UPDATING STATISTICS OF DATA

Eliezer A. Albacea[1]

1. Introduction

Most real-time data collection systems require that "running" statistics like the mean, standard deviation, correlation coefficient, and regression coefficient be computed as data accumulate. In this type of data collection system, the main concern is response time; thus, recomputing statistics from raw data as soon as new data arrive is most undesirable. Hence, such systems use the values of the statistics of previous data to compute the statistics which include the data that just came in.

Several authors have tried to solve this problem. Hanson (1975) presented an elegant and stable method of updating the mean and standard deviation of a weighted series of numbers. The metnod used by Hanson was based on matrix formulation and Givens transformation. Cotton (1975) in his remark to Hanson's paper, suggested a straightforward method for the case of unit weights. Cotton's suggestion was to make use of the elementary formula for computing the mean and standard deviation of data. Both authors presented a method of recomputing the new mean and standard deviation as one additional observation arrives.

It is unnecessary to recompute the statistics of data as soon as one or more observations arrive. Hence, this paper will present a method of recomputing the mean and standard deviation based on previous statistics, and the sum and number of observations of the newly arrived data. Moreover, the method will be extended to include a method of updating the correlation coefficient and regression coefficient for a two-variable relationship.

2. The Bases for Computing Algorithms

Following Cotton's suggestion, one can derive a formula for recomputing the mean, standard deviation, correlation coefficient and regression coefficient based on the previous value of the said statistics as soon as one observation comes in. The formulas are then extended to the case where $M > 1$ number of observations are added to the previous data.

1/ Instructor in Computer Science, Institute of Mathematical Sciences and Physics, UP at Los Banos, College, Laguna 3720, Philippines.

2.1 Mean

Let:

$$\mu = \sum_{i=1}^{N} X_i/N \qquad (2.1.1)$$

represent the mean of N observations or data points, x_1, x_2, x_3, ..., x_N. If a single observation is added to the data set, then the mean of the N + 1 points is computed as:

$$\mu' = \sum_{i=1}^{N+1} x_i/(N+1) \qquad (2.1.2)$$

However, simple algebraic manipulations lead to the following result:

$$(N+1)\mu' = \sum_{i=1}^{N+1} x_i = \sum_{i=1}^{N} x_i + X_{N+1}$$

which then gives a computational formula for the new mean which is a function of the mean of N observations and the new data point; that is

$$\mu' = (n\,\mu + x_{N+1})/(N+1)$$

If M > 1 data points are added to the data set, the mean of the N + M observations is given by:

$$\mu'' = \sum_{i=1}^{N+M} x_i/(N+M) \qquad (2.1.3)$$

Again, simple algebraic manipulations provide the following results:

$$(N+M)\mu'' = \sum_{i=1}^{N+M} x_i = \sum_{i=1}^{N} x_i + \sum_{i=N+1}^{N+M} x_i = N\mu + \sum_{i=N+1}^{N+M} x_i$$

This leads to the computational formula

$$\mu'' = (N\mu + \sum_{i=N+1}^{N+M} x_i)/(N+M) \qquad (2.1.4)$$

which allows one to compute the mean of N + M observations from the sample mean of the original N data points and the sum of the additional M data points.

2.2 Standard Deviation

The standard deviation and sum of squares, respectively, of N data points can be computed as:

$$\sigma = ((\sum_{i=1}^{N} x_i^2 /N) - \mu^2)^{1/2}$$

and

$$\sum_{i=1}^{N} x_i^2 = N(\sigma^2 + \mu^2) \qquad (2.2.1)$$

where μ is as defined in (2.1.1)

An additional observation to N data points leads to an expression for the standard deviation, as follows:

$$\sigma' = ((\sum_{i=1}^{N+1} x_i^2 /(N+1)) - \mu'^2)^{1/2}$$

where μ is as defined in (2.1.2).

This can also be expressed as:

$$\sigma' = (((\sum_{i=1}^{N} x_i^2 + x_{N+1}^2)/(N+1)) - \mu'^2)^{1/2}$$

Substituting $N(\sigma^2 + \mu^2)$ for $\sum_{i=1}^{N} x_i^2$ gives the following computational formula:

$$\sigma' = (((N(\sigma^2 + \mu^2) + x_{N+1})/(N+1)) - \mu'^2)^{1/2}$$

which gives the standard deviation of N + 1 data points as a function of the standard deviation and mean of N data points, the square of the additional observation and mean of N+1 observations.

In cases where more than one observation is added to the data set or in general, M observations, the expression for standard deviation is given by:

$$\sigma'' = ((\sum_{i=1}^{N+M} x_i^2 /(N+M)) - \mu''^2)^{1/2}$$

Where μ'' is as defined in (2.1.3). This expression is equivalent to

$$\sigma'' = (((\sum_{i=1}^{N} x_i^2 + \sum_{i=N+1}^{N+M} x_i^2)/(N+M)) - \mu''^2)^{1/2} \qquad (2.2.2)$$

Using equation (2.2.1) a computational formula for the standard deviation of N+M data points is given by:

$$\sigma'' = (((N(\sigma^2 + \mu^2) + \sum_{i=N+1}^{N+M} x_i^2) /(N+M)) - \mu''^2)^{1/2}$$

Note that the resulting expression for standard deviation is now a function of the standard deviation and mean of N points, sum of squares of M additional points and the mean of N+M data points.

2.3 Correlation Coefficient

In cases where the data consist of more than one variable, a measure of the relationships among variables may be desired. We now outline a method which can be used in recomputing the correlation coefficient between variables.

The correlation coefficient between two variables X and Y where each variable has N observations $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$ given by:

$$r = \frac{\left(\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i /N \right)}{\left(\left(\sum_{i=1}^{N} x_i^2 - \left(\left(\sum_{i=1}^{N} x_i \right)^2 /N \right) \right) \left(\sum_{i=1}^{N} y_i^2 - \left(\left(\sum_{i=1}^{N} y_i \right)^2 /N \right) \right) \right)^{1/2}}$$

or

$$r = \left(\sum_{i=1}^{N} x_i y_i - \left(\left(\sum_{i=1}^{N} x_i \right) \left(\sum_{i=1}^{N} y_i \right)/N \right) /N \right) /N\sigma_x\sigma_y \qquad (2.3.1)$$

Where σ_x and σ_y are standard deviations of the observations. Doing some manipulations on equation (2.3.1) will lead us to a simple expression for the sum of cross products between X and Y. This expression is:

$$\sum_{i=1}^{N} x_i y_i = Nr\sigma_x\sigma_y + N\mu_x\mu_y$$

where μ_x and μ_y are the means of the observations. The resulting expression for the sum of cross products between X and Y is a function of the correlation coefficient between X and Y, the standard deviations of variables X and Y and the means of variables X and Y, all based on N data points.

Suppose a pair of observations is added, that is, one observation for X and one for Y. The correlation coefficient of the N+1 (X,Y) pairs is then equal to

$$r' = \left(\sum_{i=1}^{N+1} x_i y_i - \left(\left(\sum_{i=1}^{N+1} x_i \right) \left(\sum_{i=1}^{N+1} y_i \right)/(N+1) \right) \right)/(N+1)\sigma'_x\sigma'_y \qquad (2.3.2)$$

Where σ'_x and σ'_y are standard deviations of variables X and Y with N+1 data points, respectively. Eliminating all terms involving summation in equation (2.3.2) leads to the following computational formula:

$$r' = ((Nr\sigma_x\sigma_y + N\mu_x\mu_y + x_{N+1}y_{N+1}) - ((N\mu_x + x_{N+1})(N\mu_y + y_{N+1}))/(N+1)))/(N+1)\sigma'_x\sigma'_y$$

Note that this expression for the correlation coefficient of N + 1 data points is now a function of the correlation coefficient of and the means of variables X and Y based on N data points, standards deviations of variables X and Y both on N and N + 1 data points and the additional pair of observations as well as its cross product.

In cases where M observations are added to the data set, the correlation coefficient between X and Y can be written as:

$$r'' = \left(\sum_{i=1}^{N+M} x_iy_i - \left(\left(\sum_{i=1}^{N+M} x_i\right)\left(\sum_{i=1}^{N+M} y_i\right)/(N+M)\right)\right)/(N+M)\sigma''_x\sigma''_y \qquad (2.3.3)$$

where σ''_x and σ''_y are the standard deviations of variables X and Y respectively using N+M observations. Equation (2.3.3) can be simplified to the following computational formula:

$$r'' = \left(Nr\sigma_x\sigma_y + N\mu_x\mu_y + \sum_{i=N+1}^{N+M} x_iy_i\right) - \left(\left(N\mu_x + \sum_{i=N+1}^{N+M} x_i\right)\left(N\mu_y + \sum_{i=N+1}^{N+M} y_i\right)/(N+M)\right)\Big)/(N+M)\sigma''_x\sigma''_y \qquad (2.3.4)$$

where the correlation coefficient between X and Y based on N+M data points is now a function of the standard deviations of variables X and Y based on N and N+M data points, means of variables X and Y based on N data points and the sum of the additional M observations in each variable as well as the sum of cross products of the additional M observations.

2.4 Regression Coefficient

The regression coefficient of Y on X, where each variable has N data points, is given by the expression:

$$b = \left(\sum_{i=1}^{N} x_iy_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i/N\right) / \left(\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2/N\right)$$

which can be simplified to

$$b = \left(\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i /N \right) / N\sigma_x^2 \qquad (2.4.1)$$

where σ_x^2 is the variance of variable X with N data points. Further manipulations of equation (2.4.1) lead to a simple expression for the sum of cross products of X and Y based on N data points. This resulting expression is the following:

$$\sum_{i=1}^{N} x_i y_i = Nb\sigma_x^2 + N\mu_x\mu_y \qquad (2.4.2)$$

Using equation (2.4.2), we can derive an expression for the regression coefficient of Y on X for N+1 pairs of observations. This expression is:

$$b' = \left(\sum_{i=1}^{N+1} x_i y_i - \sum_{i=1}^{N+1} x_i \sum_{i=1}^{N+1} y_i /(N+1) \right) /(N+1)\sigma_x'^2 \qquad (2.4.3)$$

where $\sigma_x'^2$ is the variance of variable X based on N+1 data points. Algebraic manipulation of equation (2.4.3) will lead us to an expression for the regression coefficient of N+1 data points as a function of the regression coefficient of Y on X based on N data points, means of variable X and Y based on N data points, variance of X based on N+1 data points and the additional pair of observations. The resulting expression is shown as follows:

$$b' = ((Nb\sigma_x 2 + N\mu_x\mu_y + x_{N+1}\, y_{N+1}) - ((N\mu_x + x_{N+1})(N\mu_y + y_{N+1}))/(N+1)))/(N+1)\sigma'^2_x$$

In cases where more than one pair of observations, say M, are added to the data set the regression coefficient of Y on X can be written as:

$$b'' = \left(\sum_{i=1}^{N+M} x_i y_i - \sum_{i=1}^{N+M} x_i \sum_{i=1}^{N+M} y_i / (N+M) \right) / (N+M)\sigma''^2_x$$

where σ''^2_x is the variance of X based on N+M observations. Simplifying this further will result to the following computational formula:

$$b'' = ((Nb\sigma_x 2 + N\mu_x\mu_y + \sum_{i=1}^{N+M} x_i y_i) - ((N\mu_x + \sum_{i=N+1}^{N+M} x_i)(N\mu_y + \sum_{i=N+1}^{N+M} y_i))/(N+M)))/(N+M)\sigma''^2_x \qquad (2.4.4)$$

Note that this produces an expression where the regression coefficient of Y on X for N+M pairs of observations is a function of the variance of X and regression coefficient based on N pairs of observations, variance of X based on N+M data points, means of variables X and Y based on N data points and sum of the M additional observations in each variable.

3. Implementation

The above algorithms have been implemented in the programming language Pascal and in C. Readers interested in the source code may communicate with the author.

Observe that equations (2.1.4), (2.2.2), (2.3.4) and (2.4.4) can easily be modified to include the situation where instead of additional observations, several observations are taken away from the original data.

4. Concluding Remarks

One consequence of extending the formula to the case of more than one additional observation is the reduction of computing time. In real-time data collection systems, the computer should be able to catch incoming data. If the computer is busy computing statistics and new data come in, there is a great possibility that some of the data that come in will be lost. However, most of these systems are provided with interrupt mechanisms whereby catching incoming data are given priority. Only when no data come in when the systems spend time computing statistics. But usually data come almost continuously in this type of data collection systems, and therefore very little time is allocated to other functions. So it might happen, especially when data keep coming in, that no statistics will be produced in long periods of time. Thus computing statistics from raw data is completely out of the question. With the reduction of time requirement for computing statistics, it is therefore possible to get an updated statistics more often and get values which are more representative of the data collected at a particular instant.

In addition, two sets of data can be combined or a set of data can be reduced and statistics can still be computed without recomputing it from raw data. The former situation always happens with time series data where data come in time intervals and statistics sometimes have to be recomputed.

5. References

Hanson, R.J. (1975). Stably updating mean and standard deviation of data Communications of the ACM. 8, 57-58.

Cotton, I.W. (1975). Remark on stably updating mean and standard deviation of data. Communications of the ACM. 18, 458.