ESTIMATIVA DO DESVIO PADRÃO [1]

TÁCITO SILVA [2]

SINOPSE.- É apresentado um estudo sobre a estimativa do desvio padrão, quanto à propriedade de não tendenciosidade, isto é, introduzindo um fator C_s que, multiplicado pela estimativa s_2, faz com que esta seja centrada, ou não tendenciosa.

Também é apresentada uma tabela de valores de C_s para ajustamento das estimativas quando o número de dados da amostra é, no máximo, igual a 25.

INTRODUÇÃO

Na análise estatística dos dados experimentais tem-se, freqüentemente, necessidade de estimar o desvio padrão e para tanto deve-se, entre as muitas operações que se podem efetuar com os dados amostrais, eleger aquele conjunto que dê a melhor estimação.

Um estimador será tanto melhor quanto em maior grau reúna as seguintes propriedades:

a) seja centrado, isto é, não tendencioso;
b) seja consistente;
c) seja eficiente, ou pelo menos suficiente.

Neste trabalho, analisa-se a primeira propriedade citada e apresenta-se uma tabela para ajustamento das estimativas, quando o número de dados não é superior a 25.

Considere-se uma variável x, com distribuição normal, sendo sua média μ e seu desvio padrão σ.

É uma distribuição contínua cuja função de densidade é:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

definida no intervalo $-\infty, +\infty$.

A função de distribuição correspondente é definida pela integral

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{X} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, d(x)$$

e corresponde à probabilidade de ocorrência de valores de x no intervalo $(-\infty, X)$ isto é, $P[x \leq X]$.

Os parâmetros dessa distribuição, sendo geralmente desconhecidos, têm de ser estimados.

Pelo método da máxima verossimilhança obtêm-se os seguintes estimadores:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X} = m \quad e$$

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

o último depende da estimação de $\overline{X}$.

Os estimadores obtidos por máxima verossimilhança não têm a propriedade de serem sempre não tendenciosos. No caso presente, o estimador

$$\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

e não tendencioso, visto que $E(\overline{X}) = \mu$.

Já o estimador

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

é um estimador tendencioso, pois $E(s) \neq \sigma$.

REVISÃO DE LITERATURA

Gomes (1963) mostra que, quando se consideram os dados observados, como uma amostra de todos os dados análogos existentes, ao se calcular a estimativa da variância, é preferível dividir a soma dos quadrados dos n desvios por n−1 e não por n.

Evidentemente, esse arranjo é feito para se eliminar a tendenciosidade da estimativa, pois equivale a multiplicar:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n} \quad \text{por} \quad \frac{n}{n-1};$$

assim,

$$E\left[\frac{n}{n-1}\right] s^2 = \frac{n}{n-1} E(s^2).$$

Como

$$E(s^2) = \frac{n-1}{n}\, \sigma^2,$$

tem-se:

$$\frac{n}{n-1} E(s^2) = \frac{n}{n-1} \cdot \frac{n-1}{n}\, \sigma^2$$

[1] Aceito para publicação em 8 jan 1973.

[2] Pesquisador em Agricultura, M.Sc., da Seção de Estatística do Instituto de Pesquisa Agropecuária do Centro-Oeste (IPEACO), Caixa Postal 151, Sete Lagoas, Minas Gerais, e bolsista do Conselho Nacional de Pesquisas.

$$E\left[\frac{n}{n-1}\right] = \sigma^2,$$

o que demonstra o acerto da operação.

Chacon (1955) demonstra, no desenvolvimento da distribuição de Helmert, que, se x é uma variável que segue a distribuição normal, com média μ e variância σ^2, $\frac{s}{\sigma}$ segue a distribuição de Helmert com média

$$C_2 = 1 - \frac{3}{4n} - \frac{7}{32n^2} \ldots\ldots$$

que tende a 1 (um) quando n cresce, e desvio padrão

$$\frac{1}{2n}\left(1 - \frac{1}{8n} - \frac{25}{128n^2} \ldots\ldots\right)$$

que tende a $\frac{1}{2n}$ ao mesmo tempo que a distribuição tende à normalidade.

A função de Helmert pode ser obtida facilmente, partindo-se da expressão $\frac{s^2 n}{\sigma^2}$, que tem distribuição de X^2 com n−1 graus de liberdade.

Fazendo-se

$$\mu = \frac{s}{\sigma} \quad e$$

$$X^2 = \mu^2 n,$$

$$f\left(\frac{s}{\sigma}\right) = \frac{2}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{n}{2}\right)^{\frac{n-1}{\sigma}} \mu^{n-2} e^{-\frac{\mu^2 n}{2\sigma^2}};$$

obtém-se daí que:

$$E\left(\frac{s}{\sigma}\right) = \left(1 - \frac{3}{4n} - \frac{7}{32n^2} \ldots\ldots\right) \quad e,$$

$$V\left(\frac{s}{\sigma}\right) = \frac{1}{2n}\left(1 - \frac{1}{4n} - \frac{1}{8n^2} \ldots\ldots\right).$$

Ainda, Chacon (1955), estudando a tendenciosidade do estimador

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

obteve o estimador $s_1 = \frac{s}{C_2}$, não tendencioso.

A tendenciosidade tende a zero à medida que n cresce.

Com efeito, $\frac{s\sqrt{n}}{\sigma}$ segue a distribuição X de Helmert com n−1 graus de liberdade.

Portanto,

$$E\left(\frac{s\sqrt{n}}{\sigma}\right) = C_2\sqrt{n} \quad e$$

$$E\left(\frac{C_2\sqrt{n}}{\sigma}\right) = C_2\,\sigma,$$

sendo que, ao crescer n, C_2 tende à unidade.

O estimador $s_1 = \frac{s}{C_2}$ será, então, não tendencioso, pois

$$E(s_1) = \frac{1}{C_2}\,E(s) = \frac{1}{C_2}\cdot C_2\,\sigma;$$

logo:

$$E(s_1) = \sigma.$$

Desenvolvimento Teórico

A estimativa da variância, obtida pelo método da máxima verossimilhança, é dada por

$$s^2 = \frac{\sum_{i=1}^{n}(Xi - \overline{X})^2}{n},$$

sendo $x \frown N(\mu, \sigma^2)$ e n o número de dados da amostra.

Tem-se, então,

$$s^2 n = \sum_{i=1}^{n}(Xi - \overline{X})^2 \quad e$$

$$\frac{s^2 n}{\sigma} = \sum_{i=1}^{n}\left(\frac{Xi - \overline{X}}{\sigma}\right)^2 - \left(\frac{\overline{X} - \mu}{\sigma\sqrt{n}}\right)^2,$$

onde

$$\sum_{i=1}^{n}\left(\frac{Xi - \overline{X}}{\sigma}\right)^2$$

tem distribuição de X^2 com n graus de liberdade e $\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2$ tem distribuição de X^2 com um grau de liberdade.

Pela propriedade da aditividade de X^2, conclui-se que $\frac{s^2 n}{\sigma^2}$ tem distribuição X^2 com n−1 graus de liberdade.

A função de X^2 é dada por

$$f(V_f) = \frac{1}{\Gamma(f/2)\,2^{f/2}}\,V_f^{\frac{f-2}{2}}\,e^{-\frac{V_f}{2}}$$

com f graus de liberdade.

Fazendo-se $V_f = \frac{s^2 n}{\sigma^2}$,

tem-se

$$f(s^2) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}\left(\frac{n}{2}\right)^{\frac{n-1}{2}} \frac{(s^2)^{\frac{n-3}{2}}}{(\sigma^2)^{\frac{n-1}{2}}}\,e^{-\frac{s^2 n}{2\sigma^2}}\,d(s^2),$$

que é a distribuição da estimativa da variância (s^2).

Daí obtém-se:

$$f(s) = \frac{2}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n}{2}\right)^{\frac{n-1}{2}} \frac{s^{n-2}}{\sigma^{n-1}} e^{-\frac{s^2 n}{2\sigma^2}} d(s),$$

isto é, a distribuição da estimativa do desvio padrão (s).

Para que s seja uma estimativa não tendenciosa de σ, é necessário que

$$E(s) \sigma.$$

Sabe-se, entretanto, que

$$E(s) = \frac{2}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n}{2}\right)^{\frac{n-1}{2}} \frac{1}{\sigma^{n-1}} \int_0^{\infty} s^{n-1} e^{-\frac{s^2 n}{2\sigma^2}} d(s).$$

Resolvendo-se a integral, tem-se

$$E(s) = \left(\frac{2}{n}\right)^{1/2} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma.$$

Logo, s é uma estimativa tendenciosa de σ e, para eliminar-se a tendenciosidade, utiliza-se o estimador s_1, sendo que

$$s_1 = \left(\frac{n}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\frac{n}{2}} s,$$

pois

$$E(s_1) = \left(\frac{n}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} E(s)$$

$$E(s_1) = \left(\frac{n}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \cdot \left(\frac{2}{n}\right)^{1/2} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma,$$

e finalmente,

$$E(s_1) = \sigma.$$

Chamando-se

$$\left(\frac{n}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = C$$

e determinando-se C para valores crescentes de n, a partir de 2, obteve-se uma tabela cujos valores são exatamente as recíprocas dos valores C_2 da tabela da ASTM (1951), reproduzida por Chacon (1955).

Sabe-se, entretanto, que

$$s^2 = \sum_{i=1}^{n} \left(\frac{X_1 - \bar{X}}{n}\right)^2,$$

sendo um estimador tendencioso de σ^2, na quase totalidade dos trabalhos, onde há necessidade da estimação de σ^2, utiliza-se o estimador s_2^2, sendo

$$s_2^2 = \frac{n}{n-1} s^2,$$

não tendencioso.

Tem-se, então,

$$s_2^2 = \sum_{i=1}^{n} \frac{(Xi - \bar{X})^2}{n-1}.$$

Considerando-se a expressão

$$s_2^2 = \frac{(n-1)}{\sigma^2},$$

que tem distribuição de X^2 com n−1 graus de liberdade, a distribuição de s_2^2 será, então,

$$f(s_2^2) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n-1}{2}\right)^{\frac{n-1}{2}} \frac{(s_2^2)^{\frac{n-3}{2}}}{(\sigma^2)^{\frac{n-1}{2}}} e^{-\frac{s_2^2 (n-1)}{2\sigma^2}} d(s_2^2),$$

Obtendo-se-daí:

$$f = (s_2) = \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n-1}{2}\right)^{\frac{n-1}{2}} \frac{s_2^{n-2}}{\sigma^{n-1}} e^{-\frac{s_2^2 (n-1)}{2\sigma^2}} d(s_2),$$

ou seja, a distribuição da estimativa do desvio padrão, quando se parte de uma estimativa não tendenciosa da variância (s_2^2).

Logo,

$$E(s_2) = \frac{2}{\Gamma\left(\frac{n-1}{2}\right)} \left(\frac{n-1}{2}\right)^{\frac{n-1}{2}} \frac{1}{\sigma^{n-1}} \int_0^{\infty} s_2^{n-1} e^{-\frac{s_2^2 (n-1)}{2\sigma^2}} d(s_2).$$

Resolvendo-se a integral, obtêm-se

$$E(s_2) = \left(\frac{2}{n-1}\right)^{1/2} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sigma,$$

isto é, s_2 é ainda uma estimativa tendenciosa de σ, pois

$$E(s_2) \neq \sigma.$$

A eliminação da tendenciosidade se faz, neste caso, utilizando-se o estimador

$$s_3 = \left(\frac{n-1}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} s_2,$$

sendo evidente que $E(s_3) = \sigma$.

Fazendo-se

$$\left(\frac{n-1}{2}\right)^{1/2} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = C_3,$$

estes valores encontram-se tabulados no Quadro 1, juntamente com os valores de C anteriormente referidos, e C_2 constantes da tabela da ASTM transcrita por Chacon (1955).

QUADRO 1. *Fatores para correção das estimativas do desvio padrão*

N	C_2[a]	$C = \frac{1}{C_2}$	C_3
2	0,5642	1,7725	1,2533
3	0,7236	1,3820	1,1284
4	0,7979	1,2533	1,0854
5	0,8407	1,1894	1,0638
6	0,8686	1,1512	1,0509
7	0,8882	1,1259	1,0424
8	0,9027	1,1078	1,0362
9	0,9139	1,0942	1,0317
10	0,9227	1,0837	1,0281
11	0,9300	1,0753	1,0253
12	0,9359	1,0684	1,0229
13	0,9410	1,0627	1,0210
14	0,9453	1,0579	1,0194
15	0,9490	1,0537	1,0180
16	0,9523	1,0501	1,0168
17	0,9551	1,0470	1,0157
18	0,9576	1,0442	1,0148
19	0,9599	1,0418	1,0140
20	0,9619	1,0396	1,0132
21	0,9638	1,0376	1,0126
22	0,9655	1,0358	1,0120
23	0,9670	1,0342	1,0114
24	0,9684	1,0327	1,0109
25	0,9696	1,0313	1,0104

[a] Fonte: American Society for Testing Materials (1951).

DISCUSSÃO

A tabela apresentada por Chacon (1955) pode ser utilizada na eliminação da tendenciosidade, quando o estimador de σ for s, isto é, tenha sido calculado pela fórmula:

$$s = \sqrt{\sum_{i=1}^{n} \frac{(Xi - \overline{X})^2}{n}}.$$

No entanto, na quase totalidade dos trabalhos no campo da pesquisa agropecuária, não se utiliza este estimador, mas sim, s_2, calculado por

$$s_2 = \sqrt{\sum_{i=1}^{n} \frac{(Xi - \overline{X})^2}{n-1}},$$

uma vez que s_2^2 é um estimador não tendencioso de σ^2. Assim sendo, os valores de C_3 podem ser utilizados para eliminação da tendenciosidade apenas multiplicando-se os valores s_2 pelos C_3 da tabela

Exemplo numérico:

Foram encontrados os seguintes valores em quilogramas, ao serem pesadas 16 crianças apresentando idades de 6 a 7 anos, nas Escolas Reunidas Anexas do IPEACO, em Sete Lagoas, MG: 19,7, 24,7, 28,3, 22,1, 23,0, 18,4, 19,5, 25,0, 20,9, 23,3, 23,8, 19,5, 20,2, 21,5, 18,6 e 22,1.

Qual será o intervalo de confiança para a média encontrada?

Com os estimadores s e s_2 obtêm-se:

$$s = 3,9745 \quad e \quad s_2 = 4,1049.$$

Ambas as estimativas são tendenciosas e, para eliminação desta tendenciosidade, usam-se os valores C e C_3 (Quadro 1), obtendo-se, assim, os estimadores $s_1 = \frac{s}{C_2}$ e $s_3 = s_2 . C_3$, ambos não tendenciosos, com n = 16.

Obtém-se, deste modo, $s_1 = 4,174$ e $s_3 = 4,174$, ou seja, $s_1 = s_2$ são estimativas não tendenciosas, obtidas por caminhos diferentes.

A partir deste valor $s_1 = s_3 = 4,174$, obtém-se o erro padrão da média:

$$s_{\overline{X}} = \frac{4,174}{\sqrt{16}} = 1,0435,$$

ao passo que $s_{\overline{X}} = \frac{4,10749}{\sqrt{16}} = 1,0262$ é uma estimativa tendenciosa de $\frac{\sigma}{\sqrt{n}}$.

O intervalo de confiança para a média terá por extremos 21,70 ± 2,21 (não tendenciosos) e 21,70 ± 2,17 (tendenciosos).

CONCLUSÕES

Do exposto se conclui:

1) as estimativas do desvio padrão, obtidas através dos estimadores s e s_2, são tendenciosas, pois $E(s) \neq \sigma$ $E(s_2) \neq \sigma$;

2) para eliminar-se a tendenciosidade, recorre-se aos valores de C_2 e C_3, obtendo-se:

$$s_1 = \frac{s}{C_2} \text{ e } s_3 = s_2 \times C_3,$$

com

$$C_2 = \left(1 - \frac{3}{4n} - \frac{7}{32n^2} \ldots\right) \text{ e}$$

$$C_3 = \left(\frac{n-1}{2}\right)^{\frac{1}{2}} \frac{\Gamma\left(\frac{n-1}{2}\right)}{\Gamma(n/2)};$$

3) quando o número de dados é grande ($n > 25$), a correção dos valores de s e s_2 se torna desnecessária, pois C_2 e C_3 se aproximam da unidade e os valores s e s_2 praticamente não se alteram ao se efetuarem os produtos $sx \frac{1}{c_2}$ e $s_2 \times C_3$.

É evidente que a tendenciosidade de s é maior que a de s_2 para um mesmo valor de n.

Referências

American Society for Testing Materials 1951. Manual on quality control of materials. Am. Soc. Test. Materials, 15 C tabela B, 2, p. 115. (Reproduzida por Chacon 1955)

Chacon, E.S.I. 1955. Curso de estatística. 2 vol. Editorial El Mensajero del Corazon de Jesus, Bilbao. 1019 p.

Gomes, F.P. 1963. Curso de estatística experimental. 2.ª ed. Esc. Sup. Agric. Luiz de Queiroz, Piracicaba, S. Paulo. 384 p.

Abstract.- Silva, T. [*The standard deviation estimative.*]. Estimativa do desvio padrão. *Pesquisa Agropecuária Brasileira, Série Agronomia* (1973) 8, 245-249 [Pt, en] IPEACO, Caixa Postal 151, Sete Lagoas, MG, Brazil.

This study deals with the unbiased proprieties of the standard deviation. The C_3 factor multiplied by s_2 estimative turns it an unbiased one. A table of C_3 values is presented, for estimative adjustments, when the number of data is 25 at the most.