2.2.1. Medidas de posição

Para comparar duas distribuições, não faz sentido comparar apenas os valores máximos de cada uma, ou os valores mínimos; é necessário encontrar um valor "típico", ou "representativo"; isto é, um valor que sozinho sirva para representar todos os dados. Este valor deve estar localizado no centro da distribuição; as *medidas de posição* servem para indicar onde se encontra este centro (são por isto também chamadas de medidas de *localização* ou de *tendência central*).

Veja por exemplo a distribuição de alturas de mulheres adultas representada pelo diagrama de pontos na Fig. 2. Que altura podemos considerar como "típica" para estas mulheres? Parece intuitivo que o valor típico é aquele localizado mais ou menos no centro da distribuição; a mulher típica seria aquela que não é nem muito alta, nem muito baixa. No caso, uma altura de cerca de 160 cm.

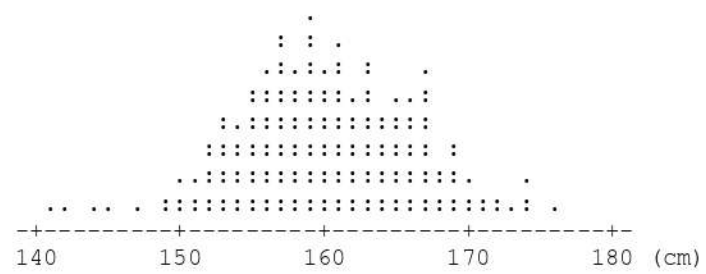

Figura 2 – Altura em uma amostra de mulheres adultas

Por que podemos considerar esta altura de 160 cm como "típica" da distribuição? Há várias maneiras de justificar a escolha deste número. Primeiro, porque é o valor em torno do qual estão concentradas mais observações. Há mais observações em torno de 160 cm do que, digamos, em torno de 140 cm; há mais mulheres com cerca de 160 cm do que mulheres com cerca de 140 cm. Segundo, porque aproximadamente metade das observações tem valores acima de 160, a outra metade tem valores abaixo; a altura de 160 portanto é exatamente a altura "nem alta, nem baixa". Terceiro, porque as observações estão distribuídas mais ou menos simetricamente em torno deste centro. Para cada mulher que está, digamos, 10 cm acima deste centro, corresponde outra que está 10 cm abaixo. O valor 160 é portanto uma espécie de "centro de equilíbrio"; se o gráfico representasse uma balança,

onde os pontos fossem pesos e o eixo horizontal fosse uma barra rígida, esta balança ficaria equilibrada se seu ponto de apoio fosse colocado no ponto com valor 160.

Estes argumentos levaram à criação das três medidas mais usadas para localizar o valor "típico" ou "central" de uma distribuição:

- Moda (*mode*): é o valor mais freqüente de uma distribuição. Num gráfico de pontos, a moda indica o ponto onde ocorre o "pico" da curva, isto é, o ponto mais alto. De acordo com o número de modas, uma distribuição pode ser classificada como *unimodal* (o gráfico tem a forma de montanha ou de um sino, com um único pico, como na Fig. **2**), *bimodal* ou *multimodal* (dois ou mais picos). Representaremos a moda de uma variável X pelo símbolo $\hat{X}$ (letra X com um acento circunflexo ou chapéu).
- Mediana (*median*): é o valor que divide a distribuição em duas metades; metade dos dados têm valor maior ou igual à mediana, a outra metade têm valor menor ou igual. Num gráfico de pontos, metade dos pontos estaria na mediana ou acima dela, enquanto que a outra metade estaria na mediana ou abaixo. No histograma, a mediana divide o gráfico em duas partes de mesma área. Representaremos a mediana de uma variável X pelo símbolo $\tilde{X}$ (letra X com um til). A mediana é uma das *separatrizes*, valores que dividem a distribuição em partes contendo a mesma quantidade de dados; outras separatrizes são os *quartis* (usados no *diagrama de Tukey*), os *decis* e os *percentis* (vistos a seguir, na seção **2.2.1.6**).
- Média aritmética (*mean*): localiza o "centro de gravidade" de uma distribuição. É fácil de calcular e é extremamente importante na Inferência Estatística. Num gráfico, é possível determinar de modo intuitivo onde se encontra aproximadamente a média de uma distribuição: basta supor que o desenho seja um sólido e procurar o ponto onde ele se equilibraria; este ponto (o fulcro de uma balança) será a média aritmética. Representaremos a média de uma variável X pelo símbolo $\overline{X}$ (letra X com uma barra em cima). Este símbolo para a média é de uso universal, ao contrário dos símbolos para mediana e moda, que variam de autor para autor. Este símbolo é normalmente lido como "X barra".

2.2.1.1. Cálculo das medidas

(i) Moda

Se uma variável pode assumir apenas alguns poucos valores discretos diferentes (como no exemplo da Fig. **3**, onde a variável pode assumir apenas os valores inteiros de 0 a 5), a moda é simplesmente o valor que foi encontrado com maior frequência. Quando a variável é contínua, porém, o idéa de *moda* não faz muito sentido. Considere, por exemplo, uma amostra de pesos de crianças ao nascer, em gramas, dos quais reproduzimos abaixo os vinte primeiros valores:

```
2900 3900 3230 3500 3650 3430 3120 3600 4240 2870
2460 2670 3210 3120 2810 3470 3700 4170 2540 3620
```

Como a variável pode assumir centenas de valores diferentes, é pouco provável que um valor apareça mais de uma vez (ou seja, é pouco provável que duas crianças nasçam com exatamente o mesmo peso). Nesta sub-amostra de 20 valores, apenas o valor 3120 foi repetido. Nestes casos, o que é feito geralmente é *agrupar* os dados, isto é, reuni-los em faixas

ou *classes* de peso (0 a 500g, 500 a 1000g, 1000 a 1500g, etc.), como na seção **2.1.4**. Poderemos então, ao invés de procurar o *valor* que teve maior frequência, procurar a *classe* que teve maior frequência; esta classe é chamada de *classe modal*. No entanto, avaliar a posição de uma distribuição com base na sua classe modal não é, em geral, uma boa idéia; o valor desta classe depende da forma como os dados foram agrupados - se fizermos um agrupamento diferente, a classe modal talvez seja diferente.

(ii) Mediana

Para encontrar a mediana de uma distribuição, basta listar os dados em ordem crescente. Se o número de dados é *ímpar*, a mediana é o valor que está localizado no meio da sequência de valores. Se há n valores, a mediana será o que está na posição $(n+1)/2$. Por exemplo, se temos uma amostra A de 17 pacientes, cujas idades são (em ordem crescente):

(A) 13 17 21 22 24 25 26 29 **32** 32 34 37 40 40 46 52 73

A mediana será a idade encontrada na posição número (17+1)/2 = 9 na sequência; isto é, a idade de 32 anos.Há oito idades abaixo dela, e oito acima.

Se o número de dados for *par*, a mediana será a média entre os dados que estão na posição $(n/2)$ e $(n/2)+1$. Por exemplo na amostra B, de 16 pacientes, cujas idades são:

(B) 20 26 32 32 35 36 41 **42 43** 45 48 48 52 57 59 61

A mediana será a média entre os valores que ocupam a 8ª e a 9ª posições na sequência (42 e 43, respectivamente); será, portanto, a idade de 42,5 anos (metade dos pacientes tem menos de 42,5 anos, a outra metade tem mais).

(iii) Média aritmética simples

A média é sem dúvida a medida estatística mais bem conhecida. Dada uma amostra, a média é calculada pela soma dos valores observados, dividida pelo número de valores. Para a amostra A do item anterior, por exemplo a média é dada por:

$$\overline{X} = \frac{13+17+21+...+46+52+73}{17} = 33{,}1$$

O somatório de um conjunto de números é usado com frequência em Estatística, e é representado por Σ (a letra grega *sigma*, maiúscula). No caso geral, para uma amostra com valores $x_1, x_2, \ldots, x_n$, a média aritmética simples é dada por :

$$\overline{X} = \frac{x_1 + x_2 + x_3 + ... + x_{n-1} + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note que a moda é igual a um dos valores da amostra; a média e a mediana, por outro lado, não precisam ser iguais a valores existentes ou possiveis na amostra. Na amostra B, a média foi de 42,3 anos e a mediana 42,5 anos, embora as idades originais sejam dada em termos de números inteiros (ninguém diz que tem "42,3" anos!).

2.2.1.2. Comparação entre as três medidas

Existem várias medidas diferentes para a *posição*, porque nenhuma medida é perfeita; se alguma delas fosse perfeita, as outras seriam desnecessárias. Toda medida têm limitações, funciona bem em algumas situações, mas não em outras. Veremos as vantagens e desvantagens de cada uma destas medidas, usando exemplos simples com dados simulados.

Suponha, por exemplo, que você esteja fazendo uma pesquisa, e coletou dados sobre o número de crianças em idade escolar nas famílias de uma bairro (ou o número de peças com defeito em um lote, o número de filhotes nos ninhos de uma espécie de ave, ou qualquer outra variável relacionada com a área de estudo que lhe interesse). Depois de estudar uma amostra de nove famílias, você encontrou a distribuição mostrada no gráfico da Fig. **3**.

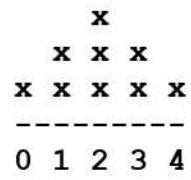

Figura 3. Exemplo de distribuição simétrica

Esta é uma distribuição inteiramente simétrica, e é fácil dizer que a família "típica" é a que tem duas crianças, já que este é o valor central da distribuição; *central* porque localiza o eixo de simetria da figura, porque aparece com maior freqüência (= *moda*), porque metade dos pontos estão distribuídos a cada lado dele (= *mediana*), porque é o centro de equilíbrio da distribuição (= *média*). Portanto, para distribuições simétricas como esta, as três medidas têm o mesmo valor. (Na distribuição da Fig. **2**, que é aproximadamente simétrica, temos *moda* = 159 cm, *média* = 160,2 cm, *mediana* = 160,0 cm).

Se no entanto você estudou 12 famílias e obteve como resultado a distribuição da Fig. **4A.** Não existe nela um valor que possamos considerar como um "centro" evidente, ou um "valor típico" óbvio; a figura não tem um eixo de simetria. Dizemos que distribuições deste tipo, em que a maioria dos valores estão acumulados no lado esquerdo, abaixo da média, e há uma cauda longa se prolongando para a direita, têm *assimetria positiva*.

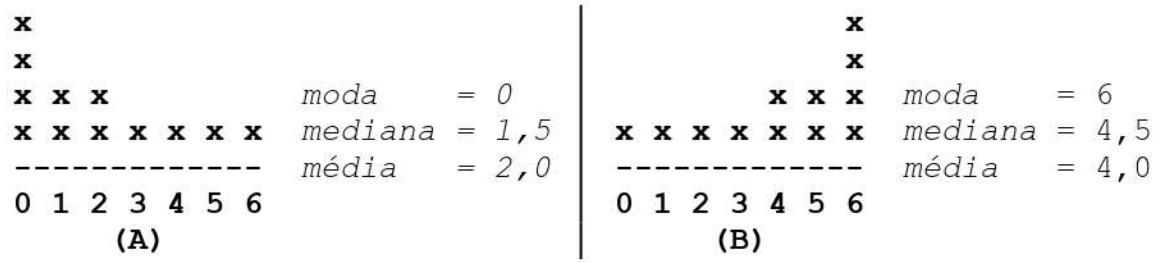

Figura 4 – exemplo de distribuições assimétricas

Se usarmos os três critérios de centro acima (*moda*, *mediana*, *média*), obteremos três valores diferentes para o centro, geralmente observando a relação:

moda < *mediana* < *média*.

Se assimetria for *negativa* (a maioria dos dados estão acumulados no lado direito do gráfico, acima da média, e a cauda se prolonga para a esquerda), observaremos a relação inversa:

$$moda > mediana > média$$

como no exemplo da Fig. **4B** (que é o mesmo gráfico da Fig. **4A**, refletido em torno de um eixo vertical).

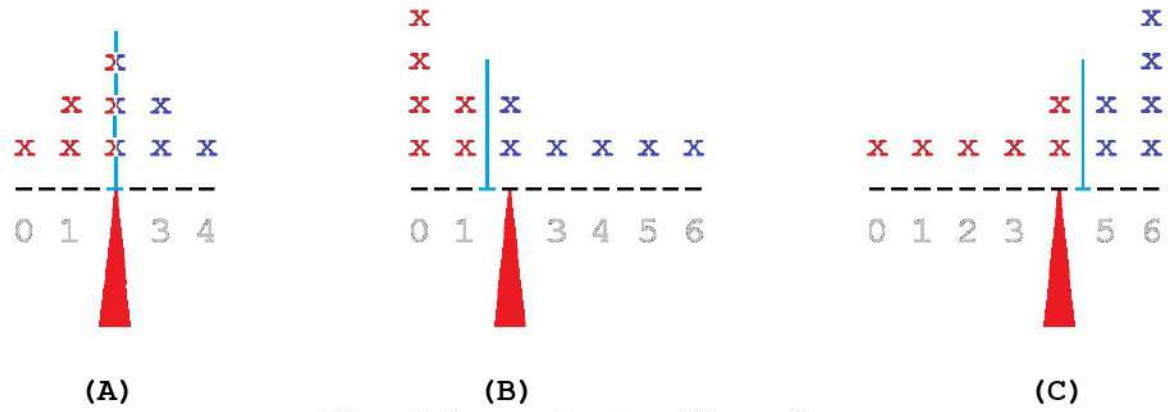

Figura 5. Comparação entre média e mediana

A Fig. **5** compara estes três gráficos. A localização das medianas é indicada por um traço vertical (azul) que divide os pontos: metade dos dados têm valor igual ou superior à mediana, a outra metade têm valor igual ou inferior. A localização das médias está indicada pelo triângulo vermelho (lembre-se de que a média indica o centróide da distribuição; isto é, o ponto onde ela se equilibraria como uma balança, se os pontos tivessem peso). Nas distribuições (B) e (C), a *moda* é a menos útil das medidas; certamente não faz sentido dizer que o centro da distribuição (B) é o valor "0", ou que o centro da (C) é o "6". A moda, portanto, só tem algum sentido quando a distribuição é simétrica e unimodal, como na distribuição (A).

2.2.1.3. Que medida usar?

Quando os dados são numéricos, a menos útil destas medidas é a moda. Entre média e mediana, a média é geralmente a mais usada, por ter propriedades estatísticas mais interessantes (isto será explicado mais tarde, seção **4.5**), e por ser mais *sensível*; basta uma modificação em apenas um dos dados de uma distribuição para que média se altere, enquanto a mediana em geral permanece a mesma.

Um exemplo disto: compare os gráficos da Fig. **6A** com o da Fig. **6B**. A única diferença entre eles está no valor extremo à direita, que foi aumentado de uma unidade. Ambas distribuições têm a mesma mediana (= 2), mas a distribuição A tem média um pouco maior (= 2,11) do que a B (= 2,00).

Portanto, se considerássemos apenas a mediana, iríamos concluir que estas duas distribuições são iguais; se considerássemos a média, veríamos que a distribuição (B) deve ter alguns valores maiores do que a (A). A média, por levar em conta os valores de todos os pontos da distribuição, é portanto mais *sensível* do que a mediana; qualquer modificação no valor dos pontos altera a média, mas não necessariamente a mediana.

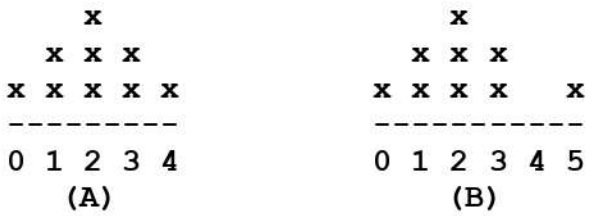

Figura 6. Duas distribuições simuladas de mesma mediana

A maior sensibilidade da média é em geral uma vantagem; às vezes, contudo, pode se tornar uma desvantagem. Se a distribuição for muito assimétrica, a média será indevidamente influenciada pelos valores extremos (discrepantes ou não), e deixará de mostrar o que é "típico" numa distribuição. Os gráficos da Fig. **7** ilustram isto, exagerando a situação encontrada na Fig. **6**.

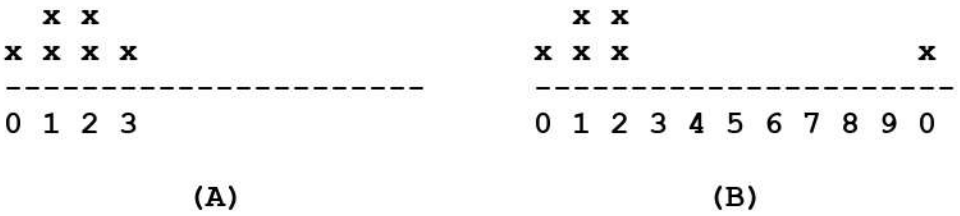

Figura 7. Duas distribuições simuladas de mesma mediana

Ambas as distribuições têm a mesma mediana (*mediana* = 1,5), mas as médias serão diferentes. A Fig. **8** mostra a posição das médias.

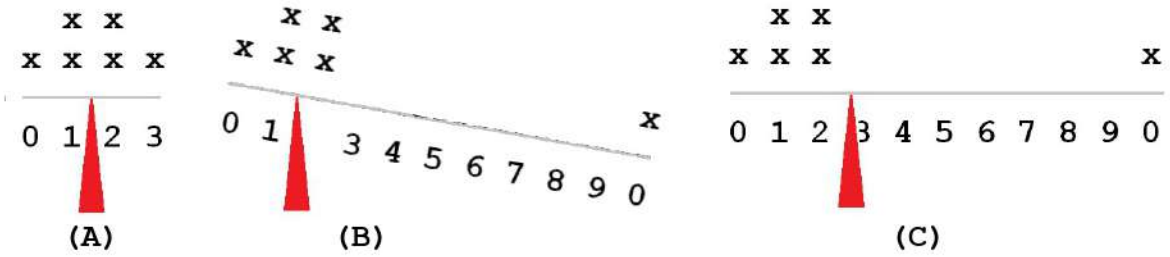

Figura 8. Posição das médias em duas distribuições

Na Fig. **8A**, é fácil ver que a média, o ponto onde a distribuição se equilibra como numa balança, é $\overline{X}$=1,5. Na Fig. **8B**, porém, a balança se desequilibra, por causa do valor X=10. A Fig. **8C** mostra o novo ponto de equilíbrio, igual a $\overline{X}$=2,67. Um único ponto, portanto (o valor X=10), causou uma grande alteração na média.

Suponha que estes números também se refiram ao tamanho de famílias, como no exemplo anterior. Se lembrarmos que o objetivo das medidas de tendência central é indicar o valor que pode ser considerado "típico" de uma distribuição, fica evidente que a média não serve para isso no caso da amostra na Fig. **8C**. Não tem sentido dizer que a família típica nesta amostra é aquela que tem 2,67 crianças; na verdade, todas as famílias, com exceção de uma, tem 2 ou menos crianças. A mediana estará mais próxima do que pode ser considerado típico: o número de crianças da maioria das famílias da amostra na Fig. **8C** é

realmente algo em torno de 1,5. Neste caso, portanto, a melhor descrição da distribuição é a fornecida pela mediana.

Em Estatística, uma medida que não é afetada pelos valores discrepantes é chamada de *robusta*. A mediana é mais *robusta* que a média aritmética. Geralmente, porém, as medidas mais *robustas* são também as menos *sensíveis*. A decisão entre usar uma medida mais robusta ou uma mais sensível depende dos objetivos da análise ou das características dos dados. Em distribuições muito assimétricas, ou que têm pontos discrepantes, a mediana provavelmente será uma alternativa mais útil que a média. Note que, nestas distribuições assimétricas, a mediana sempre cai em algum lugar entre a média e a moda; por isso faz mais sentido considerar a mediana, ao invés da média, como indicadora do "centro".

Outra alternativa é a média "podada" (*trimmed mean*): descartamos os valores extremos de cada cauda da distribuição, e tiramos a média dos restantes. No entanto, procedimentos que envolvam descartar parte dos dados em geral criam mais problemas do que resolvem. Não é fácil decidir quantos dados devemos descartar de cada cauda da distribuição. Além disso, nunca saberemos se aqueles dados descartados não tinham alguma informação útil que deveríamos ter levado em conta (talvez a resposta do problema estivesse justamente naqueles dados que decidimos jogar fora...).

Distribuições de dados reais com grande assimetria positiva e muitos pontos discrepantes são encontrados com freqüência em diversas áreas. Casos extremos podem ser encontrados, por exemplo, nos estudos de sobrevida, isto é, de quanto tempo pacientes sobrevivem depois de um determinado tratamento, ou de quanto tempo uma máquina funciona antes de apresentar defeitos. O gráfico da Fig. **9** mostra o tempo de sobrevida numa amostra de 30 pacientes com câncer, depois que a primeira entrada do paciente no hospital. Qual é a sobrevida "típica" destes pacientes? A média e a mediana são bem diferentes (383 dias, e 311 dias, respectivamente); na verdade, nenhuma destas duas medidas dá uma boa descrição da distribuição. Em vez de usá-las, os pesquisadores preferem trabalhar com *curvas de sobrevivência* (ou *sobrevida*), modelos que avaliam a probabilidade de um paciente alcançar uma certa sobrevida (esta área de pesquisa é chamada de *Análise de sobrevivência*).

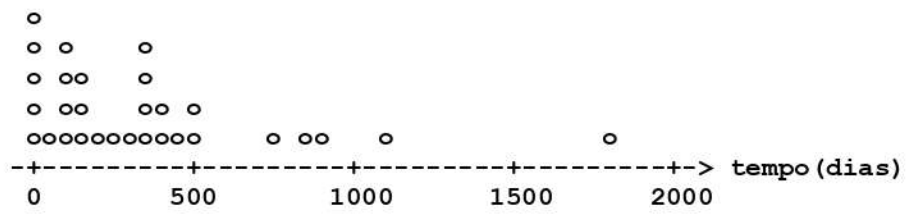

Figura 9. Tempo de sobrevida de pacientes com câncer

2.2.1.4. Outros tipos de médias

Além da média aritmética simples, há alguns outros tipos de média que às vezes são usados para resolver problemas específicos.

(i) *Média aritmética ponderada*

A média que vimos até agora é a média aritmética *simples*, na qual todos os valores tem o mesmo *peso*. Em alguns problemas, é preciso *ponderar* os valores, de forma que cada valor tenha um *peso* diferente. Isto é feito, por exemplo, quando os dados estão orga-

nizados numa tabela de distribuição de freqüência. Por exemplo, considere a Tabela **1**, que mostra o número de automóveis por família residente em um bairro.

Tabela 1 - Número de automóveis por família

carros	f	fr	fr%
0	216	0,486	48,6
1	174	0,392	39,2
2	39	0,088	8,8
3	13	0,029	2,9
4	1	0,002	0,2
5	1	0,002	0,2
totais	444	1	100,0

Se quisermos calcular o número médio de automóveis por família, não podemos simplesmente tirar a média simples dos valores de 0 a 5; o que temos que fazer é fazer uma média onde cada valor seja ponderado por sua freqüência (valores mais freqüentes terão mais peso), e dividir pelo total das frequências:

$$\overline{X} = \frac{0\times 216 + 1\times 174 + 2\times 39 + 3\times 13 + 1\times 4 + 1\times 5}{444} = 0{,}68$$

No caso geral, se temos os valores $x_1, x_2, \ldots x_n$, listados num tabela com suas frequências $f_1, f_2, \ldots f_n$, a média ponderada será dada por:

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

Podemos também ponderar pelas frequências relativas *fr*, ou pelas frequências percentuais *fr%*:

$$\overline{X} = \frac{0\times 48{,}6 + 1\times 39{,}2 + 2\times 8{,}8 + 3\times 2{,}9 + 1\times 0{,}2 + 1\times 0{,}2}{100} = 0{,}68$$

Médias ponderadas também podem ser usadas em problemas onde os pesos são atribuídos arbitrariamente, para indicar que alguns valores devem ser considerados mais importantes que outros e devem receber mais peso na média. Por exemplo, em cada disciplina na UFJF o aluno deve fazer pelo menos três avaliações ao longo do período, e o resultado final deverá ser dado pela média ponderada das notas obtidas. Suponha que uma professora dê três provas, e ache que a primeira e a segunda são mais importantes do que a terceira. Ela pode, por isso, atribuir pesos 40, 40, 20 às três provas, de forma que a média final seja dada por:

$$\overline{X} = \frac{(\text{nota 1})\times 40 + (\text{nota 2})\times 40 + (\text{nota 3})\times 20}{100}$$

(ii) *Média geométrica*

Usada em problemas relacionados às taxas de juros ou à inflação. Por exemplo, suponha um país onde a inflação seja medida trimestralmente. Em um ano, as taxas de

inflação trimestrais foram 15%, 10%, 8% e 12%. Qual foi a inflação trimestral média ao longo deste ano?

Neste país, um produto que custasse 1 real no início do ano passaria a custar

$(1 + 15\%) = 1 \times 1{,}15 = 1{,}15$ reais,

ao final do primeiro trimestre. Ao final do segundo trimestre, passaria a custar

$(1{,}15 + 10\%) = 1 \times (1{,}15 \times 1{,}10) = 1{,}265$.

Ao final do ano, o preço seria

$1 \times (1{,}15 \times 1{,}10 \times 1{,}08 \times 1{,}12) = 1{,}53$

A inflação total do ano seria de 53%, e a inflação trimestral média seria

$$\overline{X}_{geom} = \sqrt[4]{1{,}53} = 1{,}1122$$

Se esta inflação média tivesse ocorrido em todos os trimestres, o preço final teria sido igual ao que foi observado:

$1 \times (1{,}1122 \times 1{,}1122 \times 1{,}1122 \times 1{,}1122) = 1{,}53$

No caso geral, a média geométrica dos valores $x_1, x_2, \ldots x_n$ será dada por

$$\overline{X}_{geom} = \sqrt[n]{x_1 \times x_2 \times \ldots \times x_n}$$

(iii) *Média harmônica*

Usada, por exemplo, em problemas relacionados a velocidades. Suponha que você suba uma colina (5 km de extensão) de bicicleta a 5 km/h, depois desça a 60 km/h. Qual foi sua velocidade média neste percurso? A velocidade média **não** será a média aritmética das velocidades de subida e descida, (5+60)/2=32,5. Na subida a 5 km/h, você gasta uma hora para fazer o percurso; na descida a 60 km/h, gasta 5 minutos. O tempo total gasto será portanto de 65 min, e a velocidade média no percurso de 10 km será:

$$\overline{X}_{harm} = 10 \text{ km} / 65 \text{ min} = 0{,}1538 \text{ km/min} = 9{,}2 \text{ km/h}$$

A média harmônica é dada pelo *recíproco da média dos recíprocos*; no exemplo:

$$\overline{X}_{harm} = \frac{2}{\frac{1}{5} + \frac{1}{60}} \cong 9{,}2$$

No caso geral, a média harmônica dos valores $x_1, x_2, \ldots x_n$ é dada por:

$$\overline{X}_{harm} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

2.2.1.5. Medidas de posição para dados agrupados

Suponha que tenhamos apenas os dados agrupados; por exemplo, temos a pirâmide etária brasileira, e queremos calcular a partir dela a média de idade da população. Se não temos acesso aos dados brutos (a idade de cada pessoa da população), não podemos calcu-

lar os valores exatos das medidas descritivas; é possível, porém, calcularmos aproximações razoáveis destas medidas, a partir da tabela de distribuição de frequências (seção **2.1.4**). Mostraremos abaixo como isto é feito, tomando como exemplo a Tabela **2**, que mostra os pesos de 189 recém-nascidos, agrupados em 5 classes.

Tabela 2 – Pesos de recém-nascidos

pesos (g)	PM	f	fr%	F	F%
0000 ├─ 1000	500	1	0,53	1	0,53
1000 ├─ 2000	1500	18	9,52	19	10,05
2000 ├─ 3000	2500	78	41,27	97	51,32
3000 ├─ 4000	3500	83	43,92	180	95,24
4000 ├─ 5000	4500	9	4,76	189	100,00
	-	189	100	-	-

(i) Moda

Numa tabela, não tem sentido falar em moda, mas sim em *classe modal*, a classe de maior freqüência. Na Tab. **2**, a classe modal é a 3000-4000 g; se quisermos representá-la por um número único, podemos tomar o ponto médio desta classe, e dizer que a moda é igual a 3500 g. A moda, contudo, não é muito útil nestes casos, porque é imprecisa, e poderá variar dependendo da forma como os dados foram agrupados (isto é, dos limites de classes escolhidos para a tabela).

(ii) Mediana

Na tabela, como há 189 valores, a mediana será o que ocupa a posição de número
$(n+1)/2 = (189+1)/2 = 95$

Precisamos, portanto, de localizar o 95ª observação. Conferindo a coluna das frequências acumuladas F, vemos que até o limite superior da segunda classe (2000 g) estão 19 observações; até o limite da superior da terceira classe (3000 g) estão 97 observações. A 95ª observação, portanto, estará entre estes limites: entre 2000 e 3000 g (isto é, na terceira classe), provavelmente bem próxima de 3000 g.

Até o limite superior da segunda classe, há 19 observações; portanto, faltam 95-19 = 76 observação até chegarmos ao 95° valor. Na terceira classe estão 78 observações; queremos achar o ponto que separa

$$\frac{76}{78} = 0{,}974 = 97{,}4\%$$

destas observações. Se supusermos que elas estão uniformemente distribuídas ao longo do classe, procuraremos o ponto que separa igualmente 97,4% do intervalo de classe. Como a classe tem um intervalo de 1000 g, este valor estará 974 g acima do limite inferior da classe; portanto, será de 2000+974 = 2974 g.

Este processo de estimativa é conhecido como *interpolação linear*. No exemplo, a aproximação conseguida foi muito boa. A mediana exata, calculada a partir dos 189 dados original, é de 2977 g; a estimativa 2974 g, portanto, teve um erro de apenas 3 g, ou

$$\frac{2974 - 2977}{2977} = 0{,}0010 = 0{,}10\%$$

Em resumo, os passos para o cálculo da mediana de dados agrupados são:

a) Calcule a coluna de frequências acumuladas F para a tabela;

b) Determine em que posição p, na seqüência ordenada dos valores, estará a mediana. Se a distribuição tem n dados, faça

$p=(n+1)/2$ se n é ímpar

$p=n/2$ se n é par

c) Localize a classe onde estará a mediana, comparando p com os valores da coluna de freqüências acumuladas F; chamaremos esta *classe mediana*;

d) Chamando:

f_m : freqüência absoluta da classe mediana;

F_m : freqüência acumulada da classe mediana;

F_{m-1} : freqüência acumulada da classe anterior à classe mediana;

$L_{sup,m}$: limite superior da classe mediana;

$L_{sup,m}$: limite superior da classe mediana;

e) A aproximação para a mediana será então dada por:

$$\tilde{X} \approx L_{\inf,m} + \left(\frac{p - F_{m-1}}{f_m} \right) \times (L_{\sup,m} - L_{\inf,m}) \qquad (2)$$

No exemplo acima, os valores usados foram:

$$\tilde{X} \approx 2000 + \left(\frac{95 - 19}{78} \right) \times (3000 - 2000) = 2974$$

(iii) Média aritmética

Podemos conseguir uma estimativa da média de um conjunto de dados agrupados se fizermos uma simplificação: considerarmos que, dentro de cada classe, todos os valores são idênticos ao ponto médio daquela classe. Assim, a Tab. 2 ficaria como a Tab. 3.

Basta então calcular a média ponderada destes pontos médios (eq. 1):

$$\overline{X} = \frac{500 \times 1 + 1500 \times 18 + 2500 \times 78 + 3500 \times 83 + 4500 \times 9}{189} = 2929\,\text{g}$$

O valor real da média, calculado a partir dos 189 valores observados, é de 2944 g; a aproximação obtida tem portanto um erro de apenas cerca de 0,5 %.

Tabela 3 – Pesos de recém-nascidos

pesos	f	fr%
500	1	0,53
1500	18	9,52
2500	78	41,27
3500	83	43,92
4500	9	4,76
-	189	100

No caso geral, se há n classes, e cada classe i-ésima tem ponto médio PM_i e freqüência f_i, a média será dada por:

$$\overline{X} = \frac{\sum_{i=1}^{n} PM_i f_i}{\sum_{i=1}^{n} f_i} \qquad (3)$$

onde: f_i frequência da classe i-ésima
PM_i ponto médio da classe i-ésima

Podemos fazer também o cálculo usando as frequências relativas (*fr*) ou as frequências relativas percentuais (*fr%*), ao invés das frequências absolutas (*f*). Se usarmos *fr*, o denominador da razão na eq. (3) será desnecessário, portanto a média será dada simplesmente ela eq. (4):

$$\overline{X} = \sum_{i=1}^{n} PM_i fr_i \qquad (4)$$

É claro que, na realidade, os pontos não têm todos valores iguais ao PM da classe. Se a distribuição é unimodal e razoavelmente simétrica, isto não deve causar problemas para o cálculo aproximado da média, já a quantidade de pontos acima e abaixo dos PMs acabam se equilibrando.

Um problema que surge com freqüência quando trabalhamos com tabelas é o de encontrarmos classes "abertas", isto é, classes que não tem o limite inferior, ou o superior. Por exemplo, na tabela com pesos de crianças nascidas em Minas Gerais em 2018 (Tabela 5, seção 2.1.4.3), a última classe de peso era "4000 g ou mais". Para podermos calcular o ponto médio desta classe, e em seguida a média, uma solução é "fechar" esta classe, atribuindo-lhe um limite superior, por exemplo de 5000 g; é claro que pode haver crianças nascidas com peso acima deste limite, mas a proporção de tais crianças, em relação ao total de nascimentos, é tão pequena que provavelmente não irá afetar muito o resultado. Uma segunda solução é a de simplesmente abandonar a média aritmética, e calcular em vez dela a mediana, que não será afetada por estes valores extremos.

2.2.1.6. Separatrizes

Separatrizes, também chamadas de *quantis* (*quantiles*), são valores que dividem a distribuição em partes que contém a mesma quantidade de dados. Já vimos duas delas (a mediana e os quartis, usadas para construir o diagrama de Tukey na seção 2.1.3). As separatrizes usadas com mais freqüência são:

Mediana ($\tilde{X}$)– divide a distribuição em duas partes, cada uma com 50% dos dados
Quartis (*Q*) – dividem a distribuição em quatro partes, cada uma com ¼ dos dados
Decis (*D*) - dividem a distribuição em 10 partes, cada uma com 10% dos dados
Percentis (*P*) - dividem a distribuição em 100 partes, cada uma com 1% dos dados

Note que $\tilde{X} = Q_2 = D_5 = P_{50}$; ou seja, a mediana, que divide a distribuição em duas metades, é igual ao 2°. quartil (que divide a distribuição em dois quartos), ao 5°. decil (que divide em cinco décimos), e ao 50°. percentil (que divide em 50 centésimos). Da mesma forma,

$$Q_1 = P_{25}$$
$$Q_3 = P_{75}$$

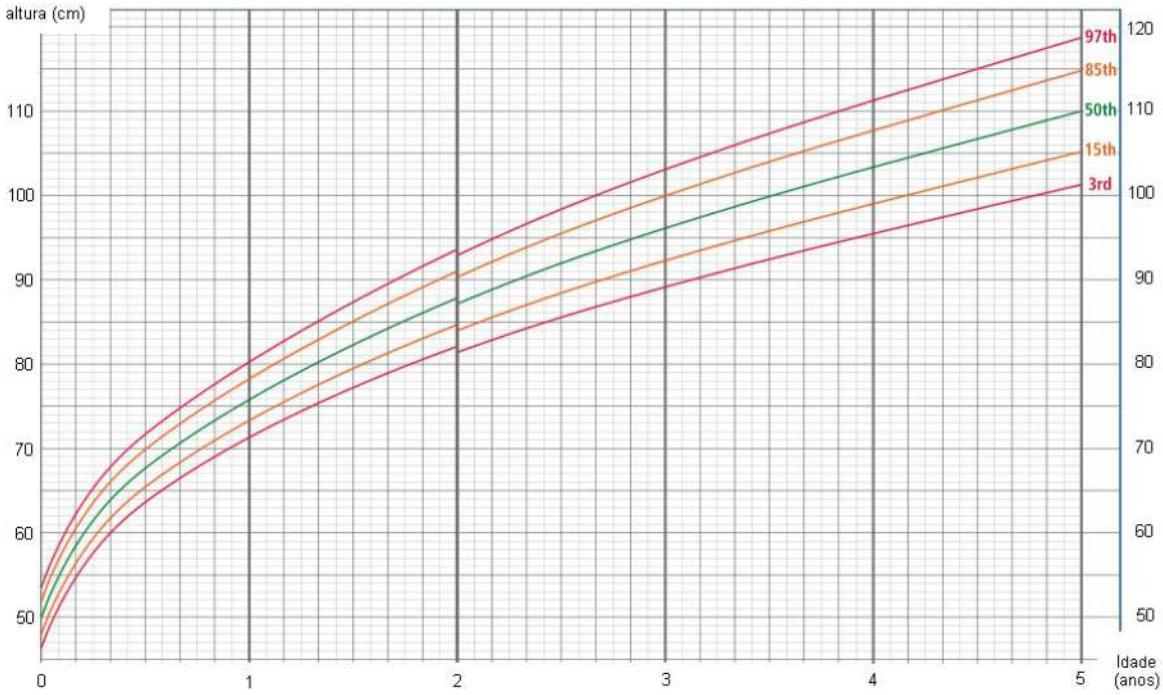

Figura 9. Curvas de crescimento para meninos até 5 anos

Separatrizes são úteis como pontos de referência, para que possamos determinar onde um determinado valor se localiza dentro de uma distribuição. Por exemplo, a Fig. 9 mostra as curvas de crescimento para meninos até 5 anos de idade, definidas pela Organização Mundial da Saúde. Para meninos de 5 anos de idade, a mediana da altura é igual a 110 cm. Nesta idade, 70% dos meninos devem ter alturas entre 105 e 115 cm, que correspondem aos percentis P_{15} e P_{85}; 94% devem ter alturas entre 101 e 119 cm, que correspondem aos percentis P_{03} e P_{97}. Estas curvas são usadas por pediatras para avaliar se um menino está crescendo normalmente; se um menino de 5 anos tiver menos de 101 cm de altura (P_{03}), por exemplo, ele terá crescido bem menos do que a grande maioria dos meninos desta idade, e a razão disto isto deverá ser investigada.