MEDIDAS DE DISPERSÃO OU DE VARIABILIDADE OU DE CONCENTRAÇÃO

- São utilizadas para determinar as variações ou oscilações dos dados individuais em torno da média, da mediana e da moda, ou qualquer outra medida de tendência central.
- Servem também para verificar a representatividade das medidas de posição.

  Ex_1: {20, 20, 20, 20, 20} $\bar{x} = 20$

  Ex_2: {15, 10, 20, 25, 30} $\bar{x} = 20$

  Claramente, no Ex_1, a média tem uma alta representatividade como medida de posição do que no Ex_2.

1. Amplitude Total

$$A = X_{máx} - X_{mín}$$

Ex_1: Diâmetro de colônias de bactérias em mm, {2, 3, 5, 8, 10, 12}

a=12-2=10 mm é a oscilação máxima nos diâmetros das colônias de bactérias

2. Desvio Médio Absoluto

- Seja $X_1, X_2, \ldots, X_n$, uma amostra aleatória.

$$DM = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n} \quad (1)$$

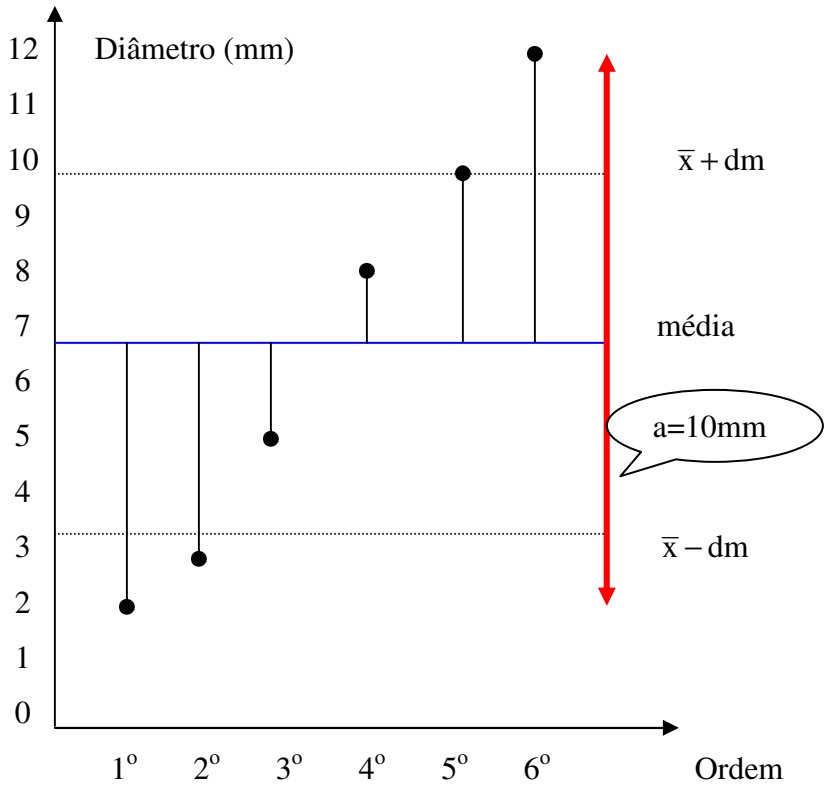
Ex_2: Diâmetro de colônias de bactérias em mm, {2, 3, 5, 8, 10, 12}

$$\bar{x} = \frac{2+3+\ldots+12}{6} = \frac{40}{6} \approx 6{,}67\,\mathbf{mm}$$

$$\text{dm} = \frac{|2-6{,}67| + |3-6{,}67| + |5-6{,}67| + |8-6{,}67| + |10-6{,}67| + |12-6{,}67|}{6}$$

$$\mathbf{dm} = \frac{|-4{,}67| + |-3{,}67| + |-1{,}67| + |1{,}33| + |3{,}33| + |5{,}33|}{6} = \frac{20}{6} \approx 3{,}33\mathbf{mm}$$

∴ O desvio médio absoluto em torno da média é de 3,33mm.



Nota: a porcentagem de diâmetros no intervalo $\bar{x} \pm dm$: $\frac{3}{6} \times 100 = 50\%$

- Cálculo do Devio Médio Absoluto através da Distribuição de Freqüências

Se $X_1, X_2, \ldots, X_k$ ocorrem com as freqüências $f_1, f_2, \ldots, f_k$, respectivamente, então,

$$DM = \frac{\sum_{i=1}^{n} f_i |X_i - \bar{X}|}{n} \quad (2)$$

Ex_3; Distribuição de freqüências das alturas (cm) de plantas de milho (1993)

Alturas (cm)	Freqüência f_J	Ponto Médio PM_j	$\lvert PM_j - \overline{X}\rvert$	$f_j\lvert PM_j - \overline{X}\rvert$
151|-158	5	154,5	17,2	86,0
159|-166	18	162,5	9,2	165,6
167|-174	42	170,5	1,2	50,4
175|-182	27	178,5	6,81	183,6
183|-|190	8	186,5	14,8	118,4
Total	100	-	-	604,0

$\overline{x} = 171{,}70\text{cm} \qquad \therefore \text{dm} = \dfrac{604}{100} = 6{,}04\,\text{cm}$

Intervalo em torno da média:

$$\overline{x} \pm \text{dm} \begin{cases} 165{,}66\text{cm} \\ 177{,}74\text{cm} \end{cases}$$

Qual a porcentagem (%) dos indivíduos pertencentes a este intervalo?

Sol.:

Número de indivíduos= $\dfrac{1}{7}(166 - 165{,}66)\times 18 + 42 + \dfrac{1}{7}(177{,}74 - 175)\times 27 \approx 53{,}4428 \approx 53$ plantas

Assim, temos $\dfrac{53}{100}\times 100 = 53\%$ das plantas nesse intervalo.

Determinação do número de pontos da 3ª classe que caem no intervalo:

$$\begin{cases} (166 - 159) - - - 18 \\ (166 - 165{,}6) - - x \end{cases}$$

3. Amplitude Semi-Interquartílica ou Desvio Quartílico

$$Q = \frac{Q_3 - Q_1}{2}$$

onde Q_1 e Q_3 são o primeiro e o terceiro quartis referentes aos dados. A amplitude interquartílica Q_3-Q_1 é empregada algumas vezes, mas a amplitude semi-interquartílica é mais comum como medida de dispersão.

4. Amplitude entre os Percentis 10-90 = P_{90}-P_{10}

5. Desvio Padrão

Seja X_1, X_2, ... , X_n uma amostra aleatória de tamanho n, então:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Obs$_1$: O valor –1 na fórmula explica-se pelo fato de ter sido estimado a média e portanto que o total era conhecido a priori. Assim, conhecido o total tem-se a liberdade de perder uma observação X da amostra e mesmo assim toda informação da amostra é recuperada.

Obs$_2$: É uma Média Quadrática dos desvios.

Para uma população de tamanho N, o desvio padrão é:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}, \quad \text{onde} \quad \mu = \frac{\sum_{i=1}^{N} X_i}{N} \text{ é a média populacional}$$

Obs$_1$: O desvio padrão S usando (n-1) em lugar de n é um estimador não viciado do verdadeiro desvio padrão da população. O s (minúsculo) é a estimativa.

Obs$_2$: Para amostras grandes (n>30), não há praticamente diferença entre as duas definições, isto é dividir por (n-1) ou n.

Ex$_4$: Diâmetro de colônias de bactérias em mm, {2, 3, 5, 8, 10, 12}

$$\mathbf{s} = \sqrt{\frac{(2-6{,}67)^2 + (3-6{,}67)^2 + \cdots + (12-6{,}67)^2}{(6-1)}} \approx 3{,}9833\mathbf{mm}$$

Quando os dados estão agrupados (temos a presença de freqüências) a fórmula do desvio padrão ficará :

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2 f_i}{\sum_{i=1}^{n} f_i - 1}}$$

6. Variância

É o quadrado do desvio padrão:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \qquad \text{(amostral)}$$

ou $$\sigma^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{N} \quad \text{(populacional)}$$

Ex_5: Diâmetro de colônias de bactérias em mm, {2, 3, 5, 8, 10, 12}

$$\mathbf{s}^2 = \frac{(2-6{,}67)^2 + (3-6{,}67)^2 + \cdots + (12-6{,}67)^2}{(6-1)} \approx 15{,}8666\mathbf{mm}^2$$

- DESVANTAGEM DA VARIÂNCIA EM RELAÇÃO AO DESVIO PADRÃO: não está na mesma unidade da variável analisada.

- PROPRIEDADES DO DESVIO PADRÃO

a) Suponha a fórmula $S^* = \sqrt{\frac{\sum_{i=1}^{n}(X_i - a)^2}{n-1}}$, em que a é uma média próxima da aritmética. De todos os possíveis a's (e portanto desvios) o menor desvio é aquele para o qual a= $\overline{x}$.

   Ex_6: Diâmetro de colônias de bactérias em mm, {2, 3, 5, 8, 10, 12}

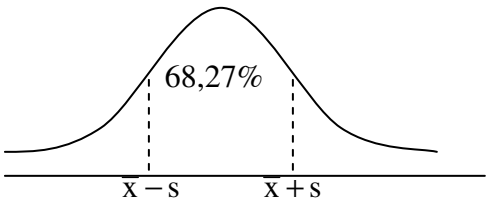
   $$\text{md} = \frac{5+8}{2} = 6{,}5 \quad \text{e} \quad s^* = \sqrt{\frac{(2-6{,}5)^2 + (3-6{,}5)^2 + \cdots + (12-6{,}5)^2}{6-1}} \approx 3{,}9874\,\text{mm}$$

   $\therefore$ s* é maior que s

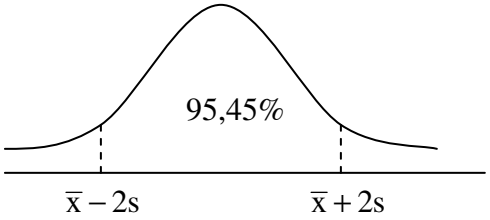
b) Para a distribuição normal (simétrica)

   1. 68,27% dos dados estão incluídos no intervalo $[\overline{x} - s \text{ e } \overline{x} + s]$
   2. 95,45% dos dados estão incluídos no intervalo $[\overline{x} - 2s \text{ e } \overline{x} + 2s]$
   3. 99,73% dos dados estão incluídos no intervalo $[\overline{x} - 3s \text{ e } \overline{x} + 3s]$

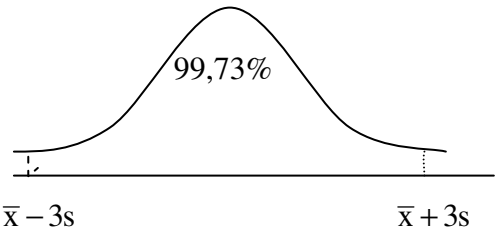
1.



2.



3.



c) Suponha que dois conjuntos de n_1 e n_2 dados de uma mesma população (ou das distribuições com freqüências totais n_1 e n_2) tenham variâncias representadas por s_1^2 e s_2^2, respectivamente e a mesma média $\bar{x}$. Então a variância conjunta ou combinada é:

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$$

que é a média ponderada das variâncias

Generalizando :

$$s^2 = \frac{n_1 s_1^2 + n_2 s_2^2 + \cdots + n_k s_k^2}{n_1 + n_2 + \cdots + n_k}, \qquad \forall\, k\,.$$

7. Dispersão Absoluta e Relativa. Coeficiente de Variação

* A variação ou dispersão real, determinada a partir do desvio padrão, ou qualquer outra medida de dispersão, é denominada dispersão absoluta.

Questão: Uma dispersão de 10cm nas medidas de árvores de sibipiúna de 50m de média é diferente de uma dispersão de 10cm nas medidas de árvores de sibipiúna de 20m de média?

Considere as alturas de quatro árvores de sibipiúna (da mesma família do pau brasil). As medidas foram feitas com o uso do aparelho suunto.

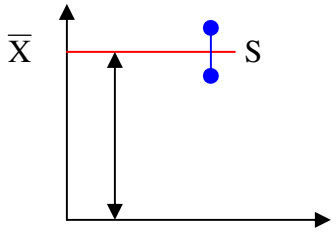
Amostras	Árvores					
	A1	A2	A3	A4	$\bar{x}$	s
A	49	51	52	48	50m	1,8257m
B	21	19	18	22	20m	1,8257m

Então: $\text{Dispersão Relativa} = \dfrac{\text{Dispersão Absoluta}}{\text{Média}}$

- Se a dispersão absoluta é o desvio padrão amostral S e a média é a média aritmética amostral $\overline{X}$, a dispersão relativa amostral é denominada COEFICIENTE DE VARIAÇÃO:

$$CV = \frac{S}{\overline{X}} \times 100$$

Note: $\begin{cases} \overline{X} \_\_\_ 100\% \\ S \_\_\_ CV \end{cases}$



Ex_1: Diâmetro de colônias de bactérias em mm, {2, 3, 5, 8, 10, 12}, em que $\overline{\mathbf{x}} = 6{,}67\mathbf{mm}$ e s=3,9833mm, têm-se que $cv = \dfrac{3{,}9833\text{cm}}{6{,}67\text{cm}} \times 100 = 59{,}72\%$

Nota: O CV é adimensional.

Ex_2: Amostras das árvores de sibipiúna:

$$cv_A = \frac{1{,}8257\text{m}}{50\text{m}} \times 100 = 3{,}6\% \quad \text{e} \quad cv_B = \frac{1{,}8257\text{m}}{20\text{m}} \times 100 = 9{,}1\%$$

Cuidado com unidades diferentes!

Ex: Uma dispersão de 10cm nas medidas de árvores de sibipiúna de 50m de média e uma dispersão de 10cm nas medidas de árvores de sibipiúna de 20m de média?

cv_A=0,20%

cv_B=0,50%

ÍNDICE DE DIVERSIDADE

As medidas de dispersão vistas anteriormente aplicam-se a variáveis quantitativas. Comunidades biologias, ou outras, entidades como coleções de organismos, podem ser estudadas sob o ponto de vista de variabilidade através dos chamados índices de diversidades, que estão para as variáveis qualitativas como a variância está para as variáveis quantitativas.

Exemplo: Foi levantado dados de restaurantes em uma localidade, em que das variáveis estudadas, a APARÊNCIA dos restaurantes, que é uma variável qualitativa, com classificação de 1 a 7, do menos favorável ao mais favorável, é mostrada a seguir:

Categorias dos restaurantes	Número de restaurantes
1	37
2	40
3	66
4	73
5	49
6	9
7	4
Total	278

Podemos construir duas colunas hipotéticas (A e B) que substituiriam "Número de restaurantes"

Categorias dos restaurantes	Número de restaurantes - A	Número de restaurantes - B
1	40	1
2	40	1
3	40	1
4	40	1
5	40	1
6	40	1
7	38	272
Total	278	278

Onde há maior incerteza com relação ao tipo de preferência dos restaurantes?

Intuitivamente percebe-se que a situação A apresenta uma diversidade maior de preferências do que a situação B, onde temos concentração em uma das categorias. Numa situação como em A temos maior incerteza com relação ao tipo preferencial de Aparência de restaurantes na localidade estudada do que naquela em B: maior incerteza, maior diversidade.

Supondo que os dados tenham sido colhidos de modo aleatório, Shannon (ver Zar, 1984), com base na chamada teoria de informação, definiu o seguinte índice H de diversidade, denominado índice de Shannon:

$$H = -\sum_{i=1}^{k} \pi_i \log(\pi_i),$$

Onde k é o número de categorias e π_i é a proporção de cada categoria. Utilizando o índice descrito para a situação original dos dados e também para as situações hipotéticas temos:

Exercício:

H=1,717;

H (A)=1,945;

H (B)= 0,142.

O valor máximo de H é H_{max}=log(k) (logaritmo na base e) e o índice de Shannon é comumente corrigido por este valor, dando origem ao chamado índice de diversidade relativa (ou de homogeneidade) cuja expressão é:

$$J = \frac{H}{H_{\max}}$$

No presente exemplo, H_{max}=log(7)=1,945 e os índices de diversidade relativa são:

J=0,882;

J (A)=1;

J (B)= 0,073.

Referências:

Zar, J.H. (1984). Biostatistical Analysis. 2ª ed., New Jersey: Prentice-Hall, Englewood Cliffs.

Botter, D.A.; Paula, G.A.; Leite, J.G.; Cordani, L.K. (1996) Noções de estatística – com apoio computacional. IME-USP.