ANÁLISE EXPLORATÓRIA DE DADOS

AULA 3

Prof. Dayane Perez Bravo

CONVERSA INICIAL

Nesta aula, continuaremos discutindo como realizar a visualização de dados a fim de determinar conclusões que, a princípio, não são facilmente visíveis para um conjunto de dados desordenados.

TEMA 1 – HISTOGRAMA E POLÍGONO DE FREQUÊNCIA

Vejamos como construir o histograma e o polígono de frequência: tipos de representação gráfica que permitem nos dar uma visualização adequada dos dados do tipo contínuo.

1.1 Histograma

Como visto ao final da aula anterior, a tabela de frequência permite descrever um conjunto de dados contínuos a partir de um processo simples de agrupamento em classes e de contagem. Discutimos que, caso isso não fosse feito, as tabelas de frequência dificilmente representariam algo relevante, visto que é improvável que um mesmo valor contínuo aconteça várias vezes ao longo de um experimento.

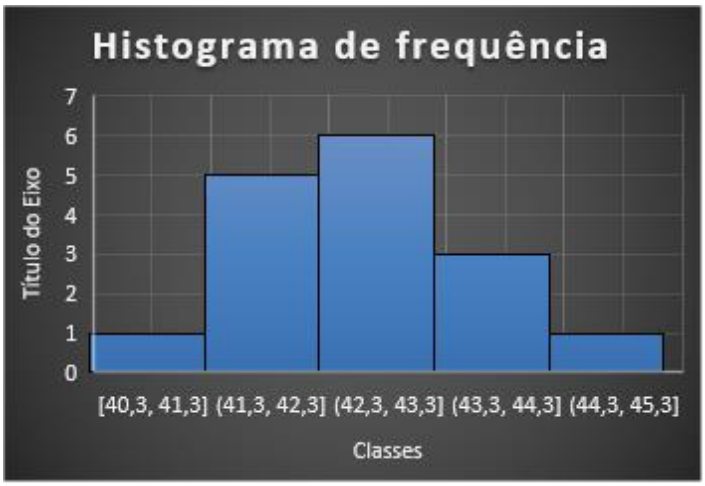
Agora, vejamos como utilizar um histograma para realizar a representação gráfica de uma tabela de frequência. Na aula anterior, geramos a tabela de frequência da classe, apresentada agora na Tabela 1, com o tempo de corrida de 16 corredores.

Tabela 1 – Tabela de frequência da classe para os tempos de 16 corredores

Classe	Frequência
$40{,}0 \vdash 41{,}0$	1
$41{,}0 \vdash 42{,}0$	2
$42{,}0 \vdash 43{,}0$	5
$43{,}0 \vdash 44{,}0$	7
$44{,}0 \vdash 45{,}0$	1
	16

Note que construímos, na Figura 1, o histograma de frequência para os dados contínuos da Tabela 1.

Figura 1 – Histograma de frequência para os dados dos 16 corredores agrupados em classes



Note que esse gráfico permite visualizar em torno de qual região os dados estão concentrados. Veja que, rapidamente, pode-se notar que existem pontos que ocorrem mais afastados da medida central; em alguns casos, podendo ser considerados como *outliers.*

1.2 Roteiro para a construção do histograma

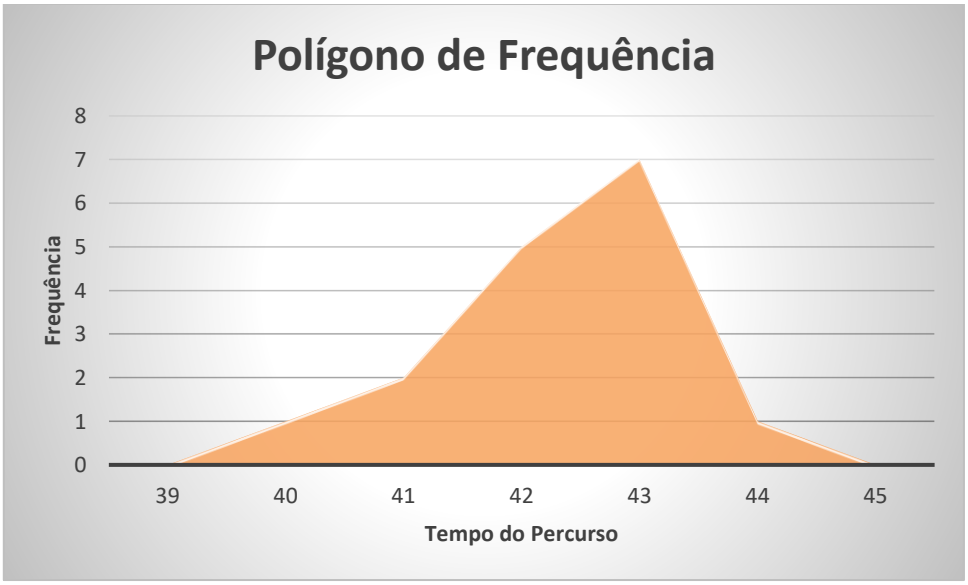
1. Para a construção do histograma, inicialmente, devemos gerar a tabela de frequência para os dados, a partir do agrupamento em classes pré-definidas.
2. Na sequência, deve-se gerar os eixos de forma proporcional a fim de obter uma visualização adequada.
3. Ao dividir o eixo horizontal, recomenda-se dividir pela quantidade de classes somado com 2 unidades. Assim, restará espaço no canto esquerdo e direito do gráfico. Esse espaço é necessário, principalmente, na construção do polígono de frequência. No eixo, deve-se marcar os limites inferiores e superiores de cada uma das classes.
4. Para o eixo vertical, deve-se atentar para a maior frequência obtida. No caso do exemplo dado, é o valor de 7. Na sequência, deve-se dividir o eixo vertical em uma quantidade razoável de divisões.
5. A largura de cada retângulo que será desenhado irá representar a largura da classe, enquanto a altura de cada um representa a frequência da classe.

6. Por fim, não deve se esquecer de adicionar o título do gráfico e as legendas dos eixos.

1.3 Polígono de frequência

O **polígono de frequência** é obtido unindo os pontos médios da parte superior, também chamada de *patamar*, de cada um dos retângulos apresentados no histograma.

Figura 2 – Polígono de frequência para o tempo do percurso para os 16 corredores analisados



Note que, na Figura 2, construímos o polígono de frequência para o tempo de percurso para os 16 corredores analisados unindo o ponto médio de cada um dos retângulos do histograma.

Essa visualização por polígono de frequência é frequentemente utilizada para um conjunto grande de dados e para verificar se uma determinada distribuição possui características de normalidade, geralmente discutidas nas disciplinas de estatística.

TEMA 2 – FREQUÊNCIA ACUMULADA E OGIVA

Para o tratamento e representação de dados de frequência acumulada para dados contínuos, utilizamos um tipo de representação conhecido como *ogiva*.

2.1 Frequência Acumulada

Antes de realizar a representação gráfica, vejamos como encontrar a frequência acumulada. Para essa análise, usaremos os dados dos 16 corredores utilizados no exemplo anterior.

Tratamos frequência acumulada para dados discretos. Quando precisamos realizar a construção para dados contínuos, definimos **frequência acumulada** como o número de dados menores que o limite superior da classe considerada. De forma similar, podemos definir a **frequência relativa acumulada** e a **porcentagem acumulada:**

$$\text{Frequência relativa Acumulada} = \frac{\text{Frequência Acumulada}}{\text{Número de Dados}}$$

$$\text{Porcentagem Relativa Acumulada} = (\text{Frequência Relativa Acumulada}) \times 100$$

Assim, note que, antes de construir a ogiva, devemos ampliar a Tabela 1, gerando a Tabela 2 com os dados acumulados.

Tabela 2 – Tabela de frequência, frequência relativa acumulada e porcentagem relativa acumulada da classe para os tempos de 16 corredores

Classe	Freq.	Freq. Acumulada	Freq. Rela. Acumulada	Porcent. Acumulada.
$40{,}0 \vdash 41{,}0$	1	1	0,06	6
$41{,}0 \vdash 42{,}0$	2	3	0,19	19
$42{,}0 \vdash 43{,}0$	5	8	0,50	50
$43{,}0 \vdash 44{,}0$	7	15	0,94	94
$44{,}0 \vdash 45{,}0$	1	16	1,00	100
	16			

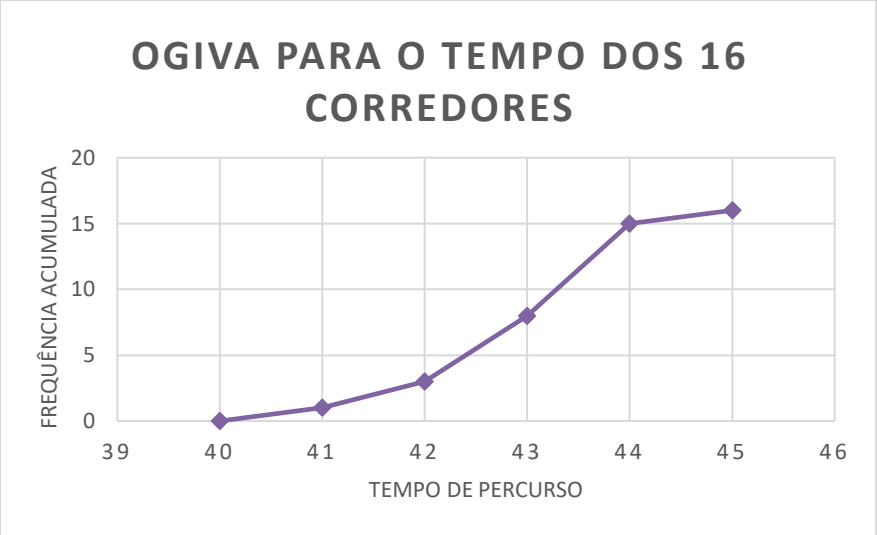
2.2 Ogiva

Para realizarmos a representação gráfica das informações da Tabela 2, utilizamos a **ogiva** a partir de segmentos de reta interligando os pontos dados pelo limite superior de cada classe e a frequência acumulada. Também precisamos utilizar o ponto definido pelo limite inferior da primeira classe e zero.

Note que, para o caso dos dados da Tabela 2, devemos colocar, portanto, os seguintes pontos:

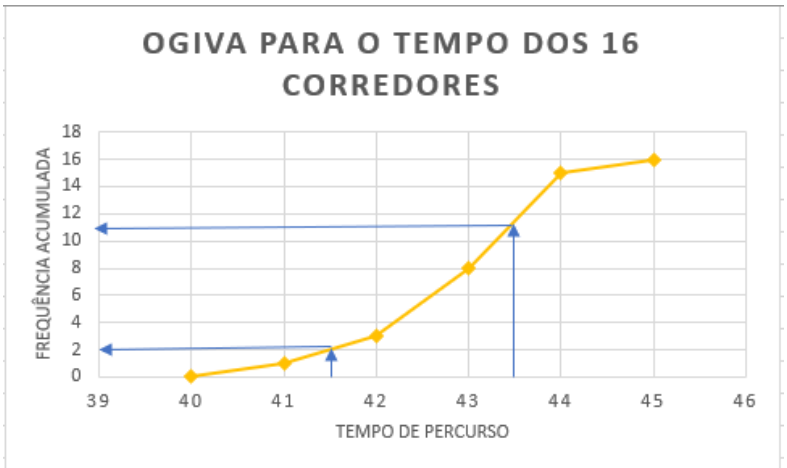
$$(40{,}0; 0), (41{,}0; 1), (42{,}0; 3), (43{,}0; 8), (44{,}0; 15), (45{,}0; 16)$$

Figura 3 – Ogiva para o tempo do percurso para os 16 corredores analisados



A Figura 3 apresenta a ogiva construída para o tempo de percurso dos 16 corredores analisados. Note que, ao visualizar o gráfico construído, rapidamente podemos estimar a frequência acumulada para qualquer outro dado. Por exemplo, o número aproximado de corredores que a terminaram em menos de 41,5 minutos é de aproximadamente 2; enquanto os que terminaram em menos de 43,5 minutos é de aproximadamente 11. Essa análise pode ser observada na Figura 4.

Figura 4 – Análise da ogiva para o tempo do percurso para os 16 corredores analisados



Note também que podemos utilizar a ogiva para **estimar** a quantidade de observações em uma determinada classe arbitrária. Com os dados da Figura 4, por exemplo, podemos estimar que 9 corredores percorreram a atividade em um tempo de percurso entre 41,5 minutos e 43,5 minutos. Veja que esse resultado

representa uma estimativa, visto que não temos certeza do número 2 e do número 11 estimados para as frequências acumuladas com o uso do gráfico.

2.3 Exemplo

Considere que você está preocupado com o tempo que leva para ir a seu trabalho. Para retirar alguma conclusão significativa, decide monitorar o tempo de viagem durante 30 dias, e os dados que encontrou estão disponíveis na Tabela 3.

Tabela 3 – Tabela de frequência da classe para o tempo que levou para ir ao trabalho durante 30 dias

Classe	Frequência
30 ⊢ 40	3
40 ⊢ 50	16
50 ⊢ 60	7
60 ⊢ 70	2
70 ⊢ 80	1
80 ⊢ 90	1
	30

Caso deseje retirar informações sobre os quartis ou sobre a mediana dos dados, pode fazer uma estimativa a partir do uso da ogiva. Então, nesse caso, expandimos a Tabela 3 para gerar a Tabela 4 contendo os dados de frequência acumulada, frequência acumulada relativa e porcentagem acumulada relativa.

Tabela 4 – Tabela de frequência acumulada da classe para o tempo que levou para ir ao trabalho durante 30 dias

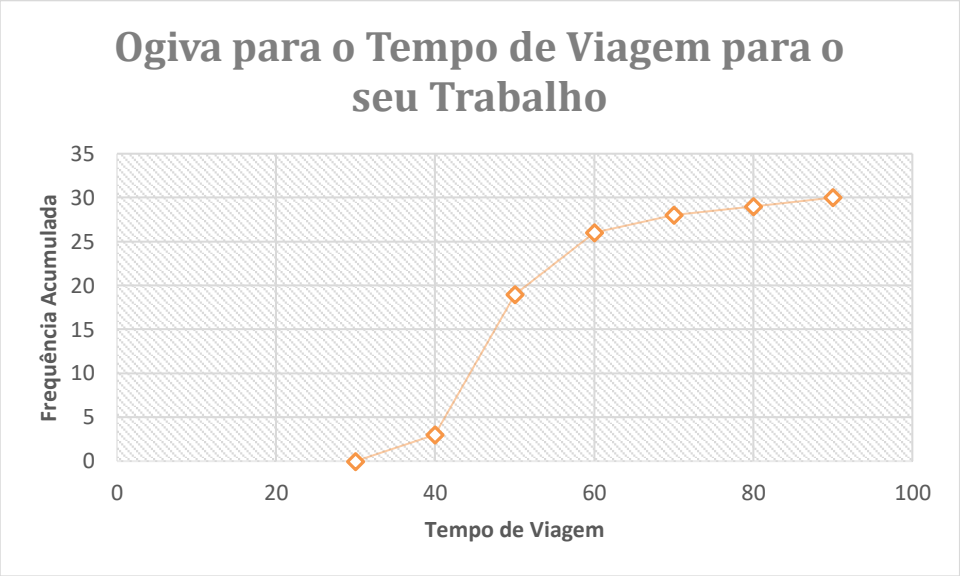
Classe	Freq.	Freq. Acumulada	Freq. Rela. Acumulada	Porcen. Acumulada.
30 ⊢ 40	3	3	0,10	10
40 ⊢ 50	16	19	0,63	63
50 ⊢ 60	7	26	0,87	87
60 ⊢ 70	2	28	0,93	93
70 ⊢ 80	1	29	0,97	97
80 ⊢ 90	1	30	1,00	100
	30			

Agora é possível criar a ogiva, representando os seguintes pontos:

$$(30,0), (40,3), (50,19), (60,26), (70,28), (80,29), (90,30)$$

Com eles, geramos o gráfico da Figura 5.

Figura 5 – Análise da ogiva para o tempo do percurso para os 16 corredores analisados



De forma similar, podemos tirar algumas conclusões: por exemplo, na maior parte dos casos, você chegará no trabalho em menos de 50 minutos!

TEMA 3 – AGRUPAMENTO DE DADOS DISCRETOS EM CLASSE

Vimos que o agrupamento de dados em classe é essencial para os dados contínuos. No caso dos dados discretos, verificamos a existência de outros tipos de gráficos. Entretanto, mesmo considerando dados discretos, em alguns casos, é mais viável realizar a representação da distribuição de frequência por meio do agrupamento em classes.

3.1 Por que agrupar dados discretos em classe

Nosso principal argumento para construir as classes para os dados contínuos foi que, dificilmente, teríamos uma quantidade de dados significativas para um mesmo valor, visto que existem uma infinidade de valores possíveis! Por mais que nos dados discretos essa quantidade não seja infinita, ainda sim, algumas observações podem indicar distribuições com poucas repetições.

Para verificar o que ocorre nesse caso, considere os dados de frequentadores de determinado bar coletado para os 50 primeiros sábados do ano:

65 93 74 102 83 80 62 75 70 96
78 55 100 86 84 77 91 88 53 70
58 71 78 72 98 65 78 83 109 95
79 97 85 68 91 102 83 108 67 80
85 101 74 106 71 97 92 99 89 82

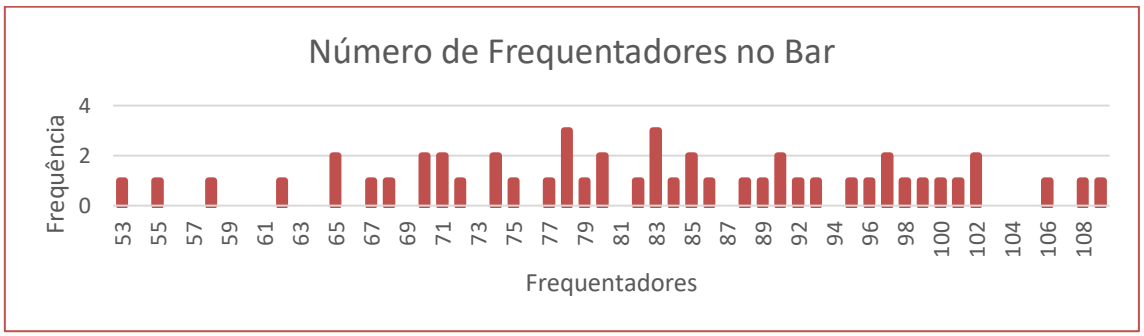
Agora vejamos, na Tabela 5, como a tabela de frequência se comporta para esse tipo de dado!

Tabela 5 – Tabela de frequência para o número de frequentadores de um bar para os primeiros 50 sábados do ano

53	**55**	**58**	**62**	**65**	**67**	**68**	**70**	**71**	**72**	**74**	**75**	**77**
1	1	1	1	2	1	1	2	2	1	2	1	1
78	**79**	**80**	**82**	**83**	**84**	**85**	**86**	**88**	**89**	**91**	**92**	**93**
3	1	2	1	3	1	2	1	1	1	2	1	1
95	**96**	**97**	**98**	**99**	**100**	**101**	**102**	**106**	**108**	**109**		
1	1	2	1	1	1	1	2	1	1	1		

Como argumentado, os dados poucas vezes se repetem. A representação a partir de um gráfico de frequência também não nos traz muita informação significativa, como pode ser observado na Figura 6.

Figura 6 – Gráfico de frequência do número de frequentadores do bar



3.2 Agrupamento de dados discretos em classes

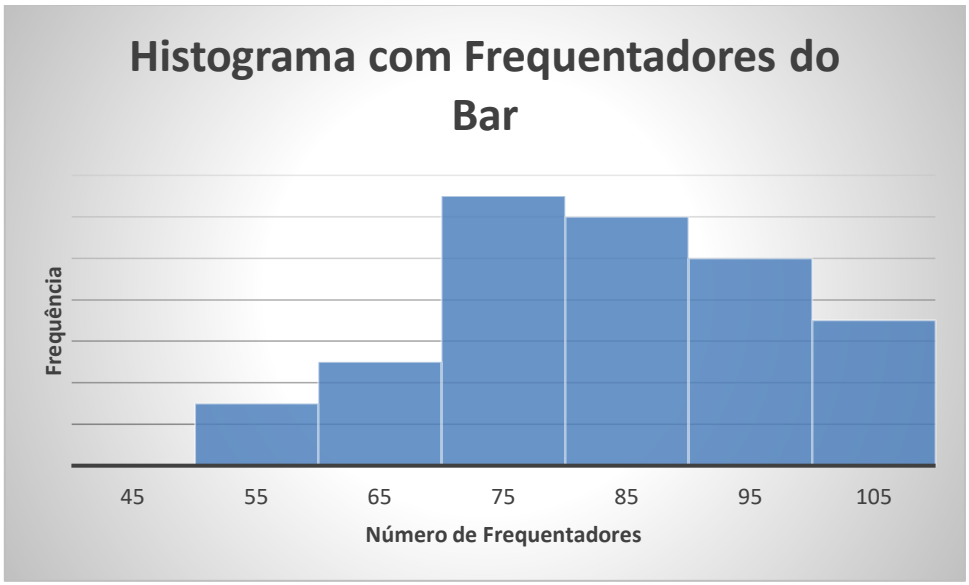
Para extrair algumas informações significativas desse problema, recomenda-se agrupá-lo em classes. Nesse caso, decidimos escolher o limite inferior da primeira classe 50, a largura da classe como 10 e um total de 6 classes. Assim, podemos gerar a Tabela 6.

Tabela 6 – Tabela de frequência da classe para a quantidade de frequentadores do bar

Classe	Frequência
50 ⊢ 60	3
60 ⊢ 70	5
70 ⊢ 80	13
80 ⊢ 90	12
90 ⊢ 100	10
100 ⊢ 110	7
	50

Esses dados podem ser representados a partir do uso de um histograma, como o representado na Figura 7.

Figura 7 – Histograma do número de frequentadores do bar



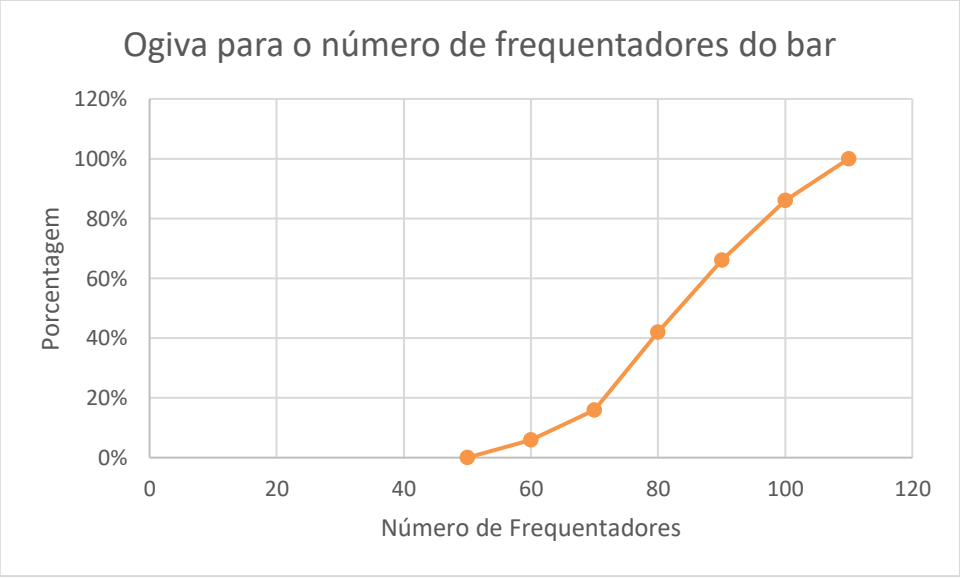
Note que, a partir dessa construção, podemos notar como os dados estão distribuídos: informação que não era possível extrair, diretamente, da figura anterior.

Também podemos expandir a Tabela 6 para gerar a Tabela 7 com os dados de frequência acumulada, frequência acumulada relativa e porcentagem acumulada relativa.

Tabela 7 – Tabela de frequência acumulada de classe para a quantidade de frequentadores do bar

Classe	Freq.	Freq. Acumulada	Freq. Rela. Acumulada	Porcen. Acumulada.
50 ⊢ 60	3	3	0,06	6
60 ⊢ 70	5	8	0,16	16
70 ⊢ 80	13	21	0,42	42
80 ⊢ 90	12	33	0,66	66
90 ⊢ 100	10	43	0,86	86
100 ⊢ 110	7	50	1,00	100
	50			

Figura 8 – Histograma do número de frequentadores do bar



Dessa informação, podemos extrair outras. Por exemplo: geralmente, nesse bar, temos mais de 70 frequentadores em pelo menos 20% dos casos; ou ainda, teremos no máximo 100 frequentadores em cerca de 85% dos finais de semana.

TEMA 4 – REGRAS PARA O NÚMERO DE CLASSES

Nem sempre a quantidade e a largura do número de classes são determinadas de forma arbitrária como fizemos no exemplo anterior. Existem estratégias consolidadas por alguns autores para esses números, as quais serão discutidas nessa seção.

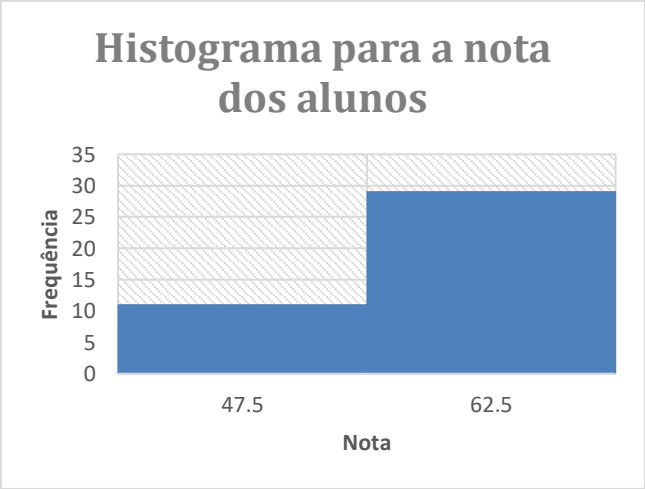
4.1 Exemplo

Para analisar as regras discutidas, considere o caso de um experimento que se coletaram as informações sobre as notas de uma avaliação realizada por 40 alunos:

60 46 58 59 68 49 63 59 50 64
55 68 55 63 56 64 43 56 56 65
57 58 53 59 60 53 61 57 62 53
62 54 57 63 50 57 51 68 55 46

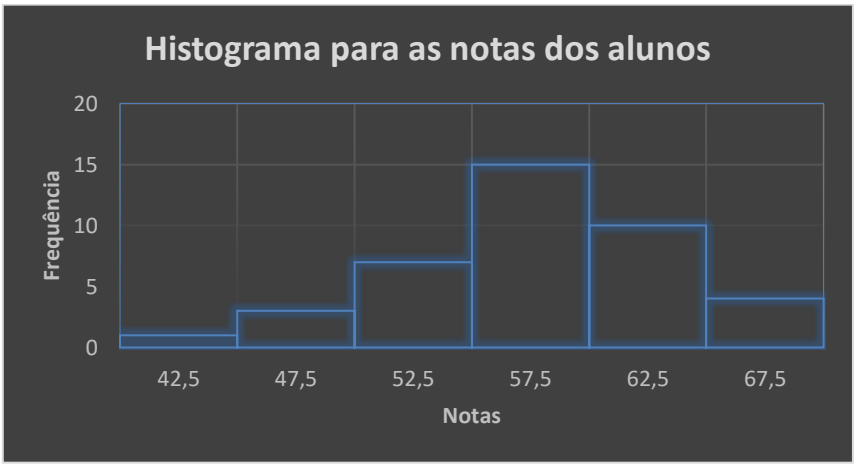
Note que esses dados poderiam ser agrupados em uma quantidade grande de classes. Se isso acontece, a representação fica similar ao que ocorreu na Figura 6 e não conseguimos tirar nenhuma interpretação razoável sobre os dados. Também não é relevante dividir os dados em, digamos, 2 classes. Se isso acontece, a visualização também fica comprometida. Veja isso na Figura 9.

Figura 9 – Histograma do número de frequentadores do bar com dados distribuídos em 2 classes



As regras que veremos na sequência nos farão concluir que devemos ter um total de 6 classes. Assim, obteremos o histograma da Figura 10.

Figura 10 – Histograma do número de frequentadores do bar com dados distribuídos em 6 classes



4.2 Regra do Logaritmo

Uma das possibilidades é utilizar a regra do logaritmo. Note que definimos K como o número de classes e n como o número de observações. Assim, temos:

$$K = 1 + 3{,}3.\log n$$

Podemos fazer alguns exemplos. No caso do exemplo anterior, que possuíamos 40 observações, temos que:

$$K = 1 + 3{,}3.\log 40 \cong 6{,}2867$$

Ao aproximarmos, verificamos que o número de classes ideal é 6.

O comportamento da função logaritmo garante que não teremos um número exagerado de classes.

4.3 Regra da Potência de 2

A outra regra que costuma ser utilizada é a regra da potência de 2. Nesse caso, temos que encontrar o menor valor inteiro de k, tal que:

$$2^k \geq n$$

No caso do exemplo anterior, que possuíamos 40 observações, temos que:

$$2^k \geq 40$$

Note que o menor valor de k também é 6, visto que $2^6 = 64$ e $2^5 = 32$.

4.4 Classes com Larguras Desiguais

Em alguns casos, categorias fazem com que as classes sejam descritas com larguras desiguais. Em alguns níveis de ensino, existem determinadas idades para cada faixa etária. Por exemplo: maternal (0 a 3 anos), educação infantil (3 a 5 anos), ensino fundamental (5 a 15 anos), ensino médio (15 a 18 anos), ensino superior (acima de 18 anos) são as idades aproximadas de cada nível de ensino. Então considere os dados da quantidade de alunos em cada nível para uma determinada instituição, apresentados na Tabela 8.

Tabela 8 – Tabela de frequência de classe para a quantidade de alunos em cada nível de ensino

Categoria	Classe	Frequência
Maternal	$0 \vdash 3$	12
Educação Infantil	$3 \vdash 5$	5
Ensino Fundamental	$5 \vdash 15$	9
Ensino Médio	$15 \vdash 18$	24
Ensino Superior	$18 \vdash 30$	12
		60

Para que essa representação fique adequada, é necessário fazer um histograma envolvendo a idade pela **densidade de frequência.** Assim, definimos:

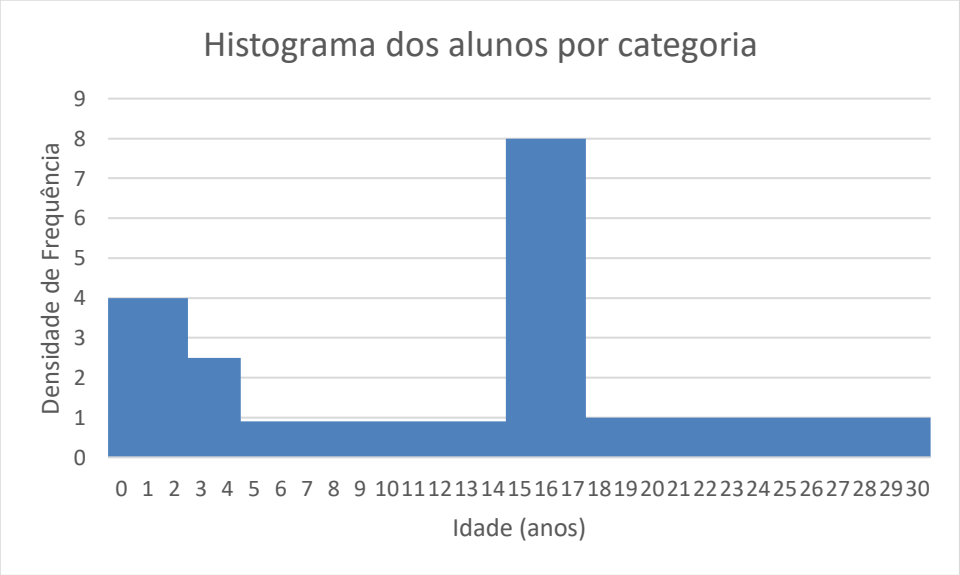
$$\text{Densidade de Frequência} = \frac{\text{Frequência da Classe}}{\text{Largura da Classe}}$$

Assim, expandimos a Tabela 8, gerando a Tabela 9 com os dados de densidade de frequência. A Figura 11 apresenta o histograma para os dados gerados.

Tabela 9 – Tabela de frequência de classe com densidades de frequências para a quantidade de alunos em cada nível de ensino

Categoria	Classe	Frequência	Largura da Classe	Densidade de Frequência
Maternal	0 ⊢ 3	12	3	4
Educação Infantil	3 ⊢ 5	5	2	2,5
Ensino Fundamental	5 ⊢ 15	9	10	0,9
Ensino Médio	15 ⊢ 18	24	3	8
Ensino Superior	18 ⊢ 30	12	12	1
		60		

Figura 11 – Histograma da quantidade de alunos por nível de ensino a partir da densidade de frequência



TEMA 5 – DIAGRAMA DE RAMOS E FOLHAS

Nem sempre o agrupamento dos dados em classes é interessante, visto que podemos perder informações acerca dos dados individuais. Uma forma de representação interessante para preservar essa informação é chamada *diagrama de ramos e folhas.*

5.1 Construção do Diagrama de Ramos e Folhas

Considere o público que frequentou uma determinada barbearia em 20 semanas diferentes:

68 82 75 43 59 80 69 60 73 54

51 93 70 32 63 61 76 87 65 71

Para construirmos o diagrama de ramos e folhas, separamos cada um dos números de acordo com algum critério em relação aos algarismos, por exemplo, a partir da dezena e da unidade. Assim, a partir desse critério, a dezena será o **ramo** enquanto a unidade será a **folha.**

No caso do número 61, o ramo é 6 e sua folha é 1; para 87, 8 é o ramo e 7 é a folha; e para 70, o ramo é 7 e a folha é 0.

Assim, devemos criar uma primeira divisão vertical indicando todos os ramos possíveis:

3|
4|
5|
6|
7|
8|
9|

Na sequência, para cada dado, adicionamos as folhas em seus respectivos ramos:

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

Note que essa representação é um agrupamento de dados em classes, visto que os dados ficaram separados de 30 a 40; 40 a 50 e assim por diante, mas é uma estratégia que facilita para uma visualização mais rápida.

5.2 Método de Ordenação de Dados

Note que a recuperação dos dados a partir do diagrama também é realizada de forma rápida.

Podemos ordenar as folhas do diagrama construído e obter os dados rapidamente:

32,43,51,54,59,61,63,65,68,69, 70,71,73,75,76,80,82,87,93

FINALIZANDO

Com esta aula, conseguimos realizar a maior parte das análises que podemos utilizar para realizar uma representação gráfica dos dados. A distribuição de frequência permite visualizar o tipo de dados que estamos verificando em determinado experimento.

REFERÊNCIAS

CASTANHEIRA, N. P. **Estatística aplicada a todos os níveis.** Curitiba: InterSaberes, 2012.

CASTANHEIRA, N. P. **Métodos Quantitativos.** Curitiba: InterSaberes, 2013.

DOWNING, D.; CLARK, J.; **Estatística aplicada.** 3. ed. São Paulo: Saraiva, 2010.

FREUND, J. E.; **Estatística aplicada.** 11. ed. Porto Alegre: Bookman, 2007.

LARSON, R.; FARBER, B.; **Estatística aplicada.** 6. ed. São Paulo: Pearson, Education do Brasil, 2015.

MONTGOMERY, D. C.; RUNGER, G. C.; **Estatística aplicada e probabilidade para engenheiros.** 5. ed. Rio de Janeiro: LTC, 2012.

MONTGOMERY, D. C.; RUNGER, G. C.; HUBELE, N. F.; **Estatística aplicada à engenharia.** 2. ed. Rio de Janeiro: LTC, 2013.

SIQUEIRA, J. O. **Fundamentos de Métodos Quantitativos.** São Paulo: Saraiva, 2011.