Aula 2 – Apresentação de dados - parte 2

Nesta aula, você aprenderá:

- a construir distribuições de freqüências agrupadas para variáveis quantitativas discretas e contínuas;
- a construir histogramas e polígonos de freqüência para representar distribuições de freqüências agrupadas;
- a construir gráficos de linha e diagramas de ramos e folhas.

Apresentação de dados quantitativos - conclusão

Variáveis quantitativas discretas

Na aula anterior, você aprendeu a construir distribuições de freqüências para variáveis discretas com poucos valores. No exemplo ali apresentado, há duas variáveis quantitativas discretas: número de dependentes e idade.
A diferença entre elas é que a idade pode assumir um número maior de valores, o que resultaria em uma tabela grande, caso decidíssemos relacionar todos os valores. Além disso, em geral não é necessário apresentar a informação em tal nível de detalhamento. Por exemplo, para as seguradoras de planos de saúde, as faixas etárias importantes - aquelas em que há reajuste por idade - são 0 a 18; 19 a 23; 24 a 28; 29 a 33; 34 a 38; 39 a 43; 44 a 48; 49 a 53; 54 a 58 e 59 ou mais. Sendo assim, podemos agrupar os funcionários segundo essas faixas etárias e construir uma *tabela de freqüências agrupadas* da mesma forma que fizemos para o número de dependentes, só que agora cada freqüência corresponde ao número de funcionários na respectiva faixa etária. Na **Tabela 2.1** temos a tabela resultante.

TABELA DE FREQÜÊNCIAS AGRUPADAS

Tabela 2.1: Distribuição de freqüência da idade de 500 funcionários

Faixa	Freqüência Simples		Freqüência Acumulada	
Etária	Absoluta	Relativa %	Absoluta	Relativa %
19 – 23	1	0,2	1	0,2
24 – 28	23	4,6	24	4,8
29 – 33	103	20,6	127	25,4
34 – 38	246	49,2	373	74,6
39 – 43	52	10,4	425	85,0
44 – 48	50	10,0	475	95,0
49 – 53	25	5,0	500	100,0
Total	500	100,0		

Atividade 2.1

Na **Tabela 2.2** temos as informações sobre idade e salário para os 15 funcionários do Departamento de Recursos Humanos. Construa uma tabela de freqüências para a idade, levando em conta as mesmas faixas etárias utilizadas acima.

Tabela 2.2: Idade e salário dos funcionários do Departamento de RH

Nome	Idade	Salário
João da Silva	36	6300
Pedro Fernandes	51	5700
Maria Freitas	26	4500
Paula Gonçalves	25	3800
Ana Freitas	29	3200
Luiz Costa	53	7300
André Souza	42	7100
Patrícia Silva	38	5600
Regina Lima	35	6400
Alfredo Souza	45	7000
Margarete Cunha	26	3700
Pedro Barbosa	37	6500
Ricardo Alves	24	4000
Márcio Rezende	31	5100
Ana Carolina Chaves	29	4500

Variáveis quantitativas contínuas

Para as variáveis quantitativas contínuas, devemos também trabalhar com distribuições de freqüências agrupadas. O processo de construção é idêntico ao visto para as variáveis discretas, mas aqui devemos tomar um cuidado especial na construção das classes. A escolha dos limites das classes deve ser feita com base na natureza, valores e unidade de medida dos dados. As únicas regras que têm que ser seguidas são as seguintes.

Regra: Definição das classes em uma distribuição de freqüências agrupadas.

1. As classes têm que ser exaustivas, isto é, todos os elementos devem pertencer a alguma classe.
2. As classes têm que ser mutuamente exclusivas, isto é, cada elemento tem que pertencer a uma única classe.

O primeiro passo é definir o número de classes desejado; esse número, de preferência, deve estar entre 5 e 25. Em seguida, devemos determinar a *amplitude* dos dados, ou seja, o intervalo de variação dos valores observados da variável em estudo.

Definição
A **amplitude** de um conjunto de dados, representada por Δ_{total}, é definida como a diferença entre os valores máximo e mínimo:

$$\Delta_{total} = V_{\text{Máx}} - V_{\text{Mín}} \tag{2.1}$$

Se queremos trabalhar com classes de mesmo comprimento (e essa é uma opção bastante comum), para determinar esse comprimento, temos que dividir a amplitude total pelo número de classes desejado. No entanto, para garantir a inclusão dos valores mínimo e máximo, podemos, como regra geral, usar o seguinte procedimento: considere o primeiro múltiplo do número de classes maior que o valor da amplitude e use esse número como a nova amplitude. Por exemplo se a amplitude é 28 e queremos trabalhar com 5 classes, vamos considerar 30 como a nova amplitude. Dividindo esse valor pelo número de classes, obtemos o comprimento de cada classe e os limites de classe podem ser obtidos somando-se o comprimento de classe a partir do valor mínimo dos dados. Continuando com o nosso exemplo, o comprimento de classe é $30 \div 5 = 6$; se o valor mínimo dos dados é 4, então os limites de classe serão

$$\begin{aligned} 4+6 &= 10 \\ 10+6 &= 16 \\ 16+6 &= 22 \\ 22+6 &= 28 \\ 28+6 &= 34 \end{aligned}$$

e as classes serão:

$$\begin{gathered} [4,10) \\ [10,16) \\ [16,22) \\ [22,28) \\ [28,34) \end{gathered}$$

Note o tipo de intervalo utilizado: para incluir o valor mínimo (4) na primeira classe, o intervalo tem que ser fechado na parte inferior: [4. Se fechássemos o intervalo no limite superior, o 10 estaria incluído na primeira classe e, portanto, não poderia estar na segunda classe. Isso resultaria em $[4,10]$ como a primeira classe e $(10,16)$ como a segunda classe. Ou seja, as duas primeiras classes estariam definidas de forma diferente, o que não é conveniente, pois dificulta a leitura da tabela. Assim, é preferível incluir o 10 na segunda classe, o que resulta nas classes apresentadas anteriormente.

Exemplo 2.1

Suponha que entre os 500 funcionários da nossa empresa, o menor salário seja 2800 e o maior salário seja de 12400. Para agrupar os dados em 5 classes devemos fazer o seguinte:

$$\begin{aligned} \Delta_{total} = V_{\text{Máx}} - V_{\text{Mín}} = &\quad 12400 - 2800 = 9600 \\ \\ \text{Próximo múltiplo de } 5 = &\quad 9605 \\ \\ \text{Comprimento de classe} = &\quad \frac{9605}{5} = 1921 \end{aligned}$$

Os limites de classe, então, são:

$$\begin{aligned} & 2800 \\ 2800 + 1921 &= 4721 \\ 4721 + 1921 &= 6642 \\ 6642 + 1921 &= 8563 \\ 8563 + 1921 &= 10484 \\ 10484 - 1921 &= 12405 \end{aligned}$$

e as classes podem ser definidas como

$[2800, 4721)$	(2800 incluído; 4721 excluído)
$[4721, 6642)$	(4721 incluído; 6642 excluído)
$[6642, 8563)$	(6642 incluído; 8563 excluído)
$[8563, 10484)$	(8563 incluído; 10484 excluído)
$[10484, 12405)$	(10484 incluído; 12405 excluído)

Essa é uma regra que resulta em classes corretamente definidas, mas nem sempre as classes resultantes são apropriadas ou convenientes. No exemplo acima, poderia ser preferível trabalhar com classes de comprimento 2000, definindo o limite inferior dos dados como 2500. Isso resultaria nas classes [2500,4500); [4500,6500); [6500,8500); [8500, 10500) e [10500, 12500), que são classes corretas e mais fáceis de ler.

Atividade 2.2

Construa uma distribuição de freqüências agrupadas em 5 classes de mesmo comprimento para os dados de salários da **Tabela 2.2**.

Histogramas e polígonos de freqüência

O histograma e o polígono de freqüências são gráficos usados para representar uma distribuição de freqüências simples de uma variável quantitativa contínua.

HISTOGRAMA

Um histograma é um conjunto de retângulos com bases sobre um eixo horizontal dividido de acordo com os comprimentos de classes, centros nos pontos médios das classes e *áreas proporcionais ou iguais às freqüências*. Vamos ilustrar a construção de um histograma usando como exemplo a distribuição de freqüência dos dados sobre salários da Atividade 2, reproduzida na **Tabela 2.3**.

Tabela 2.3: Distribuição dos salários dos funcionários do Departamento de RH

Classe	Freqüência Simples		Freqüência Acumulada	
de salário	Absoluta	Relativa %	Absoluta	Relativa %
[3200,4021)	4	26,67	4	26,67
[4021,4842)	2	1,33	6	40,00
[4842,5663)	2	1,33	8	53,33
[5663,6484)	3	20,00	11	73,33
[6484,7305)	4	26,67	15	100,00
Total	15	100,00		

Como as classes têm o mesmo comprimento, o histograma, nesse caso, pode ser construído de tal modo que as alturas dos retângulos sejam iguais às freqüências das classes. Dessa forma, as áreas serão *proporcionais* (e não iguais) às freqüências, conforme ilustrado no gráfico à esquerda na **Figura 2.1**. No gráfico à direita nessa mesma figura, a área de cada retângulo é *igual* à freqüência relativa da classe e a altura de cada classe é calculada usando-se a expressão que dá a área de um retângulo. Por exemplo, para a classe [3200,4021), a freqüência (área) é $\frac{4}{15} = 0,266667$ e a base do retângulo (comprimento de classe) é 821. Logo, a altura h do retângulo correspondente é

$$h = \frac{0,266667}{821} = 0,000325$$

DENSIDADE

O resultado dessa divisão é denominado *densidade*, uma vez que dá a concentração em cada classe por unidade da variável. Em ambos os gráficos, a forma dos retângulos é a mesma; o que muda é a escala no eixo vertical.

De modo geral, quando as classes têm o mesmo comprimento – e essa é a situação mais comum – podemos representar as alturas dos retângulos pelas freqüências das classes, o que torna o gráfico mais fácil de interpretar.

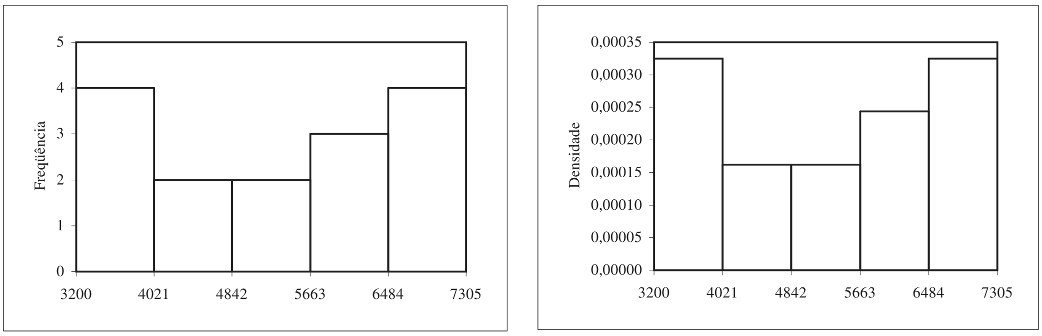

Figura 2.1: Histogramas da distribuição dos salários dos funcionários do Departamento de RH.

POLÍGONO DE FREQÜÊNCIAS

Um *p*olígono de freqüências é um gráfico de linha que se obtém unindo por uma poligonal os pontos correspondentes às freqüências das diversas classes, centradas nos respectivos pontos médios. Mais precisamente, são plotados os pontos com coordenadas (ponto médio, freqüência simples). Para obter as interseções da poligonal com o eixo, cria-se em cada extremo uma classe com freqüência nula. Na **Figura 2.2** temos o polígono de freqüências para a renda dos funcionários do Departamento de Recursos Humanos. Pode-se construir o polígono de freqüências junto com o histograma, o que facilita a visualização dos resultados.

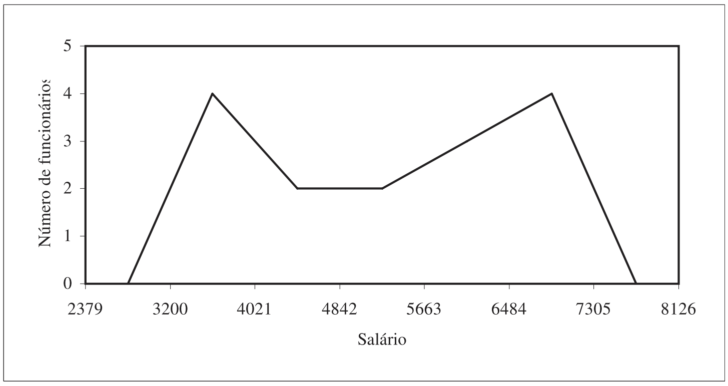

Figura 2.2: Polígono de freqüência para os salários dos funcionários do Departamento de RH.

Atividade 2.3

Na **Tabela 2.4** abaixo temos as notas de 50 alunos em uma prova. Construa uma tabela de freqüências agrupadas, usando as classes $2 \vdash 3, 3 \vdash 4, 4 \vdash 5, \cdots, 9 \vdash 10$. Construa o histograma e o polígono de freqüências. (O símbolo $\vdash$ indica que o limite inferior está incluído, mas não o superior; se quiséssemos o contrário, usaríamos o símbolo $\dashv$ para indicar que o limite superior está incluído, mas não o limite inferior.)

Tabela 2.4: Notas de 50 alunos para a Atividade 3

2,9	3,7	3,8	4,7	4,9	5,2	5,6	5,8	6,0	6,2
6,3	6,3	6,3	6,5	6,5	6,6	6,8	6,8	6,9	6,9
7,0	7,0	7,1	7,3	7,3	7,4	7,4	7,5	7,5	7,6
7,6	7,7	7,7	7,9	8,1	8,1	8,2	8,2	8,3	8,3
8,4	8,5	8,7	8,7	8,8	8,9	9,0	9,1	9,4	9,7

Diagrama de ramos e folhas

RAMO E FOLHAS

Um outro gráfico usado para mostrar a forma da distribuição de um conjunto de dados quantitativos é o diagrama de ramos e folhas, desenvolvido pelo estatístico americano John Tukey. Este gráfico é constituído de uma linha vertical, com a escala indicada à esquerda desta linha. A escala, naturalmente, depende dos valores observados, mas deve ser escolhida de tal forma que cada valor observado possa ser "quebrado" em duas partes: uma primeira parte quantificada pelo valor da escala e a segunda quantificada pelo último algarismo do número correspondente à observação. Os *ramos* do gráfico correspondem aos números da escala, à esquerda da linha vertical. Já as *folhas* são os números que aparecem na parte direita. Na **Figura 2.3** temos o diagrama de ramos e folhas para as notas de 50 alunos dadas na **Tabela 2.4**. Nesse caso, a "quebra" dos valores é bastante natural: os ramos são formados pelo algarismo inteiro e as folhas pelos algarismos decimais, o que é indicado pela escala do gráfico.

RAMOS

FOLHAS

```
Escala
 1 | 0     1,0

 2 | 9
 3 | 7 8
 4 | 7 9
 5 | 2 6 8
 6 | 0 2 3 3 3 5 5 6 8 8 9 9
 7 | 0 0 1 3 3 4 4 5 5 6 6 7 7 9
 8 | 1 1 2 2 3 3 4 5 7 7 8 9
 9 | 0 1 4 7
```

Figura 2.3: Notas de 50 alunos.

O diagrama de ramos e folhas também é útil na comparação de conjuntos de dados. Suponha que, no exemplo acima, a mesma prova tenha sido aplicada a duas turmas diferentes. Para comparar os resultados, podemos construir o diagrama que se encontra na **Figura 2.4**. Para facilitar a comparação, é usual indicar o número de dados em cada banda do diagrama. Note que, na parte esquerda do gráfico, as folhas são anotadas crescentemente da direita para a esquerda, enquanto que, na parte direita do gráfico, as folhas são anotadas crescentemente da esquerda para a direita.

Escala

1 | 0 1,0

n = 32		n = 50
8	1	
3 2 2	2	9
	3	7 8
7 5 0	4	7 9
2 1	5	2 6 8
6 5 4 3 3 2 0 0 0 0	6	0 2 3 3 3 5 5 6 8 8 9 9
2 2 2 0 0	7	0 0 1 3 3 4 4 5 5 6 6 7 7 9
4 3 3 2 1 0 0	8	1 1 2 2 3 3 4 5 7 7 8 9
5	9	0 1 4 7

Figura 2.4: Comparação das notas de 2 turmas.

Atividade 2.4

Suponha que as idades dos 23 funcionários do Departamento Financeiro sejam 27; 31; 45; 52; 33; 34; 29; 27; 35; 38; 50; 48; 29; 30; 32; 29; 42; 41; 40; 42; 28; 36; 48. Usando esses dados e aqueles apresentados na **Tabela 2.2** sobre os funcionários do Departamento de Recursos Humanos, construa um diagrama de ramos e folhas para comparar os 2 departamentos.

Gráficos de linhas

GRÁFICO DE LINHAS

SÉRIE TEMPORAL

O *gráfico de linhas* é usado principalmente para representar observações feitas ao longo do tempo, isto é, observações de uma *série de tempo*. No eixo horizontal colocam-se as datas em que foram realizadas as observações e no eixo vertical, os valores observados. Os pontos assim obtidos são unidos por segmentos de reta para facilitar a visualização do comportamento dos dados ao longo do tempo. Na **Figura 2.5** temos o gráfico que ilustra o consumo de refrigerante (em milhões de litros) no período de 1986 a 2005, conforme dados da ABIR.

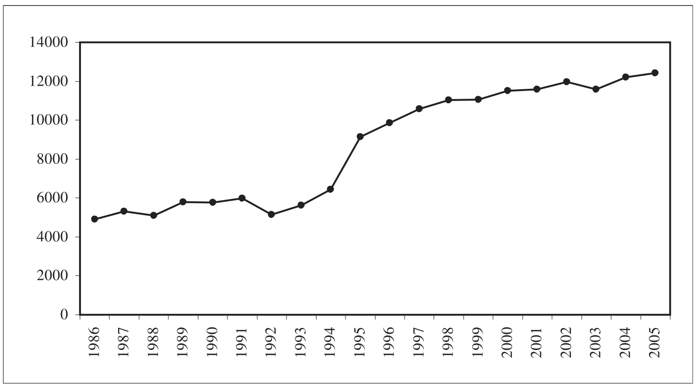

Figura 2.5: Consumo de refrigerante 1986-2005.

Na **Tabela 2.5** temos dados sobre o número de homicídios nos estados do Rio de Janeiro e de São Paulo no período de 1980 a 2002. Para efeitos de comparação, é possível construir um gráfico de linhas em que as 2 séries são representadas conjuntamente. Veja a **Figura 2.6**.

Tabela 2.5: Número de homicídios - RJ e SP - 1980 a 2002

Ano	RJ	SP	Ano	RJ	SP
1980	2946	3452	1992	4516	9022
1981	2508	4187	1993	5362	9219
1982	2170	4183	1994	6414	9990
1983	1861	5836	1995	8183	11566
1984	2463	7063	1996	8049	12350
1985	2550	7015	1997	7966	12522
1986	2441	7195	1998	7569	14001
1987	3785	7918	1999	7249	15810
1988	3054	7502	2000	7337	15631
1989	4287	9180	2001	7304	15745
1990	7095	9496	2002	8257	14494
1991	5039	9671			

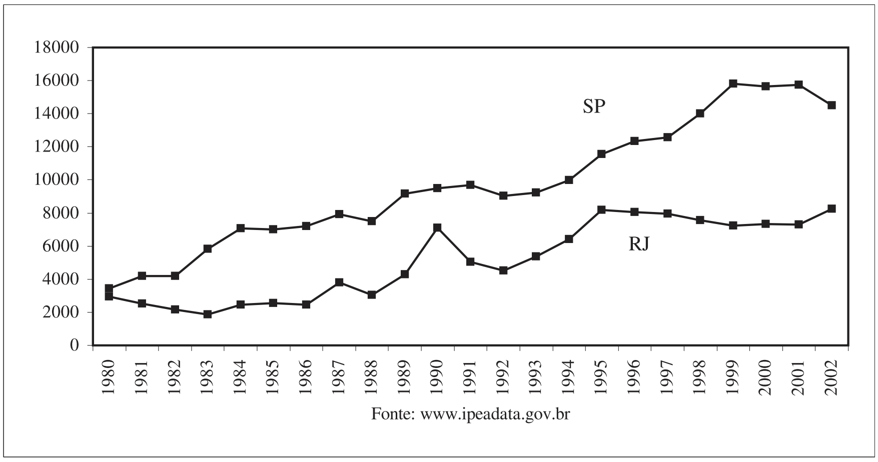

Figura 2.6: Número de homicídios nos estados do Rio de Janeiro e de São Paulo – 1980–2002.

Resumo da Aula

- Nesta aula, você completou seu estudo sobre apresentação de dados quantitativos com o estudo das distribuições de freqüências agrupadas. Certifique-se de ter compreendido a forma de obter os limites das classes, respeitando o fato de que cada observação tem que pertencer a uma única classe (as classes têm que ser exaustivas e mutuamente exclusivas). As distribuições de freqüências agrupadas podem ser representadas graficamente pelos histogramas e pelos polígonos de freqüência.
- O histograma é um gráfico representado por um conjunto de retângulos com base sobre um eixo horizontal dividido de acordo com os comprimentos das classes. As bases dos retângulos estão centradas nos pontos médios das classes e a área de cada retângulo é proporcional à freqüência da classe. Quando as classes têm comprimentos iguais, as alturas dos retângulos podem ser iguais às freqüências das respectivas classes.
- O polígono de freqüências é um gráfico em forma de uma poligonal, que une os pontos de coordenadas (ponto médio da classe, freqüência da classe).
- Na construção do diagrama de ramos e folhas, cada valor é "quebrado" em duas partes, de modo que uma das partes - a folha - corresponda ao último algarismo do valor. A outra parte é o ramo e em cada ramo colocam-se as folhas pertencentes ao ramo.
- Os gráficos de linha são utilizados na representação de séries temporais, que são dados observados ao longo do tempo. No eixo horizontal representam-se as datas em que os dados foram coletados e na escala vertical, os valores observados.

Exercícios

1. Num estudo sobre a jornada de trabalho das empresas de Produtos Alimentares foram levantados os dados da **Tabela 2.6** relativos ao total de horas trabalhadas pelos funcionários no mês de agosto (dados fictícios). Construa uma tabela de freqüências usando 5 classes de mesmo tamanho; construa também o histograma e o polígono de freqüências. Para facilitar a solução, os valores mínimo e máximo são: 1.815 e 118.800.

Tabela 2.6: Jornada de trabalho de empresas alimentares para o Exercício **2.6**

3.960	5.016	13.015	8.008	6.930	5.544	4.224	6.138
118.800	57.904	72.600	100.100	55.935	7.223	3.775	4.224
3.216	7.392	2.530	6.930	1.815	4.338	8.065	10.910
8.408	8.624	6.864	5.742	5.749	8.514	2.631	5.236
8.527	3.010	5.914	11.748	8.501	6.512	11.458	10.094
6.721	2.631	7.082	10.318	8.008	3.590	7.128	7.929
10.450	6.780	5.060	5.544	6.178	13.763	9.623	14.883
17.864	34.848	25.300	52.800	17.732	63.923	30.360	18.876
30.800	19.562	49.240	49.434	26.950	22.308	21.146	14.212
25.520	49.251	30.976	23.338	43.648	26.796	44.880	30.008
30.769	16.907	33.911	27.034	16.500	14.445	28.160	42.442
16.507	36.960	67.760	84.084	89.888	65.340	82.280	86.152
91.080	99.792	77.836	76.032				

2. Na **Tabela 2.7** temos a população dos municípios de MG com mais de 50.000 habitantes, com base nos dados do Censo Demográfico 2000. Construa uma tabela de freqüências, trabalhando com as seguintes classes (em 1.000 hab.): [50,60), [60,70), [70,80), [80,100), [100,200), [200, 500) e 500 ou mais. Note que aqui estamos trabalhando com classes desiguais, o que é comum em situações desse tipo, onde há muitos observações pequenas e poucas grandes.

Tabela 2.7: População dos municípios de MG com mais de 50.000 habitantes, para o **Exercício 2.7**

Município	População	Município	População	Município	População
Leopoldina	50.097	Timóteo	71.478	Varginha	108.998
Pirapora	50.300	Pará de Minas	73.007	Barbacena	114.126
três Pontas	51.024	Patrocínio	73.130	Sabará	115.352
São Francisco	51.497	Paracatu	75.216	Patos de Minas	123.881
Pedro Leopoldo	53.957	Vespasiano	76.422	Teófilo Otoni	129.424
Ponte Nova	55.303	Itaúna	76.862	Ibirité	133.044
S.Seb.do Paraíso	58.335	Caratinga	77.789	Poços de Caldas	135.627
Janaúba	61.651	S.João del Rei	78.616	Divinópolis	183.962
Formiga	62.907	Lavras	78.772	Sete Lagoas	184.871
Januária	63.605	Araxá	78.997	Santa Luzia	184.903
Cataguases	63.980	Itajubá	84.135	Ipatinga	212.496
Nova Lima	64.387	Ubá	85.065	Ribeirão das Neves	246.846
Viçosa	64.854	Ituiutaba	89.091	Gov.Valadares	247.131
Três Corações	65.291	Muriaé	92.101	Uberaba	252.051
Ouro Preto	66.277	Passos	97.211	Betim	306.675
João Monlevade	66.690	Cor. Fabriciano	97.451	Montes Claros	306.947
Alfenas	66.957	Itabira	98.322	Juiz de Fora	456.796
Manhuaçu	67.123	Araguari	101.974	Uberlândia	501.214
Curvelo	67.512	Cons.Lafaiete	102.836	Contagem	538.017
Unaí	70.033	Pouso Alegre	106.776	Belo Horizonte	2.238.526

Fonte: IBGE - Censo Demográfico 2000

3. Na **Tabela 2.8** temos a densidade populacional (hab/km^2) das unidades da federação brasileira. Construa um gráfico ramo-e-folhas para esses dados. Para RJ e DF, você pode dar um "salto" na escala, de modo a não acrescentar muitos ramos vazios.

Tabela 2.8: Densidade populacional dos estados brasileiros, para o Exercício **2.8**

UF	Densidade Populacional (hab/km^2)	UF	Densidade Populacional (hab/km^2)
RO	6	SE	81
AC	4	BA	24
AM	2	MG	31
RR	2	ES	68
PA	5	RJ	328
AP	4	SP	149
TO	5	PR	48
MA	17	SC	57
PI	12	RS	37
CE	51	MS	6
RN	53	MT	3
PB	61	GO	15
PE	81	DF	353
AL	102		

Fonte: IBGE - Censo Demográfico 2000

4. Construa um gráfico de linhas para os dados da inflação brasileira anual medida pelo Índice Nacional de Preços ao Consumidor (INPC) apresentados na **Tabela 2.9**.

Tabela 2.9: Índice Nacional de Preços ao Consumidor - 1995-2005

Ano	INPC (%)
1995	22,0
1996	9,1
1997	4,3
1998	2,5
1999	8,4
2000	5,3
2001	9,4
2002	14,7
2003	10,4
2004	6,1
2005	5,1