2.2.3. Medidas de assimetria e de curtose

As duas medidas que vimos, as de *posição* e de *dispersão*, são as mais importantes para a análise estatística, e as que você usará em seu trabalho ou encontrará em publicações com maior freqüência. Existem contudo algumas outras características da distribuição de um amostra que não podem ser descritas apenas em termos de localização e assimetria. Neste texto, veremos as medidas de *assimetria* e de *curtose* (ou *curvatura*).

2.2.3.1. Coeficiente de assimetria

(i) Coeficiente de Pearson

Na Seção **2.2.1**, vimos que existe uma relação entre a assimetria de uma distribuição e a posição relativa das diferentes medidas de posição. Em distribuições simétricas, média e mediana tem o mesmo valor; se a distribuição for unimodal, a moda também será igual a elas. O gráfico da Fig. **1A** ilustra esta situação; é fácil ver que

média = mediana = moda = 2

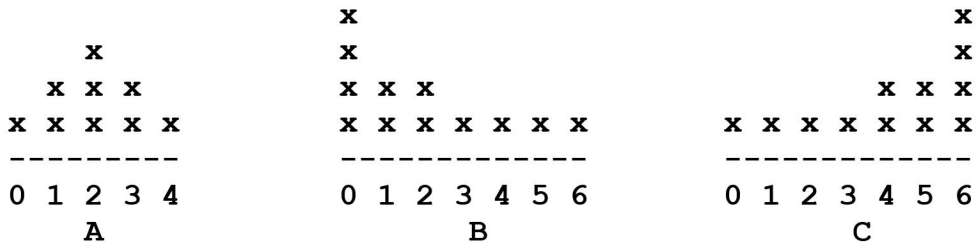

Figura 1. Exemplos de distribuições

Quando a distribuição é assimétrica, contudo, as três medidas de localização geralmente diferem entre si, e surge o problema de qual delas usar para representar o valor "típico" de uma distribuição. Se a distribuição é assimétrica positiva, com uma cauda mais longa à direita (Fig. **1B**), a média tende a ter um valor mais alto do que a mediana ou a moda, porque seu valor será influenciado pelos pontos mais afastados. Na Fig. **1B**:

média = 2 mediana = 1,5 moda = 0

Se é assimétrica negativa, por outro lado, a média será *menor* que a mediana ou a moda, como na Fig. **1C**:

média = 4 mediana = 4,5 moda = 6

Estes exemplos mostram que a distância entre a média e a mediana, ou entre a média e a moda, podem ser usadas como base para uma medida de dispersão, já que

```
média - moda = 0        → distr. simétrica
média - moda > 0        → distr. assimétrica positiva
média - moda < 0        → distr. assimétrica negativa
```

No entanto, é claro que estas distâncias não dependem apenas da *assimetria*, mas também da *dispersão* da distribuição; numa distribuição assimétrica, quanto maior a dispersão, mais afastadas da média estarão as observações, e provavelmente também mais afastada da média estará a moda. Levando isto em conta, Karl Pearson criou uma medida de assimetria onde as distâncias entre as medidas de centro são padronizadas pelo desvio-padrão, como na eq. (1).

$$A = \frac{(média - moda)}{s} \qquad (1)$$

Considerando o fato, notado empiricamente por Pearson, de que

```
média - moda ≅ 3 × (média - mediana)
```

este coeficiente também costuma ser apresentado de outra forma, que relaciona média e mediana, ao invés de média e moda, como na eq. (2).

$$A = \frac{3 \times (média - mediana)}{s} \qquad (2)$$

O coeficiente A, logicamente, terá o mesmo sinal da assimetria (positivo ou negativo), e será nulo se a distribuição for simétrica. Nos três gráficos da Fig. 1, os valores dos coeficientes de assimetria (usando a fórmula na eq. (2)) serão:

$A_A = 0$ $A_B = 0,75$ $A_C = -0,75$

(ii) Coeficiente baseado no terceiro momento

Momentos, na Física e na Matemática, são quantidades baseadas na média das potências dos desvios em relação à média aritmética de uma amostra. A *ordem* do momento indica a potência.

Suponha uma amostra com valores $x_1, x_2, ..., x_n$. O *momento de primeira ordem* m_1 é dado por:

$$m_1 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X}) = 0$$

Vimos que este momento é sempre nulo (seção 2.2.2.3); uma das propriedades básicas da média aritmética é a de ser o ponto de equilíbrio dos desvios.

O *momento de segunda ordem* m_2 é dado pela média dos quadrados dos desvios,

$$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 \qquad (3)$$

e serve como uma medida de dispersão, a variância:

$$s^2 = m_2$$

O *momento de terceira ordem* m_3 é a média dos cubos dos desvios (eq. 4):

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^3 \qquad (4)$$

e pode servir como base para uma medida de assimetria. Se uma distribuição é simétrica, para cada ponto à direita da média, com um desvio positivo, haverá um ponto à esquerda, com um desvio negativo de mesmo módulo. Se somarmos os cubos dos desvios, o resultado também será nulo, já que os cubos manterão os mesmos sinais dos momentos originais (contrários entre si). Se a distribuição for assimétrica, contudo, a soma terá um sinal que indicará o tipo de assimetria existente. Suponhamos por exemplo a distribuição mostrada na Fig. **2**.

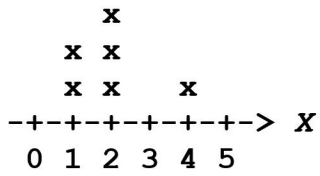

Figura 2.

A média é $x = 2$, e o somatório dos desvios é nulo, como esperado:

$$(1\text{-}2) + (1\text{-}2) + (4\text{-}2) = 0$$

Neste caso, há dois pontos à esquerda, próximos da média, que equilibram outro ponto, mais afastado à direita; isto é, há dois desvios negativos pequenos que anulam um desvio positivo maior. Se tomarmos os *cubos* destes desvios, contudo, o equilíbrio será desfeito, porque os valores maiores crescem mais rápido do que os menores, quando os elevamos ao quadrado. No exemplo, o cubo do desvio positivo será maior que a soma dos cubos dos desvios negativos, e o somatório resultante será positivo:

$$(1\text{-}2)^3 + (1\text{-}2)^3 + (4\text{-}2)^3 = -1 - 1 + 8 = 6$$

Portanto, embora o somatório dos *desvios* seja nulo, o somatório dos *cubos dos desvios* será positivo, indicando que a distribuição tem uma cauda mais longa à direita; isto é, que há alguns valores localizados muito acima da média.

Se o gráfico fosse invertido em torno de sua média, como na Fig**. 3**,

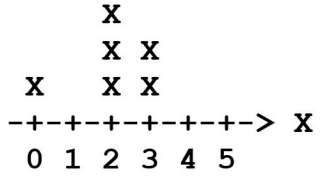

Figura 3.

o somatório dos cubos dos desvios seria negativo:

$$(0\text{-}2)^3 + (3\text{-}2)^3 + (3\text{-}2)^3 = -8 + 1 + 1 = -6$$

A partir desta constatação, foi criado o coeficiente de assimetria na eq. (**5**), baseado no terceiro momento (eq. **4**), padronizado pelo cubo do desvio-padrão, o que faz com que as unidades do denominador e do numerador se anulem e o coeficiente seja adimensional.

$$A = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^3}{s^3} = \frac{m_3}{s^3} \qquad (5)$$

Este coeficiente de assimetria é mais usado nos programas de computador do que o coeficiente de Pearson; é calculado por funções de vários pacotes do R (mas não do pacote *stats* básico). Sua interpretação é a mesma da do coeficiente de Pearson: o valor de *A* será positivo ou negativo, conforme a distribuição tenha assimetria positiva ou negativa, e nulo se a distribuição for simétrica.

2.2.3.2. Coeficiente de curtose (curvatura)

A última medida que veremos nesta seção, e a menos freqüentemente usada delas, é a que serve para comparar a "curvatura" de duas distribuições, o coeficiente de *curtose* (do grego *kurtosis*, que significa "curvatura").

Para definir este coeficiente, lançamos novamente mão do conceito de *momentos*. Usaremos o momento de quarta ordem; como em todos os momentos de ordem par, os sinais originais desaparecem, e o somatório resulta diferente de zero. Como estaremos elevando os desvios à quarta potência,, valores afastados da média (i.e., valores com grandes desvios), ainda que sejam pouco freqüentes, terão peso muito grande. Este momento é dividido pela 4ª. potência do desvio-padrão, o que torna o coeficiente unidimensional (eq. **6**).

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^4}{s^4} = \frac{m_4}{s^4} \qquad (6)$$

Como exemplo, vejamos as quatro distribuições de dados simulados na Fig. **4.**

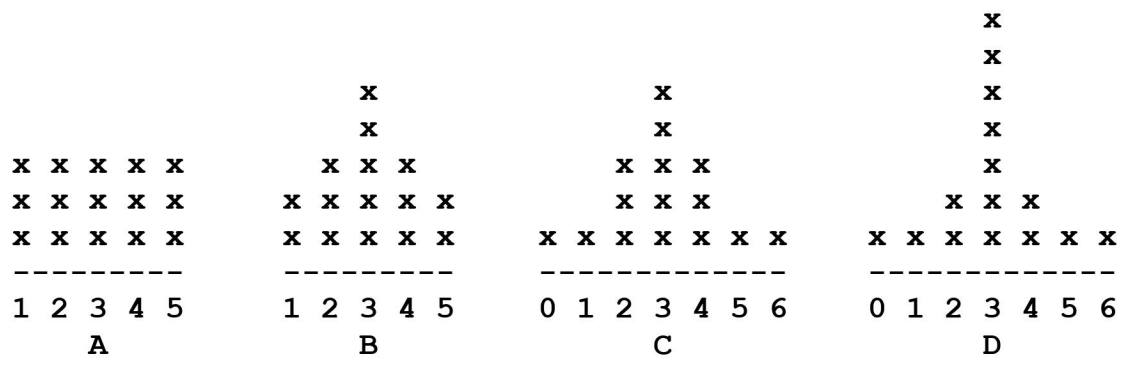

Figura 4. Distribuições com diferentes curtoses

A distribuição D é a que tem a maior concentração de observações iguais à média (o que faz o gráfico ficar "pontudo") e caudas mais finas e mais longas. Este tipo de forma é denominado de *leptocúrtica*. As distribuições A e B, por outro lado, têm um perfil mais

achatado, com caudas curtas e grossas, a parte central com pico menos acentuado; esta forma é chamada de *platicúrtica*.

A principal utilidade do coeficiente de curtose é auxiliar na comparação entre a forma da distribuição de uma amostra e a forma do modelo normal (seção **3.4.4**), que é exigido por várias técnicas da Inferência (por exemplo, os testes *t*, a análise de variância, e os modelos de regressão). Para que estas técnicas sejam usadas, é preciso verificar se a distribuição dos dados tem forma pelo menos aproximadamente igual à de uma curva normal; isto é, unimodal, mais ou menos simétrica, nem muito achatada e nem muito afilada.

Para uma curva normal, o valor do coeficiente de curtose é $k = 3$. Para distribuições leptocúrticas (pontuda, caudas finas), $k > 3$; para distribuições platicúrticas (achatadas), $k < 3$. Entre as distribuições na Fig. **4**, a que parece mais se aproximar da curva normal é a C. Os valores da curtose calculados pela fórmula na eq. (**6**) são:

$$k_A = 1{,}70 \qquad k_B = 2{,}17 \qquad k_C = 2{,}93 \qquad k_D = 3{,}30$$

o que confirma o que foi observado no gráfico: a distribuição C é a mais próxima da normal, a D é a mais leptocúrtica, e a A é a mais platicúrtica.