Some Applications of the Resampling Methods in Computational Physics

Sotiraq Marko[#1], Lorenc Ekonomi[*2]

[#1]*Physics Department, University of Korca, Albania,* [*2]*Mathematics Department, University of Korca, Albania*

***Abstract*—The statistical methods are being very important for estimating the unknown parameters in computational physics. Between them we can mention the resampling methods: the jackknife and bootstrap estimate. In our work we have done their ideas and have shown the results of some applications in physics problems of parameter estimation.**

Keywords—**Jackknife, bootstrap, bias, computational physics, Binder ratio, Lorentzian.**

I. INTRODUCTION

The physicians use various statistical methods in their works. But the books on these topics usually fall into one of two camps. At one extreme, the books for physicians don't discuss all that is needed and rarely prove the results that they quote. At the other extreme, the books for mathematicians presumably prove everything but are written in the style of lemmas, proofs and unfamiliar notation which is intimidating to physicians. For the exception, there are some works which find in a good middle group [2], [6], [10]. In the following we have treated some application of the resampling methods in physic problems. Let us see the statistical problem of the mean estimation of a random variable X.

Let suppose that $x_1, x_2,\ldots,x_n$ are independent observations from the random variable X. We note the mean of the random variable X by μ and the sample mean by $\overline{x}_n$, where $\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. If we denote X_i, i=1,2,..,n the random variable X in the i-th observation, we propose the statistic

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (1)$$

to be an estimator for the unknown parameter μ. Since $E\overline{X}_n = \mu$, the statistic (1) is a unbiased estimator for the unknown parameter μ. If we denote the variance of the random variable X by σ^2, we have $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$. We can use $\sqrt{\mathrm{Var}(\overline{X})} = \frac{\sigma}{\sqrt{n}}$ to measure the uncertainty in the sample mean or the error bar estimate. Hence our estimate of μ is

$$\overline{x} \pm \frac{\sigma}{\sqrt{n}}. \qquad (2)$$

Furthermore, $\overline{X}_n$ converges to μ. We can say that $\overline{X}_n$ become more and more accurate as the number of observations increase. Now let suppose that we want to estimate not μ, but some function of μ i.e. $g(\mu)$. In Section 2 we have done standard statistical methods for the estimation of $g(\mu)$ and have analyzed the bias of the estimation and its order. In Section 3 we have shown the idea of the jackknife and bootstrap estimations and have given some consideration about the jackknife and bootstrap estimation of the bias. We have stressed that the bias order of jackknife and bootstrap estimate is $\frac{1}{n^2}$ instead of $\frac{1}{n}$ in standard estimations. In Section 4 we have shown some simulations cases and have done a comparison between the jackknife and bootstrap methods with standard methods.

II. THE BIAS ORDER IN STANDARD METHODS

In the above conditions, we compute some statistic of interest, say $\hat{\theta} = \hat{\theta}(x_1,..., x_n)$ and define the bias in the form

$$\text{Bias} = E\hat{\theta} - \mu, \tag{3}$$

where $E\hat{\theta}$ is the mean of $\hat{\theta}$. Now let us analyze the estimation of the unknown parameter $g(\mu)$. A poor way to estimate $g(\mu)$ would be from

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \tag{4}$$

However, this is really an estimate for the mean of g(X), rather than $g(\mu)$. But, in general $Eg(X) \neq g(\mu)$. Let we evaluate the difference $Eg(X) - g(\mu)$. For the g(x) we have

$$g(x) = g(\mu) + \frac{(x-\mu)g'(\mu)}{1!} + \frac{(x-\mu)^2 g''(\mu)}{2!} + ... \tag{5}$$

We can see that the bias is equal to

$$Eg(x) - g(\mu) = \frac{g''(\mu)\text{var}X}{2!} + ... = O(1). \tag{6}$$

So, the bias does not vanish for $n \to \infty$. If g is a linear function then $g'' \equiv 0$ and Bias=0. Thus, there is no bias if g is a linear function.

We take a better result if we change in (5) g(x) with $g(\bar{x})$. In this case

$$g(\bar{x}) = g(\mu) + \frac{(\bar{x}-\mu)g'(\mu)}{1!} + \frac{(\bar{x}-\mu)^2 g''(\mu)}{2!} + \cdots. \tag{7}$$

and

$$Eg(\bar{x}) - g(\mu) = \frac{g''(\mu)\text{var}X}{n \cdot 2!} + ... = O\left(\frac{1}{n}\right). \tag{8}$$

Now, the bias is of order $\frac{1}{n}$ rather than the order 1. This bias can usually be neglected because it is smaller than statistical error in (2) of order $\frac{1}{\sqrt{n}}$. To decrease the bias order we use the jackknife and the bootstrap methods. In the following section we have given the idea of the jackknife and the bootstrap estimations.

III. THE BIAS ESTIMATE WITH RESAMPLING METHODS

The resampling methods are getting an important space in many statistical problems of estimating unknown parameters and distributions. It depends on two reasons. The first reason is the use of computers and software and the second reason is that these methods not ask any condition about the distributions. Let us see the idea of the jackknife and bootstrap estimate.

A. The Jackknife Estimate of the Bias

We define the observation mean when we have removed the observation x_i in the form $\bar{x}_{(i)} = \frac{1}{n-1}\sum_{j\neq i} x_i$, i=1,…,n. In similar way we define $g_{(i)} = g(\bar{x}_{(i)})$, i=1,…,n. The jackknife estimate of $g(\mu)$ is the average of $g_{(i)}$, i=1,…,n or

$$g_J(\mu) = \frac{1}{n}\sum_{i=1}^{n} g_{(i)} \tag{9}$$

Let us analyze the jackknife estimate. We have

$$\bar{x}_{(i)} = \mu + \frac{1}{n-1}\sum_{j\neq i}(x_i - \mu)\ , \ i=1,\ldots,n. \tag{10}$$

Then

$$g(x_{(i)}) = g\left(\mu + \frac{1}{n-1}\sum_{j\neq i}(x_i-\mu)\right) = g(\mu) + \frac{g'(\mu)}{n-1}\sum_{j\neq i}(x_i-\mu) + \frac{g''(\mu)}{(n-1)^2 2!}\sigma^2 + \ldots$$

and

$$Eg(x_{(i)}) = g(\mu) + \frac{g''(\mu)}{2(n-1)}\sigma^2 + \ldots$$

The bias of the jackknife estimate is

$$Eg_J(\mu) - g(\mu) = \frac{g''(\mu)}{2(n-1)}\sigma^2 + \ldots \tag{11}$$

We see that high order terms are at order $\frac{1}{n}$. The bias vanishes for $n \to \infty$ and it is of the same order with the bias (8) of the $g(\bar{X})$ estimate.

Sometimes we want to estimate directly the bias. To do it, let us see Quenouille's bias estimate [9]. This method is based on sequentially deleting points x_i and recomputing the statistic $\hat{\theta}$. Denoting $\hat{\theta}_{(i)} = \hat{\theta}(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$ and $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$, Quenouille's estimate of bias is

$$\hat{Bias}_J = (n-1)(\hat{\theta}_{(.)} - \hat{\theta}), \tag{12}$$

leading to the bias-corrected "jackknife estimate" of μ

$$\tilde{\theta} = \hat{\theta} - \hat{Bias}_J = n\hat{\theta} - (n-1)\hat{\theta}_{(.)}\ .$$

For many common statistics, including most maximum likelihood estimates,

$$E\hat{\theta} = \mu + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots\ , \tag{13}$$

where $a_1, a_2, \ldots$ do not depend upon n [8]. After some calculations we have

$$E\tilde{\theta} = \theta - \frac{a_2}{n(n-1)} + a_3\left(\frac{1}{n^2} - \frac{1}{(n-1)^2}\right) + \ldots . \tag{14}$$

We see that the bias order of $\tilde{\theta}$ is $O\left(\frac{1}{n^2}\right)$, compared to $O\left(\frac{1}{n}\right)$ for the original estimator (13). The jackknife estimate of variance is [11]

$$\hat{var}_J = \frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)} - \hat{\theta}_{(.)}\right)^2 \tag{15}$$

Suppose $\hat{\theta} = g(\bar{X})$, where g is some nicely behaved function (derivative g' exists continuously). Then a first order Taylor expansions gives $\hat{\theta}_{(i)} = g(\bar{x}) + g'(\bar{x})\frac{\bar{x} - x_i}{n-1}$. So, substituting this expression into (15), we have

$$\hat{\text{var}} = \frac{1}{n(n-1)}\left[g(\bar{x})\right]^2 \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 . \quad (16)$$

B. The Bootstrap Estimate of the Bias

The bootstrap [3] is conceptually a simple technique. The bootstrap, like the jackknife, is a resampling of n data points x_i. Whereas jackknife considers n new data sets, each containing all the original data points minus 1, bootstrap uses B data sets each containing n points obtained by random sampling with the same probability $\frac{1}{n}$ of the original set of n points.

Let us note such data set by $X_1^*,...,X_n^*$ and calculate $\hat{\theta}^* = (X_1^*,...,X_n^*)$. This is the bootstrap estimate for the unknown parameter. Independently we repeat this procedure a large number B of times obtaining "bootstrap replications" $\hat{\theta}^{*1},...,\hat{\theta}^{*B}$. The bootstrap estimate for the bias is

$$\hat{\text{Bias}}_B = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*b} - \hat{\theta} \quad (17)$$

and the bootstrap estimate for the variance is

$$\hat{\text{var}}_B = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*b} - \hat{\theta})^2 \quad (18)$$

The order of the bias in the bootstrap estimate is the same with the order of the jackknife estimate bias. For the quadratic functional we have

$$\hat{\text{Bias}}_J = \frac{n-1}{n}\hat{\text{Bias}}_B . \quad (19)$$

IV. EXAMPLES

In this Section we have shown the results of some applications of resampling methods in Computational Physics.

1) Example 4.1

Let us suppose we want to compute cos(E(X)), where $x_i = \frac{\pi}{3} + \varepsilon_i$ and ε_i is a Gaussian random variable with mean zero and standard deviation unity. We took a sample n=1000.

The jackknife estimate of the mean was 0.4916, with an error estimate of 0.0280 which is consistent with the exact value of 0.5. For comparison $\cos\bar{x}$ was also equal to 0.4916 to this precision. Using equation (12) to get a less biased estimate of the mean, and using the full precision of the numbers in the computer gives 0.4919. The difference between 0.4916 and 0.4919 is completely unimportant since the error bar of 0.0280 is very much larger.

With B=100 data sets we found 0.4927 ± 0.0279. This result is consistent with the exact value of 0.5 and very close to the jackknife result.

2) Example 4.2

Jackknife and bootstrap can be used to compute error bars for quite general functions of the data set. For example, one can use the jackknife and bootstrap resampling schemes to estimate parameters describing the shape of the distribution from the data set. An example, which is also used in many researches in phase transitions (where it is called the "Binder ratio"), g, defined by

$$g = \frac{E\left(X - EX\right)^4}{\left(E\left(X - EX\right)^2\right)^2} . \quad (20)$$

Since the total power of X in the numerator and denominator is the same, the kurtosis depends only on the shape of the distribution and not on its overall scale. It takes the value 3 for a Gaussian distribution.

In the example we took n=1000 points from a Gaussian distribution and computed the kurtosis using the jackknife method. For the Gaussian distribution we know that EX=0. For each of the n jackknife data sets we computed g and obtained an average and error bar using (12) above. The result is 3.090 ± 0.145 which is consistent with the exact value of 3.

The kurtosis is found for each B=100 bootstrap samples, and the mean and error obtained from (17). The result is 3.072 ± 0.1320 which agrees well with the jackknife estimate and is consistent with the exact value of 3.

3) Example 4.3 Linear regression:

Consider the linear regression model

$$y_i = x_i\beta + \varepsilon_i \ , \ i=1,\ldots,n, \tag{21}$$

where ε_i are independent random variables with identically unknown distribution F and $E(\varepsilon)=0$, x_i is a known 1xp vector of covariates when $y_i = x_i\beta + \varepsilon_i$, where β is a px1 vector of unknown parameters.

The statistic of interest is the least squares estimate of β in the form

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \ , \tag{22}$$

where $Y = \left(y_1,\ldots, y_n\right)^T$, $X = \left(X_1,\ldots,X_n\right)^T$. The usual estimate of $\text{cov}(\hat{\beta})$ is

$$\hat{\sigma}^2\left(X^T X\right)^{-1}, \tag{23}$$

where $\hat{\sigma}^2 = \frac{n-1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ and $\hat{\varepsilon}_i$ being the estimated residual $y_i - x_i\hat{\beta}$. The multivariate version of Tukey's formula is

$$\overset{\wedge}{\text{cov}}_J = \frac{n-1}{n}\sum_{i=1}^{n}\left[\hat{\beta}_{(i)} - \hat{\beta}_{(.)}\right]\left[\hat{\beta}_{(i)} - \hat{\beta}_{(.)}\right]. \tag{24}$$

If all the $\hat{\varepsilon}_i^2$ are identical in value this is about the same as the standard answer, but otherwise the two formulas are quite different. In quadratic fit of 50 points with Gaussian noise to $3-2x+x^2$, the following results were obtained

TABLE I

THE ESTIMATE RESULTS OF THE LINEAR REGRESSION WITH STANDARD AND REAMPLING METHODS

	Least square fit			**Jackknife fit**			**Bootstrap fit**		
Fit parameters	3.0121	-1.9784	1.0021	3.0122	-1.9783	1.0020	3.0123	-1.988	1.0012
Error bars	0.0212	0.0981	0.0949	0.0190	0.0847	0.0778	0.0183	0.0838	0.0878
Covariance matrix	0.0004	-0.0018	0.0015	0.0003	-0.0014	0.0011	0.0003	-0.0016	0.0012
	-0.0018	0.0096	-0.0090	-0.0014	0.0071	-0.0064	-0.0016	0.0073	-0.0082
	0.0015	-0.0090	0.0090	0.0011	-0.0064	0.0060	0.0012	-0.0082	0.0058

4) Example 4.1 The simple median

For most problems, the jackknife and bootstrap give similar results. However, there is at least one class of problems where the jackknife approach is unsatisfactory, because the data set are too similar to each other, while the bootstrap method works. We can note that the jackknife estimate of the variance fails in the case of the sample median. An estimator for the sample median is

$$\overset{\wedge}{\theta} = \begin{cases} x_{(m)} & \text{if } n = 2m-1 \\ \dfrac{x_{(m)} + x_{(m+1)}}{2} & \text{if } n = 2m \end{cases}. \tag{25}$$

From the formula (13) we have

$$\hat{\text{var}}_J = \frac{n-1}{4}\left[x_{(m+1)} - x_{(m)}\right]^2 . \tag{26}$$

Standard theory [7] shows that if the distribution F of the random variable X has a density function f then

$$n \cdot \hat{\text{var}} \xrightarrow{n\to\infty} \frac{1}{4f^2(\mu)} Y , \tag{27}$$

where $f(\mu)$ is the density at the sample median μ, $f(\mu)$ is assumed >0 and Y is a random variable with expectation 2 and variance 20.

The true variance goes to the limit [5]

$$n \cdot \text{var} \xrightarrow{n\to\infty} \frac{1}{4f^2(\mu)} \tag{28}$$

In this case, the jackknife estimate is not even a consistent estimator of the variance sample median.

From the other hand, the bootstrap estimate of the variance goes well for the sample median. The bootstrap estimate of standard deviation is shown to be asymptotically consistent for the true standard deviation [4]. In the bootstrap estimate, the B data samples are significantly different from each other, so the error in the median can be estimated. As an example, we took n=1001 data points generated from the positive half of a Lorentzian

$$f(x) = \begin{cases} \dfrac{2}{\pi\left(1+x^2\right)} & x > 0 \\ 0 & x < 0 \end{cases} . \tag{29}$$

Note that this is a very broad distribution for which even the mean is not defined. However, the median is defined and a standard integral gives the value 1. Including all 1001 values of we have x_i, we found the median to be 0.9613. Using the bootstrap with B=5000 data sets we found 0.9603 ± 0.0507. We see that the overall average and the mean of the bootstraps are very close, and the result agrees with the exact value of 1 within the error bar.

V. Conclusions

In this paper we have given some applications of jackknife and bootstrap in Computational Physics. We have shown the efficiency and facility of these methods in estimation when the distributions are unknown or asymmetric. We can see it in Examples 4.1, 4.2 and 4.3. But in some cases the resampling methods do not work. We can see it in Example 4.4, when the bootstrap method works, meanwhile the jackknife method does not work.

References

[1] V. Ambegaokar and M. Troyer M. "Estimating errors reliably in Monte Carlo simulations of the Ehrenfest model," Am. J. Phys. Vol. 78, 150, 2009.

[2] G. Bohm and G. Zech G (2010) *Introduction to statistics and data analysis for physicists.* Verlag Deutches Electronen Synchrotron, 2010.

[3] B. Efron "Bootstrap methods: another look at the jackknife" Ann. Statist., Vol. 7, pp. 1-26, 1979.

[4] B. Efron *The jackknife, the bootstrap and other resampling plans.* SIAM, 1982.

[5] M. Kendall and A. Stuart *The advanced theory of statistics.* Griffin. London, 1958.

[6] W. H. Press, S. A. Teukolsky, W. T. Wetterling and B. P. Flannery *Numerical recipes in C,* 2 end Ed. Cambridge University Press, 1992.

[7] R. Pyke "Spacings," J. Roy. Statist. Soc. Ser. B. Vol. 27, pp. 395-449, 1965.

[8] W. Schucany, H. Gray and O. Owen "On bias reduction in estimation" JASA, Vol. 66, pp. 524-533, 1971.

[9] M. Quenouille "Approximate tests of correlation time series" J. Roy. Statist. Soc. Ser. B, Vol. 11, pp 11-84, 1949.

[10] J. R. Taylor *The study of uncertainties in physical measurements.* University Science Books, Sausalito, California, 1997.

[11] J. Tukey "Bias and confidence in not quite large samples" Abstract. Ann. Math. Statist., Vol 29, pp. 614, 1958.