Estatística Descritiva

Prof. Elvis Stancanelli

Universidade Federal do Ceará
Campus de Quixadá

13 de abril de 2024

Conteúdo

No capítulo anterior vimos que há muitas maneiras de coletar dados. Agora aprenderemos que também há muitas maneiras de organizar e descrever conjuntos de dados. A Estatística Descritiva visa facilitar a compreensão dos dados e identificar medidas de tendências. Características importantes a serem observadas ao organizar e descrever um conjunto de dados são seu centro, sua variabilidade (ou dispersão) e sua forma.

1 Distribuição de frequências

Quando um conjunto de dados possui muitas entradas, pode ser difícil extrair padrões. Nesta seção, veremos como organizar os conjuntos de dados em intervalos chamados classes e formar uma distribuição de frequência.

Distribuição de frequência. Uma distribuição de frequência é uma tabela que mostra classes ou intervalos de entradas de dados com uma contagem do número de entradas dentro de cada intervalo, medida esta denominada frequência. Teremos portanto uma frequência f_k para a k-ésima classe de um total de K classes. Perceba que os intervalos não se sobrepõem; os intervalos são contíguos. Ademais, é adequado que todas as classes possuam intervalos com a mesma largura.

A seguir, o passo-a-passo para construir uma distribuição de frequência a partir de um conjunto de dados:

1. Decida qual o número de classes, K, a considerar. Tipicamente o número de classes está entre 5 e 20; caso contrário, pode ser difícil detectar quaisquer padrões. Há fórmulas empíricas que poderiam ser adotadas aqui, porém, trataremos isso tão somente como uma abordagem alternativa;

2. Encontre a largura da classe da seguinte maneira. Determine a variação dos dados e divida pelo número de classes e arredonde para cima: $\left\lceil \frac{max - min + 1}{K} \right\rceil$, em que min é o menor dos valores das entradas e max o maior;

3. Encontre os limites das classes. Você pode adotar a menor entrada dos dados, min, como o limite inferior da primeira classe. Os limites inferiores das classes seguintes são obtidos incrementando-se a largura da classe. Em seguida, encontre os limites superiores de todas as classes. Lembre-se de que as classes não podem se sobrepor;

4. Para cada classe, conte o número de entradas de dados contidas no intervalo.

Exemplo 1. O conjunto de dados a seguir lista os preços (em dólares) em uma amostra de 30 navegadores portáteis de sistema de posicionamento global (GPS). Construa uma distribuição de frequência com 7 classes.

90	130	400	200	350
70	325	250	150	250
275	270	150	130	59
200	160	450	300	130
220	100	200	400	200
250	95	180	170	150

No Exemplo 1 já foi imposto que o número de classes é $K = 7$. Desse modo, podemos obter a largura das classes de acordo com o passo descrito logo acima. A variação dos dados é obtida ao se detectar as entradas extremas, i.e, 59 é a menor das entradas enquanto 450 é a maior. A partir da variação podemos obter o número de valores compreendidos neste intervalo total. No caso, 392. Assim, a largura das classes será dada por $\left\lceil \frac{392}{7} \right\rceil = 56$.

Já podemos dizer que o limite inferior da primeira classe é 59, que é a entrada mínima, e a partir daí podemos obter todos os demais limites inferiores, simplesmente acrescentando a largura de classe, 56, à medida que avançamos entre as classes. Ou seja, os limites inferiores para a segunda classe será 115, para a terceira 171, para a quarta 227, para a quinta 283, para a sexta 339 e para a sétima, e última, o limite inferior é 395.

Também podemos encontrar os limites superiores. Para a primeira classe, sabemos que o intervalo inicia em 59 e finaliza em 114 de modo que compreenda 56 números inteiros. Subsequentemente, temos o limite superior 170 para segunda classe, 226 para terceira, 282 para a quarta, 338 para a quinta, 394 para a sexta e 450 para sétima. Perceba que nenhum intervalo se sobrepõe e que cada um deles termina imediatamente antes de o próximo intervalo iniciar. Todos esses intervalos juntos compreendem todas as entradas, que vão de 59 a 450.

Agora olhamos para cada uma das classes, e contabilizamos quantas entradas dos dados estarão compreendidas no intervalo considerado. Na primeira classe, temos o intervalo que vai de 59 a 114, incluindo os extremos. Ao verificar os dados apresentados no Exemplo 1, observamos 5 entradas compreen-

didas neste intervalo, ou seja, a frequência da primeira classe é 5. Repetimos esse procedimento para cada uma das demais classes, obtendo assim as seguintes frequências: 8 para a segunda classe; 6 para a terceira classe; 5 para a quarta classe; 2 para a quinta classe; 1 para a sexta classe e 3 para a sétima classe.

Note que, assim como unir todos intervalos faz com que todas as entradas sejam compreendidas, somar todas as frequências leva ao número de entradas total.

Agora temos todos os elementos necessários para construir a tabela de distribuição de frequências para o Exemplo 1.

classe	frequência
$59 - 114$	5
$115 - 170$	8
$171 - 226$	6
$227 - 282$	5
$283 - 338$	2
$339 - 394$	1
$395 - 450$	3
	$\sum_c f_c = 30$

Informações auxiliares podem ser disponibilizadas, como: ponto médio, frequência relativa e frequência acumulada.

Ponto médio é simplesmente o ponto central do intervalo. Trata-se de uma maneira alternativa de caracterizar as classes, mas isso acompanhado da largura da classe. Ou seja, o ponto médio da classe k será dado por

$$\frac{\text{Limite inferior}_k + \text{Limite superior}_k}{2}.$$

Frequência relativa é a razão de cada uma das frequências com o número total de ocorrências do conjunto de dados. Trata-se da porção dos dados que se encontram na classe de interesse.

Frequência acumulada de uma classe é a soma das frequências dessa classe e todas as classes anteriores. A frequência acumulada da última classe será igual ao número total de entradas no conjunto de dados.

De posse dessas informações auxiliares, podemos construir a tabela da distribuição de frequência expandida. Retomando o Exemplo 1, teremos:

classe	frequência	ponto médio	frequência relativa	frequência acumulada
59 – 114	5	86,5	0,167	5
115 – 170	8	142,5	0,267	13
171 – 226	6	198,5	0,200	19
227 – 282	5	254,5	0,167	24
283 – 338	2	310,5	0,067	26
339 – 394	1	366,5	0,033	27
395 – 450	3	422,5	0,100	30
	$\sum_c f_c = 30$		$\sum_c fr_c = 1$	

1.1 Histograma

Às vezes é mais fácil identificar padrões de um conjunto de dados observando um gráfico da distribuição de frequência. Um desses gráficos é o histograma de frequências.

Um histograma de frequência é um gráfico de barras que representa a distribuição de frequência para um conjunto de dados. Um histograma tem as seguintes propriedades:

- A escala horizontal é quantitativa e mede os valores dos dados;
- A escala vertical mede as frequências das classes;
- Barras consecutivas se tocam.

Como as barras consecutivas de um histograma devem se tocar, as barras devem começar e terminar nas fronteiras da classe ao invés dos limites da classe. Fronteiras de classe são os números que separam as classes sem deixar brechas entre elas. Se as entradas de dados forem números inteiros, subtraia 0,5 de cada limite inferior para encontrar as fronteiras inferiores da classe. Para encontrar as fronteiras superiores da classe, adicione 0,5 a cada limite superior. A fronteira superior de uma classe será igual à fronteira inferior da próxima classe. Ainda referindo-se ao Exemplo 1, poderíamos agregar duas novas colunas com os valores das fronteiras:

Agora, focando-se nas colunas *fronteira inferior*, *fronteira superior* e *frequência* podemos construir o primeiro histograma.

classe	frequência	ponto médio	fronteira inferior	fronteira superior	freq. relat.	freq. acumul.
59 – 114	5	86,5	58,5	114,5	0,167	5
115 – 170	8	142,5	114,5	170,5	0,267	13
171 – 226	6	198,5	170,5	226,5	0,200	19
227 – 282	5	254,5	226,5	282,5	0,167	24
283 – 338	2	310,5	282,5	338,5	0,067	26
339 – 394	1	366,5	338,5	394,5	0,033	27
395 – 450	3	422,5	394,5	450,5	0,100	30
	$\sum f = 30$		$\sum fr = 1$			

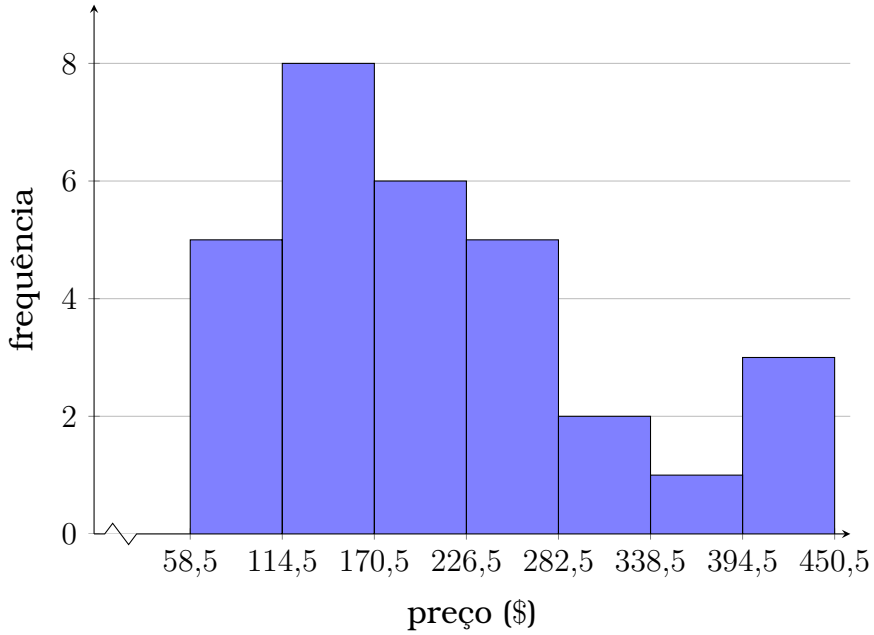

Figura 1: Histograma para os dados do Exemplo 1 construído com base nas fronteiras das classes.

Sabendo-se que os intervalos possuem largura de $\$56$, as barras também terão essa mesma largura. Assim, alternativamente, podemos construir o histograma com base nos pontos médios.

De maneira similar, podemos construir o histograma com base nas *frequências relativas* em lugar da coluna *frequência*, como ilustrado a seguir:

Lidando com números reais

Perceba que a distribuição de frequência e o histograma foram explicados com base em números inteiros. De fato a explicação não se aplica diretamente a todos os números reais, mas é possível fazer algumas adaptações nos dados para que possamos lidar com um domínio mais amplo. Para ilustrar isso, consideremos o exemplo a seguir:

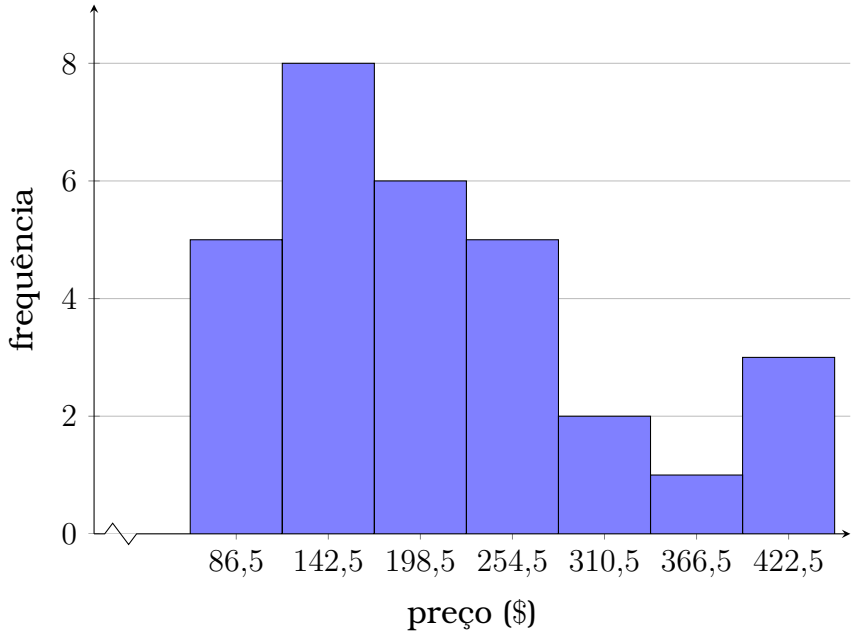

Figura 2: Histograma para os dados do Exemplo 1 construído com base nos pontos médios das classes.

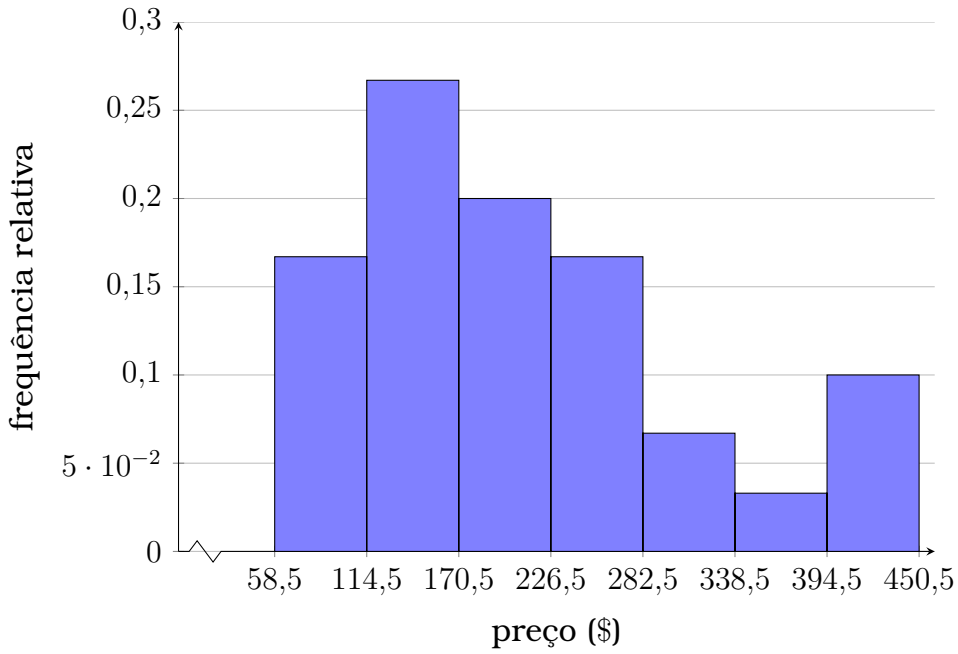

Figura 3: Histograma de frequências relativas para os dados do Exemplo 1 construído com base nas fronteiras das classes.

Exemplo 2. A longevidade (em anos) para 40 baterias de *no-breaks* está registrada nos dados abaixo. Construa uma distribuição de frequência com 7 classes.

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

Se mudarmos a unidade de medida para décimos de anos, os valores deveriam agora ser representados como:

22	34	25	33	47	41	16	43
31	38	35	31	34	37	32	45
33	36	44	26	32	38	29	32
39	37	31	33	41	30	30	47
39	19	42	26	37	31	34	35

Note que a mudança na unidade não altera qualquer medida e nem mesmo a exatidão dessas medições. Simplesmente 2,5 anos será representado como 25 décimos de anos.

Agora facilmente poderemos obter a distribuição de frequências. A entrada de maior valor é 47 e a de menor valor 16. Portanto, a largura de cada uma das sete classes será $\lceil \frac{32}{7} \rceil = 5$. Assim, já podemos definir as classes: a classe 1 vai de 16 a 20; a classe 2, de 21 a 25; a classe 3, de 26 a 30; a classe 4, de 31 a 35; a classe 5, de 36 a 40; a classe 6, de 41 a 45; e a classe 7, de 46 a 47.

Agora facilmente poderemos retornar à unidade de medida original, no caso, ano. A classe 1 vai de $1,6$ a $2,0$; a classe 2, de $2,1$ a $2,5$; a classe 3, de $2,6$ a $3,0$; a classe 4, de $3,1$ a $3,5$; a classe 5, de $3,6$ a $4,0$; a classe 6, de $4,1$ a $4,5$; e a classe 7, de $4,6$ a $4,7$. Daqui para frente, podemos proceder como já sabemos, obtendo a frequência para cada uma das classes. A tabela de distribuição é mostrada a seguir:

classe	frequência
1,6 – 2,0	2
2,1 – 2,5	2
2,6 – 3,0	5
3,1 – 3,5	15
3,6 – 4,0	8
4,1 – 4,5	6
4,6 – 5,0	2
	$\sum_c f_c = 40$

1.2 Ogiva

Se você quer descrever o número de entradas de dados que são iguais ou inferiores a um determinado valor, você pode facilmente fazê-lo construindo um gráfico poligonal de frequências acumuladas ou também conhecido como gráfico de ogiva.

Um gráfico de ogiva é construído com segmentos de reta interconectando pares coordenadas das frequência acumuladas de classes subsequentes em suas fronteiras superiores. As fronteiras superiores são marcadas no eixo horizontal e as frequências acumuladas são marcadas no eixo vertical.

Podemos descrever o passo-a-passo de como se constrói um gráfico de frequências acumuladas.

1. Construa uma distribuição de frequência que inclua frequências acumuladas em uma das colunas;

2. Especifique as escalas horizontal e vertical. A escala horizontal consiste em fronteiras superiores das classes e a escala vertical mede as frequências acumuladas;

3. Disponha no plano cartesiano os pontos que representam fronteira superior e frequência acumulada para cada uma das classes;

4. Adicione um ponto de partida disposto na fronteira inferior da primeira classe com a frequência acumulada zero;

5. Começando do ponto mais à esquerda conecte-o ao próximo ponto da direita através de um segmento de reta e repita o procedimento até se conectar ao último ponto.

Note que o gráfico começa no ponto de partida (com frequência acumulada nula) e termina no ponto em que tem-se a fronteira superior da última classe (com a frequência máxima que é igual ao tamanho da amostra). Nenhum dos segmentos de reta será decrescente. Retornemos ao Exemplo 1 para ilustrar este conceito. O gráfico de frequências acumuladas está ilustrado abaixo.

Alternativamente, podemos construir o gráfico de frequências relativas acumuladas, tal como abaixo.

2 Outras representações gráficas

Há várias outras formas de representação gráfica de dados. Algumas mais indicadas para dados quantitativos (e.g. diagrama de ramo-e-folha) e outras para dados qualitativos (e.g. gráfico de pizza e gráfico de Pareto).

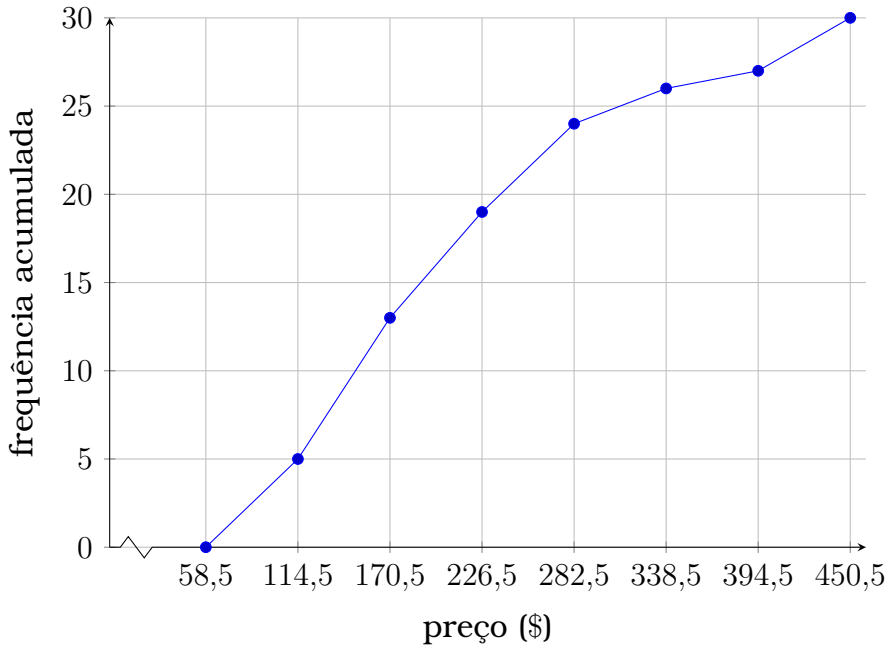

Figura 4: Gráfico de frequências acumuladas para os dados do Exemplo 1.

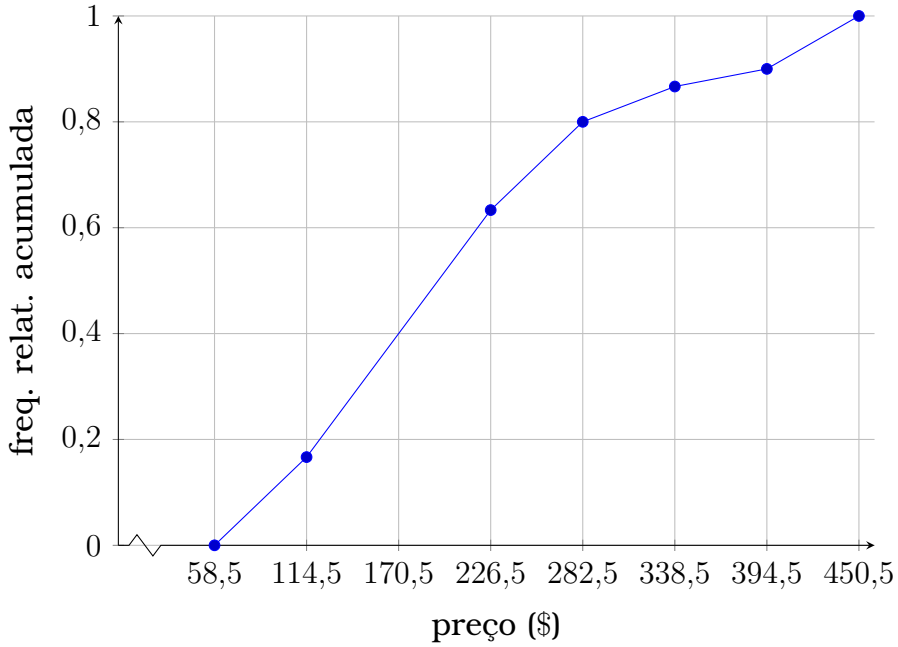

Figura 5: Gráfico de frequências relativas acumuladas para os dados do Exemplo 1.

Há ainda outras questões a se considerar na escolha da forma gráfica. Quando cada entrada em um conjunto de dados corresponde a uma entrada em um segundo conjunto de dados, os conjuntos são ditos serem dados emparelhados. Por exemplo, suponha que um conjunto de dados contenha os custos de um item e um segundo conjunto de dados contenha os valores

de vendas do item a cada custo. Como cada custo corresponde a um valor de vendas, os conjuntos de dados são pareados. Uma maneira de representar graficamente conjuntos de dados emparelhados é usar um gráfico de dispersão, em que os pares ordenados são representados graficamente como pontos em um plano de coordenadas. Um gráfico de dispersão é usado para mostrar a relação entre duas variáveis quantitativas.

Quando entradas quantitativas são tomadas em intervalos regulares de tempo, o conjunto de dados é uma série temporal, a qual poderá ser facilmente representada graficamente, configurando o eixo das abscissas com o tempo discreto.

3 Medidas de tendência central

Agora veremos como extrair características numéricas que descrevem o centro de um conjunto de dados. Uma medida de tendência central é um valor que representa uma entrada típica ou central de um conjunto de dados. As três medidas de tendência central mais usadas são a média, a mediana e a moda.

Média de um conjunto de dados é a soma das entradas de dados dividida pelo número de entradas. Para encontrar a média de um conjunto de dados, usamos uma das seguintes fórmulas.

$$\mu = \frac{\sum x}{N}$$

$$\bar{x} = \frac{\sum x}{n}$$

A letra grega minúscula μ representa a média populacional e $\bar{x}$ representa a média amostral. Observe que N representa o número de entradas em uma população e n representa o número de entradas em uma amostra.

Mediana de um conjunto de dados é o valor que fica no meio dos dados quando o conjunto de dados é ordenado. A mediana mede o centro de um conjunto de dados ordenado dividindo-o em duas partes de tamanhos iguais. Se o conjunto de dados tiver um número ímpar de entradas, a mediana é a entrada de dados do intermediária. Se o conjunto de dados tiver um

número par de entradas, a mediana será a média das duas entradas de dados intermediárias.

Moda de um conjunto de dados é a entrada de dados que ocorre com maior frequência. Um conjunto de dados pode ter uma moda, mais de uma moda ou nenhuma moda. Se nenhuma entrada é repetida, o conjunto de dados não tem moda. Se duas entradas ocorrem com a mesma frequência máxima, cada entrada é uma moda e o conjunto de dados é dito ser bimodal.

Embora a média, a mediana e a moda descrevem uma entrada típica de um conjunto de dados, elas têm particularidades e seus resultados podem ser diferentes. Há vantagens e desvantagens de usar cada uma. A média é uma medida confiável pois leva em conta cada entrada de um conjunto de dados. No entanto, a média pode ser bastante afetada quando o conjunto de dados contém entradas discrepantes.

Entradas discrepantes são aquelas que estão muito afastadas das outras entradas no conjunto de dados. Um conjunto de dados pode ter um ou mais entradas discrepantes, causando lacunas em sua distribuição de frequências. As conclusões tiradas de um conjunto de dados que contém entradas discrepantes podem ser falhas.

Um gráfico revela várias características de uma distribuição de frequências. Uma dessas características é a forma da distribuição.

Dizemos que a distribuição de frequência é simétrica quando uma linha vertical pode ser traçada no meio do gráfico da distribuição e as metades resultantes são aproximadamente imagens espelhadas. Uma distribuição de frequência é dita ser uniforme quando todas as entradas, ou classes, na distribuição têm frequências iguais ou aproximadamente iguais. Uma distribuição uniforme também é simétrica. Uma distribuição de frequência é assimétrica se a "cauda" do gráfico se alonga mais para um lado do que para o outro. Uma distribuição é assimétrica para a esquerda (assimétrica negativa) se sua cauda se estende para a esquerda. Uma distribuição é assimétrica para a direita (assimétrica positivamente) se sua cauda se estende para a direita.

Quando uma distribuição é simétrica e unimodal, a média, a

mediana e a moda são iguais. Se uma distribuição for assimétrica à esquerda, a média é menor que a mediana e a mediana geralmente é menor que a moda. Se uma distribuição for assimétrica à direita, a média é maior que a mediana e a mediana geralmente é maior que a moda.

A média sempre estará posicionada graficamente para o lado em que a cauda esteja estendida. Por exemplo, quando uma distribuição é assimétrica à esquerda, a média está à esquerda da mediana.

4 Medidas de variação

Agora veremos como extrair características numéricas que descrevem a variabilidade de um conjunto de dados. A mais simples dessas medidas é a variação calculada com base na diferença do maior valor de entrada para o menor valor de entrada. Retomando os dados apresentados no Exemplo 1, temos:

90	130	400	200	350
70	325	250	150	250
275	270	150	130	59
200	160	450	300	130
220	100	200	400	200
250	95	180	170	150

O maior valor dentre todas entradas é 450 e o menor é 59. Portanto, a variação será de $450 - 59 + 1 = 392$. Trata-se de uma medida bastante simples de ser obtida porém não tão útil quanto outras,

Outra medida dessa mesma importância é o desvio. O desvio mede o quanto o valor de cada entrada se afasta do valor médio. Portanto, temos um desvio para cada entrada dos dados. Para os dados do Exemplo 1 a média é $\$210,1333$. Desse modo, os desvios correspondentes a cada uma das entradas serão:

$-120,1333$	$-80,1333$	$189,8667$	$-10,1333$	$139,8667$
$-140,1333$	$114,8667$	$39,8667$	$-60,1333$	$39,8667$
$64,8667$	$59,8667$	$-60,1333$	$-80,1333$	$-151,1333$
$-10,1333$	$-50,1333$	$239,8667$	$89,8667$	$-80,1333$
$9,8667$	$-110,1333$	$-10,1333$	$189,8667$	$-10,1333$
$39,8667$	$-115,1333$	$-30,1333$	$-40,1333$	$-60,1333$

Perceba que nesse exemplo há desvios positivos e negativos, indicando valores acima e abaixo da média, respectivamente. Podem ocorrer também desvios nulos indicando valores idênticos à média. Há tantos desvios quanto o tamanho do seu conjunto de dados. Assim, caracterizar o conjunto de dados com a medida resumida dos desvios pode ser interessante. A primeira ideia é calcular a média dos desvios, no entanto, esse valor é sempre nulo. Uma medida alternativa que resume esses desvios é a variância obtida, respectivamente para população e amostra, por:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

A letra grega minúscula σ ao ser elevada ao quadrado representa a variância populacional σ^2, ao passo que s^2 representa a variância amostral. Note que as duas fórmulas são diferentes na essência e não apenas na notação. Isso é necessário para que a variância amostral seja não tendenciosa (enviesada).

Desse modo, no nosso exemplo temos a variância s^2 igual a $\$^2 10378,4644$. Perceba que, embora seja uma medida bastante útil, a variância altera a unidade de medida.

O desvio padrão é outra medida bastante útil e ainda preserva a unidade de medida. Respectivamente para população e amostra o desvio padrão é dado por:

$$\sigma = \sqrt{\sigma^2}$$

$$s = \sqrt{s^2}$$

Assim, o desvio padrão da amostra de preços dos GPS será de $\$101,8747$.

5 Medidas de posição

Se ordenamos o conjunto de dados, mantendo as entradas de menores valores à esquerda e as de maiores à direita, podemos observar valores intermediários que dividem esse conjunto em partes de mesmo tamanho. Quando dividimos o conjunto em duas partes de mesmo tamanho, a mediana é o valor que promove essa divisão. Teremos então duas metades do conjunto de

dados. Se tratarmos dessa mesma forma cada uma das duas metades, teremos um total de quatro quartos do conjunto de dados. O primeiro quartil é o valor que divide o primeiro e o segundo quartos; o segundo quartil, igual à mediana, é o valor que divide o segundo e o terceiro quartos; o terceiro quartil divide o terceiro e o quarto quartos. Ou seja, três quartis dividem o conjunto de dados em quatro partes de mesmo tamanho.

Novamente retornando ao Exemplo 1, o conjunto de dados ordenado será:

59	70	90	95	100	130	130	130	150	150	150	160	170	180	**200**
200	200	200	220	250	250	250	270	275	300	325	350	400	400	450

A mediana é $200, o que separa a primeira:

59	70	90	95	100	130	130	**130**	150	150	150	160	170	180	200

e a segunda metades:

200	200	200	220	250	250	250	**270**	275	300	325	350	400	400	450

Na primeira metade o valor central é $130, portanto, temos também o primeiro quartil. Na segunda metade o valor central é $270, portanto, o terceiro quartil.

Podemos representar graficamente essas informações com o diagrama de caixa-e-bigode. Trata-se de uma caixa que inicia no primeiro quartil e termina no terceiro quartil, apresentando um corte na mediana. Há uma parte do bigode que se estende do início da caixa ao menor valor da entrada dos dados e outra parte do bigode que se estende do final da caixa ao maior valor da entrada dos dados. Desse modo, a metade das entradas mais centrais está representada pela caixa, enquanto a metade das entradas mais afastadas do centro está representada no bigode. Aqui lidaremos apenas dessa forma básica do diagrama, mas vale mencionar que é comum que alguns estudiosos representem individualmente as entradas discrepantes como pontos no plano cartesiano fora da caixa e do bigode.

O diagrama de caixa-e-bigode abaixo foi construído com base nos dados de Exemplo 1.

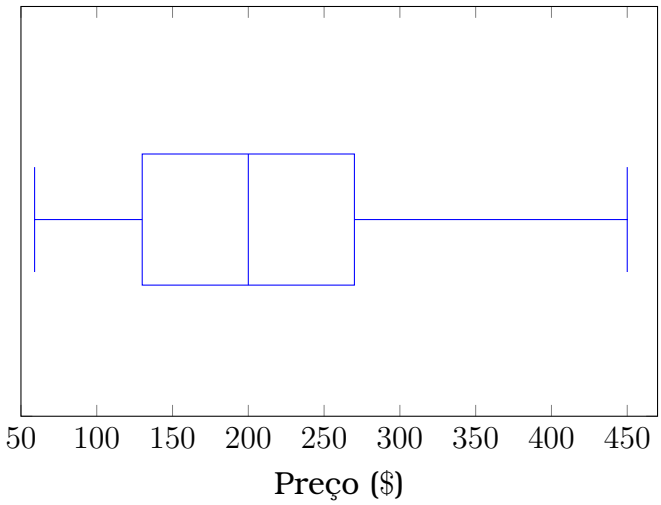

Figura 6: Diagrama de caixa-e-bigode para os dados do Exemplo 1.

6 Referência bibliográfica

1. LARSON, Ron; FARBER, Betsy. Estatística aplicada. 4. ed. São Paulo, SP: Pearson/ Prentice Hall, 2010. xiv,637 p. ISBN 9788576053729