International Journal of Management Science
Vol 9, No 1, May 2003

Assessing the Precision of a Jackknife Estimator

Daesu Park[*]
Management Research Laboratory, KT
206 Jungja-Dong, Bundang-Gu, Sungnam City, Kyunggi, 463-711, Korea,

(Received Jun. 2002, Revised Jan. 2003, Accepted Apr. 2003)

ABSTRACT

We introduce a new estimator of the uncertainty of a jackknife estimate of standard error: the jackknife-after-jackknife (JAJ). Using Monte Carlo simulation, we assess the accuracy of the JAJ in a variety of settings defined by statistic of interest, data distribution, and sample size. For comparison, we also assess the accuracy of the jackknife-after-bootstrap (JAB) estimate of the uncertainty of a bootstrap standard error. We conclude that the JAJ provides a useful new supplement to Tukey's jackknife, and the combination of jackknife and JAJ provides a useful alternative to the combination of bootstrap and JAB.

1. INTRODUCTION

The bootstrap (Efron [1], Efron and Tibshirani [4]) and the jackknife (Quenouille [11], Tukey [12]) are well known tools for estimating the standard errors of statistics. They create artificial replicates by randomly resampling or sequentially deleting data values, respectively, to simulate the sampling variability of a statistic. Kunsch [6] and Liu and Singh [7] independently proposed the moving blocks bootstrap and the moving blocks jackknife for dependent data. Park and Willemain [9] introduced the threshold bootstrap and threshold jackknife for stationary and weakly dependent time series. Park et al. [10] also established the asymptotic unbiasedness and consistency of the threshold bootstrap and threshold jackknife estimates. However, even though the bootstrap and jackknife estimates apply to a wide range of statistics and data distributions, they inevitably suffer from sampling variability. The bootstrap estimates are subject to an additional resampling

[*] Email: parkd2@kt.co.kr

variability.

In this paper, we focus on assessing the accuracy of bootstrap and jackknife estimates of standard error for statistics computed from independent and identically distributed (i.i.d.) data. For estimates computed with the bootstrap, the jackknife-after-bootstrap (JAB) (Efron [2]) can be used to assess the variance of the estimated standard errors. For estimates computed with the jackknife, we propose a new approach, the jackknife-after-jackknife (JAJ).

2. THE JACKKNIFE-AFTER-BOOTSTRAP

The JAB method uses Tukey's jackknife to estimate the variance of the bootstrap estimate of standard error. One of its virtues is that the JAB estimate can be obtained by re-using the same bootstrap replicates that were used to compute the bootstrap estimate of standard error.

Suppose that $X = \{X_1, X_2, \cdots, X_n\}$ are i.i.d. data from an unknown distribution F. Let $\hat{\theta}$ be the statistic of interest and $X^* = \{X_1^*, X_2^*, \cdots, X_n^*\}$ be the bootstrap samples of size n, which are randomly drawn with replacement from $X = \{X_1, X_2, \cdots, X_n\}$. The bootstrap replicates of the statistic $\hat{\theta}$ are $\hat{\theta}_i^* = \hat{\theta}(X_1^*, \cdots, X_n^*)$ for $i = 1, \cdots, B$, and the bootstrap estimate of the standard error of $\hat{\theta}$ is defined to be

$$\mathrm{se}^*(\hat{\theta}^*) = \left\{ \frac{1}{B-1} \sum_{i=1}^{B} (\hat{\theta}_i^* - \hat{\theta}_{\bullet}^*)^2 \right\}^{1/2}, \quad \hat{\theta}_{\bullet}^* = \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^* . \tag{2.1}$$

Now, the JAB estimate of the variance of $\mathrm{se}^*(\hat{\theta}^*)$ is computed as follows:

(1) Compute $\mathrm{se}^*(\hat{\theta}^*)_{(j)}$, $j = 1, \cdots, n$, the j^{th} jackknife replicate of $\mathrm{se}^*(\hat{\theta}^*)$, that is, the sample standard deviation of bootstrap replicates $\hat{\theta}_i^*$ computed over the subset of bootstrap samples that do not contain X_j.

(2) The JAB variance estimate of $\mathrm{se}^*(\hat{\theta}^*)$ is defined to be

$$Var_{JAB}(\mathrm{se}^*(\hat{\theta}^*)) = (n-1) \sum_{j=1}^{n} \frac{(\mathrm{se}^*(\hat{\theta}^*)_{(j)} - \mathrm{se}^*(\hat{\theta}^*)_{(\bullet)})^2}{n} \tag{2.2}$$

where $\mathrm{se}^*(\hat{\theta}^*)_{(\bullet)} = \sum_{j=1}^{n} \frac{\mathrm{se}^*(\hat{\theta}^*)_{(j)}}{n}$. Efron and Tibshirani [4] and Hill et al. [5]

showed that the JAB usually overestimates the true variance of bootstrap estimates unless the number of bootstrap replicates B is quite large.

3. THE JACKKNIFE-AFTER-JACKKNIFE

Tukey's jackknife was originally developed to estimate the standard error of a statistic. We propose the JAJ method to assess the accuracy of the jackknife estimate of standard error. The JAJ was inspired by Mosteller and Tukey's [8] description of "two simultaneous uses of leave-out-one" and by the iterated bootstrap. (Indeed, the JAJ could be called the iterated jackknife.) Once Tukey's jackknife estimate is computed, we repeat the jackknife on the original data excluding one of the data points. While the jackknife works by deleting a single datum, the JAJ works by deleting pairs of data values.

The algorithm for computing both the jackknife estimate of the standard error of $\hat{\Theta} = \hat{\Theta}(\boldsymbol{X})$ and its JAJ variance estimate is as follows:

(1) Calculate the i^{th} jackknife replicate of the statistic $\hat{\Theta} = \hat{\Theta}(\boldsymbol{X})$

$$\hat{\Theta}_{(i)} = \hat{\Theta}(\boldsymbol{X}_{(i)}) \quad \text{for } i = 1, \cdots, n \tag{3.1}$$

where $\boldsymbol{X}_{(i)}$ is defined to be $\boldsymbol{X} = \{X_1, X_2, \cdots, X_n\}$ with the i^{th} data point removed.

(2) The jackknife estimate of the standard error of $\hat{\theta}$ is

$$se_{jack}(\hat{\Theta}) = \left\{ (n-1) \sum_{i=1}^{n} \frac{(\hat{\Theta}_{(i)} - \hat{\Theta}_{(\bullet)})^2}{n} \right\}^{1/2}, \quad \hat{\Theta}_{(\bullet)} = \frac{1}{n} \sum_{i=1}^{n} \hat{\Theta}_{(i)}. \tag{3.2}$$

(3) Define $se_{jack}(\hat{\Theta})_{(i)}$ to be the jackknife estimate of the standard error of $\hat{\theta}$ with the i^{th} data point removed

$$se_{jack}(\hat{\Theta})_{(i)} = \left\{ (n-2) \sum_{j=1, j\neq i}^{n} \frac{(\hat{\Theta}_{(ij)} - \hat{\Theta}_{(i\bullet)})^2}{n-1} \right\}^{1/2}, \text{ i} = 1, \cdots, \text{n} \tag{3.3}$$

where $\hat{\Theta}_{(ij)} = \hat{\Theta}(\boldsymbol{X}_{(ij)})$, $\boldsymbol{X}_{(ij)}$ denotes $\boldsymbol{X}$ with the i^{th} and j^{th} data points deleted, and $\hat{\Theta}_{(i\bullet)} = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} \hat{\Theta}_{(ij)}$. For instance, if $\hat{\theta}$ is equal to the sample mean $\overline{X}_n$, then

$$\hat{\Theta}_{(ij)} = \frac{n\overline{X}_n - (X_i + X_j)}{n-2}, \quad \hat{\Theta}_{(i\bullet)} = \frac{n\overline{X}_n - X_i}{n-1}, \text{ and}$$

$$se_{jack}(\hat{\Theta})_{(i)} = \left\{ \sum_{\substack{j=1 \\ j\neq 1}}^{n} \frac{(n\overline{X}_n - X_i + X_j - nX_j)^2}{(n-2)(n-1)^3} \right\}^{\frac{1}{2}}.$$

(4) The JAJ variance of $se_{jack}(\hat{\Theta})$ is defined to be

$$Var_{JAJ}\,(se_{jack}(\hat{\Theta})) = (n-1)\sum_{i=1}^{n} \frac{(se_{jack}(\hat{\Theta})_{(i)} - se_{jack}(\hat{\Theta})_{(\bullet)})^2}{n} \tag{3.4}$$

where $se_{jack}(\hat{\Theta})_{(\bullet)} = \sum_{i=1}^{n} \frac{\mathrm{se}_{jack}(\hat{\Theta})_{(i)}}{n}$.

4. EMPIRICAL EVALUATION

In this section, we report the results of Monte Carlo experiments to assess the accuracy of the JAB variance estimate of bootstrap standard error $Var_{JAB}(\mathrm{se}^*(\hat{\Theta}^*))$ in Eq. (2.2) and the JAJ variance estimate of jackknife standard error $Var_{JAJ}(se_{jack}(\hat{\Theta}))$ in Eq. (3.4). We found it convenient to express the results not as variances but as standard errors, since the latter are in the same scale as the statistics they describe.

We considered three factors in our experimental design: the statistic of interest, the underlying distribution, and the sample size. We studied two statistics $\hat{\theta}$, the sample mean $\overline{\mathrm{X}}_n = \sum_{i=1}^{n} X_i \big/ n$ and sample standard deviation $S = \sqrt{\sum_{i=1}^{n}(X_i - \overline{\mathrm{X}}_n)^2 \big/ n-1}$. We drew samples from two i.i.d. distributions: the standard normal (a symmetric case) and unit exponential (an asymmetric case). We generated samples of size n = 4 (the smallest possible sample size for applying the JAJ to the sample standard deviation), n = 20 (a size for which the JAJ and JAB have equal computing costs), n = 100, and n = 200.

Given an i.i.d sample of size n, we first computed the bootstrap estimate $(\mathrm{se}^*(\hat{\theta}^*))$ and its variance estimate $Var_{JAB}(\mathrm{se}^*(\hat{\theta}^*))$; then we computed the jackknife estimate $(se_{jack}(\hat{\theta}))$ and its variance estimate $Var_{JAJ}(se_{jack}(\hat{\theta}))$. In each replication of the experiment, all computations used the same sample data.

We repeated this experiment N = 5,000 times independently. We estimated the true values of the square roots of $Var_{JAB}(\mathrm{se}^*(\hat{\theta}^*))$ and $Var_{JAJ}(se_{jack}(\hat{\theta}))$ by the sample standard deviations of the 5,000 estimates of $(\mathrm{se}^*(\hat{\theta}^*))$ and $(se_{jack}(\hat{\theta}))$, respectively. We assessed accuracy using bias, sample standard deviation, and their combination in the form of the root mean squared error (RMSE).

Besides accuracy, an important issue is the computational effort required for the JAB and JAJ. For a sample of size n, the JAJ requires n evaluations of $\hat{\theta}$ to compute $(se_{jack}(\hat{\theta}))$ and n(n-1) evaluations to compute $Var_{JAJ}(se_{jack}(\hat{\theta}))$, making a total of n^2 evaluations of $\hat{\theta}$. When n = 20, this amounts to 400 evaluations of the statistic. Accordingly, we select B = 400 bootstrap replications to match the computational effort in this case. When n = 4, the JAJ requires only 12 evaluations, but we kept B = 400 to insure good performance from the JAB.

4.1 Performance Comparison of Bootstrap and Jackknife

While our main focus is on assessing uncertainty in estimated standard errors, it is appropriate to begin with a discussion of the standard error estimates themselves. We conducted a preliminary simulation study of the performance of $(\mathrm{se}^*(\hat{\theta}^*))$ and $(se_{jack}(\hat{\theta}))$ for small sample sizes.

Table 1 shows the experimental results. Columns (1)-(3) define the eight scenarios, defined by statistic of interest, data distribution, and sample size. Columns (4)-(5) hold the Monte Carlo estimates of the true values of the mean and standard error, respectively, of the statistic. Columns (6) - (10) assess the quality of the point estimates of standard error, $(\mathrm{se}^*(\hat{\theta}^*))$ and $(se_{jack}(\hat{\theta}))$. The values in column (7) should, ideally, match the values in column (5). The sample standard deviation in column (8) is the primary focus of our paper; it is a performance measure for the point estimates of standard error. The bias estimate in column (9) is computed as column (7) minus column (5). The RMSE estimate in column (10) is the square root of the sum of the squares of columns (8) and (9).

As expected, larger sample size reduced the estimated standard deviations, biases and RMSEs in columns (8)-(10). In every scenario, the jackknife standard error was less biased but also less stable than the bootstrap standard error. The RMSE combined both bias and variability into a single summary statistic; by this measure, the bootstrap and jackknife estimates were about equally accurate, with a slight advantage to the bootstrap. However, for small samples the jackknife requires less computation.

Table 1. Experimental assessment of point estimates of standard error

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Statistic	Data	Sample Size	Sample Avg	Sample SD	SE Method	Sample Avg	Sample SD	Bias	RMSE
Mean	Normal	4	0.000	0.502	Jackknife	0.461	0.194	-0.041	0.199
					Bootstrap	0.399	0.169	-0.103	0.198
		20	-0.001	0.220	Jackknife	0.221	0.036	0.001	0.036
					Bootstrap	0.216	0.036	-0.004	0.037
	Expo	4	0.989	0.496	Jackknife	0.416	0.270	-0.080	0.282
					Bootstrap	0.360	0.235	-0.136	0.271
		20	0.994	0.224	Jackknife	0.212	0.065	-0.012	0.066
					Bootstrap	0.207	0.064	-0.017	0.066
Std Dev	Normal	4	0.921	0.392	Jackknife	0.405	0.236	0.013	0.236
					Bootstrap	0.299	0.136	-0.093	0.165
		20	0.986	0.161	Jackknife	0.160	0.047	-0.001	0.047
					Bootstrap	0.144	0.038	-0.017	0.041
	Expo	4	0.839	0.548	Jackknife	0.423	0.392	-0.125	0.411
					Bootstrap	0.295	0.225	-0.253	0.339
		20	0.955	0.286	Jackknife	0.243	0.161	-0.043	0.167
					Bootstrap	0.204	0.117	-0.083	0.144

Note 1: Results based on N = 5000 independent replications.

Note 2: Bootstrap replications B = 400 to allow for same number of statistic evaluations in JAJ and JAB.

Note 3: Bootstrap and jackknife applied to same datasets.

4.2 Performance Comparison of Jackknife-after-Bootstrap (JAB) and Jackknife-after-Jackknife (JAJ)

Now we turn to estimating the standard error of the estimated standard error. Table 2 shows the experimental results. Again, columns (1)-(3) define the scenarios. Columns (4)-(8) assess the quality of the estimates of the standard errors of the standard errors. The values in column (5) approximate the true standard errors of the standard errors. (For n = 4 and n = 20, they can be compared to the corresponding values in column (8) of Table 1 to gauge the degree of sampling variability in our results). Columns (6)-(8) assess the performance of the JAJ and JAB estimators.

Table 2. Experimental assessment of standard errors of estimated standard errors

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Statistic	Data	Sample Size	Method	True SE(SE)	Sample SD	Bias	RMSE
Mean	Normal	4	JAJ	0.180	0.127	0.032	0.131
			JAB	0.156	0.090	-0.005	0.090
		20	JAJ	0.035	0.019	0.002	0.011
			JAB	0.035	0.014	0.020	0.024
		100	JAJ	0.007	0.001	0.000	0.001
			JAB	0.008	0.006	0.038	0.039
		200	JAJ	0.004	0.000	0.000	0.000
			JAB	0.004	0.004	0.042	0.042
	Expo	4	JAJ	0.274	0.229	-0.028	0.230
			JAB	0.237	0.161	-0.063	0.173
		20	JAJ	0.064	0.038	-0.008	0.039
			JAB	0.063	0.037	0.006	0.038
		100	JAJ	0.014	0.005	-0.001	0.005
			JAB	0.014	0.009	0.034	0.035
		200	JAJ	0.007	0.002	0.000	0.002
			JAB	0.007	0.007	0.039	0.040
Std Dev	Normal	4	JAJ	0.220	0.160	0.024	0.161
			JAB	0.128	0.074	-0.010	0.075
		20	JAJ	0.045	0.033	0.000	0.034
			JAB	0.037	0.023	0.009	0.025
		100	JAJ	0.010	0.005	-0.001	0.005
			JAB	0.009	0.006	0.024	0.025
		200	JAJ	0.005	0.002	0.000	0.002
			JAB	0.005	0.004	0.028	0.028
	Expo	4	JAJ	0.405	0.312	-0.092	0.325
			JAB	0.231	0.145	-0.084	0.168
		20	JAJ	0.160	0.139	-0.036	0.144
			JAB	0.116	0.090	-0.020	0.093
		100	JAJ	0.051	0.040	-0.013	0.042
			JAB	0.043	0.033	0.021	0.039
		200	JAJ	0.031	0.023	-0.008	0.024
			JAB	0.028	0.024	0.036	0.043

Note 1: Results based on N = 5000 independent replications.

Note 2: Bootstrap replications B = 400.

Note 3: Bootstrap and jackknife applied to same datasets.

One immediate conclusion from Table 2 is that it is difficult to assess the standard error of a standard error, no matter whether one uses the jackknife/JAJ combination or the bootstrap/JAB. The RMSEs in column (8) are substantial relative to the true values in column (5); this conclusion holds even for the largest sample sizes in Table 2. A second conclusion is that the JAJ estimate, like the jackknife itself, tends to have more of a problem with variability (column (6)) than with bias (column (7)); the reverse tends to be true for the JAB estimate. A third conclusion is that, as expected, larger sample sizes always improve the performance of the JAJ. Curiously, this is not the case for the JAB: while the variability of JAB estimates (column (6)) decreases with sample size, the bias does not. Efron and Tibshirani [4] displayed data with the same behavior, though they did not remark on it. They did conclude, however, that increasing the number of bootstrap replicates B would generally reduce the bias. Finally, it is clear that the quality of estimates varies with the scenario. Generally, RMSEs are smaller when the statistic of interest is the mean rather than the standard deviation, and when the data have a normal rather than exponential distribution.

5. SUMMARY AND CONCLUSIONS

We introduced the jackknife-after-jackknife (JAJ) method for assessing the accuracy of standard errors estimated by Tukey's jackknife for i.i.d. data. The combination of jackknife and JAJ provides a new alternative to the combination of bootstrap and jackknife-after-bootstrap (JAB).

Monte-Carlo simulations show that it is difficult to accurately estimate the uncertainty of an estimated standard error, no matter which method is used. For both the JAJ and JAB, RMSEs are of roughly the same magnitude as the values being estimated. The JAJ RMSEs derive more from variability than bias, while the reverse tends to be true for the JAB. For the JAJ, larger sample sizes lead to smaller RMSEs, though the RMSE relative to the true standard error of the standard error remains substantial for samples up to $n = 200$ observations. For the JAB, larger sample sizes reduce variability but not bias, so values of B greater than the 400 used here would be required for good performance. The performance of both the JAJ and JAB is better for normal data than exponential, and better when the statistic of interest is the mean than the standard deviation.

Our simulation results are consistent with those of Hill et al. [5]. They found that the JAB usually overestimated the variability of the bootstrap standard error by a substantial amount, though the error declined with increasing numbers of bootstrap replicates B. Efron and Tibshirani [4] also noted that the JAB method is only reliable when B is sufficiently large. However, how large B should be remains uncertain. For instance, in our experiments, we used B = 400; in limited experiments not reported here using B = 1000, the JAB still performed poorly compared to the JAJ for larger sample sizes.

When both approaches have similar accuracy, the choice between jackknife and bootstrap might be made on the basis of computational cost. For n data values, the jackknife/JAJ combination requires n^2 evaluations of the statistic, while the bootstrap/JAB combination requires B evaluations. For simple statistics such as the mean and standard deviation examined here, the cost of thousands of extra computations of the statistic may be insignificant. However, in some settings, each evaluation of the statistic can be very costly; as when the statistic is the result of a Monte Carlo simulation of a complex system or solid modelling of a complicated manufactured assembly. For samples of size roughly 20 to 30, the JAJ is less costly to compute than the JAB. For larger sample sizes, the JAB should be used with a larger number of bootstrap replications than studied here, since substantial biases remain when B = 400 and $n \geq 100$.

The choice between bootstrap and jackknife estimates of standard error is now more balanced, in that the JAJ can provide good estimates of the uncertainty in jackknife standard errors, just as the JAB does for bootstrap standard errors. This expanded choice should be an advantage to the applied statistician faced with the need for computational inference.

REFERENCES

[1] Efron, B., "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics* 7 (1979), 1-26.

[2] Efron, B., "Jackknife-After-Bootstrap Standard Errors and Influence Functions (with discussion)," *Journal of the Royal Statistical Society*, Series B 54 (1992), 83-111.

[3] Efron, B., and Tibshirani, R., "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science* 1 (1986), 54-77.

[4] Efron, B., and Tibshirani, R., *An Introduction to the Bootstrap*, Chapman & Hall, Inc., New York 1993.

[5] Hill, C., Cartwright, P., and Arbaugh, J., "Jackknifing the Bootstrap: Some Monte Carlo Evidence," *Communications in Statistics: Simulation and Computation* 26 (1997), 125-139.

[6] Kunsch, H., "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics* 17 (1989), 1217-1241.

[7] Liu, R., and Singh, K., *Moving Blocks Jackknife and Bootstrap Capture Weak Dependence.* In: LePage, R., Billard, L. (Eds.), Exploring the Limit of Bootstrap. Wiley, NY (1992), 225-248.

[8] Mosteller, F., and Tukey, J., *Data Analysis and Regression: A Second Course in Statistics,* Addison - Wesley, Reading, MA 1977.

[9] Park, D., and Willemain, T., "The Threshold Bootstrap and Threshold Jackknife," *Computational Statistics and Data Analysis* 31 (1999), 187-202.

[10] Park, D., Kim, Y., Shin, K., and Willemain, T., "Simulation Output Analysis Using the Threshold Bootstrap," *European Journal of Operational Research* 134 (2001), 17-28.

[11] Quenouille, M. (1949), "Approximating Tests of Correlation in Time Series," *Journal of the Royal Statistical Society,* Series B. 11 (1949), 68-84.

[12] Tukey, J. (1958), "Bias and Confidence Interval in Not Quite Large Samples (Abstract)," *The Annals of Mathematical Statistics* 29 (1958), 614.