SCIENCE CHINA
Mathematics

• ARTICLES •

December 2014 Vol. 57 No. 12: 2609–2620
doi: 10.1007/s11425-014-4860-9

Applications of the representative points in statistical simulations

FANG KaiTai[1,2,*], ZHOU Min[1] & WANG WenJun[3]

[1]*Division of Science and Technology, Beijing Normal University-Hong Kong Baptist University, United International College, Zhuhai* 519085*, China;*
[2]*Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing* 100190*, China;*
[3]*Department of Mathematics, Hainan Normal University, Haikou* 510000*, China*

Email: ktfang@uic.edu.hk, minzhou@uic.edu.hk, 956258159*@qq.com*

Received September 12, 2013; accepted January 6, 2014; published online June 19, 2014

Abstract The paper gives a new approach to statistical simulation and resampling by the use of number-theoretic methods and representative points. Resempling techniques take samples from an approximate population. The bootstrap suggests to use a random sample to form an approximate population. We propose to construct some approximate population distribution by the use of two kinds of representative points, and samples are taken from these approximate distributions. The statistical inference is based on those samples. The statistical inference in this paper involves estimation of mean, variance, skewness, kurtosis, quantile and density of the population distribution. Our results show that the new method can significantly improve the results by the use of Monte Carlo methods.

Keywords bootstrap, kernel density estimation, normal distribution, representative points, resampling, statistical simulation

MSC(2010) 65C05, 65C50

Citation: Fang K T, Zhou M, Wang W J. Applications of the representative points in statistical simulations. Sci China Math, 2014, 57: 2609–2620, doi: 10.1007/s11425-014-4860-9

1 Stochastic simulation and representative points

Statistical simulation [10] has played an important role in statistical research and applications. The statistical simulation is also called Monte Carlo method. It takes random samples from a given population by the use of computer software, and then we calculate the statistic of the interest for statistical inference. Let X be the population random variable with the probability distribution function $F(x)$ and $x_1, \ldots, x_n$ be a random sample from the population. Let $T = T(x_1, \ldots, x_n)$ be a statistic of the interest, a function of the sample. We want to know distribution, mean, variance, quantiles of T and so on. Based on a sample we can obtain a sample, say T_1, from the population T. Repeatly taking m samples from X, we obtain samples $T_1, \ldots, T_m$ from the population T. Then we can estimate moments, quantile and distribution of T based on $T_1, \ldots, T_m$. In this paper, our study focuses only on X to be a continuous variable, i.e., X has a density function, say, $p(x)$.

*Corresponding author

Various Monte Carlo methods provide ways to generate a sample, $x_1, \ldots, x_n$, from the population, where $x_1, \ldots, x_n$ are i.i.d. and $x_i \sim F(x)$. The empirical distribution function of the sample is given by

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I\{x_i \leqslant x\} \tag{1.1}$$

that should be close to $F(x)$ in a certain sense. In the above formula

$$I\{A\} = \begin{cases} 1, & A \text{ occurs}, \\ 0, & A \text{ does not occur} \end{cases} \tag{1.2}$$

is the index function of the event A. Here, $F_n(x)$ and $F(x)$ are two functionals and we might use L_p-distance to measure their closeness, where $p \geqslant 1$ and

$$D_p(F, F_n) = \left[\int_{-\infty}^{+\infty} |F_n(x) - F(x)|^p dx \right]^{1/p}. \tag{1.3}$$

The L_1-distance and L_2-distance have been popularly used. These two distances are called L_1-F-discrepancy and L_2-F-discrepancy, respectively in the number-theoretic methods [6, 11].

Efron [3] proposed the so-called bootstrap in 1979. The bootstrap regards a given sample $x_1, \ldots, x_n$ as the support points of a discrete random variable Y, say, and Y is uniformly distributed on the support points, i.e., $P(Y = x_j) = \frac{1}{n}, j = 1, \ldots, n$; or we denote this fact as follows:

Y	x_1	x_2	$\cdots$	x_n
p	$\frac{1}{n}$	$\frac{1}{n}$	$\cdots$	$\frac{1}{n}$

The next step is to take a sample from Y, denoted by $y_1, \ldots, y_N$, where N may equal to n or not, depending on the practical situation. The key idea of the bootstrap is to use a discrete distribution, formed by a random sample, as an approximate distribution to the population distribution $F(x)$, and then we resample from this approximate distribution. Obviously, taking a random sample in physical practice needs some manpower and budget, but resampling and related statistical reference can be completed on a computer. This way can save a lot of time, manpower and budget. As a result the bootstrap technique has been widely used in the past more than 30 years. The reader can refer to [4] for the theory of bootstrap.

A natural question is: can we use other ways to construct an approximate population that has a better performance in stochastic simulations? Denote by $\mathcal{B} = \{b_1, \ldots, b_n\}$ as n support points of Y with respective probability $P(Y = b_j) = p_j, j = 1, \ldots, n$, or express as follows:

Y	b_1	b_2	$\cdots$	b_n
p	p_1	p_2	$\cdots$	p_n

(1.4)

We want to choose $\mathcal{B}$ and related probabilities such that two distribution functions $F_Y(x)$ and $F(x)$ are close to each other in a certain sense. For example, we can minimize L_2-distance between $F_Y(x)$ and $F(x)$ and the solution of $\mathcal{B}$ is called as representative points of X. In practical applications, we can relax this definition of representation. Therefore, Fang and Wang [6, Chapter 4] introduced two kinds of representative points and applications.

The first kind of the representative points of $F(x)$, RP-I for short, is constructed by the inverse transformation method and the number-theoretic methods, where

$$b_j = F^{-1}\left(\frac{2j-1}{2n}\right), \quad j = 1, \ldots, n, \tag{1.5}$$

with corresponding probability $P(Y = b_j) = 1/n$, where $F^{-1}(y)$ is the inverse function of $F(x)$ and set of points $\{\frac{2j-1}{2n}, j = 1, \ldots, n\}$ is uniformly scattered on the interval $(0, 1)$. More details can refer

to [6, Example 1.4]. When the population is the standard normal distribution,

$$F(x)=\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\mathrm{e}^{-\frac{1}{2}t^2}dt=\int_{-\infty}^{x}\phi(t)dt,$$

where $\phi(t)$ is the density function of the standard normal distribution. RP-I of $Z\sim N(0,1)$ can be calculated by

$$b_j=\Phi^{-1}((2j-1)/(2n)),\quad p_j=1/n,\quad j=1,\dots,n. \tag{1.6}$$

For the second type of the representative points of $F(x)$, RP-II for short, it requires the first two order moments of $F(x)$. RP-II was independently proposed by Cox [2], Bofinger [1], Fang and He [5] and many others. In the literature RP-II is called different names, such as "quantized" and "principal points" (see [7]). Suppose that a random variable $X\sim F(x)$ has a density function $p(x)$ with mean $\mu=E(x)$ and variance σ^2. RP-II of $F(x)$ is constructed by the following procedure.

Take $-\infty<b_1<b_2<\cdots<b_n<\infty$ and define a stepwise function

$$Q_b(x)=b_k,\quad \text{when } a_k<x\leqslant a_{k+1},\quad k=1,\dots,n, \tag{1.7}$$

where $a_1=-\infty$, $a_k=(b_k+b_{k-1})/2$, $k=2,\dots,n$, $a_{n+1}=\infty$. Define the mean square error (MSE) to measure bias between $F(x)$ and $Q_b(x)$ as follows:

$$\begin{aligned}\mathrm{MSE}(\boldsymbol{b})&=\mathrm{MSE}(b_1,\dots,b_n)=\frac{1}{\sigma^2}E(X-Q_b(X))^2\\&=\frac{1}{\sigma^2}\int_{-\infty}^{+\infty}\min_k(x-b_k)^2p(x)dx=\frac{1}{\sigma^2}\sum_{k=1}^{n}\int_{a_k}^{a_{k+1}}(x-b_k)^2p(x)dx.\end{aligned} \tag{1.8}$$

To find $\boldsymbol{b}^*=(b_1^*,\dots,b_n^*)$ such that $\mathrm{MSE}(\boldsymbol{b})$ arrives at its minimum, the solution of $\mathcal{B}^*=\{b_1^*,\dots,b_n^*\}$ is just RP-II of $F(x)$. For convenience we use $\mathcal{B}$ to replace $\mathcal{B}^*$ in the remaining text of the paper. The probability of each representative point is given by

$$P(Q_b(X)=b_i)=p_i,\quad i=1,\dots,n, \tag{1.9}$$

where

$$\begin{aligned}p_1&=\int_{-\infty}^{(b_1+b_2)/2}p(x)dx=\int_{-\infty}^{a_1}p(x)dx,\\p_i&=\int_{(b_{i-1}+b_i)/2}^{(b_i+b_{i+1})/2}p(x)dx=\int_{a_i}^{a_{i+1}}p(x)dx,\quad i=2,\dots,n-1,\\p_n&=\int_{(b_{n-1}+b_n)/2}^{\infty}p(x)dx=\int_{a_n}^{\infty}p(x)dx.\end{aligned}$$

It can be proved that $E(Q_n(X))=E(X)$, $\mathrm{Var}(Q_n(X))\to\mathrm{Var}(X)$ and $E(X-Q_n(X))^2\to 0$, where $Q_n(X)$ is $Q_b(X)$ for highlighting n. There are rich applications of RP-II in the transaction of signal. The international journal "IEEE Transaction on Information Theory" published a special issue on this research direction in 1982.

In this paper, we propose applications of two kinds of the representative points (RP-I and RP-II) into statistical simulation. A random sample from $F(x)$ can be regarded as a set of RP of $F(x)$. We denote it as RP-MC. Related to these three kinds of RP, there are three discrete distributions mentioned in the above text. The corresponding random variables are denoted by $Y_{\mathrm{MC}}, Y_{\mathrm{I}}$ and Y_{II}, respectively. These random variables construct three approximate populations to $F(x)$. Without any confusion we might use RP-MC, RP-I and RP-II for three kinds of RP, for the three approximate populations, as well as for the three methods. In the remaining of the paper, we focus on comparisons among these three methods in statistical inference. The statistical inferences discussed in this paper involve estimation of mean, variance, skewness, kurtosis, confidence interval, quantile and density estimation. Section 2 considers estimation of the mean, variance, skewness, kurtosis by the use of three kinds of RP and gives

numerical comparisons among the three classes of RP. In Section 3, we consider resampling from the three approximate populations and compare their performance in statistical inferences. Section 4 proposes some applications of RP in signal transaction by density estimation. Section 5 proposes conclusion and further studies.

2 Preliminary comparisons

In this section, we consider estimation of mean, variance, skewness and kurtosis of $X \sim N(\mu, \sigma^2)$. Without loss of any generality we can only consider the standard normal distribution $Z \sim N(0, 1)$. The three classes of RP mentioned in the previous section are used for simulation.

For a given n, RP-I can be easily obtained by formula (1.6) and RP-II can be found in Fang and He [5] for $n \leqslant 31$. However, their results can be improved because the approximate formulas for calculating the distribution function of the standard normal distribution and its inverse have been improved in many software. For example, we can find more accuracy results by the built-in functions in MATLAB. Therefore, we recalculate the RP-II of Z and related probabilities for $n \leqslant 36$ and present the results in Appendix of this paper. Two tables provide all the RP-II needed in the paper.

Denote the following statistics of X by

$$\mu = E(X), \quad \sigma^2 = \mathrm{Var}(X),$$
$$Sk(X) = \frac{E(X-\mu)^3}{\sigma^3}, \quad Ku(X) = \frac{E(X-\mu)^4}{\sigma^4} - 3.$$

It is well known that

$$E(Z) = 0, \quad \sigma^2 = 1, \quad Sk(Z) = 0 \quad \text{and} \quad Ku(Z) = 0.$$

Let Y be a discrete distribution with probability mass function (1.4) that is an approximate distribution to $\Phi(x)$. Then the above statistics are

$$\begin{aligned} &E(Y) = \sum_{i=1}^{n} b_i p_i \equiv \mu_b, \quad \mathrm{Var}(Y) = \sum_{i=1}^{n} (b_i - \mu_b)^2 p_i \equiv \sigma_Y^2, \\ &Sk(Y) = \frac{1}{\sigma_Y^3} \sum_{i=1}^{n} (b_i - \mu_b)^3 p_i, \quad Ku(Y) = \frac{1}{\sigma_Y^4} \sum_{i=1}^{n} (b_i - \mu_b)^4 p_i - 3. \end{aligned} \tag{2.1}$$

Obviously, $\mu_b, \sigma_Y^2, Sk(Y)$ and $Ku(Y)$ should be close to the respective $E(Z) = 0$, $\sigma^2 = 1$, $Sk(Z) = 0$ and $Ku(Z) = 0$, if $\mathcal{B}$ is a set of representative points. In the following comparisons, we employ RP-MC, RP-I and RP-II and consider the cases of $n = 5, 10, 15, 20, 25, 28, 31$. It is clear that RP-MC is a random sample of size n. For fair comparisons, we might take m samples and use the average of m estimators of the statistic based on those m samples. In our study, we choose $m = 1, 10, 100, 1000$.

As the standard normal density is symmetric about the origin, it is easy to see that RP-I and RP-II are also symmetric about the origin. Therefore, the corresponding $\mu_b = 0, Sk(Y) = 0$. For comparing bias of $\sigma_Y^2 - \sigma^2 = \sigma_Y^2 - 1$ and bias of $Ku(Y) - Ku(Z) = Ku(Y)$ among the three kinds of RP, Table 1 shows numerical results, where MC(m) denotes the bias between the true value of the statistic and the average value of the statistic based on m independent random samples. There are total six methods (MC(1), MC(10), MC(100), MC(1000), RP-I and RP-II) to be employed. For each statistic (variance and kurtosis) there are six estimators and corresponding six biases, among which the smallest absolute bias is marked in boldface. From results on Table 1 we may raise the following observations: (1) the estimator is more accuracy if the number of RP, n, increases, and we should choose a larger n; (2) for RP-MC method, the testing results are rather poor for smaller m ($m = 1, 10$) and we recommend to use a large m; (3) for estimation of kurtosis RP-II has the best performance; and (4) for estimation of the variance RP-II has the best result if $n \geqslant 25$, otherwise RP-MC is better if $m = 1000$ or larger.

Why does RP-II have a better performance than RP-MC and RP-I? Let us look at the empirical distribution of Y. The empirical distribution function of Y_{MC} is given by (1.1), and others, denoted by

$F_n^{\mathrm{I}}(x)$ and $F_n^{\mathrm{II}}(x)$, are as follows:

$$F_n^{\mathrm{I}}(x)=\begin{cases}0, & \text{when } x<b_1,\\ \dfrac{1}{n}, & \text{when } b_1\leqslant x<b_2,\\ \dfrac{2}{n}, & \text{when } b_2\leqslant x<b_3,\\ \vdots & \vdots\\ \dfrac{n-1}{n}, & \text{when } b_{n-1}\leqslant x<b_n,\\ 1, & \text{when } x\geqslant b_n,\end{cases}$$

$$F_n^{\mathrm{II}}(x)=\begin{cases}0, & \text{when } x<b_1,\\ p_1, & \text{when } b_1\leqslant x<b_2,\\ p_1+p_2, & \text{when } b_2\leqslant x<b_3,\\ \vdots & \vdots\\ p_1+\cdots+p_{n-1}, & \text{when } b_{n-1}\leqslant x<b_n,\\ 1, & \text{when } x\geqslant b_n.\end{cases}$$

Note that the representative points $\mathcal{B}=\{b_1,\ldots,b_n\}$ in the above two formulas are different. Denote by $(D_2(\Phi,F_n^{\mathrm{I}}))^2$ and $(D_2(\Phi,F_n^{\mathrm{II}}))^2$ the square distance between the empirical distribution function and $\Phi(x)$ (cf. (1.3)). Table 2 presents square distance for $n=10,15,20,25,28,31$ for RP-I and RP-II. It is obvious that the empirical distribution function $F_n^{\mathrm{II}}(x)$ is closer to $\Phi(x)$. An explanation for this phenomenon is given as follows:

Most of the number-theoretic methods consider construction of a set of points, $\mathcal{B}=\{\boldsymbol{b}_1,\ldots,\boldsymbol{b}_n\}$ that are uniformly scattered on a unit cube in a high dimension R^d. If the population distribution is not the uniform distribution on the unit interval/cube, the number-theoretic method suggests to use the inverse transformation method that maps points from the unit interval/cube to R^d. This inverse transformation in general is non-linear. This indirect method for generating a set of representative points is not as good as some direct method like RP-II. For example, the normal population of $N(0,1)$ is just this case. But RP-II of $N(0,1)$ can be directly found. So the discrete distribution function constructed by the RP-II is closer to $\Phi(x)$.

Table 1 Estimation bias in estimation of variance and kurtosis

n	Category	5	10	15	20	25	28	31
Variance	MC(1)	0.2683	0.3030	−0.1771	−0.1295	0.1224	−0.1535	0.0770
	MC(10)	0.2937	−0.2765	0.1195	−0.0408	0.0988	0.0018	0.0712
	MC(100)	−0.1464	−0.0456	0.0322	−0.0060	0.0076	0.0187	0.0144
	MC(1000)	−0.0126	0.0133	0.0243	0.0060	0.0063	0.0046	0.0049
	I	−0.2331	−0.1202	−0.0812	−0.0614	−0.0494	−0.0443	−0.0401
	II	−0.0799	−0.0229	−0.0107	−0.0062	−0.0040	−0.0032	−0.0027
Kurtosis	MC(1)	−1.4456	−1.1634	2.2581	−1.4309	−0.3166	−0.4627	−1.1498
	MC(10)	−1.7912	−0.9304	−0.7736	−0.3306	−0.5586	−0.4175	−0.3349
	MC(100)	−1.7552	−0.9669	−0.5986	−0.3821	−0.2737	−0.3826	−0.3257
	MC(1000)	−1.7212	−1.0220	−0.7325	−0.5488	−0.4728	−0.3807	−0.3619
	I	−1.1143	−0.7512	−0.5903	−0.4943	−0.4291	−0.3988	−0.3732
	II	−0.5753	−0.2023	−0.1033	−0.0627	−0.0421	−0.0343	−0.0284

Table 2 Square L_2-distance between the empirical distribution and $\Phi(x)$

n	10	15	20	25	28	31
$(D_2(\Phi, F_n^I))^2$	0.0428	0.0184	0.0101	0.0064	0.0050	0.0041
$(D_2(\Phi, F_n^{II}))^2$	0.0065	0.0024	0.0013	8.2606e−004	6.5931e−004	5.3934e−004

3 Resampling on representative points

The bootstrap is one of the resampling methods and was proposed by Efron in 1979. The bootstrap has been widely used in statistical inference. In traditional stochastic simulation samples are taken from the population. Resampling techniques take a sample, $y_1, \ldots, y_n$, from an approximate distribution, $F_Y(y)$ say. The bootstrap employes a random sample to form an approximate population. In fact, we may use RP-I or RP-II to form two different populations that has been introduced in Section 1. Repeatly taking N samples from Y, we can use those N samples for statistical inference. For example, if the variance is of the interest and T is a statistic for estimation of the variance, we can find $T_1, \ldots, T_N$ from the population Y. The sample mean of $T_1, \ldots, T_N$ is suggested to be an estimator of the variance. Similarly, we can choose T to be any statistic of the interest. Theory and methodology of the bootstrap can refer to [4].

Obviously, the three kinds of approximate populations are discrete. Their sampling can employ the inverse transformation method. It follows the following steps:

Step 1. Generate a random number U, i.e., $U \sim U(0,1)$, the latter is the uniform distribution on (0,1).

Step 2. Define a random variable Y by

$$Y = \begin{cases} b_1, & \text{when } U < p_1, \\ b_2, & \text{when } p_1 \leqslant U < p_1 + p_2, \\ \vdots & \vdots \\ b_n, & \text{when } \sum\limits_{i=1}^{n-1} p_i \leqslant U. \end{cases} \tag{3.1}$$

Step 3. Repeat the above two steps n times, and we have a sample of Y, $y_1, \ldots, y_n$. Calculate the given statistic T.

Step 4. Repeat the above three steps N times, and we obtain a sample of T, $T_1, \ldots, T_N$.

Step 5. We can use the mean of a sample of T to infer the statistic of the population.

Now we apply the three methods for estimation of four statistics of $Z \sim N(0,1)$: Mean, variance, skewness and kurtosis.

Tables 3–5 show estimation biases for the above four statistics, where RP-MC, RP-I and RP-II are employed involving the cases: $n = 25,\ 28,\ 31$ and $N = 1000,\ 2000,\ 5000,\ 10000$.

From visualization of the results in Tables 3–5, we have some experimental conclusions: RP-MC has a good performance for estimation of the mean (if $n = 31, N = 2000, 5000, 10000$) and for estimation of the variance (if $n = 31, N = 5000, 10000$); RP-I has the best estimation for estimation of the mean (if $n = 25,\ N = 2000, 5000$; $n = 28,\ N = 1000, 5000, 10000$), for estimation of the skewness ($n = 25,\ N = 1000, 2000$; $n = 28,\ N = 2000$; $n = 31,\ N = 1000, 2000, 5000$). In the remaining 32 cases RP-II gives the most accurate estimation. Summarizing the above comparisons in Table 6, we can see that RP-II provides more accurate estimation in most cases.

4 Applications of RP in signal transaction by density estimation

In the signal transaction one needs to estimate a density function and related statistics for the input without any prior information. When the population of the input has a continuous distribution

Table 3 $n = 25$, estimation biases by resampling

	Category	1000	2000	5000	10000
Mean	MC	−0.002429	0.004777	−0.003709	−0.003057
	I	−0.005924	−0.000038	0.001049	0.000749
	II	0.001915	0.001743	−0.004115	0.000739
Variance	MC	0.187685	0.181844	0.183003	0.187855
	I	−0.053420	−0.048240	−0.052736	−0.050859
	II	0.019990	0.001917	0.003327	−0.002831
Kurtosis	MC	0.801716	0.817543	0.891590	0.855281
	I	−0.450798	−0.464493	−0.458039	−0.451750
	II	−0.254120	−0.269414	−0.246620	−0.237732
Skewness	MC	0.630710	0.633194	0.652969	0.645178
	I	−0.015296	−0.001170	−0.006386	0.006676
	II	0.024606	0.010832	0.003904	−0.000951

Table 4 $n = 28$, estimation biases by resampling

	Category	1000	2000	5000	10000
Mean	MC	0.073298	0.077869	0.069957	0.074098
	I	−0.003045	−0.006565	0.003439	−0.000582
	II	−0.012096	−0.001874	−0.003936	−0.003599
Variance	MC	−0.170677	−0.177562	−0.178397	−0.175707
	I	−0.044419	−0.046451	−0.049236	−0.043992
	II	−0.017692	0.001960	−0.007074	−0.003750
Kurtosis	MC	−0.612533	−0.630435	−0.633608	−0.631694
	I	−0.435333	−0.435678	−0.423529	−0.431312
	II	−0.218617	−0.233815	−0.225395	−0.215011
Skewness	MC	−0.018668	−0.023468	−0.018847	−0.021584
	I	0.010544	0.008642	−0.009148	−0.003854
	II	0.004971	0.017812	0.002168	0.001850

with a unknown density function $p(x)$, the so-called Parzen-Rosenblatt window method proposed by Rosenblatt [9] and Parzen [8], provides a way to estimate the density function based on a set of samples, $x_1, \ldots, x_n$. The estimate function

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right) \tag{4.1}$$

is called kernel density estimation, where $k(\cdot)$ is the kernel function, h is the bandwidth and $k_h(y) = \frac{1}{h}k(y/h)$. Most popular kernel is the standard density function

$$k(x) = \phi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2},$$

therefore, we choose it as the kernel. In fact, n i.i.d. samples can be replaced by a set of representative points, $b_1, \ldots, b_n$ with related probabilities in (1.4). In this case, (4.1) becomes

$$\hat{p}_h(x) = \sum_{i=1}^{n} k_h(x - x_i)p_i = \frac{1}{h}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right)p_i. \tag{4.2}$$

The choice of the bandwidth is very important. There are many studies on this topic. If one is not familiar with the literature, she/he can try several h's and choose an h such that the corresponding estimate has

Table 5 $n = 31$, Estimation biases by resampling

	Category	1000	2000	5000	10000
Mean	MC	−0.008385	−0.004103	0.000779	0.000764
	I	−0.008120	0.005167	0.001466	−0.001234
	II	−0.003319	−0.007977	0.001532	−0.001053
Variance	MC	0.005258	0.007336	−0.001558	−0.000217
	I	−0.024992	−0.028539	−0.039318	−0.040546
	II	0.000968	−0.004409	−0.003516	−0.003151
Kurtosis	MC	−0.593206	−0.604982	−0.590074	−0.593610
	I	−0.407610	−0.417200	−0.403864	−0.405444
	II	−0.246208	−0.193741	−0.208364	−0.186791
Skewness	MC	0.093357	0.097954	0.100418	0.097152
	I	0.003821	−0.000494	0.000305	0.003815
	II	0.019209	−0.008148	−0.000787	0.000010

Table 6 The number of winner in statistical estimation

	MC	I	II
Mean	3	5	4
Variance	2	0	10
Kurtosis	0	0	12
Skewness	0	6	6
Total	5	11	32

a good performance in a certain sense. In our study, we choose h to have minimum L_2-distance between $\hat{p}_h(x)$ and $p(x)$.

Suppose that one needs to send a data set of a signal out. The receiver wants to estimate the density function and related statistics from the input. How to choose a good input data? Obviously, we can consider three kind of RPs, RP-MC, RP-I and RP-II as possible candidates. In this section, we still consider the normal population and choose $n = 30, 31$ (one is even and another is odd) as size of the input data. Then we estimate density function and quantiles of the input data and give comparisons among the three methods.

For $n = 30$, take $h = 0.05, 0.10, \ldots, 0.60$. Our calculation shows that $h = 0.25$ is the best when RP-I is used, and $h = 0.15$ for RP-II. When RP-MC is employed, due to randomness of RP-MC, the best h is uncertain. In our experiment, $h = 0.50$ is recommended. The three density estimators are presented on Figure 1. By visualizing Figure 1, RP-II is the best. Three L_2-distances between density estimator $\hat{p}_h(x)$ and $p(x)$ are

$$\text{RP-MC} : 0.8725, \quad \text{RP-I} : 0.0468, \quad \text{RP-II} : 0.0033$$

that is consistent with the visualization.

Note that three density estimators (when $n = 30$) give a lower estimated value at $x = 0$ as all the three RP sets do not involve $b = 0$. Therefore, we should choose n to be odd. For $n = 31$ the recommended h and L_2-distance between each density estimator and $\phi(x)$ are given in Table 7. From Figure 2, RP-II shows excellent performance among the three RP sets.

From the density estimator, we can easily find an estimator of an unknown statistic of the population, for example, mean, variance, quantile and confidence interval. Table 8 shows estimation biases for the p^{th} quantile, where $p = 0.001, 0.005, 0.01, 0.02, 0.025, 0.05, 0.1$ and $n = 30, 31$. It is clear that RP-II is better than RP-I. This conclusion can be applied to estimation of the confidence interval.

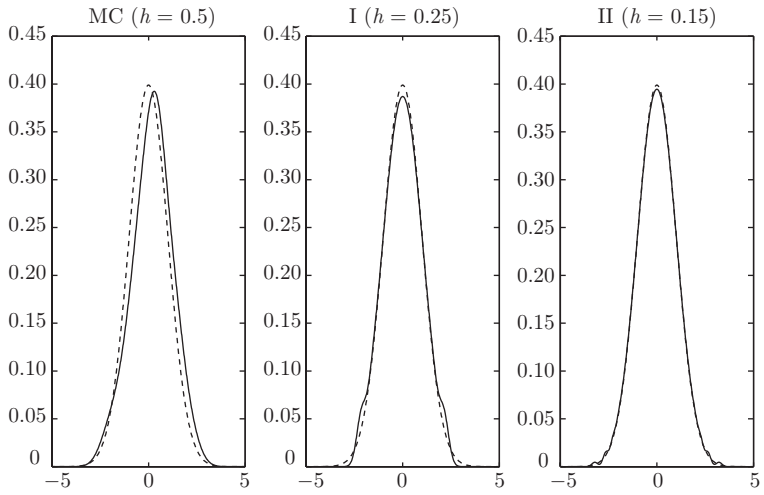

Figure 1 Kernel density estimation, $n = 30$

Table 7 Recommended h and L_2-distance under $n = 31$

	MC	I	II
h	0.4000	0.2500	0.1500
D_2^2	0.9316	0.0450	0.0031

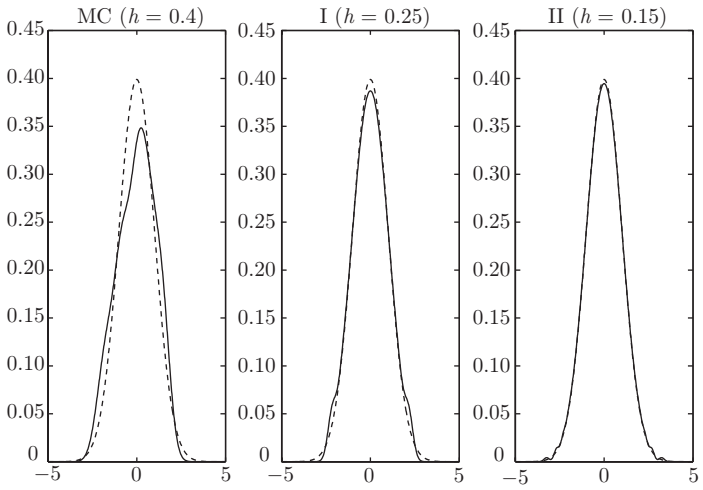

Figure 2 Kernel density estimation, $n = 31$

Table 8 The estimation biases for p-quantiles

	n	$p = 0.001$	$p = 0.005$	$p = 0.01$	$p = 0.02$	$p = 0.025$	$p = 0.05$	$p = 0.1$
I	30	−0.491730	−0.187030	−0.062250	0.036551	0.055136	0.054346	0.039448
	31	−0.482130	−0.179030	−0.055750	0.039951	0.057036	0.053846	0.039448
II	30	0.092968	0.030671	0.018752	0.020851	0.019736	0.017346	0.013748
	31	0.097368	0.032671	0.019652	0.020651	0.020036	0.017446	0.013848

5 Conclusion and further study

This paper proposes to use two kinds of representative points (RP-I and RP-II) for construction of approximate distributions and to regard a set of RP as a "random sample" in the statistical simulation. Various comparisons in estimation of the mean, variance, skewness, kurtosis, quantiles and density function among Monte Carlo method (RP-MC), RP-I and RP-II are given. Our results show that RP-II has an excellent performance among the three methods. This is a new idea and new method in the statistical simulation. This approach extends the concepts of the random sample & the bootstrap and gives more choices in the statistical simulation.

There are a lot of research topics along this direction, especially how to use two kinds of representative points for construction of approximate populations for a multivariate population. This is a challenging direction in the statistical simulation. Besides, application of two kinds of representative points in Markov chain Monte Carlo (MCMC) methods is another important research topic.

Acknowledgements This work was supported by the Special Research Foundation from the Chinese Academy of Sciences, the Beijing Normal University-Hong Kong Baptist University United International College Research (Grant No. R201409) and National Natural Science Foundation of China (Grant No. 11261016). The authors thank the valuable comments given by Professor Yuan Wang and many comments from the two referees and Associate Editor.

References

1 Boringer E. Maximizing the correlation of grouped observations. J Amer Statist Theo, 1970, 65: 1632–1638
2 Cox D R. Note on grouping. J Amer Statist Theo, 1957, 52: 543–547
3 Efron B. Bootstrap methods: Another look at the jacknife. Ann Statist, 1979, 7: 1–26.
4 Efron B, Tibshirani R J. An Introduction to the Bootstrap. New York: Chapman & Hall, 1993
5 Fang K T, He S D. The problem of selecting a specified number of representative Points from a normal population. Acta Math Appl Sin, 1984, 17: 293–306
6 Fang K T, Wang Y. Number-Theoretic Methods in Statistics. London: Chapman & Hall, 1994
7 Flury B A. Principal points. Biometrika, 1990, 77: 33–41
8 Parzen E. On estimation of a probability density function and mode. Ann Math Statist, 1962, 33: 1065–1076
9 Rosenblatt M. Remarks on some nonparametric estimates of a density function. Ann Math Stat, 1956, 27: 832–837
10 Ross S M. A Course in Simulation. New York: MacMillan Publishing Company, 1990
11 Wang Y, Fang K T. Number theoretic methods in applied statistics. Chin Ann Math Ser B, 1990, 11: 51–65

Table 9 RP of the standard normal distribution and related 1-loss function

m	1-loss function	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$	$j=11$	$j=12$	$j=13$	$j=14$	$j=15$	$j=16$	$j=17$	$j=18$
2	63.661977	0.797885																	
3	80.982596	1.224006																	
4	88.251815	0.452780	1.510418																
5	92.005887	0.764568	1.724147																
6	94.202235	0.317716	1.000106	1.893595															
7	95.599962	0.560577	1.188147	2.033369															
8	96.545224	0.245094	0.756005	1.343909	2.151946														
9	97.214674	0.443639	0.918796	1.476392	2.254664														
10	97.706295	0.199623	0.609858	1.057825	1.591340	2.345096													
11	98.078049	0.367458	0.752367	1.178826	1.692639	2.425746													
12	98.366034	0.168438	0.511847	0.876780	1.285711	1.783030	2.498435												
13	98.593694	0.313773	0.638251	0.986949	1.381263	1.864518	2.564525												
14	98.776800	0.145706	0.441321	0.750443	1.085635	1.467528	1.938612	2.625063											
15	98.926279	0.273857	0.554764	0.851134	1.174879	1.546057	2.006474	2.680866											
16	99.049899	0.128395	0.388048	0.656759	0.942340	1.256231	1.618046	2.069017	2.732590										
17	99.153306	0.242994	0.490882	0.749287	1.025597	1.330896	1.684442	2.126971	2.780762										
18	99.240683	0.114769	0.346346	0.584302	0.833862	1.102100	1.399827	1.746003	2.180927	2.825817									
19	99.315183	0.218409	0.440355	0.669797	0.911666	1.172801	1.463791	1.803345	2.231373	2.868116									
20	99.379221	0.103763	0.312791	0.526488	0.748533	0.983642	1.238467	1.523414	1.856977	2.278714	2.907961								
21	99.434669	0.198357	0.399356	0.605892	0.821442	1.050552	1.299725	1.579214	1.907323	2.323289	2.945607								
22	99.482999	0.094686	0.285199	0.479232	0.679485	0.889276	1.113019	1.357093	1.631622	1.954739	2.365386	2.981274							
23	99.525380	0.181688	0.365400	0.553325	0.748012	0.952654	1.171561	1.411007	1.681003	1.999527	2.405252	3.015150							
24	99.562751	0.087072	0.262101	0.439854	0.622370	0.812084	1.012092	1.226612	1.461834	1.727666	2.041948	2.443099	3.047398						
25	99.595872	0.167610	0.336806	0.509283	0.686972	0.872212	1.068019	1.278540	1.509886	1.771877	2.082224	2.479110	3.078159						
26	99.625364	0.080593	0.242480	0.406517	0.574288	0.747635	0.928823	1.120803	1.327657	1.555432	1.813865	2.120549	2.513445	3.107559					
27	99.651739	0.155561	0.312389	0.471823	0.635364	0.804782	0.982281	1.170756	1.374235	1.598704	1.853829	2.157094	2.546247	3.135707					
28	99.675422	0.075012	0.225602	0.377919	0.533219	0.692934	0.858775	1.032897	1.218147	1.418505	1.639905	1.891945	2.192005	2.577637	3.162701				
29	99.696767	0.145132	0.291291	0.439557	0.591119	0.747353	0.909923	1.080939	1.263209	1.460671	1.679211	1.928365	2.225415	2.607727	3.188627				
30	99.716072	0.070155	0.210928	0.353110	0.497714	0.645876	0.798927	0.958490	1.126640	1.306147	1.500912	1.716779	1.963224	2.257440	2.636614	3.213562			
31	99.733589	0.136016	0.272876	0.411464	0.552739	0.697795	0.847921	1.004710	1.170202	1.347138	1.539385	1.752745	1.996643	2.288184	2.664386	3.237577			
32	99.749533	0.065890	0.198052	0.331378	0.466700	0.604934	0.747136	0.894565	1.048783	1.211804	1.386340	1.576228	1.787233	2.028728	2.317739	2.691120	3.260732		
33	99.764087	0.127979	0.256661	0.386777	0.519114	0.654557	0.794128	0.939060	1.090888	1.251604	1.423893	1.611565	1.820352	2.059577	2.346189	2.716887	3.283086		
34	99.777407	0.062114	0.186661	0.312182	0.439366	0.568966	0.701835	0.838970	0.981581	1.131182	1.289741	1.459921	1.645507	1.852198	2.089274	2.373609	2.741752	3.304689	
35	99.789629	0.120840	0.242272	0.364907	0.489401	0.616478	0.746965	0.881839	1.022287	1.169804	1.326339	1.494535	1.678153	1.882861	2.117898	2.400065	2.765771	3.325587	
36	99.800871	0.058747	0.176513	0.295100	0.415089	0.537104	0.661847	0.790123	0.922889	1.061315	1.206878	1.361510	1.527835	1.709591	1.912420	2.145518	2.425621	2.788998	3.345823

Table 10 RP of the standard normal distribution and related 1-loss function (continuity)

m	$j=0$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$j=7$	$j=8$	$j=9$	$j=10$	$j=11$	$j=12$	$j=13$	$j=14$	$j=15$	$j=16$	$j=17$	$j=18$
2		0.500000																	
3	0.459464	0.270268																	
4		0.336851	0.163149																
5	0.297749	0.244441	0.106684																
6		0.245024	0.181007	0.073969															
7	0.220744	0.198668	0.137344	0.053616															
8		0.191656	0.161475	0.106631	0.040238														
9	0.175545	0.164360	0.132329	0.084484	0.031054														
10		0.157166	0.140649	0.109530	0.068133	0.024521													
11	0.145774	0.139344	0.120647	0.091584	0.055801	0.019738													
12		0.133125	0.123132	0.103949	0.077325	0.046321	0.016148												
13	0.124666	0.120634	0.108809	0.090038	0.065879	0.038910	0.013396												
14		0.115435	0.108938	0.096328	0.078424	0.056598	0.033028	0.011249											
15	0.108913	0.106220	0.098280	0.085515	0.068688	0.048996	0.028297	0.009548											
16		0.101882	0.097423	0.088709	0.076164	0.060482	0.042714	0.024446	0.008180										
17	0.096703	0.094816	0.089232	0.080180	0.068071	0.053530	0.037475	0.021278	0.007066										
18		0.091171	0.087980	0.081713	0.072610	0.061053	0.047605	0.033073	0.018645	0.006151									
19	0.086960	0.085587	0.081512	0.074871	0.065899	0.054949	0.042527	0.029346	0.016439	0.005390									
20		0.082494	0.080132	0.075478	0.068675	0.059949	0.049624	0.038153	0.026168	0.014575	0.004752								
21	0.079003	0.077973	0.074910	0.069898	0.063079	0.054668	0.044961	0.034365	0.023441	0.012988	0.004214								
22		0.075323	0.073526	0.069977	0.064765	0.058031	0.049973	0.040864	0.031069	0.021088	0.011629	0.003755							
23	0.072383	0.071591	0.069231	0.065357	0.060060	0.053477	0.045791	0.037251	0.028187	0.019046	0.010457	0.003362							
24		0.069298	0.067900	0.065132	0.061054	0.055756	0.049366	0.042057	0.034054	0.025656	0.017264	0.009441	0.003023						
25	0.066788	0.066166	0.064309	0.061254	0.057062	0.051821	0.045653	0.038715	0.031216	0.023424	0.015703	0.008555	0.002729						
26		0.064165	0.063055	0.060856	0.057606	0.053368	0.048225	0.042293	0.035717	0.028687	0.021449	0.014328	0.007779	0.002473					
27	0.061997	0.061499	0.060012	0.057562	0.054188	0.049954	0.044939	0.039249	0.033021	0.026428	0.019693	0.013112	0.007096	0.002249					
28		0.059739	0.058844	0.057067	0.054437	0.050995	0.046801	0.041932	0.036487	0.030590	0.024403	0.018127	0.012033	0.006492	0.002051				
29	0.057848	0.057444	0.056235	0.054239	0.051486	0.048018	0.043891	0.039180	0.033975	0.028394	0.022582	0.016726	0.011072	0.005957	0.001877				
30		0.055884	0.055151	0.053696	0.051538	0.048707	0.045244	0.041205	0.036657	0.031688	0.026404	0.020941	0.015469	0.010213	0.005480	0.001722			
31	0.054221	0.053887	0.052891	0.051245	0.048969	0.046095	0.042662	0.038723	0.034343	0.029601	0.024598	0.019458	0.014337	0.009442	0.005054	0.001584			
32		0.052496	0.051889	0.050682	0.048890	0.046533	0.043644	0.040259	0.036430	0.032217	0.027694	0.022955	0.018114	0.013315	0.008749	0.004672	0.001461		
33	0.051021	0.050744	0.049913	0.048539	0.046637	0.044229	0.041345	0.038023	0.034309	0.030261	0.025948	0.021456	0.016893	0.012391	0.008123	0.004328	0.001350		
34		0.049496	0.048987	0.047975	0.046470	0.044489	0.042053	0.039192	0.035942	0.032346	0.028459	0.024346	0.020087	0.015781	0.011551	0.007557	0.004018	0.001251	
35	0.048179	0.047945	0.047245	0.046087	0.044481	0.042445	0.040000	0.037175	0.034003	0.030527	0.026797	0.022875	0.018834	0.014767	0.010787	0.007044	0.003738	0.001161	
36		0.046820	0.046389	0.045532	0.044257	0.042575	0.040504	0.038065	0.035285	0.032197	0.028840	0.025262	0.021520	0.017685	0.013839	0.010091	0.006577	0.003483	0.001080