Universidade Federal de Lavras
Departamento de Estatística
Prof. Daniel Furtado Ferreira

4ª Aula Prática — Medidas de Dispersão

1) Os dados apresentados a seguir referem-se ao levantamento dos intervalos de parto em meses para uma amostra em $n = 20$ produtores rurais atendidos pelo plano "Panela Cheia" (Roesler, 1997), realizado na região oeste do Paraná, no município de Marechal Cândido Rondon, PR, em 1992, avaliados antes do plano ser aplicado. Os resultados dos intervalos entre partos em meses são dados por:

11,80	11,90	12,00	12,30	12,80	12,99	13,10	13,50	13,80	14,10
14,55	14,65	14,70	15,00	15,10	15,20	15,50	15,80	15,90	15,96

a) Obter a amplitude total (A). Qual é o seu significado e suas limitações?
b) Obter a variância S^2 e o desvio padrão S.
c) Determinar o coeficiente de variação CV. Qual é seu significado? Qual é a principal diferença entre o CV e o desvio padrão e a variância?
d) Erro padrão da média. A média do intervalo entre parto foi calculada com alta ou baixa precisão?
e) Se você fosse solicitado a apresentar duas medidas (estatísticas) para sintetizar os dados, quais você recomendaria?
f) Se cada dado for dividido por 12, para se obter o intervalo entre partos em anos, quais serão os novos valores da amplitude, variância, desvio padrão, CV e erro padrão da média?

2) Agrupar os dados do intervalo entre partos em classes (distribuição de frequências), resolver e responder as questões apresentadas a seguir.

(a) Determinar a média, a mediana e a moda.
(b) Calcular a amplitude, variância, desvio padrão, CV, erro padrão da média e CP.
(c) Após o programa Panela Cheia o intervalo de partos apresentou média de 13,85 e desvio padrão de 2,00 meses. Qual é a situação que apresentou maior variabilidade, anterior ou posterior ao Plano Governamental? Em qual caso a média foi calculada com maior precisão? Justifique sua resposta com os cálculos apropriados.

3) Os dados a seguir referem-se ao número empresas falidas/ano observadas em $n = 85$ anos. A amostra foi obtida em Lavras, MG.

Empresas falidas	Frequências
0	36
1	19
2	16
3	7
4	4
5	2
6	1

Determinar:

a) Calcular: a amplitude, variância, desvio padrão e o erro padrão da média.
b) Determinar: CV e CP.
c) Se os dados forem multiplicados por $k = 10$, quais são os novos valores de todas estas medidas de dispersão?

4) Diferenciar as medidas de variabilidades dadas pelo desvio padrão S e pelo erro padrão da média $S_{\bar{X}}$.

Resolução

1) As medidas de dispersão e as demais quantidades solicitadas a respeito dos dados dos intervalos de partos do município de Marechal Cândido Rondon são:

   a) A amplitude é dada por:

$$A = x_{(n)} - x_{(1)} = 15{,}96 - 11{,}80 = 4{,}16 \text{ meses.}$$

   A amplitude total representa a variação entre o menor e o maior valor, sendo simples de calcular e interpretar. Possui a limitação de tender a aumentar com o aumento da amostra, pois quanto maior a amostra maior a chance de amostrar valores extremos da população que ocorrem com baixa frequência. Também é influenciada por valores extremos, os *outliers*, pois envolve apenas o valor mínimo e máximo da amostra. Da mesma forma, por considerar apenas os dois valores extremos da amostra, pode não retratar a real variabilidade do conjunto de dados. Veja o exemplo: 2, 4, 4, 4, 4, 4, 4, 10. A amplitude total é igual a 8, mas os dados intermediários da amostra não apresentam variabilidade.

   b) A variância e o desvio padrão são:

$$\begin{aligned} S^2 &= \frac{1}{19}\left[(11{,}80^2 + \ldots + 15{,}96^2) - \frac{(11{,}80 + \ldots + 15{,}96)^2}{20}\right] \\ &= \frac{1}{19}\left[3975{,}717 - \frac{280{,}65^2}{20}\right] = 1{,}973451 \text{ mes}^2 \end{aligned}$$

   e $S = \sqrt{1{,}973451} = 1{,}404796$ mes.

   c) O coeficiente de variação CV é dado por:

$$\begin{aligned} CV &= \frac{1{,}404796}{14{,}0325} \times 100\% \\ &= 10{,}01102\%. \end{aligned}$$

   O coeficiente de variação expressa a variabilidade da amostra em porcentagem da média, sendo uma medida adimensional que não depende da grandeza dos dados. Já a variância e o desvio padrão, são medidas de variabilidade absoluta dos dados em torno da média. A diferença entre as duas medidas é que a variância é uma grandeza que está na unidade dos dados ao quadrado (meses^2) e o desvio padrão, na mesma unidade dos dados, sendo mais fácil de interpretar.

   d) O erro padrão da média é dado por:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{1{,}404796}{\sqrt{20}} = 0{,}3141219.$$

   Para responder a questão formulada, é necessário obter o coeficiente de precisão por

$$CP = \frac{S_{\bar{X}}}{\bar{X}} \times 100\% = \frac{0{,}3141219}{14{,}0325} \times 100\% = 2{,}238531\%.$$

   Como o erro padrão representou apenas 2,24% do valor médio, concluí-se que a média populacional foi estimada com alta precisão, pois o erro relativo (CP) foi muito pequeno.

   e) Para representar um conjunto de dados com duas medidas descritivas, deve-se utilizar uma medida de posição e outra de dispersão. Se a amostra possuir uma distribuição simétrica ou com

pequena assimetria apenas, deve-se utilizar a média como medida de posição. Se a distribuição for assimétrica, as medidas de posição robustas, como mediana e moda, devem ser preferidas, pois são pouco influenciadas por valores extremos. Como medida de dispersão, podemos utilizar ou a variância, ou o desvio padrão ou o coeficiente de variação, se o interesse for retratar a variabilidade entre os elementos da amostra em relação a sua média. Se por outro lado, o interesse for na precisão da estimativa da média populacional, ou o erro padrão ou o CP devem ser utilizados. A escolha entre uma medida absoluta e relativa fica a critério do pesquisador, pois podemos facilmente migrar de uma para outra.

f) As novas medidas de variabilidade após a divisão dos dados originais pela constante $k = 12$ são:

i) A nova amplitude total é:

$$A^* = \frac{A}{k} = \frac{4{,}16}{12} = 0{,}3466667 \text{ ano.}$$

ii) A nova variância é:

$$S^{2*} = \frac{S^2}{k^2} = \frac{1{,}973451}{12^2} = 0{,}01370452 \text{ ano}^2.$$

iii) O novo desvio padrão é:

$$S^* = \frac{S}{k} = \frac{1{,}404796}{12} = 0{,}1170663 \text{ ano.}$$

iv) O novo CV é:

$$\begin{aligned} CV^* =& \frac{S^*}{\bar{X}^*} \times 100\% = \frac{S/k}{\bar{X}/k} \times 100\% = CV \\ =& 10{,}01102\%. \end{aligned}$$

Isto indica que a variabilidade relativa não se altera, com a transformação de unidade, mas as variabilidades absolutas são alteradas.

v) O novo erro padrão da média e o novo CP são:

$$S^*_{\bar{X}} = \frac{S_{\bar{X}}}{k} = \frac{0{,}3141219}{12} = 0{,}02617682$$

e

$$CP^* = CP = 2{,}238531\%.$$

2) Para agrupar os dados deve-se obter:

O número de classe é dado por $k = \sqrt{n} = \sqrt{20} \approx 4$ e amplitude total por $A = X_{(20)} - X_{(1)} = 15{,}96 - 11{,}80 = 4{,}16$. Assim, a amplitude de classe é dada por $c = A/(k-1) = 4{,}16/3 \approx 1{,}39$ e o limite inferior da primeira classe por $LI_1 = x_{(1)} - c/2 = 11{,}80 - 1{,}39/2 = 11{,}11$. Os demais limites de classe são obtidos somando-se $c = 1{,}39$ aos limites anteriormente obtidos. A distribuição de frequências é:

Classes dos tempos	$\bar{X}_i$	F_i	Fr_i	$Fp_i(\%)$
$11{,}11 \vdash 12{,}50$	11,81	4	0,20	20
$12{,}50 \vdash 13{,}89$	13,20	5	0,25	25
$13{,}89 \vdash 15{,}28$	14,59	7	0,35	35
$15{,}28 \vdash 16{,}67$	15,98	4	0,20	20

a) A média aritmética é dada por:

$$\bar{X} = \frac{\sum_{i=1}^{k} F_i \bar{X}_i}{n} = \frac{11{,}81 \times 4 + 13{,}20 \times 5 + 14{,}59 \times 7 + 15{,}98 \times 4}{20} = \frac{279{,}29}{20}$$
$$= 13{,}9645 \text{ meses.}$$

A diferença encontrada para a média dos dados não agrupados (14,0325) pode ser atribuída ao agrupamento. Toda forma de representar os dados de uma maneira mais simplificada conduz a algum tipo de perda de precisão. Mas o que deve ficar claro é que apesar de menos precisa, a estimativa obtida a partir dos dados agrupados é uma "estimativa confiável" da média populacional, tanto quanto a estimativa dos dados originais. A perda de precisão é, em geral, pequena e pode ser considerada desprezível.

A mediana é obtida da seguinte maneira. A classe mediana é aquela que contém a posição número $n/2 = 20/2 = 10$. Portanto, a classe mediana é a terceira, pois as frequências acumuladas das duas primeiras classes somam apenas 9, que é inferior a 10. Logo,

$$m_d = LI_{m_d} + \frac{\frac{n}{2} - F_A}{F_{m_d}} c_{m_d} = 13{,}89 + \frac{10 - 9}{7} \times 1{,}39$$
$$= 14{,}08857 \text{ meses.}$$

Para obter a moda, é necessário determinar a classe de maior frequência, ou seja, a classe modal. A classe modal neste exercício é a terceira. A diferença das frequências da classe modal e classe anterior é $\Delta_1 = 7 - 5 = 2$ e a diferença das frequências da classe modal e classe posterior é $\Delta_2 = 7 - 4 = 3$. Assim, tem-se

$$m_o = LI_{m_o} + \frac{\Delta_1}{\Delta_1 + \Delta_2} c_{m_o} = 13{,}89 + \frac{2}{2 + 3} \times 1{,}39$$
$$= 14{,}446 \text{ meses.}$$

As três medidas, média, mediana e moda, estão muito próximas e isso é um indicativo que a distribuição dos dados deve ser aproximadamente simétrica.

b) As medidas de dispersão para os dados agrupados são dadas na sequência. A amplitude total é dada por

$$A = \bar{X}_k - \bar{X}_1 = 15{,}98 - 11{,}81 = 4{,}17 \text{ meses,}$$

a variância, por

$$S^2 = \frac{1}{n-1} \left[\sum_{i=1}^{k} \bar{X}_i^2 F_i - \frac{\left(\sum_{i=1}^{k} \bar{X}_i F_i \right)^2}{n} \right]$$
$$= \frac{1}{19} \left[11{,}81^2 \times 4 + 13{,}20^2 \times 5 + 14{,}59^2 \times 7 + 15{,}98^2 \times 4 - \right.$$
$$\left. - \frac{(11{,}81 \times 4 + 13{,}20 \times 5 + 14{,}59 \times 7 + 15{,}98 \times 4)^2}{20} \right] = \frac{1}{19} \left(3940{,}623 - \frac{279{,}29^2}{20} \right)$$
$$= 2{,}130394 \text{ meses}^2,$$

o desvio padrão, por $S = \sqrt{2{,}130394} = 1{,}459587$ meses, o CV, por

$$CV = \frac{1{,}459587}{13{,}9645} \times 100\% = 10{,}45213\%,$$

o erro padrão da média,

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{1{,}459587}{\sqrt{20}} = 0{,}3263736 \text{ mes},$$

e o CP, por

$$CP = \frac{S_{\bar{X}}}{\bar{X}} \times 100\% = \frac{0{,}3263736}{13{,}9645} \times 100\% = 2{,}337166\%.$$

c) Para responder estas questões é necessário determinar o CV e o CP, antes e após o plano panela cheia. Na tabela seguinte foram resumidas as informações necessárias.

Medida de variabilidade	Antes do plano	Após o plano
CV	10,01%	14,44%
CP	2,24%	3,23%

Como o CV do pós plano é maior do que o CV pré plano, há uma maior variabilidade dos intervalos de parto após o plano panela cheia ter sido implementado. Da mesma forma, houve uma menor precisão na estimativa da média populacional na situação pós plano, pois o erro padrão expresso em porcentagem da média (CP) foi maior do que na situação pré plano.

3) Para a variável número de empresas falidas por ano tem-se:

a) As medidas de dispersão para este conjunto de dados são apresentadas na sequência. A amplitude total é

$$A = x_{(n)} - x_{(1)} = x_{(85)} - x_{(1)} = 6 - 0 = 6 \text{ empresas falidas/ano}.$$

A variância é

$$\begin{aligned} S^2 =& \frac{1}{n-1}\left[\sum_{i=1}^{k} x_i^2 F_i - \frac{\left(\sum_{i=1}^{k} x_i F_i\right)^2}{n}\right] \\ =& \frac{1}{84}\left[0^2 \times 36 + 1^2 \times 19 + \ldots + 6^2 \times 1 - \frac{(0 \times 36 + 1 \times 19 + \ldots + 6 \times 1)^2}{85}\right] \\ =& \frac{1}{84}\left(296 - \frac{104^2}{85}\right) \\ =& 2{,}008964 \text{ (empresas falidas/ano)}^2, \end{aligned}$$

em que k é o número de categorias da variável, 7 no caso; o desvio padrão é $S = \sqrt{2{,}008964} = 1{,}417379$ empresa falida/ano e o erro padrão da média

$$S_{\bar{X}} = \frac{S}{\sqrt{n}} = \frac{1{,}417379}{\sqrt{85}} = 0{,}1537364 \text{ empresa falida/ano}.$$

b) O CV e o CP são

$$\begin{aligned} CV =& \frac{S}{\bar{X}} \times 100\% = \frac{1{,}417379}{1{,}223529} \times 100\% \\ =& 115{,}8435\% \end{aligned}$$

e

$$\begin{aligned} CP =& \frac{S_{\bar{X}}}{\bar{X}} \times 100\% = \frac{0{,}1537364}{1{,}223529} \times 100\% \\ =& 12{,}56499\%, \end{aligned}$$

respectivamente. Estes valores indicam que há uma grande variabilidade dos dados em torno da média e que a precisão da estimativa da média populacional não é muito alta, embora seja boa. Convém salientar que, tanto para o CV quanto para o CP, o pesquisador deve buscar na literatura experimentos semelhantes ao seu, ou seja, com as mesmas características utilizadas e com a mesma variável, entre outros fatores, para fazer uma comparação da variabilidade e da precisão adequadamente.

c) Utilizando a constante de multiplicação $k = 10$, tem-se:

$$\begin{aligned} A^* =& kA = 10 \times 6 = 60, & S^{2*} =& k^2 S^2 = 100 \times 2{,}008964 = 200{,}8964, \\ S^* =& kS = 10 \times 1{,}417379 = 14{,}17379, & S^*_{\bar{X}} =& kS_{\bar{X}} = 10 \times 0{,}1537364 = 1{,}537364, \\ CV^* =& CV = 115{,}8435\% \quad \text{e} & CP^* =& CP = 12{,}56499\%. \end{aligned}$$

4) O desvio padrão S é uma medida da variação que ocorre entre os elementos de uma amostra ou entre cada elemento da amostra e a média amostral. Ele se constitui numa estimativa da variabilidade que ocorre entre os elementos da população ou da variabilidade entre os elementos da população e a média populacional. Por outro lado, o erro padrão da média, $S_{\bar{X}}$, estima, a partir de resultados de uma única amostra, a variabilidade que ocorre entre as médias amostrais, se diferentes amostras de tamanho n, iguais a amostra original, fossem realizadas. Da mesma forma, podemos dizer que ele, o erro padrão da média, estima a variabilidade entre as médias amostrais obtidas em diferentes amostras de tamanho n e a média da população. Assim, quando o erro padrão da média é pequeno, as médias amostrais tendem a se concentrar em torno da média populacional. Nesse caso dizemos que temos alta precisão da estimativa da média populacional, pois a chance da estimativa de a média populacional na amostra atual estar "distante" da média populacional é pequena. Se, por outro lado, o erro padrão for grande, a distribuição das médias amostrais em torno da média populacional possui grande variabilidade e, portanto, pouco concentrado em torno de μ. Nesse caso, dizemos que temos pouca precisão, pois a chance de termos nossa estimativa "longe" da média populacional é grande.