Journal of Computer-Generated Mathematics

The least squares normal approximation to the binomial distribution

Deko Dekov

Abstract. We offer a new simple numerical method for approximation of a discrete probability distribution by a continuous probability distribution. We illustrate the method by the least squares normal approximation to the binomial distribution. The method is designed for use in high schools and colleges.

Keywords: normal approximation; least squares; binomial distribution.

The least squares approximation to a discrete probability distribution is the best fit to the discrete probability distribution. Hence, it is of interest the educators to have a simple numerical method, which allows finding the least squares approximation. In this paper we offer a new simple numerical method for finding the least squares approximation to a discrete probability distribution. The method is first described in [1], and in this paper we show that the method could be applied to the problem of finding the least squares approximation to a discrete probability distribution. We give an example in order to illustrate the method in the case of the least squares normal approximations to the binomial distribution.

The method 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 in which the objective function has two variables is straightforward.

The described numerical method has the following advantages. The method is simple, so that the school and college students could understand and use it. The method does not use the Gauss approach for finding the minimum of the objective function, so that the method could be used without preliminary studying of partial derivatives. By using the described numerical method, we can approximate a sample or a discrete probability distribution by any continuous probability distribution, provided the corresponding objective function has one or two variables. The method is fast, because it needs small numbers of iterations. In the described method each iteration adds one true digit to the answer. We receive the answer for less than 1 second, if we use a desktop personal computer. We could use the method also in the case in which the objective function has more than two variables, but in this case we must have a powerful computer.

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. A file containing calculations for the example given below is available for download as supplementary material.

Recall that the binomial distribution is given by

$$B(k;n,p) = \binom{n}{k} p^k (1-p)^{n-k}$$

where the variable k with $0 \le k \le n$ and the parameter n ($n > 0$) are integers and the parameter p ($0 \le p \le 1$) is a real quantity. The expected value of the binomial distribution is equal to $\mu = np$, and the standard deviation is equal to $\sigma = \sqrt{np(1-p)}$ .Then the normal distribution

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

is an approximation to the binomial distribution, which we will call the standard normal approximation to the binomial distribution.

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

$$f(\mu,\sigma) = \sum_{k=0}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2\sigma^2}} - \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}\right)^2$$

Example. Find the least squares normal approximation to the binomial distribution, provided n = 240 and p = 0.85.

Solution. We use the computer program. Set N = 10. As initial intervals for finding μL and σL we take the intervals [μB-0.5, μB+0.5] and [σB-0.5, σB+0.5], where μB = 204 and σB ≈ 5.5 are the rounded values of the expected value and standard deviation of the binomial distribution. If we require 4 digits after the decimal point, we receive the following answer: $\mu L \approx 204.1755$ and $\sigma L \approx 5.5316$.

Note that we could compare also the sums of the least squares in the above two cases. Let B and L are the sums of the least squares in the case of regulation and least squares normal approximations, respectively. In the above example, $B \approx 0.00004290551$ and $L \approx 0.00001725588$. We see that the quotient $B/L \approx 2.486$, that is, the least squares normal approximation about 2.5 times better fits the binomial distribution than the standard normal distribution.

The reader may download a file containing the record of calculations made by the computer program in the example given in this paper. The following file is available for download:

JCGM201210_Binomial_Distribution.pdf

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