UNIVERSIDADE DE SÃO PAULO

Faculdade de Zootecnia e Engenharia de Alimentos

ZAB1111 – Estatística Básica

Prof. César Gonçalves de Lima cegdlima@usp.br

Aula 4

Medidas de dispersão

3.2. MEDIDAS DE DISPERSÃO: procuram resumir a <u>variabilidade</u> da série de valores em torno da média.

Exemplo 1. Sabe-se que nas repúblicas A, B e C moram 4, 5 e 3 alunos da Engenharia de Biossistemas, respectivamente, e que a idade média desses alunos é 19 anos.

A média não fornece informação sobre a variabilidade dos dados em cada república.

República	Idade					Média
A	18	19	19	20		19
B	19	18	17	19	22	19
C	19	19	19			19

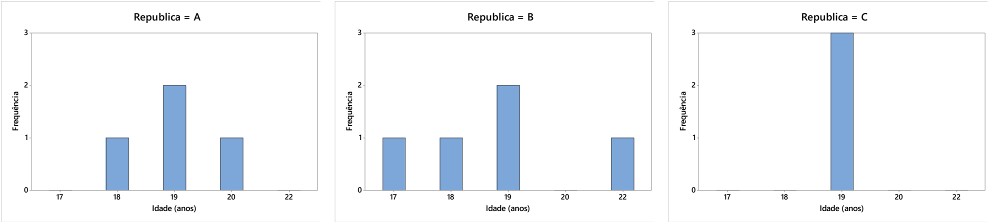

Figura 1. Histogramas da idade de alunos de três repúblicas

Avaliando os três histogramas da Figura 1 percebe-se que os grupos apresentam variabilidades distintas em relação à média, 19 anos.

Ideia: Buscar uma medida que <u>quantifique a variabilidade</u> dos dados e permita, por exemplo, identificar o grupo mais homogêneo.

Medidas de dispersão populacionais:

- Desvio médio: $DM(X) = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$
- Variância: $var(X) = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Desvio padrão: $dp(X) = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- Coeficiente de variação: $\mathrm{CV} = \frac{100\,\sigma}{\bar{x}}\,\%$

Medidas de dispersão amostrais:

- Variância: $var(X) = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$

- Desvio padrão: $dp(X) = s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$

- Coeficiente de variação: $\text{CV} = \frac{100\, s}{\bar{x}}\, \%$

Nas próximas aulas será explicado o motivo da substituição do denominador das fórmulas da variância e desvio padrão de n por $n - 1$.

Como os próximos exemplos envolvem dados amostrais, deveremos usar essas últimas fórmulas para calcular as medidas de dispersão.

Utilizando as propriedades de somatórios podemos simplificar a fórmula da variância:

$$var(X) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]$$

Mesmo assim fica difícil/trabalhoso calculá-la sem o uso de uma calculadora científica.

Na prática, se precisarmos calcular a variância usando uma calculadora científica, primeiramente calculamos o <u>desvio padrão amostral</u>, s, e elevamos este valor ao quadrado, para obter a variância s^2.

Problema: Como usar a calculadora para calcular essas medidas de dispersão.

Usando as fórmulas apresentadas podemos calcular todas as medidas de variabilidade para os três grupos de idades:

	A		B		C	
	x_i	$x_i - 19$	x_i	$x_i - 19$	x_i	$x_i - 19$
	18	-1	19	0	19	0
	19	0	18	-1	19	0
	19	0	17	-2	19	0
	20	1	19	0		
			22	3		
$\sum_{i=1}^{n}(x_i - \bar{x})$		0		0		0
$\sum_{i=1}^{n}\lvert x_i - \bar{x}\rvert$		2		6		0
$\sum_{i=1}^{n}(x_i - \bar{x})^2$		2		14		0

Aplicando as fórmulas corretamente temos:

	Grupo		
Dispersão	A	B	C
$DM(X)$	0,50	1,2	0
s^2	0,6667	3,50	0
s	0,8165	1,8708	0
$CV(\%)$	4,30%	9,85%	0%

- O grupo C é o mais homogêneo, pois apresenta os menores valores para todas as medidas de dispersão.
- O grupo B é o mais heterogêneo, pois apresenta os maiores valores para todas as medidas de dispersão.

Sobre as medidas de dispersão podemos comentar que:

- A variável com maior medida de dispersão é a mais heterogênea; a que tem menor medida de dispersão é a mais homogênea.

- O **desvio médio** é a medida **menos** usada, por conta de usar a função módulo ou valor absoluto, o que dificulta a dedução de suas propriedades estatísticas.

- O **desvio padrão** é a medida **mais** usada porque tem a mesma unidade de medida que a variável original, em estudo.

- O **coeficiente de variação** é uma medida de dispersão relativa que serve para comparar a variabilidade (relativa) de duas ou mais variáveis que tenham unidades de medidas distintas.

Se os dados da variável discreta estiverem classificados em uma distribuição de frequências usamos a fórmula:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}$$

Se os dados da variável contínua estiverem classificados em uma distribuição de frequências:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{k} (Pm_i - \bar{x})^2 f_i}$$

Não se preocupe! A calculadora científica vai ajudar no cálculo do desvio padrão, que é a medida de dispersão mais utilizada.

Exemplo: Para os dados de tamanho da ninhada (Tabela 2) tem-se:

Tamanho	f_i	$x_i - 4{,}85$
1	1	$-3{,}85$
2	2	$-2{,}85$
3	7	$-1{,}85$
4	8	$-0{,}85$
5	8	$0{,}15$
6	6	$1{,}15$
7	5	$2{,}15$
8	2	$3{,}15$
9	1	$4{,}15$
Total	40	0

Lembrando que $\bar{x} = 4{,}85$

- $Var(X) = s^2 = \frac{1}{40-1}\left[(-3{,}85)^2 + \cdots + (4{,}15)^2\right] = \frac{129{,}10}{39}$

  $= 3{,}3103\ (coelhos/ninhada)^2$

- $DP(X) = \sqrt{3{,}3103} = 1{,}8194 \cong 1{,}82\ coelhos/ninhada$

- $CV = 100\,\frac{1{,}8194}{4{,}85} = 37{,}5\%$

Este valor de CV indica uma variabilidade relativa alta do tamanho da ninhada.

Exemplo: Distribuição de frequências do peso de coelhos ao desmame, sabendo-se que $\bar{x} = 820$ gramas.

Peso (g)	P_{mi}	f_i	$(Pm_i - 820)$
$490 \vdash 590$	540	4	-280
$590 \vdash 690$	640	3	-180
$690 \vdash 790$	740	8	-80
$790 \vdash 890$	840	13	20
$890 \vdash 990$	940	6	120
$990 \vdash 1090$	1040	6	220
Total		40	

- $Var(X) = s^2 = \frac{1}{39}\left[(-280)^2(4) + \cdots + (220)^2(6)\right]$

$$= \frac{844.000}{39} = 21.641{,}03\ gramas^2$$

- $DP(X) = s = \sqrt{21.641{,}03} = 147{,}1089\ gramas$

- $CV = 100\,\frac{147{,}1089}{820} = 17{,}9\%$

Comparando-se os CV's do tamanho da ninhada e do peso ao desmame dos coelhos, pode-se concluir que o tamanho da ninhada tem uma variabilidade relativa maior que a do peso ao desmame.

Sugestão:

Resolver os primeiros exercícios até 3(b) do documento “Lista 1”, compartilhado na pasta Listas de Exercícios do Moodle.

Oportunamente a resolução dos exercícios será disponibilizada.