Journal of Computer-Generated Mathematics

A new simple numerical method for least squares approximation of a sample by a continuous probability distribution

Deko Dekov

Abstract. In this paper we offer a new simple numerical method for approximation of a sample by a continuous probability distribution. The method is suitable for use in high schools and colleges.

Keywords: Normal distribution, approximation of a sample, least squares method

Given a sample, the sample mean and the sample standard deviation define a normal distribution, which we call the standard normal approximation to the sample.

In this paper we offer a new simple numerical method, suitable for high schools and colleges. The method allows students to find the least squares normal approximation to a sample. But, if we want to receive the answer immediately, we have to use a computer program. The method is first described in [1], and in this paper we show that the method is applicable to the problem of finding the least squares normal approximation to a sample.

The numerical method, described in this paper, has the following advantages. The method uses only the definition of a function, so that the school and college students could understand and use it without studying. The method does not use the Gauss approach for finding the minimum of the objective function, so that the method does not require preliminary studying of partial derivatives and extrema of functions of many variables. The method is fast, because it needs small numbers of iterations. Since each iteration adds one true digit to the answer, we need only 100 iterations to receive an answer with 100 true digits. We receive the answer for less than 1 second, if we use a desktop personal computer.

The method, in more general framework, is as follows. Suppose data consisting of n points (x1,y1),(x2,y2),…,(xn,yn) are known and the goal is to find a function y = F(x) that fits the data reasonably well. We will use the least squares criterion. We suppose that the reader is familiar with the least squares criterion. Suppose that f(x) is the objective function, so that we have to find the minimum of f(x). We use the data set, in order to localize the minimum of f(x). Suppose that the minimum of the function f(x) is within the

segment [a,b]. We divide the segment [a,b] by N equal parts by using the points x0 = a, x1, x2 , ... , xN = b. Then we evaluate f(x0), f(x1), f(x2), ... ,f(xN), and select the minimal of these values. We use the minimal value as the midpoint of a new segment, whose length is 10 times smaller than the previous segment. The process is repeated until the minimum is found. The method works well if $N \geq 100$, but in many cases it is enough we to set smaller N.

The extension of the method to the case when the objective function has two variables is straightforward.

The method is simple, so that it allows a simple implementation. I have created a simple computer program by using PHP. The program could record the calculations, made by the computer. The file containing record of calculations for the below example is available for download as supplementary material.

Example. A sample containing the heights of 40 persons is given in the following table:

Height	Middle of the interval	Sample frequency	Sample relative frequency
150-160	155	5	0.125
160-170	165	15	0.375
170-180	175	16	0.4
180-190	185	4	0.1

Find the least squares approximation of the sample by the normal distribution.

Solution. First, we have to normalize the sample relative frequency. In order to normalize the sample relative frequency, we divide any sample relative frequency by the length of the corresponding interval. In this example, the length of all intervals is equal to 10, so that we divide each sample relative frequency by 10.

By using the computer program, which implements the above described method, we obtain the expected value μL and the standard deviation σL of the least squares normal approximation $N(\mu L,\sigma L)$ to the sample. The computer program finds μL and σL as the solution to the following problem. Find μ and σ, which minimize the function

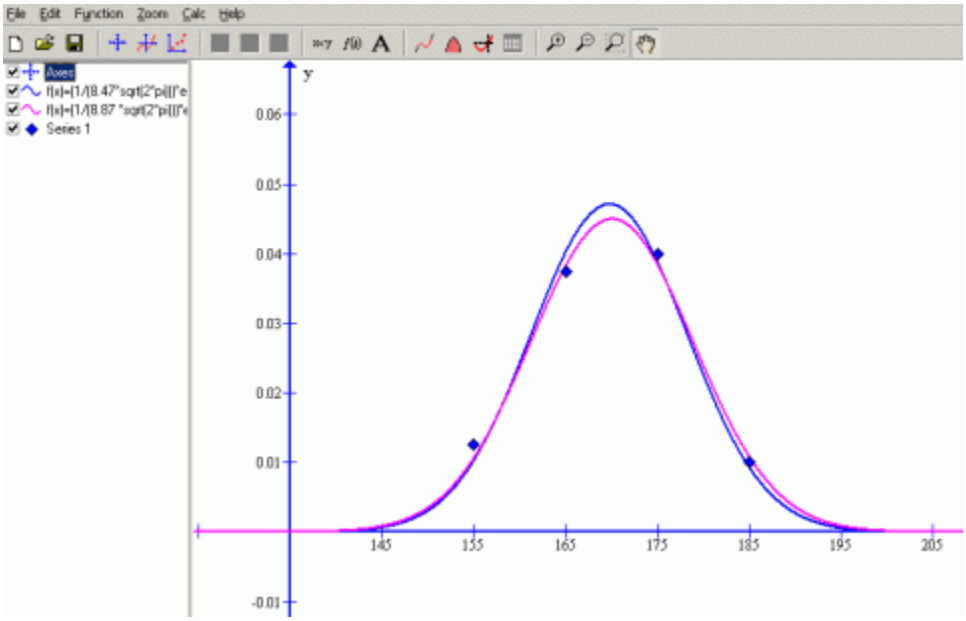
$$f(\mu,\sigma)=\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(155-\mu)^2}{2\sigma^2}}-0.0125\right)^2+\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(165-\mu)^2}{2\sigma^2}}-0.0375\right)^2$$
$$+\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(175-\mu)^2}{2\sigma^2}}-0.04\right)^2+\left(\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(185-\mu)^2}{2\sigma^2}}-0.01\right)^2$$

where μ and σ are respectively the expected value and the standard deviation of the least squares normal distribution. Set N = 10. As initial intervals for μ and σ we take the intervals $[\mu S-5, \mu S+5]$ and $[\sigma S-5, \sigma S+5]$, where $\mu S = 170$ and $\sigma S = 8$ are rounded values of the sample mean and the sample standard deviation, respectively. If we require 5 digits

after the decimal point, we receive the following answer: μL = 170.04786 and σL = 8.86524. Hence, the probability density function of the least squares normal distribution is as follows:

$$N(x) = \frac{1}{8.86524\sqrt{2\pi}} e^{-\frac{(x-170.04786)^2}{2(8.86524)^2}}$$

We could use the Ivan Johansen's computer program Graph to draw the graph of the normalized sample - the blue rhombs, the graph of the standard normal distribution defined by μS and σS – the blue curve (for the graph we take μS = 169.75 and σS = 8.47), and the graph of the least squares normal distribution defined by μL and σL – the red curve (for the graph we set μL = 170.05 and σL = 8.87):



The normal approximation, obtained by the least squares method, always better fits the sample than the standard normal approximation.

The computer program gives us also the sums of squares. For the above example, we obtain the following results: The sum of squares in the case of the standard normal approximation is equal to R = 0.00001398, and the sum of squares in the case of the least squares normal approximation is equal to L = 0.000006964. The quotient R / L is approximately equal to 2, that is, the least squares normal approximation about 2 times better fits the sample.

We could record the calculations, made by the computer. For the above example, the file containing record of calculations is available for download as supplementary material.

References

1. Deko Dekov, A numerical method for solving the horizontal resection problem in Surveying, Journal of Geodetic Science (to appear).

Dr.Deko Dekov
Zahari Knjazeski 81
6000 Stara Zagora
Bulgaria
Submitted on 1 October 2011
Publication date: 1 February 2012
Revised publication: 10 April 2012