Statstical Variable Inference Sampling in a Prediction Model

Shin Ae Kim[1,2], Juyeon Kim[3],Yunjong Lee[1,*]

[1]Korea Atomic Energy Research Institute, 29 Gumgu-gil, Jeongeup, Korea

[2]Department of Nuclear Engineering, Hanyang University, 222 Wangsimni-ro, Seongdong-gu, Seoul, Republic of Korea

3Department of Nuclear power Industry, KEPCO International Nuclear Graduate School, 658-91 Haemaji-ro, Ulju-gun, Ulsan

E-mail : yjlee@kaeri.re.kr

1. Introduction

Predicting the dispersion of contaminants when a large amount of radioactive material is released into the atmosphere, such as during a terrorist activity or an event like the Fukushima nuclear accident, is very important for effective response and emergency post-evaluation. In general, accurate local meteorlogical data can be obtained through monitors installed at local observatories. However, in the case of a sudden accident, it takes time to collect data on emission concentration, location and time. These must be inferred from the local monitor measurements. In this sutdy, to sample the source position information needed, a function was defined of which the output is the data value of x. Then, several exponetial and normal distribution functions were applied to the more general gaussian plume model and the functions compared. In addition, algorithms for various techniques such as inverse CDF, rejection sampling, and importance sampling were written and implemented. Through this, a preliminary study on the research methodology for regressively estimating the location information to know was conducted and the possibility was reviewed.

2. Method & Result

The R program (x64 3.6.2 version) was used to implement the sampling technique. R is an open-source program, a language for statistics/mining and graphing. It is mainly used for research and industry-specific applications for big data analysis. R was introduced in 1993 as a free version of S-PLUS by Ross Ihaka and Robert Gentleman of the University of Auckland.

2.1 Simulation

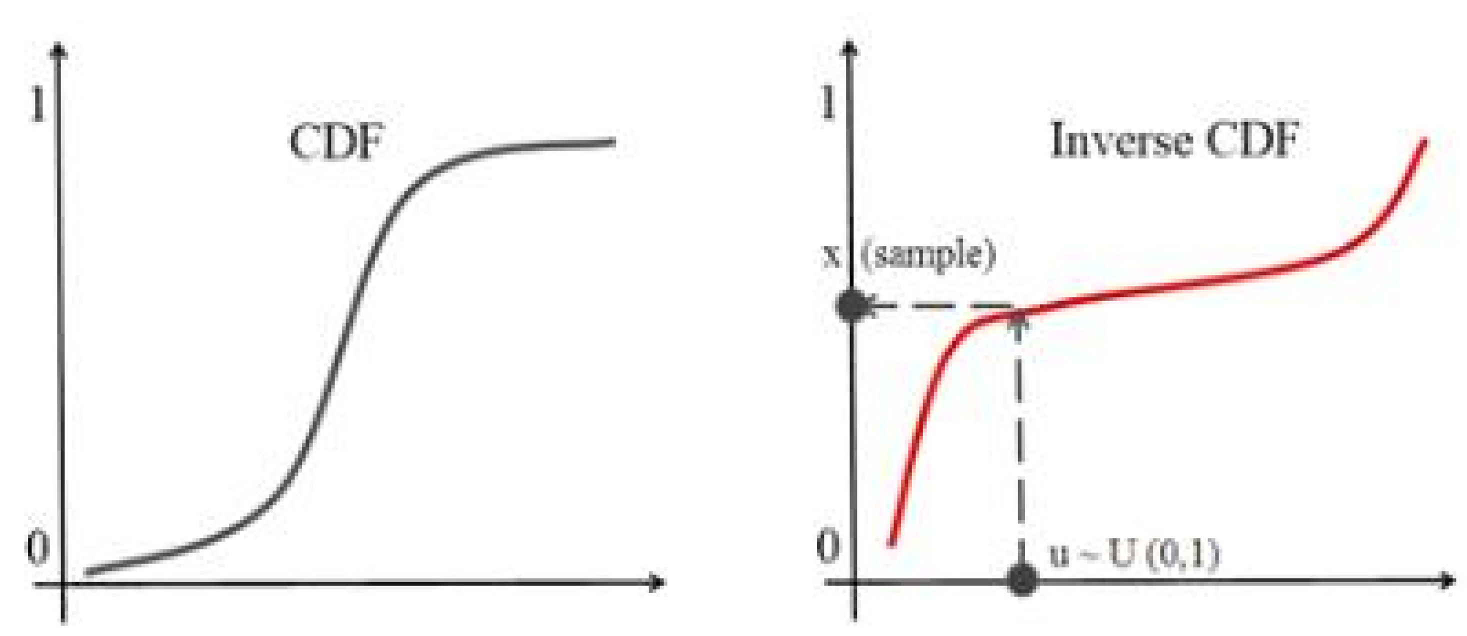

Fig. 1. Inverse Transform Function Sampling

To implement the inverse CDF method, an arbitrary function was defined and a sampling algorithm was implemented. Inverse function sampling is a solution to a relatively easy function. Random sampling times were extracted for the set function in Figure 2. It was confirmed that, as the number of times was repeated, the random value in the function set at random, approximated the exponential function, which was target function. When sampling is performed assuming a uniform distribution in this way, it is possible to extract the model variable to be found.

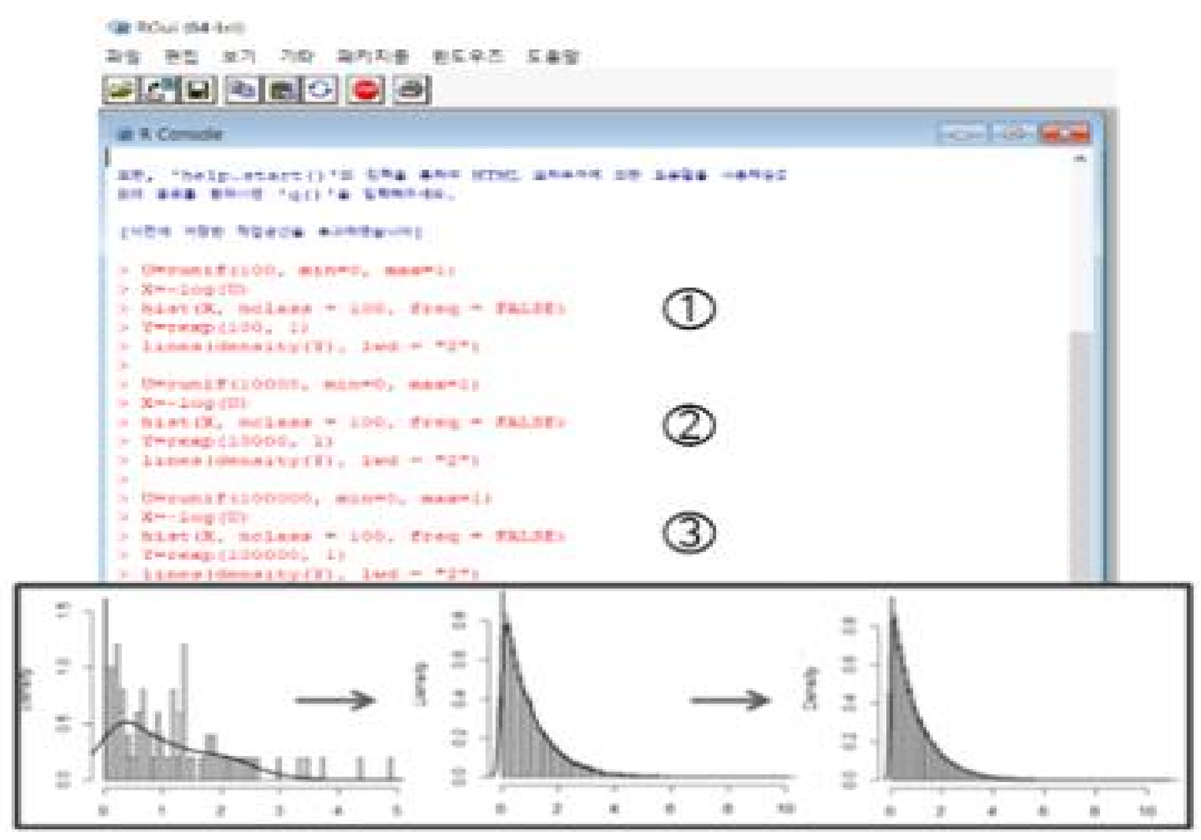

Fig. 2. Simulation of Inverse CDF in R program

Figure 3 was confirmed that the random sampling value in the set function approximates the target function. The proposed distribution is denoted as M(x), which acts as a cover function including the target distribution f(x). Rejection sampling is a technique in which a candidate to be adopted from M(x) is extracted after which the sampling probability is corrected by rejecting some samples. The proposed distribution was set using a uniform or normal distribution.

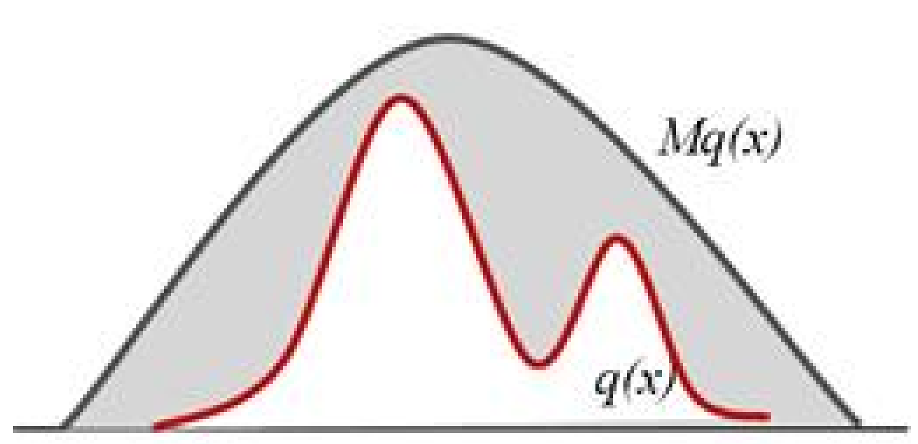

Fig. 3. It was confirmed that the random sampling in the arbitary set function approximates the target function

The proposed distribution is denoted as M(X), which acts as a cover function including the target distribution f(x). When sampling is performed assuming a uniform distribution in this way, it is possible to extract the model variable to be found.

3. Conclusion

The inverse CDF method is the most basic method and has the disadvantage of having to find the inverse function. During the sampling process, sampling near $-\infty$ and ∞ was difficult. Rejection sampling is when it is said that the purpose of obtaining a sample from a specific target function f(x) is that the distribution of the function f(x) to be sampled is not a commonly known distribution. This gets difficult. In this case, it is a method that uses g(x), which is an approximate function of the function f(x), to create another cover function M(x) that can cover all f(x). Thefore, it can be effective only when the distribution most similar to the target function f(x) is used.