Medidas de Dispersão

As medidas de tendência central não são suficientes para se caracterizar um conjunto de dados. O motivo é que existe **variação** na natureza, isto é, dados que venham de uma mesma população não serão sempre iguais. Além disso, mesmo medidas feitas de um mesmo objeto ou sujeito (pense nas medidas da altura de uma pessoa, por exemplo) estarão sujeitas à precisão do instrumento de medida, isto é, poderão variar dentro dos limites de precisão do instrumento.

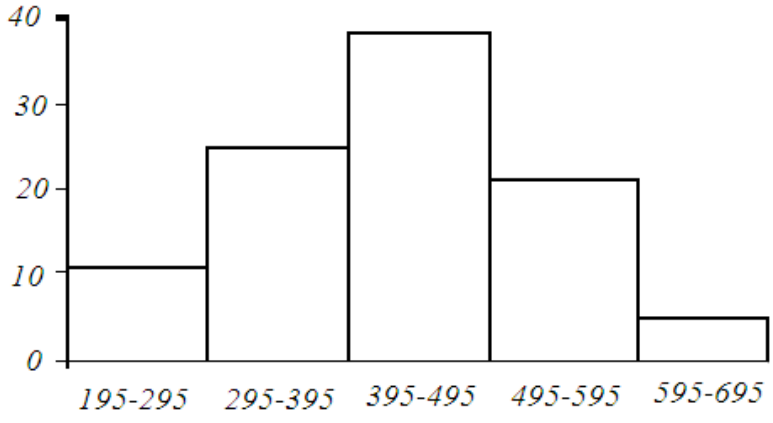
Para quantificar a variabilidade de um conjunto de dados ou medidas é que se usam medidas de dispersão. Vamos estudar algumas delas nesta aula.

A Amplitude Total dos Dados

A amplitude total dos dados de uma amostra é a diferença entre o maior e o menor número da amostra.

Por exemplo, para o conjunto de valores {2, 3, 4, 6, 6, 7, 7, 9, 9, 10, 12} a amplitude total é 12 – 2 = 10.

Já para o histograma abaixo, a amplitude total dos dados é 645 - 245 = 400. Note que esta amplitude foi calculada como a diferença entre os *pontos médios* da última e da primeira classe.



A amplitude total dos dados dá uma visão grosseira da variação, ou dispersão, dos dados. No entanto, em alguns casos é justamente esta visão grosseira sobre dispersão que se quer. Por exemplo, uma pessoa de férias no exterior e que pretende alugar um carro pode estar interessada em saber quais os valores máximo e mínimo que uma multa de trânsito pode ter no país para onde ela vai. Outro exemplo: o(a) dono(a) de uma loja pode querer saber qual o produto mais caro e qual o mais barato que ele(a) tem à venda.

O Desvio Médio, o Desvio Padrão e a Variância

O desvio médio de um conjunto de dados indica quão distantes "em média" estão os dados individuais em relação à média aritmética do grupo. Consideremos a seguinte tabela.

Número de horas vendo televisão num sábado de um grupo de 6 crianças de 12 anos

Nº da criança	Nº de horas (x_i)	$(x_i - \bar{x})$	$\lvert x_i - \bar{x}\rvert$	$(x_i - \bar{x})^2$
1	6	3	3	9
2	2	-1	1	1
3	4	1	1	1
4	1	-2	2	4
5	3	0	0	0
6	2	-1	1	1
	$\sum x_i = 18$	$\sum (x_i - \bar{x}) = 0$	$\sum \lvert x_i - \bar{x}\rvert = 8$	$\sum (x_i - \bar{x})^2 = 16$

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{18}{6} = 3 \text{ horas.}$$

A partir dos dados da segunda coluna calcula-se a média $\bar{x}$. A diferença entre um valor da amostra e a média dos valores da amostra é chamada de desvio. O desvio do *i*-ésimo elemento é definido por $x_i - \bar{x}$. A soma dos desvios dos elementos de uma amostra é *sempre* nula:

$$\sum_{i=1}^{N} (x_i - \bar{x}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \bar{x} = \sum_{i=1}^{N} x_i - N.\bar{x} = \sum_{i=1}^{N} x_i - N.\frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$$

Este fato está indicado pela terceira coluna da tabela acima. Na quarta coluna estão listados os valores absolutos dos desvios. A soma desses valores absolutos dividida pelo total de dados é o desvio médio:

$$\mathrm{DM} = \frac{\sum_{i=1}^{6} |x_i - \bar{x}|}{N} = \frac{8}{6} = 1{,}3 \text{ horas.}$$

Este resultado quer dizer que, em média, os dados estão 1,3 horas desviados (para mais e para menos) do valor médio do grupo, que vale 3 horas.

O desvio médio é muito pouco usado e só aparece aqui como artifício didático para ajudar na apresentação de uma medida similar, esta sim bastante usada, o desvio padrão.

Para obter o desvio padrão da amostra, somamos os quadrados dos desvios, ao invés dos seus valores em módulo, e dividimos o resultado por (N-1). O valor obtido é um tipo de média dos quadrados dos desvios, que é chamada de variância.

Como a variância é uma soma de quadrados, ela é expressa nas unidades da variável medida ao quadrado (no caso, horas ao quadrado).

Para voltarmos às unidades originais da variável medida (sem o quadrado), temos que tomar a raiz quadrada da variância. A raiz positiva da variância é chamada de desvio padrão.

A variância de uma amostra é designada por s^2 e o desvio padrão por s:

$$s^2 = \frac{\sum_{i=1}^{6} \left(x_i - \bar{x}\right)^2}{N-1} = \frac{16}{5} = 3{,}2 \text{ horas}^2; \qquad s = +\sqrt{s^2} = 1{,}79 \text{ horas}.$$

Para facilitar os cálculos, pode-se reescrever a fórmula para o desvio padrão através das propriedades da somatória:

$$\sum_{i=1}^{N}\left(x_i-\bar{x}\right)^2=\sum_{i=1}^{N}\left(x_i^2-2x_i\bar{x}+\bar{x}^2\right)=\sum_{i=1}^{N}x_i^2-2\bar{x}\sum_{i=1}^{N}x_i+\sum_{i=1}^{N}\bar{x}^2=$$

$$\sum_{i=1}^{N}x_i^2-2\frac{\sum_{i=1}^{N}x_i}{N}\sum_{i=1}^{N}x_i+N\left(\frac{\sum_{i=1}^{N}x_i}{N}\right)^2=\sum_{i=1}^{N}x_i^2-2\frac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}+\frac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}=$$

$$\sum_{i=1}^{N}x_i^2-\frac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}\Rightarrow s=\sqrt{\frac{\sum_{i=1}^{N}x_i^2-\frac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}}{N-1}}.$$

Observe que esta fórmula para o cálculo do desvio padrão requer apenas o conhecimento dos valores dos dados, x_i, e dos seus quadrados, x_i^2. Sendo assim, os únicos elementos que precisam ser listados na tabela de freqüência são os valores dos dados e os valores dos seus quadrados:

Nº da criança	Nº de horas (x_i)	x_i^2 (hs^2)
1	6	36
2	2	4
3	4	16
4	1	1
5	3	9
6	2	4
	$\sum x_i = 18$	$\sum x_i^2 = 70$

A partir desta tabela, o cálculo da variância e do desvio padrão é direto:

$$s^2=\frac{\sum_{i=1}^{N}x_i^2-\frac{\left(\sum_{i=1}^{N}x_i\right)^2}{N}}{N-1}=\frac{70-\frac{(18)^2}{6}}{5}=\frac{16}{5}=3{,}2\Rightarrow s=+\sqrt{3{,}2}=1{,}79\text{ horas.}$$

O desvio padrão é uma medida de dispersão. Quando temos dois conjuntos de dados e o primeiro tem uma variação em torno da média menor do que a do segundo, o desvio padrão do primeiro conjunto será menor que o do segundo conjunto.

A maneira como o desvio padrão mede dispersão é mais ou menos a mesma do desvio médio, isto é, medindo o afastamento médio dos dados em relação à média do conjunto. A diferença é que ao tomar o quadrado dos desvios, o desvio padrão faz uma espécie de média ponderada desses desvios, pois os desvios maiores entram na soma com pesos maiores que os desvios menores.

O desvio padrão, conforme foi definido, é o chamado **desvio padrão amostral**. Ele é obtido tomando-se a raiz quadrada da soma dos quadrados dos desvios dividida por ($N - 1$), o número de elementos na amostra menos um.

Existe outra definição de desvio padrão, válida para quando estamos trabalhando com uma população, ou seja, com o conjunto total de valores sendo estudado. Neste caso, o **desvio padrão populacional** é definido como a raiz quadrada da soma dos quadrados dos desvios dividida por N, o número total de dados na população,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}},$$

ou

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N}}.$$

Note que, para o caso do desvio padrão populacional, usou-se a letra grega σ (sigma) para representá-lo. Esta é a convenção adotada em estatística: o desvio padrão populacional é denotado por σ e o desvio padrão amostral é denotado por *s*.

De maneira geral, usa-se letras do alfabeto grego para representar variáveis relativas a uma população e letras do alfabeto latino para representar variáveis relativas a uma amostra (por exemplo, usa-se μ para representar a média de uma população e $\bar{x}$ para representar a média de uma amostra).

Alguém poderia perguntar por que o desvio padrão foi definido de um jeito para amostras e de outro para populações. A razão para isto está além dos objetivos deste curso. O que se pode dizer aqui é que se quisermos estimar o desvio padrão de uma população a partir do cálculo do desvio padrão de uma amostra retirada da população, o desvio padrão da amostra calculado dividindo-se por $(N - 1)$ será um melhor *estimador* do verdadeiro desvio padrão da população, σ, do que seria o desvio padrão da amostra calculado dividindo-se por *N*.

O Coeficiente de Variação

Em muitos casos é importante comparar a variabilidade relativa de muitos conjuntos de dados. Isto não pode ser feito apenas pelo exame dos desvios padrões dos conjuntos de dados, pois os conjuntos podem conter dados com magnitudes bem diferentes ou unidades diferentes.

Para fazer tal tipo de comparação, é costume expressar o desvio padrão como uma porcentagem da média aritmética. A variável definida a partir desta expressão é chamada de coeficiente de variação:

$$\mathrm{CV} = \frac{s}{\bar{x}} \cdot 100 \quad (\%).$$

Exemplo: Para um grupo de indivíduos, a temperatura corporal média é igual a 36,8°C com desvio padrão de 0,27°C e a pulsação média é igual a 78 batidas/min com desvio padrão de 9 batidas/min. Portanto, os coeficientes de variação para a temperatura e a pulsação dos indivíduos são:

$$CV_{temp.} = \frac{0,27}{36,8} \cdot 100 = 0,7\%; \qquad CV_{pulso} = \frac{9}{78} \cdot 100 = 11,5\%$$

Vemos então que a variabilidade relativa da pulsação é bem maior que a variabilidade relativa da temperatura. O coeficiente de dispersão é útil quando se quer analisar como a dispersão de um conjunto de dados varia no tempo, dado que a média dos dados também varia.

Exemplo: Suponhamos que uma pesquisa tenha sido feita comparando-se o aumento no preço de um cafezinho em seis diferentes bares da cidade entre 1994 e 2000 e os resultados sejam os dados abaixo (valores em reais).

Bar	A	B	C	D	E	F	$\bar{x}$	s	CV
1994	0,30	0,40	0,40	0,50	0,60	0,70	0,48	0,15	30,45%
2000	0,60	0,80	0,80	1,00	1,20	1,40	0,97	0,29	30,45%

Note que todos os valores dobraram de 1994 para 2000. O desvio padrão para a amostra também dobrou, indicando que a dispersão dos valores aumentou. Porém, o preço médio do cafezinho também dobrou, de maneira que o coeficiente de variação permaneceu constante. Podemos dizer que, de maneira absoluta, a dispersão dos preços do cafezinho dobrou entre 1994 e 2000; porém, de maneira relativa, ela permaneceu constante.

O Escore Padrão

Uma medida de dispersão relativa usada para caracterizar a variação de um dado em relação à media é o chamado escore padrão *z*, ou simplesmente escore *z*. Ele dá o desvio de um dado x_i em relação à média $\bar{x}$ medido em unidades de desvio padrão.

Seja um conjunto de dados com média $\bar{x}$ e desvio padrão *s*. O escore z_i do dado *i* é definido por

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Exemplo: Suponha que dois departamentos diferentes de uma empresa – por exemplo, de marketing e de recursos humanos – façam avaliações dos seus funcionários. Sejam as notas médias e os desvios padrões das avaliações dadas abaixo:

Marketing	**Recursos Humanos**
$\bar{x}_M = 6{,}5$	$\bar{x}_{RH} = 5{,}5$
$s_M = 1{,}4$	$s_{RH} = 0{,}8$

Suponha que um funcionário do Departamento de Marketing tenha recebido nota 8 e que um funcionário do Departamento de Recursos Humanos tenha recebido nota 7. Em termos absolutos, o funcionário do Departamento de Marketing teve nota mais alta, mas em termos relativos (ou seja, em comparação com os funcionários do seu próprio departamento) o funcionário do Departamento de Recursos Humanos teve um desempenho melhor, conforme atestado pelos escores *z* abaixo:

Funcionário do Departamento de Marketing	Funcionário do Departamento de RH
$z_{FM} = \frac{8{,}0 - 6{,}5}{1{,}4} = 1{,}07$	$z_{FRH} = \frac{7{,}0 - 5{,}5}{0{,}8} = 1{,}875$

O Desvio Padrão para Dados Agrupados

Assim como no caso do cálculo da média e da mediana, quando só temos acesso a uma tabela de freqüências a fórmula para o cálculo do desvio padrão passa a ser expressa em termos de uma aproximação, na qual os pontos médios dos intervalos de classe são usados como se fossem os dados verdadeiros.

Portanto, o que era

$$s = \sqrt{\frac{\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}}{N-1}},$$

passa a ser agora:

$$s = \sqrt{\frac{\sum_{i=1}^{N} f_i (PM_i)^2 - \frac{\left(\sum_{i=1}^{N} f_i PM_i\right)^2}{N}}{N-1}}.$$

Exemplo: Em um estudo para se verificar a eficácia de um novo anestésico, aplicaram-se várias doses do anestésico a 18 animais e mediram-se os tempos de duração das anestesias. Os resultados foram colocados na tabela a seguir. Calcule o desvio padrão dos valores.

Tempo de duração do efeito anestésico (min)	Ponto médio do intervalo (min) PM_i	Freqüência f_i	$f_i PM_i$	$f_i (PM_i)^2$
5 ⊢ 10	7,5	1	7,5	56,25
10 ⊢ 15	12,5	2	25	312,5
15 ⊢ 20	17,5	2	35	612,5
20 ⊢ 25	22,5	8	180	4050
25 ⊢ 30	27,5	5	137,5	3781,25
Soma		18	385	8812,5

Usando a fórmula para o desvio padrão para dados agrupados, temos:

$$s = \sqrt{\frac{\sum_{i=1}^{N} f_i (PM)^2 - \frac{\left(\sum_{i=1}^{N} f_i PM\right)^2}{N}}{N-1}} = \sqrt{\frac{8812{,}5 - \frac{385^2}{18}}{17}} = \sqrt{33{,}99} = 5{,}83 \text{ min.}$$

Exemplo Geral (medidas de tendência central e de dispersão): Um estudo para se determinar o perfil da renda dos universitários paulistanos resultou na seguinte tabela.

Faixa de Renda	**Exatas**	**Humanas**	**Biológicas**
Até 1 sal. mínimo	19%	19%	44%
1 a 3 sal. mínimos	18%	18%	24%
3 a 5 sal. mínimos	19%	21%	12%
Acima de 5 sal. mínimos	41%	38%	16%

Fonte: Perfil Sócio-Econômico do Universitário Paulista. Fórum dos Jovens Empresários (http://www.fjeacsp.com.br/SiteFJE/economico/economico.htm).

Vamos calcular a média, a mediana, a moda e o desvio padrão para os universitários da área de humanas. Deixamos os cálculos para os universitários das áreas de exatas e biológicas como exercício para casa.

A primeira coisa que devemos fazer para calcular os dados pedidos é reescrever a tabela acima colocando a informação que nos interessa, como pontos médios, freqüências acumuladas etc.

Devemos notar que a tabela não nos dá o número de estudantes pesquisados, ou seja, o valor de N. Portanto, não teremos como calcular o desvio padrão usando a fórmula para uma amostra, pois para isto teríamos que conhecer o valor de $(N - 1)$.

Porém, se supusermos que o número de estudantes na amostra foi muito grande isto não deverá causar maiores problemas, pois divisões por N ou por $(N - 1)$ resultarão em valores aproximadamente iguais. Note que embora o valor de N seja desconhecido, os valores da média e do desvio padrão podem ser calculados usando-se as fórmulas escritas em termos das freqüências relativas $f_r = f/N$.

Outro ponto importante sobre o qual devemos tomar uma decisão antes de montar a nova tabela é a definição de qual será o ponto médio do último intervalo usado. Note que este intervalo foi definido como "acima de 5 sal. mínimos". Portanto, só conhecemos o seu limite inferior. O limite superior, ou seja, a maior renda de um universitário, não é fornecido. Este é um exemplo em que a amplitude total dos dados não foi considerada relevante por quem fez a pesquisa. No entanto, para calcularmos a média e o desvio padrão *temos* que ter um valor para o ponto médio do último intervalo. Em um caso como este, a única alternativa é estimar um valor para o limite superior do último intervalo. Tal estimativa requer bom senso, pois o valor superior estimado não pode ser exageradamente alto (lembre-se que a média e o desvio padrão são bastante influenciados por valores muito altos). Para o caso em questão, vamos usar como limite superior do último intervalo o valor de 10 salários mínimos. Pode ser que existam universitários com rendas acima deste valor (com certeza existem), mas estamos supondo que eles não são muitos e não estamos querendo dar um peso muito grande a eles.

Procure fazer, como exercício para casa, este mesmo exercício usando valores diferentes para o limite superior do último intervalo; por exemplo 7 salários mínimos, 20 salários mínimos e 30 salários mínimos.

Uma vez feitas as definições acima, vamos agora montar a tabela de dados para os estudantes de humanas.

Faixa de Renda (s.m.)	**P.M.**	f_R	$f_{R.Ac.}$	f_Rx**P.M.**	f_Rx**(P.M.)^2**
0 ⊢ 1	0,5	0,19	0,19	0,09	0,05
1 ⊢ 3	2,0	0,18	0,37	0,36	0,72
3 ⊢ 5	4,0	0,21	0,58	0,84	3,36
5 ⊢ 10	7,5	0,38	0,96	2,85	21,37
Soma		0,96		4,14	25,50

O valor da média é o próprio valor da soma da coluna de f_RxP.M.:

$$\bar{x} = \sum f_R.\text{P.M.} = 4{,}14 \text{ s.m.}$$

O valor da mediana é o valor correspondente à freqüência relativa acumulada de 0,50. Note, porém, que a coluna de freqüências acumuladas nos dá um total de 0,96 (por algum motivo que não está explicado no *site* de onde os dados foram retirados). Neste caso, o valor da mediana deve corresponder à freqüência acumulada de 0,96/2 = 0,48. Portanto,

$$\text{MD} = 3 + \frac{2.(0{,}48 - 0{,}37)}{0{,}21} = 3 + 1{,}05 = 4{,}05 \text{ s.m.}$$

Note que este valor da mediana é, para o caso em questão, uma medida mais exata de tendência central do que o valor da média calculado anteriormente. Para calcular a média, fizemos uma suposição sobre o valor do extremo superior da última classe, o que pode ter induzido algum erro; já para o cálculo da mediana, este valor superior não teve qualquer influência.

A classe modal é a classe de maior freqüência, ou seja "acima de 5 s.m.".

Já o desvio padrão pode ser calculado pela fórmula:

$$s^2 = \sum f_R.(\text{P.M.})^2 - \left(\sum f_R.\text{P.M.}\right)^2 = 25{,}50 - \left(4{,}14\right)^2 = 8{,}36 \text{ s.m.} \Rightarrow$$
$$\Rightarrow \text{s} = \sqrt{8{,}36} = 2{,}89 \text{ s.m.}$$