Stony Brook University

Step Density Estimation and Bootstrap Resampling

A Dissertation Presented

by

Yeming Ma

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

December 2007

Stony Brook University
The Graduate School

Yeming Ma

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Nancy Mendell - Dissertation Advisor
Professor, Department of Applied Mathematics and Statistics

Stephen Finch - Chairperson of Defense
Professor, Department of Applied Mathematics and Statistics

Hongshik Ahn
Professor, Department of Applied Mathematics and Statistics

Haipeng Xing - Outside Member
Assistant Professor, Department of Statistics,
Columbia University

This dissertation is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Dissertation

Step Density Estimation and Bootstrap Resampling

by

Yeming Ma

Doctor of Philosophy

in

Applied Mathematics and Statistics

Stony Brook University

2007

Some failures of the nonparametric bootstrap resampling originate from the discreteness of the empirical distribution function used in the resampling process. Density estimation with smoothing kernel functions is the most suitable method to solve the problem; yet in reality density estimation had not been widely applied due to its tedious fine-tuning of smoothing width in addition to the ad hoc selection of smoothing kernel from many candidate functions. With the above restrictions in mind a novel Step Density Estimation has been devised from simple step-functions in this thesis. The step density function has been constructed and shown to be MLE and UMVUE as an estimator of the underlying distribution with a clear goal to make the density estimation objective as possible while keeping the smoothed bootstrap still as simple as it is. A well known bootstrap bias problem in small sample cases was chosen to test the success of the approach of bootstrap resampling drawn from the step density function.

Table of Contents

List of Figures vi

List of Tables viii

Preface ix

Acknowledgement xi

1. Introduction to bootstrap resampling 1

1.1 Nonparametric bootstrap resampling procedures 1

1.2 Empirical density function (*edf*) 4

1.3 Other resampling methods 6

1.4 Some bootstrap failures 12

- Local discrete density problem 12
- Global density regulation needed 15
- Small sample problem 16

2. Review of density estimation 18

2.1 Histogram 19

2.2 The naïve estimator 20

2.3 The kernel method 22

2.4 The nearest neighbor method 25

2.5 The variable kernel method 27

2.6 A short discussion 28

3. Step density estimation 30

3.1 Definition of step density function (*sdf*) 30

3.2 Two key issues 32

- Discreteness of *edf* 33
- Continuity and adaptiveness of *sdf* 34

3.3 Properties of *sdf* 35

- Easy to construct 35

- MLE 36
- UMVUE 38
- Ready for bootstrap resampling 40
- Limitations of *sdf* 41

4. Application of step density function 43

4.1 Local density estimated by step density function 43

4.2 Unique mode selection 45

4.3 First-order approximation for further smoothing 52

5. Bootstrap bias in small sample 53

5.1 Bootstrap bias for sample mean 54

5.2 Two-level structure of the bootstrap bias 57

- Bootstrap level 58
- Sample level 59

5.3 Bootstrap imputation using step density function 61

- Bootstrap level: uniform bias reduction 63
- Sample level: optimal bias reduction 64

5.4 Mechanism of bootstrap bias 66

6. Summary and future work 69

Reference 71

List of Figures

Figure.1.1 Schematic diagram of the bootstrap applied to problems with a general data structure $P \to \mathrm{x}$. The crucial step "$\Rightarrow$" produces an estimate $\hat{P}$ of the entire probability mechanism P from the observed data x. The rest of the bootstrap picture is determined by the real world: "$\hat{P} \to \mathrm{x}^*$" is the same as "$P \to \mathrm{x}$"; the mapping from $\mathrm{x}^* \to \hat{\theta}^*, s(\mathrm{x}^*)$, is the same as the mapping from $\mathrm{x} \to \hat{\theta}, s(\mathrm{x})$. 3

Figure. 1.2 The histogram with 100 bins that imitate the needle plot to show the original 50 data points ~ $U(0,1)$ on (a), and the 2000 nonparametric bootstrap replications obtained sampling with replacement from the empirical distribution function, $\widehat{F_n}$. 13

Figure. 1.3 The histogram with 100 bins that imitate the needle plot to show the 2000 parametric bootstrap replications with random sampling from uniform distribution $U(0,\hat{\theta})$ in (a) and 2000 nonparametric bootstrap replacement from the step density function, $\overline{F_n}$ in (b). 14

Figure.2.1 Two histograms with different starting point of bins of eruption lengths of the Old Faithful geyser. 20

Figure.2.2 Naïve estimate constructed from Old Faithful geyser data, h=0.25. 22

Figure. 2.3 Kernel estimates showing individual kernels. Window widths: (a) 0.2; (b) 0.4; (c) 0.8. 24

Figure.2.4 Kernel estimate for Old Faithful geyser data, window width 0.25. 24

Figure.2.5 Nerest neighbor estimate for Old Faithful geyser data, window width k = 20. 26

Figure 3.1 Needle Plot constructed on the example data set, {1, 1.2, 2, 3, 4, 5, 10, 12, 12.5, 13, 14}. 31

Figure 3.2 Step density function, $\widehat{F_n}$, for the example data set, {1, 1.2, 2, 3, 4, 5, 10, 12, 12.5, 13, 14}. 32

Figure 4.1 The histogram with 100 bins that imitate the needle plot to show the original 50 data points ~ $U(0,1)$ on (a), and the 2000 nonparametric bootstrap replications obtained sampling with replacement from the empirical distribution function, $\widehat{F_n}$. 44

Figure 4.2 The histogram with 100 bins that imitate the needle plot to show the 2000 parametric bootstrap replications obtained sampling with $U(0,\hat{\theta})$ in (a) and 2000 nonparametric bootstrap replacement from the step density function, $\overline{F_n}$ in (b). 45

Figure 4.3 Histogram of the mouse data: (top) control group, (bottom) treatment. 47

Figure 4.4 Needle plot for the mouse data: (top) control group, (bottom) treatment. 48

Figure 4.5 The unique step density estimate of the mouse data in linear scale: (top) control group, (bottom) treatment. 49

Figure 4.6 The unique step density estimate of the mouse data in logarithmic scale: (top) control group, (bottom) treatment. 50

Figure 4.7 Step density estimate of the combined mouse data in linear scale. 50

Figure 4.8 Histogram of the proportion of mice in each group with survival time > 80 days from bootstrap resampling with step density function, *B=10000*. The top panel is for the control group, and the bottom panel for the treatment group. 51

Figure 5.1 Schematic diagram of the bootstrap applied to problems with a general data structure $P \to \mathrm{x}$. The crucial step "$\Rightarrow$" produces an estimate $\hat{P}$ of the entire probability mechanism P from the observed data x. The rest of the bootstrap picture is determined by the real world: "$\hat{P} \to \mathrm{x}^*$" is the same as "$P \to \mathrm{x}$"; the mapping from $\mathrm{x}^* \to \hat{\theta}^*, s(\mathrm{x}^*)$, is the same as the mapping from $\mathrm{x} \to \hat{\theta}, s(\mathrm{x})$. 57

Figure 5.2 The uniformly imputed bootstrap resampling scheme had been used for $(\overline{Y_m} - \overline{X_n})$ with bootstrap sample size *B=5000* and Monte Carlo simulation repetitions 1000, and the performance were reported at imputation factor *I = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,* and *100*. 62

Figure 5.3 The uniformly imputed bootstrap resampling scheme had been used for $\left(\overline{Y_m} - \overline{X_n}\right) / \widehat{\sigma_{X_n}}$ with bootstrap sample size *B=5000* and Monte Carlo simulation repetitions 1000, and the performance were reported at imputation factor *I = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,* and *100*. 63

Figure 5.4 The uniformly imputed bootstrap resampling scheme had been used for $\left(\overline{Y_m} - \overline{X_n}\right) / \widehat{\sigma_{Y_m}}$ with bootstrap sample size *B=5000* and Monte Carlo simulation repetitions 1000, and the performance were reported at imputation factor *I = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,* and*100*. 65

List of Tables

Table. 1.1. Basic properties of the sampling and resampling methods. 10

Table 2.1 Eruption lengths (in minutes) of 107 eruptions of the Old Faithful Geyser. 19

Table 4.1 The mouse data. Sixteen mice were randomly assigned to a treatment group or control group. Shown are their survival times, in days, following a test surgery. Did the treatment prolong survival? 46

Table 5.1 Simulation and theoretical expectations, normal distribution. 55

Table 5.2 Simulation and theoretical expectations, normal distribution. 56

Table 5.3 Comparison of three-simulation results, normal distribution. 59

Table 5.4 Comparison of three-simulation results, normal distribution. 60

Preface

The *probability density function (pdf), or* simply the *density F*, is a fundamental concept in probably and statistics. If "Every road leads to Rome" then *probability density function* would have been the Rome in probability and statistics since every problem leads to it.

In the old days of R.A. Fisher when statisticians had little computing power (with papers and tabulated distributions), *F* was often restricted to functional families with some unknown parameters, for example the location and dispersion (μ, σ^2) family, to be determined by sample data points, and thus the term "*parametric*" method. When *F* is taking a more flexible form other than a set of functional families, it is termed as "*nonparametric*", which is less rigid and more adaptive to the observed data. The nonparametric bootstrap, also referred to as B. Efron's naïve bootstrap (Efron, 1979a), is the most frequently used among the *nonparametric* methods. *Density estimation* is to construct an estimated *density function*, $\hat{F}$, for the underlying population *F* from the observed sample.

For a random sample of size *n* from *F*,

$$F \rightarrow (x_1, x_2, \ldots, x_n), \qquad (1)$$

an *empirical density function* (*edf*) $\hat{F}_n$ is defined to be the discrete distribution that puts probability 1/n on each value x_i, which can be scalar, vector or any data unit, $i = 1, 2, \ldots, n$. In other words, $\hat{F}_n$ assigns to a set *A* in the sample space of x its empirical probability

$$\hat{P}\{A\} = \#\{x_i \in A\} / n,$$

the proportion of the observed sample **x** = $(x_1, x_2,\ldots x_n)$ occurring in *A*. The hat symbol "^" indicates quantities calculated from the observed data. The *empirical density function* (*edf)* $\hat{F}_n$ is the simplest example of nonparametric density estimation*;* and the *step density function* (*sdf)* to be discussed in this thesis may be regarded just as another special case of nonparametric density estimation. Intuitively we may view density estimation as a reverse of (1)

$$(x_1, x_2,\ldots, x_n) \rightarrow \hat{F} \ . \qquad (2)$$

The nonparametric bootstrap or B. Efron's naïve bootstrap is mainly a device to estimate sample uncertainties done by *i.i.d.* sampling from the *edf* ($\hat{F}_n$)

$$\hat{F}_n \rightarrow (x_1^*, x_2^*, ..., x_n^*). \qquad (3)$$

In practice the bootstrap sample $(x_1^*, x_2^*, ..., x_n^*)$ is simply drawn with replacement form the sample, and the only difference between (1) and (3) is that *F* is replaced by $\hat{F}_n$, which had been vividly described by Efron as the *plug-in principle*.

Some failures of the nonparametric bootstrap resampling originate from the discreteness of $\hat{F}_n$. Density estimation with smoothing kernel functions is the most suitable method to solve the problem; yet in reality density estimation had not been widely applied due to its tedious fine-tuning of smoothing width in addition to the ad hoc selection of smoothing kernel from many candidate functions. With the above restrictions in mind a novel ***Step Density Estimation*** has been devised from simple step-functions to resolve a well known bootstrap bias problem in small sample cases.

Acknowledgement

I owe infinite gratitude to Professor Nancy R. Mendell for allowing me the more than usual freedom to select the dissertation topic and for her patience with my slow progress while working at NIH. Nancy had provided me with many insightful suggestions, careful guidance and excellent ideas that helped me immensely in improving my derivations and writings. I am very grateful to Professor Stephen J. Finch who, together with Nancy, spent several afternoons listening to my various proposals. I am also very thankful to my other committee members Professor Hongshik Ahn from the department of Applied Mathematics and Statistics, Stony Brook University, and Prof. Haipeng Xing from the Department of Statistics, Columbia University, for their valuable suggestions and constructive critics.

I am obliged to Dr. Nora D. Volkow for her encouragement in applying new statistical methods in data analysis as well as her advice for me to pursue part-time study in statistics. My interest in the bootstrap resampling had been first kindled during the analysis of brain imaging data taken from Dr. Nora D. Volkow's neuroimaging laboratory at Brookhaven national Laboratory.

While writing the thesis, many people helped me either directly or indirectly. Wei, my beloved wife, deserves an accolade for her immense patience and encouragement; as do my wonderful children, Victor and Merry. I want to acknowledge the special support from my family, my parents Mr. Xijiu Ma and Ms. Qin Wei, and my brother Mr. Haiming Ma. With their love I feel solving problems and writing solutions are just as refreshing and enjoyable as walking in the sea breeze.

Last, but not least, I want to express my gratitude to the employee education program of DOE (Department of Energy) and to the intramural research program of NIH (National Institute of Health) for their financial assistance during my part-time study.

Chapter 1. Introduction to the bootstrap

The bootstrap, due to B. Efron (1979), is a computer intensive resampling method, which has been gaining popularity since its invention and is the most frequently used among all the resampling plans, such as Jackknife, subsampling or half-sampling, etc. Its root goes back to the earlier methods of Fisher's permutation tests and jackknife. Excellent reviews (Young, 1994, Hall, 2003, D. Boos, 2003) and books (Efron and Tibshirani, 1992, Shao and Tu, 1995, Davison and Hinkley, 1997) are available at different depth after three decades of intensive exploration in both theory and application. By the year of 2004, more than 1000 papers had been published on bootstrap (Efron, 2004). However one still could not tell if the unparallel bootstrap phenomenon has reached its climax or not because new problems of bootstrap arise before existing problems are resolved, and among which lays the intriguing bootstrap bias problem we would study next.

One common feature that all the resampling methods share is that they generate many "pseudo" samples from one observed sample through the resampling procedures. From the "pseudo" samples, statistical inferences, particularly measures of accuracy such as standard errors and confident intervals, could be drawn under a central assumption that these "pseudo" samples from the *empirical density function* (*edf*) $\hat{F}_n$ could be treated the same as, or very similar to, real samples that had never been collected from a true population distribution F. Therefore it might be reasonable to imagine that *edf* $\hat{F}_n$ may play a key role in the bootstrap method and even be responsible to some of its failures or "pathologies".

1.1 The bootstrap resampling procedures

Suppose we are in a common data analysis situation: a random sample **x** =$(x_1, x_2, \ldots, x_n)$ from an unknown distribution F had been observed and we wish to estimate a parameter of interest $\theta = t(F)$ on the basis of **x**. For this purpose we calculate an estimate $\hat{\theta} = s(x)$ from **x**, which could be the *plug-in* estimate of $t(F)$. The *plug-in principle* is a simple method of estimating parameters from samples. The plug-in estimate of $\theta = t(F)$ is defined by

$$\theta = t(\hat{F}).$$

When the bootstrap was first introduced in 1979 (B. Efron, 1997) as a data-based simulation method, its main purpose was to estimate the standard error of $\hat{\theta}$. It enjoyed the advantage of being completely automatic without any theoretical calculations and applicable no matter how mathematically complicated the estimation $\hat{\theta} = s(x)$ may be.

The bootstrap methods depend on the notion of a bootstrap sample, $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$, defined to be a random sample of size n drawn from the *empirical density function* (*edf*) $\hat{F}_n$,

$$\hat{F}_n \rightarrow (x_1^*, x_2^*, \ldots, x_n^*).$$

The star notion indicates that x^* is not the actual data set *x*, but rather a randomized, or resampled, version of *x*. Corresponding to a bootstrap data set, x^*, is a bootstrap replication of $\hat{\theta}$,

$$\hat{\theta}^* = s(x^*).$$

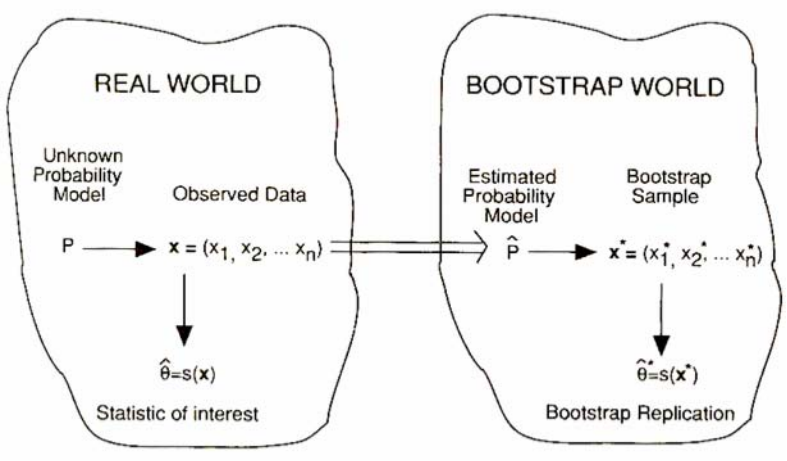


Figure.1.1 Schematic diagram of the bootstrap applied to problems with a general data structure $P \to \mathrm{x}$. The crucial step "$\Rightarrow$" produces an estimate $\hat{P}$ of the entire probability mechanism P from the observed data x. The rest of the bootstrap picture is determined by the real world: "$\hat{P} \to \mathrm{x}^*$" is the same as "$P \to \mathrm{x}$"; the mapping from $\mathrm{x}^* \to \hat{\theta}^*, s(\mathrm{x}^*)$, is the same as the mapping from $\mathrm{x} \to \hat{\theta}, s(\mathrm{x})$.

The bootstrap estimate of $se_F(\hat{\theta}) = [\mathrm{var}_F(\hat{\theta})]^{\frac{1}{2}}$, the standard error of a statistic $\hat{\theta}$, is a *plug-in* estimate that uses the *edf* $\hat{F}_n$ in place of the unknown population *pdf F*, which is called the *ideal bootstrap estimate* of $se_F(\hat{\theta})$ and defined by

$$se_{\hat{F}_n}(\hat{\theta}^*).$$

Let μ_F and σ_F^2 be the expectation and the variance of the real-valued random variable X, $X \sim (\mu_F, \sigma_F^2), \forall F$. When $\hat{\theta} = \overline{X}$, we have $se_F(\overline{X}) = [\mathrm{var}_F(\overline{X})]^{\frac{1}{2}} = \sigma_F / \sqrt{n}$. Unfortunately for virtually any estimate $\hat{\theta}$ other than the mean, there is no explicit formula like this. The bootstrap algorithm, described next, is a computational way of obtaining a good approximation to the numerical value of $se_{\hat{F}_n}(\hat{\theta}^*)$.

Bootstrap: Step 1. Fit the nonparametric MLE of *F,*

$$\hat{F}: mass\ 1/n\ at\ x, \qquad i = 1,2,...,n. \tag{1.1}$$

Bootstrap: Step 2. Select B independent bootstrap samples, $x^{*1}, x^{*2}, ..., x^{*B}$, each consisting of n data values drawn with replacement from sample x. Evaluate the bootstrap replication corresponding to each bootstrap sample, $\hat{\theta}^*(b) = s(x^{*b}),\ b = 1,2,...,B.$

Bootstrap: Step 3. Estimate the standard error $se_F(\hat{\theta})$, by the sample standard deviation of the *B* replications

$$\hat{se}_B = \{\sum_{b=1}^{B} [\hat{\theta}^*(b) - \hat{\theta}^*(.)]^2 /(B-1)\}^{\frac{1}{2}}, \text{ where } \hat{\theta}^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b)/(B-1).$$

The bootstrap procedure for standard error has been used as an example. The bootstrap sample size is recommended by Efron, B=200 for standard error estimations and B=1000 for confidence interval estimations. The bootstrap method can be easily adapted to many problems by simply modifying Step #3 to extract the statistics of interest from the bootstrap samples. The simplicity of the bootstrap procedure provided a powerful set of solutions for applied statisticians; mathematically, however, it is also highly evolved as a rich source of theoretical and methodological problems for statistics (A.C. Davison, D.V. Hinkley and G.A. Young, 2003).

1.2 The role of empirical density function (*edf*)

In the bootstrap procedure, we can see that Step #1 is the essential step, which determines the bootstrap sample distribution. The justification of the empirical distribution function as a nonparametric maximum likelihood estimate was studied by Kiefer and Wolfowitz (1956) and Scholz (1980).

The next two steps are only the mathematical processing of the bootstrap samples generated from the i.i.d. sampling from Step #1. The procedure of drawing a bootstrap sample independently with replacement from sample x has guaranteed the identical distribution from which the bootstrap samples were drawn, which is the empirical density function $(edf)\,\hat{F}_n$. Thus bootstrap method satisfies the traditional theory *i.i.d.* random sample,

$$F \xrightarrow{i.i.d.} (x_1, x_2, ..., x_n).$$

Using the 'plug-in' principle we replace $\widehat{F}_n \to F$, and have the bootstrap resampling process presented as following,

$$\hat{F}_n \xrightarrow{i.i.d.} (x_1^*, x_2^*, ..., x_n^*).$$

From the discussion on density estimation in the next chapter we would see that $\hat{F}_n$ is the MLE of F (B. Efron, 1982, p.28), and the simplest density estimator without any smoothing, that is the smoothing bandwidth, $h = 0$. The major advantage of this choice lies in the simplicity in the resampling step #2, which had thus been referred to as automatic by B. Efron. Any smoothing on $\hat{F}_n$ is to choose a density estimate with nonzero bandwidth, $h \neq 0$.

Silverman and Young had furbished a general discussion on simulation from density estimates including standard bootstrap as well as smoothed bootstrap (B.W. Silverman, 1986; B.W. Silverman and G.A. Young, 1987; D. de Angelis and G.A. Young, 1992). In Silverman's example of density estimation, the bump-hunting problem, he used bootstrap as a resampling plan to construct a test for multimodality in search for a value h, namely the critical window width or bandwidth. Our goal here is to remedy the bootstrap method, especially failures caused by the discreteness of the $edf, \hat{F}_n$, by introducing density estimation with an automatically chosen bandwidth.

One critical observation on discrete *edf* is referred to by Silverman as 'rather peculiar". That is nearly every bootstrap sample contains repeated values, and when n is large most samples contain values repeated several times. We will see that this 'peculiar property' would sometimes result in failures for the bootstrap method in our later discussion (Beran and Ducharme, 1991). Although there were more repeated values when n is large, the problem due to discreteness of the plug-in distribution, $\hat{F}_n$, is less severe as $n \to \infty$. Asymptotically both the density estimation and the bootstrap method are exact at $n \to \infty$, which is intuitive because when sample size approaches infinity, the sample itself is the true population; therefore any density estimation from the sample would approach that of the population (J. Shao, 1997). The real challenge is for small sample size, *n*, and how to provide the bootstrap with a smoothed $\widehat{F}$ to resolve the failures due to the discreteness of $\hat{F}_n$ for small sample.

1.3 Bootstrap and other resampling methods

In contrast to conventional parametric methods, resampling methods are aimed to generate a sampling distribution of a statistic by drawing random samples from the observed sample itself. This eliminates the need to assume a specific functional form for the population distribution such as normality. Obviously the sampling distribution thus obtained is, after all, only one particular realization from the population. The argument is that the error resulted from this kind of resampling distribution could be far less serious than making wrong and often unverifiable assumptions about the population distribution, *F*.

Another method closely related to resampling methods, the Monte Carlo simulation usually starts with a known *F*, and/or the probabilistic mechanism that could be realized by a computer program. A well-known Monte Carlo simulation might be the "Buffon's needle" experiment first stated by Georges Louis de Buffon in 1777 in the early days of probability and statistics. By throwing a needle of length *L* on the table with grid of parallel lines with spacing *D (D>L),* we may easily compute the chance that

the needle intersects one of the lines is $2L/\pi D$. And the proportion of "intersects" in N throws can be measured by the experiment as $\widehat{p_N}$, and the estimate of π is

$$\hat{\pi} = \lim_{N\to\infty} \frac{2L}{\widehat{p_N} \bullet D}.$$

It had been fondly used as an example to demonstrate the basic idea of Monte Carlo method with a probability mechanism completely known that can be practically used to estimate the mathematical constant of π (J. Liu, 2002).

Resampling methods are data-based simulation without the full knowledge of F or its probability mechanism but merely a random sample x drawn from F. Permutation tests, jackknife and subsampling methods are the resampling methods that were predated to bootstrap and had profound influence on the development of bootstrap method.

R.A. Fisher introduced the permutation test (randomization test) in the 1930's, a computer-intensive statistical technique before modern computers' invention, therefore it bared with a modest goal as a theoretical argument supporting Student's t-test than as a useful statistical method in its own right. Modern computers make it feasible to use permutation tests on a routine basis, mainly for the two-sample problem. The basic idea is attractively simple. These procedures were used for determining statistical significance directly from the data without recourse to some particular sampling distribution. It tests the null hypothesis

$$H_0 : F = G$$

based on two independent samples drawn from possibly different probability distributions F and G, $\quad F \xrightarrow{i.i.d.} (x_1, x_2, ..., x_{n_1})$, $G \xrightarrow{i.i.d.} (y_1, y_2, ..., y_{n_2})$.

And the achieved significance level (*ASL*) of the test defined to be the probability of observing at least that large a value under H_0,

$$ASL_{perm} = \Pr(\widehat{\theta^*} > \hat{\theta}),$$

where $\hat{\theta}$ is the observed statistic difference between the two samples, and $\widehat{\theta^*}$ is generated through permutations of the combined sample

$$(x_1, x_2, ..., x_{n_1}, y_1, y_2, ..., y_{n_2}).$$

If ASL_{perm} is smaller than some significance level α, the results are significant at the level α. A permutation test exploits special symmetry under the null hypothesis to create a permutation distribution of the test statistic. As a result of this symmetry, the ASL_{perm} from a permutation test is exact in the sense that ASL_{perm} is the exact probability of obtaining a test statistic as extreme as the one observed, having fixed the data values of the combined sample.

It would be interesting to compare the permutation test (B. Efron and R.J. Tibshirani, 1992, p.223) to the bootstrap method for hypothesis testing under the null hypothesis,

$$H_0 : F = G = \widehat{F_{n_1+n_2}}.$$

The bootstrap resampling is carried out from the combined sample $(x_1, ..., x_{n_1}, y_1, ..., y_{n_2})$ which formed the $\hat{\mathrm{F}}$ *edf* $\widehat{F_{n_1+n_2}}$. In contrast, the bootstrap explicitly estimates the probability mechanism under the null hypothesis, and then sample from it to estimate the

$$\widehat{ASL_{boot}} = \widehat{\Pr}(\widehat{\theta^*} > \hat{\theta}) = \#\{\widehat{\theta^*} > \hat{\theta}\} / B.$$

Bootstrap removes the restrictions in permutation tests, such as every permuted sample has the same combination which violates the independency of the i.i.d. aspect of the original problem.

Jackknife was first proposed by Quenouille (1949) for estimating bias. Recognizing its potential for estimating standard errors, J. Tukey coined the name "jackknife" (1958). Further development was made by Miller (1964, 1974), Gray and Schucany (1972), Hinkley (1977), Reeds (1978), Parr (1983,1985), Hinkley and Wei (1984), Sen (1988), and Wu (1986) in the linear regression setting. It has been developed from its original popular delete-1 scheme into delete-d jackknife (Shao and Wu, 1989; Shao, 1991) to amend the inconsistency for non-smooth statistics like median or percentile. Jackknife focuses on the samples that leave out one observation at a time; therefore the jackknife has a finite sample space at the same size of the original sample. Its close form is an advantage of jackknife estimate over permutation and bootstrap, both frequently need simulations with uncertainties; however the delete-*d* jackknife at $d \sim \sqrt{n}$ would make its performance similar to the bootstrap (J. Shao 1992). The tradeoff is that the jackknife sub-sample size would be necessarily reduced by deleting *d* data points than that in the original sample. In bootstrap, every bootstrap sample would enjoy the same number of observations as the original sample, which is usually the default inferential interest in most cases. Thus, the bootstrap method has an advantage of estimating the accuracy at the original sample size though it is not limited to it (Fan and Wang, 1996).

Since bootstrap's conception (B. Efron, 1978), jackknife was usually treated as a golden standard against which to compare partly due its connectivity. Theoretically Efron had shown that jackknife could be viewed as a linear approximation of bootstrap that should be applied to a linear statistics that can be written in the form of

$$\hat{\theta} = s(x) = \mu + \frac{1}{n}\sum_{i=1}^{n}\alpha(x_i),$$

where μ is a constant and $\alpha(.)$ is a function, because jackknife is inefficient for higher order statistics where bootstrap can be applied successfully.

Subsampling is also called random subsampling, which was first introduced by Hartigan (1969) as another resampling plan designed to give exact confidence intervals, rather than just the standard deviations. It may be viewed theoretically a super class of nonparametric resampling methods that may include jackknife and half-sampling etc., which were developed for dealing with special problems: that of estimating the center of a symmetric distribution on the real line. Its merits and limitations were studied in theory and in application (Hartigan, 1969, 1971, 1975; McCarthy, 1969) as well as in comparison to bootstrap (B. Efron, 1982).

Among the numerous volumes on the above nonparametric methods, works on detailed comparisons were abundant while clear delineation on the fundamental differences was much less (Efron and Tibshirani, 1992, p.216). The differences underlying the major resampling methods may be delineated from their resampling procedures. The original sample was always assumed under the *i.i.d.* condition, while in contrast the jackknife and permutation were more deviated from the *i.i.d.* condition than the bootstrap as listed in Table 1.1.

Table. 1.1. Basic properties of the sampling and resampling methods.

	Original	Bootstrap	Jackknife	Permutation
independent	*Yes*	*Yes*	*No*	*No*
Identically distributed	*Yes*	*Yes*	*No*	*Yes*
Density function	F	$\hat{F}_n$	*N.A.*	$F=G$

From the above table we may see that bootstrap resampling most closely resembles the original sampling process while the other two listed in the table are different. The fundamental differences in the probability mechanism may account for some of the excellent properties of bootstrap over the other resampling methods (B. Efron, 1982). The important difference is between the unknown density F, and the discrete MLE $\widehat{F}_n$ via the plug-in principle $\widehat{F}_n \to F$. Therefore it is reasonable to study

smoothed bootstrap beyond that of the simplest plug-in estimate $\hat{F}_n$ (B. Efron, 1982; B.W. Silverman and G.A. Young, 1987; D. de Angelis and G.A. Young, 1992).

Jackknife samples are not independent from each other. Denote a jackknife sample without one original value x_i as x_{-i}. The difference between any two jackknife samples, x_{-i} and x_{-j}, are only in one values that x_{-i} has x_j but not x_i and x_{-j} has x_i x_j but not x_j; all the remaining (n-2) values are all the same. In term of correlation these jackknife samples are strongly correlated, especially pair-wise inter-correlated. To put it in another way, the sampling is dependent so the samples are not random samples at all. It is all due to the "sample without replacement" restriction that best illustrated in the delete-d jackknife that at each step of deleting *i-th* value, the distribution is changing from the previous deleting step; therefore there is no identically distributed population space, so the *i.d.* condition was also violated. Then why jackknife had been quite successfully applied to many problems for so many years? There are two major reasons.

1. The identical distribution still held approximately after deleting only a few out of a relatively large sample size n. From the *i.i.d.* condition each sample value contributes $\frac{1}{n}$ of the total information contained in the whole sample about the density, so the majority $(1-\frac{1}{n})$ of information remained.

2. On the other hand a larger sample size n would cause even stronger correlation between samples, which had been mended by the $\sqrt{n-1}$ correction factor that were derived from the simplest sample statistic, sample mean, and expanded to all statistics without much justification.

Permutation test certainly has all its samples identical in content but with different order, which make partial use of the independency of the *i.i.d.* condition, that is the order of each observed values are random and can be randomly permuted which still makes a plausible sample observation. Correlations among permuted samples are stronger than

that among jackknife samples, which is at its maximum possible value of 1; interestingly permutation test safely gets around of the identical distribution violation in jackknife with an extremely strong hypothesis in a very specific two-sample setting of

$$H_0 : F = G.$$

The permutation method is impossible to be adopted for parameter estimation, a frequently encountered statistical problem, such as problem of estimating the standard error of one sample mean. It could not be accomplished by permutation because the entire permutated sample was identical. Thus permutation test applies mainly to inferences on at least two samples.

1.4 Some bootstrap failures

Bootstrap method has been enjoying huge success in application since its invention, and as a convenient toolbox it has been implemented in many statistical software packages, such as in SAS and Matlab, for error estimation as well as confidence interval construction (A.C. Davison and D.V. Hinkley, 1997). Its flexibility stems from its simple mathematical structure that has frequently been touted as "Such a simple idea!" However, on the other side, after nearly 30 years of enthusiastic theoretical exploration we may find that its theoretical foundation was still mainly in the asymptotic framework (J. Shao, 1995), or at most quite large sample cases. Here three bootstrap failures, or pathologies as some prefer to call it, were listed next. The first two problems were considered as partially resolved with parametric or smoothed bootstrap; while the last is to our best knowledge still considered as an open question that we are trying to provide a solution. All three problems related to the properties of the density function in either local or global scale, therefore are served to illustrate the close connection of density estimation to bootstrap in general.

- **Local discrete density problem**

This is a well-known example that B. Efron had used to illustrate situations when bootstrap failed by citing the original work of Beran and Ducharme (1991, page 23).

$$F \xrightarrow{i.i.d.} (x_1, x_2, ..., x_n),\ F \sim U(0,\theta).$$

The MLE $\hat{\theta}$ is the maximum of the sample in this example, $x_{(n)}$. A sample of 50 uniform numbers in the range $(0,1)$ is generated and computed $\hat{\theta} = 0.988$. The left panel of Fig. 1.2 shows a histogram of 50 sample points, and the right panel of 2000 bootstrap replications obtained sampling with replacement from the data. The left panel of Fig. 1.3 shows 2000 parametric bootstrap replications obtained by sampling from the uniform distribution on $U(0,\hat{\theta})$. It is evident that the right panel of histogram in Fig. 1.3 is a poor approximation to the histogram we expect. Particularly, on the right left histogram it had a large probability mass at $0.62 \times \hat{\theta}$ of the value $\theta^* = \hat{\theta}$. In general, it is easy to show that

$$\mathrm{P}(\theta^* = \hat{\theta}) = 1 - \mathrm{P}(\theta^* \neq \hat{\theta}) = 1 - (1 - \frac{1}{n})^n \xrightarrow{n \to \infty} 1 - \frac{1}{e} \approx 0.632.$$

However, in the parametric setting of the right panel, $\mathrm{P}(\theta^* = \hat{\theta}) = 0$.

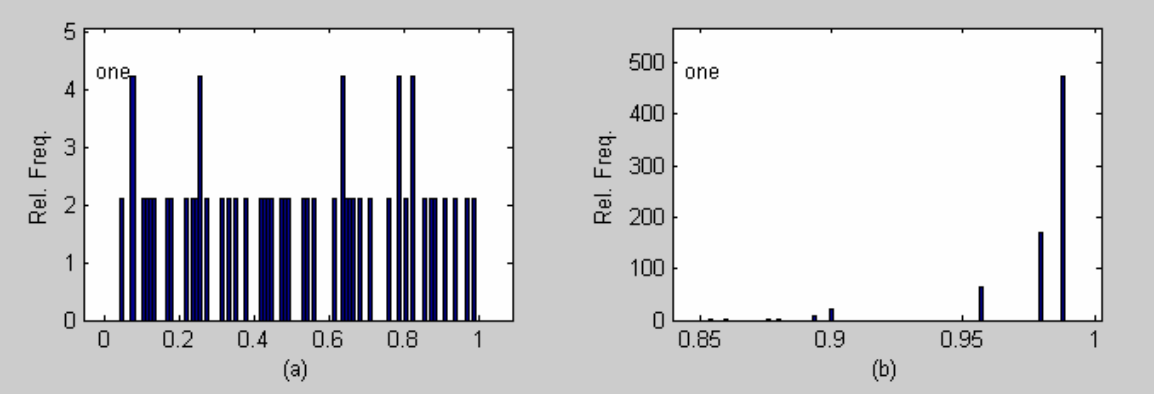


Figure. 1.2 The histogram with 100 bins that imitate the needle plot to show the original 50 data points ~ $U(0,1)$ on (a), and the 2000 nonparametric bootstrap replications obtained sampling with replacement from the empirical distribution function, $\widehat{F}_n$.

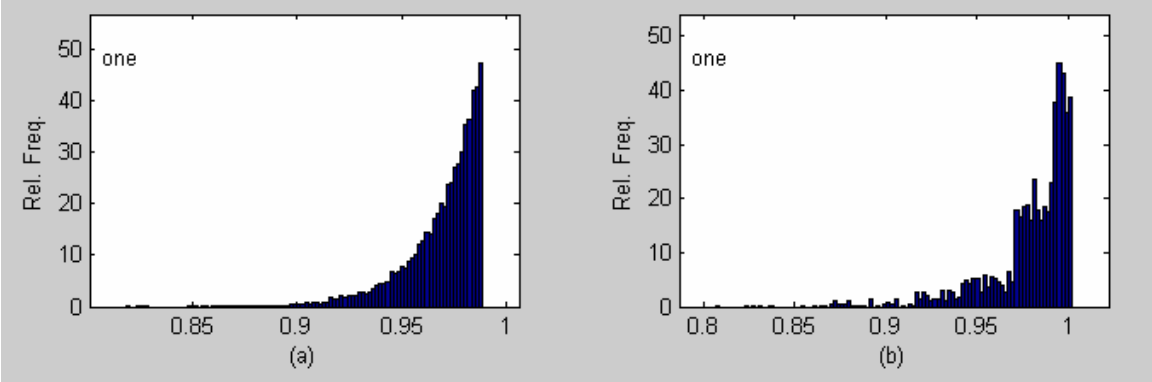


Figure. 1.3 The histogram with 100 bins that imitate the needle plot to show the 2000 parametric bootstrap replications with random sampling from uniform distribution $U(0,\hat{\theta})$ in (a) and 2000 nonparametric bootstrap replacement from the step density function, $\overline{F_n}$ in (b).

We think it is proper to the cited the exquisite insight from Efron on the issue.

"What goes wrong with the nonparametric bootstrap? The difficulty occurs because the empirical distribution function $\widehat{F}$ is not a good estimation of the true distribution F in the extreme tail. Either parametric knowledge of F or some smoothing of $\widehat{F}$ is needed to rectify matters. The nonparametric bootstrap can fail in other examples in which θ depend on the smoothness of F."

The discrete problem of the *empirical distribution function* had been recognized by Efron and applied the parametric strategy with result that was displayed in the left panel of Figure1.4 that nicely fixed the problem. As for the second suggestion of some smoothing on $\widehat{F}$, which had been the old idea that Efron introduced as "smoothed bootstrap" (B. Efron, 1982), Efron did not further pursue. However there were many studies on smoothing the discrete density and bootstrap performances on confidence interval and related applications had been done later with various degree of success (D.B. Rubin, 1981, N. Schenker, 1985, J. Farway, 1990, and A.M. Polansky and W.R. Schucany, 1997, S. Wang, 1989, 1995, Young, 1990a,).

We are taking the same strategy in our step density estimation, which is another smoothed bootstrap among the many existed except it is simpler than all of them, or the simplest smoothing of $\widehat{F}$. The unique approach of *step density estimation* lies in its automatic estimation of the smoothing bandwidth and the ease to be implemented into nonparametric bootstrap.

- **Global density regulation needed**

K.B. Athreya (1987) proved the theorem on bootstrap of the mean in the infinite variance case. The nonparametric bootstrap was referred as the naïve bootstrap in the theorem (followed some authors of the early days of bootstrap). It was shown that if the population distribution has its second moment

$$EX^2 = \infty ,$$

then the bootstrap mean would have a random distribution (given the sample) whose limit is also a random distribution implying that the naïve bootstrap could fail in the heavy tailed case. Apparently it was taken at once as a severe failure of bootstrap, which I would cite the remark in the same paper of K.B. Athreya.

"What, if any, is the significance of the Theorem? It says that if one does a naïve bootstrap on the sample mean and if the underlying population does not have a finite variance then the bootstrap distribution will not converge to the same limit as the sample mean. Thus, conducting confidence intervals on the basis of a Monte Carlo simulation of the bootstrap could yield misleading results. So unless one is reasonably sure that underlying distribution is not heavy tailed, one should hesitate to use the naïve bootstrap. In particular, in variance estimation using bootstrap could be bad if the underlying distribution has no fourth moment."

However I have a second thought 20 years after Athreya's paper that if we look up the table of common distributions, for example in Casella and Berger's textbook (2002), there is only one distribution, the Cauchy distribution that is with an unbounded

second momentum, $EX^2 = \infty$. Note that the mean of the Cauchy distribution actually does not exist strictly, but since the distribution is symmetric, it is generally taken to be μ.

$$c(x;\mu,\alpha) = \frac{1}{\pi\alpha}\frac{1}{1+(x+\mu)^2/\alpha^2},$$

$$mean = \mu, median = \mu, \text{Variance}=\infty, \text{ mode}=\mu.$$

When $EX^2 = \infty$, strictly speaking even the population mean, *EX*, does not exist. To fully appreciate the very peculiar property of $EX^2 = \infty$ in the Cauchy distribution, we may recall its sample mean $\overline{Z}$ of $Z_1, Z_2, ..., Z_n \sim$ i.i.d. Cauchy $(0,\sigma^2)$ is also Cauchy $(0,\sigma^2)$. Or in other words, $\text{var}(\overline{Z}) = \text{var}(Z)$, because $\text{var}(\overline{Z}) = \frac{\text{var}(Z)}{n}$ only holds when $EX^2 < \infty$. By definition the unavailability of sample mean under $EX^2 = \infty$ is indeed an intrinsic property of statistics that handle sample of finite size. Since it is such a "failure of statistics" (if we may call it) there would be no surprise that it is also a "failure of the bootstrap". However looking at this from another angle, we would have been surprised if the bootstrap does not fail when $EX^2 = \infty$. In other words, we may conclude that bootstrap is a generally applicable device that could be relied to for any distribution once EX^2 is bounded $(EX^2 < \infty)$.

- **Small sample**

G.A. Young and H.E. Daniels (1990) discussed the bias in the nonparametric bootstrap, which appears to have been introduced by using the *edf*, $\hat{F}_n$, in place of the true distribution *F*. It was shown by their simulation study that the bootstrap mean was biased for small sample sizes. The bias was indicated by the estimated bootstrap *pdf* of the

sample mean, $\bar{X}_B$, which was noticeably and systematically wider in comparison to the standard analytical results as a gold standard. In terms of moments of the distribution, the simulation study indicated that the bootstrap estimate of $\bar{X}_B$ had shown to be a biased estimator of variance. Due to the limitations of a numerical simulation study, the mathematical structure or the bias mechanism had been poorly understood; and the exact analysis of bias based on saddle point approximation also seems stretched to its limit for the simple statistics $\bar{X}_m - \bar{X}_n$ already. Therefore further investigation on the bias issue remains on other estimators of practical importance. The current view is that it is a failure of bootstrap method at small sample size and that favorable asymptotic property is no guarantee of good small-sample performance.

In a review (G.A. Young, 1994) the importance of the issues are raised again in a more general form including but not limited to small sample issues, by a were asked under the title "Bootstrap: more than a stab in the dark?" including "When does bootstrap work?", "When does bootstrap fail to provide valid inference?", "Are cases of failure pathological, or practically significant?". While the bootstrap is continuing to make a fundamental impact on how we carry out statistical inferences for problems without analytic solutions, the small sample bias mechanism of bootstrap still remain a mystery to date. This thesis intends to unravel the mechanism and provide some simple guidelines for applying bootstrap in small sample cases.

Chapter 2. Review of density estimation

Despite progress on the technical aspects of density estimation, it is still more of theoretical importance than application relevance. In contrast the bootstrap method is most valued for its application in real world data analysis especially when no analytic solution is available while its theoretical development is mainly restricted to asymptotic properties (J.Shao and D. Tu, 1995). The subtle gap between the two important ingredients of bootstrap has been widely recognized in an effort to investigate its internal connections (B. Silverman, 1986; B. Efron, 1992; P. Hall, 2003).

Traditional approach to density estimation is parametric by assuming that the data are drawn from one of a known parametric family of distributions, for example the normal distribution with mean μ and variance σ^2. The underlying distribution F could then be estimated by estimates of μ and σ^2 refined by the data and substituting these estimates back to F. Usually the nonparametric approach will be far more less rigid, with much less prior assumptions made about the distribution of the observed data; instead a nonparametric class of function $\hat{F}$ would be estimated from the sample more directly $(x_1, x_2, \ldots, x_n) \rightarrow \hat{F}$.

Nonparametric density estimation was first proposed as a way of freeing discriminant analysis from rigid distributional assumptions. Since then, density estimation and related ideas have been used in a variety of contexts. We will present a very brief review of the existing nonparametric density estimation methods for univariate density estimation and their most important properties. These properties will serve as the yardstick against which the step density estimator will be measured in Chapter 3.

The data set used to help illustrate the methods was the observations of eruptions of the Old Faithful geyser at the Yellowstone National Park provided by Weisberg (1980) and reproduced in Table 2.1.

Table 2.1 Eruption lengths (in minutes) of 107 eruptions of the Old Faithful Geyser.

4.37	3.87	4.00	4.03	3.5	4.08	2.25
4.7	1.73	4.93	1.73	4.62	3.43	4.25
1.68	3.92	3.68	3.10	4.03	1.77	4.08
1.75	3.2	1.85	4.62	1.97	4.50	3.92
4.35	2.33	3.83	1.88	4.60	1.80	4.73
1.77	4.57	1.85	3.52	4.00	3.70	3.72
4.25	3.58	3.80	3.77	3.75	2.50	4.50
4.10	3.7	3.80	3.43	4.00	2.27	4.40
4.05	4.25	3.33	2.00	4.33	2.93	4.58
1.90	3.58	3.73	3.73	1.82	4.63	3.50
4.00	3.67	1.67	4.6	1.67	4.00	1.80
4.42	1.9	4.63	2.93	3.5	1.97	4.28
1.83	4.13	1.83	4.65	4.2	3.93	4.33
1.83	4.53	2.03	4.18	4.43	4.07	4.13
3.95	4.1	2.72	4.58	1.9	4.50	1.95
4.83	4.12					

2.1 Histogram

The oldest and most widely used density estimator is the histogram. It bins the sample $(x_1, x_2, \ldots, x_n)$ into the intervals

$$[x_0 + mh, x_0 + (m+1)h),$$

where x_0 is the origin, m is positive or negative integers and h is the bin width. The intervals have been closed on the left and open on the right for definitiveness. The histogram is thus defined by

$$\hat{f}(x) = \frac{1}{nh} \#\{x_i \text{ in the same bin as } x\}$$

Histogram is piece-wise continuous, which is discontinuous at the boundaries between bins. The choices of an origin x_0 and bin width h were in general arbitrary despite of many rules of thumb from experience that could be followed. The problem become more severe for small sample size for example n = 10~20, as different choices of origin x_0 and bin width h would produce dramatically different density estimates and it is nearly impossible to determine which estimate would be the right choice. The choice of the bin width primarily controls the amount of smoothing inherent in the procedure. Histogram remains an excellent tool for data presentation mainly for quite large sample size. Histogram is not very sensible for small sample as it requires that the data to be grouped first.

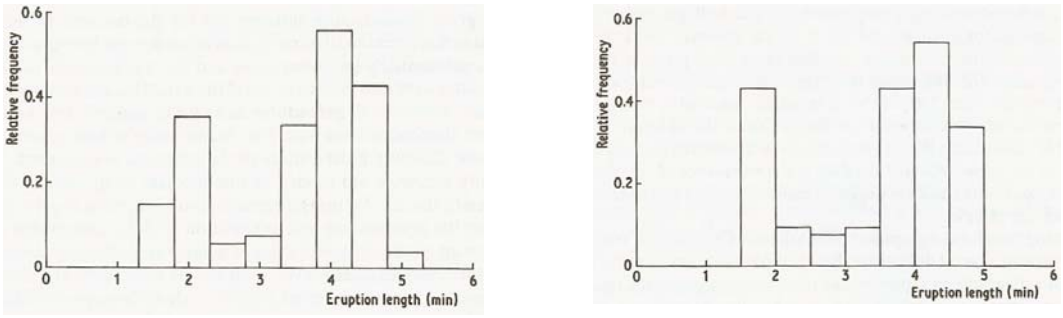


Figure.2.1 Two histograms with different starting point of bins of eruption lengths of the Old Faithful geyser.

Therefore one of the challenges for density estimation is how to obtain a unique density function that is reliable for small sample size.

2.2 The naïve estimator

By definition the *pdf* of random variable X is

$$f(x) = \lim_{h \to 0} \frac{1}{2h} P(x - h < X < x + h).$$

Thus a natural estimator $\hat{f}$ of the density is given by choosing a small number h and setting

$$\hat{f}(x) = \frac{1}{2nh} \#\{x_i \in (x - h < X < x + h)\}$$

which is called the naïve estimator. Define a weight function w by

$$w(x) = \begin{cases} \frac{1}{2} & if \mid x \mid < 1, \\ 0 & otherwise. \end{cases}$$

Then it is easy to write the naïve estimator as

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w(\frac{x - X_i}{h}).$$

Imagine that the estimate is constructed by placing a 'box' of width $2h$ and height $(2nh)^{-1}$ on each observation and then summing up to obtain the estimation.

The naïve estimator can be seen as an attempt to construct a histogram that every data point is the center of a sampling interval, thus freeing the histogram from a particular choice of bin positions. Here $\hat{f}$ is also piece-wise continuous similar to the histogram. The choice of the bin width h still remains unresolved, which controls the amount of smoothing to produce the estimate.

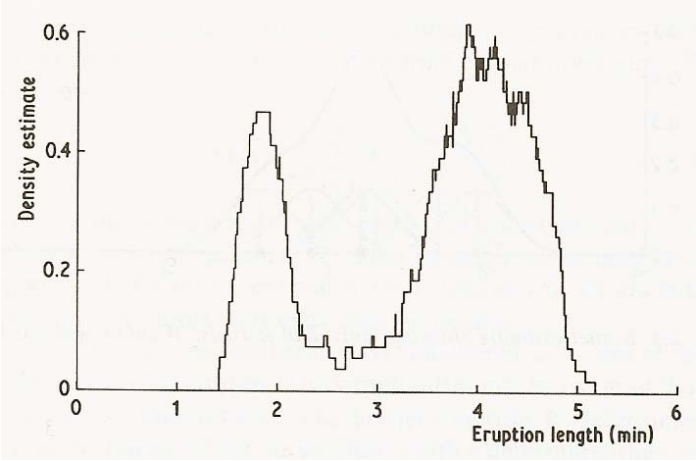


Figure.2.2 Naïve estimate constructed from Old Faithful geyser data, h=0.25.

2.3 The kernel estimator

The kernel estimator is an extension of the naïve estimator by replacing the weight function *w* by a kernel function *K* that satisfies the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1.$$

Usually, but not always, *K* will be a symmetric probability density function, the normal density, for instance, or the 'box' weight function *w* used in the definition of the naïve estimator. By analogy to the definition of the naïve estimator, the kernel estimator with kernel *K* is defined by

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh} K(\frac{x - X_i}{h}).$$

The window width *h* is also called the smoothing parameter or bandwidth. Just as the naïve estimator can be imagined as a sum of 'boxes' centered at the observations, the kernel estimator is a sum of 'bumps' placed at the observations. The kernel function *K* determines the shape of the bumps while the window width *h* determines their width. An

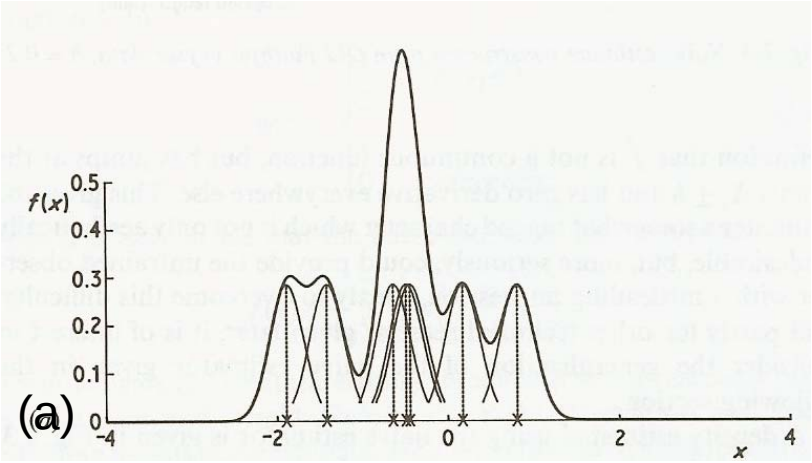
illustration is given in Figure. 2.3, individual bumps $\frac{1}{h}K(\frac{x-X_i}{h})$ with different window width h are shown as well as the estimate $\hat{f}$ constructed by adding them up. Two elementary properties of kernel estimates may follow at once

(1) $\hat{f}$ is a *pdf,* if and only if K is *pdf;*

(2) $\hat{f}$ has the same continuity and differentiability properties of the kernel K.

For example if K is the normal density function, then $\hat{f}$ will be a smooth curve having derivatives of all orders.

Apart from the histogram, the kernel estimator is probably the most commonly used density estimator and certainly the most studied mathematically. The major drawback is the fixed window width h across the entire sample. The amount of smoothing might be too much for the higher density locations, where more observations were clustered; but not enough for the lower density locations, where fewer observations were found, which are usually at the tails. Various adaptive methods have been proposed, and these are discussed in the next two sections.



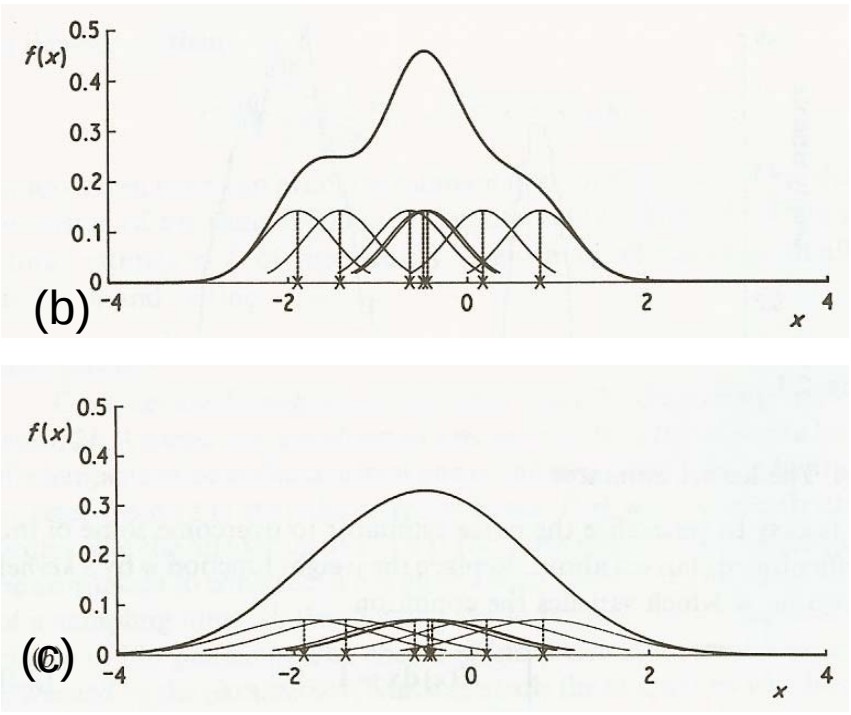


Figure. 2.3 Kernel estimates showing individual kernels. Window widths: (a) 0.2; (b) 0.4; (c) 0.8.

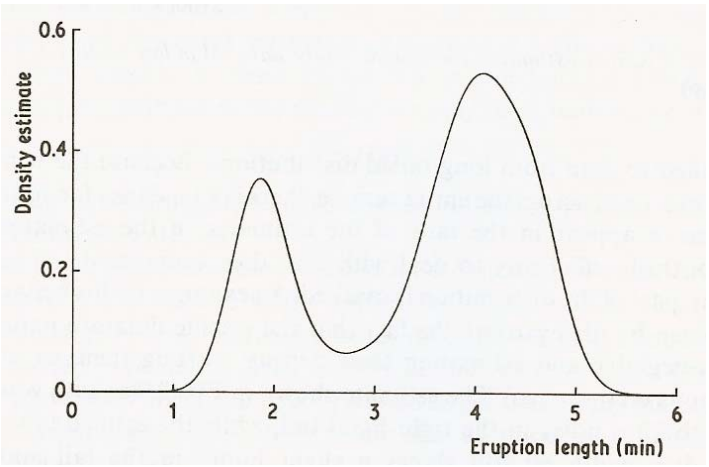


Figure.2.4 Kernel estimate for Old Faithful geyser data, window width 0.25.

2.4 The nearest neighbor method

The nearest neighbor estimator represents an attempt to adapt the amount of smoothing to the 'local' density of data. The degree of smoothing is controlled by an integer *k*, chosen to be considerably smaller than the sample size; typically $k \approx \sqrt{n}$. Define the distance $d(x, y) = |x - y|$, and for each *t* define

$$d_1(t) \le d_2(t) \le \ldots \le d_n(t)$$

to be the distances, arranged in ascending order, from *t* to the point of the sample, x_i.

The *k-th* nearest neighbor density estimator is then defined by

$$\hat{f}(t) = \frac{k}{2nd_k(t)}$$

One would expect about $2rnf(t)$ observations to fall into the interval $[t-r, t+r]$ for each $r > 0$. By definition exactly *k* observations fall into the interval $[t - d_k(t), t + d_k(t)]$, an estimate of the density at may be obtained by rearranging the following

$$k = 2d_k(t)n\hat{f}(t).$$

At the lower density locations, the distance $d_k(t)$ will be larger than that at the higher density locations of the distribution, and so the problem of under-smoothing in the tails should be reduced. The nearest neighbor estimator is not a smooth curve like the naïve estimator; to be precise, $\hat{f}$ is first order continuous with discontinuous derivative at some locations.

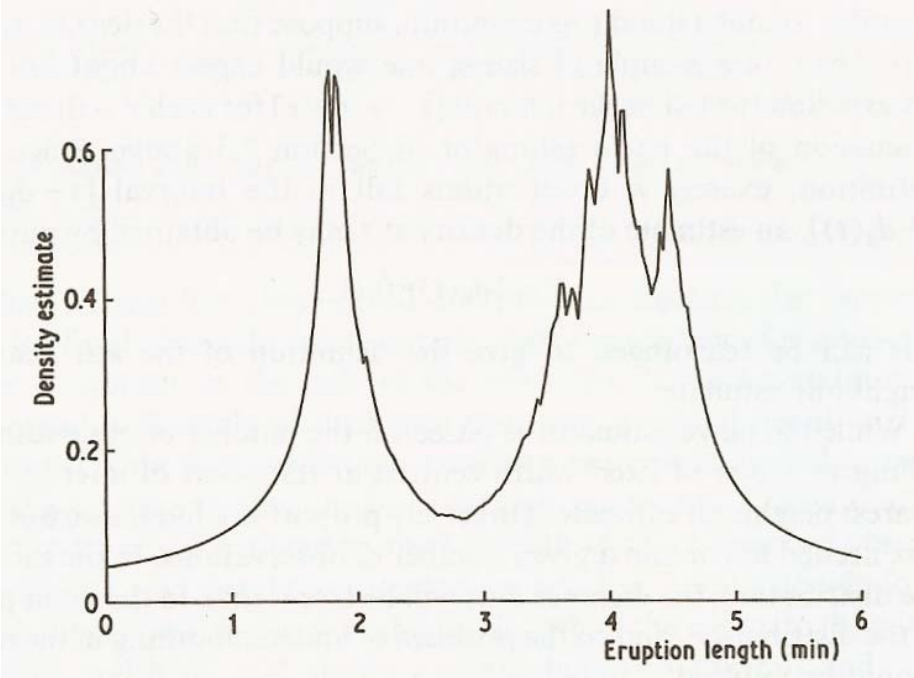


Figure.2.5 Nerest neighbor estimate for Old Faithful geyser data, window width k = 20.

In contrast to the kernel estimator, the nearest neighbor estimator does not integrate to unity therefore is not automatically a density function. Intuitively $\hat{f}(t) = \frac{k}{2nd_k(t)}$ is derived from the local property of $d_k(t)$ for parameters k and n according to the k-nearest neighbor restriction, so sophisticated computational techniques were usually necessary. An alternative solution called the generalized k-nearest neighbor estimate is defined by

$$\hat{f}(t) = \frac{1}{nd_k(t)} \sum_{i=1}^{n} K\left(\frac{t - X_i}{d_k(t)}\right).$$

It can be seen at once that this is the kernel estimator evaluated at t with different window width $d_k(t)$. The global smoothing is governed by the choice of the integer k, but the window width $d_k(t)$ used at any particular point for local smoothing depends on the density of observations near that point of t.

2.5 Variable kernel method

The variable kernel method, is somewhat related to the nearest neighbor approach and is another method that adapts the amount of smoothing to the local density of the data. The estimate is constructed similarly to the classical kernel estimate, but the smoothing parameter of the 'bumps' placed on the data points is allowed to vary from one data point to another.

Define $d_{j,k}$ to be the distance from X_j to the k-nearest point in the set comprising the other n-1 data points. Then the variable kernel method with smoothing parameter h is defined by

$$\hat{f}(t) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{hd_{j,k}}K\left(\frac{t-X_i}{hd_{j,k}}\right).$$

The window width of the kernel placed on the point X_j is proportional to $d_{j,k}$, so that data points in regions where the data are sparse will have flatter kernels associated with them. For any fixed k, the overall degree of smoothing will depend on the parameter h. The choice of k determines how responsive the window width choice will be to the very detail.

The subtle difference from the generalized k-nearest neighbor estimator may be compared with the variable kernel estimator. Note $d_k(t)$ depends on the distance from t to the data points; while $d_{j,k}$ is independent of t at which the density is estimated, and depends only on the distances between the data points.

Another important generalization of variable kernel method, the adaptive kernel method, has been developed with more flexibility to handle the complexity in reality, however based on the same common-sense notion that a natural way to deal with long-tailed densities is to use a broader kernel in regions of low density. Thus the observation in the low density part will spread out a wider range than one in the high density part.

2.6 Discussion

The theoretical importance of density estimation had been well established and an increasing number of conscientious authors have been working in the technical aspects to make it accessible to a wider audience in light of the central role of the *probability density function*, which is like the Rome that every road leads to it.

The density estimators introduced above cover most of the important ideas, from which we may see that the smoothing window width or bandwidth, *h,* plays a central role in all the methods introduced above. The details of the nonparametric density estimates and the corresponding results of data analysis were at influenced by *h*. Thus various plans for the selection of *h* have been the key ingredient that differentiates each of the density estimators. This is rather intuitive because density estimation is all about "filling up" the gaps between observations at *n* discrete data points, and *h* is the measure of this "filling up" process; and *h* should be optimized according to some criteria either globally or locally.

Density estimation with smoothing kernel functions is the most suitable method to solve the problem; yet in reality density estimation had not been widely applied due to its tedious fine-tuning of smoothing width in addition to the ad hoc selection of smoothing kernel from many candidate functions. To confirm the existence of a unique optimized *h* in the frame work of functional analysis can be a daunting task in theory as well as in application; and the lack of general criteria for optimization made the optimization rather problem-specific if not arbitrary. Sometimes after the painstaking process of selecting the *h,* we may found that it is not sensitive to the analysis results at all. This reminds us of Fisher's witty comment --- "it is not only shooting a sparrow with a cannon, it might even miss the sparrow as well." That is why density estimation has not been widely used in exploratory data analysis, such as bootstrap. However there are wider applications of smoothing other than statistical density estimation. Another major application in statistics might be the "smoother" used in regression (T. Hastie and R. Tibshirani, 1990), and in

engineering fields for such applications like noise filtering and signal smoothing whenever the bandwidth, *h,* can be effectively optimized for example with tools such as Fourier Transformation.

Chapter 3. Theory of step density estimation

In this chapter we will discuss the motivation and the criteria to construct a density estimation, the step density function (*sdf)*, to replace the discrete empirical density function (*edf)* for bootstrap resampling. The criteria were raised from studies of smoothed bootstrap, which is an old idea when bootstrap first invented (B. Efron, 1982) and researched by Silverman and Young (1987), Hall, Diciccio and Ramano (1989). In general, there is no global preference for procedures based on a smoothed version of the empirical distribution function rather than the empirical density function itself. In the majority of problems smoothing only influences the second order properties of the estimator while requiring greater computation and a suitable amount of smoothing (de Angelis and Young, 1992). However the step density function we will construct is aiming to automatically provide an objective amount of smoothing with the minimum amount of increase in the computation budget.

Throughout this chapter it will be assumed that we have a sample $X_1,...,X_n$ of independent, identically distributed observations from a continuous univariate distribution with probability density function *F*. There are many practical problems where these assumptions are not necessarily justifiable, but nevertheless they provide a standard frame work in which to discuss the properties of the step density function (*sdf)*.

3.1 Definition of step density function (*sdf*)

Step density function (*sdf*) is the result of step density estimation procedure, which produce a piece-wise continuous smoothed version of the *edf* that can be readily applied to nonparametric bootstrap resampling. Considering a random sample of size $n = 11$ from an unknown probability distribution *F*,

$$F \to \{1, 1.2, 2, 3, 4, 5, 10, 12, 12.5, 13, 14\}.$$

The observed sample is plotted via a "needle plot" in Figure 3.1, which put weight 1 on every observed data point; so a normalized needle plot is the *edf*.

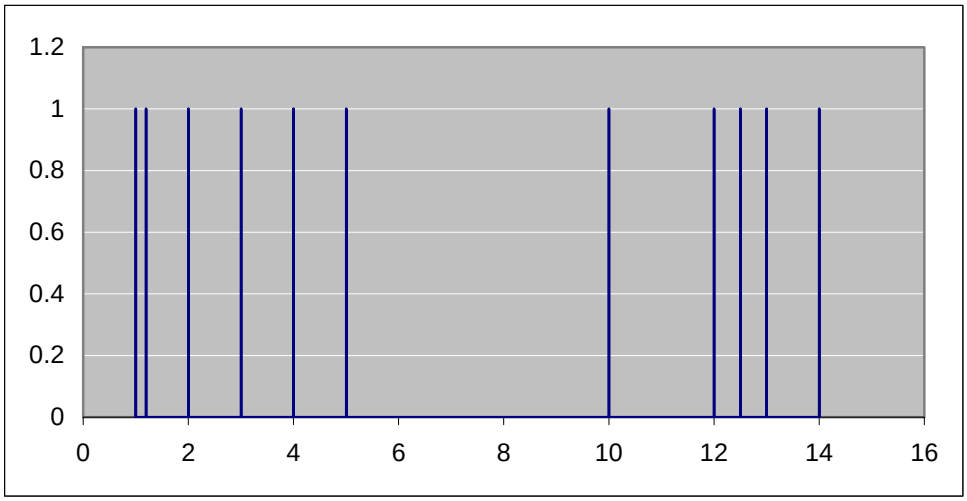


Figure 3.1 Needle Plot constructed on the example data set, {1, 1.2, 2, 3, 4, 5, 10, 12, 12.5, 13, 14}.

The Step density function (*sdf*) is defined as a class of functions that partition the domain of the sample values $[x_{(1)} - \Delta_{(1)}, x_{(n)} + \Delta_{(n)}]$ that covers the sample points with a set of bins with varied width $d_{(i)}, \{\Delta_{(1)}, d_{(2)}, ..., d_{(n)}, \Delta_{(n)}\}$; and the set of middle intervals of $\{d_{(2)}, d_{(3)}, ..., d_{(n)}\}$ and the two end intervals $\Delta_{(1)}, \Delta_{(n)}$ together form a partition of the domain $[x_{(1)} - \Delta_{(1)}, x_{(n)} + \Delta_{(n)}]$, and there would be no overlap between any two bins. It is like a histogram with variable bin width that keeps each and every bin with only one sample point in it; i.e. $x_{(i)}$ will fall into bin with width, $d_{(i)}$. There are some freedoms in choosing the way of dividing the bins by selecting the set of $\{d_{(1)}, d_{(2)}, ..., d_{(n)}\}$ and $\Delta_{(1)}, \Delta_{(n)}$; and it is the freedom of this partition scheme that we may take advantage to construct the density estimation in such a way that is computational efficient when applying to bootstrap resampling.

Once the set of $\{\Delta_{(1)}, d_{(2)}, ..., d_{(n)}, \Delta_{(n)}\}$ were determined the probability density on each bin width, $d_{(i)}$, will be fixed as a constant $\frac{1}{nd_{(i)}}$; and the *sdf* constructed this way would be shown in section 3.3 that it has the nice properties of MLE, and UMVUE as an estimator of the unknown density function, *F*. And we will use a new symbol $\widehat{F_n}$ to denote the *sdf*, with *the* hat '^' symbol representing the shape of the function. Step density function, $\widehat{F_n}$, is constructed on the example data set, {1, 1.2, 2, 3, 4, 5, 10, 12, 12.5, 13, 14}. Figure 3.2 is the $\widehat{F_n}$ constructed by defining $d_{(i)} = d_{(i)} - d_{(i-1)},\ i = 2, ..., n;$ and $\Delta_{(1)} = \frac{d_{(2)}}{2}, \Delta_{(n)} = \frac{d_{(n)}}{2}$.

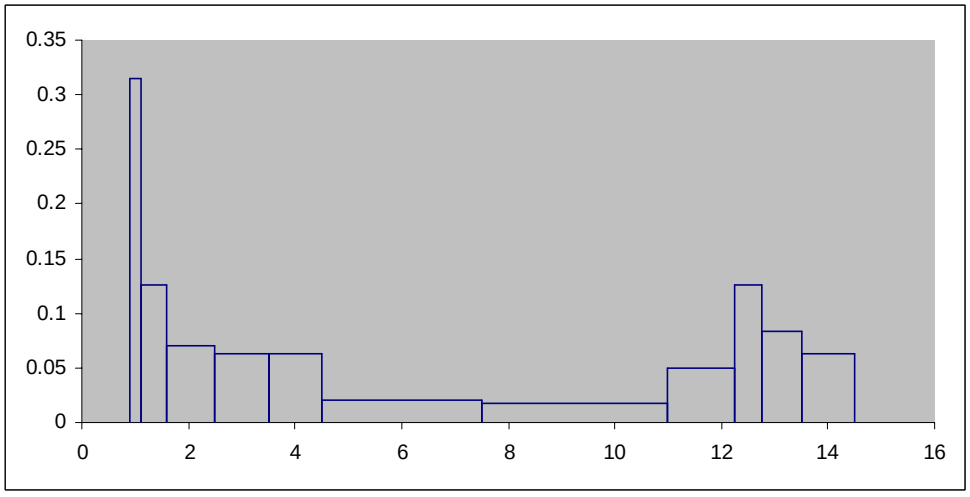


Figure 3.2 Step density function, $\widehat{F_n}$, for the example data set, {1, 1.2, 2, 3, 4, 5, 10, 12, 12.5, 13, 14}.

3.2 Two key issues

The problem we are trying to solve is the bootstrap failure in local extreme values as well as the global property of the distribution. In Chapter 1, we briefly outlined the motivation of the problem as searching for a properly constructed smoothed estimation of

the probability density function, which will not only smooth out the discreteness of *edf* but also satisfy two more criteria.

a) The first criterion is to eliminate the tedious fine-tuning of smooth bandwidth, *h*, and the ad hoc optimization rules, which compromises the objectiveness of bootstrap (bootstrap is usually considered a Frequentist method with the minimum amount of assumption beside the sample itself).

b) Secondly it is to make sure that the "smoothed" density function should add the minimum amount of computational effort to the usually computer-intensive bootstrap method.

The step density functions (*sdf*), developed in this work, will be measured according to the two criteria above.

- **Discreteness of *edf***

We start with the following natural question. Why the discreteness of empirical density function (*edf*) was not a problem for most bootstrap applications except for the extreme values at the tail or for very small sample size?

Apparently empirical density function (*edf*) is the simplest, as well as the most generally applicable density estimation. Smoothed density function is naturally available to univariate data, but is complicated to be extended to multivariate data analysis (B. Silverman, 1976) or in higher dimension geometric settings. In fact, the *edf* itself without smoothing guaranteed the wide application of bootstrap to problems with high dimension or complicated statistics without an analytic solution. The advantage of smoothing has a second order effect on the convergence of bootstrap estimator, and strongly depends on the underlying *F* and sample size *n* (D. de Angilis and G.A. Young, 1992). An adaptive bandwidth $h(x_i)$, which varies at different location x_i, is preferred to reflect the shape of

F; and the most effective case should be for small sample size *n* because the smoothing effect would diminish when $n \to \infty$.

B. Efron explained when smoothed bootstrap is necessary and simple guidelines on the discrete issue: "*the empirical distribution function is not a good estimate of the true distribution F in the extreme tail*. *Either parametric knowledge of F or some smoothing of* $\hat{F}$ *is needed to rectify matters. The nonparametric bootstrap can fail in other examples in which* θ *depend on the smoothness of F*." The step density function (*sdf*) might be a solution to this class of problems as mentioned above and we are set to study procedures that can produce a piece-wise smoothed density estimation for the empirical distribution function for nonparametric bootstrap resampling.

- **Continuity and adaptiveness of *sdf***

As we illustrated in the last example, step density function (*sdf*) that covers the sample points with a set of bin width $\{\Delta_{(1)}, d_{(2)}, ..., d_{(n)}, \Delta_{(n)}\}$ is a piece-wise continuous function. There are other density estimators that are smooth at higher orders, such as the kernel method, the nearest neighbor method, the variable kernel method as shown in Chapter 2. We avoid those more smoothed density estimators for two reasons, (a) bandwidth selection, and (b) bootstrap application. By defining $d_{(i)} = d_{(i)} - d_{(i-1)},\ i = 2, ..., n;$ and $\Delta_{(1)} = \frac{d_{(2)}}{2}, \Delta_{(n)} = \frac{d_{(n)}}{2}$ in section 3.1, the selection of $h_{(i)} = d_{(i)}$ is objective without the optimization of *h*, which was equivalent to the 1-nearest neighbor density estimation and adaptive to the local density at x_i. The piece-wise smoothness of a density function is also mathematical tractable with very little computation added to the bootstrap resampling, which we will show in Chapter 5.

Despite both the histogram and the naïve estimator are step-wise continuous like *sdf*, they were fundamentally different from the *sdf*. The first two density estimations both have the smoothing parameter selection issue, i.e., the bin size selection for

histogram and the window width for the naïve estimator. Histogram usually needs a relatively larger sample size in order to have a reasonable number of data points in each bin, and the bin boundary may have a large effect on the histogram in the small sample example given in Chapter 2.1.

The local adaptive property of the estimated density function is a property to be stressed for small sample cases while for large sample the issue is less important. However, regardless of sample size it is important for extreme value estimation, where essentially the estimation heavily relies on the very few data points in the tail portion. The way we constructed the *sdf* is dependent on the data point and its k-nearest neighbor (*k=1*), by which the local adaptiveness tends to be automatically satisfied.

3.3 Properties of *sdf*

- **Easy to construct**

Having observed a random sample of size *n* from *F,*

$$F \rightarrow (x_1, x_2, \ldots, x_n),$$

we first rank the sample $(x_1, x_2, \ldots, x_n)$ into order statistic $(x_{(1)}, x_{(2)}, \ldots, x_{(n)})$, then calculate the middle intervals i.e. the distance between $[x_{(i-1)}, x_{(i)}]$,

$$d_{(i)} = x_{(i)} - x_{(i-1)}.$$

If $x_{(i-1)} = x_{(i)}$, we define $d_{(i+1)} = d_{(i)}$, with a number of degeneracy at $x_{(i)}$ denoted as $k_{(i)} = 1$ (and $k_{(i)} = 0$ if $x_{(i-1)} \neq x_{(i)}$); and in total there would have $(n - \sum_{i=2}^{n} k_{(i)})$ distinctive

distances in the set $\{\Delta_{(1)}, d_{(2)}, ..., d_{(n)}, \Delta_{(n)}\}$. The two end intervals were defined as $\Delta_{(1)} = \frac{d_{(2)}}{2}, \Delta_{(n)} = \frac{d_{(n)}}{2}$, which is half of its neighboring middle interval.

The density between interval $[x_{(i-1)}, x_{(i)}]$ is defined as reciprocal to $d_{(i-1)}$, i.e. on $\{d_{(2)}, ..., d_{(n)}\}$ the density as $\{\frac{1}{nd_{(2)}}, ..., \frac{1}{nd_{(n)}}\}$. The sum of the total density on the middle intervals would be $1 - \frac{1}{n}$, so at the two end intervals $\Delta_{(1)}$ and $\Delta_{(n)}$ we will add two density mass, $\frac{1}{nd_{(2)}}$ on $[x_{(1)} - \Delta_{(1)}, x_{(1)}]$ and $\frac{1}{nd_{(n)}}$ on $[x_{(n)}, x_{(n)} + \Delta_{(n)}]$. Please note that there can be another convenient partition to having $(2n+1)$ intervals in the $\{\Delta_{(1)}, d_{(2)}, ..., d_{(n)}, \Delta_{(n)}\}$ by inserting a new boundary point between any two points at its middle intervals. Then the density on the left and right sides of each point is adaptive to its neighbors.

- **Maximum likelihood estimator**

Vapnik (1996) has shown that the MLE of an unknown density *F* can not be generally available without any restrictions on the density estimate, usually given by a family of distribution functions to choose from. It applies to the parametric density estimation as to the choice of a family of distribution functions for $\widehat{F}$, or to the nonparametric density estimation as to the choice of a family of distribution functions for the local atomic smoothing kernels of the $\widehat{F_n}$. (Here *F* and *f*, $\widehat{F}$ *and* $\widehat{f}$ *might be used in a exchangeable way when there is no confusion existed.*)

Theorem1. For any partition, $\{d_{(1)}, d_{(2)}, ..., d_{(j)}\}$ *of an observed discrete sample space* $\{x_{(1)}, x_{(2)}, ..., x_{(n)}\}$*, with number of observations* n_i *in each region* $d_{(i)}$ *to partition the total number of observations,* $\{n_1, n_2, ..., n_j\}$*, and* $\sum_{i=1}^{j} n_i = n$*. Assume in general* $\widehat{f}$ *belong to a*

family of step functions that have density θ_i *in each region* $d_{(i)}$, *which is a "generalized histogram" with variable bin widths* $\{d_{(1)}, d_{(2)}, ..., d_{(j)}\}$. *Then when* $\theta_i = \frac{n_i}{nd_{(i)}}$, $\hat{f}$ *is the MLE of* f.

Proof. The generalized histogram with variable bin widths $\{d_{(1)}, d_{(2)}, ..., d_{(j)}\}$ can be written as $\hat{f}(x) = \sum_{i=1}^{j} \theta_i I_{d_{(i)}}(x)$, using

indicator function: $I_{d_{(i)}}(x) = 1$, *if* $x \in d_{(i)}$,

$I_{d_{(i)}}(x) = 0$, *if* $x \notin d_{(i)}$,

where multiple parameteres (θ_i) *need to be estimated*.

The likelihood function is defined as

$$L(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \hat{f}(x_i) = \prod_{i=1}^{J} \theta_i^{n_i} = (1 - \sum_{i=1}^{J-1} \theta_i)^{n_i} \prod_{i=1}^{J-1} \theta_i^{n_i}, \ (\because \theta_1 + \theta_2 + ... + \theta_J = 1)$$

and the log-*likelihood function is*

$$l(x_1, x_2, ..., x_n) = \log L(x_1, x_2, ..., x_n) = n_j \log(1 - \sum_{i=1}^{J-1} \theta_i) + \sum_{i=1}^{J-1} n_i \log \theta_i.$$

In searching for conditions for the extrema of $l(x_1, x_2, ..., x_n)$, we take the derivative of $l(x_1, x_2, ..., x_n)$, i.e.

$$(1) \qquad \frac{\partial l}{\partial \theta_i} = 0, \Rightarrow \frac{-n_j}{1 - \sum_{i=1}^{j-1} \theta_i} + \frac{n_i}{\theta_i} = 0 \Rightarrow \theta_i = \frac{n_i}{n_j}(1 - \sum_{i=1}^{j-1} \theta_i) = \frac{\theta_j}{n_j} n_i,$$

$$\therefore 1 = \theta_1 + \theta_2 + ... + \theta_j = \frac{\theta_j}{n_j} \sum_{i=1}^{j} n_i = \frac{\theta_j}{n_j} n,$$

$$\therefore \theta_j = \frac{n_j}{n}, \theta_i = \frac{n_i}{n};$$

$$(2) \qquad \frac{\partial^2 l}{\partial \theta_i^2} = \frac{-n_j}{(1 - \sum_{i=1}^{j-1} \theta_i)^2} + \frac{-n_i}{\theta_i^2} < 0, \quad \textit{which assures that the extrema is a } \max \textit{imum.}$$

$\hat{f}(x) = \sum_{i=1}^{j} \frac{n_i}{nd_{(i)}} I_{d_{(i)}}(x)$ *is the MLE of* f. End of proof.

Theorem 2. The step density function $\{\frac{1}{nd_{(1)}},\frac{1}{nd_{(2)}},...,\frac{1}{nd_{(n)}}\}$ *defined on any partition* $\{d_{(1)},d_{(2)},...,d_{(n)}\}$ *with constant probability density mass* $\frac{1}{n}$*, is a maximum likelihood estimator (MLE) of the unknown density function f.*

Proof. In Theorem 1 let $j=n$, and $n_i=1, i=1,2,...,n,$ Theorem 2 will immediately follow. The *sdf* is a special case of the generalized histogram for a specific choice of the partition, $\{d_{(1)},d_{(2)},...,d_{(j)}\}$. Since *sdf* is a generalized histogram, any general properties of the generalized histogram should also be valid for the *sdf*.

- **UMVUE**

Theorem 3. The generalized histogram with variable bin widths $\{d_{(1)},d_{(2)},...,d_{(j)}\}$ *is an unbiased estimator for f.*

Proof. First $\widehat{f}(x)=\sum_{i=1}^{j}\theta_i I_{d_{(i)}}(x), \underline{\theta}=(\theta_1,\theta_2,...,\theta_j)$, is an unbiased estimator *for f*.

In each bin of $d_{(i)}$, and frequency $\theta_i=\frac{n_i}{n}$ is an unbiased estimator of the density function at $d_{(i)}$,

$$E\theta_i\,|_{x\in d_{(i)}}=E(\frac{n_i}{n})\,|_{x\in d_{(i)}}=\frac{En_i\,|_{x\in d_{(i)}}}{n}=\frac{E(n\int_{x\in d_{(i)}}f(x)dx)}{n}=E(\int_{x\in d_{(i)}}f(x)dx)$$. End of proof.

Theorem 4. The step density function is a uniform minimum variance unbiased estimator (UMVUE) for f.

Proof. The sample, $\underline{X}$, is a minimum sufficient statistics (MSS) for *sdf* $\hat{f}(x) = \sum_{i=1}^{n} \theta_i I_{d_{(i)}}(x)$, $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_n)$, which can be shown as the following.

The joint pdf of sample $\underline{X}$ is $g(\underline{X} \mid \underline{\theta}) = \prod_{i=1}^{n} \hat{f}(x_i \mid \underline{\theta}) = \prod_{i=1}^{n} \theta_i$, and thus for two sample points $\underline{X}$ and $\underline{Y}$ the ratio $\frac{g(\underline{X} \mid \underline{\theta})}{g(\underline{Y} \mid \underline{\theta}')} = \frac{\prod_{i=1}^{n} \theta_i}{\prod_{i=1}^{n'} \theta_i'}$ would be a constant independent of $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_n)$ *iff* $\underline{X} = \underline{Y}$.

From Lehmann and Scheffe's theorem (Casella and Berger, 2001, p. 281), the *sdf* sample $\underline{X}$ is a minimum sufficient statistics (MSS) for *f*.

From Theorem 3, the *sdf* $\hat{f}(x) = \sum_{i=1}^{n} \theta_i I_{d_{(i)}}(x)$ is a special case of the generalized histogram and is an unbiased estimator for *f ;* thus it is conditioned on the minimum sufficient statistics (MSS) $\underline{X}$. From the theorem of Rao-Blackwell (Casella and Berger, 2001. p. 342), *sdf* $\hat{f}(x) = \sum_{i=1}^{n} \theta_i I_{d_{(i)}}(x)$ is a UMVUE of the density function *f*.

End of proof.

The *edf*, $\hat{F}_n$, is a sufficient statistic for the true distribution (B. Efron and R. Tibshirani, 1992). After searching the literature for such a proof, we failed to locate a documented proof for its sufficiency, MLE or UMVUE. So we outline a brief proof here. The major steps of the proof are the family of the estimation function $\hat{f}(x) = \sum_{i=1}^{j} \theta_i \delta(x - x_i)$, the likelihood function $L(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} \hat{f}(x_i) = \prod_{i=1}^{n} \theta_i$, and $\theta_1 + \theta_2 + ... + \theta_n = 1$. The last two equations constitute a classical optimization problem with the well-known results that $L(x_1, x_2, ..., x_n)$ was at maximum when $\theta_1 = \theta_2 = ... = \theta_n = \frac{1}{n}$. However the most direct argument may be as simple as that

the complete sample, $\underline{X}$, is always a sufficient statistics, so $\widehat{F_n}$ and any functional of $\underline{X}$ is also S.S. for *f*.

- **Ready for bootstrap**

The single most important property of the *sdf* is that it can be naturally implemented into bootstrap resampling. The real hard work in smoothed bootstrap, i.e. smoothing kernel function selection and suitable amount smoothing by choosing bandwidth *h*, are eliminated automatically by the *sdf*. However, the bootstrap sample is not a "resample" with the peculiar property such as multiple observation of the same sample point (B. Silverman, 1987), but as imputed new sample points from a continuous distribution *sdf*. Therefore bootstrap resampling from *sdf* might be better called "***imputed bootstrap resampling"*** instead of "smoothed bootstrap" to stress this basic change of imputed resampling technique with an automatic *sdf,* which is almost as simple as *edf*.

Here is the computation implementation of *sdf*. Let $d_{(i-1)} = x_{(i)} - x_{(i-1)}$, add one data point, $x_{(0)}$, for programming purpose only to remove the distinction between the end intervals, $\Delta_{(1)}, \Delta_{(n)}$ and middle intervals, $\{d_{(1)}, d_{(2)}, ..., d_{(n)}\}$.

$$x_{(0)} = x_{(1)} - \frac{1}{2}(x_{(2)} - x_{(1)}) = \frac{3}{2}x_{(2)} - \frac{1}{2}x_{(1)},$$

$$d_{(0)} = \frac{1}{2}(x_{(2)} - x_{(1)}), d_{(n)} = \frac{1}{2}(x_{(n)} - x_{(n-1)}).$$

Two random variable R_1 and R_2 are drawn from uniform distribution $U[0,1]$, where $k = [R_1 \bullet n]$ is the nearest integer $< R_1 \bullet n$, $k = 0,1,2,...,n-1$. When $k = 0$, 50% chance let $k = 0$ and 50% chance let $k = n$. The bootstrap sample is generated as following

$X_i^* = x_{(k)} + R_2 \bullet d_{(k)}$.

The imputed bootstrap resampling will be used routinely in examples of the following chapters.

- **Limitations of *sdf***

Before we went on to the applications of *sdf* we think it is time to discuss the fundamental limitations of the *sdf* and their consequences on imputed resampling.

The 1st limitation is more "philosophical" and related to the freedoms in selecting the set of bin width $\{d_{(1)}, d_{(2)}, ..., d_{(n)}\}$, which is mainly restricted by the partition requirement that each bin has one and only one sample value, $x_{(i)}$. Essentially we adopted the 1-nearest neighbor approach. It stemmed from the same principle in other kernel-based density estimators that each kernel is a single-mode bell or box shaped function that weighs more on the center, which was the observation and less on its neighboring observations. The advantage of the choice was that it would automatically adapt to the local property. For example if $d_{(i-1)} < d_{(i)}$, the density on the sides of $x_{(i)}$ would be unsymmetrical with $\frac{1}{nd_{(i-1)}}$ on $[x_{(i-1)}, x_{(i)}]$ and $\frac{1}{nd_{(i)}}$ on $[x_{(i)}, x_{(i+1)}]$, where $\frac{1}{nd_{(i-1)}} > \frac{1}{nd_{(i)}}$. The unique choice of *sdf* based on the distance to the 1-nearest neighbor has the advantage in simplifying the *h*–optimization at the expenses of a less smooth density estimate by completely ignoring the 2 or 3-nearest neighbor that may improve the estimation in a less influential manner than the 1-nearest neighbor. So it is not "getting something from nothing" - we obatin what we need by giving up less relevant details in the same way of *edf* for naïve bootstrap to gain on the side of objectivity and ease of use. Yet no justification above constitutes any proof of optimality of selecting *sdf*, a result of compromise between getting around the discreteness of *edf* in an efficient way, so other better density estimation might be necessary in special situations.

And the 2nd limitation is more paradoxical in nature of the extreme value statistics. It is intuitive to see that the few data points in the tail weigh more than the bulk of the remaining observations, so the first impression was the effect of *sdf* as a global density may not be effective or even relevant. The result in the next chapter may be surprisingly in its effectiveness. Since the discreteness in the few local points near the tails, so the careful choice of local adaptive property is crucial and needs more investigation from an order statistics frame work (H.A. David and H.N. Nagaraja, 2003).

Chapter 4. Application of step density estimation

The step density function was demonstrated in (1) the estimation of local density, (2) unique mode selection, and (3) quick estimation for further smoothing. Its most important application, namely, (4) imputed bootstrap resampling, was presented in Chapter 5.

4.1 Local density estimated by step density function

Local density becomes an important issue for statistical procedures that is determined mainly by the local property of a distribution, for example the uniform distribution parameter estimation problem we used to illustrate as one of the bootstrap failures (B. Efron, 1993). Intuitively we sense it is a hard problem because a statistic, such as the extreme values, that relies on the local density property, would involve very few data point as support from the original data and the rest of data points had little to do with the accuracy of the statistic.

Step density estimation seems to be a quick choice in this situation by adding more support, infinitely many indeed, by filling up the gaps between discrete sample points with a simple uniform distribution. With additional information more complicated model can be chosen; but a non-informative choice of a uniform distribution might be less biased like a non-informative Bayesian prior.

Let's restate the well-known example that B. Efron had used to illustrate situations when bootstrap failed,

$$F \xrightarrow{i.i.d.} (x_1, x_2, ..., x_n),\ F \sim U(0, \theta).$$

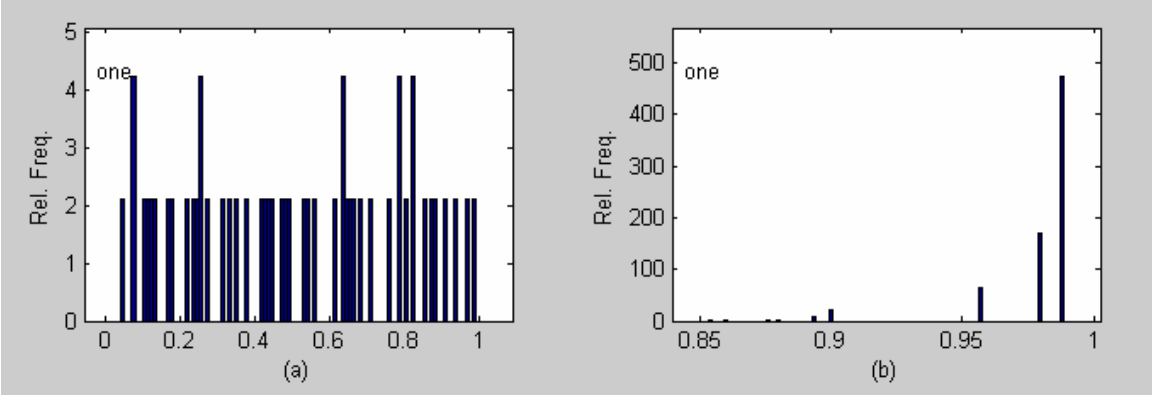


Figure 4.1 The histogram with 100 bins that imitate the needle plot to show the original 50 data points ~ $U(0,1)$ on (a), and the 2000 nonparametric bootstrap replications obtained sampling with replacement from the empirical distribution function, $\widehat{F_n}$.

The MLE $\hat{\theta}$ is the largest sample value $x_{(n)}$. A sample of 50 uniform numbers in the range $(0,1)$ is generated and computed resulting in $\hat{\theta} = 0.988$. The left panel of Figure 4.1 shows a histogram of 50 sample points, and the right panel of 2000 bootstrap replications obtained sampling with replacement from the data. The left panel of Figure 4.2 shows 2000 parametric bootstrap replications obtained by sampling from the uniform distribution on $U(0,\hat{\theta})$. It is evident that the right histogram of Figure 4.2 is an approximation to the left histogram. In particular, the left histogram has a large probability mass at $0.62 \times \hat{\theta}$ of the value $\theta^* = \hat{\theta}$. In general, it is easy to show that

$$\mathrm{P}(\theta^* = \hat{\theta}) = 1 - \mathrm{P}(\theta^* \neq \hat{\theta}) = 1 - (1 - \frac{1}{n})^n \xrightarrow{n \to \infty} 1 - \frac{1}{e} \approx 0.632 .$$

However, in the parametric setting of the right panel, $\mathrm{P}(\theta^* = \hat{\theta}) = 0$.

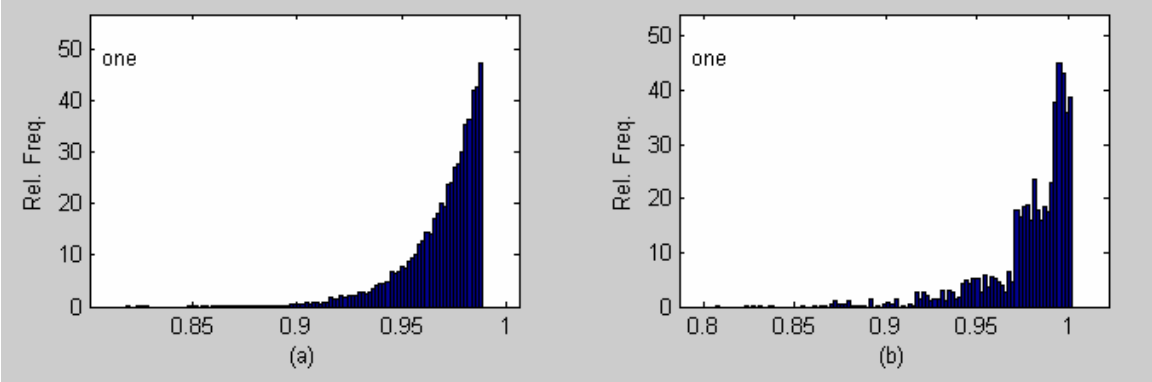


Figure 4.2 The histogram with 100 bins that imitate the needle plot to show the 2000 parametric bootstrap replications obtained sampling with $U(0,\hat{\theta})$ in (a) and 2000 nonparametric bootstrap replacement from the step density function, $\overline{F_n}$ in (b).

The improvement from the step density estimate, $\overline{F_n}$, over that from the empirical distribution function, $\widehat{F_n}$, is apparent that the previous totally discrete sample distribution of $\widehat{\theta^*}$ becomes continuous more resembling that from the parametric bootstrap. This made the accuracy measures such as $\widehat{C.I.(\theta^*)}$ and $\widehat{se(\theta^*)}$ possible from Figure 4.2 (b) and reasonably reliable even in comparison to that in Figure 4.2 (a).

The vaguely observable steps in Figure 4.2 (b) overlap reasonably with the locations of the first two order statistics $x_{(1)}, x_{(2)}$ and $x_{(3)}$ in Figure 4.1 (b), which is a good indication of how the step density estimate works. This non-smoothness character can be less observable when sample size gets larger, which left with narrower gaps between any two adjacent points in the original data.

4.2 Unique mode selection

Table 4.1 shows the results of a small experiment from B. Efron (1992), in which 7 out of 16 mice were randomly selected to receive a new medical treatment, while the remaining 9 were assigned to the non-treatment (control) group. The treatment was

intended to prolong survival after a test surgery. The table shows the survival time following surgery, in days, for all 16 mice.

Table 4.1 The mouse data. Sixteen mice were randomly assigned to a treatment group or control group. Shown are their survival times, in days, following a test surgery. Did the treatment prolong survival?

Group	Data							Sample size	$\bar{X}$	$\widehat{\sigma_X}$
Control	52	104	146	10	51	30	40	9	86.86	25.24
	27	46								
Treatment	94	197	16	38	99	141	23	7	56.22	14.14
								Difference	30.63	28.93

This set of data has been intensively analyzed by B. Efron using different methods. The standard error for the difference $(\bar{X}-\bar{Y})$ equals $28.93=\sqrt{25.24^2+14.14^2}$ (since the variance of the difference of two independent quantities is the sum of their variances). We see that the observed difference $(\bar{X}-\bar{Y})$=30.63 is only

$$\frac{\bar{X}-\bar{Y}}{\widehat{\sigma_{\bar{X}-\bar{Y}}}}=\frac{30.63}{28.93}=1.05$$

estimated standard errors greater than zeros, which yields one-side p-value ≈ 0.15 by assuming two samples were both from normal distributions. It was an insignificant result, one that could easily arise by chance even if the treatment really had no effect at all. Several nonparametric permutation and bootstrap analysis basically yield the similar insignificant results.

Because the sample size is small $(n_1, n_2)=(9,\ 7)$, it was not reliable to test the normality of the two samples. For the same reason of sample size histogram shown in

Figure 4.3 was not as intuitive as it should be for large sample to observe any modes from its sample distribution.

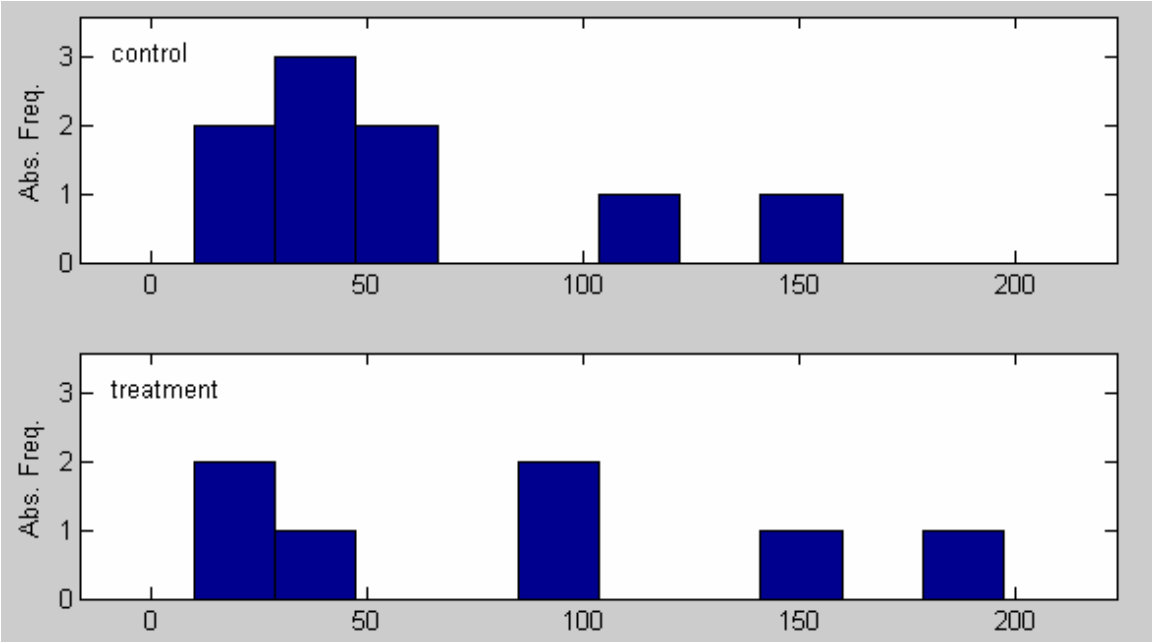


Figure 4.3 Histogram of the mouse data: (top) control group, (bottom) treatment.

The needle plot shown in Figure 4.4 indicated that within the range [10, 60] it seems to have some clustered observations especially for the control group. However it would be much intuitive to have an estimated density instead of the clustering frequency.

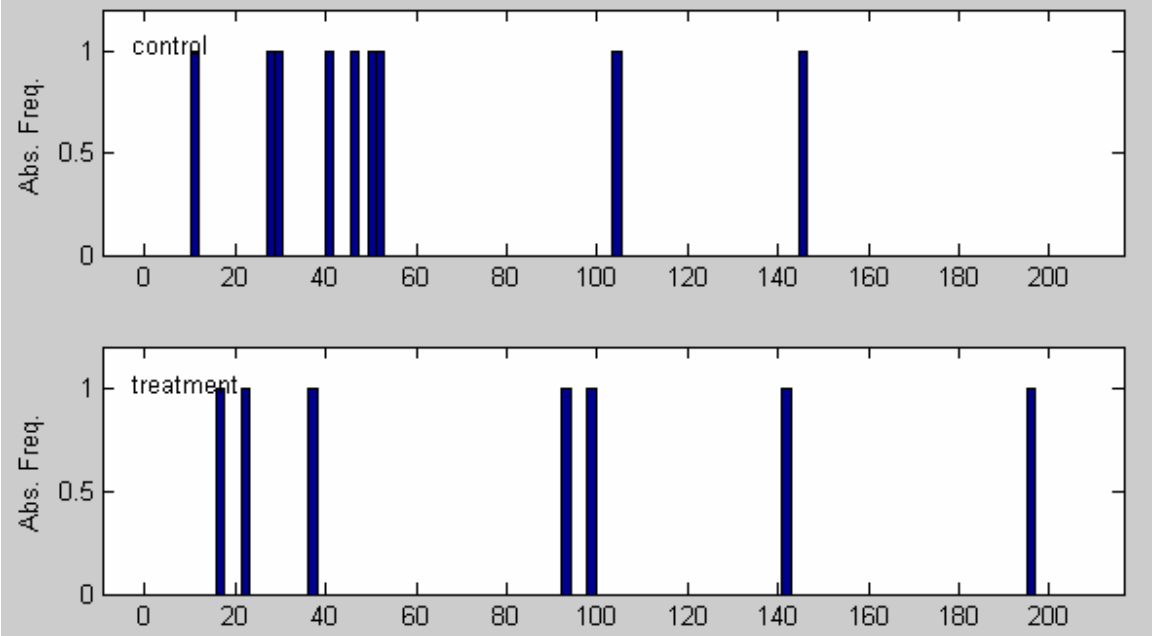


Figure 4.4 Needle plot for the mouse data: (top) control group, (bottom) treatment.

Step density estimate could be the first choice for this type of data display to lend the user a quick look because it is both unique in presentation and adaptive to local density variations. The other density estimates would either yield different estimates with window width selections, or be not adaptive to large local density variations that are usually severe for small samples. Those that can do both are usually rather cumbersome even with software. Figure 4.5 and Figure 4.6 are the same step density estimate displayed in linear as well as logarithmic scales for easy reading.

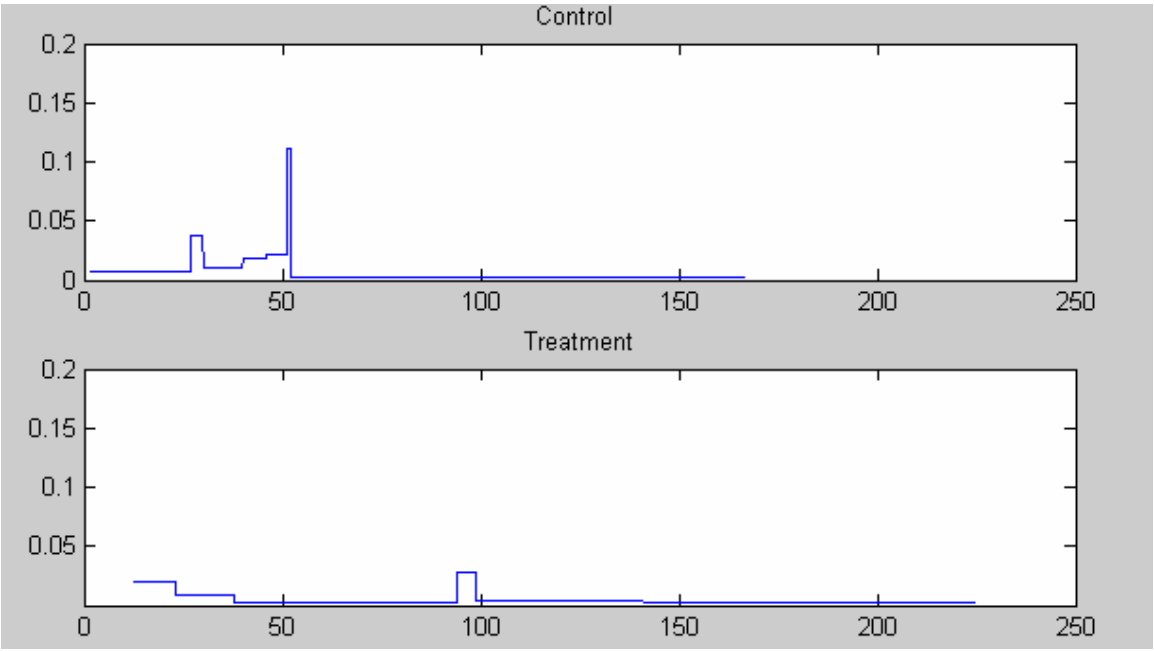


Figure 4.5 The unique step density estimate of the mouse data in linear scale: (top) control group, (bottom) treatment.

We observed that the step density estimation for the control had a mode in the range [0, 60] with a long tail in the range [60,180], and for the treatment group the mode in the range [0, 60] is significantly reduced with a new mode surging in the range [60,240] with a longer tail extended to the longer survival time. Since the data points were so scarce we combined the control and the treatment groups to see more clearly the two modes in survival time as shown in Figure 4.7, which confirmed our speculation that there might be two modes in the data structure, one concentrated mode in the range of [0, 60] and one widely-spreading mode in the range of [80,240]. The clear gap between the two modes may serve as the criteria to distinguish the mixture of the two modes.

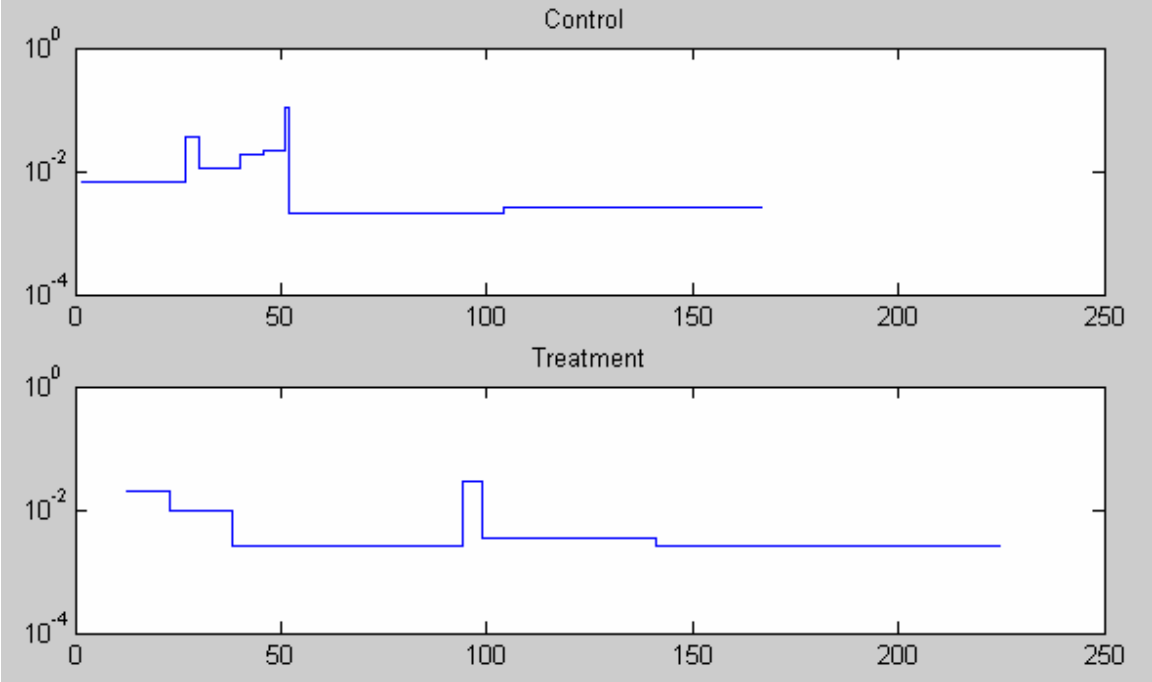


Figure 4.6 The unique step density estimate of the mouse data in logarithmic scale: (top) control group, (bottom) treatment.

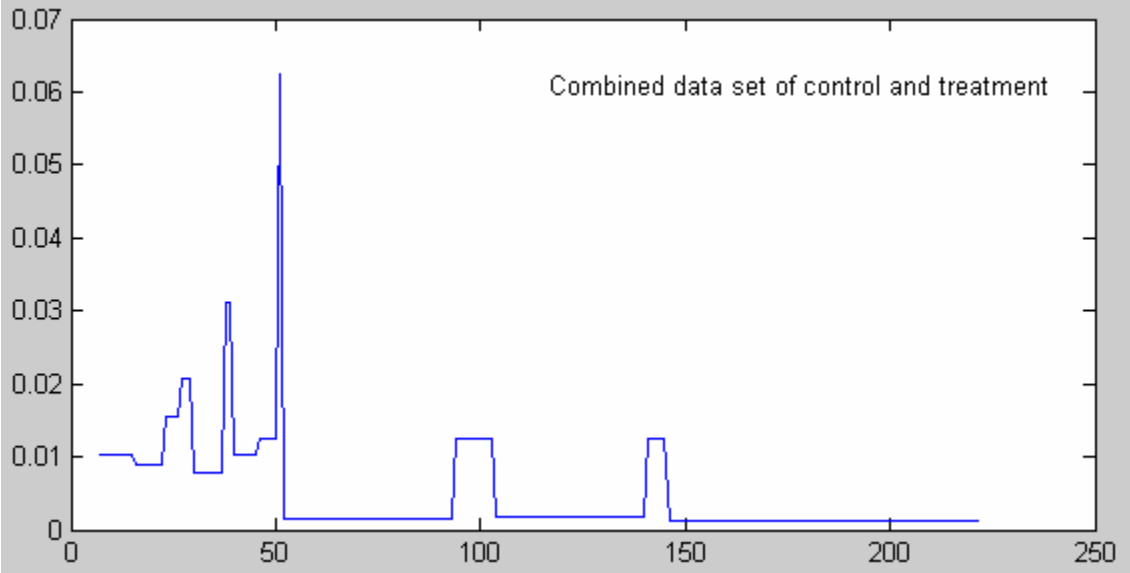


Figure 4.7 Step density estimate of the combined mouse data in linear scale.

Let's consider a simple binomial model. The proportion of the longer survival mode is denoted as *p,* the control group $\sim Bin(p_1, n_1)$ and the treatment group $\sim Bin(p_2, n_2)$, we will test the hypothesis

$$H_0 : p_1 - p_2 \geq 0 \text{ vs. } H_1 : p_1 - p_2 < 0 \ .$$

If we take *c = 80* by visual inspection of the step density function as the threshold to distinguish the two modes, $\widehat{p_1} = 2/9 = 0.222$, $\widehat{p_2} = 4/7 = 0.57$, and the test statistics $z = \dfrac{\widehat{p_1} - \widehat{p_2} - (p_1 - p_2)}{\sqrt{\dfrac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \dfrac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}} \doteq \dfrac{0.35}{0.23} \doteq 1.51$ and p-value 0.067, which is marginal to say the treatment have significant effects to prolong the survival time after surgery.

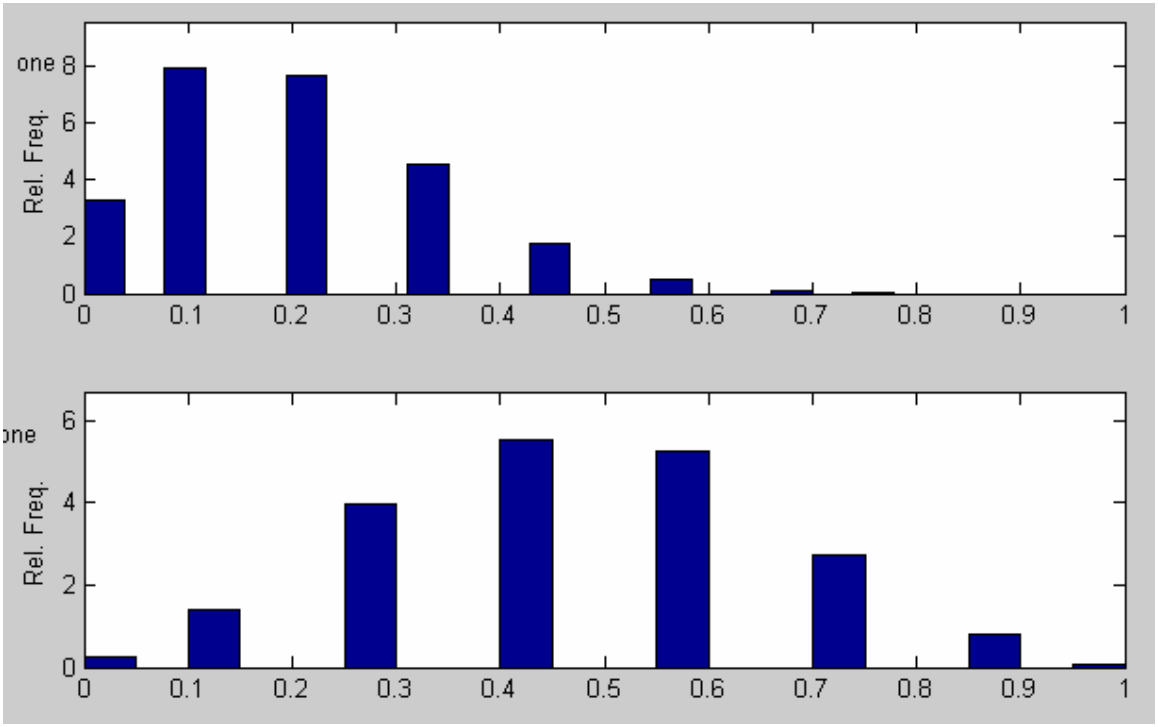


Figure 4.8 Histogram of the proportion of mice in each group with survival time > 80 days from bootstrap resampling with step density function, *B=10000*. The top panel is for the control group, and the bottom panel for the treatment group.

To confirm the stability of the test we run a bootstrap resampling for the two groups separately to estimate the proportions, the bootstrap sample size we used was B =10,000 . From the resampling data presented in Figure 4.8, $(\overline{p_1^*}, \overline{\sigma_1^*}) = (0.2019, 0.1338)$, $(\overline{p_2^*}, \overline{\sigma_2^*}) = (0.4707, 0.1881)$, and the variance of $(\widehat{p_1} - \widehat{p_2})$,

$$\widehat{\sigma_{\widehat{p_1}-\widehat{p_2}}} = \sqrt{(\overline{\sigma_1^*})^2 + (\overline{\sigma_2^*})^2} = \sqrt{0.1338^2 + 0.1818^2} = 0.231,$$

which is much closer to the simple binomial estimate of 0.23 than the traditional bootstrap estimate of S.E.($(\widehat{p_1^*} - \widehat{p_2^*})$)=0.28.

4.3 First-order approximation for further smoothing

As we saw from the review on density estimation methods in Chapter 2 the most important application of the smoothing procedure is to the empirical density (*edf*) estimator $\widehat{F_n}$. As we can imagine gaps between the isolated points need to be filled up by properly selecting the window width of the smoothing kernels, preferably with some local adaptive ability. The step density estimate (*sdf*) we constructed, $\overline{F_n}$, has been shown to be unique like the empirical density estimate, $\widehat{F_n}$, and piece-wise continuous without the gaps between isolated data points. However the *sdf* we constructed was discrete at each joint between steps, which was not differentiable. For applications where smoothness of a density function becomes necessary, we can always start the smoothing procedure from *sdf*, $\overline{F_n}$, instead of *edf* $\widehat{F_n}$. This would carry all the local adaptive property over, and greatly reduce the effort of making the kernel function to be local adaptive. The lone item remains is to select a constant smoothing window width. This would be the major potential application of *sdf* as a first-order density approximation for further smoothing.

Chapter 5. Bootstrap small sample bias

G.A. Young and H.E. Daniels (1990) studied the bias of nonparametric bootstrap estimation of a simple pivot for small sample sizes by a very simple situation, namely $F \sim N(0,1)$. It is of practical importance for statisticians in the filed, who often find bootstrap an attractive method especially in small sample situations.

$$F \xrightarrow{i.i.d.} (x_1, x_2, ..., x_n) \qquad (5.1)$$

It is required to study the distribution of the random variable, $T(X_1,...,X_m;F)$ possibly dependent on the distribution F, where $X_1,...,X_m$ is a random sample from F. The bootstrap method approximates the sampling distribution of $T(X_1,...,X_m;F)$ under F by that of $T(Y_1,...,Y_m;\widehat{F}_n)$ under $\widehat{F}_n$, where *n* is the original sample size and *m* is the bootstrap sample size. The distribution of $T(X_1,...,X_m;F)$ is denoted as

$$P(a) = pr\{T(X_1,...,X_m;F) > a \mid F\}, \qquad (5.2)$$

which was estimated by the bootstrap distribution,

$$\widetilde{P}(a) = pr\{T(Y_1,...,Y_m;\widehat{F}_n) > a \mid \widehat{F}_n\}. \qquad (5.3)$$

The bias is defined as the difference between true distribution (5.2) and bootstrap sampling distribution (5.3). The estimated density has the full information about a statistic and offers a comprehensive picture of the bias behavior more than any moment estimation. The techniques of computer algebra are used to obtain an exact analytical

assessment of the bias from a bootstrap procedure. The expectation of $\widetilde{P}(a) = pr\{T(Y_1,...,Y_m;\widehat{F}_n) > a \mid \widehat{F}_n\}$ by a simulation study agrees with the exact analytical results. Being able to determine the bias by two totally independent methods had been a unique advantage of this study, which cast mistrust on how to apply bootstrap properly with such a noticeable bias on the sampling distribution for the simplest summary statistic, the sample mean. It was difficult to extend the exact method in the study further to other statistic to verify if such bootstrap bias observed was an exceptional pathological case or a general limitation of bootstrap (G.A. Young, 1994).

However, the simulation strategy of Young and Daniels' could be applied to other statistics without the exact analytical confirmation step, which was mathematically less vigorous but application-wise more flexible. Through simulation within the framework of Young and Daniels we found that the bootstrap-t was much less biased than the bootstrap-z. Perhaps more importantly, our study illustrated the first-order mechanism of the bootstrap bias. In small sample cases bootstrap bias may not be an issue by judicially selecting proper statistics. Furthermore, the step density function can be used as a second-order means of imputation that will further reduce the bootstrap bias for small sample cases.

5.1 Bootstrap bias for sample mean

The bias is defined as the difference between bootstrap distribution $\widetilde{P}(a) = pr\{T(Y_1,...,Y_m;\widehat{F}_n) > a \mid \widehat{F}_n\}$, where $T(Y_1,...,Y_m;\widehat{F}_n) = \overline{Y_m}$, the estimated bootstrap sample mean, and $P(a) = pr\{T(X_1,...,X_m;F) > a \mid F\}$, the true distribution. In principle, $\widetilde{P}(a)$ should be constructed by considering all n^m possible bootstrap samples. In practice, $\widetilde{P}(a)$ is estimated by drawing a large number of bootstrap samples from $\widehat{F}_n$. The bootstrap sample size used here was, *B*, which equals *50000*. The sampling procedure was repeated over different $\widehat{F}_n$ to estimate $E(\widetilde{P}(a))$ via a Monte Carlo simulation with 1000 repetitions. Young's results were summarized in Table 5.1 and

Table 5.2 for two combinations of data and bootstrap sample sizes: *(m, n) = (5,10)* and (20,20), where *P(a)* is the expected theoretical value computed from *N(0, 1/5)*, and the simulation results were listed as $E(\widetilde{P}(a))$ (sim.Young). The simulation had been performed on a HP9000/330 UNIX workstation, which was considered a fast computer back in 1990 when the original work was done.

Table 5.1 Simulation and theoretical expectations, normal distribution.

m,n	*a*	*P(a)*	$E(\widetilde{P}(a))$ $(exact)$	$E(\widetilde{P}(a))$ $(sim.Young)$	$E(\widetilde{P}(a))$ $(dup.Ma)$	*Est.Error* $(Young)$
5,10	0.1	0.41153	0.40138	0.40066	0.4006	0.00103
	0.3	0.25116	0.22901	0.22807	0.2273	0.00186
	0.5	0.13177	0.11220	0.11246	0.1106	0.00174
	0.7	0.05876	0.04851	0.04928	0.0478	0.00120
	0.9	0.02208	0.01909	0.01937	0.0189	0.00070
	1.1	0.00695	0.00702	0.00735	0.0071	0.00036
	1.3	0.00182	0.00247	0.00266	0.0025	0.00018
	1.5	0.00040	0.00085	0.00094	0.0009	0.00009
	1.7	0.00007	0.00029	0.00033	0.0003	0.00004
Max(Δ)				0.0231	0.0239	0.00186
mean(Δ)				0.0075	0.0080	0.00080
std(Δ)				0.0088	0.0093	0.00070

We define $\Delta = |E(\widetilde{P}(a)) - P(a)|$, an absolute distance measure between the bootstrap estimation and the true distribution.Young used the techniques of computer algebra to obtain exact assessment of $E(\widetilde{P}(a))$ that was listed in the table as $E(\widetilde{P}(a))$ (exact), which may be served as a validation of the simulation results in the study. The estimated error was used by Young to indicate the bias is significant relative to the estimated standard error.

We had duplicated Young's simulation results and listed as $E(\widetilde{P}(a))$ *(dup.Ma)* in Table 5.1, which indicated that within the simulation error our simulation results were in good agreement to Young's original results. Therefore we may infer that we have implemented the equivalent simulation procedure, and we may apply the simulation procedure to evaluate other pivotal in studying of other statistics.

Table 5.2 Simulation and theoretical expectations, normal distribution.

m,n	*a*	*P(a)*	$E(\widetilde{P}(a))$ $(exact)$	$E(\widetilde{P}(a))$ $(sim.Young)$	$E(\widetilde{P}(a))$ $(dup.Ma)$	*Est.Error* $(Young)$
20,20	0.1	0.41153	0.40138	0.31806	0.3176	0.00094
	0.2	0.25116	0.22901	0.17403	0.1735	0.00124
	0.3	0.13177	0.11220	0.08226	0.082	0.00103
	0.4	0.05876	0.04851	0.03406	0.0342	0.00065
	0.5	0.02208	0.01909	0.01254	0.0128	0.00034
	0.6	0.00695	0.00702	0.00417	0.0044	0.00016
	0.7	0.00182	0.00247	0.00129	0.0014	0.00007
	0.8	0.00040	0.00085	0.00037	0.0004	0.00002
	0.9	0.00007	0.00029	0.0001	0.0001	0.00001
Max(Δ)				0.0115	0.0120	0.00124
mean(Δ)				0.0036	0.0038	0.00050
std(Δ)				0.0046	0.0048	0.00048

One of the conclusions from the original study is that favorable asymptotic property is no guarantee for good small sample performance. This conclusion had been fully supported by the simple and yet convincing work of Young and thus casting a long shadow in the application of bootstrap resampling methods in small sample size situations, where nonparametric methods such as bootstrap resampling are most needed [Young, 1994 review]. It also played down the importance of the theoretical asymptotic work, which had dominated publications in bootstrap theory [J. Shao].

5.2 Two-level structure of the bootstrap bias

As we discussed in Section 5.1, Young and Daniels had demonstrated the bias in bootstrap method in small sample sizes by a simple and yet convincing example. However neither their analytical calculation nor their simulation study can explain the mechanism that caused the bias. Therefore the structure and nature of the bootstrap bias had been a mystery, which we try to unravel as the following.

We found in the schematic diagram in Figure 5.1 the bootstrap resampling was represented in two levels. Traditional statistical inference is done with the original sample on the 1st level in the "real world", while resampling methods is applied to the duplicated samples on the second level created by the original sample in the "bootstrap world". The "bootstrap world" appears to possess all the information of the "real world", and intuitively the "bootstrap world" would asymptotically approach the "real world" when sample size $n \to \infty$.

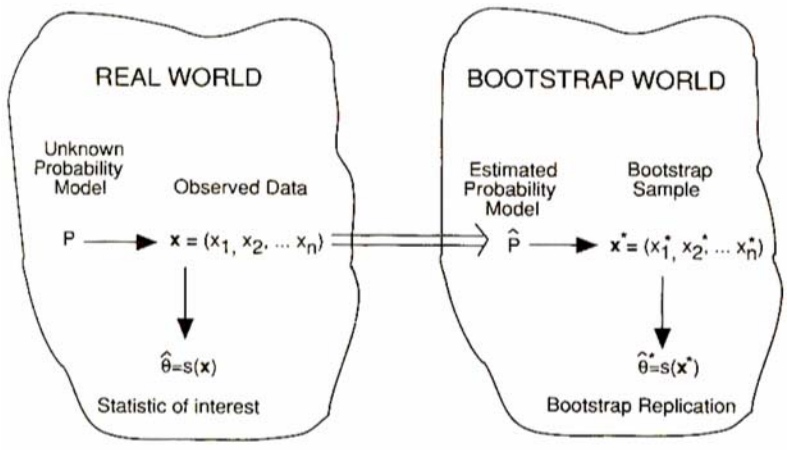


Figure 5.1 Schematic diagram of the bootstrap applied to problems with a general data structure $P \to \mathrm{x}$. The crucial step "$\Rightarrow$" produces an estimate $\hat{P}$ of the entire probability mechanism P from the observed data x. The rest of the bootstrap picture is determined by the real world: "$\hat{P} \to \mathrm{x}^*$" is the same as "$P \to \mathrm{x}$"; the mapping from $\mathrm{x}^* \to \hat{\theta}^*, s(\mathrm{x}^*)$, is the same as the mapping from $\mathrm{x} \to \hat{\theta}, s(\mathrm{x})$.

However our problem is associated with the small sample size, and our solution is to apply the statistical information between the "bootstrap world" and the "real world" to correct the well known bias, such as that between the population variance, σ^2, and the sample variance, $\sigma^2_{\overline{Y}_m}$.

From the moment generating function of $\widehat{F}_n$, $M_Y(t \mid x_1, x_2, ..., x_n) = \frac{1}{n}\sum_{i=1}^{n} e^{x_i t}$ and

$$M_{\overline{Y}_m}(t \mid x_1, x_2, ..., x_n) = [M_Y(\frac{t}{m} \mid x_1, x_2, ..., x_n)]^m,$$

we may derive the following bootstrap expectations,

$$\overline{Y_m} = EX = \mu_X,\ \sigma^2_{\overline{Y}_m} = \frac{\sigma^2_X}{m},$$

noting that $X \sim (\mu_X, \sigma^2_X)$, which is not necessarily from normal $N(\mu_X, \sigma^2_X)$. Apparently the bootstrap mean $\overline{Y_m}$ is unbiased but not the bootstrap sample variance, which is biased from the well-known result $\widehat{\sigma_{\overline{X}_m}} = \frac{\sigma^2_X}{m-1}$. The first major term of bias in a distribution function in terms of the moments is the variance, or the 2nd moment; therefore our first step towards bias correction is using the bootstrap-t statistic as described next.

- **Bootstrap level**

There are two levels in Young's bootstrap simulation, the "real world" and the "bootstrap world", and correspondingly there are two ways of implementing the bootstrap-t statistics. We may study the statistic $\overline{Y_m} - \overline{X_n}$ through the bootstrap-t at the bootstrap level

$$\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{Y_m}}.$$

Table 5.3 Comparison of three-simulation results, normal distribution.

m, n	*a*	*P(a)*	$E(\widetilde{P}(a))$ $\left(\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{X_n}}\right)$	*t*(a)* *Scaled t-table*	$E(\widetilde{P}(a))$ $\left(\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{Y_m}}\right)$	$E(\widetilde{P}(a))$ $(\overline{Y_m} - \overline{X_n})$	*Est.Error* (*Young*)
5,10	0.1	0.41153	0.40066	0.4163	0.4149	0.4006	0.00103
	0.3	0.25116	0.22807	0.2695	0.2676	0.2273	0.00186
	0.5	0.13177	0.11246	0.1628	0.1659	0.1106	0.00174
	0.7	0.05876	0.04928	0.0966	0.104	0.0478	0.0012
	0.9	0.02208	0.01937	0.0575	0.0678	0.0189	0.0007
	1.1	0.00695	0.00735	0.0352	0.0463	0.0071	0.00036
	1.3	0.00182	0.00266	0.0221	0.0331	0.0025	0.00018
	1.5	0.00040	0.00094	0.0144	0.0246	0.0009	0.00009
	1.7	0.00007	0.00033	0.0096	0.0188	0.0003	0.00004
Max(Δ)			0.01110(Δ) 43.0%		0.01110(Δ') 46.5%	0.0231	0.00186
mean(Δ)			0.00729(Δ) 55.3%		0.00729(Δ') 91.5%	0.00796	0.00080
std(Δ)			0.00404(Δ) 43.4%		0.00404(Δ') 43.3%	0.00934	0.00070

- **Sample level**

Alternatively we may study the statistic $\overline{Y_m} - \overline{X_n}$ through the bootstrap-t at the sample level

$$\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{X_n}}.$$

The simulations were performed under the same conditions as B=5000, Monte Carlo simulation at 1000 repetitions, except two bootstrap-t statistics were used and the results were summarized in the following two tables 5.3 and 5.4.

Table 5.4 Comparison of three-simulation results, normal distribution.

m, n	*a*	*P(a)*	$E(\widetilde{P}(a))$ $\left(\frac{\overline{Y_m}-\overline{X_n}}{\widehat{\sigma_{X_n}}}\right)$	*t*(a)* *Scaled t-table*	$E(\widetilde{P}(a))$ $\left(\frac{\overline{Y_m}-\overline{X_n}}{\widehat{\sigma_{Y_m}}}\right)$	$E(\widetilde{P}(a))$ $(\overline{Y_m}-\overline{X_n})$	*Est.Error* (*Young*)
20,20	0.1	0.32736	0.3238	0.3304	0.3292	0.3176	0.00094
	0.2	0.18555	0.1799	0.1909	0.1907	0.1735	0.00124
	0.3	0.08986	0.0845	0.0984	0.0983	0.082	0.00103
	0.4	0.03682	0.033	0.0451	0.0462	0.0342	0.00065
	0.5	0.01267	0.0107	0.0186	0.0204	0.0128	0.00034
	0.6	0.00365	0.0029	0.0074	0.0088	0.0044	0.00016
	0.7	0.00087	0.0006	0.0028	0.0038	0.0014	0.00007
	0.8	0.00017	0.0001	0.0011	0.0016	0.0004	0.00002
	0.9	0.00003	0	0.0004	0.0007	0.0001	0.00001
Max(Δ)			0.005647(Δ) 46.9%		0.0018(Δ') 14.9%	0.012	0.00124
mean(Δ)			0.002386(Δ) 63.2%		0.0008(Δ') 22.4%	0.0038	0.00050
std(Δ)			0.002267(Δ) 47.6%		0.0006(Δ') 12.5%	0.00048	0.0048

The scaled t-table, t*(a), in both tables were transformed from the standard t-table to the normal-distribution scale we used in the remaining of the simulation. We denote $\Delta = |E(\widetilde{P}(a)) - P(a)|$, when $P(a)$ follow normal distribution; $\Delta' = |E(\widetilde{P}(a)) - P(a)| = |E(\widetilde{P}(a)) - t^*(a)|$ when $P(a)$ follow t-distribution.

The bootstrap-t can be regarded as the variance-corrected bootstrap-mean, $\overline{Y_m} - \overline{X_n}$. Just as we expected from the previous analysis on the variance bias in the two-level structure, the bootstrap-t statistic showed about 50% reduction in its distribution for $(m,n) = (5,10)$, and $\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{Y_m}}$ performed even better by reducing the bias by about 80% in the larger sample case of $(m,n) = (20,20)$. From a distance measure of simulation distribution deviation, $(\text{Max}(\Delta),\text{mean}(\Delta),\text{std}(\Delta)) = (0.0018,\ 0.0008,\ 0.0006)$, the bias is marginally observable in comparison to intrinsic statistical error estimated by Young, which was $(0.00124,\ 0.00050,\ 0.0048)$; therefore further improvement might be masked by simulation errors and we would use the $(m,n) = (5,10)$ case to study further bias reduction. The trend also agrees with the bias behavior of sample variance that when *m* gets larger, we will have $\left(\frac{\sigma_X^2}{m-1}\right) \to \left(\frac{\sigma_X^2}{m}\right)$.

5.3 Imputed bootstrap resampling

The step density function may be readily used as an imputation technique to enlarge the bootstrap original sample size. It seems a natural idea to enlarge the original sample size to combat problems caused by the small sample size, the equivalence to inserting more points to remedy discreteness. But to our best knowledge from intensive literature review, there is no precedence to such effort for improving bootstrap performance in small sample scenarios. The lack of practical density estimation method may explain this phenomenon because a density function is needed in order to generate such imputed, extra data points from the original sample. With the step density estimation the task becomes easy since any bootstrap duplicated sample point may be used as the imputed data points.

We define an integer, *I*, the imputation factor indicating the imputed sample with size n', where $n' = n \bullet I$, because fractional imputation factor seems to be of little practical importance. We have applied a uniformly imputed bootstrap resampling scheme that

would force the imputed data point to be equally inserted into each possible density step, i.e. *(I-1)* imputed new data points added in each step, to guarantee the newly imputed data set with minimum deviation from the original sample. Within each density step the *(I-1)* imputed new data points were drawn randomly form a uniform distribution. Then the (m,n) resampling scheme will be executed as a (m,n') resampling scheme, where $n' = n \bullet I$.

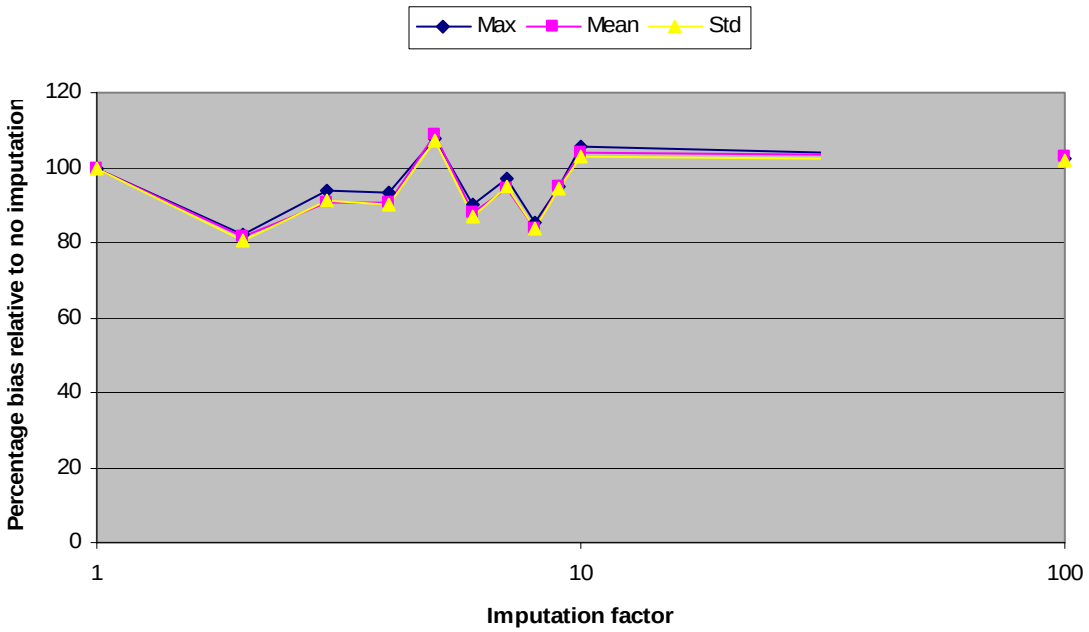


Figure 5.2 The uniformly imputed bootstrap resampling scheme had been used for $(\overline{Y_m} - \overline{X_n})$ with bootstrap sample size *B=5000* and Monte Carlo simulation repetitions 1000, and the performance were reported at imputation factor *I = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,* and*100*.

The imputed bootstrap was applied to the $(m,n) = (5,10)$ case with different imputation factors from 1 to100. The pivotal statistic, $(\overline{Y_m} - \overline{X_n})$, was used and the results summarized in terms of (Max(Δ),mean(Δ),std(Δ)) were shown in Figure 5.2. It appears that the uniformly imputed bootstrap resampling scheme did not change the bias in any consistent pattern at different imputation factors from 1 to 100 for the 2nd-order biased

statistic $(\overline{Y_m} - \overline{X_n})$ in small sample situation. A large range of imputation schemes can be further studied; and we would first investigate the performance of the uniformly imputed bootstrap resampling scheme for the already partially bias-corrected statistics, $\left(\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{X_n}}\right)$ and $\left(\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{Y_m}}\right)$ next.

- **Bootstrap level: uniform bias reduction**

When the same uniformly imputed bootstrap resampling scheme was applied to the $(m,n) = (5,10)$ case with different imputation factors from 1 to 100, the pivotal statistic with bias-corrected at the bootstrap level, $\left(\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{\overline{X}_n}}\right)$, was studied and the results were summarized in terms of (Max(Δ),mean(Δ),std(Δ)) in Figure 5.3.

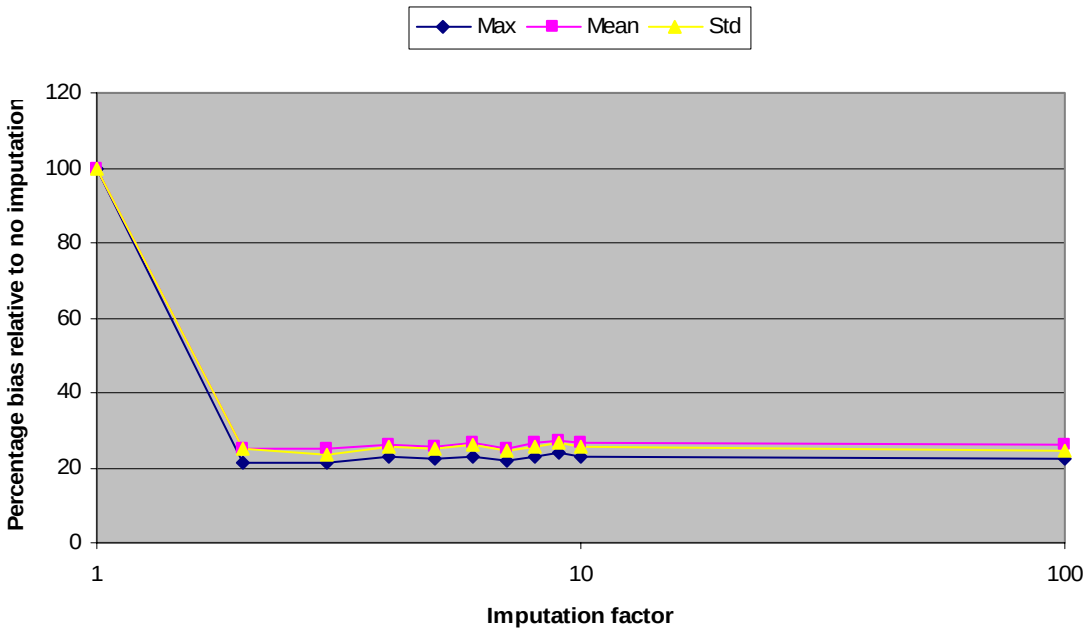


Figure 5.3 The uniformly imputed bootstrap resampling scheme had been used

for $\left(\frac{\overline{Y_m}-\overline{X_n}}{\widehat{\sigma_{X_n}}}\right)$ with bootstrap sample size *B=5000* and Monte Carlo simulation repetitions 1000, and the performance were reported at imputation factor *I = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,* and*100*.

The apparently uniform bias reduction for imputation factors starting immediately from 2 is surprising and we think it needs further confirmation from different aspects; and if in deed confirmed we would also need a through study to understand the probability mechanism behind such peculiar effect. However from the parallel but independent simulation study on $\left(\frac{\overline{Y_m}-\overline{X_n}}{\widehat{\sigma_{X_n}}}\right)$, the qualitatively similar surprisingly good results led us to future work on the full mechanism of the imputed bootstrap resampling in small sample sizes.

- **Sample level: optimal bias reduction**

The same uniformly imputed bootstrap resampling scheme was also applied to the $(m,n)=(5,10)$ case to the pivotal statistic bias-corrected at the bootstrap level, $\left(\frac{\overline{Y_m}-\overline{X_n}}{\widehat{\sigma_{Y_m}}}\right)$, with different imputation factors from 1 to100. The results were surprisingly positive and effective: first of all, even with the lowest imputation factor, 2, the bias had be reduced to 20% of the level without any correction, and at its optimal performance at the imputation factor ~ 4 and 5 the bias has been reduced to 12% of the un-corrected level, essentially within the statistical error.

From what is summarized in terms of (Max(Δ),mean(Δ),std(Δ)) in Figure 5.4 we may easily see that the curve was quite similar to that in Figure 5.3 with the bias reduction reaching saturation at around the 20% reduction level, which mean too large an imputed sample is not of much help. However from the sample size behavior of the Student-t statistic, we know that at sample size of 5 (i.e. degree of freedom of 4) the t-

distribution deviates significantly from the normal distribution; yet if the sample size is increased by a factor of two to 10 (i.e. degree of freedom of 9), the amount of deviation of the t-distribution from the normal distribution would be greatly reduced; and finally at the sample size of 20~30, the difference between the two distribution becomes negligible. We believe this explains what we have observed with the imputation effect in the last two sections.

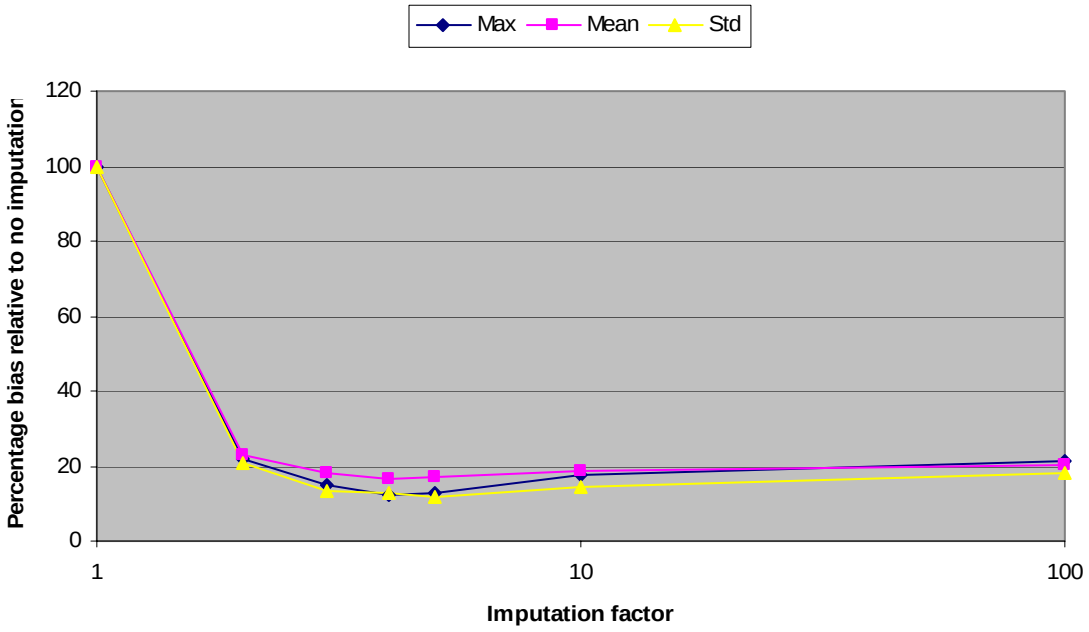


Figure 5.4 The uniformly imputed bootstrap resampling scheme had been used for $\left(\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma}_{Y_m}}\right)$ with bootstrap sample size *B=5000* and Monte Carlo simulation repetitions 1000, and the performance were reported at imputation factor *I = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,* and*100*.

The strength of studying bootstrap bias by Monte Carlo simulation is a proper framework for the problem. Density is at the heart of many statistical problems, especially for nonparametric bootstrap resampling, while other measures may be viewed as different aspects of the density. Density is mathematically less tractable than a measure of it, such as mean, variance or higher order moments. Yet a Monte Carlo simulation,

which is a tally of outcome summarized into histogram or frequency, generates an unbiased estimate of the interested density.

5.4 Mechanism of bootstrap bias

The error in approximating the sample distribution of $\overline{X_n} - \mu$ by that of $\overline{Y_n} - \bar{x}_n$ is $O(n^{-\frac{1}{2}})$ (J. Hartigan, 1986), and Table 5.2 shows that this bias is appreciable even for n=20 in the case of a normal population. Further investigation on the bootstrap bias and its relationship to a finite support yields to an exact theoretical distribution, which was a linear combination of standard densities, such as normal in this case (G.A. Young, 1990).

The two alternative statistics we studied, $\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma_{Y_m}}}$ and $\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma_{X_n}}}$, via simulations reduced about 80% of the distribution bias. It indicated that t-pivotal is a better statistics for small sample bootstrapping. G.A. Young take the normal distribution as the gold standard, where the population variance is a known quantity.

$$
\begin{aligned}
P(a) &= pr\{T(X_1,...,X_m;F) > a \mid F\} \\
&= pr\{\bar{X}_m - \mu > a \mid F\} \\
&= pr\left\{ \frac{\bar{X}_m - \mu}{\frac{\sigma}{\sqrt{m}}} > \frac{a}{\frac{\sigma}{\sqrt{m}}} \mid F \right\} \sim N(0,1),
\end{aligned}
$$

However, the population variance is always an unknown quantity in bootstrap, where the only sample variance is a known quantity in the "bootstrap world". Therefore it is proper to take the t-distribution as the gold standard.

$$P(a) = pr\{T(X_1,...,X_m;F) > a \mid F\}$$
$$= pr\{\bar{X}_m - \mu > a \mid F\}$$
$$= pr\left\{\frac{\bar{X}_m - \mu}{\frac{s}{\sqrt{m}}} > \frac{a}{\frac{s}{\sqrt{m}}} \mid F\right\} \sim t(m-1),$$

Accordingly the bootstrap estimator would be the following,

$$\tilde{P}(a) = pr\{T(Y_1,...,Y_m;\hat{F}_n) > a \mid \hat{F}_n\}$$
$$= pr\{\bar{Y}_m - \bar{X}_n > a \mid \hat{F}_n\}$$
$$= pr\left\{\frac{\bar{Y}_m - \bar{X}_n}{\frac{s}{\sqrt{m}}} > \frac{a}{\frac{s}{\sqrt{m}}} \mid \hat{F}_n\right\}$$

In the simulation we sample form the population distribution ~ *N(0,1)*, so in the last step can be reduced by letting *s=1* on the right side of the inequality and two forms of standard deviation were both simulated.

$$\tilde{P}(a) = pr\{T(Y_1,...,Y_m;\hat{F}_n) > a \mid \hat{F}_n\}$$
$$= pr\{\bar{Y}_m - \bar{X}_n > a \mid \hat{F}_n\}$$
$$= pr\left\{\frac{\bar{Y}_m - \bar{X}_n}{\frac{s}{\sqrt{m}}} > \frac{a}{\frac{s}{\sqrt{m}}} \mid \hat{F}_n\right\}$$
$$= pr\left\{\frac{\bar{Y}_m - \bar{X}_n}{s} > \frac{a}{s} \mid \hat{F}_n\right\}$$
$$= pr\left\{\frac{\bar{Y}_m - \bar{X}_n}{\hat{\sigma}_{Y_m}} > a \mid \hat{F}_n\right\} \quad or\ pr\left\{\frac{\bar{Y}_m - \bar{X}_n}{\hat{\sigma}_{X_n}} > a \mid \hat{F}_n\right\}.$$

The results in Figure 5.3 for $\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma_{X_n}}}$, and Figure 5.4 for $\frac{\overline{Y_m} - \overline{X_n}}{\widehat{\sigma_{Y_m}}}$, were very similar without imputation, which reflect the relationship between $\widehat{\sigma_{X_n}}$ and $\widehat{\sigma_{Y_m}}, E\widehat{\sigma_{Y_m}} = \widehat{\sigma_{X_n}}$. However, the difference between $\widehat{\sigma_{X_n}}$ and $\widehat{\sigma_{Y_m}}$, were further delineated in bootstrap imputation from *sdf*. Figure 5.2 indicated that *sdf* can not remedy the situation when the gold standard to be compared was not selected properly, which should be regarded a pathological case.

Through this specific case of bootstrap mean we might tough the bottom of the problem. Other statistics $\theta(X, F)$ of finite sample has an exact bootstrap sampling distribution based on the combinatoric structure of finite sample bootstrap. Yet the sampling distribution might not always feasible to be compute numerically, therefore an approximation of the exact sampling distribution via a relatively simple, standard distribution is needed, such as a generalized t-pivotal,

$$t^* = \frac{\theta^*(\underline{X}) - \widehat{\theta(\underline{X})}}{\widehat{\sigma_{\theta(\underline{X})}}}.$$

Chapter 6. Summary and future work

Some failures of bootstrap were related to the discrete nature of the *empirical distribution function*. Density estimation may overcome the discreteness of the *empirical distribution function* via smoothing techniques. The common difficulty involved in density estimation is how to select an optimal smoothing window width. We devised a *Step Density Function*, named after the step-like shape of a *histogram*. The implementation of the *step density function* would result in a piece-wise continuous density function. This *step density function* is both a *MLE* and a *UMVUE*. Its most attractive feature, however, is that it can be readily applied to the imputed bootstrap resampling, an alternative bootstrap method developed in this thesis. Several examples have been provided to illustrate immediate application of univariate *step density function*, and it can be generalized to multivariate cases in the future to handle a wider spectrum of small sample problems.

Small sample bias in terms of the density function is essentially the residue of its approximated density to its exact distribution, which is theoretically available but difficult to compute in reality. We found that imputed bootstrap resampling may further reduce the bias of the small sample mean $\overline{X_m}$ after it was approximated by a t- instead of the normal distribution. We would raise the following very general hypothesis.

(1) A generalized t-pivotal exists in the form of $t^* = \dfrac{\theta^*(\underline{X}) - \widehat{\theta(\underline{X})}}{\widehat{\sigma_{\theta(\underline{X})}}}$, where sample statistic $\theta(\underline{X}, F)$ in small sample scenario and $\widehat{\sigma_{\theta(\underline{X})}}$ can be estimated via bootstrapping;

(2) $t^* \sim t(n-1)$ is the best 1st-order approximation;

(3) Imputed bootstrap resampling via *sdf* would reduce the bootstrap residues/bias of t^* from order $O(n^{-\frac{1}{2}})$ to $O(n^{-1})$.

The above general hypothesis on the t-pivotal will serve small sample bootstrap analysis in a similar role as the Law of Large Number for asymptotic analysis. An analytical proof might be quite difficult because $\theta(X,F)$ was proposed in a very general form; still proof for specific cases of $\theta(\underline{X},F)$, if attained, can be helpful in understanding the general properties or restrictions on $\theta(\underline{X},F)$. Historically R.A. Fisher' advocations of relaxing the normality assumption on general applications of the t-distribution have been supported by permutation study that the t-distribution was not sensitive to most foreseeable distributions with minimum regulatory conditions. Although an analytical proof has been elusive, we are in general faithful believers of Fisher's observations. However, one has to admit that only an analytical proof can fully clear all the barriers to the use of t-distributions for reliable analysis in small sample scenarios. If our general hypothesis on the t-pivotal could be proven, it would signify a substantial advancement in statistical inference based on small samples.

REFERENCES

Athraya, K.B. (1987) Bootstrap of the mean in the infinite in the infinite variance case. Ann. Statist. 15, 724-731.

Beran, R. and Ducharme, G. (1991) Asymptotic theory for bootstrap methods in statistics, Centre de Researes Mathematiques, Univ. of Montreal

Beran, R. and Srivastava, M.S. (1985) Bootstrap tests and confidence regions for functions of a covariance matrix. Ann. Statist. 13, 95-115.

Boos, D.D. (2003) Introduction to the bootstrap world. Statist. Science, 18, 168-174
Breiman, L. (1992) The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. J. Amer. Statist. Assoc. 87, 738-754.

Breiman, L., Freedman, J., Olshen, R., and Stone, C. (1984) Classification and Regression Trees. Wadsworth, Belmont, CA.

Breiman, L. and Spector, P. (1992) Submodel selection and evaluation in Regression. The X-Random Case. Int. Stat. Rev. 60, 291-319.

Buckland, S.T. (1983) Monte Carlo methods for confidence interval estimation using the bootstrap technique. Bull. Appl. Statist. 10, 194212.

Davison, A.C. and Hinkley, D.V. (1997) Bootstrap methods and their application. Cambridge University Press

Davison, A.C. Hinkley, D.V. and G.A. Young (2003) Recent development of Bootstrap methodology. Stat.Science, Vol.18, No.2. Silver Anniversary of the Bootstrap, 141-157.

De Angelis, D. and Young, G.A. Smoothing the bootstrap. (1992) International Statistical Review/Revue Internationale de Statistique, Vol.60,No.1 45-56.

Efron, B. (1979a) Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1-26.
Efron, B. (1979b) Computers and the theory of statistics: thinking the unthinkable. SIAM Review 21, 460-480.

Efron, B. (1981a) Nonparametric standard errors and confidence intervals. (With discussion.) Can. J. Statist. 9, 139-172.

Efron, B. (1981b) Nonparametric estimates of standard error: the jackknife, the bootstrap, and other methods. Biometrika 68, 589-599.

Efron, B. (1982) The jackknife, the bootstrap and other resampling plans. Volume 38 of CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvements on cross-validation. J. Amer. Statist. Assoc. 78, 316-331.

Efron, B. (1985) Bootstrap confidence intervals for a class of parametric problems. Biometrika 72, 45-58.

Efron, B. (1986) How biased is the apparent error rate of a prediction rule? J. Amer. Statist. Assoc. 81, 461-70.

Efron, B. (1987) Better bootstrap confidence intervals. (with discussion.) J. Amer. Statist. Assoc. 82, 171-200.

Efron, B. (1988) Bootstrap confidence intervals: good or bad? (With discussion.) Psychol. Bull. 104, 293-296.

Efron, B. (1990) More efficient bootstrap computations. J. Amer. Statist. Assoc. 85, 79-89.

Efron, B. (1991) Regression percentiles using asymmetric squared error loss. Statistica Sinica 1, 93-125.

Efron, B. (1992a). Six questions raised by the bootstrap. Exploring the Limits of Bootstrap (Eds. R. LePage and L. Billard), John Wiley and Sons, 99-126, New York.

Efron, B. (1992b) Jacknife-after-bootstrap standard errors and influence functions. J. Royal. Statist. Soc. B 54, 83-127.

Efron, B. (1992c) Bayes and likelihood calculations from confidence intervals. Tech, rep., Dept. of Statistics, Stanford Univ.

Efron, B. and Feldman, D. (1991) Compliance as an explanatory variable in clinical trials. J. Amer. Statist. Assoc. 86, 9-26.

Efron, B. and Gong, G. (1983) A leisurely look at the bootstrap, the jackknife and cross-validation. Amer. Statistician 37, 36-48.

Efron, B. and Stein, C. (1981) The jackknife estimate of variance, Ann. Statist. 9, 586-596.

Efron, B. and Tibshirani, R. (1985) The bootstrap method for assessing statistical accuracy. Behaviormetrika 17, 1-35.

Efron, B. and Tibshirani, R. (1986) Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1, 54-77.

Efron, B. and Tibshirani, R. (1991) Statistical data analysis in the computer age. Science 253, 390-395.

Efron, B. (1982) The jackknife, the bootstrap and other resampling plans. Vol. 38 of CBMS-NSF regional conference series in applied mathematics, SIAM.

Efron, B. and Tibshirani, R. (1986) Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1, 54-77.

Efron, B. and Tibshirani, R. (1993) An introduction to the bootstrap. New York: Chapman & Hall/CRC

Efron, B. (2004) A second thought of Bootstrap. Stat.Science, Vol.18, No.2. Silver Anniversary of the Bootstrap, 41-57.

Fang, K.T. and Wang, Y. (1994) Number-theoretic methods in statistics. London: Chapman & Hall

Faraway, J. and Jhun, M. (1990) Bootstrap choice of bandwidth for density estimation. J. Amer. Statist. Assoc. 85, 1119-1122.

Freedman, D.A. (1981) Bootstrapping regression models. Ann. Statist. 9,1218-1228.

Freedman, D.A. and Peters, S.C. (1984) Bootstrapping a regression equation: Some empirical results. J. Amer. Statist. Assoc. 79, 97106.

Gray, H.L.and Schucany, W.R. (1972) The generalized jackknife statistics. Marcel Dekker, New York.

Hall, P. (1986a) On the bootstrap and confidence intervals. Ann. Statist. 14, 1431-1452.

Hall, P. (1986b) On the number of bootstrap simulations required to construct a confidence interval. Ann. Statist. 14, 1453-1462.

Hall, P. (1987) On the bootstrap and likelihood-based confidence intervals. Biometrika 74, 481-493.

Hall, P. (1988x) Theoretical comparison of bootstrap confidence intervals. (with discussion.) Ann. Statist. 16, 927-953.

Hall, P. (1988b) On symmetric bootstrap confidence intervals. J. Royal. Statist. Soc. B 50, 35-45.

Hall, P. (1989x) On efficient bootstrap simulation. Biometrika 76, 613617.

Hall, P. (1989b) Antithetic resampling for the bootstrap. Biometrika 76, 713-724.

Hall, P. (1990) Performance of bootstrap balanced resampling in distribution function and quantile problems. Prob. Th. Rel. Fields 85, 239-267.

Hall, P. (1991) Bahadur representations for uniform resampling and importance resampling, with applications to asymptotic relative efficiency. Ann. Statist. 19, 1062-1072.

Hall, P. (1992) The Bootstrap and Edgeworth Expansion. SpringerVerlag, New York, Berlin, Heidelberg, London, Paris, Toyko, Hong Kong, Barcelona, Budapest.

Hall, P., DiCiccio, T.J. and Romano, J.P. (1989) On smoothing and the bootstrap. Ann. Statist. 17, 692-704.

Hartigan, J.A. (1969) Using subsample values as typical values, J. Amer.Statist.Assoc., 69, 383-393.

Hall, P. and Titterington, D. (1987) Common structure of techniques for choosing smoothing parameters in regression problems. J. Royal. Statist. Soc. B 49, 184-198.

Hall, P. and Titterington, M. (1988) On confidence bands in nonparametric density estimation and regression. J. Mult. Anal. 27, 228-254.

Hall, P. and Titterington, M. (1989) The effect of simulation order on level accuracy and power of Monte Carlo tests. J. Royal. Statist. Soc. B 51, 459-467.

Hall, P. and Wilson, S.R. (1991) Two guidelines for bootstrap hypothesis testing. Biometrics 47, 757-762.

Hardle, W. (1990) Applied Non-parameterc Regression. Oxford University Press.

Hardle, W. and Bowman, A. (1988) Bootstrapping in nonparametric regression: local adaptive smoothing and confidence bands. J. Amer. Statist. Assoc. 83, 102-110

Hartigan, J.A. (1969) Using subsample values as typical values. J. Amer. Statist. Assoc. 64, 1303-1317.

Hartigan, J.A. and Forthythe, A. (1970)Efficiency and confidence intervals generated by repeated subsample calculations, Biometrika, 57, 629-640.

Hartigan, J.A. (1971) Error analysis by replaced samples. J. Royal. Statist. Soc. B 33, 98-110.

Hartigan, J.A. (1975) Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. Ann. Statist. 3,573-580.

Hartigan, J.A. (1986) Discussion of Efron and Tibshirani (1986). Statistical Science 1, 75-77.

Hastie, T. and Tibshirani, R. (1990) Generalized additive models. Chapman and Hall, London.

Hinkley, D.V. (1977) Jackknifing in unbalanced situations. Technometrics 19, 285-292.

Hinkley, D.V. and Wei, B.C. (1984) Improvement of jackknife confidence limit methods. Biometrika, 71, 331-339.

Kiefer, J. and Wolfowitz, J. (1956) Robust properties of likelihood ratio tests. Biometrika 69, 19-27.

Liu. J. (2002) Monte Carlo strategies in scientific computing. Springer series instatistics.

Miller, R.G. (1964) A trustworthy jackknife. Ann. Math. Statist. 39, 1594-1605.

Miller, R.G. (1974) The jackknife - a review. Biometrika 61, 1-17.

Parr, W.C. (1983) A note on the jackknife, the bootstrap and delta method estimators of bias and variance. Biometrika 70, 719-722.

Parr, W.C. (1985) Jackknifing differentiable statistical functionals, J.Royal. Statist.Soc. B 47, 56-66.

Polansky, A.M. and Schucany, W.R. (1996) Kernel smoothing to improve bootstrap confidence intervals. J. Royal. Statist. Soc. B 59, 821-838.

Quenouille, M. (1949) Approximate tests of correlation in time series. J. Royal. Statist. Soc. B 11, 18-44.

Reeds, J.A. (1978) Jackknifing maximum likelihood estimates. Ann. Statist. 6, 727-739.

Rubin, D.B. (1981) The Bayesian bootstrap. Ann. Statist. 9, 130-134.

Schenker, N. (1985) Qualms about bootstrap confidence intervals. J. Amer. Statist. Assoc. 80, 360-361.

Scholz, F.W. (1980) Towards a unified definition of maximum likelihood. Can. J. Statist. 8, 193-203.

Sen, P.K. (1988) Functional jackknifing: rationality and general asymptotics. Ann. Statist. 16, 450-469.

Shao, J. and Wu, C.F.J. (1989) A general theory for jackknife variance estimation. Ann. Statist. 17, 1176-1197.

Shao, J. (1991) Consistency of jackknife variance estimators. Statistics 22, 49-57.

Shao, J. and Tu, D. (1995) The jackknife and bootstrap. New York: Springer.

Shao, J. (1996) Bootstrap model selection. J. Amer. Satist. Assoc. 91, 655-665.

Silverman, B.W. and Young, G.A. (1987) The bootstrap: to smooth or not to smooth? Biometrika 74, 469-479.

Silverman, B.W. (1986) Density estimation for statistics and data analysis. Chapman & Hall/CRC

Hastie, T. and Tibshirani, R. (1990) Generalized additive models. Chapman and Hall, London.

Tukey, J.W. (1958) Bias and confidence in not quite large samples. (Abstract.) Ann. Math. Statist. 29, 614.

Wang, S. (1989) On the bootstrap and the smoothed bootstrap. Communs. Statist.Theory Math., 18, 3949-3962.

Wang, S. (1995) Optimizing the smoothed bootstrap. Ann. Inst. Statist. Math., 47, 65-80.

Wu, C.F.J. (1986) Jackknife, bootstrap and other resampling plans in regression analysis (with discussion.) Ann. Statist. 14, 1261-1350.

Young, G.A. (1990a) Alternative smooothed bootstrap. J. R. Statist. Soc. B 52:3, 477-84.

Young, G.A. and Daniels, H.E. (1990b) Bootstrap Bias. Biometrika 77, 179-85.

Young, G.A. (1994) Bootstrap: more than a stab in the dark. Statist. Science, 9, 382-415.