COMPUTATION OF VARIANCE AND CORRELATION COEFFICIENT

Engr. Imran Anwar Ujan
Lecturer, Department of Statistics
University of Sindh, Jamshoro

Raja M.Ilyas
Associate Professor, Department of Statistics
University of Sindh, Jamshoro
and

Azizullah Memon
Associate Professor, Department of Statistics
University of Sindh, Jamshoro

ABSTRACT

Analysing data is not a straight forward procedure like solving a quadratic equation. Data are investigated in various ways. The tools used are: tables, graphs, summary statistics and statistical tests. The data may be modified, as the analysis proceeds, by rejecting some items or by the use of mathematical transformation. Thus computer program in which there is a fixed sequence of algorithms is not likely to be convenient statistical analysis. We here suggest a form of program that allows an interactive type of analysis.

1. VARIANCES BY METHODS ON STANDARD FORMULAE

Text on statistics usually gives two alternative and equivalent formula for calculating the estimated population variance S^2, based on sample of n observation:

$$S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2/(n-1); \qquad (1)$$

$$S^2 = (\sum_{i=1}^{n} x_i^2 - (\sum x_i)^2/n)/(n-1) \qquad (2)$$

Note that we are using a divisor(n-1). This is appropriate when estimating the variance of a population, given a sample of n observations from the population, and is what we usually require. If the n observations do actually constitute a complete population, the variance is given by similar formula in which n replaces (n-1). Formula (1) is the direct definition of variance, while in Formula (2) the sum of squares $\Sigma(x_i - \bar{x})^2$ has been expanded and rearranged in a form that is often more convenient for desk calculation. The term $(\Sigma x_i)^2/n$ is called a correction term and the numerator $\Sigma x_i^2 - (\Sigma x_i)^2/n$ is called a corrected sum of squares.

Taking the square root of the variance gives the standard deviation.

Suppose we wish to estimate the variance from a sample consisting of the numbers 1000, 1003, 1006, 1007, 1009, using a computer which express floating point numbers corrected to 6 decimal digits. In table 1 we follow through the calculation of the variance using formula (2), assuming (a) that the numbers are truncated, or (b) that they are rounded, in order to fit into the space available for each number. We give also the exact calculation; and we see that the result is disconcerting. The errors in the result (7.5 or 15, instead of the correct value of 12.5) are substantial; and other sets of numbers could be chosen that would give even more dramatic errors. An operator using a desk calculator would immediately 'code' the data by subtracting 1000 from each observation before calculating a variance, and this would avoid the error. But one cannot rely on data being first scrutinised like this when a computer is used. We must conclude that the method based on formula (2) can be unreliable and so should not be used.

Table 1

	EXACT	TRUNCATED	ROUNDED
x_1^2	1000000	100000 E1	100000 E1
x_2^2	1006009	100600 E1	100601 E1
x_3^2	1012036	101203 E1	101204 E1
x_4^2	1014049	101404 E1	101405 E1
x_5^2	1018081	101808 E1	101808 E1
$A = \sum x^2$	5050175	505015 E1	505018 E1
$(\sum x^2)$	25250625	252506 E2	252506 E1
$B = (\sum x)^2)/5$	5050125	505012 E1	505012 E1
A-B	50	3 E1= 30	6 E1 = 60
VARIANCE, (A-B)/4	12.5	7.5	1.5

We recommend the two-pass deviation method as the best method for general use, An improvement, in accuracy, in situations where the standard deviation is small compared with the mean, may be obtained by making correction to the sum of squares of deviation. The formula is:

$$S^2 = \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 - \left(\sum_{i=1}^{n} x_i - \bar{x}\right)^2 /n\right) /(n-1).$$

The correction is zero in exact compuation, but in normal computation is a good approximation to the error in the first term.

There may be situations in which it is considered important to calculate the variance in a single pass. We give a method of doing this discuss the merits of alternative methods of calculating the variance.

2. AN UPDATING PROCEDURE FOR CALCULAING MEAN AND VARIANCE

This method updates the value of the mean and the sum of squares of deviation as each observation is introduced into the calculation.

Write m_i for mean the first i observations $\left(\sum_{r=i}^{i} Xr/i\right)$ and S_i for the sum of squares of deviation of the first i observation about their mean $\left[\sum_{r=1}^{i} (X_r - m_i)^2\right]$.

3. PROOF OF RECURRENCE RELATIONS FOR MEAN AND VARIANCE

We require to prove the two relations given below.

$$m_i = [(i-1)m_{i-1} + X_i)]/i;$$

$$S_i = S_{i-1} + (i-1)\,(X - m_{i-1})^2/i.$$

The proof of that for m_i is straight forward. If we multiply both sides of the equation by i, we have $im_i = (i-1)m_{i-1} + X_i$

Each side is now the sum of the i observation, since

$$im_i = \sum_{r=1}^{i} X_r \text{ and } (i-1)\, m_{i-1} = \sum_{r=1}^{i-1} X_r$$

Before proving the relation for S_i note that

$$\begin{aligned} i(m_i - m_{i-1}) &= im_i - m_{i-1} \\ &= (i-1)m_{i-1} + X_i - im_{i-1} \text{ (from the result for means)} \\ &= X_i - m_{i-1} \end{aligned}$$

Now

$$S_{i-1} = \sum_{r=1}^{i-1} (X_r - m_{i-1})^2 = \sum_{r=1}^{i} (X_r - m_{i-1})^2 - (X_r - m_{i-1})^2$$

$$= \sum_{r=1}^{i} [(Xr - m_i)^2 + (m_i - m_{i-1})^2] - (X_i - m_{i-1})^2$$

$$= \sum_{r=1}^{i} (Xr - m_i) + 2\sum_{r=1}^{i} (m_i - m_{i-1})^2 - (Xi - m_{i-1})^2$$

Here the cross-product term disappears since

$$= \sum_{r=1}^{i} (Xr - m_i)(m_i - m_{i-1}) = (m_i - m_{i-1}) \sum_{r=1}^{i} (Xi - m_{i-1})$$

$$= (m_i - m_{i-1})X\ 0 = 0$$

Thus

$$S_{i-1} = S_i + \sum_{r=1}^{i} (X_i - m_{i-1})^2/i^2 - (X_i - m_{i-1})^2$$

$$= S_i + (X_i - m_{i-1})^2/i - (X_i - m_{i-1})^2$$

$$= S_i - (i-1)(X_i - m_{i-1})^2/i$$

Which leads at once the result.

4. CALCULATION OF SUMS OF SQUARE AND PRODUCT OF TWO VARIATES

To Calculate the product-moment correlation coefficient of two varietes X and Y, we need the correct sums of squares.SSx = $\Sigma(x_i - \bar{x})^2$ and SSy = $\Sigma(y_i - \bar{y})^2$, and the corrected sum of products SPxy =$\Sigma(x_i - \bar{x})^2 (y_i - \bar{y})^2$.

To extend the updating method, we need to find a recurrence relation for the sum of products. By analogy with the definition and relation for s_i, we define the sum of products of i pairs of observation by i

$$P_i = \sum_{r=i} (x_r - m_{1,i})(y_i - m_{2,i})$$

and it can be shown that

$$P_i = P_{i-1} + (i-1)(x_i - m_{1,i-1})(y_i - m_{2,i-1})/i$$

5. CALCULATION OF PRODUCT-MOMENT CORRELATION COEFFICIENT

To obtain the correlation coefficient r is a one-line calculation when we have the results of an algorithm for finding sums of squares and products;

```
r = SPxy / sqrt (SSx * SSy);
```

It is always wise, before calculating a correlation coefficient, to plot a scatter diagram to check whether the points representing the data pairs (x_i, y_i) are approximately linearly placed, or whether there is any systematic curved trend.

6. CONCLUSION

The essence of the method is that program offers a list of procedures, commonly called a menu, of which one is chosen, carried out and followed by a return to the menu. Each time the menu is consulted another procedure may be chosen or the analysis stopped. We present in Program, Mean Var Package, two procedures are offered: (1) input of data; (2) calculation of mean and variance. The program might be used to calculate the mean and variance of each of a number of data.The program should be self-explanatory.

REFERENCES

1. Abramowitz, M, and Stegum, I A, Handbook of Mathematical Functions, Dover, New York (1972)

2. Ashby, T, A modification to paulson's approximation to the variance ratio distribution. The Computer Journal, 11, 209-10 (1968).

3. Chan, TF, Golub, Gh, and Leveque, RJ, Algorithms of computing the sample variance: analysis and recommendations. American Statistics, 37,238-43 (1980)

4. Paulson, E, An Approximate normalization of the analysis of variance distribution, Annals of Mathematical Statistics, 13, 233-5 (1942).

5. Turkey, JW, Exploratory Data Analysis, Addision-Wesley, Reading, Mass (1977).

```
Program Mean VarPackage (input,output);
    Const   MaxSampleSize = 50;
            Noofprocedure =2;
    Type    Characterset = set of char;
            Unit = 1 .. MaxSampleSize;
            DataVector= array [units] of real;
    Var     Choice:integer;
            Mean real;
            n: Units;
            ReplySet: CharacterSet;
            Response: Char;
            Variance: Real;
            WantMenu: Boolean;
            x: Data Vector;
            Procedure PrintMenu (Var choice :integer);
            Var Response: char;
            begin
               repeat
               write ('Type number of procedure
               required and Return');
               readln (Choice);
               end;
procedure inputData (Var x: DataVector: Var n:Units);
    Var i : integer;
    begin
    Write ('State number of observation');
    readln(n);
    Writeln('Input observation separated by spaces');
    Writeln('Check data on each line before pressing
    Return');
    for i:= 1 to n do
    read(x[i]);
    readln;
    end;
Procedure FindMean Vari(x: Data Vector; n Units;
                        Var Mean, Variance:real);
          Var     Dev:real;
                  i: integer;
                  Sum: real;
          begin
                  Sum:= 0.0;
                  for i:= 1 to n do
                  Sum:= Sum+x[i];
                  Mean:= Sum/n;
                  Sum:=0.0;
                  begin
                  Dev:=x[i]-Mean;
                  Sum:=Sum+Dev*Dev
                  end;
                  Variance:= Sum/(n-1)
                  end;
```

```
begin
ReplySet:=['Y','y','n','N'];
repeat
repeat
Writeln;writeln('Do you want the Menu?');
Write('Type Y for yes or n for no and Return');
          readln(response)
    Until Response in ReplySet;
    if want Menu:= Response in ['Y','y'];
    then begin
           PrintMenu (Choice);
           Case choise of
           1: InputData (x,n);
           2: begin

          FindMenuVari(x,n,Mean,Variance);
          Writeln('Number of observation=',n);
          Writeln('Mean = ',Mean);
          Writeln('Variance = ',Variance);
          end
          end;
          end;
          until not WantMenu;
   end.
```

```
C:\>cd tp
C:\>tp>cd bin
C:\TP\BIN>turbo
Turbo Pascal Version 7.0 Copy right (c) 1983,92,
Borland International
Do you want to the Menu?
Type y for yes and n for no and return y
Type number of procedure required and return I
state number of observation = 5
Input observation separated by spaces
Check data on each line before pressing Return
1.0  2.0  3.0  5.0  9.0
Do you want to the Menu?
Type y for yes and n for no and return y
Type number of procedure required and return 2
state number of observation = 5
Mean        = 4.00000000000E+00
Variance    = 6.25000000000E+00
Do you want to the Menu?
Type y for yes and n for no and return.
```