ANÁLISE EXPLORATÓRIA DE DADOS

AULA 2

Prof. Dayane Perez Bravo

CONVERSA INICIAL

Ao realizar uma análise exploratória de dados, diversas vezes necessitamos de uma representação gráfica que nos permita verificar algum padrão ou algum resultado a partir de um conjunto de dados. Na aula anterior, verificamos que a construção das tabelas de frequência permite tirar conclusão de dados que, a princípio, estão desordenados. Nesta aula, iremos verificar o uso de alguns gráficos e como construí-los de maneira adequada.

TEMA 1 – GRÁFICO DE BARRAS

Vejamos como construir os gráficos de barras para dois tipos de dados: qualitativos ou discretos.

1.1 Gráfico de barras para dados qualitativos

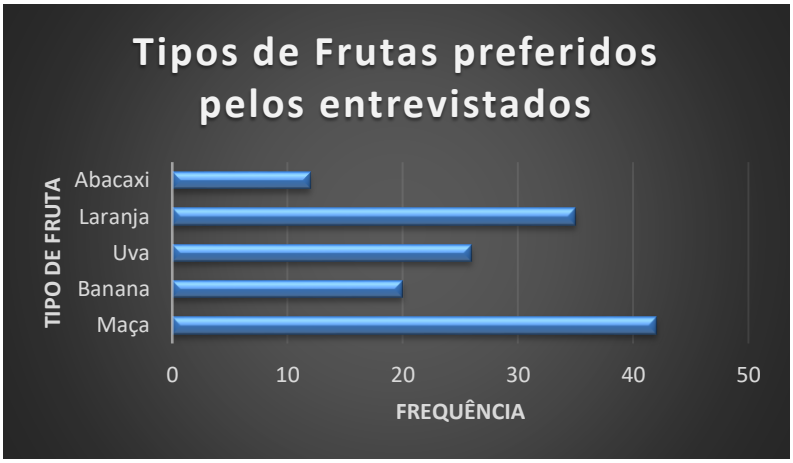
Para construir o gráfico de barras para dados qualitativos, iremos partir da tabela de frequência que construímos na aula anterior. Nesse caso, considere os dados de 135 indivíduos sobre os tipos de frutas preferidas, de acordo com a Tabela 1.

Tabela 1 – Tabela de frequência com tipos de frutas preferidas entre 135 entrevistados

Tipo de Fruta	Frequência
Maça	42
Banana	20
Uva	26
Laranja	35
Abacaxi	12
	135

A Figura 1 apresenta o gráfico de barras para essa tabela, originada, automaticamente, pelo Excel. Para sua construção, utilizamos as opções:

Figura 1 – Gráfico de Barras para os tipos de frutas preferidos pelos entrevistados



Inserir → Gráficos → Ver Todos os Gráficos →Todos os Gráficos → Barras.

1.2 Gráfico de barras para dados discretos

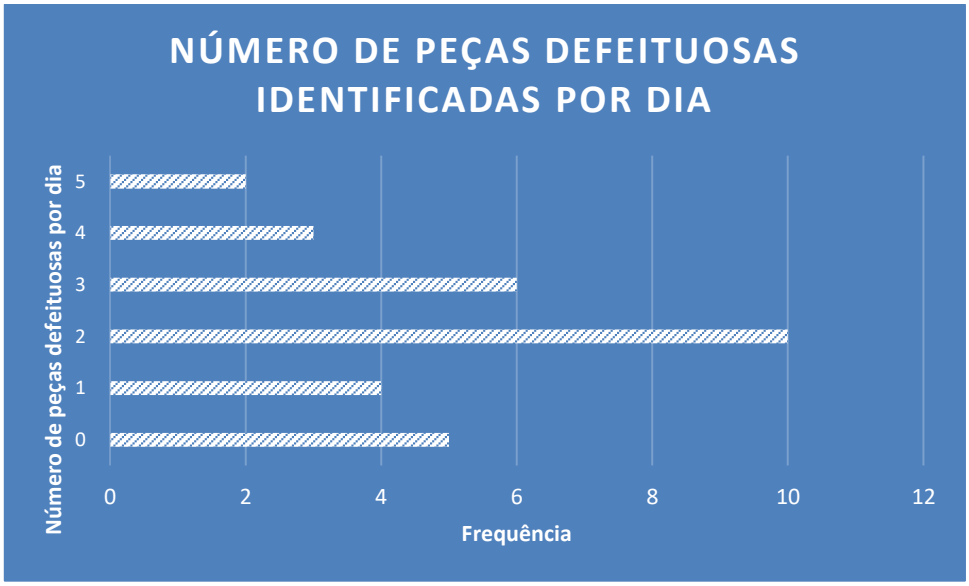
Vejamos, agora, o caso em que o conjunto de dados é discreto. Podemos considerar dados acerca da quantidade de peças defeituosas identificadas em um dia, durante um período de 30 dias. Os resultados se encontram na tabela de frequência 2.

Tabela 2 – Tabela de frequência com número de peças defeituosas por dia coletadas durante 30 dias

N. de peças defeituosas por dia	Frequência
0	5
1	4
2	10
3	6
4	3
5	2
	30

De forma equivalente, construímos a Figura 2 representando o gráfico de barras para os dados discretos.

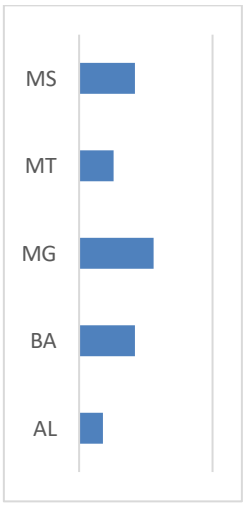
Figura 2 – Gráfico de Barras para o número de peças defeituosas identificadas por dia



1.3 Roteiro para a construção do gráfico de barras

Como percebeu, o software constrói o gráfico de forma praticamente automática. Entretanto, é necessário compreender quais são os critérios utilizados por ele para a construção do gráfico de barras para que se tenha facilidade e entendimento dos critérios adotados na construção do gráfico a partir de outros softwares, principalmente aqueles que exigem determinado nível de programação.

Figura 3 – Criação do eixo vertical



Inicialmente, deve-se construir os dois eixos de forma ortogonal possuindo um tamanho adequado. Note que, se os eixos forem desproporcionais, a visualização pode ficar comprometida.

Para o gráfico de barras, o eixo vertical deve ser dividido para cada uma das variáveis do problema, i.e., 5 para o primeiro caso e 6 para o segundo caso. Note que, ao tratar o problema como uma função, a variável independente está alocada no eixo vertical (ao contrário da representação usual de funções).

Na sequência, deve colocar nas divisões, de baixo para cima, os valores que a variável independente pode assumir. Note que, na Figura 3, foi criado o eixo vertical, atentando-se para a restrição imposta de manter a divisão do eixo vertical em partes iguais.

Na sequência, deve-se criar o eixo horizontal. Note que, para que a representação se torne adequada, é necessário apresentar, de forma visível, todo o intervalo de frequência analisado. No caso do primeiro problema, é essencial que o eixo assuma um valor de pelo menos 42. Entretanto, não é recomendado esse como o valor máximo para ter uma visualização mais adequada. Note que, na Figura 1, escolhemos como o maior valor do eixo 50, enquanto na Figura 2, escolhemos 12. Depois disso, divide-se o eixo horizontal em partes iguais colocando um número correspondente em cada divisão. É comum vermos uma tentativa de colocar, na legenda dos eixos, apenas os dados encontrados. Então, no segundo caso, nosso eixo teria marcações para os números 2, 3, 4, 5, 6 e 10, mas isso não traria uma representação adequada das informações.

Somente depois de criar as divisões dos eixos que podemos criar as barras horizontais. Note que o comprimento corresponde à frequência de cada dado! É importante que as barras tenham a mesma largura e estejam centralizadas em relação à divisão do eixo vertical. Além disso, recomenda-se que exista determinado afastamento entre cada barra, principalmente quando não se pretende utilizar cores.

Lembre-se também que o gráfico deve permitir uma visualização e uma compreensão imediata dos dados do problema. Assim, é relevante que, ao final, você adicione um título para o gráfico, além de legendas para os eixos.

1.4 Exemplo

Considere o caso dos números de veículos vendidos por dia em um período de 20 dias por uma revendedora:

3 5 1 4 5 6 2 3 4 2 5 5 4 5 3 5 4 6 3 5

Podemos construir o gráfico de barras para representar esse problema ao ordenar os dados e construir a tabela de frequência, descrita na Tabela 3. Assim, temos:

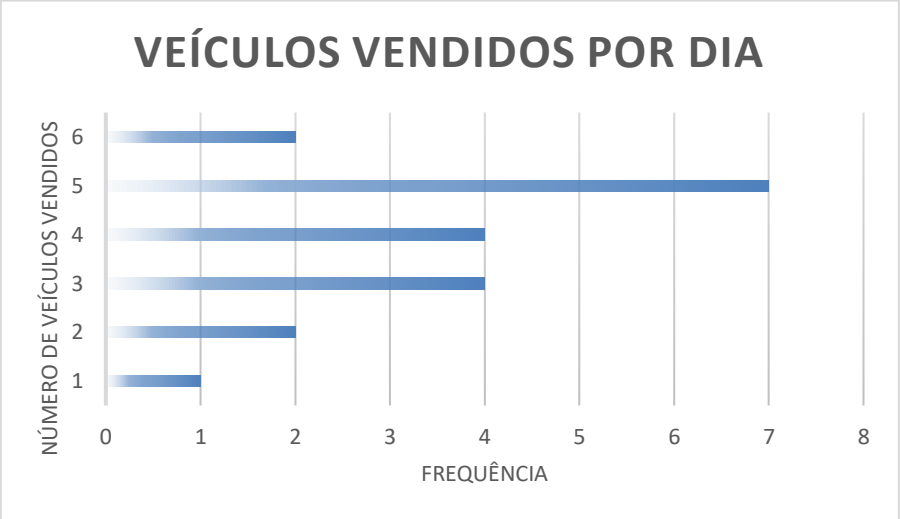
1 2 2 3 3 3 3 4 4 4 4 5 5 5 5 5 5 5 6 6

Tabela 3 – Tabela de frequência com número de veículos vendidos por dia em um período de 20 dias por uma revendedora

N. de veículos vendidos	Frequência
1	1
2	2
3	4
4	4
5	7
6	2
	20

Veja que a representação dessa tabela de frequência a partir de um gráfico de barras é dada pela Figura 4.

Figura 4 – Gráfico de barras com número de veículos vendidos por dia em um período de 20 dias por uma revendedora



Note que verificamos as condições que colocamos ao apresentar o roteiro para a construção do gráfico de barras.

TEMA 2 – GRÁFICO DE COLUNAS

Outro tipo de gráfico comum é o gráfico de colunas. A construção é muito similar ao gráfico de barras, com a diferença de que os valores da variável são colocados no eixo horizontal, enquanto as frequências são colocadas no eixo vertical. Note que, ao tentar representar por um gráfico de colunas os dados qualitativos, como o exemplo da fruta, poderemos ter dificuldade para colocar, de forma adequada, os rótulos na legenda do eixo. Assim, costumamos utilizar esse tipo de gráfico para os dados do tipo discreto.

2.1 Gráfico de Colunas para Dados Discretos

Considere um jogo de boliche com 5 pinos. A tabela de frequência apresentada na Tabela 4 representa a frequência que determinada quantidade de pinos foi derrubada para cada jogada.

Tabela 4 – Tabela de frequência com número de pinos derrubados por jogada para 13 jogadas distintas

N. de pinos derrubados	Frequência
0	2
1	3
2	4
3	2
4	1
5	1
	13

A Figura 5 mostra o gráfico de colunas gerado a partir do software Excel, conforme o roteiro anterior, mas considerando a inversão dos eixos.

Figura 5 – Gráfico de colunas com o número de pinos derrubados em 13 jogadas distintas



2.2 Gráfico de colunas para Dados Qualitativos

Como discutido, geralmente, o gráfico de colunas não é adequado para a representação de dados qualitativos em virtude da representação dos eixos. Entretanto, em alguns casos, essa representação é viável, como no caso a seguir.

Considere uma pesquisa realizada com moradores selecionados aleatoriamente, da cidade de Curitiba, que não nasceram nela. A Tabela 5

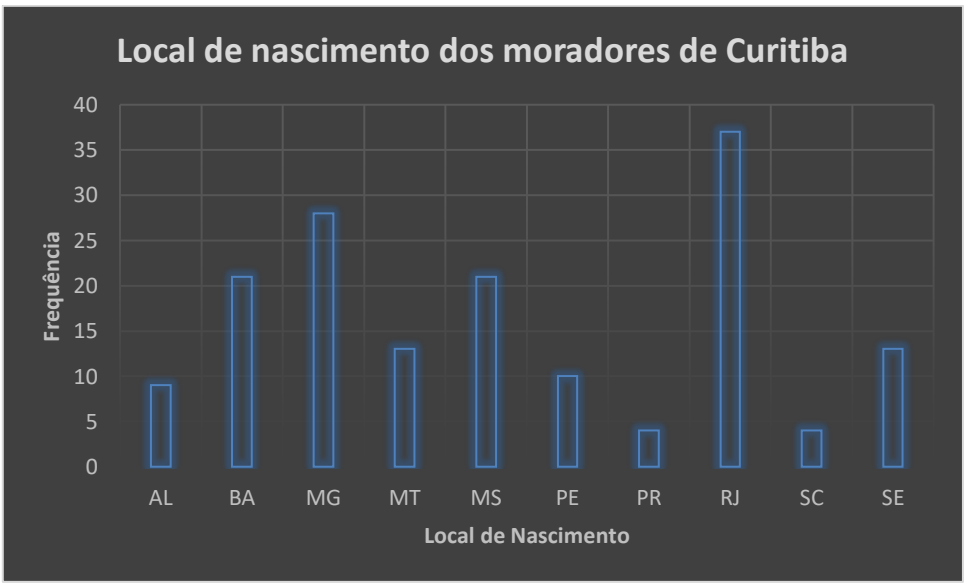
apresenta a localização de cada um desses moradores. No total foram entrevistadas 160 pessoas.

Tabela 5 – Local de nascimento de 160 moradores da cidade de Curitiba que não nasceram lá

Estado	*AL*	*BA*	*MG*	*MT*	*MS*	*PE*	*PR*	*RJ*	*SC*	*SE*	
Frequência	9	21	28	13	21	10	4	37	4	13	160

A Figura 6 apresenta o gráfico de colunas para esse conjunto de dados. Note que, ao utilizar a sigla dos estados, em vez de seus nomes completos, foi possível realizar uma representação adequada nos rótulos do eixo horizontal.

Figura 6 – Gráfico de colunas para o local de nascimento dos moradores de Curitiba não nascidos na cidade



2.2 Gráfico de Múltiplas Colunas

Em alguns casos, gostaríamos de ter ferramentas para comparar duas ou mais séries distintas de dados para uma mesma variável. Nesse caso, é interessante utilizar um gráfico de múltiplas colunas que representa uma pequena variação do gráfico de colunas.

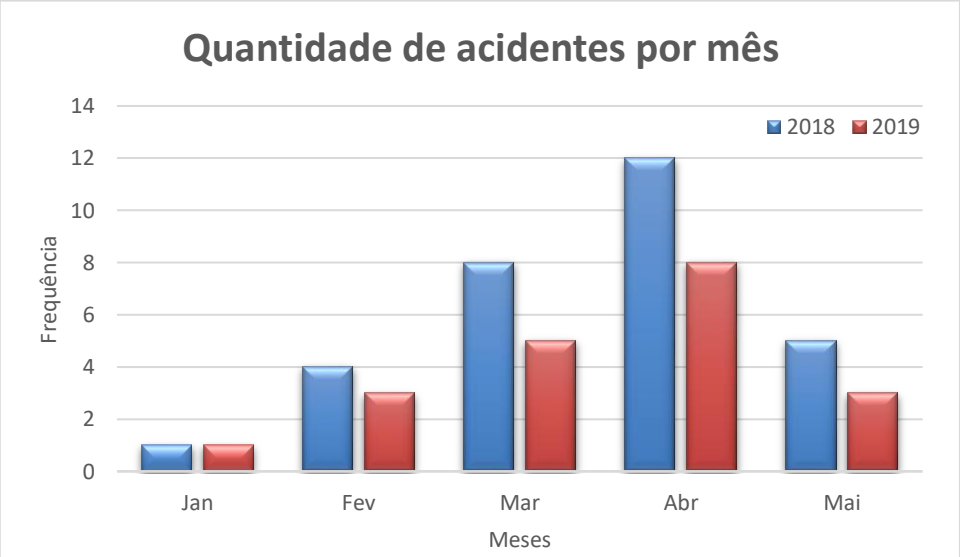
Nesse caso, considere o número de acidentes de trânsito que ocorreram nos cinco primeiros meses dos anos de 2018 e 2019. Os dados estão contidos na Tabela 6.

Tabela 6 – Número de acidentes de trânsito que ocorreram nos cinco primeiros meses dos anos de 2018 e 2019

Mês	Frequência	Mês	Frequência
Jan	1	**Jan**	1
Fev	4	**Fev**	3
Mar	8	**Mar**	5
Abr	12	**Abr**	8
Mai	5	**Mai**	3
	30		20

Note que a quantidade de acidentes no ano de 2018 foi 30, enquanto no ano de 2019 foi 20. Esses dados podem ser colocados na forma de um gráfico de colunas, conforme apresentado na Figura 7. Note que colocamos legendas para cada um dos anos e os diferenciamos utilizando cores diferentes.

Figura 7 – Comparação entre a quantidade de acidentes de trânsito para os 5 primeiros meses do ano de 2018 e 2019



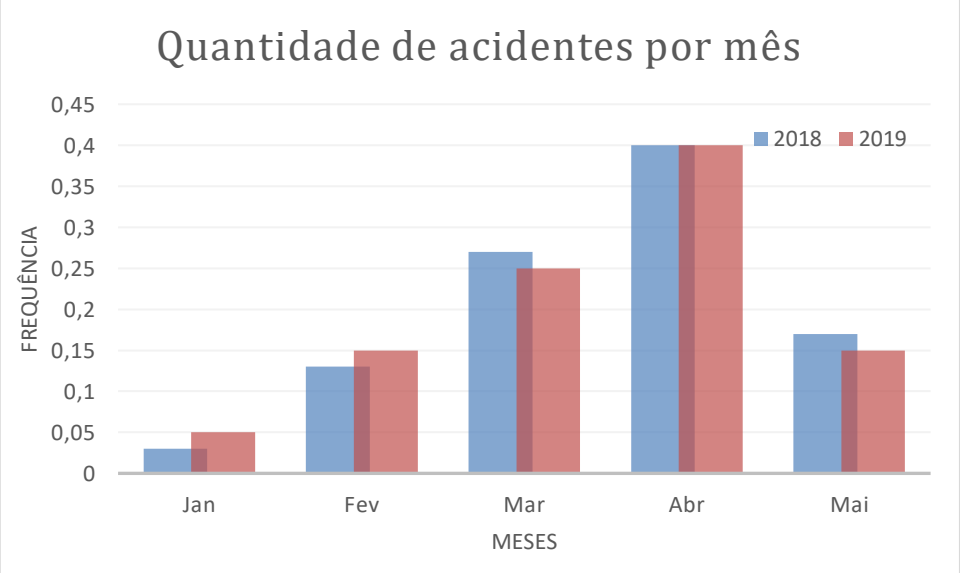
Em alguns casos, você quer ter uma alternativa para comparar quão pior ou melhor foi cada mês do ano seguinte em relação ao anterior. Dessa forma, é necessário utilizar um gráfico de múltiplas colunas considerando a frequência relativa ao invés da frequência absoluta. Assim, ampliamos a Tabela 6, obtendo a Tabela 7 a seguir.

Tabela 7 – Número relativo de acidentes de trânsito que ocorreram nos cinco primeiros meses dos anos de 2018 e 2019

Mês	Frequência	Frequência Relativa	Mês	Frequência	Frequência Relativa
Jan	1	0,03	**Jan**	1	0,05
Fev	4	0,13	**Fev**	3	0,15
Mar	8	0,27	**Mar**	5	0,25
Abr	12	0,40	**Abr**	8	0,40
Mai	5	0,17	**Mai**	3	0,15
	30			20	

Assim, geramos a Figura 8 realizando a mesma comparação, entretanto, agora, a partir dos dados normalizados.

Figura 8 – Comparação entre a quantidade relativa de acidentes de trânsito para os 5 primeiros meses do ano de 2018 e 2019



Note que, por exemplo, comparando os dois gráficos, verificamos que o mês de abril não teve uma melhora significativa em relação à melhora obtida pelo ano, assim como em janeiro.

TEMA 3 – GRÁFICO DE SETORES

Outra representação comum de gráficos é o uso de gráficos de setores, também conhecido como *gráfico de pizza*. A principal vantagem do uso desse tipo de gráfico é visualizar a proporção de um dado em relação ao todo.

3.1 Construção do gráfico de setores

Considere, a título de análise, uma pesquisa realizada com 60 entrevistados considerando os estados nordestinos para viagem preferidos. Você pode conferir o resultado dessa pesquisa na Tabela 8.

Tabela 8 – Tabela de frequência sobre a preferência de 60 entrevistados sobre os estados nordestinos para viagem preferidos

Estados	*Frequência*
Maranhão (MA)	3
Piauí (PI)	5
Ceará (CE)	33
Rio Grande do Norte (RN)	7
Pernambuco (PE)	12
	60

Para a construção do gráfico de setores, construímos a Tabela 9 apresentando a tabela de frequência relativa e a porcentagem em relação ao todo de cada um desses dados:

Tabela 9 – Tabela de frequência relativa e porcentagem sobre a preferência de 60 entrevistados sobre os estados nordestinos para viagem preferidos

Estados	*Frequência*	*Freq. Relativa*	*Porcentagem (%)*
Maranhão (MA)	3	0,55	55
Piauí (PI)	5	0,20	20
Ceará (CE)	33	0,08	8
Rio Grande do Norte (RN)	7	0,05	5
Pernambuco (PE)	12	0,12	12
	60	1	100

Note que, para a construção do gráfico de setores, cada um dos dados ocupará uma região do círculo. Para encontrar o ângulo de abertura necessário para a representação de cada dado, basta lembrar que o todo é equivalente a uma abertura de $360°$. Assim, ampliamos a tabela 9, obtendo a Tabela 10.

Tabela 10 – Cálculo do ângulo do setor para cada dado

Estados	*F*	*FR*	*%*	*Ângulo do setor*
Maranhão (MA)	3	0,55	55	$0{,}55.360° = 198°$
Piauí (PI)	5	0,20	20	$0{,}20.360° = 72°$
Ceará (CE)	33	0,08	8	$0{,}08.360° = 29°$
Rio Grande do Norte (RN)	7	0,05	5	$0{,}05.360° = 18°$
Pernambuco (PE)	12	0,12	12	$0{,}12.360° = 43°$
	60	1	100	$360°$

A Figura 9 representa o gráfico de setores para o caso analisado. Note como é fácil notar a relação entre as proporções de cada estado selecionado.

Figura 9 – Gráfico de setores para a preferência dos estados para turismo no Nordeste



3.2 Exemplo

Considere os dados obtidos na Tabela 11 que representam as cores de meias preferidas pelos consumidores de uma determinada loja.

Tabela 11 – Preferência de cores de meias entre 80 clientes entrevistados

Cores	Frequência
Azul	7
Bege	14
Branca	45
Marrom	4
Outras	9
	80

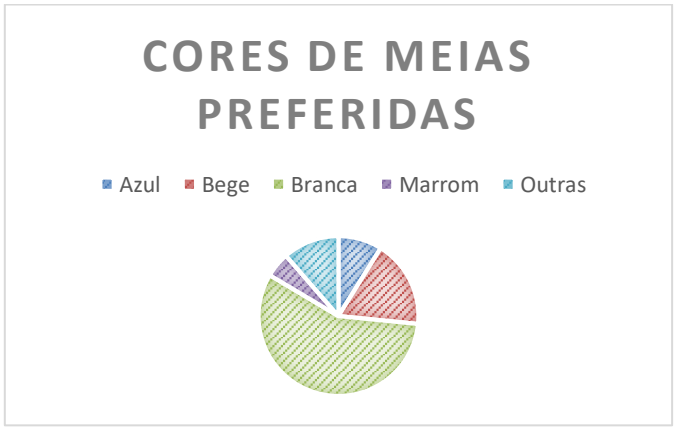
Note que tivemos um total de 80 clientes entrevistados. Na sequência, para a construção do gráfico de setores, construímos a Tabela 12 apresentando as frequências relativas, a porcentagem e o ângulo do setor.

Tabela 12 – Frequência relativa, porcentagem e ângulo do setor calculados para os dados da Tabela 11

Cores	F	FR	%	Ângulo
Azul	7	0,0875	8,75	31,5°
Bege	14	0,175	17,5	63°
Branca	45	0,5625	56,25	202,5°
Marrom	4	0,05	5	18°
Outras	9	0,1125	11,25	40,5°
	80	1	100	360°

Assim, construímos o gráfico de setores apresentado na Figura 10.

Figura 10 – Gráfico de setores apresentando as cores de meias preferidas entre 80 consumidores entrevistados



TEMA 4 – OUTROS GRÁFICOS

Existem diversas possibilidades de gráficos utilizados para a representação de um conjunto de dados. Nesta seção, veremos um, entre esses tipos, que costuma ser utilizado em algumas análises estatísticas. É o gráfico de frequência acumulada.

4.1 Gráfico de Frequência Acumulada

Considere os dados acerca dos números de gols obtidos por partida para uma série de 20 observações. Esses dados estão descritos na Tabela 13. Note que adicionamos algumas linhas para representar a frequência acumulada, a frequência acumulada relativa e a porcentagem acumulada.

Tabela 13 – Frequência, frequência acumulada, frequência acumulada relativa e porcentagem acumulada para o número de gols em 20 partidas observadas

Gols	0	1	2	3	4	5	
Frequência	4	8	3	3	1	1	20
Frequência Acumulada	4	12	15	18	19	20	
Frequência Acumulada Relativa	0,20	0,60	0,75	0,90	0,95	1,00	
Porcentagem Acumulada	20	60	75	90	95	100	

Note que as Figuras 11 e 12 apresentam os gráficos de frequência e porcentagem acumuladas, respectivamente. Elas apresentam resultados similares, com a diferença dos valores adotados no eixo vertical. É possível verificar que os dados vão se aproximando da totalidade e como eles estão distribuídos. Esse tipo de gráfico permite tirar outras interpretações que serão abordadas em análises posteriores.

Figura 11 – Gráfico de frequência acumulada para o número de gols em 20 partidas observadas

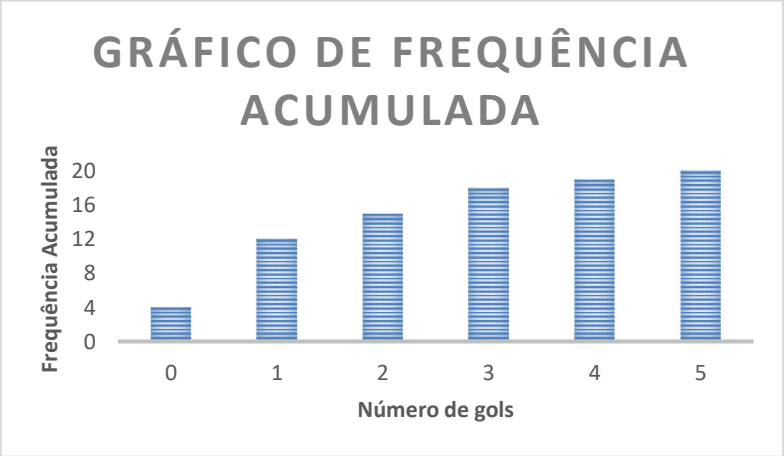
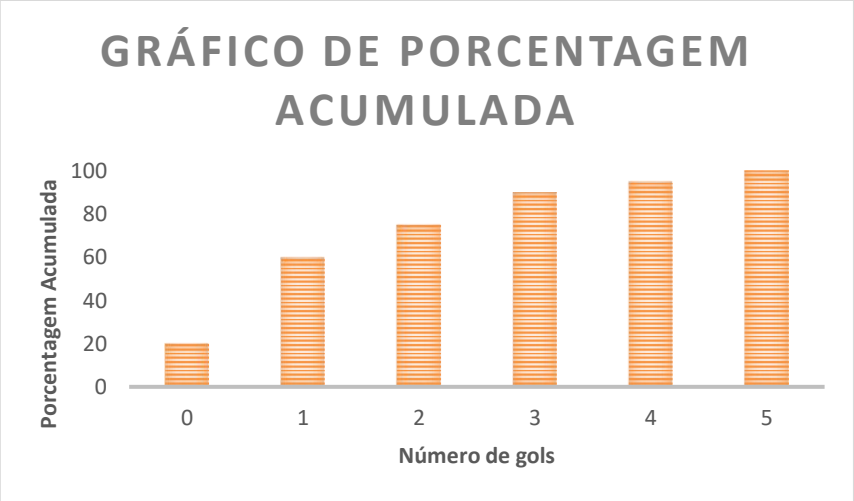


Figura 12 – Gráfico de porcentagem acumulada para o número de gols em 20 partidas observadas



4.2 Exemplo

Considere um caso em que o número de habitantes por residência em determinado condomínio foi coletado e o resultado foi dado por:

4 4 1 2 4 5 2 1 4 4 5 2 3 4 4

2 4 1 4 5 3 3 3 2 2 4 4 4 3 2

4 3 2 2 4 5 4 1 3 4 1 5 2 4 4

Ao ordenar e realizar a contagem dos dados, geramos a tabela de frequência apresentada na Tabela 14.

Tabela 14 – Frequência do número de habitantes por residência em determinado condomínio

Habitantes	**1**	**2**	**3**	**4**	**5**	
Frequência	5	10	7	20	6	48

Assim, ampliamos a Tabela 14 gerando a Tabela 15 encontrando a frequência acumulada.

Tabela 15 – Frequência e frequência acumulada do número de habitantes por residência em determinado condomínio

Habitantes	**1**	**2**	**3**	**4**	**5**	
Frequência	5	10	7	20	6	48
Frequência Acumulada	5	15	22	42	48	

E a partir delas, geramos o gráfico de frequência da Figura 13 e da frequência acumulada da Figura 14.

Figura 13 – Gráfico de frequência com o número de habitantes em determinado condomínio residencial

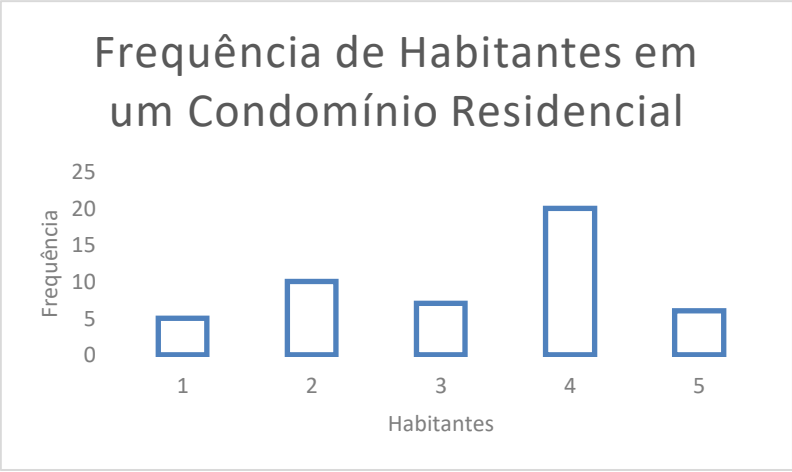
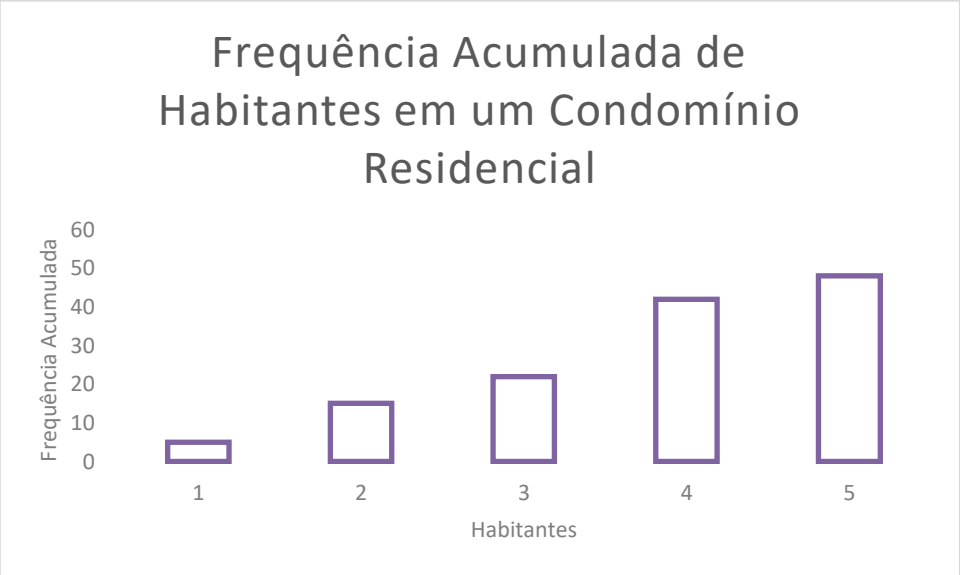


Figura 14 – Gráfico de frequência acumulada com o número de habitantes em determinado condomínio residencial



TEMA 5 – TABELA DE FREQUÊNCIA PARA DADOS CONTÍNUOS

Uma das principais análises de dados é realizada para dados contínuos a partir da construção de sua tabela de frequência. Entretanto, veremos que precisamos fazer algumas adaptações para construir essa tabela de forma adequada.

5.1 Agrupamento de Dados em Classes

Considere o caso de dados coletados para 16 corredores sobre o tempo que levaram para completar determinado circuito. Os dados coletados foram:

$$43{,}1\ 41{,}4\ 43{,}7\ 42{,}0\ 43{,}9\ 42{,}3\ 43{,}2\ 42{,}6$$
$$41{,}7\ 44{,}6\ 43{,}0\ 42{,}1\ 43{,}2\ 40{,}3\ 43{,}5\ 42{,}5$$

Antes de analisarmos os dados, devemos colocá-los de forma ordenada. Assim, obtemos:

$$40{,}3\ 41{,}4\ 41{,}7\ 42{,}0\ 42{,}1\ 42{,}3\ 42{,}5\ 42{,}6$$
$$43{,}0\ 43{,}1\ 43{,}2\ 43{,}2\ 43{,}5\ 43{,}7\ 43{,}9\ 44{,}6$$

Note, inicialmente, que dificilmente nenhum dos dados ocorre mais de uma vez (com exceção do $43{,}2$). Isso se deve à característica dos dados contínuos que podem assumir uma infinidade de valores dentro de determinado intervalo. Assim, se tentássemos criar uma tabela de frequência diretamente a

partir dos dados coletados, encontraríamos 1 (ou 2) para cada um dos dados e isso não nos traria nenhuma interpretação interessante.

Para contornar essa dificuldade, tratamos os dados contínuos selecionando intervalos de valores da variável analisada. Para o caso analisado, podemos, por exemplo, separar em 5 intervalos:

Intervalo 1: de 40,0 a 40,99
Intervalo 2: de 41,0 a 41,99
Intervalo 3: de 42,0 a 42,99
Intervalo 4: de 43,0 a 43,99
Intervalo 5: de 44,0 a 44,99

Note que esses cinco intervalos incluem todos os dados observados, visto que o maior valor encontrado foi $44{,}6$ enquanto o menor valor foi de $40{,}3$. Em estatística, chamamos cada um desses intervalos de *classe* e a frequência que os dados pertencem a uma determinada classe é chamada de *frequência de classe.* Em alguns casos, simplesmente a chamamos de *frequência*, gerando algumas dificuldades de compreensão ao diferenciar o conceito de frequência e frequência de classe. Além disso, utilizamos uma notação específica para especificar o limite inferior e o limite superior de cada classe:

$$(\text{Limite Inferior}) \vdash (\text{Limite Superior})$$

Veja que cada classe possui, a partir da segunda, limite inferior igual ao limite superior da classe anterior. Assim, vale ressaltar que essa notação é equivalente a dizer que:

$$(\text{Limite Inferior}) \leq x < (\text{Limite Superior})$$

A esse processo de organização dos dados, damos o nome de *agrupamento de dados em classes*. Gerando as classes, podemos, finalmente, gerar a tabela de frequência como a apresentada na Tabela 16.

Tabela 16 – 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 a construção dessa tabela de frequência é gerada pelo mesmo processo de contagem, com o cuidado de verificar a quantidade de dados que pertencem a uma determinada classe.

5.2 Número de Classes

Existem vários processos para determinar a quantidade ideal de classes para representar determinado conjunto de dados. Geralmente, o número de classes é arbitrário, mas podem existir estratégias que facilitam a determinação desse número, as quais serão discutidas na aula seguinte. Inicialmente, sugere-se utilizar uma quantidade entre 5 e 20 classes.

5.3 Largura das Classes

Como cada classe é definida pelo seu limite inferior e pelo seu limite superior, definimos:

$$\text{Largura da Classe} = \text{Limite Superior} - \text{Limite Inferior}$$

Geralmente, utilizamos classes de mesma largura, salvo algumas exceções quando os intervalos são muito extensos, mas os dados costumam estar concentrados em uma determinada região do domínio.

No caso em que a largura é igual, determinamos a largura da classe a partir de:

$$\text{Largura da Classe} = \frac{\text{Máximo} - \text{Mínimo}}{\text{Número de Classes}}$$

Em que o máximo e o mínimo se referem aos valores da série de dados.

Recomenda-se, também, não utilizar essa expressão ao pé da letra e utilizar dados razoáveis para determinar o tamanho da classe. Note que, ao usar a expressão no problema inicial, obteríamos:

$$\text{Largura da Classe} = \frac{43{,}5 - 39{,}0}{5} = 0{,}9$$

Mas utilizamos a largura da classe como $1{,}0$ para tornar a representação mais compreensível.

5.4 Ponto Inicial

Definimos o ponto inicial como o menor valor que os dados podem assumir. No caso do problema analisado, seria o $39{,}0$.

5.5 Frequência Relativa

De forma similar ao tratamento realizado aos dados qualitativos ou aos dados discretos, podemos definir a frequência relativa da classe realizando a proporção entre a quantidade de dados pertencente a classe pelo total de dados. Assim, geramos a Tabela 17 ampliando a Tabela 16.

Tabela 17 – Tabela de frequência da classe para os tempos de 16 corredores

Classe	Frequência	Frequência Relativa
$40{,}0 \vdash 41{,}0$	1	0,06
$41{,}0 \vdash 42{,}0$	2	0,13
$42{,}0 \vdash 43{,}0$	5	0,31
$43{,}0 \vdash 44{,}0$	7	0,44
$44{,}0 \vdash 45{,}0$	1	0,06
	16	1

FINALIZANDO

Com essas análises, aprendemos como realizar o tratamento e a visualização de um conjunto de dados a partir do uso de alguns gráficos. Você deve se atentar ao cuidado que precisa ter para que sua visualização fique adequada e represente, adequadamente, o experimento que está analisando.

Também aprendemos como iniciar a análise de dados contínuos a partir do conceito de classes. Esses conceitos serão amplamente utilizados nas próximas aulas.

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.