Journal of Statistical Planning and Inference 143 (2013) 1039–1048

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Data-driven nonparametric prediction intervals

Jesse Frey

Department of Mathematics and Statistics, Villanova University, Villanova, PA 19085, USA

ARTICLE INFO

Article history:
Received 7 June 2012
Received in revised form
28 November 2012
Accepted 16 January 2013
Available online 29 January 2013

Keywords:
Order statistics
Probability integral transform
Standard uniform distribution

ABSTRACT

Standard nonparametric prediction intervals for a single future observation are obtained by taking the interval between two pre-specified order statistics from the initial sample. In this paper, we consider the alternate approach of taking the shortest interval that contains a pre-specified number of the subintervals between the order statistics of the initial sample. We develop a method for determining exact confidence coefficients for such intervals, and we show that these data-driven prediction intervals outperform standard equal-tailed nonparametric prediction intervals. Specifically, they are much shorter than the standard intervals when the underlying distribution is skewed, and they are only slightly longer when the underlying distribution is symmetric. We also obtain the asymptotic approximation that to achieve exact confidence coefficient $1-\alpha$ when using the new data-driven prediction intervals with initial sample size n, approximately $n(1-\alpha)+1.12\sqrt{n\alpha}$ of the subintervals between the order statistics of the initial sample must be included.

1. Introduction

Let $X_1,\ldots,X_n$ be a simple random sample from a distribution with continuous cumulative distribution function F. If $X_{(1)} < \cdots < X_{(n)}$ are the order statistics from this sample and r and s are pre-specified integers such that $1 \le r < s \le n$, then the interval $(X_{(r)},X_{(s)})$ is a $100((s-r)/(n+1))\%$ prediction interval for an independent future observation X_{n+1} from the same distribution. Prediction intervals of this form were discussed by Wilks (1941), and Fligner and Wolfe (1979) used intervals of the same form as nonparametric prediction intervals for the median of a future sample. The values r and s need not satisfy any constraints other than the constraint that $1 \le r < s \le n$, but they must be chosen before seeing the data. Thus, for a given desired confidence level, there will typically be multiple (r,s) pairs that offer equal confidence.

One way to select a single pair (r,s) is to insist that $s=n+1-r$ so that the prediction interval is equal-tailed. However, if the underlying distribution F is not symmetric, then an equal-tailed prediction interval may be unnecessarily wide. For example, if F is right-skewed, then we can make the interval shorter by choosing r and s so that $r+s<n+1$, and if F is left-skewed, then we can make the interval shorter by choosing r and s so that $r+s>n+1$. If we do not know ahead of time how the distribution is shaped, however, we would likely use an equal-tailed interval. Thus, we take equal-tailed prediction intervals of the form $(X_{(r)},X_{(n+1-r)})$ as the reference intervals when evaluating the performance of the alternate intervals that we propose in this paper.

To obtain nonparametric prediction intervals that adapt to the shape of the distribution F, we use the data-driven approach of taking the shortest interval that contains a prespecified number of the subintervals between the order statistics of the original sample. Specifically, for an appropriately chosen integer k such that $1 \le k \le n-1$, we take as our

E-mail address: jesse.frey@villanova.edu

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.01.004

prediction interval the shortest interval of the form $(X_{(r)},X_{(r+k)})$, where $1 \le r \le n-k$. The confidence coefficient for the resulting prediction interval is then an increasing function of k, and we would choose k to meet our coverage probability requirements.

Such a prediction interval clearly adapts to the symmetry or skewness of the distribution F. However, the confidence coefficient now differs from the value $k/(n+1)$ that we would have obtained had we picked the bounds r and $r+k$ ahead of time. It turns out, in fact, that the coverage probability differs from one distribution F to another. Thus, in the remainder of this paper, we carefully distinguish between the nominal confidence coefficient, the exact confidence coefficient, and the coverage probability of the prediction interval for a specific underlying distribution F.

In general, a $100(1-\alpha)\%$ prediction interval for an independent future observation X_{n+1} from the same distribution is a random interval $(L(X_1,\ldots,X_n),U(X_1,\ldots,X_n))$ such that

$$\inf_{F\in\mathcal{F}} P_F(L(X_1,\ldots,X_n) < X_{n+1} < U(X_1,\ldots,X_n)) \ge 1-\alpha, \tag{1}$$

where $\mathcal{F}$ is the set of distributions under consideration. Note that the probability in Eq. (1) is computed over the joint distribution of $X_1,\ldots,X_n,X_{n+1}$. We call the probability $P_F(L(X_1,\ldots,X_n) < X_{n+1} < U(X_1,\ldots,X_n))$ the coverage probability of the prediction interval at F, and we call the infimum in Eq. (1) the exact confidence coefficient for the prediction interval method. The coverage probability may vary from one F to another, and the exact confidence coefficient may be strictly larger than the nominal confidence coefficient $1-\alpha$.

When the set $\mathcal{F}$ from Eq. (1) is a parametric family, then $(L(X_1,\ldots,X_n),U(X_1,\ldots,X_n))$ is a parametric prediction interval, and when $\mathcal{F}$ is the set of all continuous distributions, then the interval is a nonparametric prediction interval. Prediction intervals of both types were discussed by Hahn and Meeker (1991) and Vardeman (1992), but our focus here is on the nonparametric case. Thus, in what follows, we initially take $\mathcal{F}$ to be the set of all continuous distributions. We then show, however, that if the prediction intervals are taken as closed intervals, then the prediction intervals may also be used when the distribution function F is not continuous. Thus, the prediction intervals may be applied when there are tied data values.

Random intervals obtained by taking the shortest interval of the form $(X_{(r)},X_{(r+k)})$ for $1 \le r \le n-k$ were proposed as nonparametric tolerance intervals by Di Bucchianico et al. (2001), and Frey (2010) developed a method for finding exact confidence coefficients for such intervals when they are used as nonparametric tolerance intervals. Some of the results that we obtain here are related to this earlier work, but most of the results are entirely different because of the differing probability requirements for prediction and tolerance intervals.

The prediction intervals that we develop here would be applied when one is interested in making inference on a single future observation without assuming a particular parametric model. As one example of an application, both Hahn and Meeker (1991, p. 31) and Vardeman (1992) discuss the case of a customer purchasing a car. The customer would be more interested in a prediction interval for the miles per gallon of the particular car that he or she purchases than in a confidence interval for the average miles per gallon for all cars of the same type. Similarly, a seriously ill patient deciding whether to adopt a new medical treatment would be more interested in a prediction interval for his or her own survival time than in a confidence interval for the average survival time for all patients who adopt the treatment.

In Section 2, we develop the theory needed for determining exact confidence coefficients for data-driven nonparametric prediction intervals of the type described. In Section 3, we compute 90% nonparametric prediction intervals of both types for a data set consisting of heights of 396 trees, and we find that the data-driven prediction interval is nearly 15% shorter than the standard equal-tailed prediction interval. In Section 4, we use simulation to make a more comprehensive comparison of the two types of prediction intervals in terms of average length, and we also examine the extent to which the data-driven intervals are conservative. In Section 5, we obtain the asymptotic approximation that to achieve exact confidence coefficient $1-\alpha$ when using the new data-driven prediction intervals with initial sample size n, approximately $n(1-\alpha)+1.12\sqrt{n\alpha}$ of the subintervals between the order statistics of the initial sample must be included. We conclude with a discussion in Section 6.

2. Exact confidence coefficients

When we take as our prediction interval the shortest interval of the form $(X_{(r)},X_{(r+k)})$ for $1 \le r \le n-k$, we clearly do not get the same exact confidence coefficient that we would get if we selected an interval containing k subintervals ahead of time. Theorem 1 shows that we can obtain the exact confidence coefficient by determining the coverage probability of the prediction interval when the underlying distribution is the uniform distribution on the interval [0,1].

Theorem 1. *Let the sample size n be fixed, and let $a_1,\ldots,a_l$ and $b_1,\ldots,b_l$ be fixed integers that satisfy $1 \le a_i < b_i \le n$ for $i=1,\ldots,l$. If we obtain our prediction interval by choosing the shortest of the intervals $(X_{(a_i)},X_{(b_i)})$ for $i=1,\ldots,l$, then the exact confidence coefficient for the prediction interval is achieved when the parent distribution is a uniform distribution.*

Proof. The shortest interval has at least as much probability content as the interval with the least probability content. Thus, for any distribution F, the coverage probability achieved by always taking the interval with the least probability content provides a lower bound on the coverage probability obtained by always taking the shortest interval. By the probability integral transform, the joint distribution of the probability contents for the l intervals does not depend on the particular continuous F that we choose. Thus, the coverage probability obtained by always taking the interval with the least probability content provides a lower bound on the exact confidence coefficient. If the parent distribution is a uniform distribution, then the shortest interval and the interval with the least probability content coincide. Thus, the exact confidence coefficient for the prediction interval is the coverage probability achieved when the parent distribution is uniform. □

To apply Theorem 1 to our prediction intervals, we let $a_1,\ldots,a_l$ be the values $1,\ldots,n-k$ and $b_1,\ldots,b_l$ the values $1+k,\ldots,n$. Theorem 1 then says that the exact confidence coefficient, over the class $\mathcal{F}$ of all continuous F, is the coverage probability achieved when F is uniform. The coverage probability for non-uniform F is then at least as high as the coverage probability for uniform F, which means that when computing the exact confidence coefficient, it suffices to compute the coverage probability when the underlying distribution is the uniform distribution on [0,1]. In other words, the uniform distribution is the worst case scenario from the point of view of coverage.

Theorem 2 shows that the process of finding exact confidence coefficients can be simplified further. The proof of this theorem uses well-known properties of standard uniform order statistics and standard uniform spacings, many of which were discussed by Pyke (1965, Section 2.1) and by David and Nagaraja (2003, Section 6.4).

Theorem 2. *Let the sample size n be fixed, and let k be an integer that satisfies* $(n-1)/2 \le k \le n$. *The nonparametric prediction interval obtained by choosing the shortest of the intervals* $(X_{(r)},X_{(r+k)})$ *for* $r=1,\ldots,n-k$ *has exact confidence coefficient*

$$\frac{1}{n+1}(2k-n+1+2(n-k-1)c_{n-k-1}), \tag{2}$$

where c_l *is the expected length of the shortest of the intervals* $(0,U_{(l)}),(U_{(1)},U_{(l+1)}),\ldots,(U_{(l)},1)$ *when* $U_{(1)}<\cdots<U_{(2l-1)}$ *are the order statistics from a standard uniform sample of size* $2l-1$.

Proof. By Theorem 1, it suffices to find the coverage probability of the prediction interval when the underlying distribution F is standard uniform. Let $U_1,\ldots,U_n$ be a simple random sample from the standard uniform distribution, and let $U_{(1)}<\cdots<U_{(n)}$ be the corresponding order statistics.

Since $k\ge(n-1)/2$, we know that $k+1\ge n-k$. This means that when we choose the shortest of the intervals $(U_{(1)},U_{(k+1)}),\ldots,(U_{(n-k)},U_{(n)})$, the interval $(U_{(n-k)},U_{(k+1)})$ is automatically included, and it contributes coverage probability $(k+1-(n-k))/(n+1)=(2k-n+1)/(n+1)$. The shortest of the intervals $(U_{(1)},U_{(k+1)}),\ldots,(U_{(n-k)},U_{(n)})$ is then the union of $(U_{(n-k)},U_{(k+1)})$ and the shortest of the sets $(U_{(1)},U_{(n-k)}),(U_{(2)},U_{(n-k)})\bigcup(U_{(k+1)},U_{(k+2)}),(U_{(3)},U_{(n-k)})\bigcup(U_{(k+1)},U_{(k+3)}),\ldots,(U_{(k+1)},U_{(n)})$, which each include $n-k-1$ subintervals. By properties of uniform order statistics, the gaps $U_{(1)}-0,U_{(2)}-U_{(1)},\ldots,1-U_{(n)}$ are exchangeable. Thus, the expected length for the shortest of the sets $(U_{(1)},U_{(n-k)}),(U_{(2)},U_{(n-k)})\bigcup(U_{(k+1)},U_{(k+2)}),(U_{(3)},U_{(n-k)})\bigcup(U_{(k+1)},U_{(k+3)}),\ldots,(U_{(k+1)},U_{(n)})$ is the same as the expected length for the shortest of the sets $(0,U_{(n-k-1)}),\ldots,(U_{(n-k-1)},U_{(2n-2k-2)})$. When $U_{(2n-2k-2)}$ is given, the random variables $U_{(1)}/U_{(2n-2k-2)},\ldots,U_{(2n-2k-3)}/U_{(2n-2k-2)}$ are distributed like standard uniform order statistics from a sample of size $2n-2k-3=2(n-k-1)-1$. Thus, the expected length for the shortest of the sets $(0,U_{(n-k-1)}),\ldots,(U_{(n-k-1)},U_{(2n-2k-2)})$ is given by $c_{n-k-1}E[U_{(2n-2k-2)}]=c_{n-k-1}\cdot 2(n-k-1)/(n+1)$. Adding this expected length to the expected length of the automatically included interval $(U_{(n-k)},U_{(k+1)})$ gives formula (2). □

Example 1. Suppose that $n=9$ and $k=7$ so that we obtain our nonparametric prediction interval by taking the shorter of the intervals $(X_{(1)},X_{(8)})$ and $(X_{(2)},X_{(9)})$. By Theorem 2, the exact confidence coefficient for this prediction interval is $\frac{1}{10}(6+2c_1)$, where c_1 is the expected length for the shorter of the intervals $(0,U_1)$ and $(U_1,1)$ when U_1 is a single draw from the standard uniform distribution. If $U_1<1/2$, then the length of the shorter interval is U_1, and if $U_1>1/2$, then the length of the shorter interval is $1-U_1$. Thus,

$$c_1=\int_{u=0}^{1/2}u\,du+\int_{u=1/2}^{1}(1-u)\,du=\frac{1}{8}+\frac{1}{8}=0.25,$$

which means that the exact confidence coefficient is $\frac{1}{10}(6+2(0.25))=0.65$. Had we instead decided on either the interval $(X_{(1)},X_{(8)})$ or the interval $(X_{(2)},X_{(9)})$ before seeing the data, the exact confidence coefficient would have been 0.7.

Using the same sort of logic used in Example 1, we also computed that $c_2=9/32$. To obtain approximate c_l values for larger l, we used simulation. For each l, we generated 200,000 standard uniform samples of size $2l-1$. For each such sample, we found the length of the shortest of the intervals $(0,U_{(l)}),(U_{(1)},U_{(l+1)}),\ldots,(U_{(l)},1)$. We then averaged the 200,000 lengths to obtain an estimate for c_l. To assess the accuracy of our estimates, we also computed the simulation standard error associated with each estimated c_l value by computing $s_l/\sqrt{200{,}000}$, where s_l is the standard deviation of the 200,000 simulated lengths that we obtained for a particular l. In each case, this simulation standard error was less than 0.0001. Simulated values of c_l for $l=3,\ldots,99$ are given in Table 1, where the rows correspond to the tens digit of l and the columns correspond to the ones digit. We see from the table that, as expected, c_l is strictly increasing in l. In the limit as l goes to infinity, c_l converges to 0.5 from below.

Our results thus far apply when the set $\mathcal{F}$ of distributions under consideration is the set of all continuous distributions. However, our results may also be extended to the case where $\mathcal{F}$ is the set of *all* distributions. Specifically, if we think of the prediction intervals as closed intervals rather than as open intervals, then it follows from an argument of Scheffé and Tukey (1945, Section 5) that the coverage probability for discontinuous F is never smaller than the minimum coverage

Table 1
Simulated and calculated (indicated with an asterisk) values of c_l for $l=1,\ldots,99$. Each simulated value is based on 200,000 runs, and all simulation standard errors are 0.0001 or less.

Tens digit	Ones digit									
	0	1	2	3	4	5	6	7	8	9
0		0.2500*	0.2812*	0.3021	0.3173	0.3292	0.3388	0.3470	0.3537	0.3597
1	0.3650	0.3696	0.3737	0.3774	0.3809	0.3840	0.3868	0.3895	0.3920	0.3943
2	0.3964	0.3984	0.4004	0.4021	0.4038	0.4054	0.4070	0.4083	0.4098	0.4111
3	0.4124	0.4135	0.4147	0.4158	0.4169	0.4179	0.4188	0.4198	0.4206	0.4216
4	0.4225	0.4232	0.4240	0.4248	0.4255	0.4263	0.4270	0.4276	0.4283	0.4290
5	0.4296	0.4302	0.4307	0.4313	0.4320	0.4324	0.4330	0.4335	0.4340	0.4345
6	0.4350	0.4354	0.4359	0.4364	0.4368	0.4373	0.4377	0.4381	0.4385	0.4389
7	0.4393	0.4397	0.4400	0.4404	0.4408	0.4412	0.4415	0.4418	0.4422	0.4425
8	0.4428	0.4431	0.4434	0.4438	0.4441	0.4444	0.4447	0.4449	0.4452	0.4455
9	0.4458	0.4461	0.4463	0.4466	0.4468	0.4471	0.4474	0.4476	0.4478	0.4481

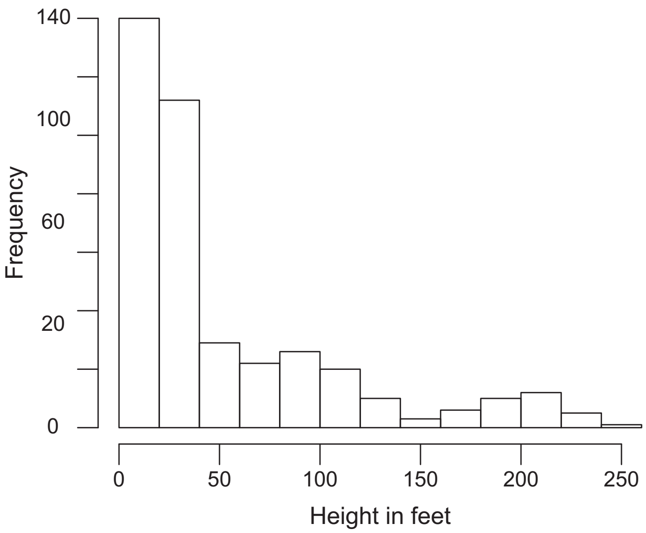

Fig. 1. Histogram of 396 heights (in feet) for long-leaf pines.

probability for continuous F. Thus, our prediction intervals may be applied to discrete distributions and to data sets where there are ties. The data example in Section 3 provides an illustration.

3. Data example

Platt et al. (1988) collected data on the heights and the diameters at chest height for long-leaf pines in the Wade Tract in Thomas County, Georgia. Heights (in feet) and diameters at chest height (in centimeters) for 396 of these long-leaf pines appear in Appendix B of Chen et al. (2004). We think of the 396 tree heights as our initial sample, and we seek a 90% prediction interval for the height of the next independent long-leaf pine to be sampled. Since there are ties in the data set, we will use closed intervals rather than open intervals to ensure that the desired coverage probability is achieved.

Since there are 397 total subintervals, a nonparametric prediction interval of the form $[X_{(r)},X_{(s)}]$ must satisfy $s-r \geq 397(0.9)=357.3$ in order to have exact confidence coefficient at least 90%. Thus, at least 358 of the subintervals between order statistics must be included. If we use the standard equal-tailed nonparametric prediction interval, then our prediction interval is $[X_{(19)},X_{(378)}]=[4,198]$. However, as we see from the histogram of the tree heights given in Fig. 1, the data are highly skewed to the right. Had we anticipated this skewness, we might have decided ahead of time to use the interval $[X_{(1)},X_{(359)}]$ or some other interval that relies less heavily on the values in the long right tail. Fig. 2 plots the interval length against r for all prediction intervals of the form $[X_{(r)},X_{(r+358)}]$ for $r=1,\ldots,38$. The figure also includes a dashed horizontal line that represents the length (194 feet) of the equal-tailed interval $[X_{(19)},X_{(378)}]$.

Had we chosen r and s before seeing the data, we could have used any one of the intervals shown in Fig. 2, and the interval $[X_{(1)},X_{(359)}]$ would have given the best performance. However, if we first look at the data and then chose the shortest interval that contains a certain number of subintervals, we need to include more subintervals to achieve the same exact confidence coefficient. Table 2, obtained using the method described in Section 2, shows the exact confidence

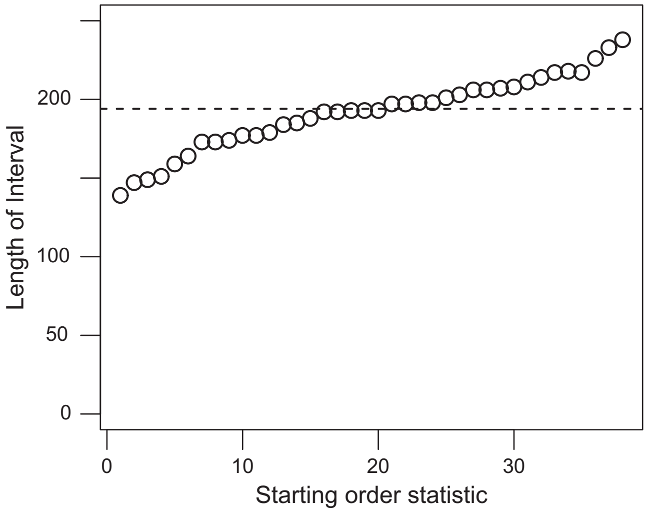

Fig. 2. Lengths of 90% prediction intervals as a function of the rank of the order statistic used as the lower bound. The dashed line gives the length of the 90% equal-tailed prediction interval.

Table 2
Simulated exact confidence coefficients as a function of k when $n=396$.

Number of subintervals k	359	360	361	362	363
Exact confidence coefficient	0.890	0.892	0.895	0.898	0.901

coefficient as a function of the number of subintervals that we include. We see that if we ignored the fact that we were choosing the shortest interval and simply used the interval $[X_{(1)},X_{(359)}]$, then the exact confidence coefficient would be 89.0% rather than 90%. If we include 363 subintervals, then the exact confidence coefficient exceeds 90%. Taking the shortest interval that includes 363 subintervals gives the prediction interval $[X_{(1)},X_{(364)}]=[1,167]$, which is nearly 15% shorter than the standard equal-tailed interval. Thus, for these tree data, the data-driven prediction interval is substantially shorter than the standard equal-tailed prediction interval.

4. Performance comparisons

To compare the performance of the new data-driven prediction intervals to that of the standard equal-tailed intervals, we did a simulation study. We considered both symmetric distributions F and skewed distributions F. The symmetric distributions that we considered were the standard uniform and standard normal distributions, and the skewed distributions that we considered were the standard exponential and Gamma(2,1) distributions. We considered sample sizes of 100 and 800, and we considered exact confidence coefficients between 50% and 100%. We then compared prediction intervals of the two types in terms of average length. Since the data-driven prediction intervals are slightly conservative when the data are not uniform, we also estimated the true coverage probability for each interval so that we could assess the extent to which the data-driven prediction intervals are conservative. Figs. 3 and 4 show the results of the length comparisons when the exact confidence coefficient is approximately 90%, and Figs. 5 and 6 show the extent to which each interval is conservative when the exact confidence coefficient is approximately 90%.

In Figs. 3–6, each dot represents the performance of one available prediction interval. For the standard equal-tailed intervals, the possible prediction intervals are the intervals $(X_{(r)},X_{(n+1-r)})$ for $r<(n+1)/2$. For the data-driven intervals, each choice of k provides one possible prediction interval.

Fig. 3 shows the results of the length comparison when $n=100$, and Fig. 4 shows the results when $n=800$. In each figure, average lengths for standard equal-tailed prediction intervals are shown with solid dots, and average lengths for data-driven prediction intervals are shown with open dots. We see from Fig. 3 that when the distribution F is standard uniform, the two types of prediction intervals are comparable in terms of length as expected. Indeed, the difference between the intervals in this case is only that more (and different) exact confidence coefficients are available for the data-driven intervals. When the distribution is standard normal, the equal-tailed prediction intervals are slightly shorter than the standard intervals, but the advantage is relatively small. For the two skewed distributions, the data-driven intervals are substantially shorter than the standard intervals. The 90% data-driven intervals are roughly 15% shorter than the equal-tailed intervals for the standard exponential distribution, and they are also roughly 10% shorter than the equal-tailed intervals for the less skewed Gamma(2,1) distribution.

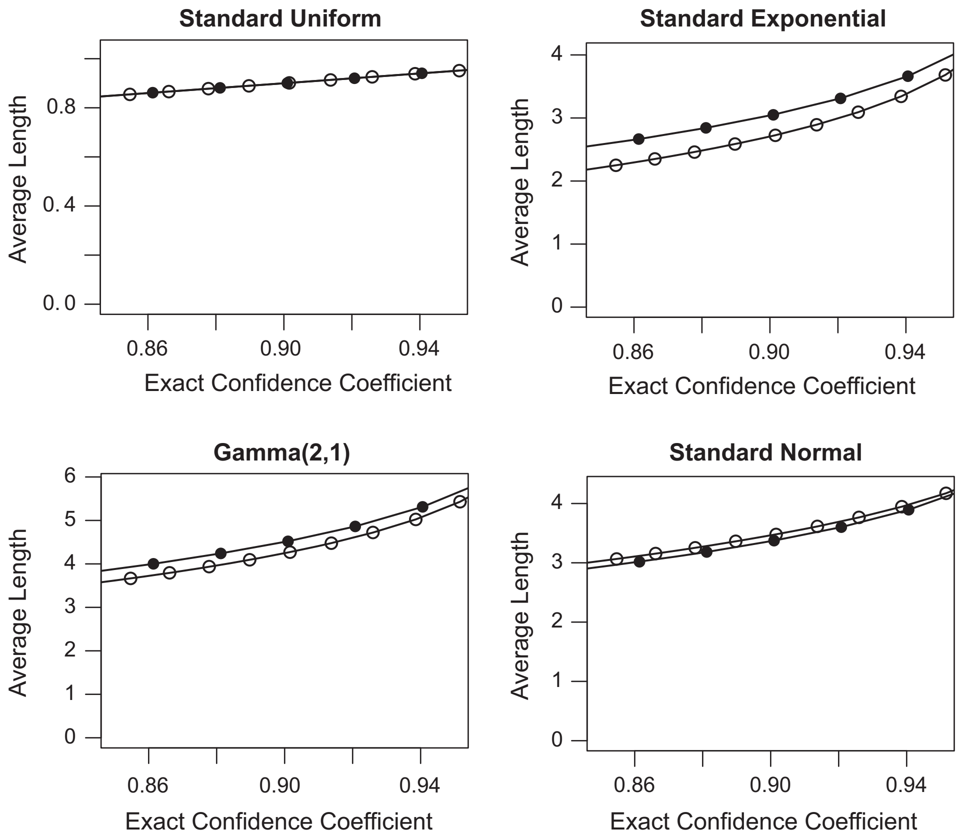

Fig. 3. Simulated average length versus simulated exact confidence coefficient for the standard nonparametric prediction intervals and the new data-driven prediction intervals when the exact confidence coefficient is near 90% and the sample size is 100. Solid dots indicate standard intervals and open dots indicate data-driven intervals.

Fig. 4, which shows results for the case where $n=800$, gives much the same pattern that we saw in Fig. 3. The two types of prediction intervals are comparable in terms of length when F is uniform, and there is a slight advantage for the equal-tailed intervals when F is standard normal. However, there is a substantial advantage for the data-driven prediction intervals when the underlying distribution is skewed. Comparing Fig. 4 to Fig. 3 shows that the advantage for the data-driven intervals does not go away as the sample size n increases. Instead, it seems clear that if we let n increase while keeping the desired confidence coefficient fixed, the data-driven intervals will be either shorter than (for skewed distributions) or comparable in length to (for symmetric distributions) the standard equal-tailed intervals.

Fig. 5 shows the excess coverage probability (above the exact confidence coefficient) for the data-driven prediction intervals when $n=100$, and Fig. 6 shows the excess coverage probability when $n=800$. We see from both figures that when F is uniform, there is no excess coverage probability at all, as we know from Theorem 1. For the three non-uniform distributions, there is excess coverage. Among the distributions considered, the excess coverage is largest for the standard exponential distribution and smallest for the standard normal distribution. When F is exponential and $n=100$, the excess coverage for prediction intervals with exact confidence coefficient near 90% is as high as 2.5%, and the corresponding numbers for Gamma(2,1) and normal data are 1.9% and 1.0%. Comparing Fig. 6 to Fig. 5 shows that when the sample size increases, the excess coverage decreases in all three non-uniform cases.

Considering both Figs. 5 and 6, we see that the amount of excess coverage is both small and decreasing with n. Moreover, the excess coverage is largest for precisely those highly skewed distributions where the data-driven prediction intervals have the biggest advantage over the standard intervals in terms of average length.

5. Asymptotic approximations

The method from Section 2 allows us to find the exact confidence coefficient for the data-driven nonparametric prediction interval corresponding to a particular choice of k and n. However, it is desirable to have compact approximate formulas both for c_l and for the k needed to get an exact confidence coefficient of $1-\alpha$ or more when n is given. In this section, we use empirical process theory to develop such formulas.

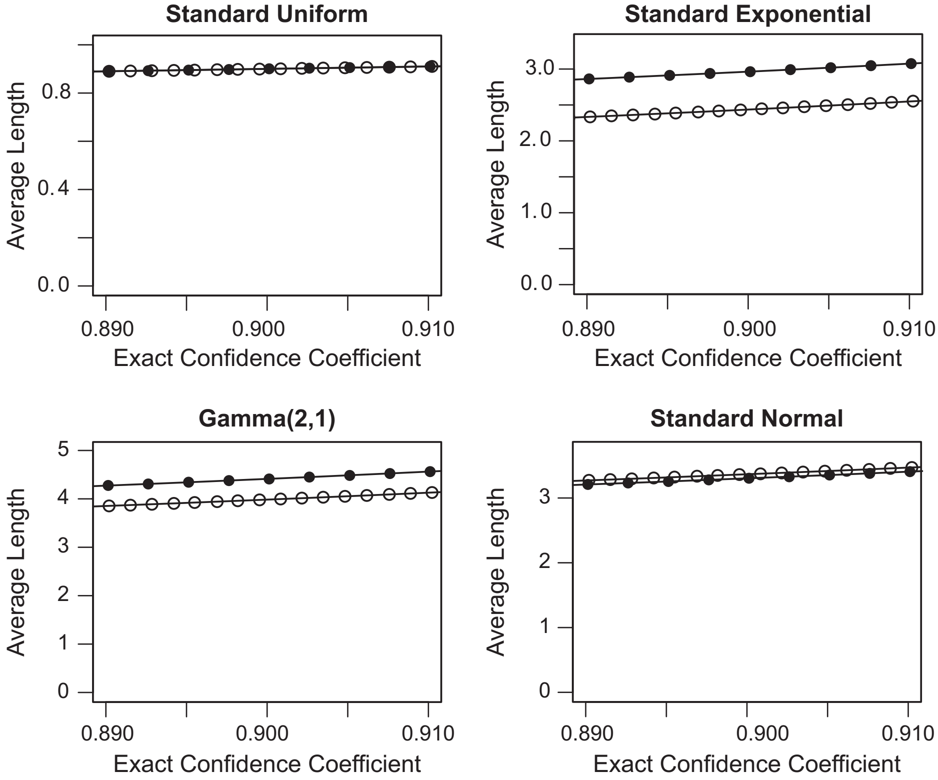

Fig. 4. Simulated average length versus simulated exact confidence coefficient for the standard nonparametric prediction intervals and the new data-driven prediction intervals when the exact confidence coefficient is near 90% and the sample size is 800. Solid dots indicate standard intervals and open dots indicate data-driven intervals.

We first obtain an asymptotic approximation for c_l when l is big. Set $m=2l-1$, suppose that $U_1,\ldots,U_m$ is a simple random sample from the standard uniform distribution, and let $\hat{U}(t)$ be the empirical distribution function for the sample. It then follows from a standard result that as $m\to\infty$, $\sqrt{m}(\hat{U}(t)-t)$ converges on (0,1) to a Brownian bridge process $R(t)$. Define

$$c\equiv E\left[\max_{0\le t\le 1/2} R(t+1/2)-R(t)\right].$$

It then follows that if we look at all intervals of the form $(t,t+1/2)$ for $t\in(0,1/2)$ and m large, the expected empirical probability content for the interval with highest empirical probability content is approximately $1/2+c/\sqrt{m}$. Here the empirical probability content for an interval $(a,b)\subset(0,1)$ is $\hat{U}(b)-\hat{U}(a)$, and the true probability content is $U(b)-U(a)=b-a$.

Now define a function $e(\cdot)$ on the interval (0,1) by setting

$$e(d)\equiv E\left[\max_{0\le t\le 1-d} R(t+d)-R(t)\right].$$

It is clear that $e(d)$ is continuous in d, and we also have from the definition of c that $e(1/2)=c$. Thus, when we look at all intervals of the form $(t,t+d)$ for a fixed d near 1/2, the expected amount by which the empirical probability content of the interval with highest empirical probability content exceeds the true probability content d is approximately $c/\sqrt{m}$. This means that to get expected empirical probability content 1/2, we need to use a d value of approximately $1/2-c/\sqrt{m}$. Thus, when we choose the shortest interval from the list $(0,U_{(l)}),(U_{(1)},U_{(l+1)}),\ldots,(U_{(l)},1)$, the expected length c_l is approximately $1/2-c/\sqrt{m}\approx 1/2-c/\sqrt{2l}$. We now estimate c.

To estimate c, we simulated realizations from a Brownian bridge process on (0,1). Since c is the expected value of a maximum over the interval (0,1/2), our estimate will be sensitive to the number of points that we use when simulating the Brownian bridge. Thus, we ran the simulation multiple times, using an increasingly fine grid of points each time. For each of $w=10$, $w=100$, $w=1000$, $w=10{,}000$, and $w=100{,}000$, we simulated 100,000 realizations of a Brownian bridge on the points $0,1/(2w),\ldots,1$. We then estimated c as

$$\frac{1}{100{,}000}\sum_{i=1}^{100{,}000}\max_{t\in\{0,\frac{1}{2w},\ldots,1/2\}} R_i(t+1/2)-R_i(t),$$

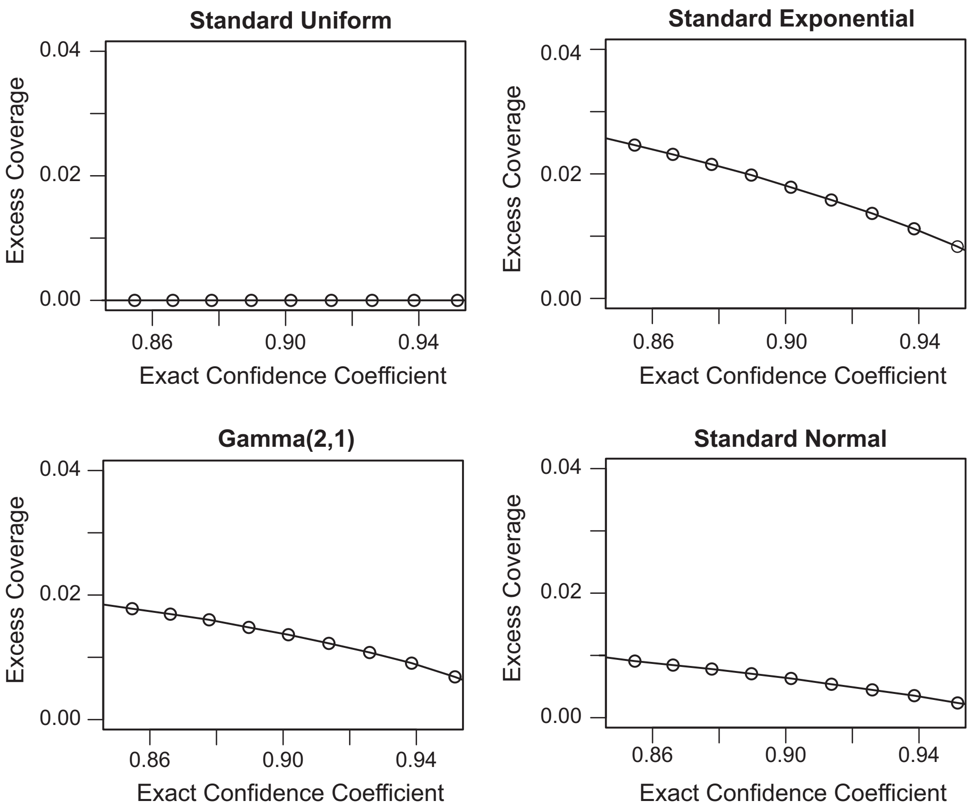

Fig. 5. Simulated excess coverage versus simulated exact confidence coefficient for the new data-driven intervals when the exact confidence coefficient is near 90% and the sample size is 100.

where $R_i(\cdot)$ is the ith simulated realization of a Brownian bridge. Results from these simulations appear in Table 3. We see from the table that our estimates of c change considerably as we increase w from 10 to 100 to 1000. Once w is 10,000 or more, however, the estimates are relatively stable, and we take the estimate $c \approx 0.795$ for $w = 100{,}000$ as a satisfactory approximation. Combining this with our earlier work gives the following approximation for c_l.

Approximation 1. For l large, $c_l \approx 1/2 - c/\sqrt{2l} \approx 1/2 - 0.56/\sqrt{l}$.

Suppose now that we wish to find an integer k so that the data-driven prediction interval based on k and sample size n has coverage probability $1-\alpha$, where $\alpha < 1/2$. If we include $\lceil (1-\alpha)(n+1) \rceil$ of the subintervals between the order statistics in the initial sample, then Theorem 2 and Approximation 1 imply that the exact confidence coefficient is

$$\frac{1}{n+1}\left(2\lceil(1-\alpha)(n+1)\rceil - n + 1 + 2(n - \lceil(1-\alpha)(n+1)\rceil + 1)c_{n-\lceil(1-\alpha)(n+1)\rceil+1}\right)$$

$$\approx 2(1-\alpha) - 1 + 2\alpha(1/2 - 0.56/\sqrt{n\alpha}) = 1 - \alpha - 1.12\sqrt{\alpha/n},$$

which falls short of $1-\alpha$ by $1.12\sqrt{\alpha/n}$. Since including an additional interval increases the exact confidence coefficient by approximately $1/n$, we need an additional $1.12\sqrt{n\alpha}$ intervals. We thus obtain the following approximation.

Approximation 2. For $n\alpha$ large, the value of k that gives exact confidence coefficient $1-\alpha$ is approximately $n(1-\alpha) + 1.12\sqrt{n\alpha}$.

To test the accuracy of Approximation 1, we did a simulation study. For each of several choices of l ranging from $l = 100$ to $l = 10{,}000$, we estimated c_l using the same method used in computing Table 1. For each l, we generated 100,000 standard uniform samples of size $2l-1$ and found the length of the shortest of the intervals $(0, U_{(l)}), (U_{(1)}, U_{(l+1)}), \ldots, (U_{(l)}, 1)$. We averaged the 100,000 lengths to obtain an estimate of c_l, and we also computed the simulation standard error, which in each case was less than 0.0001. We then compared the simulated c_l values to the values given by Approximation 1. Results of this comparison are shown in Table 4. We see from the table that Approximation 1 is very good, but slightly low for small l. This means that when Approximation 1 is used in combination with Theorem 2 to estimate the exact confidence coefficient for a given choice of n and k, the estimate will tend to be conservative.

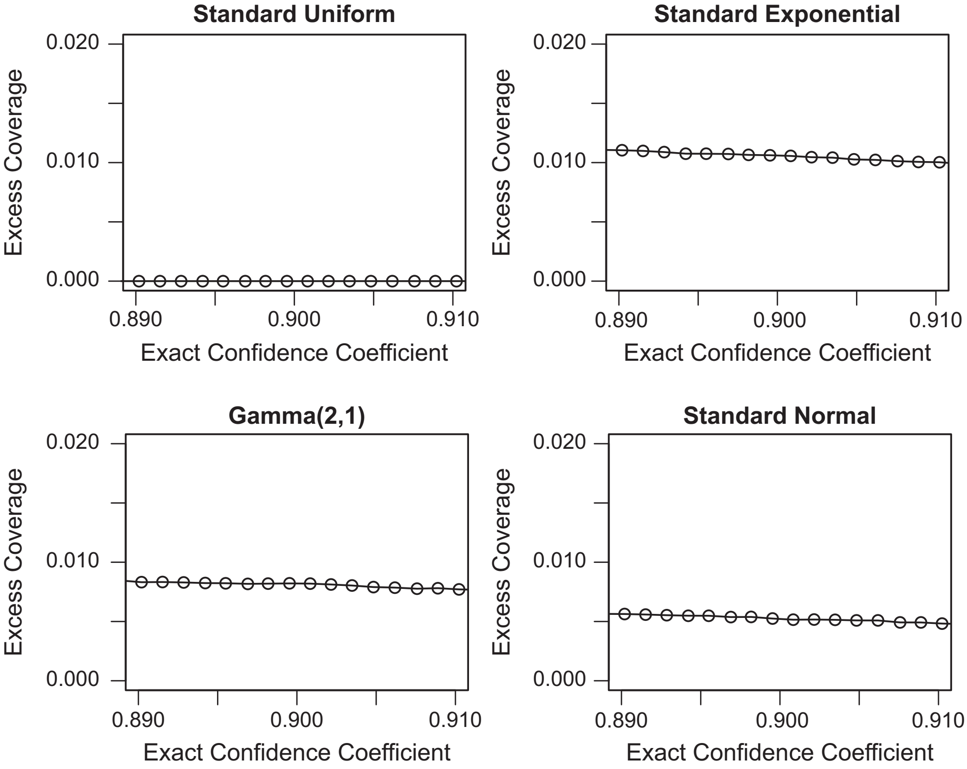

Fig. 6. Simulated excess coverage versus simulated exact confidence coefficient for the new data-driven intervals when the exact confidence coefficient is near 90% and the sample size is 800.

Table 3
Results of the simulation study used to estimate c. Each estimate was based on 100,000 runs, and each simulation standard error was less than 0.001.

Points	10	100	1000	10,000	100,000
Estimate	0.627	0.741	0.780	0.793	0.795

Table 4
Comparison of simulated c_l values and approximate values obtained from Approximation 1. Each simulation standard error was 0.0001 or less.

l	Simulated c_l	Approximate c_l
100	0.4484	0.4440
200	0.4624	0.4604
500	0.4757	0.4750
1000	0.4827	0.4823
2000	0.4876	0.4875
5000	0.4921	0.4921
10,000	0.4944	0.4944

6. Discussion

We have proposed new data-driven nonparametric prediction intervals for a single future observation, and we have shown through a simulation study that these intervals outperform the usual equal-tailed nonparametric prediction intervals. The data-driven prediction intervals are slightly longer than the equal-tailed intervals when the underlying distribution is symmetric, but they are much shorter than the equal-tailed intervals when the underlying distribution is skewed.

The intervals can be used when the set $\mathcal{F}$ of distributions under consideration is the set of all continuous distributions, and an argument from Scheffé and Tukey (1945, Section 5) shows that, as is typical for nonparametric statistical intervals, the intervals can also be used when $\mathcal{F}$ is the set of all distributions, provided that we think of the intervals as closed intervals instead of open intervals. Thus, our prediction intervals may be applied to discrete distributions and to data sets where there are ties.

In this paper, we have applied Theorem 1 only for the case where we are choosing the shortest of the intervals of the form $(X_{(r)},X_{(r+k)})$ for $1 \le r \le n-k$. However, the theorem can be applied more generally to expand the number of available exact confidence coefficients. For example, suppose that the initial sample size is $n=99$. The exact confidence coefficients that are available if we use the standard equal-tailed prediction intervals are 98% (for the interval $(X_{(1)},X_{(99)})$), 96%, 94%, and so on. However, if we take the shorter of two intervals that are offset by one subinterval, we obtain the additional exact confidence coefficients 96.5% (for choosing the shortest of $(X_{(1)},X_{(98)})$ and $(X_{(2)},X_{(99)})$), 94.5%, 92.5%, and so on. This idea also extends to cases where we choose the shortest of more than two intervals, and as long as the number of intervals from which we choose is kept small, there is very little excess coverage of the sort that we see in Figs. 5 and 6. Thus, Theorem 1 provides an alternate way to obtain a nonparametric prediction interval with a desired confidence coefficient. Unlike the interpolated intervals described by Hall and Rieck (2001), the intervals obtained using Theorem 1 offer guaranteed coverage.

Acknowledgment

The author thanks the two anonymous reviewers for helpful comments and suggestions.

References

Chen, Z., Bai, Z., Sinha, B.K., 2004. Ranked Set Sampling: Theory and Applications. Springer, New York.

David, H.A., Nagaraja, H.N., 2003. Order Statistics, third ed. Wiley, New York.

Di Bucchianico, A., Einmahl, J.H.J., Mushkudiani, N.A., 2001. Smallest nonparametric tolerance regions. Annals of Statistics 29, 1320–1343.

Fligner, M.A., Wolfe, D.A., 1979. Nonparametric prediction intervals for a future sample median. Journal of the American Statistical Association 74, 453–456.

Frey, J., 2010. Data-driven nonparametric tolerance sets. Journal of Nonparametric Statistics 22, 169–180.

Hahn, G.J., Meeker, W.Q., 1991. Statistical Intervals: A Guide for Practitioners. Wiley, New York.

Hall, P., Rieck, A., 2001. Improving coverage accuracy of nonparametric prediction intervals. Journal of the Royal Statistical Society, Series B 63, 717–725.

Platt, W.J., Evans, G.W., Rathbun, S.L., 1988. The population dynamics of a long-lived conifer (*Pinus palustris*). The American Naturalist 131, 491–525.

Pyke, R., 1965. Spacings. Journal of the Royal Statistical Society, Series B 27, 395–449.

Scheffé, H., Tukey, J.W., 1945. Non-parametric estimation. I. Validation of order statistics. Annals of Mathematical Statistics 16, 187–192.

Vardeman, S.B., 1992. What about the other intervals? The American Statistician 46, 193–197.

Wilks, S.S., 1941. Determination of sample sizes for setting tolerance limits. Annals of Mathematical Statistics 12, 91–96.