2.2.2. Medidas de dispersão

Depois de localizar os valores "centrais" ou "típicos" de uma distribuição, a segunda tarefa mais importante das medidas descritivas é geralmente a de quantificar a *dispersão* das observações, isto é, medir quão dispersas elas estão, em relação ao centro da distribuição. Dispersão é a razão de ser da Estatística; se ela não existisse (se todas as observações fossem iguais), a Estatística não seria necessária.

Estas medidas servem, primeiro, para comparar numericamente a dispersão de duas distribuições. Isto é importante em vários problemas; por exemplo, no controle de qualidade num processo de fabricação. Suponha uma fábrica que produza peças do tipo A, que depois devem ser encaixadas em peças do tipo B. É claro que todas as peças A devem ser feitas exatamente do mesmo tamanho; se uma delas sair maior (ou menor) do que o projetado, não será possível encaixá-la exatamente na B, e a peça terá que ser descartada, ou ajustada – o que, de qualquer forma, causará prejuízo. Se existirem dois processos para fabricar estas peças, o melhor será aquele que conseguir produzir peças com *menor* dispersão nas dimensões; precisamos, portanto, de uma medida para esta dispersão.

Outras áreas nas quais medidas de variação são extremamente importantes são as ligadas à Economia e às Finanças. Isto pode ser entendido por meio de um exemplo simples: se os rendimentos de um pessoa (salário, investimentos, etc.) são mais ou menos constantes a cada mês, esta pessoa pode planejar seus gastos e poupanças mensais; se os rendimentos porém variam muito de mês para o outro, de forma aleatória, fica muito difícil planejar algo.

Segundo, as medidas de dispersão fornecem um medida de *escala,* isto é, fornecem uma unidade com a qual devemos medir as distâncias entre as observações, ou entre as observações e a média ou mediana. Por exemplo, suponha que duas turmas fizeram provas de uma mesma disciplina, e em ambas a mediana das notas foi 70. Se um aluno tirou nota 60, ele ficou abaixo da mediana - mas *quanto* abaixo? Esta nota 60 deve ser considerada *regular* ou *ruim*? Isto vai depender da dispersão da distribuição das notas. Veja os gráficos da Fig. **1**: para a turma B (que teve grande dispersão), uma nota de 60 pode ser considerada *regular*; para a turma A (que teve pequena dispersão), esta mesma nota seria considerada *ruim*. Embora a nota esteja 10 pontos abaixo da mediana, nas duas turmas, diremos que a *distância estatística* entre a nota e a mediana é maior na turma A.

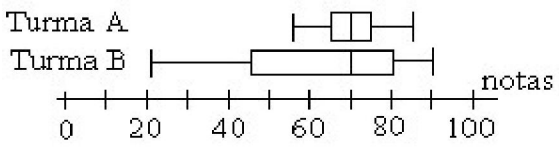

Figura 1

2.2.2.1. Amplitude total (AT)

A *amplitude total* (*range*) é a diferença entre os valores extremos de uma distribuição:

AT = valor máximo – valor mínimo

Esta diferença indica a distância em que estes pontos estão um do outro; quanto mais afastados estiverem entre si, mais dispersa será a distribuição. A idéia de avaliar a dispersão de uma distribuição através de seus valores extremos é a que empregamos informalmente, quando dizemos, por exemplo, que "entre as provas que fiz este ano, minha menor nota foi 55, a maior foi 90" (a amplitude, portanto, é de 90-55=35)". A AT é por isso uma medida muito fácil de compreender, e sua interpretação gráfica também é simples.

A Fig. **2** mostra três distribuições que tê mesmo número de pontos (n=12), mesma média (igual a 3), são unimodais e simétricas. As três diferem contudo na dispersão.

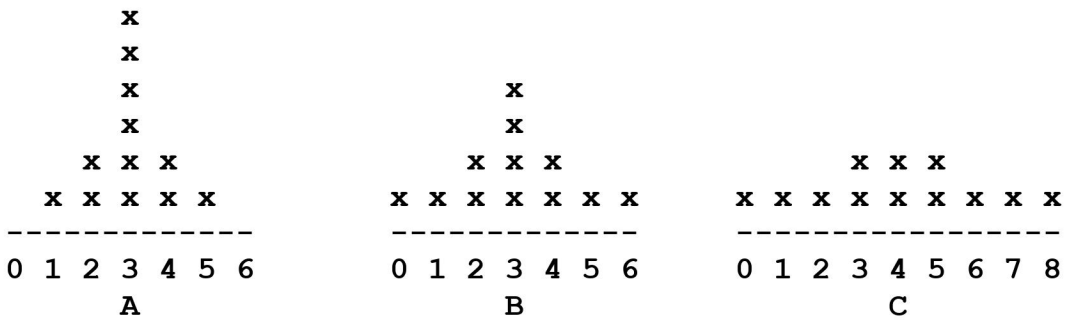

Figura 2

Calculando os valores de AT para as três distribuições, obtemos

$AT_A = 5-1 = 4$ $AT_B = 6-0 = 6$ $AT_C = 8-0 = 8$

o que confirma que a distribuição A é a menos dispersa, e a C é a mais dispersa.

Contudo, como vimos ao comparar a média com mediana, é sempre importante em Estatística não confiar muito em medidas descritivas cujos valores dependem de apenas algumas observações. A AT é uma destas medidas suspeitas, pois seu valor depende apenas dos dois extremos e não leva em conta nenhuma das observações intermediárias. Pode, por isto, ter seu valor exagerado pelo efeito de um ponto discrepante. Os três gráficos da Fig. **3** ilustram este problema.

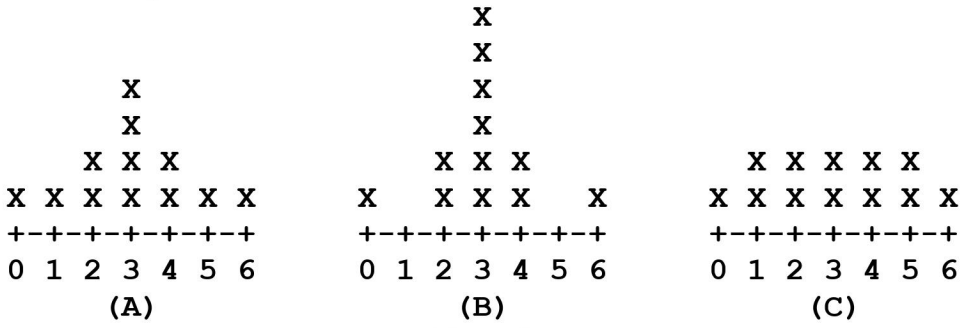

Figura 3

Estas três distribuições são simétricas, têm 12 pontos cada, têm mesma AT, média e mediana. É evidente porém que suas dispersões não são iguais; a mais dispersa é a distri-

buição *C*. A menos dispersa é a distribuição *B*; sua AT, porém, é igual a das outras distribuições, porque foi inflacionada pelos dois pontos discrepantes (0 e 6); se não houvesse estes dois pontos, sua AT seria a menor de todas, confirmando que ela tem menor dispersão.

2.2.2.2. Intervalo quartílico (IQ)

O *intervalo quartílico* (*interquartile range*) é a diferença entre os valores do terceiro e do primeiro quartil de uma distribuição:

$$IQ = Q_3 - Q_1$$

A idéia básica é a mesma da AT; usar como medida de dispersão a distância entre dois pontos de referência na distribuição. A vantagem do IQ sobre a AT está na sua maior *robustez*: o valor da AT depende de apenas duas observações (os dois extremos), e é portanto muito influenciado pelos pontos discrepantes; o valor do IQ depende da posição dos 50% de pontos que estão mais próximos do centro da distribuição (da mediana), e é portanto imune ao efeito dos pontos discrepantes. A Fig. **4** ilustra isto.

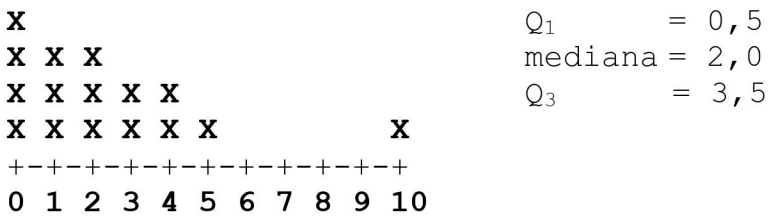

Figura 4

Suponha que o ponto x=10 tenha seu valor alterado para 5, por exemplo. O ponto deixaria de ser discrepante, e a AT seria alterada, de AT=10 para AT=5. O IQ, contudo, permaneceria constante (IQ = 3,5 - 0,5 = 3,0).

O IQ é portanto uma medida *robusta* (o conceito de *robustez* foi discutido ao compararmos média e mediana, na Seção **4.2.1.3**), que tem ainda a vantagem de ser fácil de interpretar graficamente: no diagrama de Tukey, o IQ é simplesmente a largura da "caixa" que contém a metade central dos dados, i.e., aqueles localizados mais próximos da mediana da distribuição. Por isso, o IQ é uma medida bastante usada em Estatística Descritiva, principalmente quando trabalhamos com medianas e diagramas de Tukey.

Dissemos acima que as medidas de dispersão são úteis como medidas de *escala*. A regra sugerida por Tukey para identificação de valores discrepantes (seção **2.1.3**) faz justamente isto: usa o IQ como medida de escala, e considera discrepantes os valores que estiverem a mais de 1,5×IQ afastados da caixa central do diagrama:

acima de $Q_3 + 1{,}5 \times IQ$ ou
abaixo de $Q_1 - 1{,}5 \times IQ$

Um outro exemplo do uso do IQ como medida de escala pode ser visto no exemplo que mencionamos no início deste texto (Fig. **1**). Dissemos que, para a turma B, uma nota de 60 pode ser considerada *regular*; para a turma A, esta mesma nota considerada *ruim*. Se

calcularmos a distância destas notas em relação à mediana, usando como unidade o IQ, teremos:

Turma A → IQ_A = 75-65=10; $dist_A$ = (60-70)/10 = -1,00
Turma B → IQ_B = 80-45=35; $dist_B$ = (60-70)/35 = -0,29

Portanto, a *distância estatística* entre a nota e a mediana é maior na turma A do que na turma B; por isso, consideramos que a nota 60 indica um pior resultado na turma A, porque a nota então mais afastada (estatisticamente) da mediana.

Voltando aos exemplos da Fig. **3** : para os três gráficos, os IQs são:

(A) Q_1= 2 Q_3= 4 → IQ_A= 4-2 = 2
(B) Q_1= 2,5 Q_3= 3,5 → IQ_B= 3,5-2,5 = 1
(C) Q_1= 1,5 Q_3= 4,5 → IQ_C= 4,5-1,5 = 3

o que confirma que a distribuição mais dispersa é a C, e a menos dispersa é a B.

2.2.2.3. Desvio médio (DM)

O "desvio" de um valor, em Estatística, é a diferença entre este valor e a média da distribuição, isto é:

$$desvio = x - \overline{X}$$

Se uma distribuição é considerada "dispersa" quando seus pontos se afastam muito da média, parece natural que uma medida de dispersão possa ser criada a partir dos desvios. Uma primeira idéia seria a de tomar simplesmente a média destes desvios:

$$\frac{\sum_{i=1}^{n}(x_i - \overline{X})}{n}$$

Uma média alta indicaria que os pontos estão em geral muito afastados do centro da distribuição; a distribuição seria portanto muito dispersa. Se tentarmos colocar esta idéia em prática, contudo, veremos que ela não funciona. Suponhamos uma distribuição com os valores (0, 2, 2, 2, 3, 3, 5, 7), representados no gráfico da Fig. **5**.

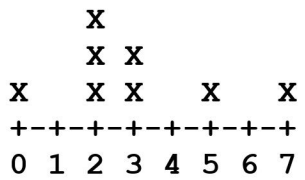

Figura 5

A média desta distribuição é $\overline{X} = 3$. A média dos desvios é

$$média\ dos\ desvios = \frac{(0-3)+(2-3)+(2-3)+(2-3)+(3-3)+(3-3)+(5-3)+(7-3)}{8} = 0$$

A média será nula porque o somatório dos desvios negativos (desvios das observações que estão abaixo da média) será idêntico ao somatório dos desvios positivos (desvios das observações que estão acima da média). Este fato não é uma peculiaridade desta distribuição, mas sim algo que acontece com qualquer distribuição: os desvios positivos sempre equilibram os negativos – é exatamente por isso que a média é considerada o "centro", já que os desvios à direita equilibram os desvios à esquerda.

Para evitar que os desvios negativos anulem os positivos, podemos simplesmente eliminar os sinais dos desvios e considerar apenas os valores absolutos. Se fizermos isto, obtemos uma medida chamada de "desvio absoluto médio" ou simplesmente "desvio médio", que é expressa da forma:

$$DM = \frac{\sum_{i=1}^{n} |x_i - \overline{X}|}{n} \qquad (1)$$

Calculando o valor desta medida para os dados do exemplo acima, obtemos

$$DM = \frac{|0-3|+|2-3|+|2-3|+|2-3|+|3-3|+|3-3|+|5-3|+|7-3|}{8} = 1{,}5$$

Um DM elevado indica simplesmente que as observações tendem a estar muito afastadas do centro da distribuição, tanto para um lado quanto para o outro; a distribuição, em conseqüência, pode ser considerada muito dispersa. É importante aqui chamar a atenção para a definição do conceito de "dispersão". Sem a definição exata, não é possível dizer qual das duas distribuições na Fig. **6** é mais dispersa.

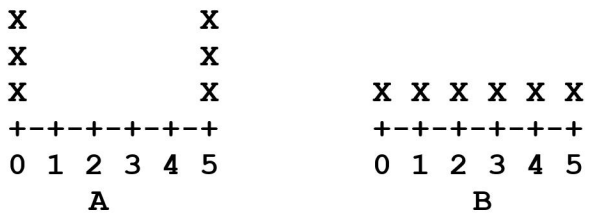

Figura 6

Como a *dispersão* é definida como função do afastamento dos pontos em relação à média, a distribuição mais dispersa é A, pois todos os pontos estão afastados ao máximo. Poderia ser argumentado que na distribuição B os pontos estão mais "espalhados" ao longo do eixo do que na distribuição A; isto implica que esta distribuição tem maior *entropia*, mas este conceito, originário da Teoria da Informação, não é usado na Estatística básica.

O DM tem como vantagem a facilidade de interpretação, mas tem por outro lado uma desvantagem técnica. Medidas de dispersão serão fundamentais para o desenvolvimento da *Inferência Estatística,* i.e., dos métodos que nos permitem tirar conclusões sobre uma população a partir de uma amostra. Estes métodos se baseiam, resumidamente, na análise da diferença entre os valores encontrados na amostra e os valores previstos por um modelo teórico; as medidas de dispersão serão fundamentais nesta análise, e suas fórmulas estão na base dos modelos. O problema com o DM é a presença de módulos na sua fórmula. A função *módulo* não tem boas propriedades algébricas. Por exemplo, não tem a *propriedade distributiva*; se temos uma expressão com parênteses da forma

a(b+c)

podemos simplificá-la, eliminando o parênteses:

a(b+c) = ab+ ac

Se temos uma expressão com módulo, contudo, não há como prosseguir, algebricamente:

a |b+c| = ?

Além disse, a função módulo não tem derivada contínua, que será essencial mais tarde. Estas dificuldades limitam a aplicação do DM; é uma medida simples de se calcular e de interpretar graficamente, mas que terá pouca utilidade nos capítulos mais avançados da Estatística.

Voltando aos exemplos da Fig.**3**; para as três distribuições, os DMs são:

```
DM_A= 1.17      DM_B= 0.83      DM_C= 1.50
```

o que confirma novamente que a distribuição mais dispersa é a C, e a menos dispersa é a B.

2.2.2.4. Variância (s^2) e desvio-padrão (s)

A *variância* (*variance*) é outra medida que se baseia nos desvios das observações para quantificar a dispersão. Como vimos, a média desses desvios não serve como medida, pois será sempre nula. A média dos módulos destes desvios pode ser usada, e dá origem ao DM, mas não é conveniente, do ponto de vista algébrico.

A dificuldade, portanto, é: como eliminar os sinais dos desvios, sem usar a função módulo? Este problema é encontrado em vários outros ramos da Estatística (por exemplo, no desenvolvimento de *modelos de regressão*) e a solução é sempre a mesma: elevar os desvios ao quadrado. Com isso, os valores negativos desaparecerão, já que qualquer número ao quadrado se torna positivo. Na fórmula do DM, simplesmente substituímos os módulos por quadrados:

$$DM = \frac{\sum_{i=1}^{n} |x_i - \bar{X}|}{n} \qquad \rightarrow \; s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^2}{n} \qquad (2)$$

A medida resultante é chamada de *variância*, simbolizada por s^2. Aplicando esta medida sobre os dados do exemplo anterior (Fig. **5**), temos:

$$s^2 = \frac{(0-3)^2 + (2-3)^2 + (2-3)^2 + (2-3)^2 + (3-3)^2 + (3-3)^2 + (5-3)^2 + (7-3)^2}{8} = 4$$

A variância é extremamente importante; é com certeza a medida mais importante e mais usada da Estatística, depois da média aritmética. Iremos por isso mencioná-la em praticamente todos os capítulos subseqüentes.

No entanto, mesmo a variância apresenta alguns problemas. O primeiro deles é o das *unidades* empregadas. No exemplo acima, suponhamos que as observações se referem ao número de crianças em famílias de um bairro. A média aritmética tem a mesma unidade das observações, portanto será igual a 3 *crianças*. A unidade da variância, contudo, será o quadrado do unidade original, já que os desvios foram elevados ao quadrado. Portanto,

s^2 = 4 crianças^2

Esta unidade, "crianças ao quadrado", não faz nenhum sentido, e dificulta a interpretação da medida. A saída aqui é criar uma nova medida de dispersão, dada pela raiz quadrada da variância. Esta medida será representada por s, e denominada *desvio-padrão* (*standard deviation*). Sua expressão é, portanto:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{n}} \qquad (3)$$

A unidade desta medida será a mesma unidade das observações; no exemplo acima, o desvio-padrão será

$$s = \sqrt{s^2} = \sqrt{4} = 2\,\text{crianças}$$

O segundo problema no uso, tanto da variância quando do desvio-padrão, é que nenhuma delas tem interpretação geométrica óbvia; isto é, não é fácil descobrir como elas se relacionam com o gráfico de uma distribuição. As medidas que vimos anteriormente (AT, IQ, DM) são medidas de *distâncias*, que podem ser facilmente identificadas em qualquer gráfico:

AT: distância entre os pontos extremos da distribuição
IQ: distância entre o terceiro e o primeiro quartil
DM: distância média entre as observações e o centro da distribuição

e são por isso conceitos fáceis de serem compreendidos. Na variância, por outro lado, os desvios foram elevados ao quadrado, e o resultado é uma média de distâncias ao quadrado, o que não faz muito sentido geometricamente.

Existe em geral uma relação entre os gráficos de uma distribuição e o valor de suas medidas. A partir de um gráfico de pontos ou de ramo-e-folhas, por exemplo, podemos estimar quais serão aproximadamente a média a e mediana da distribuição, além de sua AT. Com um pouco de prática, podemos também ter uma idéia dos valores de seu IQ e DM. A recíproca também é verdadeira; a partir das medidas, podemos ter uma idéia aproximada de como seria o gráfico de uma distribuição. Com a variância e o desvio-padrão, contudo, isto não acontece. É difícil entender a relação que existe entre um gráfico e estas medidas; é difícil ter uma idéia de qual seria o desvio padrão de uma distribuição, dado o seu gráfico, ou de imaginar como seria o gráfico, dado o desvio-padrão.

Há ainda um terceiro problema: como a variância e o desvio-padrão são baseados na soma dos quadrados dos desvios, são medidas muito sensíveis aos valores discrepantes; os desvios destes valores serão grandes e, quando elevados ao quadrado ficariam ainda maiores, e pesariam muito na soma.

Em compensação (e isto é o mais importante!), ambas as medidas têm excelentes "propriedades amostrais"; ou seja, é fácil demonstrar matematicamente a relação existente entre a variância de uma amostra e a variância da população de onde esta amostra foi retirada. Por causa destas propriedades, a variância e o desvio-padrão são as medidas de dispersão que estão na base de todas as técnicas mais usadas de Inferência Estatística, como veremos nos próximos capítulos..

Voltando aos exemplos da Fig.**3**; para as três distribuições, as variâncias e desvios-padrões são:

$s^2_A = 2,50$ $s^2_B = 1,83$ $s^2_C = 3,17$
$s_A = 1,58$ $s_B = 1,35$ $s_C = 1,78$

(i) Desvio-padrão como medida de escala

Veremos mais tarde que o desvio-padrão desempenha um papel fundamental como medida de *escala*, isto é, como unidade para a medição de distâncias entre as observações e o centro de uma distribuição (no modelo *normal*, seção **3.4.4**). Por enquanto, como primeiro exemplo desta utilização, citaremos um teorema que relaciona a proporção de observações de uma distribuição que se encontram além de um certo ponto, e a distância deste ponto em relação ao centro da distribuição. O teorema, demonstrado pelo matemático russo Tchebishev, afirma que:

> A proporção de observações de uma distribuição que diferem da média em mais de *k* desvios-padrões é de, no máximo, $1/k^2$

Isto quer dizer, basicamente, que grandes afastamentos da média são pouco prováveis. Se denotarmos a média e o desvio-padrão pelas letras gregas μ (mu) e σ (sigma), respectivamente, este teorema é expresso da forma:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Portanto, quanto maior o afastamento $|X-\mu|$, menor a proporção. A proporção de observações a mais de 2 desvios-padrões acima ou abaixo da média será de no máximo ¼; a proporção a mais de 3 desvios-padrões, será de no máximo 1/9:

$$P(|X-\mu| \geq 3\sigma) \leq \frac{1}{3^2} = \frac{1}{9}$$

Este teorema é um exemplo da utilização do desvio-padrão como *medida de escala*: o afastamento dos pontos em relação à média foi medido em termos da quantidade de desvios-padrões acima ou abaixo da média; o desvio-padrão foi portanto usado como unidade. (Voltaremos a falar dele quando estudarmos a distribuição *normal*, na qual o desvio-padrão também é usado como medida de escala, seção **3.4.4**).

(ii) Variância e desvio padrão corrigidos

Na Eq. **2**, a variância foi definida pela média dos quadrados dos desvios das observações em relação à média aritmética. Existe, contudo, um problema com esta versão da variância: ela não é muito útil para a *inferência estatística*, isto é, para tirar conclusões sobre uma população a partir de uma amostra. Se a usarmos para calcular a variância de uma amostra, e depois quisermos usar o valor obtido como estimativa da variância da população, esta estimativa será *subestimada*; isto é, será menor do que o valor verdadeiro.

Para corrigir isto, pode ser demonstrado que uma estimativa melhor é obtida usando a variância *corrigida*, dada por:

$$s^2_{corr} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2 \qquad (4)$$

O mesmo acontece com o desvio-padrão; sua versão corrigida é dada por:

$$s_{corr} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} \qquad (5)$$

A diferença entre as eqs. (2) e (4), e as eqs. (3) e (5), é que usamos $n-1$ no denominador, em vez de n. É claro que, se a amostra for grande, isto não vai haver muita diferença; por exemplo, se n=100, dividir por 100 ou dividir por 99 dá quase no mesmo. Quando a amostra for pequena, no entanto, é preciso tomar cuidado. Se n=5, por exemplo, dividir por 5 ou dividir por 4 leva a uma diferença de 20% no resultado.

Alguns programas de Estatística permitem calcular as duas versões da variância e do desvio-padrão; outros calculam apenas uma. O *R* calcula apenas as versões corrigidas. Verifique, portanto, quando usar um programa, qual das das versões da variância e do desvio-padrão estão sendo calculadas.

(iii) Cálculo da variância de dados agrupados

Como acontece com a média (seção **2.2.1.5**), a variância de uma distribuição pode também ser calculada aproximadamente, se os dados estão agrupados em uma tabela de distribuição de frequências. Se fizermos a suposição de que todos as observações em uma classe têm valor igual ao ponto médio *PM* daquela classe, a variância pode ser aproximada pela eq. (**6**):

$$S^2 = \frac{\sum_{j=1}^{NC}(PM_j - \overline{X})^2 f_j}{\sum_{j=1}^{NC} f_j} \qquad (6)$$

onde: PM_j ponto médio da classe j
f_j freqüência da classe j
NC número de classes

É claro que, na realidade, os pontos não têm todos valores iguais ao PM da classe. Se a distribuição for unimodal e razoavelmente simétrica, isto não deve causar problemas para o cálculo aproximado da média, já os desvios dos pontos acima e abaixo dos PMs acabam se equilibrando. No gráfico da Fig. **7**, por exemplo, os dados estão agrupados em 5 classes, como na Tabela **1**. Na segunda classe (X=3 a X=5), há cinco pontos acima do PM, e três abaixo; na quarta classe (X=9 a X=11), a situação se reverte, o que faz com que os erros desta classe compensem os da outra.

Para o cálculo da variância, contudo, isto não funciona, já que os desvios são elevados ao quadrado; o resultado é que a variância aproximada calculada pela eq. (**6**) tende sempre a ser maior do que a variância verdadeira.

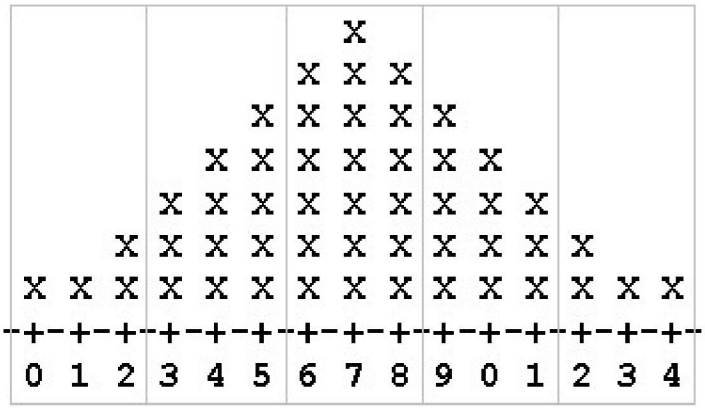

Figura 7

Tabela 1

X	f
0 – 2	4
3 – 5	12
6 – 8	19
9 – 11	12
12 – 14	4
Σ	51

Para evitar isto, pode ser usada a *correção de Sheppard*, que introduz na eq. (**6**) um fator de correção que é função do intervalo de classe IC (eq. **7**):

$$S^2_{corr} = \frac{\sum_{j=1}^{NC}(PM_j - \overline{X})^2 f_j}{\sum_{j=1}^{NC} f_j} - \frac{IC^2}{12} \qquad (7)$$

onde *IC*: intervalo de classe. (Esta forma de correção, contudo, só funciona bem se a distribuição for unimodal e simétrica, como a do exemplo; a fórmula tem que ser modificada, para outros tipos de distribuição).

2.2.2.5. Coeficiente de variação (CV)

Todas as medidas de dispersão vistas até agora sofrem de duas limitações. Primeiro: só servem para comparar variáveis que tenham a mesma unidade. Se queremos verificar se a variação é maior no *peso* do que na *altura* dos indivíduos de uma amostra, por exemplo, não podemos usar as medidas AT, IQ ou desvio-padrão, pois a dispersão do peso estará expressa em *quilogramas*, a da altura em *centímetros*; estaríamos comparando medidas em duas unidades diferentes, o que não faz sentido. O ideal é usar uma medida *adimensional*, isto é, uma que não dependa das unidades das variáveis.

Segundo: só podemos comparar através destas medidas distribuições que tenham aproximadamente a mesma média. Para ilustrar isto, suponha que tenhamos duas amostras, uma de crianças (distribuição A) e outra de homens idosos (distribuição B), cujas idades estão representadas nos gráficos da Fig. **8**:

Figura 8.

Estas distribuições têm exatamente a mesma forma, unimodal e simétrica, e têm médias iguais a 2 anos e 62 anos, respectivamente. Se compararmos as medidas de dispersão destas duas distribuições, seremos levados a concluir que elas têm a mesma dispersão, já que em ambas AT = 4 anos, IQ = 2 anos, s = 1,15 anos. Contudo, não podemos deixar de reconhecer que o grupo das crianças é muito mais heterogêneo do que o dos homens; existe muito mais diferença entre uma criança de 0 e uma de 4 anos, por exemplo, do que entre um homem de 60 e um de 64 anos; precisamos, de alguma forma, relacionar estas medidas de dispersão ao *nível* da distribuição.

Isto pode ser feito, por exemplo, se tomarmos as razões entre o desvio e a médias: a criança mais velha tem uma idade que é o dobro da média (4/2 = 2), enquanto o homem mais velho tem uma idade que é apenas 3% maior que a média (64/62 = 1,03). O *coeficiente de variação* (*coefficient of variation*) é uma medida que aproveita esta idéia, e compara o desvio-padrão e a média, como na eq. (6).

$$CV = 100 \times \frac{s}{\overline{X}} \tag{6}$$

Se comparamos as dispersões das distribuições A e B usando esta medida, obtemos:

$$CV_A = \frac{s}{\overline{X}} = 100 \times \frac{1{,}15}{2} = 58\ \% \qquad CV_B = \frac{s}{\overline{X}} = 100 \times \frac{1{,}15}{62} = 1{,}8\ \%$$

Estes valores confirmam a idéia de que a variabilidade é muito maior no grupo das crianças (A) que no grupo dos homens (B).

Por usar a razão entre duas quantidades (desvio-padrão e média) que têm a mesma unidade, o CV também resolve a primeira limitação mencionada anteriormente: a unidade do denominador anula a do numerador, e o CV fica portanto adimensional. Podemos, por exemplo, comparar as dispersões dos pesos e das alturas de homens em uma amostra. Num grupo de 205 alunos de Medicina, com idade entre 19 e 25 anos, foram encontradas as seguintes medidas descritivas: para os pesos, média de 71,6 kg e desvio-padrão de 10,6 kg; para as alturas, média de 176,7 cm e desvio-padrão de 6,1 cm.

Intuitivamente, é fácil perceber que a variação do peso das pessoas deva ser maior que a das alturas. É possível, por exemplo, encontrarmos uma amostra de estudantes onde o mais pesado tenha o dobro do peso do estudante mais leve (por exemplo, um com 110 kg, o outro com 55 kg); é praticamente impossível, contudo, encontrar uma amostra onde o estudante mais alto tenha o dobro da altura do mais baixo. Não podemos comparar diretamente os desvios-padrões do peso (em kg) com o da altura (em cm); podemos porém comparar os CVs:

$$CV_{altura} = 100 \times \frac{s}{\overline{X}} = 100 \times \frac{6{,}1\ cm}{176{,}7\ cm} = 3{,}4\,\%$$

$$CV_{peso} = 100 \times \frac{s}{\overline{X}} = 100 \times \frac{10{,}6\ kg}{71{,}6\ kg} = 14{,}8\ \%$$

Concluímos então, com base nos CVs, que a dispersão do *peso*, nesta amostra, é bem maior que a da *altura*.

Como qualquer medida, o CV também tem suas limitações. A mais importante delas é que, por expressar o desvio-padrão como uma porcentagem da média, o CV não pode ser usado quando a média é nula (não podemos ter zero no denominador da razão).

2.2.2.6. Vantagens e desvantagens de cada medida

- Amplitude Total (AT): é a menos útil das medidas, pois é influenciada demais pelos pontos extremos (discrepantes ou não).
- Intervalo Quartílico (IQ): útil quando trabalhamos com diagramas de Tukey. Não é muito usado para Inferência Estatística.
- Variância (s^2) e desvio-padrão (s): são as medidas mais usadas para comparar a dispersão de duas variáveis de mesma unidade e aproximadamente a mesma média; muito úteis na *Inferência Estatística*.
- Coeficiente de Variação (CV): permite comparar as dispersões de distribuições de variáveis que tenham unidades diferentes (pesos e alturas, por exemplo) ou médias diferentes.