A Sequential Boothstrap/Resampling Method

Silvia Shamrna[1], Mian Arif Shams Adnan[2]
[1,2]Bowling Green State University, Bowling Green, OH, 43402

Abstract
Since a rare observation resembles not only an unknown probability distribution but also their unknown characteristics, it is better to construct a basket of several characteristics based on subsets of observations. In this paper we offer a sequential-resampling method for generating several correlated observations from the same distribution from where the original sample is drawn. It is a kind of check and balance method for resampling each successive observation.

Key Words: Average Log Likelihood Function ; Combination ; Dummy Missing Value ; Likelihood Rate ; Simple Random Sample.

1. Introduction

In statistics, bootstrapping is any test or metric that relies on random sampling with replacement. Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using random sampling methods. Generally, it falls in the broader class of resampling methods.

Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution. One standard choice for an approximating distribution is the empirical distribution function of the observed data. In the case where a set of observations can be assumed to be from an independent and identically distributed population, this can be implemented by constructing a number of resamples with replacement, of the observed dataset (and of equal size to the observed dataset).

It may also be used for constructing hypothesis tests. It is often used as an alternative to statistical inference based on the assumption of a parametric model when that assumption is in doubt, or where parametric inference is impossible or requires complicated formulas for the calculation of standard errors.

Mian et al (2017) proposed a missing value estimation method which is very similar to bootstrap method. I may also be called as Sequential Bootstrap Resampling Method.

2. Methodology

Let there be n observations and 2 missing observations (2002). We want to resample paired missing observations if those were not missing. We know nothing about missing value or the distribution of observation from where the observations are drawn. So, we know nothing about the new generating paired values, or

the distribution of the observations or the parameters of the distribution or other characteristics like mean, median, mode, variance, skewness, kurtosis, and higher order moments of the distribution. In this situation we will estimate all the aforesaid characteristics and their volatility due to the change of sample size. We will also measure the deviation of the estimated characteristics from those of the missing values. So, we adjust our estimates of various characteristics due to the exact sample size and bandwidth of each of the characteristics. Later all the estimated characteristics will be used to find out several relations among themselves to predict the probability distribution. The parameters will also be estimated under the predicted probability distribution. Later on the deviation of the theoretically estimated characteristics and practically observed characteristics can be found to check how better the predicted distribution was by checking the equivalence of the theoretical and observed characteristics. Average Maximum Likelihood function and the consistent rate of the mean sum of squares of error can be found to confirm that the performance of the estimated generated (or missing) values and the error conducted due to the estimated missing values is the least.

2.1 Generating First Resampled Observation (Estimating First Missing Value) from a Sample of Size n

Let the observations $x_1, x_2, \ldots, x_{n-2}$ be non-missing and two observations to be resampled. Let the generated observation be y and z. We want to estimate y and z. So out of $(n-2)$ non-missing observations $n-2_{C_{n-2-2}}$ samples each of size $(n-2-2)$ can be drawn. We can generate $n-2_{C_{n-4}}$ samples each of which is consisting of $(n-4)$ non-missing observations pretending the rest non-missing observations as the missing observation. So, the $n-2_{C_{n-4}}$ generated samples are as below:

$n-2_{C_{n-4}}$ **samples each of size** $(n-4)$	**Assumed missing observation**
$x_1, x_2, \ldots, x_{n-2}$	x_{n-1}, x_n
$\ldots$	$\ldots$
$x_1, x_3, \ldots, x_{n-1}$	x_2, x_n
$x_3, \ldots, x_n$	x_1, x_2

So, we have calculated a class of characteristics (demonstrated in Table 1) to develop and observe several relationships among themselves (characteristics). For each of these characteristics, we will observe its deviation from the same characteristic with the presence of two dummy missing observations. Let us at first explain the easiest characteristic say sample mean and its sample standard deviation from the assumed missing value as addressed in Table 2.

Now,
$$L = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \ldots f(x_{n-2}; \bar{x}, S^2)$$

$$\log(L) = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \ldots f(x_{n-2}; \bar{x}, S^2)]$$

$$\log(L) = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2)) + \cdots + log(f(x_{n-2}; \bar{x}, S^2))$$

$$\therefore \frac{1}{n-2}\log(L) = \frac{1}{n-2}\sum_{i=1}^{n-2} \log(f(x_i; \bar{x}, S^2))$$

which can be termed as the average expected log likelihood function or expected log likelihood rate. Now, we should generate short incremented (various) values for x form the following range

$$\left(\frac{1}{n-2}\sum_{i=1}^{n-2} x_i - \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}}, \frac{1}{n-2}\sum_{i=1}^{n-2} x_i + \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}}\right).$$

Here k may be 0.50 or 1 or 2 or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values could be as below

$$\frac{1}{n-2}\sum_{i=1}^{n-2} x_i - \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}},$$

$$\frac{1}{n-2}\sum_{i=1}^{n-2} x_i - \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}} + h,$$

$$\frac{1}{n-2}\sum_{i=1}^{n-2} x_i - \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}} + 2h,$$

$$\frac{1}{n-2}\sum_{i=1}^{n-2} x_i - \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}} + 3h,$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots,$$

$$\frac{1}{n-2}\sum_{i=1}^{n-2} x_i + \mathrm{k}\frac{\overline{|x_1 - \overline{x_1}'|} + \overline{|x_2 - \overline{x_2}'|} + \cdots + \overline{|x_{n-2C_{n-4}} - \overline{x}_{n-2C_{n-4}}'|}}{n-2C_{n-4}}.$$

If we assume one of the aforesaid two observations to be resampled as the estimate of the *n-1*th pretended missing observation, and (if we consider) the available original observations $x_{1,}x_{2,} \dots, x_{n-2}$ as the $(n-2)$ other non-missing observations then the consecutive Maximum Likelihood Function or Likelihood Rate will be

$$L' = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_{n-2}; \bar{x}, S^2) f(x_{n-1}; \bar{x}, S^2)$$

$$\log(L') = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_{n-2}; \bar{x}, S^2) f(x_{n-1}; \bar{x}, S^2)]$$

$$\log(L') = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2) + \cdots + log(f(x_{n-1}; \bar{x}, S^2))$$

$$\frac{1}{n-1}\log(L') = \frac{1}{n-1}\sum_{i=1}^{n-1}\log(f(x_i; \bar{x}, S^2))$$

We will search the incremented value of the *n-1*th observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\frac{1}{n-2}\log(L) = \frac{1}{n-2}\sum_{i=1}^{n-2}\log(f(x_i; \bar{x}, S^2)) \cong \frac{1}{n-1}\log(L') = \frac{1}{n-1}\sum_{i=1}^{n-1}\log(f(x_i; \bar{x}, S^2)).$$

The incremented value of the *n-1*th observation for which the likelihood functions are same, will be an efficiently-estimated value of the n-1th resampled observation.

However, if we get more than two estimates of the resampled observation, we can check for which estimate of the resampled observation the first two moments are close to those of the original $(n-2)$ observations. Hence, we will find the closer estimate of the resampled observation. Therefore, if we get more than two or three or more estimates of a resampled observation, we can use all the estimates to estimate that resampled observation. Hence, we will estimate the (n-1)th resampling observation which is the estimate of one resampled value out of two resampled observation.

So, we have described how $n-2_{C_{n-2-2}}$ samples have been generated assuming two non-missing observations as two missing ones in each case and calculated their sample averages to find out a bandwidth for the first missing value. Here the missing value has been determent adding the half of the bandwidth of the 1st missing value with the average of all of the available non-missing values. Similarly, several sample characteristics and their bandwidth can be calculated to find out different characteristics of the missing data as well as the distribution from which the sample (consisting of the 1st missing value and non-missing value) has been drawn. So, sample variance, sample higher order moments, sample median, mode, skewness, kurtosis, tail behaviors, etc. can be found using their respective bandwidth. Several relationships can be explored from the aforesaid estimated characteristics to recognize the pattern of the distribution and its relevant features. The relevant features, estimated parameters and the predicted distribution are used to fit the observed sample data. So least square fitting or least deviation fitting or any sort of other goodness of fit can be used to check the performance of the predicted probabilistic model along-with the bandwidth based estimated parameters and the characteristics. After checking the fitting performance of the predicted model for the observed data, we can observe whether the average log-likelihood function for both the non-missing and the first missing or resampled observation is equivalent that of the average log-likelihood rate for the all non-missing values.

After estimating the first resampled observation, we will estimate the 2nd as well as the last resampled observation based on the non-missing values and the estimated 1st resampled observation. Hence, we will repeat the previously developed method of resampling one observation as follows.

2.2 Generating Last Resampled Observation (Estimating Last Missing Value) from a Sample of Size *n* using the First Generated Observation

Suppose there are n observations out of which $(n-1)$ non-missing observations and one observation to be generated. We also suppose that observations $x_1, x_2, \ldots, x_{n-1}$ are non-missing and one observation x_n is to be resampled. We want to estimate x_n. So out of $(n-1)$ non-missing observations, $(n-1)$ samples each of which is of size $(n-2)$ can be drawn assuming each sample has one missing observation. Assuming one non-missing observation as a missing one we can generate $(n-1)$ samples each of which is consisting of $(n-2)$ non-missing observations pretending the rest non-missing observations as the missing observation. So, the $(n-1)$ generated samples are as below:

$(n-1)$ samples each of size $(n-2)$	Assumed missing observation
$x_1, x_2, \ldots, x_{n-2}$	x_{n-1}
$x_1, x_2, \ldots, x_{n-1}$	x_{n-2}
$\ldots$	$\ldots$
$x_1, x_3, \ldots, x_{n-2}$	x_2
$x_2, x_3, \ldots, x_{n-1}$	x_1

So, we have calculated a class of characteristics (demonstrated in Table 1) to develop and observe several relationships among themselves (characteristics). For each of these characteristics, we will observe its deviation from the same characteristic with the presence of dummy missing observation. Let us at first explain the easiest characteristic say sample mean and its deviation from the assumed missing value as addressed in Table 2. Now,

$$L = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_{n-1}; \bar{x}, S^2)$$

$$\log(L) = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_{n-1}; \bar{x}, S^2)]$$

$$\log(L) = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2)) + \cdots + log(f(x_{n-1}; \bar{x}, S^2))$$

$$\therefore \frac{1}{n-1}\log(L) = \frac{1}{n-1}\sum_{i=1}^{n-1} \log(f(x_i; \bar{x}, S^2))$$

which can be termed as the average expected log likelihood function or expected log likelihood rate. Now, we should generate short incremented (various) values for x form the following range

$$\left(\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1}, \frac{1}{n-1}\sum_{i=1}^{n-1} x_i + \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} \right).$$

Here *k* may be 0.50 or 1 or 2 or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values could be as below

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1},$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} + h,$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} + 2h,$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i - \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1} + 3h,$$

$$\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots,$$

$$\frac{1}{n-1}\sum_{i=1}^{n-1} x_i + \mathrm{k}\frac{\substack{|\overline{x_1}-x_{n-1}|+|\overline{x_2}-x_{n-2}|+\cdots \\ +|\overline{x_{n-2}}-x_2|+|\overline{x_{n-1}}-x_1|}}{n-1}.$$

If we assume any of the aforesaid observations as the estimate of the n^{th} pretended missing observation, and (if we consider) the available original observations $x_1, x_2, \dots, x_{n-1}$ as the $(n-1)$ other non-missing observations then the consecutive Maximum Likelihood Function or Likelihood Rate will be

$$L' = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_n; \bar{x}, S^2)$$

$$\log(L') = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) \dots f(x_n; \bar{x}, S^2)]$$

$$\log(L') = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2) + \cdots + log(f(x_n; \bar{x}, S^2))$$

$$\frac{1}{n}\log(L') = \frac{1}{n}\sum_{i=1}^{n} \log(f(x_i; \bar{x}, S^2))$$

We will search the incremented value of the n^{th} observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\frac{1}{n-1}\log(L) = \frac{1}{n-1}\sum_{i=1}^{n-1} \log(f(x_i; \bar{x}, S^2)) \cong \frac{1}{n}\log(L') = \frac{1}{n}\sum_{i=1}^{n} \log(f(x_i; \bar{x}, S^2)).$$

The incremented value of the n^{th} observation for which the likelihood functions are same, will be an efficiently-estimated value of the missing or resampled observations. However, if we get more than two estimates of the resampled observation, we can check for which estimate of the resampled observation the first two moments are close to those of the original $(n-1)$ observations. Hence, we will find the closer estimate of the resampled observation. Therefore, if we get more than two or three or more estimates of a missing or resampled observation, we can use all the estimates to estimate that resampled observation. So, we have described how $(n-1)$ samples have been generated assuming one non-missing observation as a missing one in each case and calculated their sample averages to find out a bandwidth for the resampled observation. Here the resampled observation has been determent adding the half of the bandwidth of the resampled observation with the average of all the available non-missing values. Similarly, several sample characteristics and their bandwidth can be calculated to find out different characteristics of the resampled data as well as the distribution from which the sample (consisting of resampled value and non-missing or non-resampled observation) has been drawn. So, sample variance, sample higher order moments, sample median, mode, skewness, kurtosis, tail behaviors, etc. can be found using their respective bandwidth. Several relationships can be explored from the aforesaid estimated characteristics to recognize the pattern of the distribution and its relevant features. The relevant features, estimated parameters and the predicted distribution are used to fit the observed sample data. So least square fitting or least deviation fitting or any sort of other goodness of fit can be used to check the performance of the predicted probabilistic model along-with the bandwidth based estimated parameters and the characteristics. After checking the fitting performance of the predicted model for the observed data, we can observe whether the average log-likelihood function for both the non-missing and resampled observations is equivalent that of the average log-likelihood rate for the all non-missing values.

2.3 Generating First Resampled Observation from a Sample of Size 6

For more clarification let $n = 6$. So, there are 4 non-missing observations and 2 dummy missing observations. The non-missing observations are x_1, x_2, x_3, x_4 and the dummy missing observations are x_6 and x_5. Assuming two non-missing observations as missing ones we can generate 6 samples each of which is consisting of 2 non-missing observations assuming the rest non-missing observations as the missing observations. So, the 6 samples are as below:

Samples of size 2	Assumed missing observations
x_1, x_2	x_3, x_4
x_1, x_3	x_2, x_4
x_1, x_4	x_2, x_3

x_2, x_3	x_1, x_4
x_2, x_4	x_1, x_3
x_3, x_4	x_1, x_2

From table A1 of appendix, we have calculated a class of characteristics to develop and observe some relationships among them (characteristics). For each of these characteristics we will observe its deviation from the same characteristic with the presence of assumed missing observation. Let us at first explain the easiest characteristics say sample mean and its deviation from the assumed missing value in the table A2.

Now, $$L = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) f(x_3; \bar{x}, S^2) f(x_4; \bar{x}, S^2)$$

$$\log(L) = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) f(x_3; \bar{x}, S^2) f(x_4; \bar{x}, S^2)]$$

$$\log(L) = log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2)) + log(f(x_3; \bar{x}, S^2)) + log(f(x_4; \bar{x}, S^2))$$

$$\frac{1}{4}\log(L) = \frac{1}{4}\sum_{i=1}^{4} \log(f(x_i; \bar{x}, S^2))$$

which can be termed as the average expected likelihood or expected likelihood rate.

Now, we should generate short incremented various values form the range

$$\left(\begin{matrix} \frac{1}{4}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1} - \overline{x_1}'| + |\overline{x_2} - \overline{x_2}'| + |\overline{x_3} - \overline{x_3}'| + |\overline{x_4} - \overline{x_4}'| + |\overline{x_5} - \overline{x_5}'| + |\overline{x_6} - \overline{x_6}'|}{6}, \frac{1}{4}\sum_{i=1}^{4} x_i + \\ \mathrm{k}\frac{|\overline{x_1} - \overline{x_1}'| + |\overline{x_2} - \overline{x_2}'| + |\overline{x_3} - \overline{x_3}'| + |\overline{x_4} - \overline{x_4}'| + |\overline{x_5} - \overline{x_5}'| + |\overline{x_6} - \overline{x_6}'|}{6} \end{matrix} \right).$$

Here k may be 0.50 or 1 or 2 or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values the values could be

$$\begin{aligned} &\frac{1}{4}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1} - x_4| + |\overline{x_2} - x_3| + |\overline{x_3} - x_2| + |\overline{x_4} - x_1|}{4}, \\ &\frac{1}{4}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1} - x_4| + |\overline{x_2} - x_3| + |\overline{x_3} - x_2| + |\overline{x_4} - x_1|}{4} + h, \\ &\frac{1}{4}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1} - x_4| + |\overline{x_2} - x_3| + |\overline{x_3} - x_2| + |\overline{x_4} - x_1|}{4} + 2h, \\ &\frac{1}{4}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1} - x_4| + |\overline{x_2} - x_3| + |\overline{x_3} - x_2| + |\overline{x_4} - x_1|}{4} + 3h, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, \\ &\frac{1}{4}\sum_{i=1}^{4} x_i + \mathrm{k}\frac{|\overline{x_1} - x_4| + |\overline{x_2} - x_3| + |\overline{x_3} - x_2| + |\overline{x_4} - x_1|}{4}. \end{aligned}$$

If we assume any one of the two afore said observations as the 5th observation and the four other observations are the given original observations x_1, x_2, x_3, x_4; then the consecutive average observed likelihood or observed likelihood rate will be

$$L' = f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) f(x_3; \bar{x}, S^2) f(x_4; \bar{x}, S^2)\, f(x_5; \bar{x}, S^2)\, f(x_6; \bar{x}, S^2)$$

$$\log(L') = log[f(x_1; \bar{x}, S^2) f(x_2; \bar{x}, S^2) f(x_3; \bar{x}, S^2) f(x_4; \bar{x}, S^2) f(x_5; \bar{x}, S^2) f(x_6; \bar{x}, S^2)]$$

$$\begin{aligned} \log(L') = {} & log(f(x_1; \bar{x}, S^2)) + log(f(x_2; \bar{x}, S^2)) + log(f(x_3; \bar{x}, S^2)) \\ & + log(f(x_4; \bar{x}, S^2)) + log(f(x_5; \bar{x}, S^2) + log(f(x_6; \bar{x}, S^2)) \end{aligned}$$

$$\frac{1}{6}\log(L') = \frac{1}{6}\sum_{i=1}^{6}\log(f(x_i;\bar{x},S^2))$$

We will search the incremented value of the 5th observation for which the expected likelihood rate and the observed likelihood rate will be same i.e.

$$\frac{1}{4}\log(L) = \frac{1}{4}\sum_{i=1}^{4}\log(f(x_i;\bar{x},S^2)) \cong \frac{1}{6}\log(L') = \frac{1}{6}\sum_{i=1}^{6}\log(f(x_i;\bar{x},S^2)).$$

The incremented value of the 5th and 6th observations for which the likelihood functions are same, will be the estimated values of the first of the two missing or resampled observations.

If we get more than two estimates of the resampled observation (since we get two values of the 5th observation for whom the likelihood rates are same), we can check for which estimate of the resampled observation the first two moments are close to those of the original 4 observations. Hence, we will find the estimate of the resampled observations.

If we get more than two or three or more estimates of each of the resampled observations, we can have the corresponding averages all the estimates of the resampled observations and can assume that as the estimate of that resampled observation. Hence, we have derived the 5th observation. We will now estimate the 6th (last) observation.

2.4 Generating Last Resampled Observation from a Sample of Size 6

Now let $n = 6$. So there are 5 non-missing observations and one dummy missing observation. The non-missing observations are x_1, x_2, x_3, x_4, x_5 and the dummy missing observation is x_6. So, assuming one non-missing observation as a missing one we can generate 5 samples each of which is consisting of 4 non-missing observations assuming the rest non-missing observations as the missing observation. So, the 5 samples are as below:

Samples of size 4	**Assumed missing observation**
x_1, x_2, x_3, x_4	x_5
x_1, x_2, x_3, x_5	x_4
x_1, x_2, x_4, x_5	x_3
x_1, x_3, x_4, x_5	x_2
x_2, x_3, x_4, x_5	x_1

So, we have calculated a class of characteristics (Table A3) to develop and observe some relationships among them (characteristics). For each of these characteristics we will observe its deviation from the same characteristic with the presence of assumed missing observation. Let us at first explain the easiest characteristics say sample mean and its deviation from the assumed missing value in the Table A4.

Now, $$L = f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)f(x_5;\bar{x},S^2)$$

$$\log(L) = log[f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)f(x_5;\bar{x},S^2)]$$

$$\log(L) = log(f(x_1;\bar{x},S^2)) + log(f(x_2;\bar{x},S^2)) + log(f(x_3;\bar{x},S^2)) + log(f(x_4;\bar{x},S^2) + log(f(x_5;\bar{x},S^2))$$

$$\frac{1}{5}\log(L) = \frac{1}{5}\sum_{i=1}^{5}\log(f(x_i;\bar{x},S^2))$$

which can have been termed as the average expected log likelihood or expected log likelihood rate. Now, we should generate short incremented various values form the range

$$\begin{pmatrix} \frac{1}{5}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5}, \\ \frac{1}{5}\sum_{i=1}^{4} x_i + \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5} \end{pmatrix}.$$

Here k may be 0.50 or 1 or 2 or so on. The increment h can take the value 0.01 or 0.05 or 0.10 and so on. The values the values could be

$$\frac{1}{5}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5},$$

$$\frac{1}{5}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5} + h,$$

$$\frac{1}{5}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5} + 2h,$$

$$\frac{1}{5}\sum_{i=1}^{4} x_i - \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5} + 3h,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots,$$

$$\frac{1}{5}\sum_{i=1}^{4} x_i + \mathrm{k}\frac{|\overline{x_1}-x_4|+|\overline{x_2}-x_3|+|\overline{x_3}-x_2|+|\overline{x_4}-x_1|+|\overline{x_5}-x_5|}{5}.$$

If we assume any of the afore said observations as the 6th observation and the four other observations are the given original observations x_1, x_2, x_3, x_4, x_5; then the consecutive maximum likelihood function or observed likelihood rate will be

$$L' = f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)\, f(x_5;\bar{x},S^2)f(x_6;\bar{x},S^2)$$

$$\log(L') = log[f(x_1;\bar{x},S^2)f(x_2;\bar{x},S^2)f(x_3;\bar{x},S^2)f(x_4;\bar{x},S^2)f(x_5;\bar{x},S^2)f(x_6;\bar{x},S^2)]$$

$$\log(L') = log(f(x_1;\bar{x},S^2)) + log(f(x_2;\bar{x},S^2)) + log(f(x_3;\bar{x},S^2)) + log(f(x_4;\bar{x},S^2)) + log(f(x_5;\bar{x},S^2)) + log(f(x_6;\bar{x},S^2)))$$

$$\frac{1}{6}\log(L') = \frac{1}{6}\sum_{i=1}^{6}\log(f(x_i;\bar{x},S^2))$$

We will search the incremented value of the 6th observation for which the expected log likelihood rate and the observed log likelihood rate will be same i.e.

$$\frac{1}{5}\log(L) = \frac{1}{5}\sum_{i=1}^{5}\log(f(x_i;\bar{x},S^2)) \cong \frac{1}{6}\log(L') = \frac{1}{6}\sum_{i=1}^{6}\log(f(x_i;\bar{x},S^2)).$$

The incremented value of the 5th observation for which the likelihood functions are same, will be the estimated value of the missing or resampled observations. If we get more than two estimates of the resampled observation (since we get two values of the 5th observation for whom the likelihood rates are same), we can check for which estimate of the resampled value the first two moments are close to those of the original 4 observations. Hence, we will find the estimate of the resampled observations.

3. Real Life Examples

We like to simulate a couple of random observations of size n from a probability distribution with specified parameters. Later we will treat two observations completely missing and pull these out from the original sample. Hence the original sample turns to a sample of size $n-2$. Out of $n-2$ available observations of the sample, we will draw samples each of which is of size $n-2$. For each of the $n_{C_{n-2}}$ samples of size $n-2$, we will assume the two absent observations as two dummy missing values of the sample. So, for each of the $n_{C_{n-2}}$ samples, there are $n-2$ available observations and two dummy missing values. From each of the $n_{C_{n-2}}$ samples, we will have one absolute dispersion between the average of $n-2$ available observations and the average of the two dummy missing observations. So, we will have $n_{C_{n-2}}$ absolute between differences for $n_{C_{n-2}}$ pairs of averages and dummy missing values. Averaging the $n_{C_{n-2}}$ absolute differences, we will calculate average absolute difference. Based on the average absolute difference, we will generate a possible range of the original missing value. We will generate several values of that range starting from the lower limit and will get several valued for fixed increment upto to upper limit of that range. We will check whether the average likelihood of the $n-2$ original observations is similar for which n-1th, n^{th} observed missing values or the resampled observations from the generating range and the $n-2$ observations.

Let $n=10$. So there are 8 non-missing observations and two missing (assumed totally missing) observations. The non-missing observations (from Normal with mean 5 and standard deviation 2) are 1.729466, 3.547037, 3.6597, 5.814905, 3.817457, 6.333606, 4.05684, 3.748781, and the missing observations are 3.608116, 2.671239. The average of these eight non-missing observations are 4.09. Now, assuming two non-missing observations as two missing ones we can generate 28 samples each of which is consisting of 6 non-missing observations assuming the rest two non-missing observations as two dummy missing observations.

So, the 28 samples (as addressed in table A3) each consisting of 6 non-missing values are as given in the next page (the bold numbers in the last row are representing here the assumed missing value for each sample).

The Expected Log Likelihood Rate for 9 observations (8 non-missing and one from the generating interval) is -0.743. By using the formula shown above, we get the range as (2.1262, 6.1262); where k=2. Let the increment, h=0.1. For each increment we will get average likelihood rate for 9 observations. And for the incremented value = 2.726, we get the same value for the Expected Average Likelihood and Observed Average Likelihood. So, our estimated value of the 1st missing or resampled observation is 2.726.

Now depending on the 1st missing value and the missing value based, or 9 observations based mean and variance, the likelihood function and likelihood rate for 10 observations have been found. The Expected Log Likelihood Rate is -0.741. By using the formula shown above, we get the range as (2.6188619, 5.2688619); where k = 1.2. Let the increment, h=0.05. For each increment we will get average likelihood rate for 10 observations (8 non-missing, one estimate of the 1st missing or resampled and one from the generating interval for the 2nd resampled value or resampled observation). And for the incremented value = 2.62, we get the same value for the Expected Average Likelihood and Observed Average Likelihood. So, our estimated value of the 2nd resampled observation is 2.62.

So, the estimates of these two missing values or resampled observations are 3.608116, 2.671239 are 2.726 and 2.62.

Conclusion

The missing technique is a kind of check and balance method in estimating the missing value. It can also be termed as Sequential Bootstrap Method. In each step it checks the fluctuation due to sample size and balance it by capturing the dispersion of the estimate of the known data from the assumed unknown data which is really known. So, this method is trying to find the original rate of change of the deviation from the missing value for the exact size of the realized sample. So, from two directions, one direction from sample size and other direction for the deviation from the missing values, the missing technique has been aided to estimate the missing value efficiently maintaining a good performance through several goodness of fit tests. This paper demonstrates a resampling method for generating 1 or 2 correlated observations from the same distribution from where the original sample is drawn. This paper can also be extended to get a resampling method for ($n > 2$) three or more correlated observations.

References

Sharna, S, Adnan, M. A. S., and Imon, R. 2016. A Missing Technique for Estimating a missing value. In JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 398-409.

Sharna, S, Adnan, M. A. S., and Imon, R. 2016. A Missing Technique for Estimating a missing value. In JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 398-409.

Sharna, S. and Adnan, M. A. S. et al. (2017). A Missing Technique for Estimating Univariate Multiple Missing Values: An Advanced Resampling Method for Correlated Observations in JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association. 2522-2535.

Little. R. J. A, Rubin. D. B. (2002). Statistical Analysis with Missing Data. 2nd edition. Wiley Publishers.

Appendix

Table A1: Sample means and sample variances for several samples.

	Sample Mean	**Sample Variance**
	$\overline{x_1} = \frac{x_1+x_2}{2}$	$S_1^2 = \frac{(x_1-\overline{x_1})^2+(x_2-\overline{x_1})^2}{2-1}$
	$\overline{x_2} = \frac{x_1+x_3}{2}$	$S_2^2 = \frac{(x_1-\overline{x_2})^2+(x_3-\overline{x_2})^2}{2-1}$
	$\overline{x_3} = \frac{x_1+x_4}{2}$	$S_3^2 = \frac{(x_1-\overline{x_3})^2+(x_4-\overline{x_3})^2}{2-1}$
	$\overline{x_4} = \frac{x_2+x_3}{2}$	$S_4^2 = \frac{(x_2-\overline{x_4})^2+(x_3-\overline{x_4})^2}{2-1}$
	$\overline{x_5} = \frac{x_2+x_3}{2}$	$S_5^2 = \frac{(x_2-\overline{x_5})^2+(x_3-\overline{x_5})^2}{2-1}$
	$\overline{x_6} = \frac{x_3+x_4}{2}$	$S_6^2 = \frac{(x_3-\overline{x_6})^2+(x_4-\overline{x_6})^2}{2-1}$
Average	$\bar{x} = \frac{\overline{x_1}+\overline{x_2}+\overline{x_3}+\overline{x_4}+\overline{x_5}+\overline{x_6}}{6}$	$S^2 = \frac{S_1^2+S_2^2+S_3^2+S_4^2+S_5^2+S_6^2}{6}$

Table A2: Sample mean difference for several samples.

Sample Mean of size 3	**Assumed Missing Values**	**Difference**	$\|\boldsymbol{Difference}\|$
$\overline{x_1} = \frac{x_1+x_2}{2}$	x_3, x_4	$\overline{x_1} - \frac{x_3+x_4}{2}$	$\overline{\lvert x_1 - x_1' \rvert}$
$\overline{x_2} = \frac{x_1+x_3}{2}$	x_2, x_4	$\overline{x_2} - \frac{x_2+x_4}{2}$	$\overline{\lvert x_2 - x_2' \rvert}$
$\overline{x_3} = \frac{x_1+x_4}{2}$	x_2, x_3	$\overline{x_3} - \frac{x_2+x_3}{2}$	$\overline{\lvert x_3 - x_3' \rvert}$
$\overline{x_4} = \frac{x_2+x_3}{2}$	x_1, x_4	$\overline{x_4} - \frac{x_1+x_4}{2}$	$\overline{\lvert x_4 - x_4' \rvert}$
$\overline{x_5} = \frac{x_2+x_3}{2}$	x_1, x_3	$\overline{x_5} - \frac{x_1+x_3}{2}$	$\overline{\lvert x_5 - x_5' \rvert}$
$\overline{x_6} = \frac{x_3+x_4}{2}$	x_1, x_2	$\overline{x_6} - \frac{x_1+x_2}{2}$	$\overline{\lvert x_6 - x_6' \rvert}$
Total			$\overline{\lvert x_1 - x_1' \rvert} + \overline{\lvert x_2 - x_2' \rvert} + \overline{\lvert x_3 - x_3' \rvert} + \overline{\lvert x_4 - x_4' \rvert} + \overline{\lvert x_5 - x_5' \rvert} + \overline{\lvert x_6 - x_6' \rvert}$
Average			$\frac{\overline{\lvert x_1 - x_1' \rvert} + \overline{\lvert x_2 - x_2' \rvert} + \overline{\lvert x_3 - x_3' \rvert} + \overline{\lvert x_4 - x_4' \rvert} + \overline{\lvert x_5 - x_5' \rvert} + \overline{\lvert x_6 - x_6' \rvert}}{6}$

Table A3: The 28 samples each consisting of 6 non-missing values.

Sample	**Non-Missing Part**						**Missing Part**	
1	1.73	3.55	3.66	5.81	3.82	6.33	4.06	3.75
2	1.73	3.55	3.66	5.81	3.82	4.06	6.33	3.75
3	1.73	3.55	3.66	5.81	4.06	6.33	3.82	3.75
4	1.73	3.55	3.66	4.06	3.82	6.33	5.81	3.75
5	1.73	3.55	4.06	5.81	3.82	6.33	3.66	3.75
6	1.73	4.06	3.66	5.81	3.82	6.33	3.55	3.75
7	4.06	3.55	3.66	5.81	3.82	6.33	1.73	3.75
8	1.73	3.55	3.66	5.81	3.82	3.75	4.06	6.33
9	1.73	3.55	3.66	5.81	3.75	6.33	4.06	3.82
10	1.73	3.55	3.66	3.75	3.82	6.33	4.06	5.81
11	1.73	3.55	3.75	5.81	3.82	6.33	4.06	3.66
12	1.73	3.75	3.66	5.81	3.82	6.33	4.06	3.55
13	3.75	3.55	3.66	5.81	3.82	6.33	4.06	1.73
14	1.73	3.55	3.66	5.81	4.06	3.75	3.82	6.33
15	1.73	3.55	3.66	4.06	3.82	3.75	5.81	6.33
16	1.73	3.55	4.06	5.81	3.82	3.75	3.66	6.33
17	1.73	4.06	3.66	5.81	3.82	3.75	3.55	6.33
18	4.06	3.55	3.66	5.81	3.82	3.75	1.73	6.33
19	1.73	3.55	3.66	4.06	3.75	6.33	5.81	3.82
20	1.73	3.55	4.06	5.81	3.75	6.33	3.66	3.82
21	1.73	4.06	3.66	5.81	3.75	6.33	3.55	3.82
22	4.06	3.55	3.66	5.81	3.75	6.33	1.73	3.82
23	1.73	3.55	4.06	3.75	3.82	6.33	3.66	5.81
24	1.73	4.06	3.66	3.75	3.82	6.33	3.55	5.81
25	4.06	3.55	3.66	3.75	3.82	6.33	1.73	5.81
26	1.73	4.06	3.75	5.81	3.82	6.33	3.55	3.66
27	4.06	3.55	3.75	5.81	3.82	6.33	1.73	3.66
28	4.06	3.75	3.66	5.81	3.82	6.33	1.73	3.55

Table A4: The Bandwidth for each of the 28 samples.

Sample #	Sample Mean	First Missing Value	Second Missing Value	Absolute Difference or Bandwidth
1	4.15	4.06	3.75	0.247551
2	3.77	6.33	3.75	1.270293
3	4.19	3.82	3.75	0.40714
4	3.86	5.81	3.75	0.924492
5	4.22	3.66	3.75	0.512311
6	4.24	3.55	3.75	0.58742
7	4.54	1.73	3.75	1.799134
8	3.72	4.06	6.33	1.475665
9	4.14	4.06	3.82	0.201767
10	3.81	4.06	5.81	1.129865
11	4.17	4.06	3.66	0.306939
12	4.18	4.06	3.55	0.382047
13	4.49	4.06	1.73	1.593761
14	3.76	3.82	6.33	1.316077
15	3.43	5.81	6.33	2.647709
16	3.79	3.66	6.33	1.210905
17	3.80	3.55	6.33	1.135797
18	4.11	1.73	6.33	0.075917
19	3.85	5.81	3.82	0.970276
20	4.21	3.66	3.82	0.466527
21	4.22	3.55	3.82	0.541636
22	4.53	1.73	3.82	1.75335
23	3.87	3.66	5.81	0.865105
24	3.89	3.55	5.81	0.789996
25	4.19	1.73	5.81	0.421718
26	4.25	3.55	3.66	0.646807
27	4.55	1.73	3.66	1.858521
28	4.57	1.73	3.55	1.93363
Average	4.09			0.981156