Note on a Method for Calculating Corrected Sums of Squares and Products
Author(s): B. P. Welford
Source: *Technometrics*, Vol. 4, No. 3 (Aug., 1962), pp. 419-420
Published by: American Statistical Association and American Society for Quality
Stable URL: http://www.jstor.org/stable/1266577
Accessed: 17/03/2010 12:50

Notes

Note on a Method for Calculating Corrected Sums of Squares and Products

B. P. Welford

Imperial Chemical Industries Limited, Pharmaceuticals Division. Alderley Park, Macclesfield, Cheshire, England.

In many problems the "corrected sum of squares" of a set of values must be calculated i.e. the sum of squares of the deviations of the values about their mean. The most usual way is to calculate the sum of squares of the values (the "crude" sum of squares) and then to subtract a correction factor (which is the product of the total of the values and the mean of the values). This subtraction results in a loss of significant figures and if a large set of values is being handled by a computer, this can result in a corrected sum of squares which has many fewer, accurate significant figures than the computer uses in calculations.

Various alternative schemes are available to combat this. One method is to scale the values to an arbitrary origin which is approximately equal to the mean: if successful, this will reduce the loss in significant figures. An alternative method is to first calculate the mean and then sum the powers of the deviations from the mean. This involves each value being considered twice: first in evaluating the mean and then when calculating its deviation from the mean. If the set of values is large and is being handled by a computer this can involve either storing the data in a slow speed store or reading the same data into the computer twice. A third method which is less cumbersome than either of these is outlined below.

The basis of the method is an iteration formula for deriving the corrected sum of squares for n values from the corrected sum of squares for the first $(n - 1)$ of these. We are given a set of x_i's $(i = 1, \cdots, k,)$ for which we require the corrected sum of squares.

$$S = \sum_{i=1}^{k} (x_i - \bar{x})^2 \quad \text{where} \quad \bar{x} = \sum_{i=1}^{k} x_i/k$$

We define

$$m_n = 1 \sum_{i=1}^{n} x_i/n \qquad n = 1, \cdots, k$$

and

$$S_n = \sum_{i=1}^{n} (x_i - m_n)^2 \qquad n = 1, \cdots, k$$

Thus

$$S_k = S$$

The following identities hold:—

$$m_n = \frac{n-1}{n} m_{(n-1)} + \frac{1}{n} x_n \tag{1}$$

For

$$i < n, \qquad x_i - m_n = x_i - m_{(n-1)} - \frac{1}{n}(x_n - m_{(n-1)}) \tag{2}$$

$$x_n - m_n = \frac{n-1}{n}(x_n - m_{(n-1)}) \tag{3}$$

for $n = 1, 2, \cdots, k$

$$\therefore\ S_n = \sum_{i=1}^{n} (x_i - m_n)^2$$

$$= \sum_{i=1}^{n-1} \left[(x_i - m_{(n-1)}) - \frac{1}{n}(x_n - m_{(n-1)}) \right]^2 + \left(\frac{n-1}{n}\right)^2 (x_n - m_{(n-1)})^2$$

$$= \sum_{i=1}^{n-1} (x_i - m_{(n-1)})^2 + \left[\frac{n-1}{n^2} + \frac{(n-1)^2}{n^2}\right](x_n - m_{(n-1)})^2$$

$$= S_{(n-1)} + \left(\frac{n-1}{n}\right)(x_n - m_{(n-1)})^2 \qquad I$$

Using this formula, the corrected sum of squares is computed using for each value its deviation from the mean of all the previous values. At no stage are significant figures lost and each value is only used once and need not be stored. Formula I is similar in form to that derived by Box and Hunter (1959) for computing the change in residual sum of squares after each cycle of an 'evolutionary operation' design.

A similar formula can be derived for calculating corrected sums of products, viz.

$$S_n = \sum_{i=1}^{n} (x_i - m_n)(y_i - m_n')$$

$$= S_{(n-1)} + \left(\frac{n-1}{n}\right)(x_n - m_{(n-1)})(y_n - m'_{(n-1)}) \qquad II$$

where

$$m_n' = \sum_{i=1}^{n} y_i/n$$

Similar formulae to I can also be derived for the iterative calculation of corrected sums of higher powers of values although these are a little more complex since they involve the corrected sums of the lower powers. However, since these corrected sums would usually be required for calculating moments about means and one would generally require all moments up to a certain order, this would not be a drawback.

Define

$$S_n^{(r)} = \sum_{i=1}^{n} (x_i - m_n)^r \qquad n = 1, \cdots k,$$

and $S_k^{(r)} = S^{(r)}$, the required corrected sum of rth powers of deviations about the mean.

Then, using identities (2) and (3), it can be shown that

$$S_n^{(r)} = \sum_{j=0}^{r} {}^rC_j (1/n)^{r-j} (m_{(n-1)} - x_n)^{r-j} S_{(n-1)}^{(j)} + \left(\frac{n-1}{n}\right)^r (x_n - m_{(n-1)})^r \qquad III$$

Remembering that

$$S_n^{(0)} = n \quad \text{for} \quad n = 1, \cdots, k$$

$$S_n^{(1)} = 0 \quad \text{for} \quad n = 1, \cdots, k$$

it is easily verified that III reduces to I when $r = 2$.

Reference

G. E. P. Box and J. S. Hunter, "Condensed calculations for Evolutionary Operation Programs," *Technometrics* Vol. 1, No. 1, February 1959, pp. 77–95.