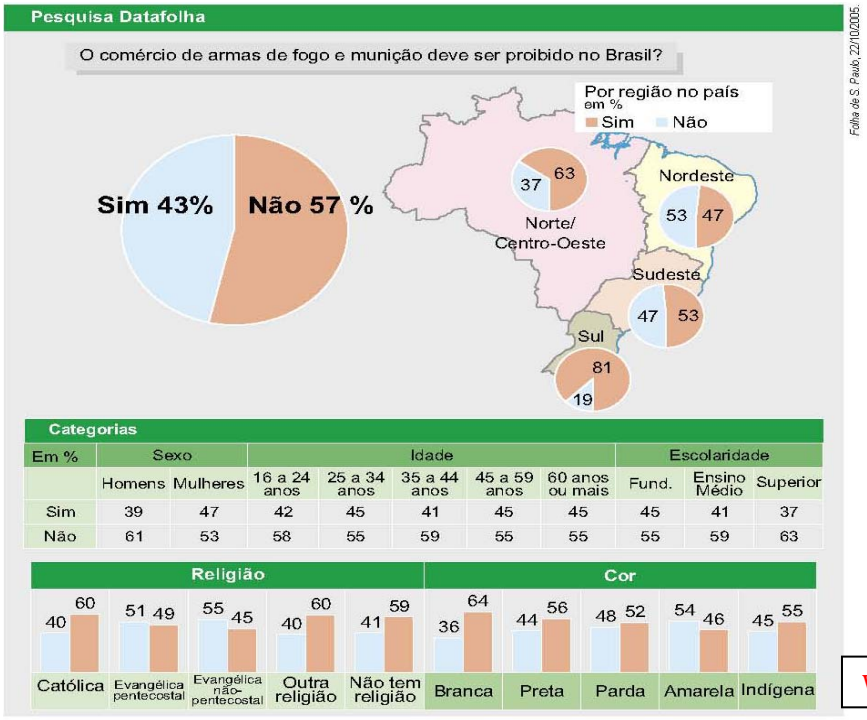
IMES-FAFICA

Tópicos de Matemática Aplicada

Estatística Aplicada no Excel

Ciência da Computação

Bertolo, L.A.



Versão BETA

Agosto 2008

Capítulo 2 – Medidas Estatísticas

A redução dos dados através de tabelas de freqüências e gráficos fornece muito mais informações sobre o comportamento de uma variável do que a própria série original de dados. Contudo, muitas vezes queremos resumir ainda mais esses dados, apresentando um ou alguns valores que sejam representativos da série toda. Quando usamos um só valor, obtemos uma redução drástica dos dados.

As principais medidas estatísticas (ou simplesmente estatísticas) referem-se às **medidas de posição** (locação ou tendência central) ou às **medidas de dispersão** (ou variabilidade):

2.1 – Medidas de Posição (ou tendência central)

Mostram o valor representativo em torno do qual os dados tendem a agrupar-se com maior ou menor freqüência.

A medida de tendência central é um número que está representando todo o conjunto de dados; nas pesquisas tal número pode ser encontrado a partir da **média aritmética, da moda ou da mediana**, e o uso de cada uma delas é mais conveniente de acordo com o nível de mensuração, o aspecto ou forma da distribuição de dados e o objetivo da pesquisa.

2.1.1 – Média Aritmética Simples ($\bar{x}$)

É a medida de centralidade mais comum, porém deve ser usada em dados representados por intervalos, pois não haveria sentido utilizá-la em uma distribuição em que a variável fosse, por exemplo, time de futebol ou sexo. A média representa, ainda, o ponto de distribuição no qual se equilibram as discrepâncias (diferenças) positivas e negativas de cada dado, ou seja, as discrepâncias positivas somadas se anulam com as negativas somadas.

Definida da seguinte forma:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

é a soma de todos os números dividida pelo número de parcelas. É uma das medidas de tendência central de maior emprego.

Usada em dados não agrupados.

EX: 4 15 20 20 24 27 30

$$\bar{x} = \frac{4 + 15 + 20 + 20 + 24 + 27 + 30}{7} = 20$$

Observe que: (20-4) + (20-15) + (20 – 24) + (20 – 27) + (20 – 30) = 0

2.1.2 – Média Aritmética Ponderada ($\bar{\bar{x}}$)

È um tipo de média aritmética de vários valores com pesos diferentes, dada por:

$$\bar{x} = \frac{p_1x_1 + p_2x_2 + p_3x_3 + \ldots + p_nx_n}{p_1 + p_n + \ldots\ldots + p_n} = \frac{\sum_{i=1}^{n} p_ix_i}{\sum_{i=1}^{n} p_i}$$

p_i = peso da amostra x_i.

Para dados agrupados em classes, temos:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_in_i}{\sum_{i=1}^{n} n_i} = \frac{1}{N}\sum_{i=1}^{n} x_in_i = \sum_{i=1}^{n} x_if_i$$

A média aritmética simples pode ser vista como a média ponderada com todos os pesos iguais. Para efeito de nomenclatura sempre trataremos a média aritmética simples ou ponderada simplesmente por média representada por ($\bar{x}$).

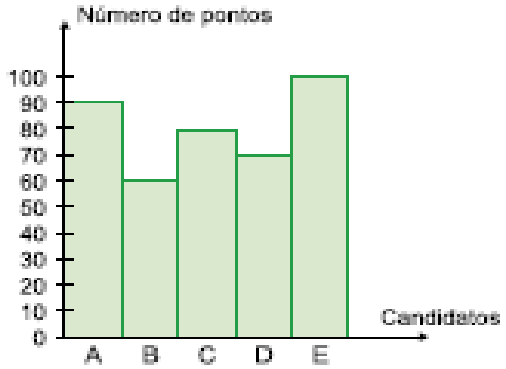
2.1.2 – Média Geométrica ($\bar{x}_G$)

È definida como a raiz de ordem n do produto desses números.

$$\bar{x}_G = \sqrt[n]{x_1.x_2.x_3 \ldots . x_n}$$

2.1.3 – Exercícios Modelos

01. Temos um gráfico que nos mostra o desempenho dos 5 melhores classificados em um determinado concurso, no qual a pontuação varia de zero a cem pontos.
a) Qual é a soma dos pontos dos candidatos A, B, C, D e E?
b) Determine a média aritmética dos pontos dos candidatos discriminados no gráfico.
c) Mostre qual o candidato que fez mais e o que fez menos pontos.



02. Um professor de uma determinada disciplina resolveu que suas provas bimestrais terão pesos diferentes em cada bimestre e que seus alunos, só no final do 4º bimestre, receberão a média final. Escolhendo aleatoriamente um aluno desse professor, vamos, de acordo com suas notas e respectivos pesos, verificar sua média final.
O aluno no primeiro bimestre tirou 6 e a prova tinha peso 2, no 2º bimestre tirou 5 e o peso era 4, no 3º bimestre o aluno tirou 3 e o peso era 2 e, finalmente, no 4º bimestre tirou 10 e o peso era 4. Calcule sua média final.

03. A tabela a seguir apresenta a distribuição de freqüências dos salários de um grupo de 50 empregados de uma empresa, num certo mês.

Número da classe	Salário do mês em reais	Número de empregados
1	1.000 |— 2.000	20
2	2.000 |— 3.000	18
3	3.000 |— 4.000	9
4	4.000 |— 5.000	3

O salário médio desses empregados, nesse mês, foi de:
a. R$ 2.637,00
b. R$ 2.500,00
c. R$ 2.420,00
d. R$ 2.400,00

04. Calcule a média geométrica da série (2, 4, 8)

2.1.4 – Mediana ($\tilde{x}$)

É o valor "do meio" de um conjunto de dados, quando os dados estão dispostos em ordem crescente ou decrescente, cortando, assim, a distribuição em duas partes com o mesmo número de elementos.

É uma medida separatriz definida e exata, de fácil compreensão. Ela serve para análise comparativa e é representada por $\tilde{x}$

Para dados não agrupados em classes:

Se *n* é **ímpar** $\rightarrow \tilde{x} = \left(\frac{n+1}{2}\right)^{o}$ termo

Com os números:

20 27 30 24 20 15 4

devemos, colocá-los em ordem:

4 15 20 20 24 27 30

Mediana igual a 20

Se *n* é **par** $\rightarrow \tilde{x} = \frac{\left(\frac{n}{2}\right)^{o} termo + \left(\frac{n}{2}+1\right)^{o} termo}{2}$ temos que a mediana será a média aritmética dos dois elementos centrais, após todos os elementos serem colocados em ordem.

Com os números:

20 27 30 24 20 15

Deve-se colocá-los em ordem:

15 20 20 24 27 30

Mediana igual a $\tilde{x} = \frac{20+24}{2} = 22$

Exercício: {35, 36, 37, 38, 40, 40, 41, 43, 46} $\Rightarrow \tilde{x} = 40$

{12, 14, 14, 15, 16, 16,17, 20} $\Rightarrow \tilde{x} = \frac{15+16}{2} = 15{,}5$

02. Em um colégio, estão matriculados numa determinada classe 21 alunos. Durante o 1º bimestre foi feito um levantamento da freqüência destes alunos e foram observadas as seguintes faltas: 0, 0, 3, 5, 7, 9, 0, 1, 2, 3, 11, 2, 3, 5, 6, 4, 10, 12, 0, 1, 2. Qual a mediana $\tilde{x}$ das faltas?

03. As idades dos atletas amadores de uma determinada modalidade esportiva são 14, 12, 16, 13, 17, 16 anos. Encontre a mediana da série.

2.1.5 – Média x Mediana

A média é muito sensível a valores extremos de um conjunto de observações, enquanto a mediana não sofre muito com a presença de alguns valores muito altos ou muito baixos. A mediana é mais "robusta" do que a média.
Devemos preferir a mediana como medida sintetizadora quando o histograma do conjunto de valores é assimétrico, isto é, quando há predominância de valores elevados em uma das caudas.

Ex.: { 200, 250, 250, 300, 450, 460, 510 }
$\bar{x}$=345,7 $\tilde{x}$=300
Tanto como $\tilde{x}$ são boas medidas de posição.

Ex.: { 200, 250, 250, 300, 450, 460, 2300 }
$\bar{x}x$ = 601 $\tilde{x}$ = 300
Devido ao valor 2300, $\tilde{x}$ é preferível a $\bar{x}$.

2.1.6 – Percentis

" O percentil de ordem p, 0 ≤ p≤ 100, de um conjunto de valores dispostos em ordem crescente é um valor tal que p% das observações estão nele ou abaixo dele e (1 - p)% estão nele ou acima dele."

Ex: Para valores de 51 a 100, ordenados crescentemente:

P_{25} = *25* deixa 25% dos dados (12,5 ⇒ 13 valores) nele ou abaixo dele e 75% dos dados (37,5 ⇒ 38 valores) nele ou acima dele. Assim: P_{25} = 63.

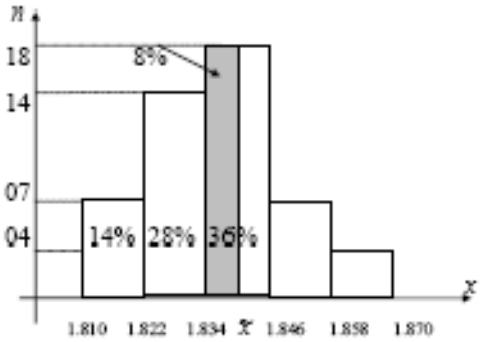
Similarmente, P_{80} deixa 80% dos dados (40 valores) nele ou abaixo dele e 20% dos dados (10 valores) nele ou acima dele. Assim: $P_{80} = \frac{90+91}{2} = 90{,}5$

Para dados agrupados em classes, os percentís podem ser obtidos por interpolação linear (regra de três simples).

Ex.: Dada a distribuição de freqüência de uma variável *X* qualquer:

X			x_i	N_i	N_i
1,810	|—	1,822	1,816	7	7
1,822	|—	1,834	1,828	14	21
1,834	|—	1,846	1,840	18	39
1,846	|—	1,858	1,852	7	46
1,858	|—	1,870	1,864	4	50

Temos que, para P_{50} (50% de 50) será o 25º elemento, está na terceira classe. Isto porque a segunda classe contém 21 elementos e a terceira, 39 elementos. Logo, o 25º elemento estará na 3ª



$\frac{1.846-1.834}{36\%} = \frac{\tilde{x}-1.834}{8\%} \Rightarrow \tilde{x} = 1.837$

Um outro processo gráfico pode ser usado para o cálculo desses percentis. (Veja Ogiva de Galton). Tal processo exige rigor no traçado e deve-se preferir papel milimetrado.
Obs.: As calculadoras geralmente não fornecem mediana e percentis.

2.1.7 – Moda

É o valor que ocorre com maior freqüência em um conjunto de observações individuais. Para dados agrupados temos a **classe modal**. Em alguns casos pode haver mais de uma moda. Assim temos uma distribuição bimodal, trimodal, etc...
A moda é o valor em torno do qual os dados estatísticos tendem a estar mais pesadamente concentrados e é representada por Mo, também conhecida pelo nome de norma ou modo.
O termo moda foi introduzido por Pearson.

Exemplos

01 Em um grupo de pessoas cujas idades são: 3, 2, 5, 2, 6, 2, 4, 4, 2, 7, 2 anos, a moda é **2** anos (Mo = 2). Portanto, denomina-se unimodal.

02 Algumas pessoas freqüentaram a escola por estes números de anos: 5, 3, 7, 5, 5, 8, 5, 3, 1, 1, 3, 3, 10, 3, 5. Nesta série de números, podem-se ter duas modas: $M_o = \begin{cases} 3 \\ 5 \end{cases}$ Portanto bimodal

Para exemplificar tomamos dados observados e colocamos em uma tabela.

03 Temos um grupo de pessoas cujas idades são: 3, 2, 5, 2, 6, 2, 4, 4, 2, 7, 2 anos:

Idade	2	3	4	5	6	7
Freqüência	5	1	2	1	1	1

Fica claro que a moda é 2 anos.

04 Tempo, em anos, que um grupo de pessoas freqüentou a escola.

Tempo de Escolaridade	
Tempo em anos de permanência na escola	**Freqüência**
1	**2**
3	**5**
5	**5**
7	**1**
8	**1**
10	**1**

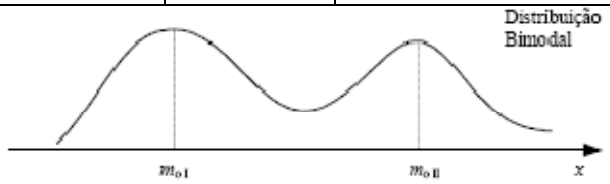
Nesse exemplo, afirmamos que há duas modas, 3 e 5, portanto o conjunto de dados é bimodal.

Nota importante

Quando não houver repetição de números, não haverá moda (o conjunto de dados é **amodal**).

Quando os dados estão agrupados em classes,

X	x_i	n_i	
10 |— 20	15	2	
20 |— 30	25	4	
30 |— 40	35	10	⇒ Classe Modal
40 |— 50	45	6	
50 |— 60	55	2	



Exemplo

Nesta série, 1, 7, 9, 12, 17, não há moda, pois não há repetição de número. Observe a resolução deste exemplo.

Considere os números 621, 310, 621, 201 e calcule:

a) a média aritmética ($\bar{x}$);

b) a média aritmética ponderada ($\bar{x}_p$), com pesos 2, 3, 1 e 2, respectivamente;

c) a moda (Mo).

Resposta

Primeiramente, monta-se a tabela:

Números	621	310	201
Freqüência	2	1	1

a. $\bar{x} = \frac{621+310+621+201}{4} = \frac{1.753}{4} = 438{,}25$

b. $\bar{x} = \frac{621.2+310.3+621.1+201.2}{2+3+1+2} = \frac{3.195}{8} = 399{,}375$

c. Observando a tabela com os dados do exercício, verificamos que o número 621 aparece 2 vezes. Essa é a maior freqüência de acordo com a tabela, portanto Mo = 621.

Exercícios de Aplicação

01 UFRN-RN Uma prova foi aplicada em duas turmas distintas. Na primeira, com 30 alunos, a média aritmética das notas foi 6,40. Na segunda, com 50 alunos, foi 5,20. A média aritmética dos 80 alunos foi:
a) 5,65 c) 5,75
b) 5,70 d) 5,80

02 Fuvest-SP
Uma prova continha cinco questões, cada uma valendo dois pontos. Em sua correção, foram atribuídas a cada questão apenas as notas 0 ou 2, caso a resposta estivesse, respectivamente, errada ou certa. A soma dos pontos obtidos em cada questão forneceu a nota do aluno. Ao final da correção, produziu-se a seguinte tabela, contendo a porcentagem de acertos em cada questão.

Questão	Porcentagem de acerto
1	30%
2	10%
3	60%
4	80%
5	40%

De acordo com a tabela construída no exercício anterior, encontrar a freqüência percentual de cada salário e repetir a sua tabela acrescentando mais uma coluna, de valores percentuais.
Logo, a média das notas da prova foi:
a) 3,8 d) 4,4
b) 4,0 e) 4,6
c) 4,2

03 TCU - Considere a distribuição de freqüências dos tempos de auditoria:

Tempo de auditoria (min)	Freqüência
10 ... 19	10
20 ... 29	20
30 ... 39	40
40 ... 49	20
50 ... 59	10

Assinale a opção **incorreta**.
a) O intervalo de classe modal é dado por [30; 39].
b) O tempo médio de auditoria é dado por 34,5 min.
c) A mediana, a moda e a média da distribuição são coincidentes.
d) A distribuição é assimétrica.
e) 30% das auditorias demoram menos de trinta minutos.

2.2 – Medidas de Dispersão ou Variabilidade

Vimos que a **moda**, a **mediana** e a **média aritmética** possuem a função de representar, a partir de um único número, a seqüência a ser analisada. Porém, tal método ainda é muito incompleto para que nós possamos tirar alguma conclusão sobre o trabalho. É necessário que possamos enxergar algo mais nessa seqüência que estamos analisando, como, por exemplo, certa "personalidade" da seqüência. Observe a seguinte situação: quatro turmas, uma de cada um dos cursos Ciência da Computação, Matemática, Ciências Contábeis e Fisioterapia, fizeram uma prova de estatística e quando o professor verificou a média das notas de cada turma, constatou que, em cada uma das quatro turmas, a média dos alunos foi igual a 6,0. E aí? Será que podemos concluir que o desempenho das quatro turmas foi o mesmo? Será que todos os alunos, de todas as turmas, tiraram nota 6,0 na prova? É óbvio que, nesse momento, o bom senso fala mais alto e podemos, no mínimo, desconfiar de que não. Pois é exatamente aí que reside a tal "personalidade" que podemos atribuir a cada turma em relação ao comportamento das notas. O que quero dizer é que, com as **medidas de dispersão**, seremos capazes de verificar que, por mais que a média das turmas na prova de estatística tenha sido 6,0, poderemos com tais medidas determinar as turmas que tiveram um comportamento homogêneo, em que os alunos tiraram notas próximas de 6,0, como também determinar as turmas que tiveram um comportamento heterogêneo em relação à nota 6,0, ou seja, por mais que a média tenha sido 6,0, as notas não foram próximas de 6,0. Em outras palavras, torna-se necessário estabelecer medidas que indiquem o grau de dispersão em relação ao valor central. Algumas medidas de dispersão que sintetizam essa variabilidade são:

2.2.1 – Amplitude (H)

É uma medida de dispersão muito rápida e, ao mesmo tempo, **muito imprecisa**, pois consiste simplesmente em verificar a diferença entre o maior valor e o menor valor obtido na coleta de dados. Essa é nossa velha conhecida. Mesmo assim um exemplo

Pessoas	Peso (kg)
Agulha	30
Aderbal	15
Corá	55
Renato	52
Guilherme	60
Bruno	53
Bertolo	75
Alexandre	20
Fábio Thomáz	40

Na tabela ao lado, temos o peso das pessoas de um determinado grupo analisado e podemos verificar que a amplitude total foi de:AT = 75 – 15 = 60

2.2.2 – Desvio Médio

Como a palavra desvio está associada à diferença, temos que, no contexto da nossa matéria, o desvio deve ser empregado com a diferença do elemento analisado em relação à média, ou seja, o quanto o elemento se afasta da média da seqüência. Daí é importante perceber que essa diferença deve ser necessariamente trabalhada em módulo, pois não tem sentido a distância negativa[1]. E o desvio médio, então, passa a ser encontrado a partir da média aritmética de todos os desvios.

$$\text{Desvio Médio} = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + |x_3 - \bar{x}| + \cdots + |x_N - \bar{x}|}{N} = \frac{\sum_{i=1}^{N}|x_i - \bar{x}|}{N}$$

Exemplo: Com os dados do exercício anterior, temos:

$$\bar{x} = \frac{30 + 15 + 55 + 52 + 60 + 53 + 75 + 20 + 40}{9} = 44{,}4$$

$$\text{Desvio Médio} = \frac{|30 - 44{,}4| + |15 - 44{,}4| + |55 - 44{,}4| + |52 - 44{,}4| + |60 - 44{,}4| + |53 - 44{,}4| + |75 - 44{,}4| + |20 - 44{,}4| + |40 - 44{,}4|}{9}$$

$$= 16{,}17$$

2.2.2 – Variância

A variância é uma medida de dispersão muito parecida com o desvio médio, a única diferença em relação a este é que, na variância, ao invés de trabalharmos em módulo as diferenças entre cada elemento e a média, tomamos os quadrados das diferenças. Isso se dá pelo fato de que, elevando cada diferença ao quadrado, continuamos trabalhando com números não negativos, como também pelo fato de que, em procedimentos estatísticos mais avançados, tal método facilita futuras manipulações algébricas.

$$\text{Variância } \sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

Exemplo: Com os dados do exercício anterior, temos:

$$\text{Variância} = \frac{(30 - 44{,}4)^2 + (15 - 44{,}4)^2 + (55 - 44{,}4)^2 + (52 - 44{,}4)^2 + (60 - 44{,}4)^2 + (53 - 44{,}4)^2 + (75 - 44{,}4)^2 + (20 - 44{,}4)^2 + (40 - 44{,}4)^2}{9}$$

$$= 345{,}57$$

2.2.3 – Desvio-padrão

[1] E também porque é fácil ver que a soma dos desvios, é identicamente nula e que portanto não serve como medida de dispersão: $\sum_{j=1}^{N}(x_j - \bar{x}) = \sum_{j=1}^{N}(x_j) - \sum_{j=1}^{N}(\bar{x}) = N\,\bar{x} - N\,\bar{x} = 0$. Por isso temos duas opções: a) considerar os desvios em valor absolutos ou b) considerar os quadrados dos desvios.

Para entendermos o procedimento para o cálculo do desvio-padrão, é interessante percebermos que, no cálculo da variância, tal como vimos no tópico anterior, cometemos um "erro técnico" que será corrigido pelo desvio-padrão, ou seja, no momento em que elevamos ao quadrado as dispersões (diferenças) de cada elemento em relação à média, automaticamente alteramos a **unidade** de trabalho. Por exemplo: se estivermos trabalhando com a coleta das alturas, em metro, das pessoas de uma determinada comunidade, a unidade da variância encontrada será o m^2 (metro quadrado), que representa áreas. E é aí que entra o desvio-padrão, ou seja, extraindo a raiz quadrada da variância.

$$\text{Desvio} - \text{padrão}\ \sigma\ = \sqrt{\text{Variância}}$$

Então, se no exemplo do item anterior a variância encontrada foi 345,57, temos que o desvio-padrão foi de $\sqrt{345{,}57} = 18{,}58$

Observação: O uso do Desvio Médio pode causar dificuldades quando comparamos conjuntos de dados com números diferentes de observações:

Exemplo: Em A = {3,4,5,6,7} temos o Desvio Médio (DM) como 6/5 = 1,2 e σ^2 = 10/5 = 2

Em D = {3,5,5,7} temos o Desvio Médio (DM) = 1,0 e σ^2 = 2

Assim, podemos dizer que, segundo o Desvio Médio, o grupo *D* é mais homogêneo (tem menor dispersão) do que *A*, enquanto que ambos têm a mesma homogeneidade segundo a variância. O desvio médio possui pequena utilização em estatística e em geral vale 0,8 vezes o desvio padrão

O cálculo do desvio padrão exige o cálculo prévio da variância e uma fórmula alternativa para S^2 é dada por:

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

Relacionados à inferência estatística, alguns autores usam (*n* - 1) como divisor para a variância:

$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$, e isto será visto adiante (tendenciosidade)

Obs.: Muitas calculadoras científicas possuem duas medidas para desvio padrão. Uma associada à divisão por *n* (simbolizada geralmente por σ ou σ_n) e outra associada à divisão por *n* - 1 (chamada também de não-polarizada, simbolizada geralmente por *S* ou σ_{n-1}). Verifique a simbologia usada pela sua calculadora, caso você possua uma!

Para dados agrupados em classes, a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2 \,.\, n_i}{N} = \sum_{i=1}^{N}(x_i - \bar{x})^2 \,.\, f_i$$

2.2.4 – Momentos de uma distribuição de freqüências

Definimos o momento de ordem *t* de um conjunto de dados como:

$$M_t = \frac{\sum_{i=1}^{N}(x_i)^t}{N}$$

Definimos o momento de ordem *t* centrado em relação a uma constante *a* como

$$M_t = \frac{\sum_{i=1}^{N}(x_i - a)^t}{N}$$

Especial interesse tem o caso do momento centrado em relação a $\bar{x}$, dado por:

$$m_t = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^t}{N}$$

Conforme já vimos nos casos da média e da variância, as expressões precedentes podem ser reescritas levando-se em consideração as freqüências dos diferentes valores existentes. Temos então respectivamente,

$$M_t = \frac{\sum_{i=1}^{N}(x_i)^t \,.\, f_i}{N}$$

$$M_t = \frac{\sum_{i=1}^{N}(x_i - a)^t \,.\, f_i}{N}$$

$$m_t = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^t \cdot f_i}{N}$$

É fácil ver que $M_1 = \bar{x}$; $m_1 = 0$; $m_2 = \sigma^2$.

2.2.5 – Coeficiente de variação (CV)

O coeficiente de variação exprime a variabilidade em termos relativos. É uma medida adimensional e sua grande utilidade é permitir a comparação das variabilidades em diferentes conjuntos de dados.

$$CV = \frac{\sigma}{\bar{x}}$$

Exemplo: Testes de resistência à tração, aplicados a dois tipos diferentes de aço:

	$\bar{x}$ (kg/mm^2)	σ (kg/mm^2)
Tipo I	27,45	2,0
Tipo II	147,00	17,25

$CV_I = 2/27{,}45 = 7{,}29\%$
$CV_{II} = 17{,}25/145 = 11{,}73\%$
Assim, apesar do Tipo I ser menos resistente, é ele mais estável, mais consistente.

O uso do coeficiente de variação pode ser pensado considerando a questão: Um desvio padrão de 10 se a média é 10.000 é bem diferente se a média é 100!

Nas questões matemáticas não se compreende a incerteza nem a dúvida, assim como tampouco se podem estabelecer distinções entre verdades médias e verdades de grau superior.
(Hilbert)