## 2.1.4. Tabelas de distribuição de freqüências



---

### 2.1.4.1. Introdução

"Tabelas de distribuição de freqüências", como o nome indica, são aquelas que mostram como se distribuem as *freqüências* de cada valor de uma variável, num conjunto de dados, e indicam quais valores são mais comuns e quais são mais raros. Nem todas as tabelas são assim; a Tab. **1**, por exemplo, que lista os nove países que compõem a Comunidade dos Países de Língua Portuguesa e alguns dados estatísticos sobre eles, não mostra uma distribuição de freqüências. Esta tabela inclui os três tipos mais comuns de variáveis: uma variável nominal (o nome do país), uma ordinal (o IDH), e duas quantitativas (população e área). Tabelas destes tipo são muito úteis, mas não apresentam nenhuma dificuldade especial, e não serão estudadas neste capítulo.

**Tabela 1. Países da Comunidade dos Países de Língua Portuguesa**

| País | População (milhões de hab.) | Área (x 1000 km$^2$) | Índice de Desenvolvimento Humano |
|---|---|---|---|
| Brasil | 210,1 | 8.515,8 | alto |
| Moçambique | 30,0 | 801,6 | médio |
| Angola | 31,1 | 1246,7 | médio |
| Portugal | 10,3 | 92,2 | muito alto |
| Guiné Bissau | 1,6 | 36,1 | baixo |
| Guiné Equatorial | 1,3 | 28,0 | médio |
| Cabo Verde | 0,5 | 4,0 | médio |
| São Tomé e Príncipe | 0,2 | 1,0 | médio |
| Timor Leste | 1,4 | 15,0 | médio |

(fonte: Wikipedia, 2020)

### 2.1.4.2. Tabelas de freqüências com dados não-agrupados

Quando a variável é *qualitativa*, ou é *quantitativa discreta* mas assume poucos valores diferentes, a tabela de distribuição de freqüências pode reportar diretamente a freqüência de cada valor. A Tab. **2**, por exemplo, mostra dados sobre as mortes no trânsito do Brasil, em 1996 e 2011, de acordo com a categoria do meio de transporte. A variável é a *categoria*, uma variável qualitativa; as freqüências reportadas são as *freqüências relativas percentuais*; isto é, a porcentagem do total de mortes que ocorreu em cada categoria. (Estes dados serão usados como exemplos de gráficos na seção **2.1.6**).

Um exemplo usando uma variável quantitativa é a Tab. **3**, que dá as idades dos estudantes (de ambos os sexos) de Estatística, numa turma do curso de Medicina da UFJF.

A variável é a *idade*, que assumiu nesta amostra apenas valores entre 17 e 26 anos. A *freqüência absoluta* (*f*) indica o número de vezes em que cada valor foi encontrado na população ou amostra.

**Tabela 2. Mortes no trânsito no Brasil, por categoria de meio de transporte**

| Categoria | *fr%* | |
|---|---|---|
| | **1996** | **2011** |
| Pedestre | 69,8 | 27,3 |
| Automóvel | 20,4 | 28,7 |
| Motocicleta | 4,0 | 33,9 |
| Caminhão | 2,2 | 3,6 |
| Bicicleta | 1,8 | 4,4 |
| Ônibus | 0,4 | 0,6 |
| Outros | 1,4 | 1,5 |
| Total | 100,0 | 100,0 |

fonte: Waiselfisz (2013)

Na turma, havia 15 estudantes com 20 anos de idade; a freqüência absoluta deste valor é portanto $f = 15$. A *freqüência relativa* (*fr*) é a razão entre a freqüência absoluta de um valor e o total da coluna: no exemplo, a freqüência relativa da idade de 20 anos é $fr = 15 / 66 = 0{,}227$. Se multiplicarmos este valor por 100, obtemos a *freqüência relativa percentual*, ou simplesmente *freqüência percentual* (*fr%*); no exemplo, $fr\% = 27{,}7\ \%$. Note que o somatório das freqüências relativas deve ser igual a 1, e o das freqüências relativas percentuais deve ser igual a 100.

A coluna da *freqüência acumulada F* dá a soma das freqüências absolutas de cada linha com as das linhas anteriores; indica portanto o número de observações que tiveram valores iguais ou inferiores ao daquela linha. Por exemplo, 49 destes alunos desta amostra tem 20 anos de idade ou menos. A coluna da *freqüência percentual acumulada F%* indica, por sua vez, a soma das freqüências percentuais de cada linha com as das linhas anteriores; nesta amostra, 74,2% dos estudantes tem 20 anos de idade ou menos.

Todas estas freqüências podem ser deduzidas umas das outras; a partir da *freqüência absoluta* pode-se calcular por exemplo a *freqüência percentual*, ou vice-versa. Não é necessário que uma tabela reporte todas estas freqüências, uma vez que elas não são mais que maneiras alternativas de reportar uma mesma informação.

**Tabela 3 – Idades dos alunos de Estatística, Curso de Medicina, UFJF.**

| idades | *f* | *fr* | *fr%* | *F* | *F%* |
|---|---|---|---|---|---|
| 17 | 1 | 0,015 | 1,5 | 1 | 1,5 |
| 18 | 15 | 0,227 | 22,7 | 16 | 24,2 |
| 19 | 18 | 0,273 | 27,3 | 34 | 51,5 |
| 20 | 15 | 0,227 | 22,7 | 49 | 74,2 |
| 21 | 10 | 0,152 | 15,2 | 59 | 89,4 |
| 22 | 4 | 0,061 | 6,1 | 63 | 95,4 |
| 23 | 1 | 0,015 | 1,5 | 64 | 97,0 |
| 24 | - | - | - | 64 | 97,0 |
| 25 | 1 | 0,015 | 1,5 | 65 | 98,5 |
| 26 | 1 | 0,015 | 1,5 | 66 | 100 |
| soma | 66 | 1,000 | 100 | - | - |

fonte: o autor

### 2.1.4.3. Tabelas com dados agrupados

Se a variável for *quantitativa contínua* (podendo por isso assumir infinitos valores diferentes, pela própria definição de continuidade), ou for uma variável *discreta* que pode assumir uma grande quantidade de valores diferentes, será preciso *agrupar* os dados. A Tab. **4** reporta as idades das mulheres que tiveram filhos em 2018, no Brasil. Estas idades variaram entre 10 e 69 anos; para descrever sua distribuição, não podemos simplesmente fazer uma tabela atribuindo uma linha para cada idade, pois esta tabela teria mais de 50 linhas, e seria demasiado longa. O que podemos fazer é dividir as idades em *classes* com intervalos de, por exemplo, cinco anos, desta forma: 10 a 14 anos; 15 a 19 anos; 20 a 24 anos, etc., e registrar o número de mães com idades dentro de cada classe, isto é, a *freqüência* de cada classe.

**Tabela 4. Idades das mães de crianças nascidas vivas (Brasil, 2018)**

| Idade da mãe | freqüência absoluta | freqüência percentual (%) |
|---|---|---|
| 10 – 14 | 21,172 | 0,72 |
| 15 – 19 | 434,956 | 14,77 |
| 20 – 24 | 723,352 | 24,56 |
| 25 – 29 | 696,559 | 23,65 |
| 30 – 34 | 611,715 | 20,77 |
| 35 – 39 | 365,814 | 12,42 |
| 40 – 44 | 85,978 | 2,92 |
| 45 – 49 | 4,920 | 0,17 |
| 50 – 54 | 323 | 0,01 |
| 55 – 59 | 50 | 0,00 |
| 60 – 64 | 21 | 0,00 |
| 65 – 69 | 3 | 0,00 |
| ignorada | 69 | 0,00 |
| Total | 2.944.932 | 100,00 |

(fonte: datasus.gov.br)

"Agrupar" os dados significa portanto dividir o domínio da variável em subintervalos contíguos, denominados *classes*, e registrar na tabela o número de observações que pertencem a cada classe, isto é, a *freqüência* de cada classe. Como exemplo do agrupamento de variáveis contínuas, a Tab. **5** mostra a distribuição dos pesos ao nascer de 263.640 crianças nascidas em Minas Gerais durante o ano de 2018.

**Tabela 5 - Pesos de nascidos vivos em Minas Gerais, 2018**

| Peso ao nascer (g) | f | fr | fr(%) | F | F% |
|---|---|---|---|---|---|
| 0 \|-- 500 | 378 | 0,0014 | 0,14 | 378 | 0,14 |
| 500 \|-- 1000 | 1.494 | 0,0057 | 0,57 | 1.872 | 0,71 |
| 1000 \|-- 1500 | 2.130 | 0,0081 | 0,81 | 4.002 | 1,52 |
| 1500 \|-- 2500 | 20.727 | 0,0786 | 7,86 | 24.729 | 9,40 |
| 2500 \|-- 3000 | 66.022 | 0,2504 | 25,04 | 90.751 | 34,42 |
| 3000 \|-- 4000 | 163.050 | 0,6185 | 61,85 | 253.801 | 96,27 |
| 4000 ou mais | 9.831 | 0,0373 | 3,73 | 263.632 | 100,00 |
| Ignorado | 8 | 0,0000 | 0,00 | 263.640 | 100,00 |
| Totais | 263.640 | 1,0000 | 100,0 | --- | --- |

(fonte: datasus.gov.br)

Os dados estão agrupados em sete *classes*. Note que para definir cada classe foi usado o símbolo matemático ⊢ (intervalo fechado à esquerda e aberto à direita); voltaremos a ele mais abaixo. Cada classe é limitada por dois números, chamados de *limites* da classe; a distância entre eles é chamado de *intervalo de classe*. A classe 0 ⊢ 500 tem um intervalo de 500 g; a classe 3000 ⊢ 4000, um intervalo de 1000 g. A média entre os limites da classe é chamada de *ponto médio* da classe; para a classe 0 ⊢ 500, o ponto médio é 250; para a classe 3000 ⊢ 4000, o ponto médio é 3500 g

Às vezes, o agrupamento dos dados é imposto pela forma de coletar os dados. Por exemplo, quando um levantamento é feito por meio de questionários, e os respondentes devem indicar a que faixa etária pertencem, ou a que faixa de salários, etc.; neste caso, a forma de agrupar os dados foi determinada por quem elaborou o questionário, e decidiu quais faixas de idade, salário, etc., seriam consideradas.

Antes da introdução dos computadores pessoais, nos anos 1980s, o objetivo destas tabelas para dados contínuos era, principalmente, o de facilitar os cálculos. Se queremos conhecer o peso médio das crianças da Tab. **5**, por exemplo, é mais fácil obter uma estimativa aproximada através da tabela do que obter o valor exato, somando manualmente os pesos de 263.640 crianças. O trabalho do estatístico, por isso, começava quase sempre pela organização da tabela; a seguir, qualquer *medida estatística* de interesse, como a média ou o desvio-padrão (seções **2.2** e **2.3**) poderia ser obtido aproximadamente, de forma rápida, através da tabela, e não mais dos dados originais.

Atualmente, não é mais necessário usar tabelas de distribuição de freqüências como ferramenta para facilitar os cálculos. Depois que os dados foram inseridos num computador, a média ou qualquer outra medida pode ser obtida quase que instantaneamente com o apertar de uma tecla. As tabelas, contudo, continuam sendo importantes. Em primeiro lugar, porque são a melhor forma de publicar os dados; se tenho dados sobre os pesos de 263.640 crianças, ao invés de imprimir todos estes números, posso fazer uma tabela que resume a informação em menos de uma página (p. ex., a Tab. **5**). Em segundo lugar, porque uma tabela, justamente por trazer a informação resumida, pode destacar as características principais da distribuição de uma variável (localização e dispersão, assimetria, etc.), que não ficariam evidentes se estivéssemos trabalhando com os dados originais.

### 2.1.4.4. Como construir uma tabela

Para fazer o agrupamento dos dados, você precisa tomar algumas decisões:
(i) os intervalos de classes serão iguais entre si ou não?
(ii) quantas classes devem ser criadas? (ou seja, quantas linhas terá a tabela?)
(iii) qual será o limite inferior da primeira classe?
(iv) os intervalos de classe serão fechados à direita ou à esquerda?

Evidentemente, a maior parte dos programas de computador podem atualmente fazer as tabelas de forma automática, usando respostas *default* para estas decisões. O resultado, porém, nem sempre é o melhor possível; e iremos por isso discutir a seguir cada um destes itens.

*(i) Intervalos de classes iguais ou diferentes?*

Na maioria das vezes, as distribuições de freqüências são organizadas com os dados agrupados em classes de mesmo intervalo. Isto permite que o agrupamento seja feito auto-

maticamente por computadores, e facilita a interpretação da tabela; comparando as freqüências de cada classe, o leitor pode ter uma idéia da forma da distribuição, verificar se ela é unimodal ou não, simétrica ou assimétrica. Se as classes têm intervalos diferentes, esta interpretação da tabela é mais difícil.

Tabelas com classes de diferentes intervalos podem ser necessárias se a distribuição for muito assimétrica (exemplos disto serão vistos na seção **2.1.5.4**), ou se for intenção dos pesquisadores dar mais destaque a certos valores ou intervalos da variável do que a outros. Por exemplo, na Tab. **5**, algumas classes têm IC de 500 g, outras têm de 1000 g. Isto foi feito para destacar os valores de referência 1500 g e 2500 g (crianças com menos de 2500 g são ditas de *baixo peso ao nascer*; com menos de 1500 g, são de *muito baixo peso*; o baixo peso ao nascer significa um risco para a saúde da criança). Um outro exemplo: na Tab. **4**, poderíamos estar interessados em conhecer a proporção de mulheres que tiveram filhos antes dos 18 anos, ou dos 21 anos; neste caso, teríamos que reorganizar a tabela e modificar os limites das classes, de forma a deixar os valores 18 e 21 como limites superiores de duas classes, fazendo algo como,

10 ⊢ 15
15 ⊢ 18
18 ⊢ 21
21 ⊢ 25

A Tab. **5** mostra um problema bastante comum: sua última classe é "aberta". Classes *abertas* são aquelas que não tem definido o limite superior, ou o inferior. Apesar de estas classes serem encontradas com freqüência, é melhor evitá-las; já que não há limites definidos, não podemos completar o gráfico para estes dados (o *histograma*, ver seção **2.1.5**), e também não poderemos estimar a média ou a variância da distribuição a partir da tabela (seções **2.2** e **2.3**), pois não conheceremos o ponto médio das classes (qual é o ponto médio da classe "4000 g ou mais"?).

*(ii) Quantas classes devem ser criadas?*

Organizar uma distribuição de freqüências manualmente era uma tarefa tediosa, que tomava muito tempo. Por isto, antes da introdução dos microcomputadores era importante definir, logo ao início do trabalho, qual seria o número de classes (NC) adequado, já que não poderíamos simplesmente experimentar diferentes valores até encontrar o que desse o melhor resultado. Várias regras empíricas para a seleção do NC foram sugeridas, e podem ser encontradas em livros de Estatística publicados antes de 1980. Em geral, estas regras sugeriam um valor para NC em função do número $n$ de observações disponíveis; quanto mais observações, maior o NC. Exemplos destas regras são a que sugere que o NC não deve ser maior que cinco vezes o logaritmo de $n$, isto é:

$$NC \leq 5 \times log(n)$$

ou a mais conhecida delas, a chamada *regra de Sturges*:

$$NC \cong 1 + 3{,}3 \times log(n)$$

Para uma amostra com $n$ =100 observações, por exemplo, a primeira regra sugere

$$NC \leq 5 \times log(100) = 10$$

enquanto a regra de Sturges sugere:

$$NC \approx 1 + 3{,}3 \times log(n) = 1 + 3{,}3 \times log(100) \approx 8$$

Hoje em dia, já que praticamente todo trabalho estatístico é feito com computadores, estas regras perderam muito de sua importância, e devem ser encaradas apenas como um ponto de partida para a escolha do NC. Em geral, o NC costuma estar entre 5 e 20, mas isto depende não só da quantidade de dados, mas também dos objetivos da pesquisa. Se você quer dar apenas uma idéia geral da forma de uma distribuição, talvez uma tabela simples, com 5 classes, seja suficiente (serve para mostrar se a distribuição é unimodal, simétrica, etc.); se quer mostrar algo com mais detalhes, talvez precise de um NC maior. De qualquer forma, não há uma maneira simples de escolher o NC, que sirva para todos os problemas. É melhor experimentar; se você tem os dados, faça diversas tabelas, com diferentes NCs, compare os histogramas, e veja qual delas parece lhe parece mais interessante, e deixa mais evidente aquilo que você quer mostrar.

*(iii) Qual é o limite inferior da primeira classe?*

Decidido o número de classes ou o intervalo de classe, será preciso escolher o limite inferior da primeira classe, que será o ponto inicial para a marcação dos intervalos. Em geral, é preferível escolhê-lo de modo que os limites das classes sejam números "redondos", para facilitar a leitura da tabela. Por exemplo, na Tab. **5**, que representa a idade de mães, usamos ICs de 5 anos. Como a mais nova das mães tinha 13 anos de idade, poderíamos construir uma tabela usando este valor como limite inferior, e criar classes 13-17, 18-22, 23-27, etc. Também poderíamos iniciar a tabela a partir de valores menores que 13, e começar a partir de 12 anos (12-16, 17-21, 22-26, etc.) ou de 11 anos (11-15, 16-20, 21-25, etc.). O mais natural, contudo, seria começar com o valor 10, como foi feito na tabela, já que estamos geralmente habituados a raciocinar com faixas etárias de 5 ou de 10 anos. A mesma escolha de números "redondos" foi feita, por exemplo, na Tab. **5**.

*(iv) Os intervalos de classe serão fechados à direita ou à esquerda?*

Se compararmos a maneira como os limites de classe foram definidos nas Tabs. **4** e **5**, veremos que eles seguem padrões diferentes. A maneira usada na Tab. 4 é mais usada em publicações em língua inglesa e talvez seja mais simples de entender, para o leitor leigo. A maneira usada na Tab. **5** é encontrada em publicações brasileiras. Nela, os valores limites são repetidos nas classes contíguas; por exemplo, o valor 1000 g é o limite superior da classe na segunda linha, e o limite inferior da classe na terceira linha. Se houver uma criança que pese exatos 1000 g, será preciso decidir se ela deve ser contada na segunda ou na terceira linha da tabela. Neste exemplo, seria na terceira linha, o que é mostrado pelo símbolo matemático $\vdash$ (intervalo fechado à esquerda e aberto à direita) na definição das classes. O valor 1000 g não será incluído na classe definida como $500 \vdash 1000$, mas sim na classe seguinte, definida como $1000 \vdash 1500$.

Esta maneira de representar os intervalos pode parecer um pouco mais complicada que a anterior, mas ela tem suas justificativas. Em primeiro lugar, serve para enfatizar que as classes são contíguas. Não existe espaço entre uma classe e outra. no gráfico que representa a tabela (o histograma), as classes são representadas por retângulos justapostos, reforçando esta idéia de continuidade. Em segundo lugar, esta maneira torna mais fáceis quaisquer operações que devam ser feitas sobre os limites de classe. Por exemplo, o inter-

valo de uma classe pode ser obtido subtraindo-se o limite superior do limite inferior desta classe; se tentarmos fazer isto na Tab. **4**, estaremos obtendo resultados errados. Outro exemplo: se quisermos determinar o *ponto médio* de uma classe, simplesmente tomamos a média dos limites da classe; o sistema brasileiro 50 ⊢ 60 deixa claro que o ponto médio é 55, enquanto que o americano 50 — 59 pode levar a crer que é 54,5. Estes pontos médios serão úteis quando quisermos calcular alguma estatística da distribuição, como sua média ou desvio-padrão (ver seções **2.2** e **2.3**).

*(v) Que fazer com as classes cujas frequências são "ignoradas"?*

As Tabs. **4** e **5** mostram um problema que é muito encontrado: houve nascimentos para os quais a "idade da mãe" e o "peso ao nascer" não foram registrados. A freqüência destes nascimentos está representada na última linha das tabelas, com o rótulo "ignorado". Como fazer para calcular as freqüências relativas : incluir ou não estes nascimentos no total? Nas Tabs. **4** e **5**, as freqüências relativas foram calculadas usando o total de valores "válidos", isto é, excluindo os valores ignorados. Nestes exemplos, não faz muita diferença incluir ou não os valores ignorados, pois eles representam uma fração insignificante do total. Pode haver situações, porém em que a quantidade de valores ignorados seja tão grande que sua inclusão no total chegue a afetar a interpretação dos resultados. Suponha, por exemplo, uma pesquisa feita para estimar as proporções de eleitores que apoiam os candidatos concorrentes A e B em uma eleição. Os resultados são mostrados na Tabela **6**.

**Tabela 6**

| candidato apoiado | *freqüência absoluta* | *freqüência relativa* | |
|---|---|---|---|
| | | todos os valores | valores válidos |
| A | 200 | 20,0 | 33,3 |
| B | 400 | 40,0 | 66,7 |
| ignorado | 400 | 40,0 | - |
| Total | 1000 | 100,0 | 100,0 |

Não seria correto, neste caso, dizer que o candidato B é apoiado por 66,7% dos eleitores, e portanto tem a maioria dos votos; a quantidade de eleitores de candidato "ignorado" foi tão grande que eles podem vir a afetar o resultado da eleição. Este tipo de raciocínio, contudo, às vezes é encontrado mesmo em publicações científicas. Goldacre cita o caso de um experimento que visava avaliar a eficácia da arnica homeopática como anestesia em cirurgia, usando uma amostra de 59 pacientes, dos quais uma parte recebia arnica, e outra recebia um placebo. Os pesquisadores concluíram que os pacientes com arnica sentiram "significativamente" menos dor do que os do placebo. O problema é que houve 41 pacientes que abandonaram o tratamento antes do fim, e a opinião deles não foi registrada; não podemos concluir nada, se não soubermos *por que* estes pacientes decidiram abandonar o tratamento (Goldacre, 2008, p. 51).

---

**referências**

Waiselfisz, Julio Jacobo (2013). *Mapa da Violência 2013 - Acidentes de Trânsito e Motocicletas*. Rio de Janeiro: CEBELA / FLACSO, 2013

Goldacre, Ben (2008). *Bad Science*. London: Fourth State.