac{1}{\frac{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}}{n}} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + ... + \frac{1}{x_n}} = \frac{n}{\sum \frac{1}{x_i}}$$

Observando a expressão do cálculo da média harmônica pode-se verificar que ela é definida como sendo: *O inverso da média aritmética dos inversos.*

Exemplo:

Calcular as médias dos seguintes conjuntos de dados:

(a) 1 9 (b) 4 6 (c) 1/2 4/5 3/2 7/4

Para o conjunto em (a) tem-se:

$$\bar{x} = (1 + 9) / 2 = 5 \quad m_g = \sqrt{1.9} = \sqrt{9} = 3 \quad m_h = 2 / (1 + 1/9) = 18/10 = 1,80$$

Para o conjunto em (b) tem-se:

$$\bar{x} = (4 + 6) / 2 = 5 \quad m_g = \sqrt{4.6} = \sqrt{24} = 4,90 \quad m_h = 2 / (1/4 + 1/6) = 24/5 = 4,80$$

Para o conjunto em (c) tem-se:

$$\bar{x} = [1/2 + 4/5 + 3/2 + 7/4] / 4 = 91/80 = 1,14 \quad m_g = \sqrt[4]{\frac{1}{2}.\frac{4}{5}.\frac{3}{2}.\frac{7}{4}} = \sqrt[4]{\frac{84}{80}} = 1,02$$

$$m_h = \frac{4}{\frac{2}{1} + \frac{5}{4} + \frac{2}{3} + \frac{4}{7}} = \frac{4}{\frac{377}{84}} = \frac{336}{377} = 0,89$$

(d) Relação entre as três médias

As três médias mantém a seguinte relação entre elas, desde que os valores sejam positivos e diferentes entre si. $\bar{x} > m_g > m_h$

(e) A média aritmética ponderada

A média aritmética ponderada do conjunto x_1, x_2, ..., x_k, com pesos w_1, w_2, ..., w_2, é representada por m_p e calculada por:

$$m_p = (x_1 w_1 + x_2 w_2 + ... + x_n w_k) / (w_1 + w_2 + ... + w_k) = \frac{\sum x_i w_i}{\sum w_i}$$

Exemplo

A média da primeira prova de Estatística da turma 135 foi de 6,0 e foi realizada por 55 alunos. Na segunda prova compareceram 50 alunos que tiveram uma média de 6,5. A terceira prova realizada por 40 alunos teve média de 5,5. Qual a média geral das 3 provas?

Solução:

$$m_p = \frac{\sum x_i w_i}{\sum w_i} = (6,0.55 + 6,5.50 + 5,5.40) / (55 + 50 + 40) = 875 / 145 = 6,03.$$

2.2.2. A MEDIANA

A mediana de um conjunto **ordenado** de valores, anotada por m_e, é definida como sendo o valor que separa o conjunto em dois subconjuntos do mesmo tamanho. Assim se "n" (número de elementos) é ímpar a mediana é o valor central do conjunto. Caso contrário a mediana é a média dos valores centrais do conjunto. Tem-se:

$m_e = \mathbf{x_{(n+1)/2}}$ se "n" é ímpar e $m_e = \mathbf{[x_{(n/2)} + x_{(n/2)+1}] / 2}$ se "n" é par

Exemplo

Para o conjunto: 15 18 21 32 45 46 49

A mediana é:

$m_e = x_{(7+1)/2} = x_4 = 32,$

Ou seja, a mediana é o quarto valor na seqüência ordenada de elementos.

Se o conjunto acima fosse: 15 18 21 32 45 46

Então a mediana seria:

$me = [x_{(n/2)} + x_{(n/2)+1}] / 2 = [x_{(6/2)} + x_{(6/2)+1}] / 2 = (x_3 + x_4) / 2 = (21 + 32) / 2 = 53/2 = 26,50$

2.2.3. A MODA

A moda de um conjunto de valores, anotada por m_o, é definida como sendo "*o valor (ou os valores) do conjunto que mais se repete*". Convém lembrar que a moda ao contrário da mediana e da

média pode não ser única, isto é, um conjunto pode ser bimodal, trimodal, etc. ou mesmo amodal (sem moda).

Exemplo:

Dado o conjunto: 1 2 2 3 3 4 4 4 7 9 15

A moda será: mo = 4, pois este valor se repete 3 vezes no conjunto e qualquer outro se repete duas ou menos vezes.

2.3. MEDIDAS DE VARIABILIDADE OU DISPERSÃO

2.3.1. A AMPLITUDE

A mais simples das medidas de dispersão é a amplitude, anotada por "h", e definida como sendo a diferença entre os valores extremos do conjunto, isto é:

$$h = x_{max} - x_{min}$$

Exemplo:

A amplitude do conjunto: -5 4 0 3 8 10, vale:

$$h = x_{max} - x_{min} = 10 - (-5) = 15.$$

2.3.2. O DESVIO MÉDIO (ABSOLUTO)

A amplitude é uma medida simples e fácil de calcular. Tem a virtude de dar uma idéia da variabilidade do conjunto. No entanto ela não leva em consideração todos os valores do conjunto como seria desejável. Assim prefere-se, em geral, trabalhar com medidas que utilizam toda a informação disponível. Uma destas medidas é o desvio médio absoluto ou simplesmente desvio médio. O desvio médio é representado por "dma" e definido como sendo "*a média das distâncias que os valores do conjunto se encontram da média*".

$$dma = [\ |x_1 - \bar{x}| + |x_2 - \bar{x}| + ... + |x_n - \bar{x}|\ ] / n = \sum \frac{|x_i - \bar{x}|}{n}$$

Exemplo:

Calcular o dma do conjunto: -7 4 0 3 8 10

A média é $\bar{x} = (-7 + 4 + 0 + 3 + 8 + 10) / 6 = 18/6 = 3$

Então o desvio médio será:

$dma = [|-7 - 3| + |4 - 3| + |0 - 3| + |3 - 3| + |8 - 3| + |10 - 3|] / 6 = (10 + 1 + 3 + 0 + 5 + 7) / 6 = 26/6 = 4{,}33$

2.3.3. A VARIÂNCIA

O desvio médio apesar de intuitivamente fácil de interpretar e simples de calcular não é muito utilizado em Estatística. O que de fato é a medida de dispersão usual é a variância e principalmente sua raiz quadrada que é denominada de desvio padrão. A variância é anotada por s^2 e definida como sendo "*a média dos quadrados dos desvios em relação a média aritmética*." Por desvio entende-se a diferença entre um valor do conjunto e a média.

$$s^2 = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2] / n = \sum \frac{(x_i - \bar{x})^2}{n}$$

Nem sempre esta expressão é a mais indicada para ser utilizada. Quando a média é um valor decimal não exato ela não é muito prática, uma vez que entrará no cálculo "n" vezes aumentando os erros de arredondamento que ocorrem. Neste caso é melhor se valer de uma expressão alternativa que pode ser derivada da expressão acima desenvolvendo o quadrado dentro do somatório e fazendo algumas simplificações.

Trabalhando inicialmente apenas com o numerador da fórmula acima vem:

$$\sum (x_i - \bar{x})^2 = \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum x_i^2 - 2\bar{x}\sum x_i - \sum \bar{x}^2$$

Observando que $\bar{x} = \sum \frac{x_i}{n}$ tem-se que: $\sum x_i = n\bar{x}$ e ainda que: $\sum \bar{x}^2 = n\bar{x}^2$ vem:

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

Dividindo este resultado por "n" e simplificando a segunda parcela vem:

$$s^2 = \sum \frac{(x_i - \bar{x})^2}{n} = \sum \frac{x_i^2}{n} - \bar{x}^2$$

Esta é uma segunda expressão para o cálculo da variância e em muitas situações é mais vantajosa de ser usada. Neste caso a variância pode ser caracterizada como sendo: "*a média dos quadrados menos o quadrado da média*".

2.3.4. O DESVIO PADRÃO

A variância por ser um quadrado não permite comparações com a unidade que se está trabalhando. Para se ter uma medida de variabilidade com a mesma unidade do conjunto utiliza-se a raiz quadrada da variância, que é denominada de desvio padrão. Assim a expressão para o desvio é:

$$s = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n}} = \sqrt{\sum \frac{x_i^2}{n} - \bar{x}^2}$$

Exemplo:

Calcular a variância e o desvio padrão do conjunto: -7 4 0 3 8 10

A média é $\bar{x}$ = (-7 + 4 + 0 + 3 + 8 + 10) / 6 = 18/6 = 3

Então variância será:

$s^2 = [(-7 - 3)^2 + (4 - 3)^2 + (0 - 3)^2 + (3 - 3)^2 + (8 - 3)^2 + (10 - 3)^2] / 6 =$

$= (100 + 1 + 9 + 0 + 25 + 49) / 6 = 184 / 6 = 30,67$

E o desvio padrão: s = 5,54

2.3.5. A VARIÂNCIA RELATIVA

A variância relativa, representada por g^2 é o quociente entre a variância absoluta e o quadrado da média. Isto é:

$g^2 = s^2 / \bar{x}^2$

2.3.6. O COEFICIENTE DE VARIAÇÃO

O coeficiente de variação é a raiz quadrada da variância relativa. Isto é: $g = s / \bar{x}$

Exemplo:

Calcular a variância relativa e o coeficiente de variação do conjunto:

-7 4 0 3 8 10

A média é $\bar{x}$ = (-7 + 4 + 0 + 3 + 8 + 10) / 6 = 18/6 = 3

Então variância será:

$s^2 = [(-7 - 3)^2 + (4 - 3)^2 + (0 - 3)^2 + (3 - 3)^2 + (8 - 3)^2 + (10 - 3)^2] / 6 =$

$= (100 + 1 + 9 + 0 + 25 + 49) / 6 = 184 / 6 = 30,67$

O desvio padrão será: s = 5,54

Então a variância relativa será:

$g^2 = (184/6) / 9 = 3,41$

E o coeficiente de variação será: $g = s / \bar{x}$ = 5,54 / 3 = 184,59%

3. RESUMO DE GRANDES CONJUNTOS DE DADOS.

3.1. INTRODUÇÃO

Para se trabalhar com grandes conjuntos de dados é necessário inicialmente agrupar estes dados. O agrupamento é feito em tabelas, denominadas de **distribuições de freqüências**. Para se construir uma distribuição de freqüências é comum fazer a distinção entre dois tipos de variáveis. A variável (ou conjunto) discreta (valores que são resultados de contagem) e a variável (ou conjunto) contínuo (valores que são resultados de uma medida). Em geral variáveis discretas são agrupadas em **distribuições por ponto ou valores** e variáveis contínuas em **distribuições por classes ou intervalos**. A separação não é rígida e depende basicamente dos dados considerados. Poderá ser necessário usar uma distribuição por classes ou intervalos mesmo quando a variável é discreta.

3.2. DISTRIBUIÇÕES POR PONTO OU VALORES.

Considere-se um conjunto de valores resultados de uma contagem. Poderia ser, por exemplo, o número de irmãos dos alunos da turma U, disciplina de Estatística.

Número de irmãos dos alunos da turma U - disciplina Estatística

0	1	1	6	3	1	3	1	1	0
4	5	1	1	1	0	2	2	4	1
3	1	2	1	1	1	1	5	5	6
4	1	1	0	2	1	4	3	2	2
1	0	2	1	1	2	3	0	1	0

Esta coleção de valores, que apresentadas desta forma não é informação, pode ser transformada em informação mediante sua representação em uma distribuição de freqüências por pontos ou valores. Para tal, coloca-se o conjunto em uma tabela em que a coluna da esquerda é representada pêlos diferentes números ordenados (os pontos ou valores) e a coluna da direita pelo número de vezes que cada valor se repetiu (as freqüências simples ou absolutas). Para o exemplo, na tabela três, tem-se:

Tabela 03 - Distribuição de freqüências por ponto ou valores do número de irmãos dos alunos da turma U. Disciplina Estatística.

Número de irmãos	Número de alunos
0	7
1	21
2	8
3	5
4	4
5	3
6	2
Total	**50**

3.3. DISTRIBUIÇÕES POR CLASSES OU INTERVALOS

Considere-se um conjunto de valores resultados de uma medida. Poderia ser, por exemplo, a idade dos alunos da turma U da disciplina de Estatística.

Idade (em meses) dos alunos da turma U - Disciplina Estatística

230	234	276	245	345	240	270	310	368	369
334	268	288	336	299	236	239	355	330	247
287	344	300	244	303	248	251	265	246	266
240	320	308	299	312	324	289	320	264	275
252	298	315	255	274	264	263	230	303	281

Este conjunto de valores, obviamente não pode ser apresentado da mesma forma que o anterior, pois quase não há repetições. Neste caso é necessário construir uma tabela denominada de ”distribuição de freqüências por classes ou intervalos”. Evidentemente haverá perda de informação neste processo, mas o ganho obtido pela facilidade de interpretação e compreensão dos dados compensa.

O procedimento para construir uma distribuição de freqüências por classes ou intervalos segue os seguintes passos:

- Determinar a amplitude dos dados: $h = x_{max} - x_{min}$.
- Decidir sobre o número de classes “k“ a ser utilizado. Recomenda-se um número de classes entre 5 e 15. Para que a decisão não seja totalmente arbitrária pode-se usar a raiz quadrada do número de valores como o número de classes, ou seja, $k \cong \sqrt{n}$.

- Determinar a amplitude de cada classe. Sempre que possível manter todas as amplitudes iguais. Para tanto deve-se dividir a amplitude dos dados "h" pelo número de classes "k", arredondando para mais, ou seja, $h_i \cong h / k$.
- Contar o número de valores pertencentes a cada classe. Em geral, utiliza-se a simbologia (|---), para indicar um intervalo fechado à esquerda e aberto à direita. Também poderia ser utilizado o intervalo aberto à esquerda e fechado à direita (---|), aberto de ambos os lados (---) ou ainda fechado de ambos os lados (|---|).

Um exemplo de uma distribuição por classes ou intervalos é apresentado na tabela 04.

Tabela 04 - Idades dos alunos da turma U - Disciplina Estatística.

Idades	Número de alunos
230 |---- 250	12
250 |---- 270	9
270 |---- 290	8
290 |---- 310	7
310 |---- 330	6
330 |---- 350	5
350 |---- 370	3
Total	50

3.4. ELEMENTOS DE UMA DISTRIBUIÇÃO DE FREQÜÊNCIAS

Além da freqüência simples ou absoluta pode-se definir ainda:

3.4.1. A FREQÜÊNCIA RELATIVA OU PERCENTUAL

A freqüência relativa simples ou percentual é definida como sendo o quociente entre a freqüência simples "f_i" e o total de dados "n".

$$fr_i = f_i / n$$

Exemplo:

Na tabela três tem-se:

$fr_3 = 8 / 50 = 0,16 = 16\%$, significando que 16% dos alunos da turma possuem 2 irmãos.

Na tabela quatro tem-se:

$fr_2 = 9 / 50 = 0{,}18 = 18\%$, significando que 18% dos alunos possuem idades maiores ou iguais a 250 meses porém menores do que 270 meses.

3.4.2. A FREQÜÊNCIA ACUMULADA SIMPLES OU ABSOLUTA.

A freqüência acumulada simples ou absoluta da linha "i" é definida como sendo a soma das freqüência simples ou absolutas até a linha "i ".

$$F_i = f_1 + f_2 + ... + f_i$$

Exemplo:

Na tabela três tem-se:

$F_4 = f_1 + f_2 + f_3 + f_4 = 7 + 21 + 8 + 5 = 41$, significando que 41 alunos da turma possuem até 3 irmãos.

3.4.3. A FREQÜÊNCIA ACUMULADA RELATIVA OU PERCENTUAL

A freqüência acumulada relativa ou percentual da linha "i" é definida como sendo a soma das freqüência relativas ou percentuais até a linha "i ".

$Fr_i = fr_1 + fr_2 + ... + fr_i$, ou então, como sendo o quociente da freqüência acumulada simples pelo total de dados. $Fr_i = F_i / n$

Exemplo:

Na tabela quatro tem-se:

$Fr_2 = (12 + 9) / 50 = 42\%$, isto é, 42% dos alunos possuem idades menores do que 270 meses.

3.4.4. OUTROS ELEMENTOS

(i) Na tabela três os valores da coluna da esquerda são denominados de pontos ou valores. Cada um deles é representado por $\mathbf{x_i}$, onde "i" varia de 1 até k, sendo "k" o número de linhas da tabela.

(ii) Na tabela quatro os valores da coluna da esquerda são denominados de classes ou intervalos. As classes, também, variam de 1 até k.

(iii) Limite inferior da classe "i". Anota-se por li_i. Na tabela 4 o limite inferior da terceira classe é: 270.

(iv) Limite superior da classe "i". Anota-se por ls_i. Na tab. 4 o limite superior da quinta classe é: 330.

(v) Amplitude da classe "i". Anota-se por $\mathbf{h_i}$ e é calculada como a diferença entre os limites superior ou inferior da classe "i". Assim $\mathbf{h_i = ls_i - li_i}$. Na tabela quatro a amplitude da classe quatro é: $h_4 = ls_4 - li_4 = 310 - 290 = 20$ meses.

(vi) Ponto médio da classe. Como não é possível trabalhar com classes é necessário escolher um representante da classe. Este representante é denominado de ponto médio da classe. É representado por $\mathbf{x_i}$ e calculado por: $\mathbf{x_i = (li_i + ls_i) / 2}$ ou então $\mathbf{x_i = li_i + h_i / 2}$. Na tabela quatro o ponto médio da terceira classe é: $x_3 = (li_3 + ls_3) / 2 = (270 + 290) / 2 = 280$ meses.

Exemplo

Na tabela 05, abaixo, estão ilustrados os cálculos das freqüências relativas percentuais, da freqüência acumulada simples e da freqüência acumulada percentual.

Tabela 05 - Exemplos de freqüências

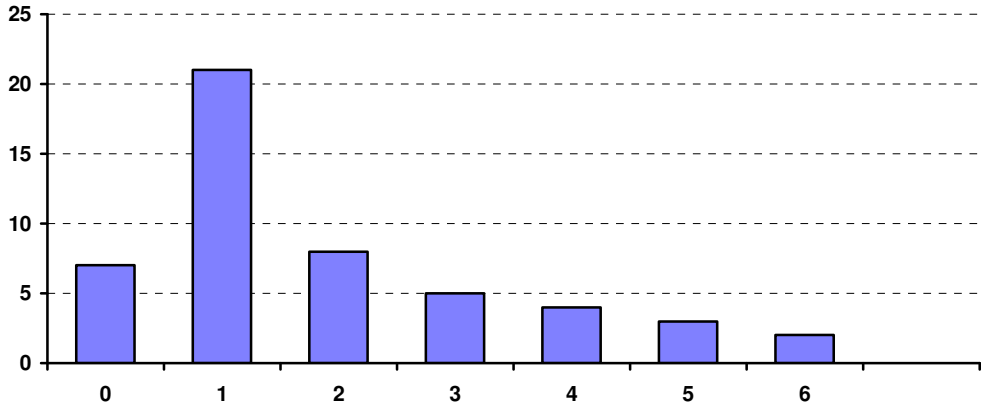
Número de irmãos	Número de alunos	fr_i	F_i	Fr_i
0	7	14	7	14
1	21	42	28	56
	8	16	36	72
3	5	10	41	82
4	4	8	45	90
5	3	6	48	96
6	2	4	50	100
Total	**50**	**100**	**----**	**----**

3.5. APRESENTAÇÃO DE UMA DISTRIBUIÇÃO DE FREQÜÊNCIAS

3.5.1. 3DISTRIBUIÇÃO DE FREQÜÊNCIAS POR PONTOS OU VALORES.

Uma distribuição de freqüências por pontos ou valores é apresentada graficamente através de um diagrama de linhas ou colunas, onde a variável "x_i" é representada no eixo das abcissas (horizontal) e as freqüências (que podem ser de qualquer tipo) no eixo das ordenadas (vertical). Veja-se um exemplo de diagrama de colunas simples na figura 01.

Figura 01 - Diagrama de colunas simples do número de irmãos dos alunos da turma U - Disciplina de Estatística



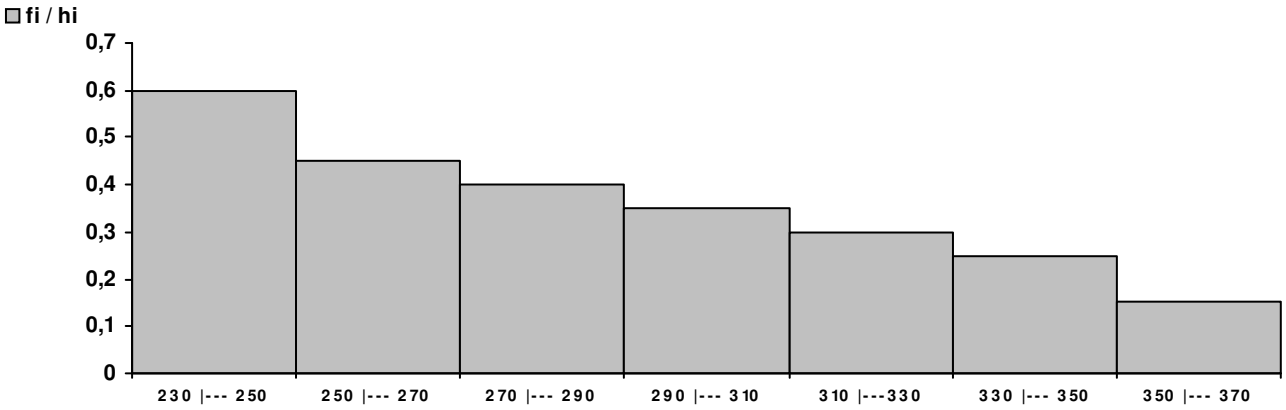
3.5.2. DISTRIBUIÇÃO DE FREQÜÊNCIAS POR CLASSES OU INTERVALOS

Uma distribuição de freqüências por classes ou intervalos é apresentada graficamente através de um diagrama denominado de histograma. Um histograma é um gráfico de retângulos justapostos onde a base de cada retângulo é a amplitude de cada classe e a altura é proporcional a freqüência (simples ou relativa) de modo que a área de cada retângulo seja igual a freqüência considerada. Desta forma a altura de cada retângulo será igual a: f_i / h_i ou então fr_i / h_i. Veja-se o cálculo das alturas na tabela 06 e o exemplo na figura 02. Também pode ser construído um histograma utilizando-se as freqüências acumuladas. Neste caso o diagrama resultante é denominado de **ogiva**. Se os pontos médios de cada classe de um histograma forem unidos através de segmentos de retas teremos então um diagrama denominado de **polígono de freqüências**.

Tabela 06 - Cálculo das ordenadas do histograma

Idades	Número de alunos	f_i / h_i
230 |---- 250	12	0,60
250 |---- 270	9	0,45
270 |---- 290	8	0,40
290 |---- 310	7	0,35
310 |---- 330	6	0,30
330 |---- 350	5	0,25
350 |---- 370	3	0,15
Total	50	----

Figura 02 - Histograma de freqüência simples do número de irmãos dos alunos da turma U - Disciplina de Estatística



3.6. RESUMO DE UMA DISTRIBUIÇÃO DE FREQÜÊNCIAS

3.6.1. MEDIDAS DE POSIÇÃO OU TENDÊNCIA CENTRAL

(i) A média aritmética

A média aritmética de uma distribuição de freqüências por pontos ou valores ou ainda por classes ou intervalos é dada por:

$$\bar{x} = (f_1x_1 + f_2x_2 + ... + f_nx_n) / (f_1 + f_2 + ... + f_n) = \sum \frac{f_i x_i}{n}$$

Exemplos:

A média da distribuição da tabela três, utilizando a tabela 07 para fazer os cálculos será:

Tabela 07 - Cálculo da média de uma distribuição por pontos ou valores

Número de irmãos	Número de alunos	$f_i.x_i$
0	7	0
1	21	21
2	8	16
3	5	15
4	4	16
5	3	15
6	2	12
Total	50	95

$$\bar{x} = \sum \frac{f_i x_i}{n} = 95 / 50 = 1{,}90 \text{ irmãos.}$$

Ou seja, o número médio de alunos da turma U, de Estatística, é de 1,90.

Já para a tabela quatro é necessário primeiro obter os valores dos pontos médios de cada classe ou intervalo. Fazendo os cálculos na tabela 08, vem:

Tabela 08 - Cálculo da média de uma distribuição por classes

Idades	Número de alunos	x_i	$f_i x_i$
230 |---- 250	12	240	2880
250 |---- 270	9	260	2340
270 |---- 290	8	280	2240
290 |---- 310	7	300	2100
310 |---- 330	6	320	1920
330 |---- 350	5	340	1700
350 |---- 370	3	360	1080
Total	50	----	14260

Deste modo a média das idades será:

$$\bar{x} = \sum \frac{f_i x_i}{n} = 14\,260 / 50 = 285{,}20 \text{ meses, ou seja, 285 meses e 6 dias.}$$

(ii) A mediana

(a) A mediana de uma distribuição de valores ou pontos é obtida da mesma forma que para dados não agrupados, isto é:

$m_e = \mathbf{x_{(n+1)/2}}$ se "n" é ímpar e $m_e = \mathbf{[x_{(n/2)} + x_{(n/2)+1}] / 2}$ se "n" é par

Observação: Neste caso deve-se trabalhar como se o conjunto não estivesse agrupado.

Exemplo

Para os valores da tabela três a mediana é:

$m_e = [x_{50/2} + x_{(50/2)+1}] / 2 = [x_{25} + x_{26}] / 2 = (1 + 1) / 2 = 1$, pois da oitava posição até a vigésima oitava posição todos os valores são iguais a um, e a mediana é a média entre os valores que se encontra na vigésima quinta e vigésima sexta posição.

(b) A mediana de uma distribuição de freqüências por classes ou intervalos é dada pela seguinte expressão:

$$m_e = li_i + h_i\left[\frac{n/2 - F_{i-1}}{f_i}\right], \text{ onde}$$

li_i = limite inferior da classe mediana, isto é, a classe que contém o ou os valores centrais;

h_i = amplitude da classe mediana;

f_i = freqüência simples da classe mediana;

F_{i-1} = freqüência acumulada simples da classe anterior à classe mediana.

Exemplo

Considerando que a classe mediana, na tabela quatro, é a que contém os valores x_{25} e x_{26}, isto é, a terceira classe, vem:

$$m_e = li_3 + h_3\left[\frac{n/2 - F_2}{f_3}\right] = 270 + 20[(25 - 21) / 8] = 270 + 10 = 280 \text{ meses.}$$

(iii) A moda

(a) A moda de uma distribuição de valores ou pontos é obtida da mesma forma que para dados não agrupados, ou seja, observando o valor ou os valores que mais se repetem.

m_o = valor da linha com maior freqüência (se existir apenas uma).

Exemplo

Para os valores da tabela três a moda é 1, pois este valor com uma freqüência de 21 é o que mais se repete.

(b) A moda de uma distribuição de freqüências por classes ou intervalos é dada pelas seguintes expressões:

$$m_o = li_i + h_i\left[\frac{f_{i+1}}{f_{i-1} + f_{i+1}}\right], \text{ denominada de moda de King ou, então, por:}$$

$$m_o = li_i + h_i\left[\frac{f_i - f_{i-1}}{2f_i - f_{i-1} - f_{i+1}}\right], \text{ denominada de moda de Kzuber, onde:}$$

li_i = limite inferior da classe modal, isto é, a classe de maior freqüência;

h_i = amplitude da classe modal;

f_i = freqüência simples da classe modal;

f_{i-1} = freqüência simples da classe anterior à classe modal;

f_{i+1} = freqüência simples da classe superior à classe modal.

Exemplo

Considerando que a classe de maior freqüência, a classe modal, na tabela quatro, é a primeira, vem:

$$m_o = li_1 + h_1\left[\frac{f_2}{f_2}\right] = 230 + 20 = 250 \text{ meses.}$$

$$m_o = li_1 + h_1\left[\frac{f_1}{2f_1 - f_2}\right] = 230 + 20[12 / (24 - 9)] = 230 + 16 = 246 \text{ meses.}$$

(iv) Relação entre as três medidas de posição

Karl Pearson estabeleceu a seguinte relação aproximada entre as três medidas de posição:

$\bar{x} - m_o = 3\ (\bar{x} - m_e)$,

Ou seja, em uma distribuição de freqüências à diferença entre a média e a moda é 3 vezes maior do que a diferença entre a média e a mediana.

3.6.2. MEDIDAS DE VARIABILIDADE OU DISPERSÃO

(a) A amplitude

A amplitude de uma distribuição de freqüências é definida como sendo a diferença entre os valores extremos da distribuição, isto é:

$h = x_{max} - x_{min}$, para a distribuição por pontos ou valores e

$h = ls_k - ls_1$, para a distribuição por classes ou intervalos.

Exemplo:

A amplitude da distribuição da tabela três é:

$h = x_{max} - x_{min} = 6 - 0 = 6$ irmãos

Já a amplitude da distribuição da tabela quatro vale:

$h = ls_7 - li_1 = 370 - 230 = 140$ meses

(b) O desvio médio (absoluto)

O desvio médio absoluto de uma distribuição de freqüências é dado por:

$$dma = [\ f_1|x_1 - \bar{x}| + f_2|x_2 - \bar{x}| + \ldots + f_k|x_n - \bar{x}|\ ] / n = \sum \frac{f_i|x_i - \bar{x}|}{n}$$

Exemplo:

O dma da distribuição da tabela três utilizando a tabela 09 para os cálculos, vale:

Tabela 09 - Cálculo do desvio médio absoluto

Número de irmãos	Número de alunos	$f_i\|x_i - \bar{x}\|$
0	7	7|0 - 1,90| = 13,30
1	21	21|1 - 1,90| = 18,90
2	8	8|2 - 1,90| = 0,80
3	5	5|3 - 1,90| = 5,50
4	4	4|4 - 1,90| = 8,40
5	3	3|5 - 1,90| = 9,30
6	2	2|6 - 1,90| = 8,20
Total	**50**	**64,40**

$$dma = \sum \frac{f_i \left| x_i - \bar{x} \right|}{n} = 64,40 / 50 = 1,29 \text{ irmãos}$$

(c) A variância

A variância de uma distribuição de freqüências pode ser avaliada por qualquer uma das seguintes expressões: $s^2 = [f_1(x_1 - \bar{x})^2 + f_2(x_2 - \bar{x})^2 + ... + f_k(x_k - \bar{x})^2] / n = \sum \frac{f_i(x_i - \bar{x})^2}{n} = \sum \frac{f_i x_i^2}{n} - \bar{x}^2$

(d) O desvio padrão

O desvio padrão de uma distribuição de freqüências é determinado extraindo-se a raiz quadrada da variância. Assim, do desvio padrão é: $s = \sqrt{\sum \frac{f_i(x_i - \bar{x})^2}{n}} = \sqrt{\sum \frac{f_i x_i^2}{n} - \bar{x}^2}$.

Exemplo:

A variância e o desvio padrão da distribuição da tabela 04.

Tabela 10 - Ilustração do cálculo da variância

Idades	Número de alunos	x_i	$f_i x_i$	$f_i x^2$
230 |---- 250	12	240	2880	691200
250 |---- 270	9	260	2340	608400
270 |---- 290	8	280	2240	627200
290 |---- 310	7	300	2100	630000
310 |---- 330	6	320	1920	614400
330 |---- 350	5	340	1700	578000
350 |---- 370	3	360	1080	388800
Total	**50**	**----**	**14260**	**4138000**

A variância da distribuição será:

$$s^2 = \sum \frac{f_i x_i^2}{n} - \bar{x}^2 = 4\ 138\ 000 / 50 - 285{,}20^2 = 82760 - 81339{,}04 = 1420{,}96$$

O desvio padrão vale:

$$s = \sqrt{\sum \frac{f_i x_i^2}{n} - \bar{x}^2} = 37{,}70$$

A variância relativa: $g^2 = s^2 / \bar{x}^2 = 0{,}0175$

O coeficiente de variação vale: $g = s / \bar{x} = 0{,}132\ 2 = 13{,}22\%$

3.6.3. MEDIDAS DE ASSIMETRIA

A **assimetria** (*skewness*) de um conjunto de dados, agrupados ou não, pode ser avaliada por meio da seguinte relação devida a Karl Pearson:

$$a_1 = 3(\bar{x} - m_e) / s$$

Se a_1 for igual a zero então a distribuição (ou conjunto) é dito simétrico. Se $a_1 > 0$ então a assimetria é positiva significando que o gráfico da distribuição tem uma cauda alongada à direita. Caso a_1 seja negativo a cauda do gráfico será alongada à esquerda.

Se uma distribuição de freqüências é simétrica então as 3 medidas de posição coincidem, isto é: $\bar{x} = m_e = m_o$.

Se a distribuição é positivamente assimétrica então $\bar{x} > m_e > m_0$

E se a distribuição é negativamente assimétrica então $\bar{x} < m_e < m_o$

Outra forma de avaliar a assimetria é utilizando o momento centrado de terceira ordem. Isto é:

$$g_1 = \sum \frac{f_i (x_i - \bar{x})^3}{n} / s^3 = \sum f_{r_i} (x_i - \bar{x})^3 / s^3 .$$

A assimetria é um número puro (positivo, zero ou negativo). Ela não tem unidade e não deve ser expressa em percentual.

3.6.4. MEDIDA DE CURTOSE

A **curtose** (*kurtosis*) de um conjunto de dados, agrupados ou não, pode ser avaliada por meio do momento centrado de quarta ordem, isto é:

$$g_2 = \sum \frac{f_i (x_i - \bar{x})^4}{n} / s^4 - 3 = \sum f_{r_i} (x_i - \bar{x})^4 / s^4 - 3 .$$

A curtose é um número puro (positivo, zero ou negativo). Ela não tem unidade e não deve ser expressa em percentual. A subtração do valor 3 é para que, a exemplo da assimetria, o coeficiente varie em torno de zero. A curtose é uma medida de achatamento e compara um conjunto de dados com a curva normal que teria um coeficiente de curtose (achatamento) igual a zero (três segundo algumas referências). Essa interpretação não é unânime e alguns autores dizem que a curtose mede a concentração dos valores nas caudas da distribuição e não o achatamento. De qualquer forma conforme o valor do coeficiente de curtose um conjunto de dados pode ser:

Mesocúrtico (com curtose igual a curva normal) se o coeficiente for igual a zero;

Leptocúrtico (mais pontiagudo que uma normal) se o coeficiente for maior do que zero e

Platicúrtico (mais achatado do que uma normal) se o coeficiente for menor do zero.

3.7. PROPRIEDADES DAS MEDIDAS

3.7.1. MEDIDAS DE POSIÇÃO

(i) Se todos os valores de um conjunto de dados forem somados a uma constante então as medidas de posição aumentam desta constante. Em símbolos. Dado um conjunto de dados x e somando a este conjunto uma constante "c". Então para y = x + c, tem-se:

$$\overline{y} = \overline{x} + c$$

O mesmo acontece com a mediana e a moda.

(ii) Se todos os valores de um conjunto de dados forem multiplicados a uma constante então as medidas de posição ficam multiplicadas por esta constante. Em símbolos. Se um conjunto de dados x for multiplicado por uma constante "c". Então para y = cx, tem-se:

$$\overline{y} = c\overline{x}$$

O mesmo acontece com a mediana e a moda.

3.7.2. MEDIDAS DE DISPERSÃO

(i) Se todos os valores de um conjunto de dados forem somados a uma constante então as medidas de dispersão não se alteram. Em símbolos. Dado um conjunto de dados x e somando a este conjunto uma constante "c". Então para y = x + c, tem-se:

$$s_y = s_x$$

O mesmo vale para a variância e para o dma. O coeficiente de variação e a variância relativa são exceções, pois são medidas derivadas, que combinam uma medida de posição a média no

denominador que se altera e uma medida de dispersão o desvio padrão ou a variância no numerador que não se altera.

(ii) Se todos os valores de um conjunto de dados forem multiplicados a uma constante então as medidas de posição ficam multiplicadas por esta constante, sendo que a variância fica multiplicada pelo quadrado desta constante. Em símbolos. Se um conjunto de dados x for multiplicado por uma constante "c". Então para y = cx, tem-se:

$$s_y = cs_x$$

O mesmo vale para a o dma. Já a variância que é um quadrado fica multiplicada pelo quadrado da constante. O coeficiente de variação e a variância relativa são exceções, pois são medidas derivadas, que combinam uma medida de posição, a média no denominador que se altera, e uma medida de dispersão, o desvio padrão ou a variância no numerador, que também se altera. Como tanto o numerador quanto o denominador se alteram na mesma proporção, então a razão entre as duas alterações passará a ser um. Portanto tanto a variância relativa quanto o coeficiente de variação são indiferentes a uma multiplicação do conjunto de valores por uma constante.

4. EXERCÍCIOS

(01) É possível encontrar a seguinte série de desvios tomados em relação a média aritmética: 4, -3, 2, -7 e 5? Justifique.

(02) Dados dois grupos de pessoas, o grupo A com 10 elementos e o grupo B com 40 elementos. Se o peso médio do grupo A for de 80 kg e o do grupo B for de 70 kg então é verdade que o peso médio dos dois grupos considerados em conjunto é de 75 kg? Justifique.

(03) Um concurso realizado simultaneamente nos locais A, B e C, apresentou as médias: 70, 65 e 45, obtidos por 30, 40 e 30 candidatos, nessa ordem. Qual foi a média geral do concurso?

(04) Para um dado concurso, 60% dos candidatos eram do sexo masculino e obtiveram uma média de 70 pontos em determinada prova. Sabendo-se que a média geral dos candidatos (independente de sexo) foi de 64 pontos, qual foi a média dos candidatos do sexo feminino?

(05) Determinar a moda dos seguintes conjuntos:

(05.1) 1, 6, 9, 3, 2, 7, 4 e 11

(05.2) 6, 5, 5, 7, 5, 6, 5, 6, 3, 4, e 5

(05.3) 8, 4, 4, 4, 4, 6, 9, 10, 10, 15, 10, 16,e 10

(05.4) 23, 28, 35, 17, 28, 35, 18, 18, 17, 18, 18, 18, 28, 28 e 18

(06) Determinar a mediana dos seguintes conjuntos:

(06.1) 9 14 2 8 7 14 3 21 1

(06.2) 0,02 0,25 0,47 0,01 -0,30 -0.5

(06.3) 1/2 3/4 4/7 5/4 -2/3 -4/5 -1/5 3/8

(07) Para os conjuntos abaixo, determinar com aproximação centesimal, as seguintes medidas:

(a) A amplitude **(b)** O desvio médio **(c)** A variância **(d)** O desvio padrão **(e)** O coeficiente de variação.

(7.1) 0,04 0,18 0,45 1,29 2.35

(7.2) -7/4 -1/3 3/5 7/20 1 4/3

(08) Dados os seguintes conjuntos de valores:

(a) 1 3 7 9 10 **(b)** 20 60 140 180 200 **(c)** 10 50 130 170 190.

Calculando a média e o desvio padrão do conjunto em **(a)**, determinar, através das propriedades, a média e o desvio padrão dos conjuntos em **(b)** e **(c).**

(09) Quarenta alunos da UFRGS foram questionados quanto ao número de livros lidos no ano anterior. Foram registrados os seguintes valores:

4	**2**	**1**	**0**	**3**	**1**	**2**	**0**	**2**	**1**
0	**2**	**1**	**1**	**0**	**4**	**3**	**2**	**3**	**5**
8	**0**	**1**	**6**	**5**	**3**	**2**	**1**	**6**	**4**
3	**4**	**3**	**2**	**1**	**0**	**2**	**1**	**0**	**3**

(09.1) Organize os dados em uma tabela adequada.

(09.2) Qual o percentual de alunos que leram menos do que 3 livros.

(09.3) Qual o percentual de alunos que leram 4 ou mais livros.

(09.4) Classifique a variável e o tipo de distribuição utilizada.

(10) O conjunto de dados abaixo representa uma amostra de 40 elementos:

3,67	**1,82**	**3,73**	**4,10**	**4,30**	**1,28**	**8,14**	**2,43**	**4,17**	**2,88**
5,36	**3,96**	**6,54**	**5,84**	**7,35**	**3,63**	**2,93**	**2,82**	**8,45**	**4,15**
5,28	**5,41**	**7,77**	**4,65**	**1,88**	**2,12**	**4,26**	**2,78**	**5,54**	**6,00**
0,90	**5,09**	**4,07**	**8,67**	**0,90**	**6,67**	**8,96**	**4,00**	**2,00**	**2,01**

(10.1) Agrupe os dados em uma distribuição de freqüências, considerando o limite inferior igual a zero, o superior igual a 10 e utilizando cinco classes de mesma amplitude.

(10.2) Construa um histograma de freqüências relativas.

(10.3) Una os pontos médios de cada retângulo, obtendo o polígono de freqüências relativas e classifique o conjunto quanto à assimetria.

(11) A tabela registra simultaneamente 200 aluguéis de imóveis urbanos e 100 de imóveis rurais.

Aluguéis	**Zona Urbana**	**Zona Rural**
1 |----- 3	10	30
3 |----- 5	40	50
5 |----- 7	80	15
7 |----- 9	50	05
9 |----- 11	20	00
Σ	**200**	**100**

(11.1) Calcule e interprete fr_2 para cada caso.

(11.2) Calcule e interprete F_3 para cada caso.

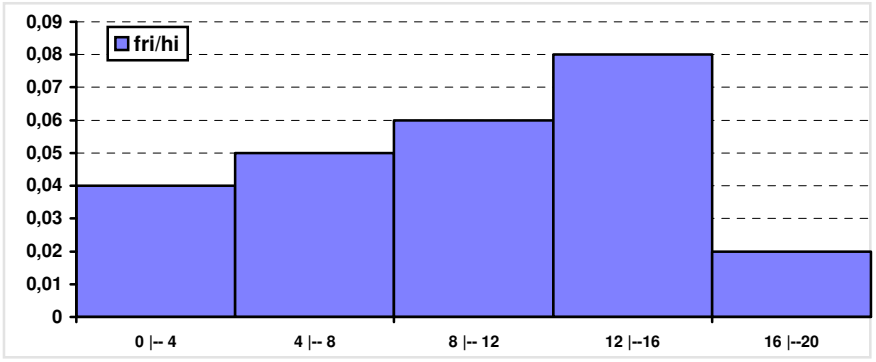
(11.3) Calcule e interprete Fr_4 - Fr_2 para cada caso.

(12) O histograma abaixo representa os salários, em unidades monetárias (u.m.) dos 100 empregados de uma empresa:

(12.1) Que percentual de empregados recebem 8 u.m. ou mais?

(12.2) Quantos empregados recebem de 4 a 16 u.m.?

(12.3) Quantos empregados recebem menos que 4 u.m. ou mais que 12 u.m.?



(13) Um livro com 50 páginas apresentou um número de erros de impressão por página conforme tabela:

Erros	Nº de páginas
0	25
1	20
2	3
3	1
4	1
Total	**50**

(13.1) Qual o número médio de erros por página?

(13.2) Qual o número mediano de erros por página?

(13.3) Qual o número modal de erros por página?

(13.4) Qual o desvio padrão do número de erros por página?

(14) Durante certo período de tempo o rendimento de 10 ações foram os que a tabela registra.

Ação	Taxa (%)
1	**2,59**
2	**2,64**
3	**2,60**
4	**2,62**
5	**2,57**
6	**2,55**
7	**2,61**
8	**2,50**
9	**2,63**
10	**2,64**

(14.1) Calcule o rendimento médio.

(14.2) Calcule o rendimento mediano.

(14.3) Calcule o rendimento modal.

(14.4) Calcule o desvio padrão do rendimento.

(14.5) Calcule o coeficiente de variação do rendimento.

(15) Uma região metropolitana tem 50 quarteirões com os seguintes números de casas por quarteirão:

2	**2**	**3**	**10**	**13**	**14**	**15**	**15**	**16**	**16**
18	**18**	**20**	**21**	**22**	**22**	**23**	**23**	**25**	**25**
26	**27**	**29**	**29**	**30**	**32**	**36**	**42**	**44**	**45**
45	**46**	**48**	**52**	**58**	**59**	**61**	**61**	**61**	**65**
66	**66**	**68**	**75**	**78**	**80**	**89**	**90**	**92**	**97**

(15.1) Construa, com os dados, uma distribuição de freqüências por intervalos fazendo com que as classes tenham amplitudes igual a 14.

(15.2) Calcule o número médio de casas por quarteirão.

(15.3) Determine o número mediano de casas por quarteirão.

(15.4) Calcule a variância do número de casas por quarteirão.

(15.5) Calcule, pelos dois processos, o número modal de casas por quarteirão.

(16) De um levantamento feito entre 100 famílias resultou a tabela ao lado. Determine:

(16.1) O número médio de filhos.

(16.2) O número mediano de filhos.

(16.3) O número modal de filhos.

(16.4) O desvio padrão do número de filhos.

Número de filhos	Número de famílias
0	18
1	23
2	28
3	21
4	7
5	3
Total	**100**

(17) As informações abaixo dizem respeito a distribuição de três variáveis. Indique, justificando, qual delas tem média mais representativa.

Distribuição A	Distribuição B	Distribuição C
$n = 200$	$n = 50$	$\bar{x} = 8$
$\sum fx = 5000$	$\sum fx = 500$	$\sum fx = 3200$
$\sum fx^2 = 130000$	$\sum fx^2 = 5450$	$\sum fx^2 = 32000$

(18) Identifique, justificando, qual a variável mais homogênea.

Distribuição A	Distribuição B
$n = 100$	$\bar{x} = 50$
$\sum fx = 5000$	$\sum fx = 10000$
$\sum fx^2 = 256400$	$\sum f(x - \bar{x})^2 = 7200$

(19) Uma variável x tem média igual a 10 e variância igual a 16. Calcule a média e a variância da variável dada por $y = (3x + 5) / 2$

(20) Uma variável x tem média igual a 5 e desvio padrão igual a 3. Calcule o coeficiente de variação da variável $y = 4x + 4$

(21) Uma variável x tem média igual a 6 e coeficiente de variação igual a 0,50. Calcule o coeficiente de variação da variável $y = (5x - 2) / 2$

(22) Os operários de um setor industrial têm, em uma época 1, um salário médio de 5 salários mínimos (sm) e desvio padrão de 2 sm. Um acordo coletivo prevê, para uma época 2, um aumento linear de

60%, mais uma parte fixa correspondente a 70% de um salário mínimo. Calcule a média e o desvio padrão dos salários na época 2.

(23) Uma variável x assume valores no intervalo [10; 30].

(23.1) Sabendo que x tem uma distribuição assimétrica positiva você diria que a média de x é: 20, menor que 20 ou maior que 20. Justifique.

(23.2) E se x tiver uma distribuição simétrica?

(24) O que se pode dizer se fosse dada a informação de que o salário mediano de um conjunto de profissionais é de 6 sm?

(25) Um a comunidade A tem 100 motoristas profissionais cujo salário médio é de 5 sm. A comunidade B, com 300 desses profissionais, remunera-os com uma média de 4 sm.

(25.1) É correto afirmar que A remunera melhor seus motoristas profissionais que B?

(25.2) Diante das informações disponíveis há garantia que os 100 salários individuais de A são maiores que os 300 de B? Por que?

(26) Abaixo você encontra duas distribuições que refletem os comportamentos de x e y (tamanhos de famílias) em duas comunidades, sendo que uma de base cultural alemã e outra italiana. Utilize tais informações para uma análise que indique qual das duas comunidades tem famílias maiores.

X	f
2	25
3	30
4	48
5	111
6	98
7	88

Y	f
3	48
4	51
5	48
6	41
7	32
8	14
9	6

(27) O departamento de pessoal de certa firma fez um levantamento dos salários dos 120 funcionários do setor administrativo, obtendo os resultados da tabela:

Faixa salarial (em s.m.)	Percentual de funcionários
1 |---- 3	0,25
3 |---- 5	0,40
5 |---- 7	0,20
7 |--- 10	0,15
Total	**1,00**

(27.1) Determine o salário médio dos funcionários

(27.2) Determinar a variância e o desvio padrão dos salários.

(27.3) Determinar o salário mediano.

(27.4) Determinar o salário modal pêlos critérios de King e Czuber.

(27.5) Se for dado um aumento de 20% para todos os funcionários, qual será o novo salário médio e o novo desvio padrão dos salários?

(27.6) Se for dado um abono de 0,5 s.m. a todos os funcionários como fica a média e o desvio padrão dos salários?

(28) O que acontece com a média e o desvio padrão de um conjunto de dados quando:

(28.1) Cada valor é multiplicado por 2.

(28.2) Soma-se o valor 10 a cada valor.

(28.3) Subtrai-se a média de cada valor.

(28.4) De cada valor subtrai-se a média e em seguida divide-se pelo desvio padrão

(29) A média aritmética entre dois valores é igual a 5 e a média geométrica igual a 4. Qual a média harmônica entre estes dois valores?

(30) Identifique os tipos de escalas utilizadas para cada uma das seguintes características das unidades de observação, retiradas de uma tabela do Guia do Usuário do aplicativo Microsoft Excel: mês, tipo de produto, vendedor, região do país, unidades vendidas e total de vendas.

5. RESPOSTAS DOS EXERCÍCIOS

(1) Não, pois a soma dos desvios é diferente de zero.

(2) Não, pois deve ser utilizada a média ponderada e não da média aritmética simples.

(3) 60,50

(4) 55

(5) **(5.1)** Amodal **(5.2)** 5 **(5.3)** 4 e 10 **(5.4)** 18

(6) **(6.1)** 8 **(6.2)** 0,02 **(6.3)** 7/16

(7) **(7.1)** **(a)** 2,31 **(b)** 0,77 **(c)** 0,74 **(d)** 0,86 **(e)** 99,90%

(7.2) **(a)** 37/12 = 3,08 **(b)** 149/180 = 0,83 **(c)** 1,03 **(d)** 1,02 **(e)** 508,01%

(8) Observe que o conjunto em (b) é igual ao conjunto em (a) multiplicado por 20 e o conjunto em (c) é igual ao conjunto em (a) multiplicado por 20 e subtraído de 10 unidades.

(9) **(9.1)** **(tabela ao lado)**

(9.2) 24/40 = 60%

(9.3) 9/40 = 22,5%

(9.4) Distribuição por ponto ou valores

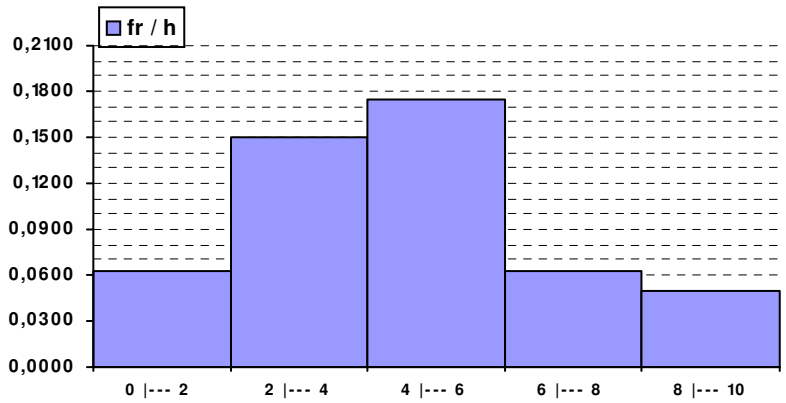
Variável quantitativa discreta.

N^0 de livros	**N^0 de alunos**
0	7
1	9
2	8
3	7
4	4
5	2
6	2
8	1
Total	40

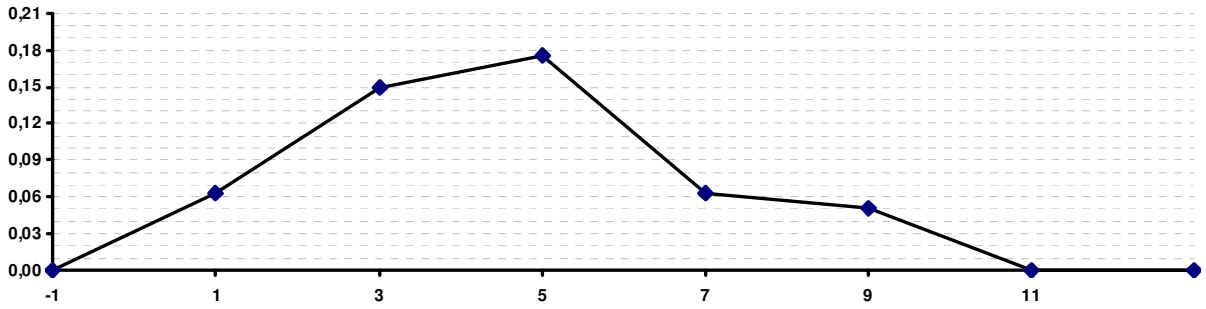
(10) **(10.1)**

Variável	**Freqüências**
0 |----- 2	5
2 |----- 4	12
4 |----- 6	14
6 |----- 8	5
8 |----- 10	4
Σ	**40**

(10.2)



(10.3) Assimétrico positivo (leve cauda para à direita)



(11) **(11.1)** Zona urbana: $fr_2 = 0{,}20 \Rightarrow 20\%$ dos aluguéis observados estão entre 3 e 5.

Zona rural: $fr_2 = 0{,}50 \Rightarrow 50\%$ dos aluguéis investigados ente entre 3 e 5.

(11.2) Zona urbana: $F_3 = 130 \Rightarrow$ 130 aluguéis investigados são menores do que 7.

Zona rural: $F_3 = 95 \Rightarrow$ 95 aluguéis investigados são menores do que 7.

(11.3) Zona urbana: $Fr_4 - Fr_2 = 0{,}90 - 0{,}25 = 0{,}65 = 65\%$ dos aluguéis estão entre 5 e 9.

Zona rural: $Fr_4 - Fr_2 = 1{,}00 - 0{,}80 = 0{,}20 = 20\%$ dos aluguéis estão entre 5 e 9.

(12) **(12.1)** 64% **(12.2)** 76 **(12.3)** 56

(13) **(13.1)** 0,66 erros **(13.2)** 0,50 erros **(13.3)** Zero erros **(13.4)** 0,84 erros

(14) **(14.1)** 2,60% **(14.2)** 2.60% **(14.3)** 2,64% **(14.4)** 0,04% **(14.5)** 1,63%

(15) (15.1)

Número de casas por quarteirão	Número de quarteirões
02 |----- 16	8
16 |----- 30	16
30 |----- 44	4
44 |----- 58	6
58 |----- 72	9
72 |----- 86	3
86 |----- 100	4
Σ	**50**

(15.2) 41,76 casas **(15.3)** 33,50 casas **(15.4)** 686,86 casas **(15.5)** 20,67 e 21,60 casas

(16) **(16.1)** 1,85 filhos **(16.2)** 2 filhos **(16.3)** 2 filhos **(16.4)** 1,30 filhos

(17) É a variável da distribuição A cujo coeficiente de variação é 0,20, o menor dentre as 3 distribuições.

(18) É a variável da distribuição B cujo coeficiente de variação é 0,12, o menor dentre as 2 distribuições.

(19) $\bar{x}$ = 17,50 s^2 = 36

(20) g = 50%

(21) g = 0,5357 = 53,57%

(22) $\bar{x}$ = 8,70 sm s = 3,20 sm

(23.1) É maior do que 20. **(23.2)** Seria 20, pois 20 é o ponto médio do intervalo total de x.

(24) Que metade dos profissionais recebem até 6 sm.

(25) (25.1) Sim, em média.

(25.2) Não, pois se teria que, no mínimo, conhecer os dois desvios padrões para poder fazer alguma afirmação a este respeito.

(26) A comparação pretendida deve ser feita pelas médias. As famílias de base cultural alemã têm, em média, 5,23 membros, enquanto que os de base italiana têm 5,10. Então as de base alemã tem média famílias levemente maiores.

(27) **(27.1)** 4,58 sm **(27.2)** 4,51 e 2,12 sm **(27.3)** 4,25 sm **(27.4)** 3,89 sm e 3,86 sm

(27.5) 5,49 e 2,55 sm **(27.6)** 5,08 e 2,12 sm

(28) **(28.1)** A média e o desvio padrão ficam multiplicados por 2.

(28.2) A média fica somada de 10 e o desvio padrão não se altera.

(28.3) A média fica igual a zero e o desvio padrão não se altera.

(28.4) A média fica igual a zero e o desvio padrão fica igual a um.

(29) 3,2

(30) Mês (Qualitativa ordinal)**;** Tipo de produto (Qualitativa nominal); Vendedor (Qualitativa nominal); Região do país (Qualitativa ordinal); Unidades vendidas (Quantitativa discreta); Total de vendas (Quantitativa contínua).

6. REFERÊNCIAS

AZEVEDO, Amilcar Gomes de, CAMPOS, Paulo Henrique Borges de. *Estatística Básica.* São Paulo. Livros Técnicos e Científicos Editora, 1981.

BUSSAB, Wilton O, MORETTIN, Pedro A. *Estatística Básica*. 3° ed. São Paulo, Atual, 1986.

HOFFMAN, Rodolfo. *Estatística para Economistas*. São Paulo. Livraria Pioneira Editora, 1980.

NETO, Pedro Luiz de Oliveira Costa. *Estatística*. São Paulo, Edgard Blücher, 1977.

MASON, Robert D., DOUGLAS, Lind A. *Statistical Techniques in Business And Economics*. IRWIN, Boston, 1990.

SPIEGEL, Murray R. *Estatística*. São Paulo. McGraw-Hill do Brasil, 1985.

STEVENSON, William J. *Estatística Aplicada à Administração*. São Paulo. Editora Harbra, 1981.

WONNACOTT, Ronald J., WONNACOTT, Thomas. *Fundamentos de Estatística*. Rio de Janeiro. Livros Técnicos e Científicos Editora S. A., 1985.