AN ABSTRACT OF THE THESIS OF

Thomas M. Kincaid for the degree of Doctor of Philosophy in Statistics presented on November 25, 1997. Title: Estimating Absence.

Abstract approved: Redacted for Privacy

W. Scott Overton

The problem addressed is absence of a class of objects in a finite set of objects, which is investigated by considering absence of a species and absence in relation to a threshold. Regarding absence of a species, we demonstrate that the assessed probability of absence of the class of objects in the finite set of objects given absence of the class in the sample is either exactly or approximately equal to the probability of observing a specific single object from the class of objects given the protocol for observation, where probability is interpreted as a degree of belief. Regarding absence in relation to a threshold, we develop a new estimator of the upper confidence bound for the finite population distribution function evaluated at the threshold and investigate its properties for a set of finite populations. In addition we show that estimation regarding the initial ordered value in the finite population has limited usefulness.

ESTIMATING ABSENCE

by

Thomas M. Kincaid

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented November 25, 1997
Commencement June 1998

Doctor of Philosophy thesis of Thomas M. Kincaid presented on November 25, 1997.

APPROVED:

Redacted for Privacy

Major Professor, representing Statistics

Redacted for Privacy

Chair of Department of Statistics

Redacted for Privacy

Dean of Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for Privacy

Thomas M. Kincaid, Author

ACKNOWLEDGEMENTS

I am grateful to Dr. W. Scott Overton for insights, guidance, and support offered to me during preparation of this thesis. I thank Dr. David A. Thomas for numerous discussions and helpful suggestions and Dr. Jeffrey L. Arthur for insightful questions and recommendations. For the other members of my committee, Dr. Dawn Peters and Dr. Murray Levine, I appreciate your patience. I offer a special thank you to Dr. N. Scott Urquhart for encouragement and timely assistance. For my parents, who always have demonstrated faith in me despite my nonlinear path through life, I say thank you. Finally, I thank Susan, whose love has sustained me during the final stage of this endeavor.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

Table **Page**

LIST OF TABLES (Continued)

LIST OF TABLES (Continued)

Estimating Absence

1. INTRODUCTION

1.1 The Problem to be Addressed

Suppose one is interested in absence of a class of objects in a finite set of objects. We will refer to the finite set of objects as a universe. A finite population is then a function on the universe that assigns a value to each object in the universe. These population values provide the basis for identifying the class of objects that is the focus of this investigation. As an example of a class of objects, consider a specific species of fish among the fish of a single pool. For this example the universe is the set of fish in the pool; the population is categorical, assigning fish to specific species, and the class of interest is a particular species. As a second example, consider a quantitative attribute such as pH that was measured at a specific time in the center of each lake in a finite set of lakes. For this example the universe is the finite set of lakes; the population is continuous, assigning a pH value to each lake; and let the class of interest be defined by values of pH that are in a specific range, say, less than 5.

These two examples of this problem will be considered in subsequent chapters: (a) absence of a species and (b) absence in relation to a threshold. Absence of a species will be addressed in Chapter 2, and absence in relation to a threshold will be addressed in Chapters 3 and 4.

1.2 Absence of a Species

In order to assess absence in the universe of the class of objects, a probability sample of objects will be selected from the universe. It will be given that the class of objects is absent in the sample. Our goal may be described as follows: given absence of the class in the probability sample, we want to infer presence or absence of the class of objects in the universe. As will be demonstrated, inferring absence of a class of objects in a universe from a sample of objects poses certain difficulties.

We will express inferred absence as a probability. If the class of objects is not observed in the sample, then we propose to assess absence as a probability that the class of objects is absent given absence in the sample. This probability will be assessed as a property of the sampling design; the greater the effort that was expended in searching for the class, the greater the probability of absence given that none was found. Furthermore, the probability is clearly to be interpreted as a degree of belief. Interpretation of probability as a measure that summarizes the strength of conviction of an observer regarding occurrence of an event is supported by extensive probability theory (see de Finetti, 1937; Savage, 1954); further discussion regarding degree of belief is provided in Section 1.4.

Note that this problem does not have a parameter in the usual sense of that term. Since the problem does not have an identifiable parameter, the inferred probability should not be interpreted as an estimator in the usual sense. Rather, the probability is an assessment (assessed value) of our degree of belief that the class of objects is absent in

the universe given that the class is absent in the sample and the sampling effort. Thus, as will be developed, the degree of belief that the class is absent in the universe given absence in the sample is a property of the sampling design and the observation protocol. Consider, for example, a simple random sample of the objects in the universe. Given absence of the class in the sample, we will find that the appropriate degree of belief that the class is absent in the universe is given by the sampling fraction. The foundation of this rule will be a major development of this thesis.

Absence of a fish species in a single pool in a stream reach follows the general formulation outlined above. Although a finite sampling approach could be used for this problem, from a practical viewpoint it is extremely difficult to obtain a probability sample of fish in the pool. Therefore, a modeled approach using an explicit sampling protocol will be used. Given that no individuals of the species are observed during sampling, the inference goal is to assess the probability that the species is absent in the pool.

Absence of a species also will be considered in terms of absence of a specific species of fish in a stream reach, where the reach is composed of a finite set of pools, each of which could contain the species. In order to assess absence in the reach, the probability of absence will be developed at two levels: (a) the probability of absence of the species in the set of pools in the reach given absence in a sample of pools, and (b) the probability of absence of the species in an individual sampled pool given that no individuals of the species were observed in the pool. The probability of absence

developed for those two levels will be used to assess the probability of absence in the reach given that none were observed.

1.3 Absence in Relation to a Threshold

Absence in relation to a threshold will consider absence of objects in a universe that belong to a class of objects defined by values of a quantitative attribute in a specific range, say, less than a low threshold. Specifically, this example will consider a universe composed of a finite set of lakes, where the class of objects will be identified via values of a chemical attribute. For this example a finite sampling approach will again be used. If the threshold value is less than the initial ordered value in the finite population, this example involves a class that is absent, and one can take an approach identical to the approach taken in Chapter 2 for absence of a species. The general case, however, poses other issues that are not identifiable as inferring absence of a species. Specifically, these issues are associated with inference with respect to the extreme lower tail of the distribution function of the population.

Two approaches will be considered for inference with respect to the lower tail of the population distribution function: (a) inference in terms of the initial ordered value in the finite population, and (b) inference relative to the distribution function evaluated at the threshold value. The first approach will be addressed in Chapter 3, and the second approach will be addressed in Chapter 4.

1.4 Degree of Belief

As previously mentioned, the assessed probability that the target class is absent in the universe will be interpreted as a degree of belief in its absence. That is, degree of belief will serve as the interpretational basis of the probability measure. This section will provide some discussion pertaining to the concept of degree of belief. It will be given that the probability measures being used in this development obey the group of axioms codified by Kolmogorov (1933).

Two viewpoints may be employed as a means of providing a basis for interpretation of probability. These viewpoints are usually referenced as the objective and subjective viewpoints.

First, consider the objective viewpoint for interpretation of probability. Although this viewpoint is often developed by means of a relative frequency argument, a stochastic process is the appropriate orientation for the problem under consideration. That is, one would proceed as if presence or absence of the class of objects in the universe is the result of an underlying dynamic stochastic process. The difficulty with this approach is that, rather than being interested in the probability of an "absent" outcome by the stochastic process, we are interested in an assessment of the probability that class of objects is absent in the realized population that exists at the time of sampling.

The subjective viewpoint for interpretation of probability received major contributions by de Finetti (1937) and Savage (1954), among others. For this case

probability is interpreted from the viewpoint of summarizing the degree of belief (strength of conviction) of a rational observer in regard to the likelihood of occurrence of a particular event. Degree of belief will be taken to be synonymous with the term subjective probability. For the problem being considered, the subjective viewpoint implies that the assessed probability value summarizes our degree of belief that the class is absent in the universe given absence in the sample.

We will derive the probability of absence from a fiducial probability distribution, as introduced by Fisher (1930). It is clear that Fisher believed greater evidence was provided by the likelihood function than simply reported by the maximum likelihood estimate. Also, note that the dictionary definition of fiducial includes "founded on faith or trust". Thus, we will assess the probability of absence of the class of objects in the universe given absence in the sample as a fiducial probability that will be interpreted as a degree of belief.

2. ABSENCE OF A SPECIES

2.1 Introduction

If the class is observed in the sample, then presence of the class in the universe is established unambiguously. Conversely, failing to observe the class does not establish absence of the class in the universe unless the sample included every object in the universe. For the example of fish in a pool, if every fish in the pool was observed and the species was not present, then absence of the species is unambiguously established in the pool. If every fish in the pool was not observed, then failing to observe an individual belonging to the species does not establish absence of the species in the pool.

This chapter will address absence of a species in terms of absence of a specific species of fish in a stream reach, with assessment via a sample from a universe composed of the set of pools contained in the reach. Thus, it will be necessary to determine the probability of absence at two levels. First, the probability of absence in the universe of pools in the reach given absence in a sample of pools will be developed. Second, the probability of absence in an individual sampled pool given that no individuals of the species were observed in the pool will be developed. The estimated probability of absence in the reach given that none were observed will then be developed using the probability of absence in the universe of pools given absence in a sample of pools and the probability of absence for each sampled pool given that no individuals of the species were observed in that pool.

Absence of a species brings to mind the related concept of species extinction. Any discussion of species extinction in the context of this thesis implies that the universe constitutes the entire domain for the species and that the species was known to have been present in the domain at some time in the past. If the universe of objects does not constitute the entire domain of the species, then local extirpation of a species rather than species extinction is the relevant issue. In the latter case, the appropriate frame of reference is local, i.e., absence of the species is only applicable to the particular universe of objects from which the sample was taken. For example absence of a fish species in a lake would constitute local extirpation rather than species extinction, if the species previously had existed in that lake, except in the case where the lake encompasses the entire domain for the species. Neither extinction nor extirpation is a subject of this thesis, however, and attention will be focused simply on presence or absence of the class of objects in a specific universe at a specific time.

2.2 Absence of a Species in the Universe of Pools in a Reach

Investigation of absence of a species of fish in the universe of pools in a reach may be modelled as follows. Recall that a finite sampling approach will be applied to this problem, where the sample of pools will be a simple random sample of the pools in the reach. Let N equal the number of pools in the reach, n equal the number of pools in the sample, K equal the number of pools in the reach with the species present, and X equal the number of pools in the sample with the species present. Note that N and n are fixed known quantities and K is an unknown parameter. Then, for a given value of K,

X may be identified as a hypergeometric random variable. The probability of a particular value of X given K is furnished by:

$$P(X = x \mid K) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

Thus, the likelihood of a particular value of K given the observed value of X is provided by:

$$L(K = k \mid X = x) = P(X = x \mid K = k)$$

Note that the value of interest for X is zero, i.e., the species was absent from the sample of pools. If X was greater than zero, then the species was present in the sample.

Conventional inference tools can be applied to estimation of K, the number of pools in the reach with the species present. As an initial step, one could calculate a point estimate for K. The maximum likelihood estimator (MLE) for K is the integer value of K that maximizes the likelihood for the observed value of X, i.e., the integer, say d, that maximizes $P(X = x \mid K = d)$. Since $P(X = 0 \mid K = 0)$ is equal to one and $P(X = 0 \mid K = k)$ is less than one for k greater than zero, the MLE for K is zero when X is equal to zero, i.e., for observed absence in the sample of pools, the MLE supports the inference that the species is absent in the universe of pools.

As a second step, one could calculate an upper confidence bound for K, U(x). Given X=x, a 100(1-α)% upper confidence bound for K is given by:

$$U(x) = \text{the largest integer value of K such that } P(X \leq x \mid K) > \alpha$$

When X=0, the upper confidence bound is given by:

$$U(0) = \text{the largest integer value of K such that } P(X = 0 \mid K) > \alpha$$

Consider the case where N equals 100 and n equals 20. A 95% upper confidence bound for K is 12, which is a value that does not well support the inference that the species is absent in the universe of pools. Although the confidence bound can be made closer to zero by choosing a larger value of α or using a larger sample size, the bound will remain greater than zero for typical values of α and a realistic sample size. The conclusion to be reached here is that conventional estimation does not yield useful inference with respect to absence.

In light of these considerations, it would be useful to develop other tools to apply to this inference problem. One approach is to determine the odds that the species is absent in the universe of pools. Given observed absence of the species in the sample, the odds may be obtained by dividing the likelihood that K equals zero by the likelihood that K is greater than zero. Recall that $L(K = 0 \mid X = 0)$ is equal to one. Thus, the odds that the species is absent from the universe of pools in the reach given absence in the sample is provided by the following expression:

$$\frac{L(K=0\,|\,X=0)}{\sum_{k>0} L(K=k\,|\,X=0)} = \frac{1}{\sum_{k=1}^{N-n} L(K=k\,|\,X=0)}$$

The odds can be converted to a probability by dividing the odds by one plus the odds. Let ψ equal the probability that the species is absent from the universe of pools given absence in the sample of pools. Then ψ is assessed as follows:

$$\psi = \frac{1}{\sum_{k=0}^{N-n} L(K=k\,|\,X=0)}$$

Some discussion regarding this probability is warranted. Note that ψ is a normed likelihood, i.e., the likelihood in the numerator for K equal to zero is divided by the sum of the likelihoods for the complete set of allowable values of K. We think of ψ as a fiducial probability that will be interpreted as a degree of belief. Given that the only information available regarding K is observed absence of the species in the sample, ψ summarizes our degree of belief that the species is absent in the universe of pools. The same result could be obtained by using a uniform prior probability distribution for K in Bayes Theorem. We prefer, however, the fiducial probability viewpoint.

Two cases were investigated to explore the behavior of this assessment of the probability that the species is absent in the universe of pools given absence in the sample. For the first case n was set equal to 10, and ψ was calculated for the following range of values of N: {20, 30, 40, 50, 60, 70, 80, 90, 100}. For the second case N was set equal

to 100, and ψ was calculated for the following range of values of n: {5, 10, 15, 20, 25, 30, 35, 40, 45, 50}. Results are presented in Table 1, where the first case is given on the left half of the table and the second case is given on the right half of the table. Several patterns can be seen in Table 1. First, for a fixed value of n, the magnitude of ψ decreases as N increases. Second, for a fixed value of N, the magnitude of ψ increases as n increases. Lastly, and most importantly, the magnitude of ψ is always approximately equal to the sampling fraction, $\frac{n}{N}$, regardless of the values of n and N. In order to visualize that fact, the value of the sampling fraction is provided in Table 1 for each of the cases considered. An exposition of the logic that establishes that ψ is always approximately equal to the sampling fraction is provided in the Appendix; the exact value of ψ is given by: $\frac{n + 1}{N + 1}$. Thus, whenever the value of n is appreciably greater than one, the value of ψ will be approximately equal to the sampling fraction.

Evidence that ψ is approximately equal to the sampling fraction comes also from an additional source. Wright (1990) investigated upper confidence bounds on the number of defective units in a simple random sample from a finite universe of units and established upper bounds on the confidence coefficients associated with fixed-sized upper confidence bounds. For the specific case where the sample contained zero defective units, Wright showed that the maximum value for the confidence coefficient associated with an upper confidence bound of zero is equal to the sampling fraction. Further discussion of Wright's result is provided in the next section.

Table 1. Values of ψ, the probability of absence of a fish species in the universe of pools given absence in the sample of pools, where N is the number of pools in the reach, n is the sample size, and n/N is the sampling fraction.

n = 10			N = 100		
N	ψ	n/N	n	ψ	n/N
20	0.524	0.500	5	0.059	0.050
30	0.355	0.333	10	0.109	0.100
40	0.268	0.250	15	0.158	0.150
50	0.216	0.200	20	0.208	0.200
60	0.180	0.167	25	0.257	0.250
70	0.155	0.143	30	0.307	0.300
80	0.136	0.125	35	0.356	0.350
90	0.121	0.111	40	0.406	0.400
100	0.109	0.100	45	0.455	0.450
110	0.099	0.091	50	0.505	0.500

It is essential to comprehend that the result derived for the probability that the species is absent from the universe of pools given absence in the sample, ψ, and Wright's result are equivalent. That is, the two cases supply independent evidence in support of using the sampling fraction to represent the degree of belief that the species is absent from the universe of pools. Therefore, for simple random sampling, we conclude that the sampling fraction summarizes the degree of belief that the species is absent in the universe given absence in the sample. Although it is tempting to conclude that absence of the species in the universe is established given absence in the sample, that conclusion is not well supported by this result unless the sampling fraction is appreciably large.

Moreover, as we will later discuss, a certain conclusion that a class of objects is absent from a finite set of objects requires exhaustive sampling.

2.3 General Result

A general result now can be stated regarding the probability of absence in the universe of objects given absence in a sample from the universe. First, we will revisit the result of Wright (1990). Let K be the number of defective units in universe and X equal the number of defective units in a simple random sample. For the general case Wright established that the maximum value of the confidence coefficient, $1-\alpha$, for an upper confidence bound equivalent to the observed value of X is given by:

$$\max(1-\alpha) = 1 - P(X \leq x \mid K = x+1)$$

where x is the observed value of X. When X=0, the maximum value of the confidence coefficient for an upper bound equal to zero is given by:

$$\begin{aligned} \max(1-\alpha) &= 1 - P(X \leq x \mid K = x+1) \\ &= 1 - P(X = 0 \mid K = 1) \\ &= P(X = 1 \mid K = 1) \end{aligned}$$

Couched in terms of absence of a class of objects, $P(X = 1 \mid K = 1)$ is the probability of observing presence in the sample given that the universe contains a single object belonging to the class of objects. For probability sampling the inclusion probability is the probability that a specific object in the universe occurs in a sample. For simple

random sampling the inclusion probability for an object is the sampling fraction, and thus $P(X=1 \mid K=1)$ is equal to the sampling fraction for simple random sampling. Thus, ψ is approximately equal to $P(X=1 \mid K=1)$ for the simple random sampling model.

2.3.1 Another Model for Sampling

Suppose that X is distributed such that $P(X=0 \mid K=k) = \xi^k$, where ξ is the complement of the inclusion probability, and $k = 0, 1, \cdots, \infty$. It will be given that the inclusion probability for each object in the universe is constant. For this sampling model let ζ equal the probability of absence of the class of objects in the universe given absence of the class in the sample. Using the likelihood approach discussed in Section 2.2, ζ is assessed by:

$$\zeta = \frac{1}{\sum_{k=0}^{\infty} \xi^k}$$

$$= 1 - \xi$$

The result follows from the fact that the sum of the likelihood values in the denominator takes the form of an infinite geometric series. For the distribution being discussed, $P(X=1 \mid K=1)$ is equal to $1 - \xi$. Thus, ζ is exactly equal to $P(X=1 \mid K=1)$ for this distribution. Therefore, the probability that the class of objects is absent in the universe given absence of the class in the sample is exactly equal to $P(X=1 \mid K=1)$ for the

distribution presented in this section and is approximately equal to $P(X = 1 \mid K = 1)$ for the hypergeometric distribution.

2.3.2 Statement of the General Result

The results presented in the preceding sections allow the conclusion that the assessed probability of absence of the class of objects in the universe given absence of the class in the sample is either exactly or approximately equal to the probability of observing a single object from the class of objects given the protocol for observation, where probability is interpreted as a degree of belief. As discussed previously in this chapter, when the sampling design is simple random sampling, this probability is approximately equal to the sampling fraction.

2.4 Absence of a Species in an Individual Pool in a Reach

Pools will be surveyed using an explicit pool sampling protocol. For example the survey procedure could involve an individual wearing snorkeling equipment making one or more sampling passes through the pool and recording the number of fish belonging to the class observed during each sampling pass. The pool will be surveyed in such a manner that each fish in the pool has equal probability of being observed during a sampling pass and the several sampling passes are assumed independent. Let m equal the number of sampling passes and p equal the probability of observing any individual fish in a single sampling pass. Then 1- p is the probability of not observing any individual

fish in a single sampling pass, $(1-p)^m$ is the probability of not observing any individual fish in m passes, and $1-(1-p)^m$ is the probability of observing any individual fish in m passes. That is, $1-(1-p)^m$ is the inclusion probability for an individual fish. Let θ equal the probability of absence (no fish in the pool) given that none were observed in the pool by the survey protocol. Using the general result that was stated in Section 2.3.2, θ is assessed by $1-(1-p)^m$. We note that this sampling model does not require knowledge of the number of objects in the universe, as does the model based on a simple random sample. Although the model does require knowledge of the probability of observing any individual fish in a sampling pass, an estimate of that probability often can be obtained from the sampling procedure via maximum likelihood estimation. In other cases it may be necessary to use hypothetical values of the observation probability.

The probability that the species is absent from the pool also can be assessed using the likelihood approach employed for the simple random sampling model discussed previously. Let D equal the hypothetical, maximum number of individuals of the species in the pool and X equal the observed number of individuals in the sample. Regarding D, note that a finite pool cannot contain an infinite number of fish of finite size. Then the probability of absence in the pool given that none were observed during sampling is:

$$\theta = \frac{P(X=0 \mid K=0)}{\sum_{k=0}^{D} P(X=0 \mid K=k)} = \frac{1}{\sum_{k=0}^{D} \left((1-p)^m\right)^k} = \left(\frac{1 - \left((1-p)^m\right)^{D+1}}{1 - (1-p)^m} \right)^{-1}$$

$$= \left(1 - (1-p)^m\right) + R \approx 1 - (1-p)^m$$

Given the model employed, it is clear in this derivation that θ is approximately equal to the inclusion probability for an individual fish.

Values of θ are provided in Table 2 for a range of values of m and p. It is seen that θ, the probability of absence in an individual pool given that none were observed by the survey protocol, typically is much greater than ψ, the probability of absence in the universe of pools given absence in the sample of pools (see Table 1). For the range of observation probabilities, p, given in Table 2, θ can be made sufficiently large, e.g., greater than 0.97, by appropriate choice of m, the number of sampling passes in the pool. Whereas ψ is bounded by the proportion of pools in the universe that are included in the sample, the same restriction does not apply to θ.

2.5 Absence of a Species in a Reach

The operational survey of a reach involves two stages of "error" in concluding absence of a species. Specifically, one may miss the pools containing the species in selection of the sample of pools, or one may miss the species in surveying a selected

Table 2. Values of θ, the probability of absence of a fish species in a pool given that none were observed, where m equals the number of sampling passes and p equals the probability of observing any individual fish during a sampling pass.

m	p				
	0.5	0.6	0.7	0.8	0.9
1	0.500	0.600	0.700	0.800	0.900
2	0.750	0.840	0.910	0.960	0.990
3	0.875	0.936	0.973	0.992	0.9990
4	0.938	0.974	0.992	0.998	0.9999
5	0.969	0.990	0.998	0.9997	0.99999
6	0.984	0.996	0.999	0.9999	0.999999

pool. Suppose that a simple random sample of n pools from the universe of N pools in the reach was selected in order to infer presence or absence of the species in the reach. Let S represent the set of sampled pools, and for $u \in S$ let ${}^{I}\theta_{u}$ be the assessed value of the probability of absence given that none were observed for pool u, where I indicates the first tier of sampling, within a pool. Assuming that the same sampling protocol was used in each sampled pool, then ${}^{I}\theta_{u}$ will equal ${}^{I}\theta$ for each pool in the sample. Since sampling was conducted independently for each sampled pool, the value of the probability that the species is absent from the sample of pools given that none were observed is furnished by:

$$\prod_{u \in S} {}^{I}\theta_{u} = \prod_{u \in S} {}^{I}\theta = {}^{I}\theta^{n}$$

Then, let Θ equal the probability that the species is absent in the reach given that none were observed and ${}^{II}\theta$ equal the probability of absence of the species in the universe of pools in the reach given absence in the sample of pools, where *II* indicates the second tier of sampling. Then Θ is assessed by the product of the probability of absence of the species in the sample of pools given that none were observed and ${}^{II}\theta$. Thus, Θ is given by:

$$\Theta = \left(\prod_{u \in S} {}^{I}\theta_u \right) {}^{II}\theta$$

$$= {}^{I}\theta^{n} \; {}^{II}\theta \qquad \text{Assuming that } {}^{I}\theta \text{ is constant across pools}$$

$$\approx \left(1 - (1-p)^{m}\right)^{n} \frac{n}{N} \qquad \text{Assuming that p is constant across pools and fish}$$

Even though the value of ${}^{I}\theta$ can be made close to one, Θ will still be limited by the value of ${}^{II}\theta$, e.g., the sampling fraction, $\frac{n}{N}$, for simple random sampling. In order for the degree of belief that the species is absent in the reach to be high, an exhaustive sampling effort ($n \rightarrow N$) is required. Note, however, that ${}^{I}\theta^{n}$ will decrease as n approaches N unless ${}^{I}\theta$ is also very close to one, i.e., unless the sampling effort within the selected pools is also exhaustive. Thus, there is a tradeoff between m and n in relation to Θ.

To explore this tradeoff, Table 3 provides values of Θ for several choices of m and n, where the product of m and n is approximately equal to 100 and N is fixed at 100. As would be expected, for the largest values of m, Θ is controlled solely by the value of ${}^{II}\theta$. In Table 3 for each value of p, the maximum value of Θ is underlined, from which

Table 3. Values of Θ, the probability of absence of a fish species in a stream reach given that none were observed, where n equals the number of pools in the sample, m equals the number of sampling passes per selected pool, p equals the probability of observing any individual fish during a sampling pass, and N, the number of pools in the reach, equals 100. For each value of p, the largest value of Θ is underlined.

n	m	p				
		0.5	0.6	0.7	0.8	0.9
5	20	5.00e-02	5.00e-02	5.00e-02	5.00e-02	5.00e-02
6	17	6.00e-02	6.00e-02	6.00e-02	6.00e-02	6.00e-02
7	14	7.00e-02	7.00e-02	7.00e-02	7.00e-02	7.00e-02
8	12	7.98e-02	8.00e-02	8.00e-02	8.00e-02	8.00e-02
9	11	8.96e-02	9.00e-02	9.00e-02	9.00e-02	9.00e-02
10	10	9.90e-02	9.99e-02	1.00e-01	1.00e-01	1.00e-01
11	9	1.08e-01	1.10e-01	1.10e-01	1.10e-01	1.10e-01
12	8	1.14e-01	1.19e-01	1.20e-01	1.20e-01	1.20e-01
14	7	1.25e-01	1.37e-01	1.40e-01	1.40e-01	1.40e-01
17	6	1.30e-01	1.59e-01	1.68e-01	1.70e-01	1.70e-01
20	5	1.06e-01	1.63e-01	1.91e-01	1.99e-01	2.00e-01
25	4	4.98e-02	1.31e-01	2.04e-01	2.40e-01	2.49e-01
33	3	4.03e-03	3.72e-02	1.34e-01	2.53e-01	3.19e-01
50	2	2.83e-07	8.18e-05	4.48e-03	6.49e-02	3.03e-01
100	1	7.89e-31	6.53e-23	3.23e-16	2.04e-10	2.66e-05

a clear pattern can be seen. As p increases, the maximum value of Θ occurs for a larger value of n and a smaller value of m, i.e., for a fixed level of effort, as p increases, the degree of belief that the species is absent in the reach is maximized by surveying more pools and spending less time surveying each selected pool.

2.6 Generalization

This section will address means by which the results that have been presented in this chapter could be generalized. To begin, suppose that one does not know, or have an estimate of, the number of objects in the universe. As an example, consider inference regarding presence or absence of a species from a non-mobile resource such as trees in a forest. One could divide the forest into N quadrats of equal size and select a simple random sample of n quadrats. Then, as established by previous results, the probability that the species is absent from the forest given absence in the sample is assessed by the sampling fraction of quadrats. Application to a mobile resource will require certain assumptions regarding the interaction of the observer and the resource.

Suppose that one is interested in absence of a species from a mobile resource and cannot divide the domain into physical sampling units. For example, in sampling a population of fish in a lake, quadrats are not feasible as units for counting fish. However, a measure analogous to the sampling fraction of quadrats could be employed to assess the probability of absence. As an example, one could define effective effort as the effort expended in obtaining the sample relative to the effort that would be expended in conducting a complete census of the fish in the lake. The value of effective effort then would be employed in assessing the probability of absence of the species.

3. ABSENCE IN RELATION TO A THRESHOLD I

3.1 Introduction

This chapter and the next chapter will address absence of objects in a universe that belong to a class of objects defined by values of a quantitative attribute in a specific range, say, less than a low threshold. As an example of this type of problem, consider a universe composed of a finite set of lakes. Suppose that a sample of lakes was selected from the universe, and a chemical attribute was measured for each lake in the sample. Given that none of the sample values of the chemical attribute are less than a threshold value, the inference goal is to assess whether any of the lakes in the universe have values of the chemical attribute that are less than the threshold value. Finally, note that methodology developed for inference regarding a low threshold can be applied to inference regarding a high threshold, so that it suffices to study only the low threshold.

Two inferential approaches can be employed for this problem. The first approach will be addressed in this chapter, and the second approach will be addressed in Chapter 4. For the first approach we will investigate the initial ordered value in the population, e.g., the smallest value of the chemical attribute for the lakes in the universe. In Chapter 4 the estimated distribution function will be evaluated at the threshold value.

Let α reference the initial ordered value in the population. Then, estimates of α, both point estimates and interval estimates, can be compared to the threshold value in order to provide information regarding whether any lakes in the universe possess values

of the attribute that are less than the threshold value. Inference regarding α will employ two types of estimation methodology: (1) the jackknife, the bootstrap, and sample spacings and (2) extreme value theory.

Prior to examining inference about α, some comments regarding the ordered population values will be provided. For a finite universe, such as a set of lakes, the ordered population values represent a function defined on the population. That is, the ordered population values are a fixed set of quantities that characterize the finite population. Note the analogy between the ordered population values and the sample order statistics, i.e., the ordered values in a sample. As the name implies, however, the sample order statistics are indeed statistics, i.e., functions that are defined on a sample. Although the ordered population values are analogous to sample order statistics, it is imperative to remember that the ordered population values are not statistics.

3.2 Inference Using the Jackknife, Bootstrap, and Sample Spacings

In this section methodology based on application of the first order jackknife, the bootstrap, and sample spacings will be investigated for estimation of α. Regarding the jackknife and bootstrap, estimators developed in this section represent standard application of that methodology. Further information about the jackknife and bootstrap is available in Efron (1982). For sample spacings, existing theory is utilized to develop a new estimator of α. Further information about sample spacings is available in Pyke (1965).

Ensuing development will assume that a simple random sample of lakes has been obtained from the universe of lakes. Again, let N equal the total number of lakes in the universe, and n equal the sample size. Consideration will be given to procedures relating to the maximum likelihood estimator (MLE) of α, i.e., the first order statistic in the sample. Analysis will utilize various configurations of the Uniform and Normal distributions.

3.2.1 Uniform Distribution

In this section estimators of the initial ordered value in a finite population, i.e., α, will be developed using theory applicable to the infinite Uniform(α, β) distribution, where α is the lower bound and β is the upper bound of the distribution. Alpha is being used as the lower bound for the Uniform distribution in order to emphasize the connection with the initial ordered value in a finite population. Theory for the estimators will be illustrated using the specific example of a sample of size 16 selected from a Uniform(80, 120) distribution. Results from simulations examining performance of the estimators will then be presented. The simulations will utilize a fixed finite population of size 1,000 selected from the Uniform distribution.

Assume that $y_1, y_2, \cdots, y_n$ are independent and identically distributed (*iid*) Uniform(α, β) random variables. Let $y_{(1)}, y_{(2)}, \cdots, y_{(n)}$ be the associated sample order statistics. Then the MLE of α, $\hat{\alpha}_{MLE}$, and the expected value of $\hat{\alpha}_{MLE}$ are given by:

$$\hat{\alpha}_{MLE} = y_{(1)}$$

$$E[\hat{\alpha}_{MLE}] = \alpha + \frac{(\beta - \alpha)}{(n+1)}$$

Thus, the bias of $\hat{\alpha}_{MLE}$ is given by:

$$E[\hat{\alpha}_{MLE} - \alpha] = \frac{(\beta - \alpha)}{(n+1)}$$

For the specific example the expected value of $\hat{\alpha}_{MLE}$ is 82.35, and the bias of $\hat{\alpha}_{MLE}$ is 2.35.

The jackknife estimator of the bias of $\hat{\alpha}_{MLE}$ and its expected value are given by:

$$\text{Jackknife Estimator of Bias} = \frac{(n-1)}{n}(y_{(2)} - y_{(1)})$$

$$E[\text{Jackknife Estimator of Bias}] = \frac{(n-1)(\beta - \alpha)}{n(n+1)}$$

Then, the jackknife estimator of α, which will be referenced as Jackknife, $\hat{\alpha}_J$, and the expected value of $\hat{\alpha}_J$ are given by:

$$\hat{\alpha}_J = \hat{\alpha}_{MLE} - \frac{(n-1)}{n}(y_{(2)} - y_{(1)})$$

$$E[\hat{\alpha}_J] = \alpha + \frac{(\beta - \alpha)}{n(n+1)}$$

Thus, the bias of $\hat{\alpha}_J$ is given by:

$$E[\hat{\alpha}_J - \alpha] = \frac{(\beta - \alpha)}{n(n+1)}$$

It is seen that bias of $\hat{\alpha}_J$ is smaller than the bias of $\hat{\alpha}_{MLE}$ by a factor of n^{-1}. For the specific example the expected value of the jackknife estimator of the bias of $\hat{\alpha}_{MLE}$ is 2.21, the expected value of $\hat{\alpha}_J$ is 80.15, and the bias of $\hat{\alpha}_J$ is 0.15. It is clear that the jackknife estimator performs very well in terms of bias for the case of a Uniform distribution. The jackknife procedure also provides an estimator of the standard error of $\hat{\alpha}_{MLE}$, which is given by:

$$\textbf{Jackknife Standard Error Estimator} = \frac{(n-1)}{n}(y_{(2)} - y_{(1)})$$

Note that this estimator is the same as the jackknife estimator of the bias of $\hat{\alpha}_{MLE}$.

The bootstrap procedure may be described as follows. Given a set of sample values of size n, a large number of bootstrap samples are created. Each bootstrap sample consists of a random sample with replacement from the n original sample values. Typically, the bootstrap sample is also of size n. For each bootstrap sample an estimate of the statistic calculated from the original sample is determined. The bootstrap estimator consists of the mean of the estimates from the set of bootstrap samples. For estimating α the estimate determined for each bootstrap sample is the smallest value in the bootstrap sample. The expected value of the bootstrap estimator of α, $\hat{\alpha}_B$, for any distribution is given by:

$$E[\hat{\alpha}_B] = \sum_{i=1}^{n} \left(E[y_{(i)}] \cdot \left(\left(\frac{n-i+1}{n} \right)^n - \left(\frac{n-i}{n} \right)^n \right) \right)$$

where the terms in the summation are the product of the expected value of an order statistic from the original sample and the probability that the order statistic will be the smallest value in a bootstrap sample. For the specific case of a Uniform(α, β) distribution, the expected value of the bootstrap estimator of α is given by:

$$E[\text{Bootstrap Estimator of } \alpha] = \sum_{i=1}^{n} \left(\left(\alpha + \frac{i(\beta - \alpha)}{(n+1)} \right) \cdot \left(\left(\frac{n-i+1}{n} \right)^n - \left(\frac{n-i}{n} \right)^n \right) \right)$$

Note that the bootstrap procedure also provides an estimator of the standard error of $\hat{\alpha}_{MLE}$, which consists of the standard deviation of the estimates from the set of bootstrap samples.

An estimator of the bias of $\hat{\alpha}_{MLE}$ can be obtained from the bootstrap procedure and consists of the bootstrap estimator of α, $\hat{\alpha}_B$, minus the maximum likelihood estimator (see Efron, 1982). The bootstrap estimator of bias can then be subtracted from $\hat{\alpha}_{MLE}$ to produce a bias-adjusted estimator, which will be referenced as Bootstrap I, $\hat{\alpha}_{B1}$. Note that the value of $\hat{\alpha}_{B1}$ is given by: $2\,\hat{\alpha}_{MLE} - \hat{\alpha}_B$. For the specific example the expected value of $\hat{\alpha}_B$ is 83.58, the expected value of the bootstrap estimator of the bias of $\hat{\alpha}_{MLE}$ is 1.23, the expected value of $\hat{\alpha}_{B1}$ is 81.12, and the bias of $\hat{\alpha}_{B1}$ is 1.12, which is close to an order of magnitude greater than the bias of $\hat{\alpha}_J$.

Sample spacings are the differences between successive sample order statistics. A sample of size n thus defines a set of size n-1 sample spacings. For a Uniform(α, β) distribution the expected value of the sample spacings, and therefore the expected value of the average of the sample spacings, is given by: $\frac{(\beta - \alpha)}{(n + 1)}$ (see Pyke, 1965). Note that this quantity was earlier shown to be the bias of $\hat{\alpha}_{MLE}$. Let $\bar{s}$ be the average of the sample spacings. A fourth estimator of α, which will be referenced as Average Spacing, $\hat{\alpha}_S$, can be defined by: $\hat{\alpha}_S = \hat{\alpha}_{MLE} - \bar{s}$. By construction $\hat{\alpha}_S$ is unbiased for α.

A fifth estimator of α, which will be referenced as Bootstrap II, $\hat{\alpha}_{B2}$, was obtained by applying the bootstrap procedure to $\hat{\alpha}_S$. This bootstrap estimator was included in order to estimate the standard error of $\hat{\alpha}_S$. Note that $\hat{\alpha}_{B2}$ is also unbiased for α.

Simulation results for the estimators based on the Uniform distribution are provided in Table 4. A finite population was created by selecting 1,000 values from the Uniform(80, 120) distribution, and samples were selected from this fixed finite population. The value of α for the selected finite population was 80.05. The simulations consisted of 1,000 replications, where a simple random sample of size 16 was selected for each replication and 500 bootstrap samples per replication were used for the two bootstrap estimators. The means, standard deviations, and root mean square errors are presented in the Table 4 for the estimates and the estimated standard errors. Note that the standard deviation of the estimates for $\hat{\alpha}_{MLE}$ and $\hat{\alpha}_S$ assess the standard errors for those

Table 4. Means, standard deviations (SD) and root mean square errors (RMSE) for estimates and estimated standard errors of the initial ordered value in a finite population of size 1,000 that was selected from the Uniform distribution, where 1,000 replications were used in the simulations and the initial ordered value in the population was 80.05.

Procedure	Estimate			Est. Standard Error		
	Mean	SD	RMSE	Mean	SD	RMSE
MLE	82.36	2.11	3.13			
Jackknife	80.15	3.06	3.06	2.20	2.06	2.06
Bootstrap I	81.14	2.34	2.58	2.07	1.04	1.04
Avg. Spacing	80.01	2.25	2.25			
Bootstrap II	80.18	2.24	2.24	2.09	1.04	1.05

estimators. Regarding the estimates, $\hat{\alpha}_S$ has the smallest bias, although $\hat{\alpha}_J$ and $\hat{\alpha}_{B2}$ also performed well in term of bias. In terms of root mean square error of the estimates, $\hat{\alpha}_S$ and $\hat{\alpha}_{B2}$ performed better than the other estimators. Regarding bias of the estimated standard errors, $\hat{\alpha}_{B1}$, $\hat{\alpha}_{B2}$, and $\hat{\alpha}_J$ all had minimal bias. The root mean square error of the estimated standard error was much larger for $\hat{\alpha}_J$ in comparison to $\hat{\alpha}_{B1}$ or $\hat{\alpha}_{B2}$. Overall $\hat{\alpha}_S$ performed best among the estimators for the Uniform distribution.

3.2.2 Normal Distribution

In this section the estimators that were developed for the Uniform distribution will be applied to data from the Normal distribution. Performance of the estimators will be examined using simulations that utilized a fixed finite population of size 1,000 selected

from the Normal distribution with a mean of 100 and a standard deviation of 10. The value of α for the selected finite population was 65.64. Note that a parameter analogous to α does not exist for the infinite Normal distribution.

Simulation results are provided in Table 5. The simulations consisted of 1,000 replicate samples of size 16 using 500 bootstrap samples per replication. Although performance of the estimators was similar to that observed for the Uniform distribution, bias was a more severe problem for the Normal distribution. Although $\hat{\alpha}_J$ had the least bias of the estimates, the amount of bias was still substantial. The jackknife estimator also had the smallest root mean square error of the estimators of α. Root mean square error of the estimated standard errors was smallest for $\hat{\alpha}_{B1}$ followed by $\hat{\alpha}_{B2}$ and $\hat{\alpha}_J$. The considerable bias of the estimates of α, however, indicates that these estimators based on the infinite Uniform distribution would not be very valuable for estimating the initial ordered value in a finite population selected from the Normal distribution.

3.3 Extreme Value Theory

In this section use of extreme value theory to estimate α will be investigated. For the following discussion it will be assumed that several (say r) independent simple random samples are available. Assuming the existence of an asymptotic distribution for the minimum value in a sample of size n from a common distribution F, Fisher and Tippett (1928) proved that the asymptotic distribution can take only three forms. These

Table 5. Means, standard deviations (SD) and root mean square errors (RMSE) for estimates and estimated standard errors of the initial ordered value in a finite population of size 1,000 that was selected from the Normal distribution, where 1,000 replications were used in the simulations and the initial ordered value in the population was 65.64.

Procedure	Estimate			Est. Standard Error		
	Mean	SD	RMSE	Mean	SD	RMSE
MLE	81.83	5.75	17.18			
Jackknife	77.17	9.10	14.69	4.66	4.23	4.37
Bootstrap I	79.61	7.02	15.63	3.39	2.09	3.15
Avg. Spacing	79.43	6.12	15.09			
Bootstrap II	79.73	6.04	15.33	3.41	2.09	3.42

three forms in turn can be combined into a single generalized extreme-value (GEV) distribution given by the following:

$$L(x) = \begin{cases} 1 - \exp(-(1+\rho(x-\gamma)/\beta)^{1/\rho}) & \text{for } \rho \neq 0 \\ 1 - \exp(-\exp((x-\gamma)/\beta)) & \text{for } \rho = 0 \end{cases}$$

where γ is a location parameter, β is a scale parameter, and ρ is a shape parameter. The value of the shape parameter ρ divides the GEV distribution into the three forms identified by Fisher and Tippett: (1) $\rho>0$, for which the value of x is bounded below; (2) $\rho=0$, which is the Gumbel distribution and for which x is unlimited; and (3) $\rho<0$, for which the value of x is bounded above. The case $\rho=0$ is interpreted as the limit of L(x) as ρ approaches zero. For a given set of data, the parameters of the GEV and the Gumbel distributions can be estimated by the method of probability-weighted moments

(see Hosking, Wallis, and Wood (1985)). Note that the Gumbel distribution is the applicable distribution for the minimum value in a sample selected from the infinite Normal distribution.

Recall that our interest in this chapter concerns estimation of the initial ordered value in a finite population using a sample from the population. Extreme value theory, conversely, is applicable to samples selected from an infinite population defined by a particular distribution function. Applied to a finite population selected from, say, the Normal distribution, the GEV and Gumbel distributions are approximations to the distribution of the minimum value in a sample selected from the given finite population.

Application of extreme value theory to our problem will be illustrated with a particular case. A finite population of size 1,000 was selected from a Normal distribution with a mean of 100 and a standard deviation of 10. A set of r independent simple random samples of size 16 was selected from the fixed finite population, where r was a member of the set: {5, 10, 15, 20, 25, 50}. For each of the r simple random samples, the minimum value was determined to create the set of values $\underline{m} = \{m_i, i = 1, 2, \cdots, r\}$. The set $\underline{m}$ was utilized to fit the parameters of the GEV and Gumbel distributions.

Samples were selected from the fixed finite population for the range of values of r indicated previously, and parameters of the GEV and Gumbel distribution were estimated. Plots of the fitted Gumbel distributions are provided in Figure 1 for each value

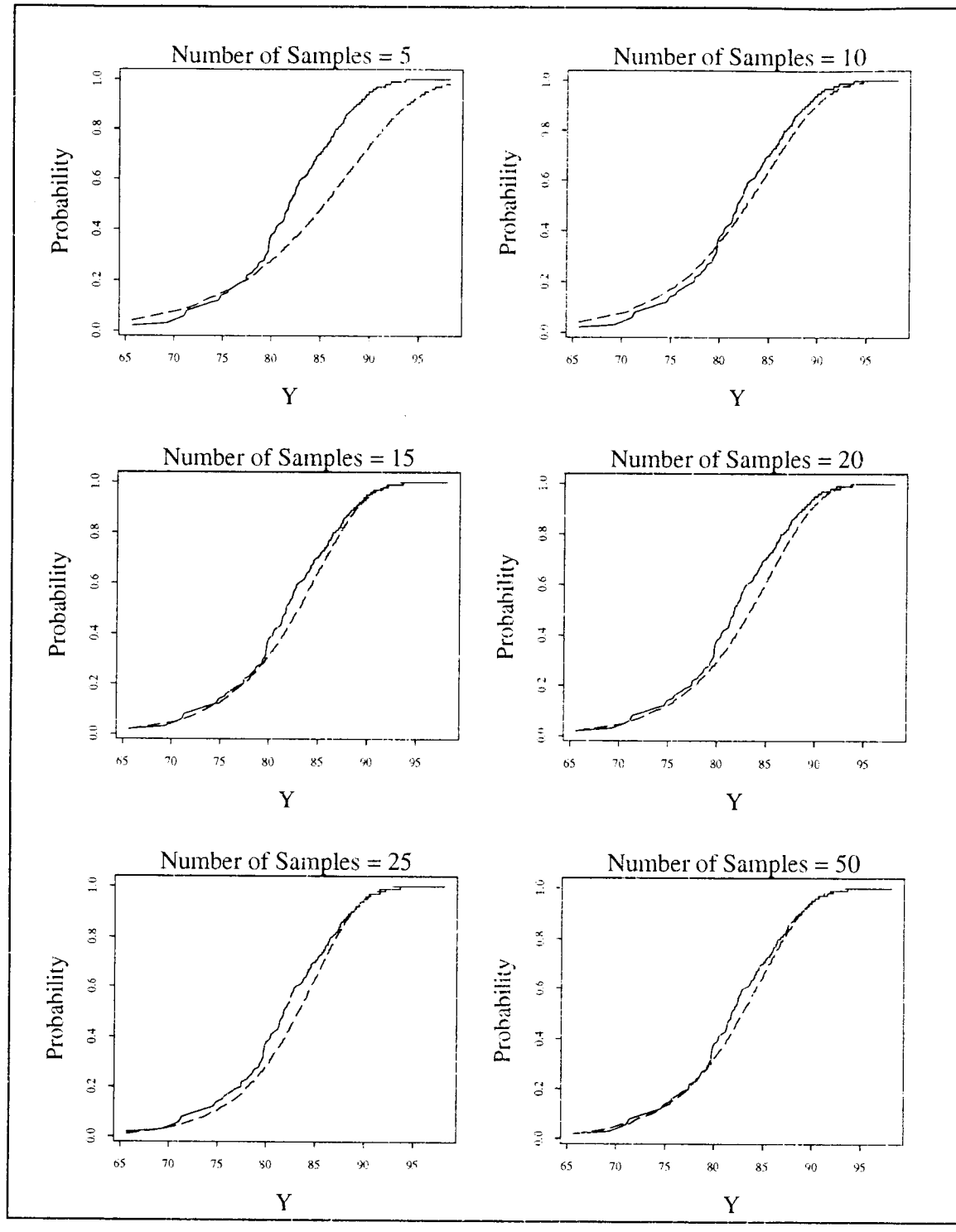

Figure 1. Exact and predicted (Gumbel) distributions of the minimum value in a sample of size 16 from a finite population of size 1,000 that was selected from a Normal distribution with mean 100 and standard deviation 10, where each plot represents a different value of the number of samples, the exact distribution is the solid line, and the predicted distribution is the dashed line.

of r. The exact distribution of the minimum value in a sample of size 16 selected from the finite population is also included in the plots. Note that for any of the ordered values in a finite population of size N, say $y_{(i)}$, the probability of that specific value being the first order statistic in a sample of size n is provided by the following:

$$\text{Probability} = \frac{\binom{N-i}{n-1}}{\binom{N}{n}}$$

These probabilities can be used to construct the exact distribution for the minimum value in a sample selected from the finite population. Although the amount of agreement between the fitted Gumbel distribution and the exact distribution improves as r increases, agreement is quite good even for the case r=10.

The GEV distribution, and thus the Gumbel distribution, has a simple form for the inverse distribution (quantal) function, $Q(\eta)$, which is given by:

$$Q(\eta) = \begin{cases} \gamma + \beta\left(\left(-\log(1-\eta)\right)^{\rho} - 1\right)/\rho & \text{for } \rho \neq 0 \\ \gamma + \beta \log\left(-\log(1-\eta)\right) & \text{for } \rho = 0 \end{cases}$$

Since α, the parameter of interest in this chapter, is the initial ordered value in the finite population, there is a known probability associated with α being the first order statistic in a simple random sample from the population. That probability is equal to $\frac{n}{N}$, which for the case under consideration equals 0.016. Therefore, a reasonable estimator for α is given by the quantile associated with a probability of 0.016, i.e., $Q\left(\frac{n}{N}\right)$.

Simulation results are presented in Table 6. For each choice of r, estimates of α were calculated for both the GEV and Gumbel distributions. The simulations used 1,000 replications for each value of r. Means, standard deviations, and root mean square errors of the estimates for both distributions are provided in Table 6. In addition the bootstrap procedure was applied to the estimates from both distributions. In Table 6 Bootstrap I and Bootstrap II reference the bootstrap procedure applied to the GEV and Gumbel distributions, respectively. The bootstrap procedures used 500 bootstrap samples per replication. Means, standard deviations, and root mean square errors of the bootstrap estimates of α and the bootstrap estimated standard errors are provided in Table 6. Regarding estimation of α, the Gumbel-based estimates performed better than the GEV-based estimates. For all cases except r=5, the Gumbel-based estimate had smaller bias, and the root mean square error of the estimate was smaller for the Gumbel distribution for all cases. For both distributions and for all values of r, the mean of the bootstrap estimate of standard error was biased downward. Since the bootstrap procedure would be employed to calculate confidence intervals for α, the latter fact is an impediment to use of the estimators. Root mean square error for the bootstrap estimate of standard error consistently was smaller for the Gumbel distribution in comparison to the GEV distribution. As mentioned previously, the Gumbel distribution is the appropriate asymptotic distribution for the minimum value in a sample from an infinite Normal distribution, so the superior performance of the Gumbel-based estimator for a finite population selected from a Normal distribution was not surprising.

Table 6. Means, standard deviations (SD) and root mean square errors (RMSE) for estimates and estimated standard errors of the initial ordered value in a finite population of size 1,000 that was selected from the Normal distribution, where 1,000 replications were used in the simulations and the initial ordered value in the population was 65.64.

Number of samples per replication = 5						
Procedure	Estimate			Est. Standard Error		
	Mean	SD	RMSE	Mean	SD	RMSE
GEV	66.48	10.29	10.32			
Bootstrap I	67.73	7.75	8.03	6.82	3.61	5.01
Gumbel	65.00	7.47	7.50			
Bootstrap II	66.27	5.94	5.97	4.83	2.07	3.35
Number of samples per replication = 10						
Procedure	Estimate			Est. Standard Error		
	Mean	SD	RMSE	Mean	SD	RMSE
GEV	66.99	6.47	6.61			
Bootstrap I	68.31	5.27	5.91	4.32	1.91	2.88
Gumbel	64.99	4.99	5.03			
Bootstrap II	66.26	4.09	4.14	3.30	1.08	2.01
Number of samples per replication = 15						
Procedure	Estimate			Est. Standard Error		
	Mean	SD	RMSE	Mean	SD	RMSE
GEV	67.28	4.98	5.24			
Bootstrap I	68.61	4.08	5.05	3.43	1.31	2.03
Gumbel	65.09	3.83	3.87			
Bootstrap II	66.32	3.18	3.25	2.65	0.76	1.40

Table 6. (Continued)

Procedure	Estimate			Est. Standard Error		
	Mean	SD	RMSE	Mean	SD	RMSE
Number of samples per replication = 20						
GEV	67.51	4.32	4.71			
Bootstrap I	68.81	3.62	4.81	2.91	0.98	1.72
Gumbel	65.14	3.46	3.50			
Bootstrap II	66.33	2.84	2.92	2.28	0.61	1.33
Number of samples per replication = 25						
GEV	67.54	3.91	4.35			
Bootstrap I	68.82	3.25	4.55	2.57	0.81	1.57
Gumbel	65.11	3.03	3.08			
Bootstrap II	66.32	2.46	2.55	2.04	0.51	1.11
Number of samples per replication = 50						
GEV	67.66	2.61	3.30			
Bootstrap I	68.94	2.24	3.99	1.82	0.44	0.90
Gumbel	65.12	2.10	2.16			
Bootstrap II	66.32	1.72	1.85	1.46	0.30	0.71

4. ABSENCE IN RELATION TO A THRESHOLD II

4.1 Introduction

In this chapter the estimated distribution function (cdf) of a quantitative attribute for the objects in a finite universe of objects will be utilized for inference relative to a threshold. Specifically, the estimated cdf will be evaluated at a threshold value in the extreme lower tail of the population distribution function. The estimated cdf, evaluated at any threshold value, is an estimator of the proportion of objects in the universe that have values of the attribute less than or equal to that threshold value. Thus, an upper confidence bound for the cdf will provide an upper bound on the proportion of objects in the universe that have values of the attribute less than or equal to that threshold value.

Model-based methodology for estimation of the cdf will be investigated in this chapter. Since interest lies only in the extreme lower tail of the cdf, the model is required only to provide an adequate fit to that portion of the cdf. The model should have sufficient flexibility to be able to accommodate a variety of shapes of the lower tail of the cdf. In addition the model should possess sufficient robustness to avoid excessive sensitivity to violations of model assumptions. Thus, some form of model will be utilized in order to provide a point estimator and an upper confidence bound for the finite population cdf evaluated at the threshold value. In addition the measured values for objects in the sample will be treated as fixed, i.e., observed without error, from which it follows that the number of sampled values less than the threshold value will be a known quantity. The sampled values will be used to estimate the parameters of the model.

Then, conditional on the sample and on the model, a point estimator and estimators of the upper confidence bound for the finite population cdf will be developed. Estimation of the cdf at the threshold value and its associated upper confidence bound will include the case when a predictor variable is available and the case when a predictor variable is not available.

The upper tail of the cdf is also a potential site for inference involving the same issues and considerations as we deal with in the lower tail. Since these can be addressed by inverting the cdf, only the lower tail case will be investigated.

Regarding the upper confidence bound for the cdf evaluated at the threshold value, a typical bound is the binomial upper confidence bound. The conventional estimator of the cdf, i.e., the proportion of sample values less than or equal to each sample order statistic, and the associated binomial upper confidence bound are provided in Figure 2 for a sample of size 16 from a finite population. Note that, for all values of the threshold less than the first sample order statistic, the binomial upper confidence bound remains fixed at a value greater than zero, which we consider unreasonable performance for the upper confidence bound. We will seek an approach for the upper confidence bound that allows the bound to decrease as the threshold value continues to decrease in magnitude relative to the observed values.

A growing literature exists regarding estimation of the finite population cdf. An early paper was contributed by Sedransk and Sedransk (1979), who considered use of

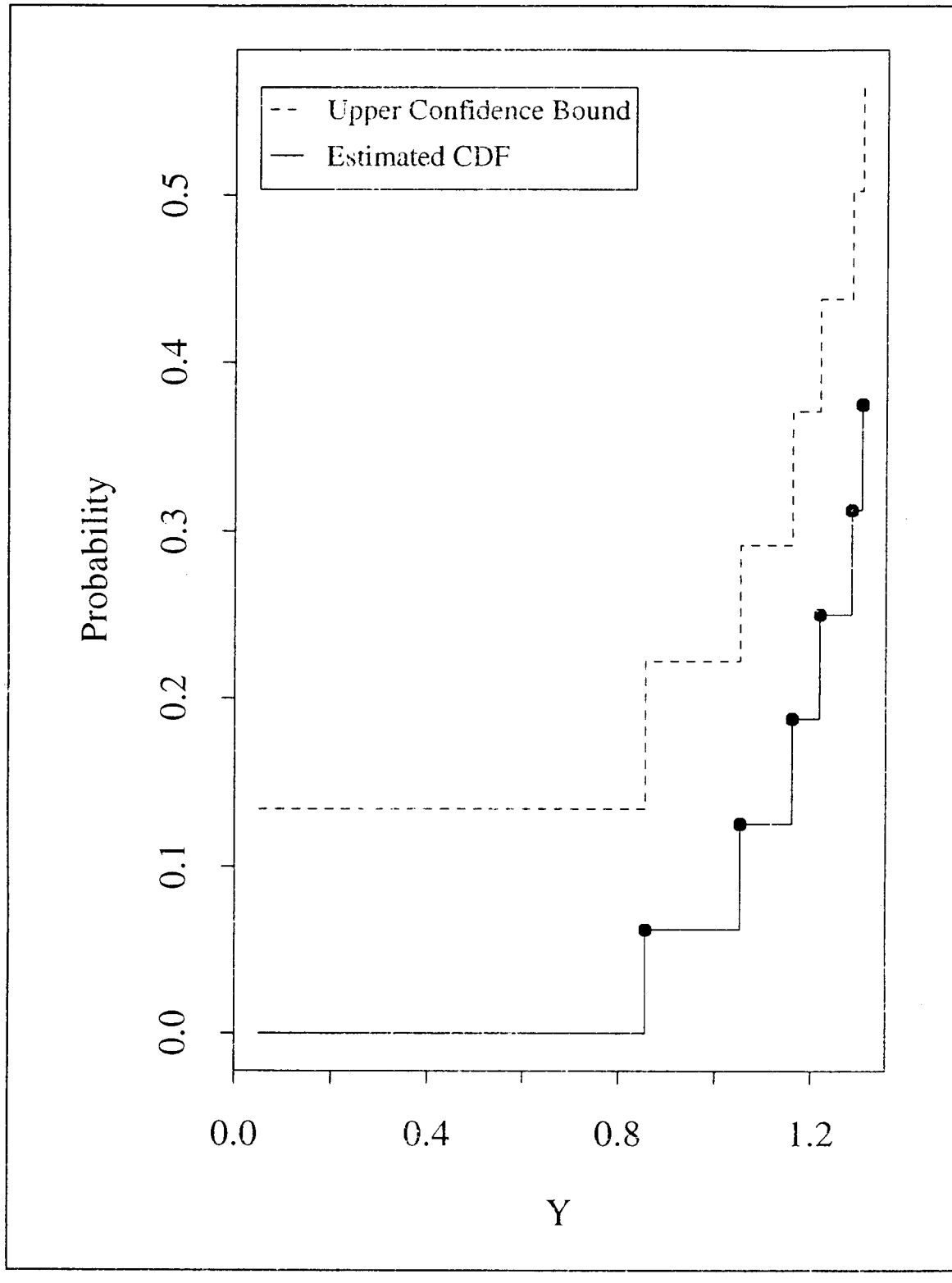

Figure 2. The conventional estimator of the finite population distribution function (cdf) and its associated binomial upper confidence bound for a sample of size 16 from a finite population.

design-based estimators of the cdf in the context of stratified sampling. Design-based estimators of the cdf have been used extensively by the U.S. Environmental Protection Agency in the Eastern Lakes Survey (Linthurst et al, 1986) and the National Stream Survey (Messer et al, 1988) as well as in the Environmental Monitoring and Assessment Program (Overton, Stevens, and White, 1990). In a seminal paper Chambers and Dunstan (1986) provided a model-based approach for estimation of the finite population cdf when a predictor variable is available. Rao, Kovar, and Mantel (1990) extended this approach to provide design-based estimators of the cdf when a predictor variable is available. Dorfman (1993) contrasted the rival estimators of the finite population cdf and concluded that, when standard regression methodology was employed for model selection, the model-based estimator was better than Rao, Kovar, and Mantel extension. Robust estimation of the cdf has been addressed by Kuo (1988), Chambers, Dorfman, and Wehrly (1993), Dorfman and Hall (1993), and Kuk (1993). Other recent papers include Kuk (1988) and Bolfarine and Sandoval (1993, 1994).

4.2 Estimation with a Predictor Variable

In this section it will be assumed that values of a predictor variable are known for every object in the finite population. The approach utilized is based on the method of Chambers and Dunstan (1986), utilizing a predictor variable in estimation of the cdf of the finite population. Chambers and Dunstan employed linear regression through the origin to model the relationship between the variable of interest and the predictor variable. They assumed that the variance of each value of the variable was a simple function of the

associated value of the predictor variable. Specifically, the following model was assumed by Chambers and Dunstan:

$$y_u = x_u \beta + \varepsilon_u^* \quad \text{for } u = 1, 2, \cdots, N$$
$$\varepsilon_u^* = g(x_u)\, \varepsilon_u$$

where $g(x_u)$ is a known positive function of x_u and the ε_u are *iid* random variables from a distribution function $G(\cdot)$ with mean zero.

Assume that *S* is a simple random sample of size n selected from the N objects in the universe, *U*. Based on the sample values of the variable and the predictor variable, linear regression through the origin is employed to calculate an estimator, $\hat{\beta}$, of the slope parameter of the regression model. Chambers and Dunstan (1986) proposed the following estimator of the finite population cdf:

$$\hat{F}_{CD}(t) = \frac{\sum_{u \in S} I_{(-\infty, t]}(y_u) + \sum_{u \in U-S} \left(\frac{1}{n} \sum_{v \in s} I_{(-\infty,\, (t - x_u \hat{\beta})/g(x_u)]}(e_v) \right)}{N}$$

where t is the threshold value, $I_A(x)$ is the indicator function for inclusion of x in the set A, and e_v are the sample residuals defined by:

$$e_v = \frac{y_v - x_v \hat{\beta}}{g(x_v)} \qquad \text{for } v \in S$$

Two modifications to the Chambers and Dunstan estimator were incorporated in our investigation. The first modification involved inclusion of an intercept term in the

regression model. The second modification concerned the form of the distribution employed for the residuals. Since the error distribution employed in the Chambers and Dunstan estimate was not specified, their procedure may be considered nonparametric. Rather than assume that the ε_u are *iid* random variables from an unspecified distribution $G(\cdot)$, we assume the ε_u to be *iid* random variables from a Normal(0, σ^2) distribution. Thus, the modified model used in our investigation is given by:

$$y_u = \beta_0 + x_u \beta_1 + \varepsilon_u^* \quad \text{for } u = 1, 2, \cdots, N$$
$$\varepsilon_u^* = g(x_u)\, \varepsilon_u$$

where $g(x_u)$ is a known positive function of x_u, and the ε_u are *iid* Normal(0, σ^2) random variables. Let $\hat{\beta}_0$ be the estimator of β_0 and $\hat{\beta}_1$ be the estimator of β_1 obtained using simple linear regression. Then, the modified estimator of the finite population distribution function is given by:

$$\hat{F}_N(t) = \frac{\sum_{u \in s} I_{(-\infty,\, t]}(y_u) + \sum_{u \in U-s} \Phi\left(\dfrac{t - \hat{y}_u}{\hat{\sigma} \cdot g(x_u)}\right)}{N}$$

$$\hat{y}_u = \hat{\beta}_0 + x_u \hat{\beta}_1$$

where $\hat{y}_u$ is the predicted value of y_u and $\hat{\sigma}$ is the estimator of σ calculated from the sample residuals given by:

$$e_v = \frac{y_v - \hat{y}_v}{g(x_v)} \qquad \text{for } v \in S$$

For any value t at which the cdf was evaluated, the predicted value for the cdf is given by:

$$\hat{F}_{Pred}(t) = \frac{\sum_{u \in S} I_{(-\infty, t]}(y_u) + \sum_{u \in U-S} I_{(-\infty, t]}(\hat{y}_u)}{N}$$

Thus, the predicted value is the estimate of the cdf obtained without adding error to the values $\hat{y}_u$.

The modified Chambers and Dunstan estimator, $\hat{F}_N(t)$, can be decomposed into two parts. The first part is composed of the summation across the sample objects of the indicator function that indicates whether an observed value of y_u is less than or equal to the threshold value t. The second part is composed of the summation across the non-sample objects of estimates of the probability, $P(y_u \leq t)$, that y_u is less than or equal to t. Let p_u equal $P(y_u \leq t)$; an estimate of the modeled probability p_u is given by:

$$\hat{p}_u = \hat{P}(y_u \leq t) = \Phi\left(\frac{t - \hat{y}_u}{\hat{\sigma} \cdot g(x_u)}\right).$$

Note that p_u may be interpreted as the success probability for a Bernoulli random variable. Thus, the second term in the modified Chambers and Dunstan estimator is a

sum of estimated probabilities for independent and not identically distributed Bernoulli random variables. Although an exact confidence bound is theoretically available for the Chambers and Dunstan estimator, it is computationally more tractable to employ a simulation approach to obtain the desired bound.

An upper confidence bound can be developed as follows. For each of a sequence of B simulations, conduct N-n Bernoulli trials with associated success probabilities $\hat{p}_u$, $u \in U-S$. Let k_b equal the sum of the values from the Bernoulli trails for a particular simulation b, where $1 \le b \le B$ and b is an integer. Order the B values of k_b, and let $\tilde{r}$ be the value of B multiplied by $(1-\alpha)$. Then define r as follows: r equals $\tilde{r}$ when $\tilde{r}$ is an integer, and r equals one plus the integer portion of $\tilde{r}$ when $\tilde{r}$ is not an integer. Conditional on the sample and on the model, a conservative $100(1-\alpha)\%$ upper confidence bound is provided by $(k_s + k_r)/N$, where k_s is the number of sample values less than or equal to the threshold value and k_r is the r^{th} value among the values in the ordered set: $\{k_b: b = 1, \cdots, B\}$. Due to the discrete nature of the procedure, the conservative upper bound will have a nominal coefficient no smaller than $100(1 - \alpha)\%$ given that the sample is observed without error and the assumed model is correct.

Two additional upper confidence bounds for the cdf can be developed. Let k_q equal either k_r-1 or 0, whichever value is larger. Consider the two proportions: $Pr(k \le k_q)$ and $Pr(k \le k_r)$, where the proportions are calculated from the values in $\{k_b\}$. The first proportion is anti-conservative in nature whereas the second proportion

is conservative. The observed values of the proportions $\mathbf{Pr(k \leq k_q)}$ and $\mathbf{Pr(k \leq k_r)}$ in conjunction with the desired confidence coefficient, $(1 - \alpha)$, can be used to produce a randomized upper confidence bound having the desired proportion. That is, a random trial is conducted to select either k_q or k_r, where the values of $\mathbf{Pr(k \leq k_q)}$, $\mathbf{Pr(k \leq k_r)}$, and $(1 - \alpha)$ are used to specify the parameters for the random trial. In repeated application of the randomization procedure, the mean of the percentage of values in the set $\{\mathbf{k_b}\}$ that are less than or equal to the value selected by the randomization procedure will be exactly equal to $100(1 - \alpha)\%$.

The values of $\mathbf{Pr(k \leq k_q)}$, $\mathbf{Pr(k \leq k_r)}$, and $(1 - \alpha)$ can also be used to create weights for $\mathbf{(k_s + k_q)/N}$ and $\mathbf{(k_s + k_r)/N}$ that can be employed to produce a weighted average upper confidence bound. The weight for $\mathbf{(k_s + k_q)/N}$ is given by:

$$\frac{\mathbf{Pr(k \leq k_r)} - (1 - \alpha)}{\mathbf{Pr(k \leq k_r)} - \mathbf{Pr(k \leq k_q)}}$$

The weight for $\mathbf{(k_s + k_r)/N}$ is given by one minus the weight for $\mathbf{(k_s + k_q)/N}$. The weighted average bound represents the average value of the randomized bound that would be achieved during repeated application of the randomization procedure. Due to greater precision, the weighted average bound is preferred over the randomized bound in practice. Coverage of the weighted average bound, however, will be equal to the coverage achieved by the smaller of the two values used in calculating the bound. Therefore, coverage of

the weighted average bound should be estimated from coverage of the randomized bound and not from coverage of the weighted average bound.

In addition to simple linear regression, two other procedures were employed to calculate estimators of the parameters of the regression model for predicting y_u: robust locally weighted least squares (lowess: Cleveland, 1979) and piecewise simple linear regression. A description of the piecewise simple linear regression procedure follows. First, the objects in the sample were ordered based on values of the predictor variable. The ordered objects were then divided into three groups of equal size (or as close to equal size as allowed by the size of the sample) such that the first group contained the smallest values of the predictor variable, the second group contained the middle-sized values of the predictor variable, and the third group contained the largest values of the predictor variable. In addition the objects were allocated to the groups such that the first and second group had a single object in common, and the second and third group had a single object in common. The latter condition was imposed to eliminate ambiguity of the prediction procedure for y_u. Simple linear regression was applied independently to each of the three groups to produce three sets of estimates of the regression model parameters.

Recall that estimates of the regression model parameters are used to calculate predicted values for y_u. When piecewise simple linear regression was employed to produce the parameter estimates, the Chambers and Dunstan procedure was modified to allow usage of the three sets of parameter estimates for predicting y_u. If a non-sampled value of the predictor variable was less than or equal to the value of the predictor variable

of the ordered sample object held in common by groups one and two, then the first set of regression model parameter estimates was employed in calculating the predicted value for that non-sampled object. If a non-sampled value of the predictor variable was greater than or equal to the value of the predictor variable of the ordered sample object held in common by groups two and three, then the third set of parameters estimates was employed to calculate the predicted value. Otherwise, the second set of parameter estimates was used for prediction.

4.3 Estimation without a Predictor Variable

In this section it will be assumed that a predictor variable is not available. For this case a model for the finite population variable will be assumed. Based on sample values of the variable, the parameters of the assumed model will be estimated. Then, conditional on the sample and on the model, the overall proportion, say $\hat{p}$, less than the threshold value will be estimated. An analytical solution employing the Binomial distribution with success probability $\hat{p}$ can then be employed to produce the desired point estimator and upper confidence bound for the population cdf evaluated at the threshold value. Let k_s be the number of values in the sample less than or equal to the threshold value. Then, under the Binomial model, the expected number of non-sampled values less than or equal to the threshold is given by: $(N-n) * \hat{p}$. Thus, a point estimator of the cdf evaluated at the threshold value, t, is given by:

$$\hat{F}(t) = \frac{k_s + (N-n) * \hat{p}}{N}$$

Conditional on the model, let k_u be the smallest integer that satisfies the following inequality:

$$P\left(\hat{F}(t) \le \frac{k_u}{N} \mid k_s\right) = \sum_{k=k_s}^{k_u} \binom{N-n}{k-k_s} \hat{p}^{\,k-k_s} (1-\hat{p})^{(N-n)-(k-k_s)} \ge (1-\alpha)$$

Then the conservative 100(1-α)% upper confidence bound is given by k_u/N. Given that the assumed model is correct, the conservative upper bound will have a confidence coefficient no smaller than (1-α)%. As discussed in the previous section, the observed values of the probabilities:

$$P\left(\hat{F}(t) \le \frac{k_u}{N} \mid k_s\right) \quad \text{and} \quad P\left(\hat{F}(t) \le \frac{k_u - 1}{N} \mid k_s\right)$$

in conjunction with the desired confidence coefficient, (1-α), can be used to produce randomized and weighted average upper confidence bounds. As mentioned previously, due to greater precision, the weighted average bound is preferred over the randomized bound in practice.

Three specific models were investigated: (1) a Normal model, (2) a Normal model censored at the median, i.e., a half Normal model (see Johnson and Kotz, 1970), and (3) a Gamma model. For the half Normal model, the sample was ordered and values less than the median of the sample were employed to estimate model parameters using the standard procedures for a censored Normal distribution. For the Gamma model a three-parameter Gamma distribution, i.e. a distribution with location, scale, and shape

parameters, was employed. The approach suggested by Bowman and Shenton (1988) was utilized to estimate the parameters of the Gamma model.

4.4 Preliminary Simulation Results and Discussion

The estimation procedures were applied to a group of standardized populations: PADDY, STREAM, DATAA, DATAB, DATAC, DATAG, DATAGNB, and DATAUNB. Each population was composed of one hundred objects for each of which there was a response variable value and an associated value of a predictor variable. The population correlation, ρ, between the variable and the predictor variable varied among the populations as follows: PADDY (ρ = 0.79), STREAM (ρ = 0.86), DATAA (ρ = 0.77), DATAB (ρ = 0.97), DATAC (ρ = 0.75), DATAG (ρ = 0.80), DATAGNB (ρ = 0.79), and DATAUNB (ρ = 0.84). In each of the populations and for both the response variable and the predictor variable, the mean was two and the standard deviation was one. Descriptions of PADDY and STREAM are provided in Stehman and Overton (1994). The other populations are described in Overton and Stehman (1993).

For the purpose of preliminary analysis, one hundred samples of size sixteen (i.e., n = 16) were selected from each of the eight populations, and the three predictor variable procedures and three procedures without a predictor variable were applied to the samples. For each of the predictor variable procedures, three values of the function g(x) were considered: $g(x) = 1$, $g(x) = x$, and $g(x) = x^2$. For populations in which values of both the response variable and the predictor variable were strictly positive, a fourth case was considered that consisted of taking the natural logarithm of the response variable and

predictor variable values and applying the procedures using g(x) = 1, which will be referenced as the Log case. For all three of the predictor variable procedures, one hundred simulations were employed for determination of the weighted average confidence bounds. For the procedures without a predictor variable, a second case was considered for the Normal and half Normal models, which consisted of taking the natural logarithm of the response variable values and applying the procedures. The initial ordered value of the response variable in each of the populations was utilized to evaluate performance of the procedures. Since each population was composed of one hundred values, the true value of the cdf evaluated at the initial ordered value in the population was equal to 0.01. For calculating the upper confidence bound, α was equal to 0.10, producing 90% upper bounds. For each configuration of the procedures, sample means of the cdf estimates and the upper confidence bound estimates were determined. In addition the confidence bound coverage, i.e., the proportion of the confidence bounds that included the known true value 0.01, was calculated for each configuration of the procedures.

Each of the eight populations will be discussed separately. For each population a table displaying the means of the cdf estimates, standard deviations of the cdf estimates, means of the weighted average 90% upper confidence bounds, and coverage of the upper confidence bounds is provided. For purposes of discussion, acceptable bias will be taken to mean that bias was no greater than 0.02. Adequate confidence bound coverage will be taken to mean that coverage was between 80% and 95%, inclusive. In order to provide graphical illustration of results, plots of the actual cdf, means of the cdf estimates, means of the weighted average 90% upper confidence bounds, and means of the predicted

values are provided for the piecewise simple linear regression procedure using $g(x) = 1$. In producing the plots, the cdf was evaluated at a set of values ranging from a number much less than the smallest ordered value in the population through the sixth ordered value for the population. Finally, for reasons that will be developed in the following discussion, special attention will be given to the $g(x) = 1$ version of the piecewise simple linear regression procedure.

For PADDY the conditional distribution of the response variable given the predictor variable exhibits a large increase in variability as the value of the predictor variable increases for the natural scale and a minor increase in variability as the value of the predictor variable increases for the log scale (Figure 3). The first six ordered values in PADDY are: 0.812, 0.856, 0.885, 0.928, 0.957, and 0.994. Results for PADDY are provided in Table 7. Acceptable bias of the estimates was achieved by the Log case of all three predictor variable procedures in addition to the half Normal model using the log scale and the Gamma model. Acceptable coverage was achieved by the Log case of the robust regression procedure, the $g(x) = x$ and $g(x) = x^2$ cases of the piecewise regression procedure, and the half Normal model using both the natural and log scales. Note that coverage for the $g(x) = 1$ version of the piecewise linear regression procedure was 96%, which is marginally greater than the acceptable range. Plots of the actual cdf, means of the estimated cdf, means of the weighted average confidence bounds, and means of the predicted values are provided in Figure 4 for the $g(x) = 1$ version of the piecewise linear regression procedure. For this procedure the estimated cdf exhibits significant positive bias for all values in the lower tail of the cdf for PADDY.

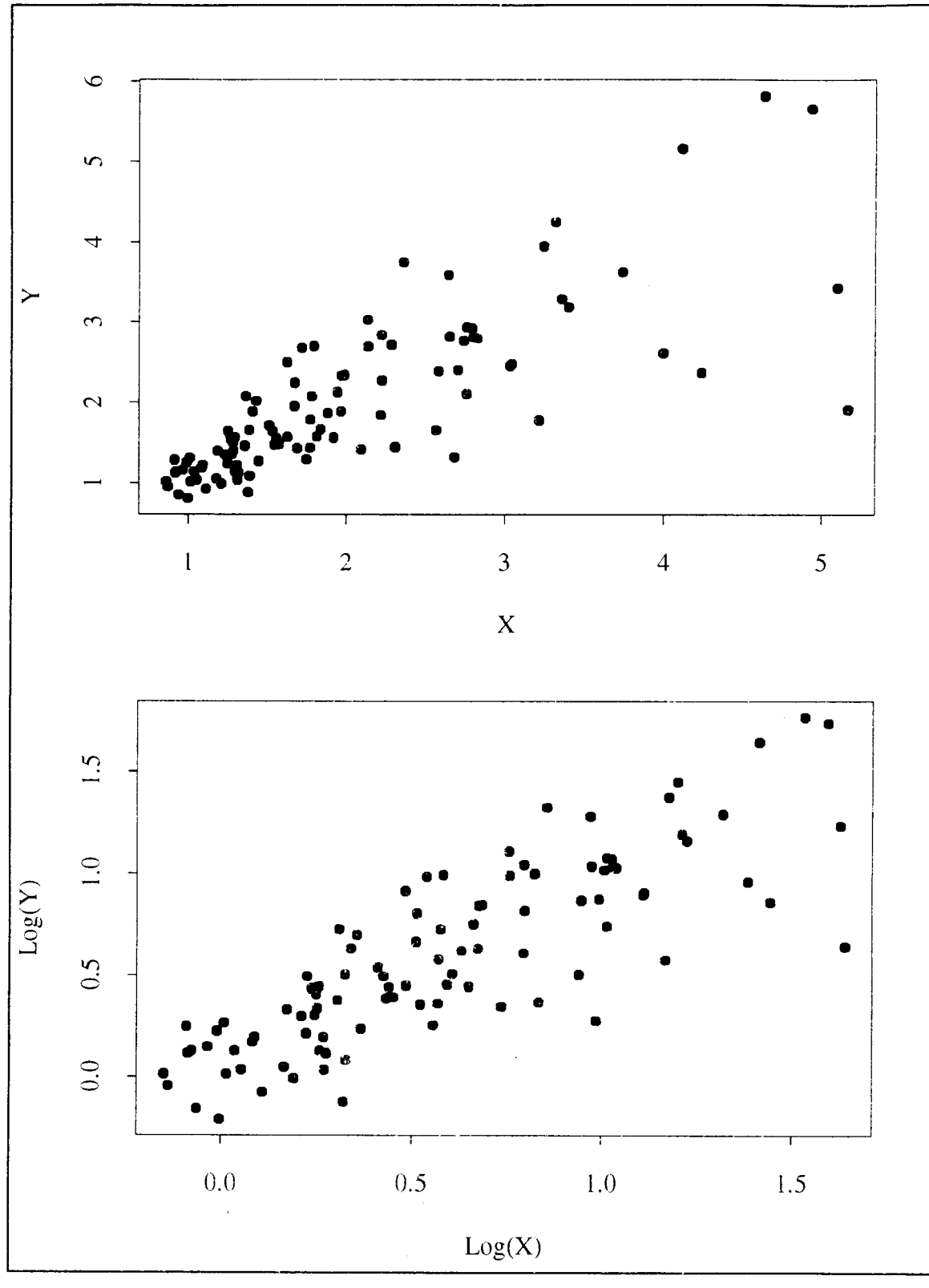

Figure 3. Scatter plots of Y versus X using the natural and log scales for population PADDY.

Table 7. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population PADDY, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0710	0.0356	0.0964	100%
$g(x) = x$	0.0292	0.0210	0.0448	97%
$g(x) = x^2$	0.0472	0.0318	0.0678	99%
Log	0.0211	0.0144	0.0336	96%
Robust Regression:				
$g(x) = 1$	0.0791	0.0423	0.1050	100%
$g(x) = x$	0.0318	0.0222	0.0482	98%
$g(x) = x^2$	0.0349	0.0242	0.0521	96%
Log	0.0239	0.0173	0.0370	94%
Piecewise Linear:				
$g(x) = 1$	0.0407	0.0325	0.0568	96%
$g(x) = x$	0.0333	0.0325	0.0472	95%
$g(x) = x^2$	0.0371	0.0352	0.0521	95%
Log	0.0219	0.0272	0.0309	77%
Normal:				
Natural	0.0893	0.0397	0.1201	100%
Log	0.0321	0.0223	0.0489	97%
Half Normal:				
Natural	0.0222	0.0172	0.0352	91%
Log	0.0142	0.0142	0.0231	80%
Gamma:				
Natural	0.0085	0.0146	0.0132	42%

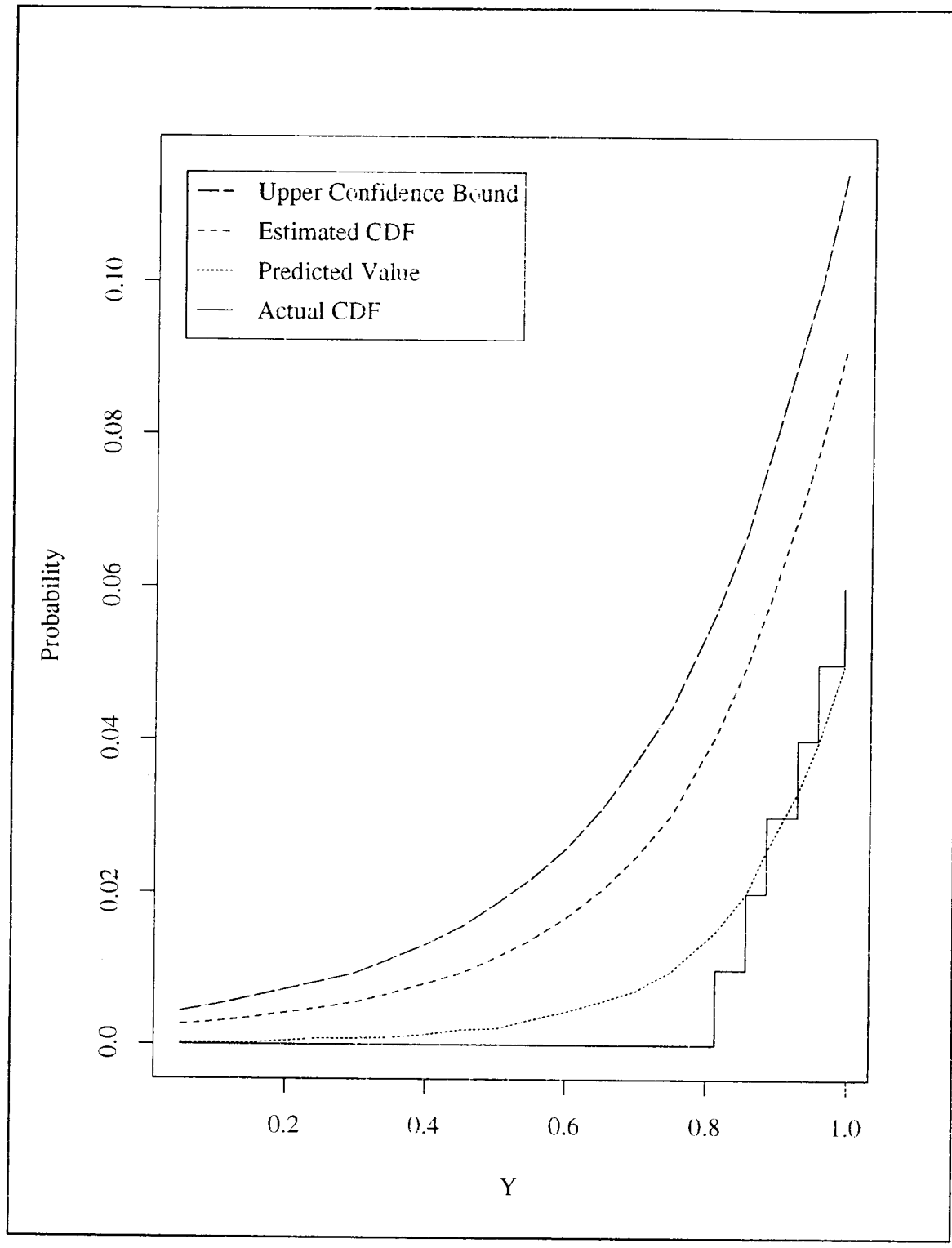

Figure 4. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population PADDY, where 100 replications were used in the simulations.

For STREAM the conditional distribution of the response variable given the predictor variable exhibits a large increase in variability as the value of the predictor variable increases for the natural scale and a minor increase in variability as the value of the predictor variable increases for the log scale (Figure 5). The first six ordered values in STREAM are: 0.805, 0.831, 0.866, 1.010, 1.013, and 1.045. Results for STREAM are provided in Table 8. Acceptable bias of the estimates was achieved by all procedures except the $g(x) = 1$ case of the linear and robust regression procedures and the Normal model using the natural scale. Acceptable coverage was achieved by the $g(x) = x^2$ case for all three predictor variable procedures, the $g(x) = 1$ case of the piecewise regression procedure, and the Normal model using the log scale. Plots of the actual cdf, means of the estimated cdf, means of the weighted average confidence bounds, and means of the predicted values are provided in Figure 6 for the $g(x) = 1$ version of the piecewise regression procedure. Although the estimated cdf is moderately biased for the initial ordered value in STREAM, the estimated cdf performs very well for the other values in the lower tail of STREAM.

For DATAA the conditional distribution of the response variable given the predictor variable exhibits a moderate increase in variability as the value of the predictor variable increases for the natural scale and no increase in variability as the value of the predictor variable increases for the log scale (Figure 7). The first six ordered values in DATAA are 0.643. 0.656, 0.721, 0.777, 0.826, and 0.888. Results are provided in Table 9. Acceptable bias was achieved by the Log case of all three predictor variable procedures and all cases of the procedures without a predictor variable except the Normal

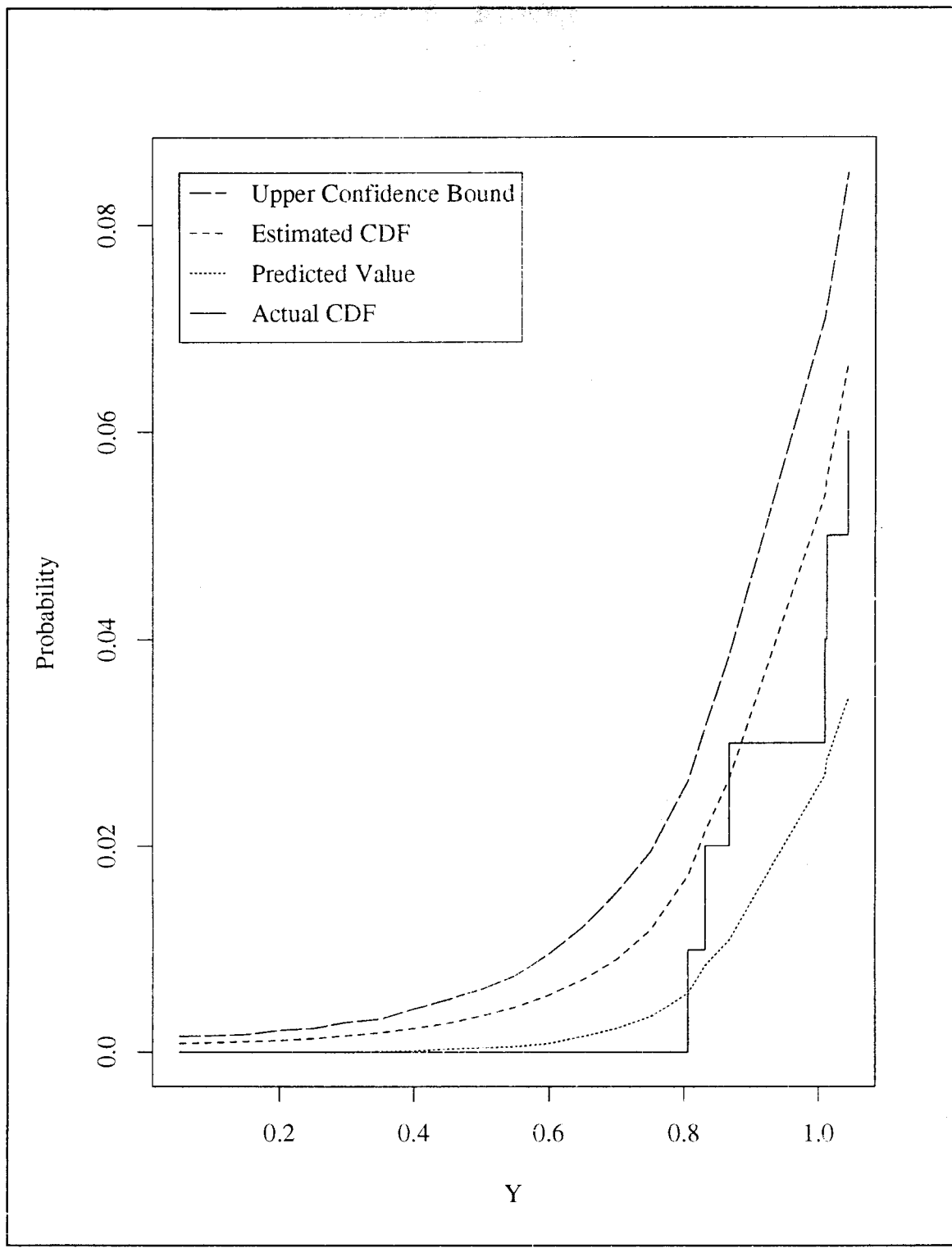

Figure 5. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population STREAM, where 100 replications were used in the simulations..

Table 8. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population STREAM, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0456	0.0400	0.0642	98%
$g(x) = x$	0.0153	0.0246	0.0237	69%
$g(x) = x^2$	0.0218	0.0268	0.0337	90%
Log	0.0075	0.0104	0.0127	64%
Robust Regression:				
$g(x) = 1$	0.0452	0.0407	0.0643	99%
$g(x) = x$	0.0117	0.0137	0.0193	73%
$g(x) = x^2$	0.0160	0.0146	0.0269	90%
Log	0.0066	0.0084	0.0112	57%
Piecewise Linear:				
$g(x) = 1$	0.0171	0.0150	0.0263	85%
$g(x) = x$	0.0137	0.0137	0.0210	74%
$g(x) = x^2$	0.0165	0.0158	0.0251	81%
Log	0.0085	0.0108	0.0123	53%
Normal:				
Natural	0.0856	0.0432	0.1156	100%
Log	0.0228	0.0181	0.0362	92%
Half Normal:				
Natural	0.0136	0.0126	0.0227	79%
Log	0.0080	0.0099	0.0133	61%
Gamma:				
Natural	0.0044	0.0095	0.0070	30%

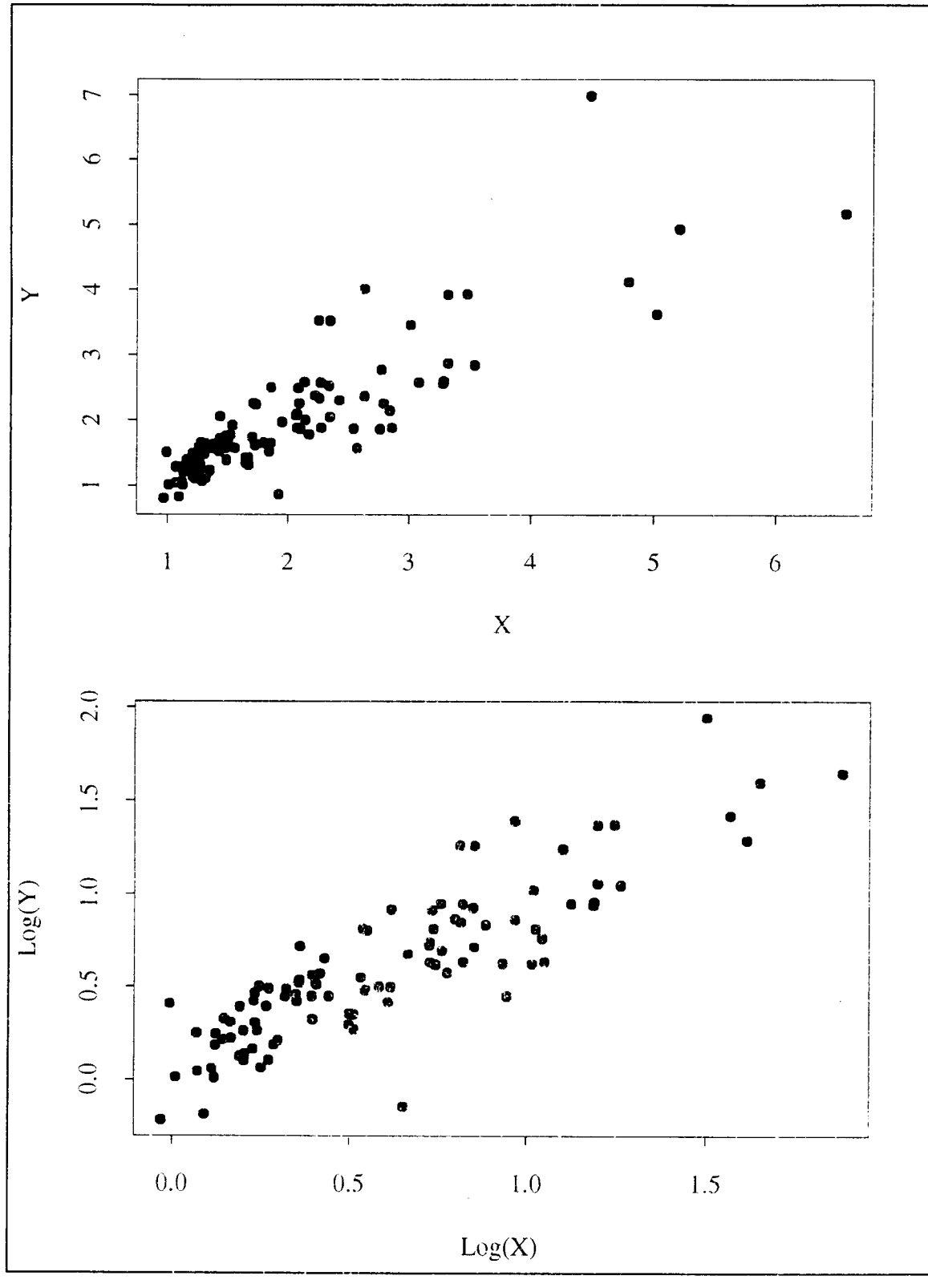

Figure 6. Scatter plots of Y versus X using the natural and log scales for population STREAM.

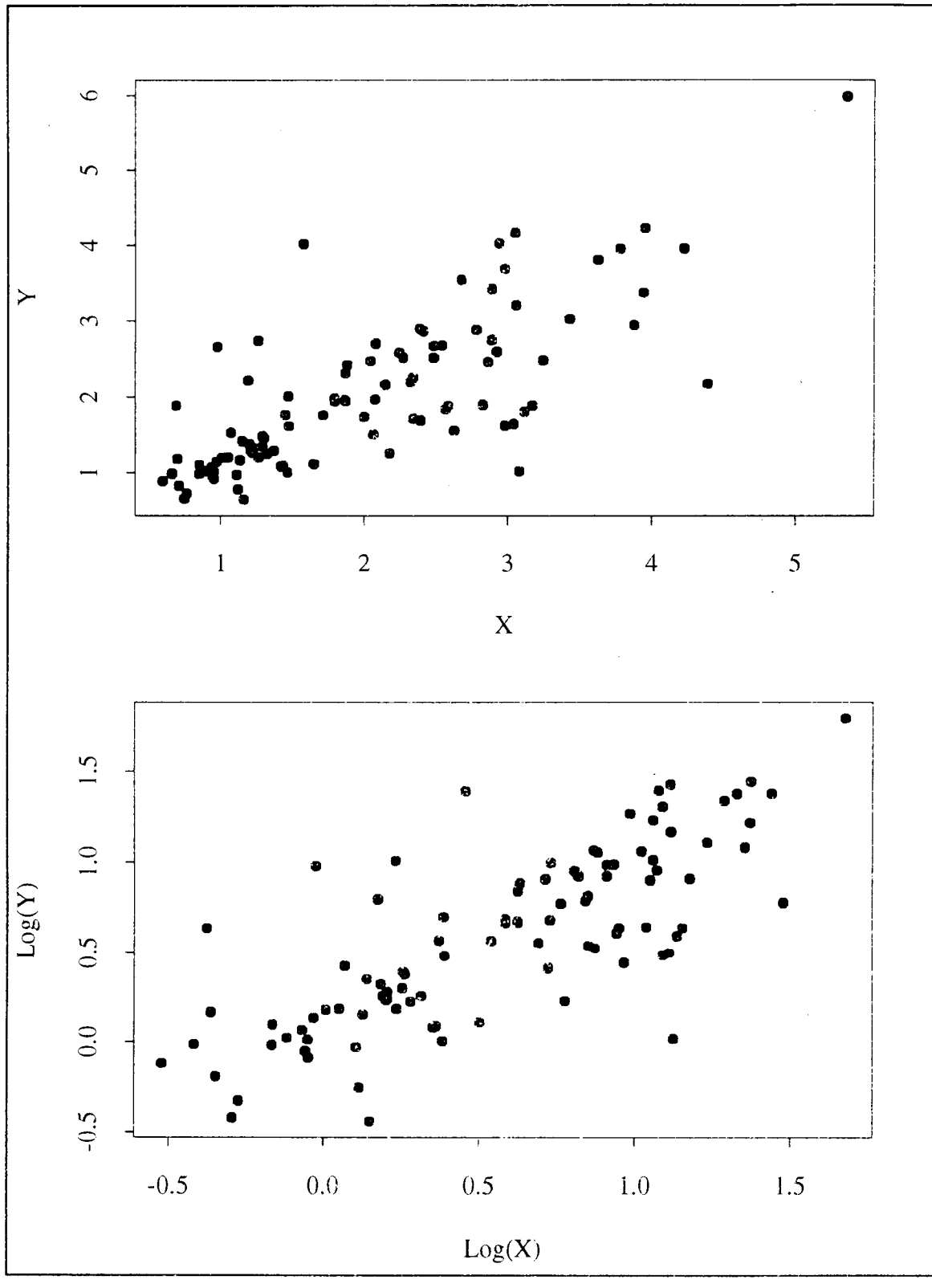

Figure 7. Scatter plots of Y versus X using the natural and log scales for population DATAA.

Table 9. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population DATAA, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0613	0.0277	0.0851	100%
$g(x) = x$	0.0430	0.0293	0.0622	97%
$g(x) = x^2$	0.1037	0.0604	0.1335	99%
Log	0.0153	0.0125	0.0250	88%
Robust Regression:				
$g(x) = 1$	0.0736	0.0328	0.0993	100%
$g(x) = x$	0.0559	0.0426	0.0778	96%
$g(x) = x^2$	0.1071	0.0733	0.1368	97%
Log	0.0199	0.0141	0.0315	91%
Piecewise Linear:				
$g(x) = 1$	0.0413	0.0272	0.0583	96%
$g(x) = x$	0.0396	0.0281	0.0557	93%
$g(x) = x^2$	0.0473	0.0327	0.0649	96%
Log	0.0186	0.0182	0.0277	80%
Normal:				
Natural	0.0719	0.0319	0.0997	100%
Log	0.0174	0.0152	0.0284	92%
Half Normal:				
Natural	0.0198	0.0147	0.0321	92%
Log	0.0098	0.0107	0.0164	68%
Gamma:				
Natural	0.0060	0.0114	0.0096	34%

model using the natural scale. Adequate coverage was achieved by the Log case of all three predictor variable procedures, the g(x) = x case of the piecewise regression procedure, the Normal model using the log scale, and the half Normal model using the natural scale. Note that coverage for the g(x) = 1 version of the piecewise regression procedure was 96%. Figure 8 provides plots for the g(x) = 1 version of the piecewise regression procedure. For this procedure the estimated cdf exhibits significant positive bias for all values in the lower tail of the cdf for DATAA.

For DATAB the conditional distribution of the response variable given the predictor variable exhibits no increase in variability as the value of the predictor variable increases for both the natural and log scales (Figure 9). In addition there is a very strong linear relationship between the response variable and the predictor variable for both the natural and log scales. Note, however, that values of the response variable are more evenly spaced on the log scale in comparison to the natural scale. The first six ordered values in DATAB are 0.677, 0.681, 0.692, 0.709, 0.764, and 0.765. Results are provided in Table 10. Acceptable bias was achieved by the g(x) = x, g(x) = x^2, and Log cases of all three predictor variable procedures, the g(x) = 1 case of the piecewise regression procedure, and all cases of the procedures without a predictor variable except the Normal model using the natural scale. Adequate coverage was achieved by the g(x) = x and g(x) = x^2 cases of the robust and piecewise regression procedures, the Log case of the piecewise regression procedure, the Normal model using the log scale, and the half Normal model using both the natural and log scale. Coverage for the g(x) = 1 version of the piecewise regression procedure was 97%. Plots are provided in Figure 10 for the

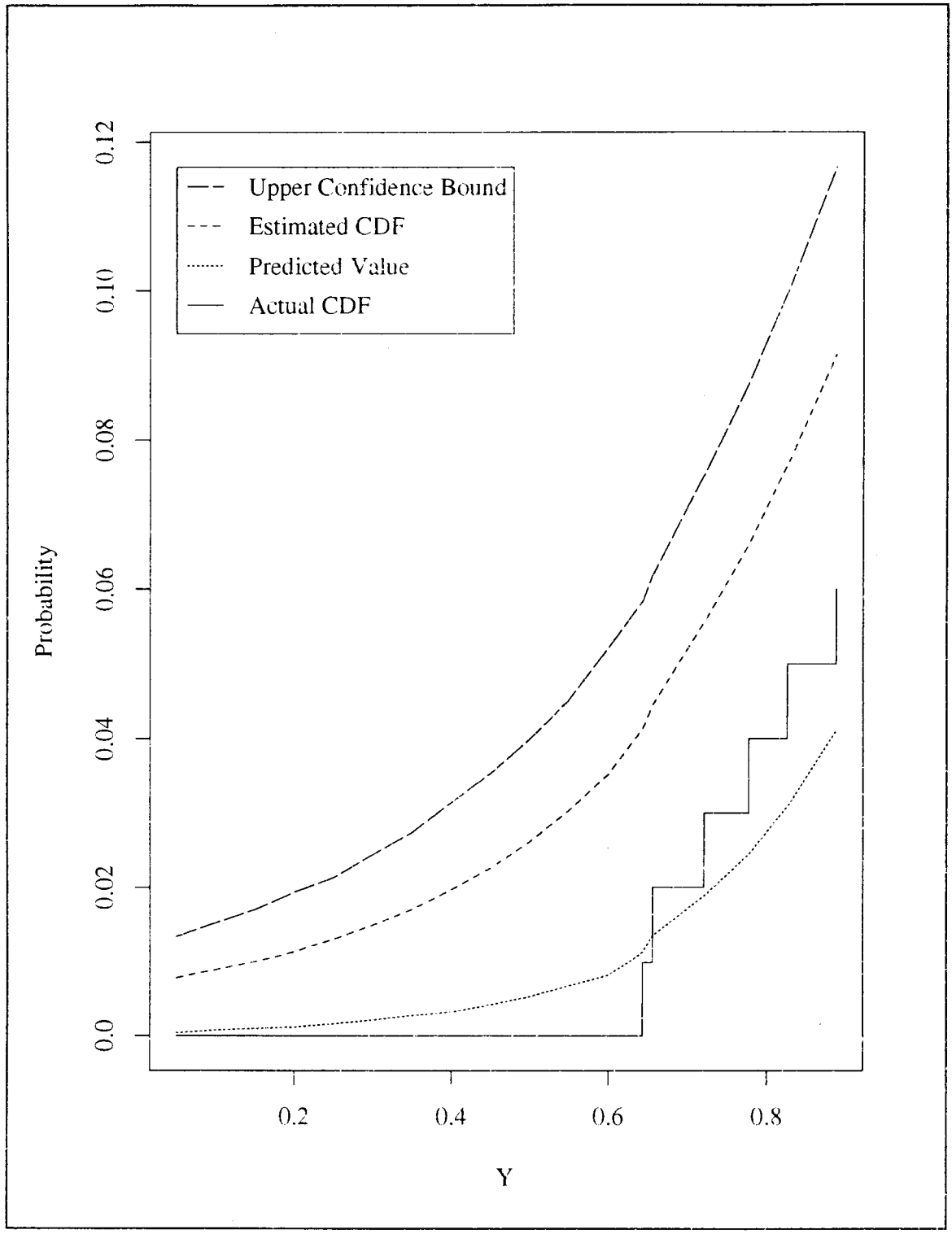

Figure 8. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population DATAA, where 100 replications were used in the simulations.

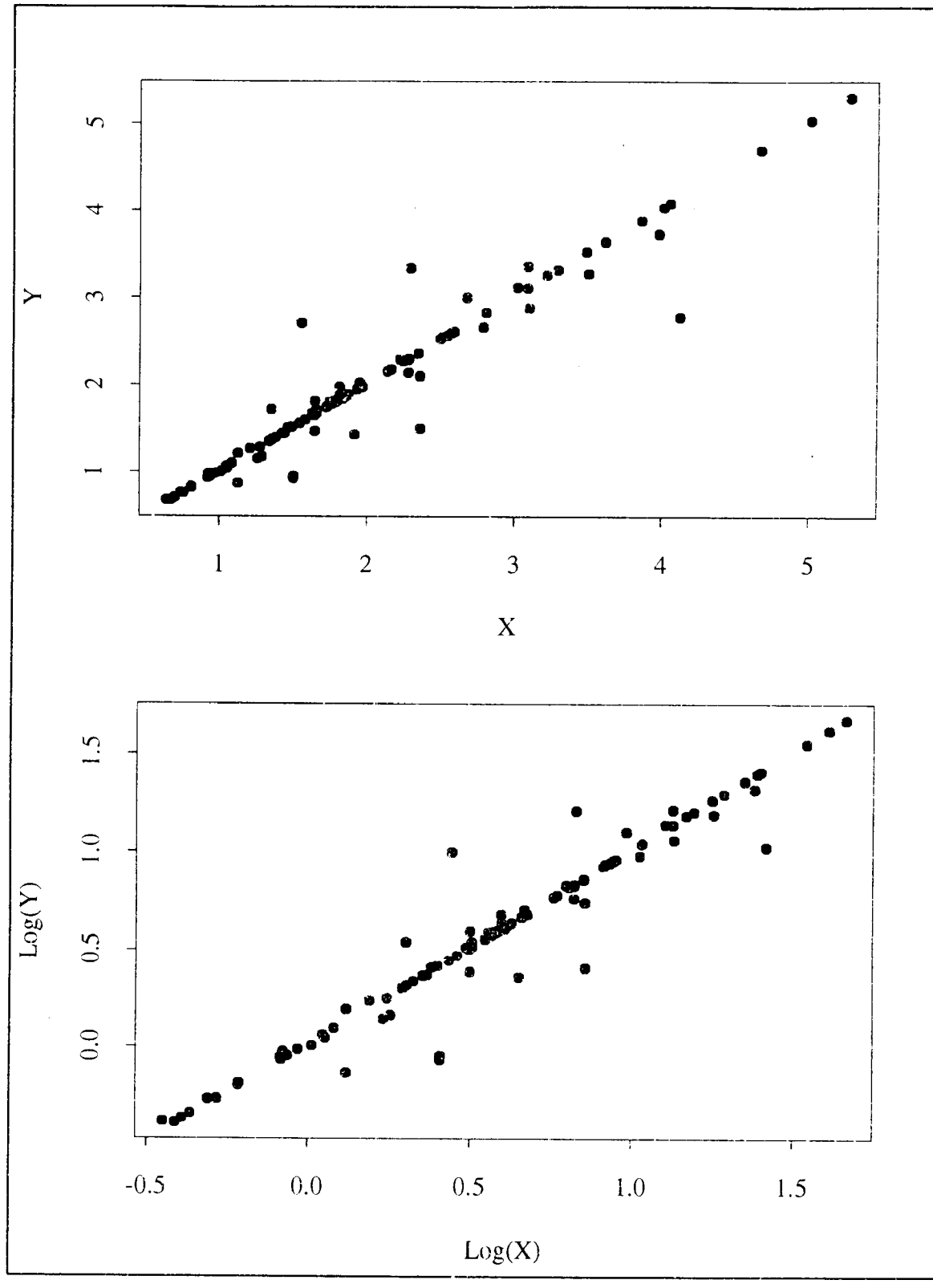

Figure 9. Scatter plots of Y versus X using the natural and log scales for population DATAB.

Table 10. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population DATAB, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0327	0.0146	0.0467	100%
$g(x) = x$	0.0213	0.0166	0.0295	89%
$g(x) = x^2$	0.0217	0.0172	0.0281	90%
Log	0.0220	0.0121	0.0310	99%
Robust Regression:				
$g(x) = 1$	0.0349	0.0135	0.0500	100%
$g(x) = x$	0.0212	0.0087	0.0314	99%
$g(x) = x^2$	0.0178	0.0094	0.0254	96%
Log	0.0207	0.0070	0.0302	99%
Piecewise Linear:				
$g(x) = 1$	0.0280	0.0209	0.0390	97%
$g(x) = x$	0.0237	0.0207	0.0315	87%
$g(x) = x^2$	0.0218	0.0215	0.0272	83%
Log	0.0205	0.0150	0.0282	92%
Normal:				
Natural	0.0767	0.0323	0.1055	100%
Log	0.0245	0.0201	0.0387	93%
Half Normal:				
Natural	0.0300	0.0217	0.0462	95%
Log	0.0197	0.0191	0.0314	85%
Gamma:				
Natural	0.0083	0.0148	0.0133	44%

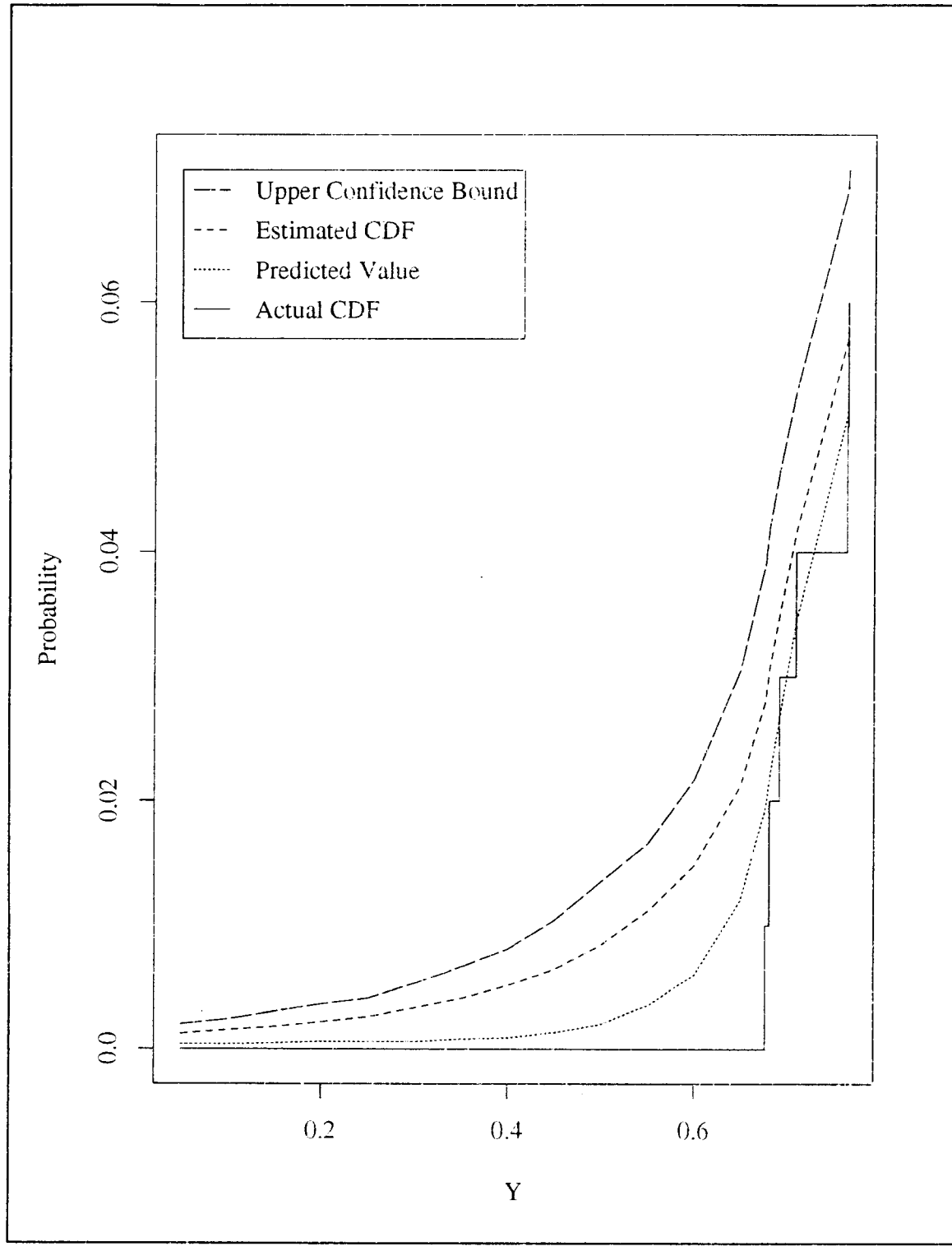

Figure 10. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population DATAB, where 100 replications were used in the simulations.

g(x) = 1 version of the piecewise regression procedure. Given the small spread among the values in the lower tail of the cdf for DATAB, the estimated cdf performs very well for this procedure.

For Population DATAC the conditional distribution of the response variable given the predictor variable exhibits a moderate decrease in variability as the value of the predictor variable increases for the natural scale (Figure 11). Due to the occurrence of negative values, DATAC was not analyzed using the log scale or the Gamma model. The first six ordered values in DATAA are -0.317, -0.163, -0.116, 0.296, 0.304, 0.354. Results are provided in Table 11. Acceptable bias was achieved by the g(x) = 1 and g(x) = x cases of all three predictor variable procedures, the g(x) = x^2 case of the piecewise regression procedure, and the Normal and half Normal models. None of the procedures provided adequate coverage. Coverage for the piecewise regression procedure using g(x) = 1 was 57%. Plots of the actual cdf, means of the estimated cdf, means of the weighted average confidence bounds, and means of the predicted values are provided in Figure 12 for the g(x) = 1 version of the piecewise linear regression procedure. For this procedure the cdf was virtually unbiased for the smallest value in the lower tail of the cdf and exhibited small negative bias for the other values in the lower tail of the cdf for DATAC.

For DATAG the conditional distribution of the response variable given the predictor variable exhibits a small increase in variability as the value of the predictor variable increases for the natural scale and no increase in variability as the value of the predictor variable increases for the log scale (Figure 13). The first six ordered values in

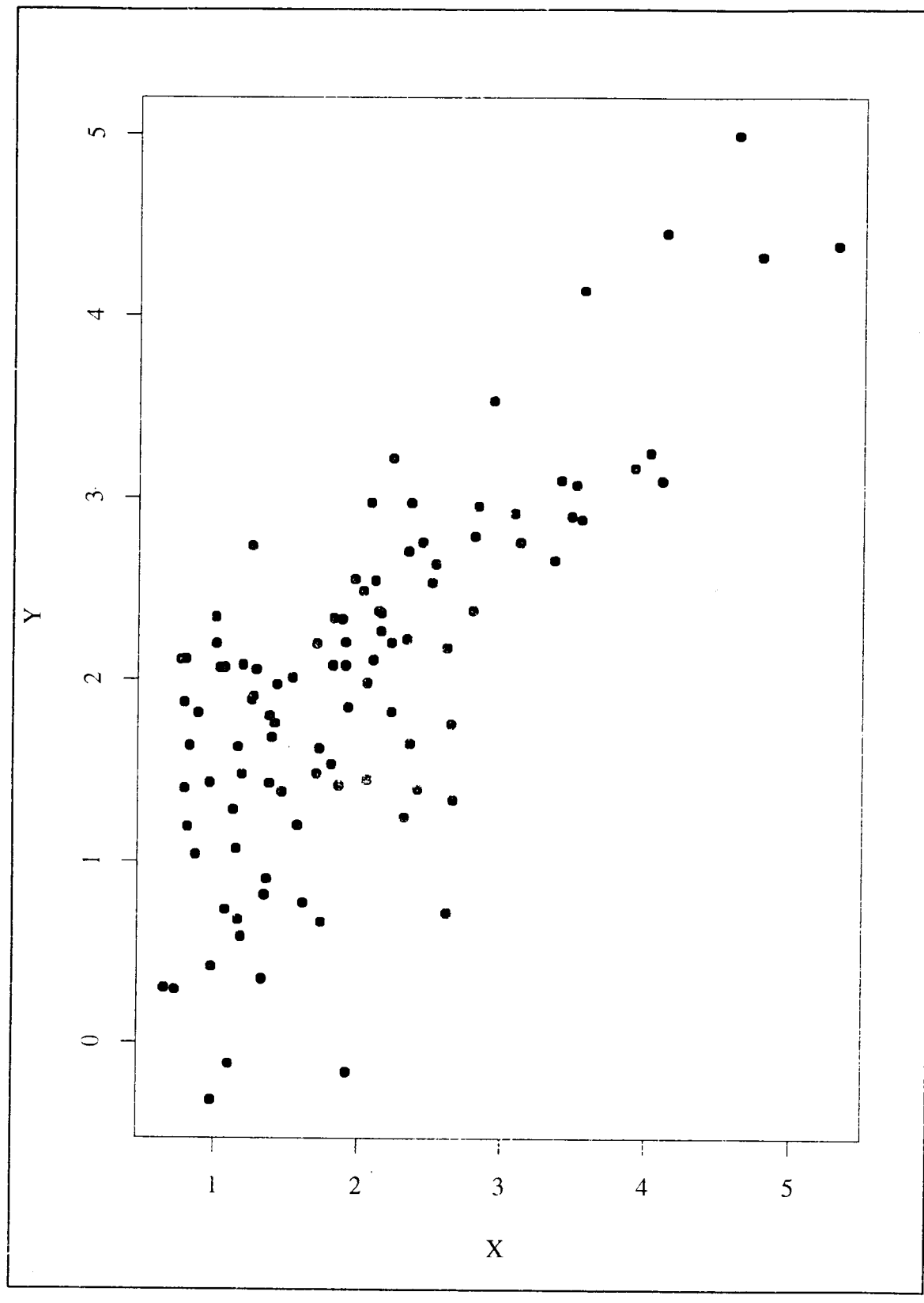

Figure 11. Scatter plot of Y versus X using the natural scale for population DATAC.

Table 11. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population DATAC, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0048	0.0068	0.0079	43%
$g(x) = x$	0.0119	0.0136	0.0202	73%
$g(x) = x^2$	0.0834	0.0458	0.1104	99%
Robust Regression:				
$g(x) = 1$	0.0036	0.0057	0.0057	32%
$g(x) = x$	0.0111	0.0133	0.0189	67%
$g(x) = x^2$	0.0779	0.0484	0.1033	99%
Piecewise Linear:				
$g(x) = 1$	0.0099	0.0214	0.0156	57%
$g(x) = x$	0.0119	0.0224	0.0186	61%
$g(x) = x^2$	0.0212	0.0282	0.0316	78%
Normal:				
Natural	0.0104	0.0118	0.0176	76%
Half Normal:				
Natural	0.0085	0.0101	0.0143	60%

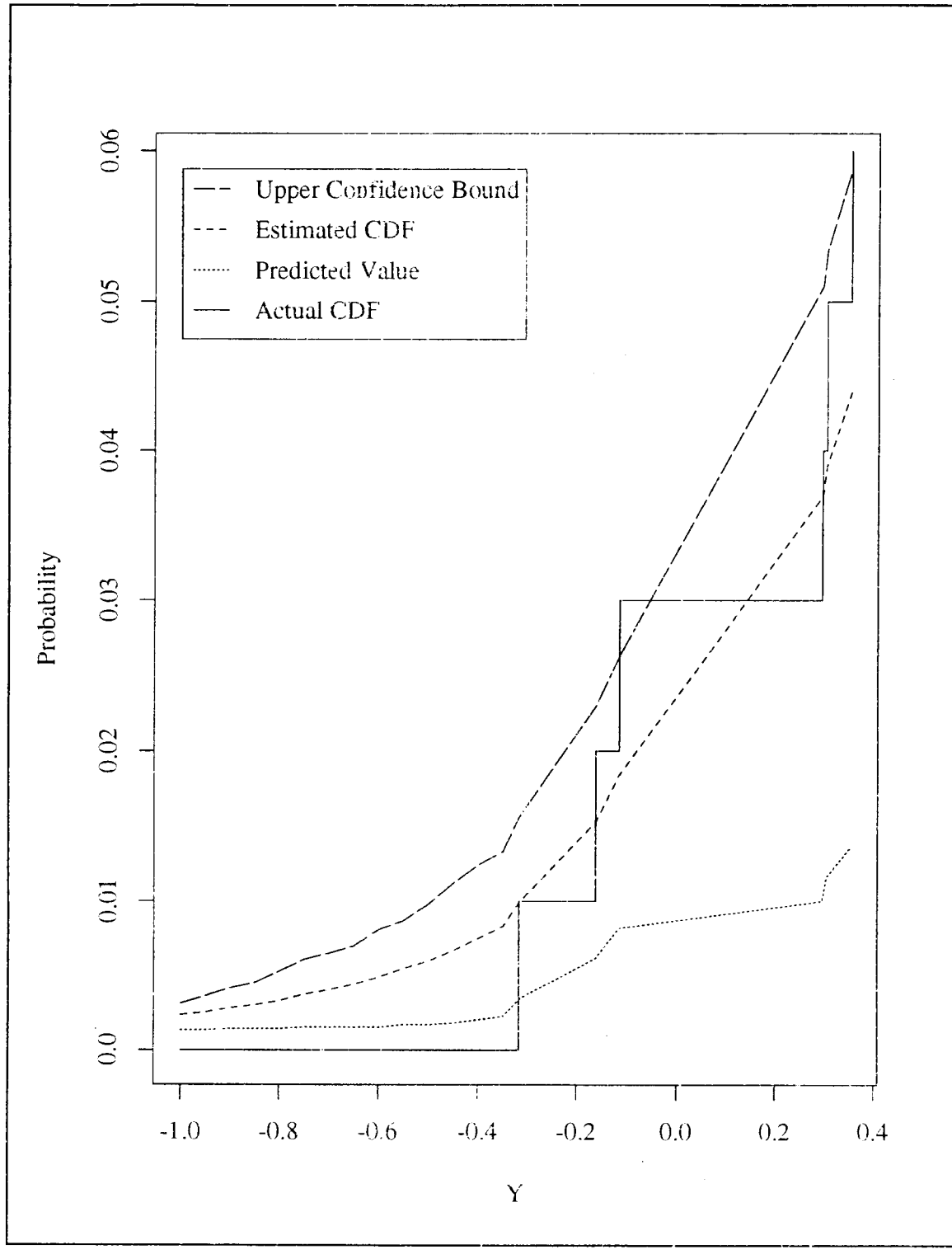

Figure 12. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population DATAC, where 100 replications were used in the simulations.

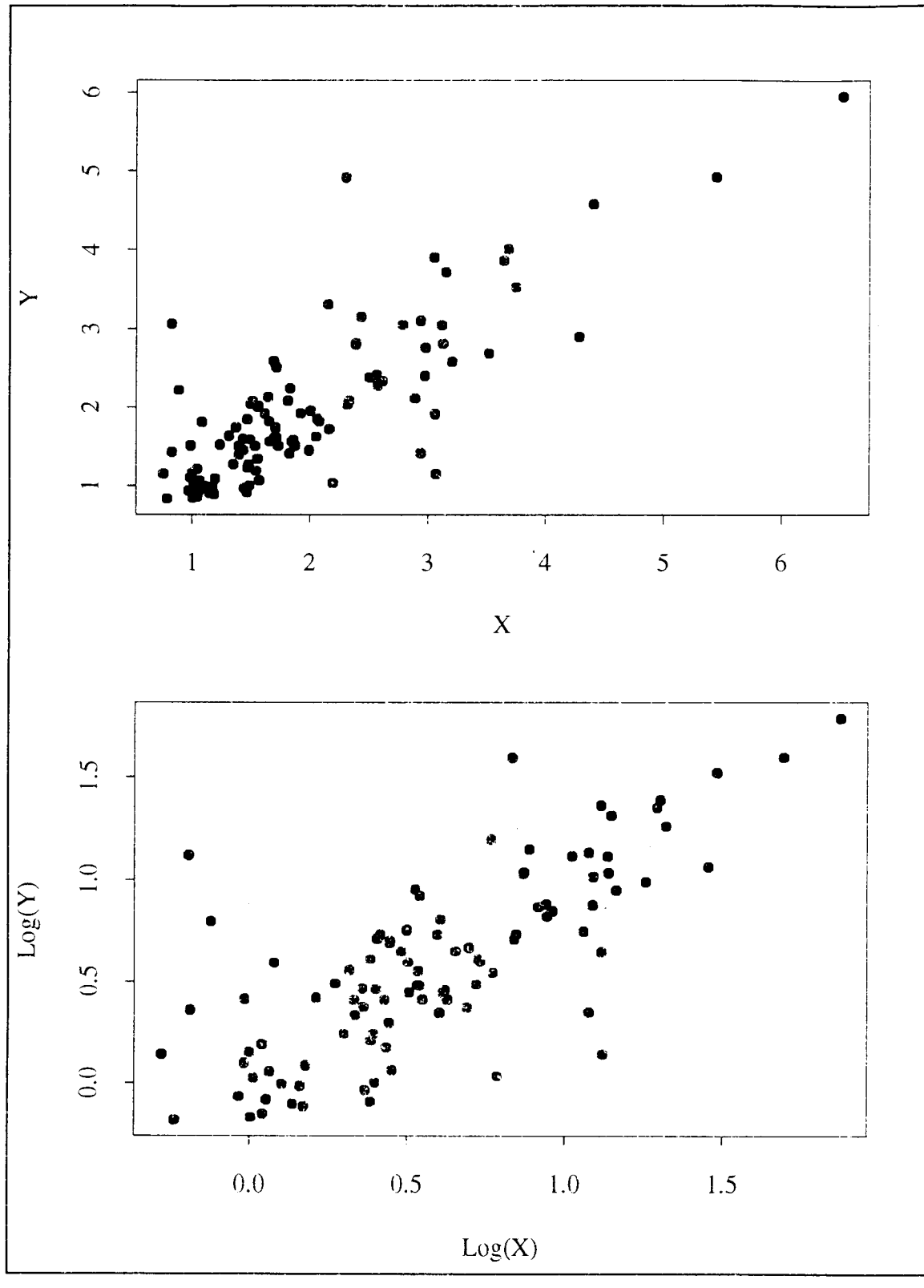

Figure 13. Scatter plots of Y versus X using the natural and log scales for population DATAG.

DATAG are 0.833, 0.844, 0.859, 0.891, 0.902, and 0.912. Results are provided in Table 12. Only the half Normal model using the log scale and the Gamma model achieved acceptable bias. Acceptable coverage was achieved by the half Normal model using the log scale. Although none of the predictor variable procedures achieved acceptable coverage, the Log case of the piecewise regression procedure was just outside the acceptable range (96%). Plot for the g(x) = 1 version of the piecewise regression procedure are provided in Figure 14. The estimated cdf produced extensive positive bias for this procedure.

For DATAGNB the conditional distribution of the response variable given the predictor variable exhibits a large increase in variability as the value of the predictor variable increases for the natural scale and a moderate increase in variability as the value of the predictor variable increases for the log scale (Figures 15). The first six ordered values in DATAGNB are 1.014, 1.089, 1.097, 1.100, 1.121, and 1.127. Results are provided in Table 13. Acceptable bias was achieved by the Half Normal model using the natural and log scales and the Gamma model. Although none of the predictor variable procedures achieved acceptable bias, the Log case of the piecewise regression procedure was marginally greater than the acceptable range for bias (0.0208). Only the half Normal using the natural and log scale achieved adequate coverage. Although none of the predictor variable procedures achieved acceptable coverage, the Log case of the piecewise regression procedure was marginally greater than the acceptable range for coverage (96%). Coverage for the piecewise regression procedure using g(x) = 1 was 100%. Plots are provided in Figure 16 for the g(x) = 1 case of the piecewise regression procedure.

Table 12. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population DATAG, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0705	0.0316	0.0956	100%
$g(x) = x$	0.0621	0.0417	0.0856	99%
$g(x) = x^2$	0.1110	0.0769	0.1410	100%
Log	0.0344	0.0206	0.0510	100%
Robust Regression:				
$g(x) = 1$	0.0802	0.0404	0.1062	100%
$g(x) = x$	0.0757	0.0587	0.1010	100%
$g(x) = x^2$	0.1213	0.0918	0.1523	100%
Log	0.0396	0.0280	0.0564	99%
Piecewise Linear:				
$g(x) = 1$	0.0564	0.0348	0.0773	100%
$g(x) = x$	0.0569	0.0359	0.0772	100%
$g(x) = x^2$	0.0645	0.0382	0.0863	100%
Log	0.0380	0.0301	0.0531	96%
Normal:				
Natural	0.0934	0.0365	0.1252	100%
Log	0.0380	0.0222	0.0571	98%
Half Normal:				
Natural	0.0312	0.0174	0.0482	97%
Log	0.0224	0.0158	0.0359	95%
Gamma:				
Natural	0.0102	0.0157	0.0160	47%

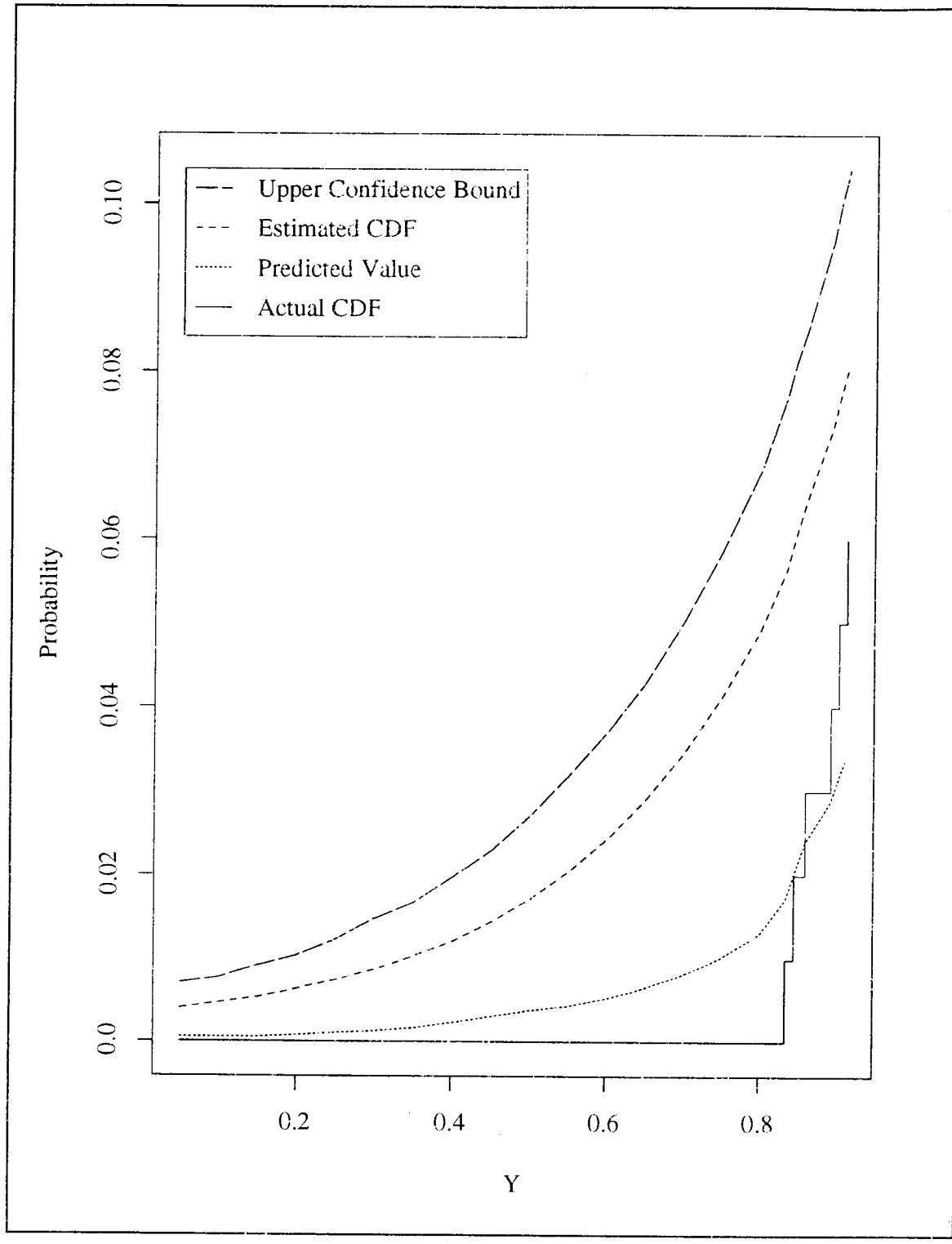

Figure 14. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using $g(x) = 1$ for population DATAG, where 100 replications were used in the simulations.

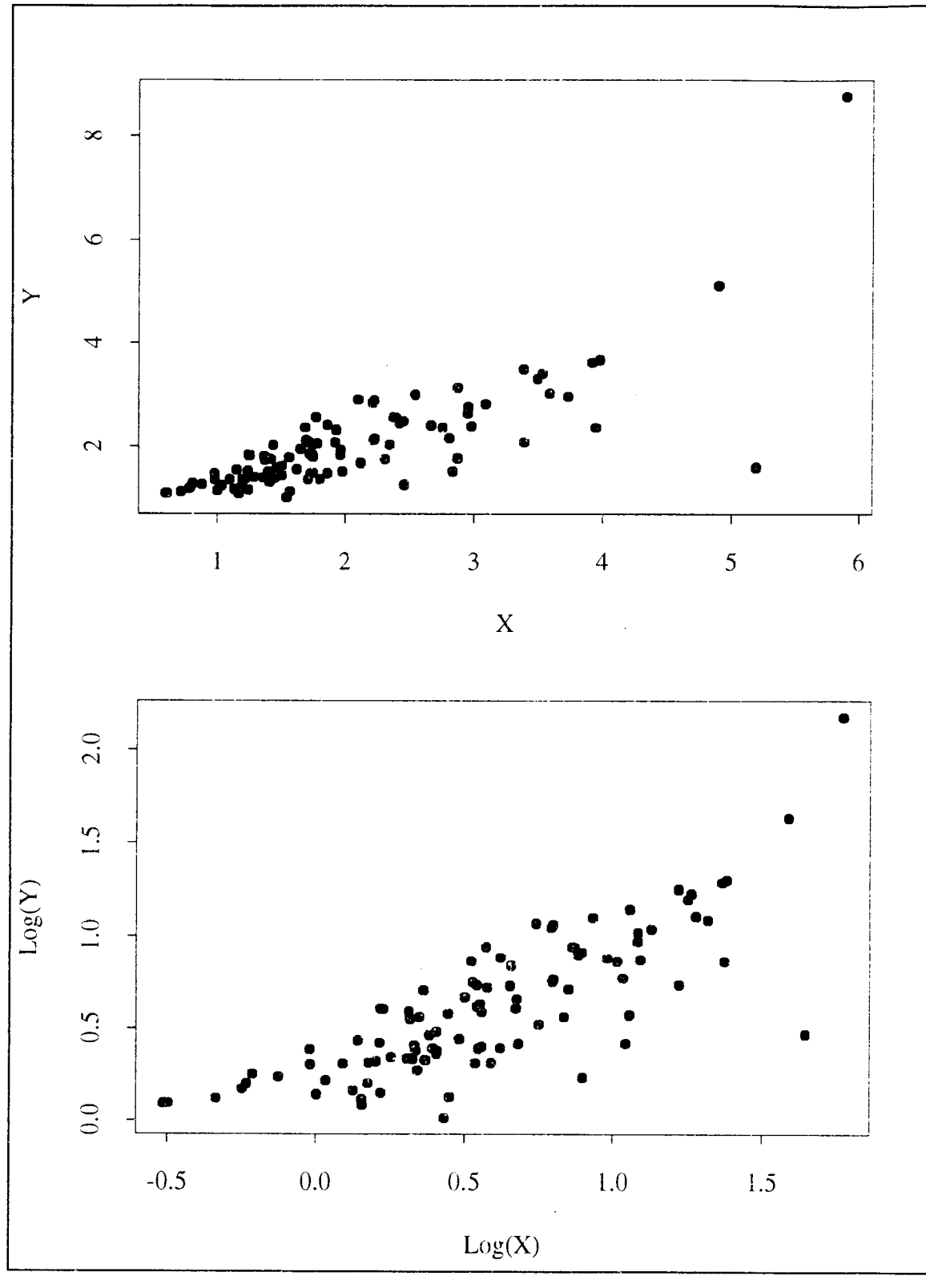

Figure 15. Scatter plots of Y versus X using the natural and log scales for population DATAGNB.

Table 13. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population DATAGNB, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0978	0.0646	0.1242	100%
$g(x) = x$	0.0810	0.0726	0.1017	98%
$g(x) = x^2$	0.1272	0.1005	0.1559	100%
Log	0.0527	0.0272	0.0696	100%
Robust Regression:				
$g(x) = 1$	0.0774	0.0439	0.1022	100%
$g(x) = x$	0.0424	0.0276	0.0588	99%
$g(x) = x^2$	0.0548	0.0373	0.0739	100%
Log	0.0426	0.0223	0.0586	100%
Piecewise Linear:				
$g(x) = 1$	0.0457	0.0328	0.0635	100%
$g(x) = x$	0.0403	0.0314	0.0558	99%
$g(x) = x^2$	0.0439	0.0336	0.0600	100%
Log	0.0308	0.0273	0.0429	96%
Normal:				
Natural	0.1068	0.0552	0.1399	100%
Log	0.0486	0.0266	0.0703	99%
Half Normal:				
Natural	0.0248	0.0193	0.0385	92%
Log	0.0183	0.0168	0.0289	84%
Gamma:				
Natural	0.0079	0.0147	0.0118	33%

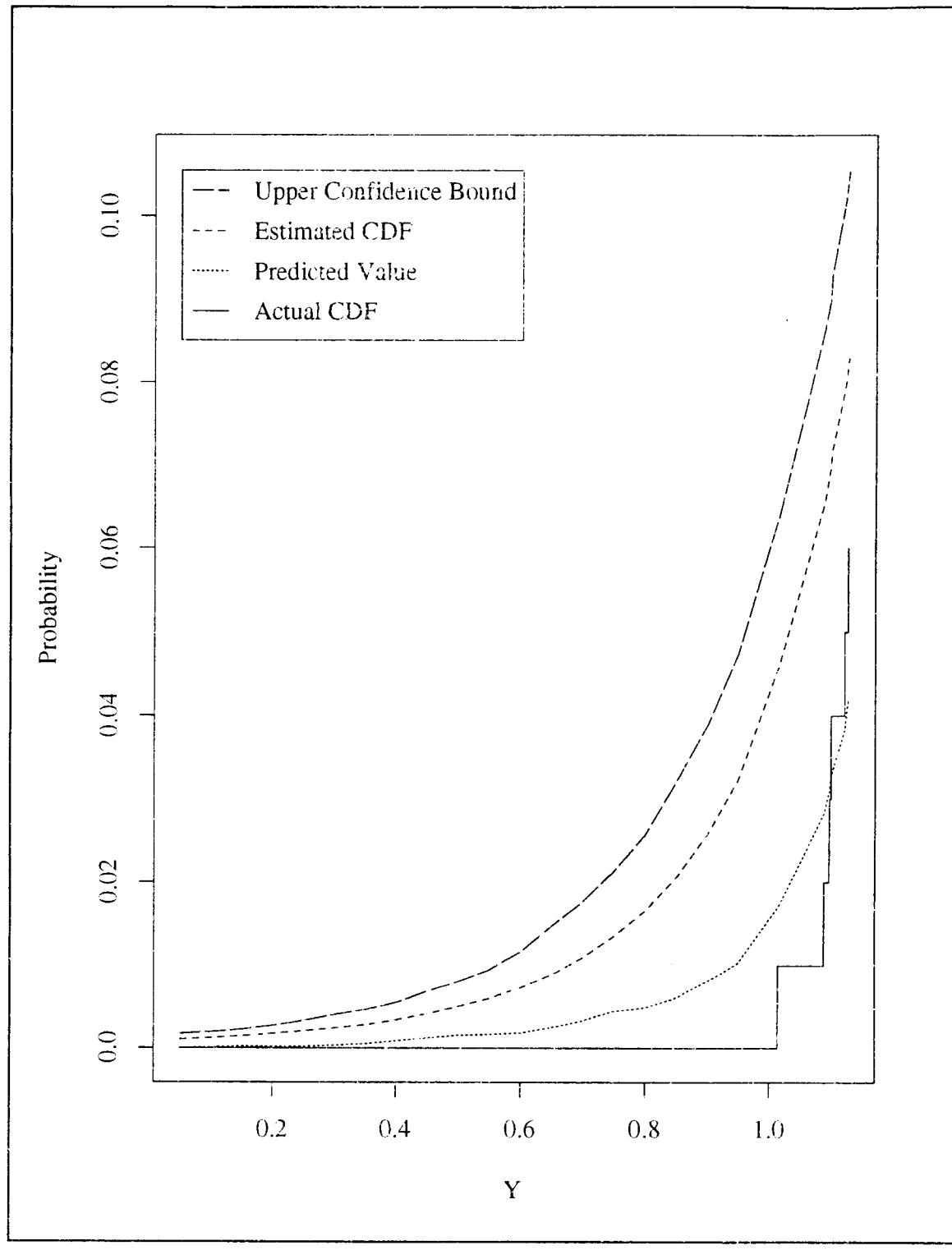

Figure 16. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population DATAGNB, where 100 replications were used in the simulations.

The estimated cdf exhibited moderate to large positive bias for values in the lower tail of the cdf for this procedure.

For DATAUNB he conditional distribution of the response variable given the predictor variable exhibits a moderate increase in variability as the value of the predictor variable increases for the natural scale and a small increase in variability as the value of the predictor variable increases for the log scale (Figure 17). The first six ordered values in DATAUNB are 0.508, 0.521, 0.542, 0.589, 0.640, and 0.689. Results are provided in Table 14. Acceptable bias was achieved by the g(x) = x and Log cases of the robust and piecewise regression procedures, the g(x) = 1 case of the piecewise regression procedure, and all cases of the procedures without a predictor variable except the Normal model using the natural scale. Adequate coverage was achieved by the g(x) = x case of the simple linear regression and piecewise regression procedures and the g(x) = x^2 version of the piecewise regression procedure. Coverage for the g(x) = 1 version of the piecewise regression procedure was 99%. Plots of results for the g(x) = 1 version of the piecewise regression procedure are provided in Figure 18. Although the estimated cdf produced moderately large positive bias for the smallest values in the lower tail of the cdf, overall the estimated cdf achieved very good performance in the lower tail.

Some discussion of the preliminary simulation results will be offered in the following paragraphs. No single procedure dominated the results for all eight populations. In addition procedures that performed well in terms of bias did not necessarily provide acceptable performance in terms of coverage. For that reason bias and coverage will be

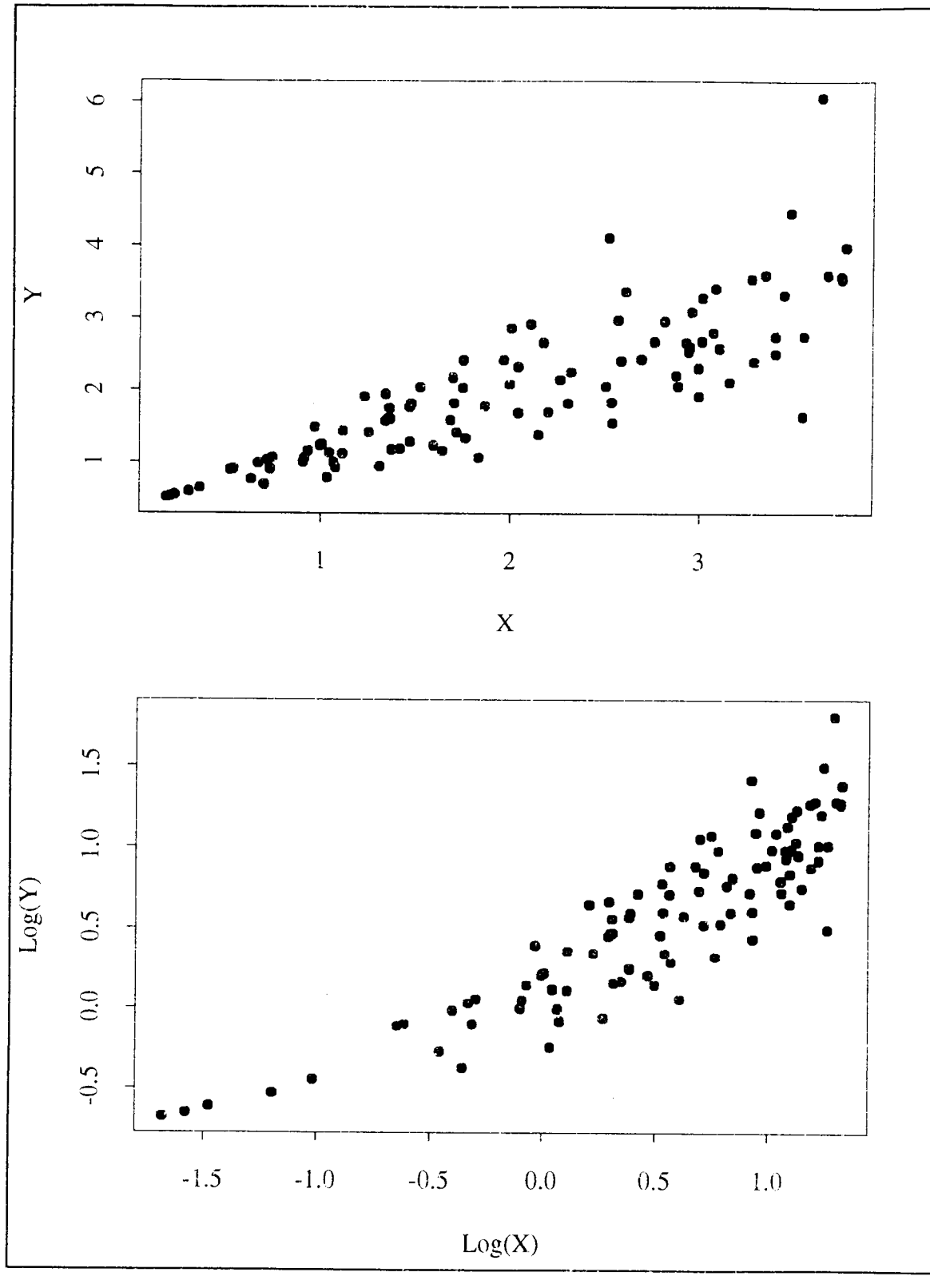

Figure 17. Scatter plots of Y versus X using the natural and log scales for population DATAUNB.

Table 14. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the initial ordered value in population DATAUNB, where 100 replications were used in the simulations and the actual value of the cdf was 0.01.

Model	Estimate	Std. Dev.	Bound	Coverage
Linear Regression:				
$g(x) = 1$	0.0540	0.0284	0.0731	100%
$g(x) = x$	0.0343	0.0393	0.0437	81%
$g(x) = x^2$	0.1231	0.1177	0.1486	99%
Log	0.0338	0.0131	0.0410	100%
Robust Regression:				
$g(x) = 1$	0.0518	0.0244	0.0712	100%
$g(x) = x$	0.0196	0.0193	0.0259	77%
$g(x) = x^2$	0.0828	0.0762	0.1040	96%
Log	0.0274	0.0150	0.0352	99%
Piecewise Linear:				
$g(x) = 1$	0.0287	0.0195	0.0407	99%
$g(x) = x$	0.0225	0.0232	0.0295	85%
$g(x) = x^2$	0.0375	0.0360	0.0485	91%
Log	0.0251	0.0173	0.0322	98%
Normal:				
Natural	0.0533	0.0277	0.0767	99%
Log	0.0098	0.0107	0.0167	74%
Half Normal:				
Natural	0.0219	0.0147	0.0354	96%
Log	0.0097	0.0110	0.0165	68%
Gamma:				
Natural	0.0049	0.0089	0.0080	36%

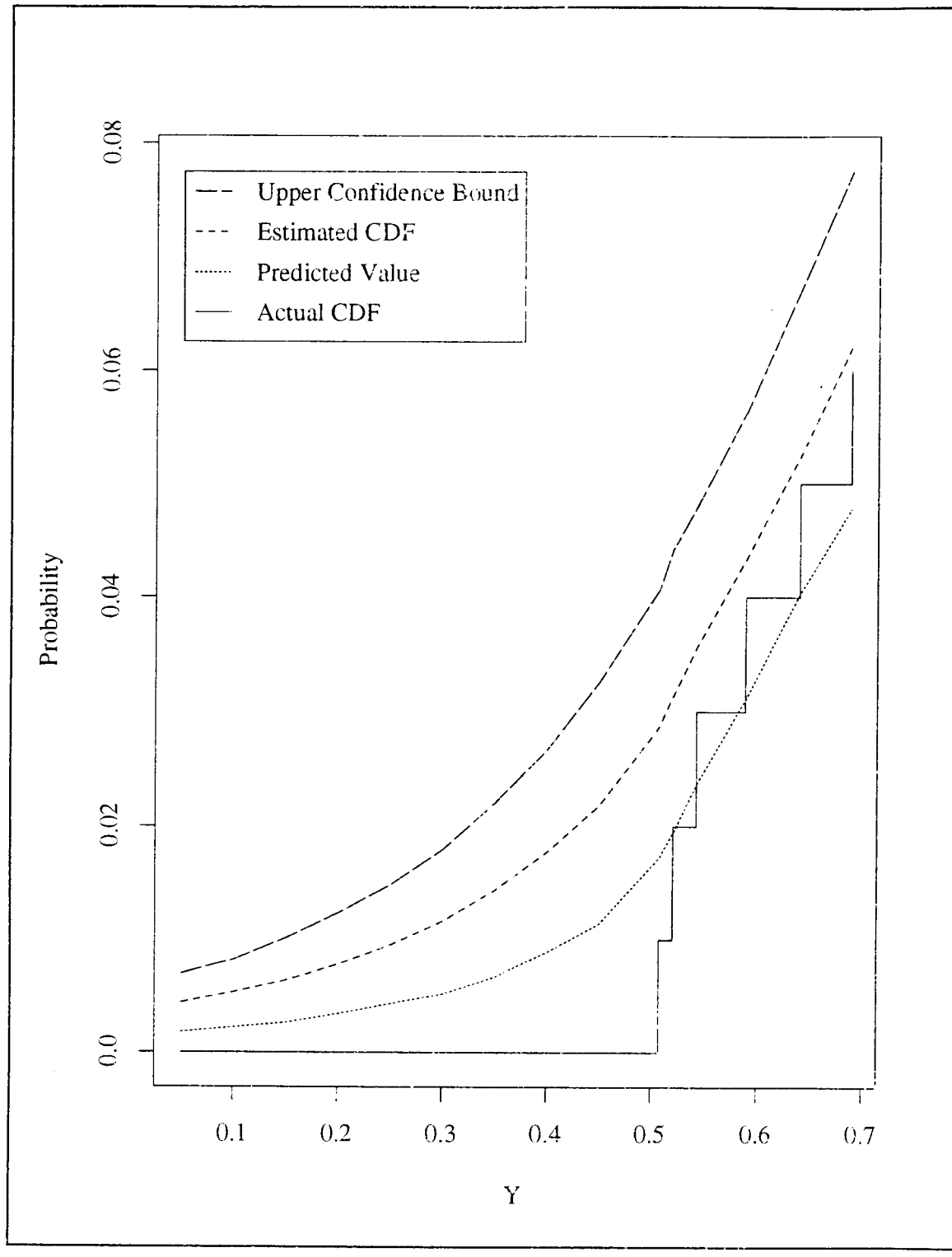

Figure 18. Plots of the actual lower tail of the cdf, means of the estimates of the lower tail of the cdf, means of the predicted values, and means of the weighted 90% upper confidence bounds for the piecewise linear regression procedure using g(x) = 1 for population DATAUNB, where 100 replications were used in the simulations.

discussed separately. In terms of bias, the procedures without a predictor variable often performed better than the predictor variable procedures. The predictor variable procedures exhibited moderate to large positive bias for most of the populations. Among the procedures without a predictor variable, the half Normal and Gamma models were notable for having small bias for many of the populations.

In terms of coverage, the procedures using a predictor variable were usually superior to the procedures without a predictor variable. Among the procedures without a predictor variable, the half Normal model performed best regarding coverage. Among the predictor variable procedures, the piecewise simple linear regression procedure performed best overall. For the four cases that were investigated for the piecewise linear regression procedure, the $g(x) = 1$ and $g(x) = x$ cases produced the most notable performance in terms of coverage. The only population for which the piecewise simple linear regression procedure produced inadequate coverage was DATAC, a population for which all of the procedures produced inadequate coverage results.

4.5 Further Simulation Results and Discussion

Based on the preliminary simulation results presented in Section 4.4, additional simulations were performed. The simulations employed all four versions of the piecewise linear regression procedure for each of the eight populations. One thousand samples of size sixteen were selected from each population, and the four versions of the piecewise regression procedure were applied to the sample values. Results of the simulations are provided in Tables 15 to 22 for the eight populations, respectively. In the tables the

Table 15. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population PADDY using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0341	0.0426	0.0498	0.0614	0.0708	0.0836
$g(x) = x$	0.0289	0.0374	0.0446	0.0563	0.0660	0.0794
$g(x) = x^2$	0.0330	0.0416	0.0490	0.0611	0.0710	0.0849
Log	0.0182	0.0264	0.0335	0.0454	0.0552	0.0689
Std. Dev.:						
$g(x) = 1$	0.0314	0.0343	0.0367	0.0399	0.0415	0.0442
$g(x) = x$	0.0313	0.0348	0.0375	0.0413	0.0433	0.0463
$g(x) = x^2$	0.0329	0.0367	0.0398	0.0441	0.0464	0.0498
Log	0.0256	0.0301	0.0335	0.0382	0.0406	0.0443
Bound:						
$g(x) = 1$	0.0489	0.0592	0.0676	0.0815	0.0917	0.1065
$g(x) = x$	0.0417	0.0522	0.0605	0.0745	0.0854	0.1009
$g(x) = x^2$	0.0472	0.0577	0.0662	0.0803	0.0915	0.1076
Log	0.0262	0.0368	0.0456	0.0603	0.0714	0.0875
Coverage:						
$g(x) = 1$	96.0%	88.9%	87.7%	85.8%	83.6%	86.8%
$g(x) = x$	92.3%	82.2%	81.6%	80.2%	76.9%	81.8%
$g(x) = x^2$	94.5%	86.3%	84.3%	81.1%	78.7%	82.1%
Log	72.3%	63.9%	65.6%	65.7%	65.6%	72.7%

Table 16. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population STREAM using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0181	0.0220	0.0278	0.0565	0.0590	0.0696
$g(x) = x$	0.0148	0.0185	0.0241	0.0525	0.0549	0.0657
$g(x) = x^2$	0.0177	0.0214	0.0270	0.0558	0.0583	0.0693
Log	0.0082	0.0117	0.0169	0.0456	0.0481	0.0594
Std. Dev.:						
$g(x) = 1$	0.0168	0.0197	0.0230	0.0347	0.0350	0.0379
$g(x) = x$	0.0157	0.0188	0.0222	0.0348	0.0351	0.0384
$g(x) = x^2$	0.0179	0.0210	0.0245	0.0373	0.0376	0.0409
Log	0.0110	0.0148	0.0187	0.0338	0.0342	0.0380
Bound:						
$g(x) = 1$	0.0280	0.0327	0.0398	0.0750	0.0772	0.0895
$g(x) = x$	0.0229	0.0275	0.0345	0.0696	0.0720	0.0844
$g(x) = x^2$	0.0271	0.0316	0.0385	0.0737	0.0762	0.0888
Log	0.0125	0.0171	0.0240	0.0603	0.0627	0.0760
Coverage:						
$g(x) = 1$	86.8%	70.9%	67.5%	84.4%	76.0%	80.5%
$g(x) = x$	79.7%	60.3%	59.6%	79.4%	70.9%	76.8%
$g(x) = x^2$	85.4%	64.6%	62.8%	80.2%	72.9%	78.3%
Log	57.8%	41.6%	44.1%	70.7%	62.1%	69.5%

Table 17. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population DATAA using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0379	0.0409	0.0514	0.0623	0.0740	0.0900
$g(x) = x$	0.0365	0.0395	0.0503	0.0617	0.0739	0.0907
$g(x) = x^2$	0.0446	0.0476	0.0588	0.0706	0.0832	0.1007
Log	0.0183	0.0213	0.0323	0.0442	0.0571	0.0751
Std. Dev.:						
$g(x) = 1$	0.0302	0.0310	0.0337	0.0367	0.0388	0.0412
$g(x) = x$	0.0312	0.0321	0.0352	0.0383	0.0406	0.0433
$g(x) = x^2$	0.0357	0.0366	0.0395	0.0423	0.0446	0.0471
Log	0.0225	0.0241	0.0294	0.0345	0.0383	0.0425
Bound:						
$g(x) = 1$	0.0539	0.0573	0.0697	0.0827	0.0958	0.1140
$g(x) = x$	0.0517	0.0553	0.0680	0.0815	0.0951	0.1142
$g(x) = x^2$	0.0617	0.0650	0.0783	0.0918	0.1061	0.1256
Log	0.0270	0.0306	0.0442	0.0588	0.0738	0.0948
Coverage:						
$g(x) = 1$	95.2%	87.8%	89.2%	87.7%	86.8%	91.6%
$g(x) = x$	92.7%	84.4%	82.6%	85.8%	84.5%	90.3%
$g(x) = x^2$	95.4%	87.3%	89.1%	88.0%	87.8%	92.6%
Log	77.9%	59.7%	67.8%	69.3%	71.4%	81.1%

Table 18. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population DATAB using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0267	0.0293	0.0338	0.0406	0.0566	0.0582
$g(x) = x$	0.0229	0.0256	0.0306	0.0380	0.0552	0.0569
$g(x) = x^2$	0.0216	0.0244	0.0296	0.0374	0.0552	0.0569
Log	0.0200	0.0228	0.0279	0.0355	0.0528	0.0544
Std. Dev.:						
$g(x) = 1$	0.0199	0.0192	0.0187	0.0178	0.0166	0.0166
$g(x) = x$	0.0200	0.0194	0.0191	0.0185	0.0178	0.0178
$g(x) = x^2$	0.0209	0.0203	0.0202	0.0198	0.0192	0.0193
Log	0.0142	0.0133	0.0132	0.0128	0.0125	0.0126
Bound:						
$g(x) = 1$	0.0368	0.0395	0.0444	0.0509	0.0677	0.0693
$g(x) = x$	0.0303	0.0333	0.0387	0.0461	0.0641	0.0657
$g(x) = x^2$	0.0272	0.0305	0.0361	0.0438	0.0625	0.0643
Log	0.0272	0.0303	0.0361	0.0433	0.0613	0.0629
Coverage:						
$g(x) = 1$	96.0%	87.9%	90.0%	78.6%	92.0%	82.1%
$g(x) = x$	91.0%	80.0%	80.6%	71.9%	89.6%	76.4%
$g(x) = x^2$	85.9%	75.0%	74.1%	67.3%	87.1%	71.8%
Log	93.0%	83.3%	84.1%	74.5%	90.3%	75.8%

Table 19. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population DATAC using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0106	0.0156	0.0186	0.0386	0.0407	0.0456
$g(x) = x$	0.0116	0.0170	0.0200	0.0408	0.0430	0.0480
$g(x) = x^2$	0.0193	0.0254	0.0288	0.0516	0.0538	0.0591
Std. Dev.:						
$g(x) = 1$	0.0181	0.0222	0.0244	0.0368	0.0375	0.0399
$g(x) = x$	0.0184	0.0228	0.0250	0.0386	0.0394	0.0419
$g(x) = x^2$	0.0222	0.0268	0.0289	0.0424	0.0434	0.0459
Bound:						
$g(x) = 1$	0.0164	0.0232	0.0268	0.0534	0.0555	0.0612
$g(x) = x$	0.0183	0.0253	0.0288	0.0563	0.0584	0.0643
$g(x) = x^2$	0.0296	0.0375	0.0413	0.0700	0.0723	0.0783
Coverage:						
$g(x) = 1$	56.8%	47.3%	44.8%	58.4%	50.4%	50.5%
$g(x) = x$	61.5%	49.9%	46.6%	60.1%	52.0%	53.0%
$g(x) = x^2$	80.8%	66.2%	62.1%	71.4%	63.5%	63.1%

Table 20. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population DATAG using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0567	0.0605	0.0652	0.0739	0.0779	0.0817
$g(x) = x$	0.0577	0.0617	0.0665	0.0755	0.0796	0.0836
$g(x) = x^2$	0.0669	0.0710	0.0759	0.0852	0.0894	0.0934
Log	0.0376	0.0418	0.0469	0.0568	0.0612	0.0655
Std. Dev.:						
$g(x) = 1$	0.0344	0.0355	0.0368	0.0392	0.0406	0.0419
$g(x) = x$	0.0357	0.0369	0.0382	0.0407	0.0421	0.0432
$g(x) = x^2$	0.0386	0.0397	0.0409	0.0431	0.0443	0.0454
Log	0.0291	0.0308	0.0327	0.0363	0.0381	0.0397
Bound:						
$g(x) = 1$	0.0771	0.0815	0.0865	0.0964	0.1006	0.1048
$g(x) = x$	0.0780	0.0824	0.0874	0.0979	0.1023	0.1066
$g(x) = x^2$	0.0890	0.0935	0.0986	0.1093	0.1138	0.1181
Log	0.0528	0.0574	0.0632	0.0749	0.0798	0.0846
Coverage:						
$g(x) = 1$	98.8%	97.2%	96.0%	94.0%	88.9%	87.7%
$g(x) = x$	98.7%	96.2%	94.8%	93.2%	88.1%	87.1%
$g(x) = x^2$	99.4%	97.3%	96.2%	94.2%	90.4%	90.0%
Log	95.9%	89.9%	86.1%	82.7%	75.4%	74.8%

Table 21. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population DATAGNB using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
g(x) = 1	0.0405	0.0592	0.0631	0.0654	0.0736	0.0771
g(x) = x	0.0339	0.0532	0.0573	0.0597	0.0685	0.0721
$g(x) = x^2$	0.0376	0.0576	0.0618	0.0643	0.0736	0.0774
Log	0.0262	0.0446	0.0486	0.0509	0.0593	0.0628
Std. Dev.:						
g(x) = 1	0.0323	0.0375	0.0379	0.0379	0.0400	0.0401
g(x) = x	0.0315	0.0373	0.0378	0.0378	0.0401	0.0403
$g(x) = x^2$	0.0334	0.0399	0.0406	0.0406	0.0431	0.0433
Log	0.0267	0.0332	0.0336	0.0336	0.0362	0.0365
Bound:						
g(x) = 1	0.0570	0.0792	0.0835	0.0859	0.0955	0.0990
g(x) = x	0.0480	0.0712	0.0759	0.0783	0.0886	0.0924
$g(x) = x^2$	0.0527	0.0765	0.0813	0.0837	0.0945	0.0987
Log	0.0372	0.0601	0.0645	0.0669	0.0768	0.0805
Coverage:						
g(x) = 1	98.2%	96.8%	95.7%	91.5%	87.7%	85.3%
g(x) = x	94.7%	93.0%	92.2%	85.5%	83.1%	80.2%
$g(x) = x^2$	95.6%	94.4%	92.9%	86.4%	83.0%	81.7%
Log	91.8%	91.5%	89.2%	79.2%	75.6%	73.0%

Table 22. Means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds for the first six ordered values in population DATAUNB using the piecewise simple linear regression procedure, where 1,000 replications were used in the simulations and the actual value of the cdf was 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06 for the six ordered values, respectively.

Model	Ordered Population Value					
	First	Second	Third	Fourth	Fifth	Sixth
Estimate:						
$g(x) = 1$	0.0285	0.0315	0.0356	0.0431	0.0521	0.0616
$g(x) = x$	0.0236	0.0270	0.0317	0.0405	0.0508	0.0612
$g(x) = x^2$	0.0404	0.0440	0.0492	0.0589	0.0704	0.0814
Log	0.0239	0.0270	0.0312	0.0384	0.0463	0.0545
Std. Dev.:						
$g(x) = 1$	0.0201	0.0201	0.0202	0.0208	0.0216	0.0235
$g(x) = x$	0.0232	0.0234	0.0236	0.0244	0.0252	0.0273
$g(x) = x^2$	0.0382	0.0389	0.0395	0.0405	0.0419	0.0439
Log	0.0173	0.0167	0.0164	0.0167	0.0174	0.0197
Bound:						
$g(x) = 1$	0.0401	0.0433	0.0478	0.0561	0.0660	0.0765
$g(x) = x$	0.0304	0.0340	0.0391	0.0492	0.0608	0.0727
$g(x) = x^2$	0.0516	0.0559	0.0612	0.0721	0.0848	0.0971
Log	0.0303	0.0334	0.0374	0.0447	0.0536	0.0629
Coverage:						
$g(x) = 1$	97.5%	91.6%	88.8%	85.5%	83.4%	85.9%
$g(x) = x$	84.4%	73.5%	74.7%	75.5%	78.1%	79.8%
$g(x) = x^2$	89.1%	80.4%	80.1%	80.9%	82.7%	83.3%
Log	95.5%	88.5%	88.6%	77.4%	70.1%	63.7%

means of the cdf estimates, standard deviations of the estimates, means of the weighted 90% upper confidence bounds, and coverage of the weighted 90% upper confidence bounds are provided for the first six ordered values in each population.

Regarding the means of the cdf estimates, the Log version of the piecewise regression procedure was superior to the versions using the original scale. Exclusive of populations DATAG and DATAGNB, the Log case performed very well with no consistent pattern in terms of bias. Even for populations DATAG and DATAGNB, the Log case did well for the fifth and sixth ordered values in the populations (Tables 20 and 21). Among the versions using the original scale, the $g(x) = x$ version usually performed marginally better than the $g(x) = 1$ and $g(x) = x^2$ versions. With a few exceptions, the versions using the original scale showed positive bias for all of the first six ordered values in each population. In addition, for most of the populations, bias of the estimates for the versions using the original scale was relatively constant for the first six ordered values in each population.

Regarding standard deviation of the estimates, the Log version consistently had smaller values than the three versions using the original scale. Among the three version using the original scale, standard deviations for the $g(x) = x$ version were somewhat smaller in most populations than values for the other two versions. Exclusive of population DATAB for which there was a decrease, standard deviations increased across the first six ordered population values. Note that the amount of increase was very small for population DATAUNB (Table 22).

Results for means of the upper confidence bounds were very similar to results for means of the cdf estimates. Analogous to the estimates, the amount by which the upper bound exceeded the actual cdf tended to remain close to constant for the versions using the original scale for most of the populations.

Regarding coverage of the upper confidence bounds, the three versions using the original scale produced superior performance in comparison to the Log version. Among the versions using the original scale, coverage was usually better for the second through the six ordered population values in comparison to the first ordered value. For example for the $g(x) = 1$ version and population DATAA, coverage was 95.2% for the first ordered value but ranged from 87.7% to 91.1% for the other ordered values (Table 17). Overall, the $g(x) = 1$ version performed best among the three versions using the original scale. The $g(x) = 1$ case performed well for all populations except DATAC and, to a lesser extent, STREAM. Population DATAC was the only population for which coverage was consistently inadequate for all four versions and all of the six ordered population values (Table 19). For population STREAM the $g(x) = 1$ case produced adequate coverage for three of the six ordered population values, while the $g(x) = x^2$ case produced adequate coverage for two ordered population values and the other two cases failed to produce adequate coverage for any of the ordered population values (Table 16).

5. CONCLUSIONS

5.1 Summary

This thesis has addressed absence of a specific class of objects in a finite set of objects, where the term universe was used to reference the finite set. A species of fish is an example of a class of objects, and the finite set of fish in a pool within a stream reach is an example of a universe. The problem that was addressed may be described as follows. A universe has been defined, and a probability sample has been selected from that universe. For each object in the sample, membership in a class of objects has been determined. Given absence of the class in the sample, the objective is to infer presence or absence of the class of objects in the universe. Two examples of this group of problems were considered: (a) absence of a species and (b) absence in relation to a threshold. Absence of a species was addressed in Chapter 2. Absence in relation to a threshold was addressed in Chapters 3 and 4.

Absence of a species was considered in terms of absence of a specific species of fish in a stream reach, with assessment via a sample of pools from the universe composed of the set of pools contained in the reach. For this example inference was in terms of the assessed probability that the species is absent in the universe given absence of the species in the sample, where the assessed probability is interpreted as a degree of belief. The probability of absence was developed at two levels: (a) the probability of absence of the species in the universe of pools in the reach given absence in a sample of pools, and (b) the probability of absence of the species in an individual sampled pool given that no

individuals of the species were observed in the pool. The probability of absence developed for those two levels was used to assess the probability of absence in the reach given that none were observed in the sample.

Absence in relation to a threshold considered absence of objects in a universe that belong to a class of objects defined by values of a quantitative attribute in a specific range, say, less than a low threshold. Specifically, this example considered a universe composed of a finite set of lakes, where the class of objects was lakes with values of a chemical attribute less than a low threshold value. Two inferential approaches were considered: (a) inference in terms of the initial ordered value in the finite population, e.g., the smallest value of the chemical attribute for the lakes in the universe; and (b) inference in terms of the threshold value, where inference utilized the estimated distribution function evaluated at the threshold value. The first inferential approach was addressed in Chapters 3, and the second inferential approach was addressed in Chapter 4.

Regarding absence of a species, our results demonstrate that, for the case of simple random sampling, the assessed probability of absence of the class of objects in the universe given absence of the class in the sample is either exactly or approximately equal to the probability of observing a specific single object from the class of objects given the protocol for observation. Using a modelled approach for the universe of fish in an individual pool, the assessed probability of absence given that none were observed can be made arbitrarily close to one for the range of observation probabilities considered.

Using the finite sampling approach for the universe of pools in a stream reach, the assessed probability of absence in the universe given absence in the sample of pools is no greater than the proportion of the pools in the reach that are sampled. Combining the two approaches to produce an assessment of the probability of absence in a reach given that none were observed, we demonstrated that the assessed probability is bounded by the sampling fraction, $\frac{n}{N}$. We conclude that the degree of belief the species is absent in the reach given that none were observed is no stronger that the amount of effort expended in sampling, and certainty of belief that the species is absent requires exhaustive sampling.

To further explore the primary result from Chapter 2, consider adding a third tier to the sampling design by obtaining a simple random sample of size m reaches from the M reaches in a stream. Using the finite sampling approach, the assessed probability that the species is absent from the universe of reaches in the stream given that the species was absent from the sample of reaches is approximately equal to $\frac{m}{M}$. The probability that the species is absent from the stream given that none were observed, therefore, is bounded by $\frac{n}{N} \cdot \frac{m}{M}$. Thus, the degree of belief that the species is absent from the stream given that none were observed is bounded by the product of the sampling fractions for the second and third tiers of sampling. Adding a fourth tier of sampling to assess absence of the species in the universe of stream in a stream basin will have an analogous effect on the assessed probability of absence given that none were observed. The general conclusion to be reached is that a conclusive (high probability) statement regarding absence of a species requires an exhaustive sampling effort at all levels below the lowest

level, say, a pool. Furthermore, weakness in the assessed probability increases as the size of the reporting unit increases from, say, a reach to a stream to a basin, etc. Thus, due to the monumental amount of sampling that is required, concluding that a species is absent from a domain of appreciable geographic extent is unlikely to be a realistic goal for a monitoring program. Applied to assessing absence of an endangered species based on a sampling protocol, one should maintain a healthy dose of skepticism regarding a conclusion that the species is absent from a geographic region.

The inferential approaches considered in Chapters 3 and 4 are alternative means for assessing absence of objects with values of a quantitative attribute less than a threshold value in the lower tail of the population distribution function. In Chapter 3 the estimators presented support point estimation and interval estimation regarding the initial ordered value in the finite population. Although such estimation is of interest, it does not provide assessment of the probability that the class of objects is absent from the universe. Conversely, the cdf estimators discussed in Chapter 4 provide estimates of the proportion of objects in the universe that have values of the quantitative attribute less than the threshold value. The upper confidence bounds for the estimated cdf that were discussed in Chapter 4 provide a mechanism for assessing the probability that the class of objects is absent in the universe.

In the context of inference regarding the initial ordered value in the finite population, the estimators in Chapter 3 that were based on the MLE produced mixed results. When the finite population can be approximated as a sample from the Uniform

distribution, the estimators performed very well. When the finite population can be approximated as a sample from the Normal distribution, however, the positive bias of the estimators made them effectively useless.

Within the confines of the sampling design that was presented, the estimators in Chapter 3 based on extreme value theory performed very well. Recall, however, that those estimators were predicated on the existence of several samples. In order to apply the methodology we have developed to an actual sampling situation, one must allocate available resources to create a set of r independent simple random samples. Conversely, one could create a composite simple random sample by combining the r simple random samples and eliminating repeat units. Using these two sampling designs, performance of the estimators developed using extreme value theory is very unlikely to equal performance of the estimators based on the MLE or the cdf estimators discussed in Chapter 4. For that reason the extreme value theory estimators would not be used in practice.

Regarding the cdf estimators discussed in Chapter 4, if the design goal was to produce unbiased estimates, then the best choice was to employ estimation without the predictor variable using the half Normal model or the Gamma model. For most of the populations, the procedures with a predictor variable produced consistent bias in the estimates.

Recall, however, that our goal was production of upper confidence bounds with acceptable performance in terms of coverage. Given that goal, the best overall choice

was to employ estimation with the predictor variable using the piecewise simple linear regression procedure. If the sample size was sufficiently large, some gain in performance would be expected by employing standard model fitting techniques to determine the best choice for the function g(x) and the value of utilizing transformations of scale. For sample sizes similar to those employed in the simulations, however, it is unlikely that one can distinguish among models, i.e., among versions of the procedures. In light of that fact, the consistent performance of the g(x) = 1 version of the piecewise linear regression procedure means that, in spite of an inability to distinguish among models, use of the that version of the piecewise linear regression procedure would not result in a significant decrease in performance.

5.2 Extensions

Three extensions to the problem that was addressed in this thesis will be discussed in this section. The first extension concerns the course of action to follow when samples are missing from a sampling design for detecting presence/absence of a species. The second extension concerns presence of a specific species of fish in a stream reach. Given a simple random sample of pools from the universe of pools in a reach, suppose that presence of the species was established for at least one sampled pool. Under these conditions we will consider assessment of the proportion of pools in the reach that contain the species. The third extension concerns assessment of the probability that a class of objects will become absent from a universe. This topic will be examined by considering a species of fish in a lake. Conditional on the number of individuals belonging to the species that was observed in the sample of fish from the lake, where the number of

individuals is not necessarily zero, the inference goal is to assess the probability that the species will become absent in the lake. An explicit length of time within which the species will become absent is not included in this definition.

5.2.1 Missing Samples

Consider the sampling design presented in Section 2.5 for assessing presence/absence of a species of fish in a stream reach. An issue that needs to be addressed is the correct manner in which to proceed when samples are missing, e.g., it was not possible to sample a selected pool. Two approaches for dealing with this situation will be considered. For the first approach the universe of pools is redefined by eliminating the non-sampled pools. In some situations the subset of non-sampled pools will constitute a distinguishable subset, in which case the probability would carry a proviso that the assessed value does not apply to that subset.

For the second approach, it will be taken as given that the missing samples are to be treated as missing at random. In most cases evidence in support of the missing at random designation will be required. For this approach inference would apply to the entire universe; this assumption results in the reduced sample still being a probability sample. Thus, estimation would proceed as discussed previously in Chapter 2, using the realized sample.

5.2.2 Proportion of Pools in a Reach That Contain a Species

Employing the sampling design presented in Section 2.5 for assessing presence or absence of a species of fish in a stream reach, suppose that presence was established for at least one sampled pool. Recall that observing the species in a pool established presence of the species in that pool. Conversely, some of the pools for which presence was not established may contain the species. Thus, the observed proportion of sampled pools with the species present is biased for the true proportion of pools in the universe with the species present. In Chapter 2 we interpreted the probability measure θ for a sampled pool as the degree of belief that the species was absent in the pool given that none were observed.

An alternative approach to assessment under a specified survey protocol is to define the species as being absent in the habitat unit when no individuals of the species are observed during sampling. This definition of absence will be referenced as statutory absence. Strict application of the prescribed survey protocol provides statutory validity for a determination that the species is absent in the habitat unit given that none were observed. The concept of statutory absence is commonly employed, e.g., Azuma et al (1990) used statutory absence to "establish" absence of spotted owls in a habitat unit.

For each sampled pool $u \in S$, let $\mathbf{y}_u \in \{0, 1\}$ represent the observation, where 0 indicates statutory absence and 1 indicates observed presence. Our development provides a method of determining a probability measure to associate with such a statutory

definition of absence. Specifically, we relate statutory absence to the assessed probability of absence in a pool given that none were observed by letting 0 represent 1 - θ. Then, since presence and absence are complementary, 1 - θ is interpreted as the assessed probability that the species is present in a pool given that none were observed. Also, recall that we have assumed that θ is constant across sampled pools.

The conventional estimator of the proportion of pools with the species present is given by $\hat{P} = \frac{t_y}{n}$, where $t_y = \sum_s y_u$. Since t_y represents the number of pools in the sample for which statutory presence was established rather than the number of sampled pools in which the species truly was present, $\hat{P} = \frac{t_y}{n}$ is biased for the true proportion of pools in the reach that contain the species. Define a modified observation as $y_u' \in \{1-\theta, 1\}$, where the 0 is replaced by 1 - θ. A bias-corrected estimator of the proportion of pools with the species present is given by $\hat{P}' = \frac{t_y'}{n}$ where $t_y' = \sum_s y_u'$. Let n_1 equal the number of pools in the sample for which presence was established and n_0 equal the number of sampled pools for which presence was not established, where $n = n_1 + n_0$. Then $t_y = n_1$ and $t_y' = n_1 + n_0(1-\theta) = n - n_0\theta$. Bias correction is provided by the term $n_0(1-\theta)$, which is an estimator of the number of sampled pools in which the species was assessed as statutorily absent but in which the species actually was present.

5.2.3 Probability of a Species Becoming Absent

One means by which the species of fish could become absent in the lake is due to a stochastic process that governs the number of individuals of the species in the lake. Initially, the finite sampling approach considered in Chapter 2 will be utilized. Let K equal the number of fish of the species in the lake, X equal the number of fish of the species in the sample, and $L(K=k \mid X=x)$ equal the likelihood function. In addition let V equal the safe population size for the species, i.e., if the lake contains at least V individuals of the species, then the species will not become absent due to the stochastic process governing the number of individuals of the species in the lake. Let p(x) equal the assessed probability that the species will become absent in the lake given that X=x individuals of the species were observed in the sample. We can write p(x) as:

$$p(x) = \sum_{k=x}^{N-n+x} P(K=k \mid X=x) * \phi(k)$$

where $P(K=k \mid X=x)$, the fiducial probability that the population contains K=k individuals of the species given that X=x individuals were observed in the sample, is given by:

$$P(K=k \mid X=x) = \frac{L(K=k \mid X=x)}{\sum_{k'=x}^{N-n+x} L(K=k' \mid X=x)}$$

and $\phi(k)$ is the extinction model, i.e., the probability that a population containing K=k individuals of the species will become absent due to the modelled stochastic process.

Recall that $\mathbf{P(K=k \mid X=x)}$ is interpreted as a degree of belief. Conversely, $\phi(k)$ is interpreted as a physical probability resulting from the assumed stochastic process. As discussed in Savage (1954), the product of a subjective probability (degree of belief) and an objective probability (physical probability) is a subjective probability. Thus, p(x) is interpreted as summarizing the degree of belief that the species will become absent in the lake given that X=x individuals of the species were observed in the sample.

Any of a large number of functions could be employed as the extinction model. As a specific example, consider the following formula for $\phi(k)$, which was derived from one presented by Goel and Richter-Dyn (1974):

$$\phi(K) = \begin{cases} 1 - \dfrac{(1-\tau^{K})}{(1-\tau^{V})} & \text{for } K<V \\ 0 & \text{for } K \geq V \end{cases}$$

where τ is a parameter for the probability function, a constant contained in the range (0, 1) that is specific for the species of interest. For sake of reference, note that the formula for $\phi(K)$ is related to the model of MacArthur and Wilson (1967) for extirpation of a colonizing species. The MacArthur and Wilson model is a stochastic birth and death process for which per capita birth and death rates depend linearly on the size of the population. Richter-Dyn and Goel (1972) demonstrated that the probability of extirpation of a colonizing species for this model is given by $\phi(k)$, where τ is the ratio of per capita death rate to per capita birth rate in the underlying model. We propose the ad hoc use of this model, which does not imply belief in its applicability.

Using the model that has been developed, the value of p(x) was determined for the following values of N, n, τ, and V: N = {500, 1,000, 2,500, 10,000}, n = {50, 100, 150, 200, 250, 300, 350, 400, 450, 500}, τ = {0.10, 0.20, 0.50, 0.60, 0.75, 0.90, 0.95, 0.99}, and V = {10, 25, 50}. Results are presented in Tables 23 - 26 for the four valuesof N, respectively. The value of p(x) increased for both of the following cases: (a) increasing value of n and fixed values of N, τ and V; and (b) increasing value of τ and fixed values of N, n and V. For increasing value of V and fixed values of N and n, p(x) was constant for fixed τ within the set {0.10, 0.20}, was nondecreasing for fixed τ within the set {0.50, 0.60, 0.75}, and was increasing for fixed τ within the set {0.90, 0.95, 0.99}. The value of p(x) decreased for increasing value of N and fixed values of n, τ and V.

It is unlikely that an estimate of N, the number of fish in the lake, will be available. Even if an estimate of N was available, it is not likely that an investigator will have an extinction model for $\phi(K)$ to employ in assessing p(x). For that reason, it was decided to choose a set of standard parameter values and to interpret values of p(x) as an index related to the probability that the species will become absent in the lake given that x individuals were observed in the sample. The standardized values chosen were as follows: N=1,000, τ=0.5, and V=25. Results for the standardized set of parameter values and the range of values n = {50, 100, 150, 200, 250, 300, 350, 400, 450, 500} and x = {0, 1, 2, 3, 4, 5, 6, 7, 8, 9} are presented in Table 27. Two trends can be discerned from Table 27. For fixed value of n, the index p(x) decreased for increasing value of x. For fixed value of x, the index p(x) increased for increasing value of n.

Table 23. Values of the probability that the species will become absent in the lake, where n equals the sample size, τ equals the intrinsic extinction factor, and V equals the safe population size for the species. Note that N, the number of fish in the lake, equals 500, and X, the number of individuals of the species in the sample, equals 0.

n	τ = 0.10			τ = 0.20		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.112	0.112	0.112	0.124	0.124	0.124
100	0.219	0.219	0.219	0.240	0.240	0.240
150	0.324	0.324	0.324	0.350	0.350	0.350
200	0.427	0.427	0.427	0.456	0.456	0.456
250	0.527	0.527	0.527	0.557	0.557	0.557
300	0.626	0.626	0.626	0.653	0.653	0.653
350	0.722	0.722	0.722	0.745	0.745	0.745
400	0.817	0.817	0.817	0.834	0.834	0.834
450	0.909	0.909	0.909	0.919	0.919	0.919
n	τ = 0.50			τ = 0.60		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.185	0.185	0.185	0.218	0.221	0.221
100	0.335	0.336	0.336	0.384	0.387	0.387
150	0.463	0.464	0.464	0.517	0.519	0.519
200	0.573	0.573	0.573	0.624	0.627	0.627
250	0.668	0.668	0.668	0.714	0.715	0.715
300	0.751	0.751	0.751	0.789	0.790	0.790
350	0.824	0.824	0.824	0.853	0.854	0.854
400	0.889	0.889	0.889	0.909	0.909	0.909
450	0.947	0.948	0.948	0.957	0.958	0.958

Table 23. (Continued)

n	τ = 0.75			τ = 0.90		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.286	0.312	0.313	0.367	0.500	0.531
100	0.477	0.503	0.503	0.582	0.696	0.716
150	0.613	0.634	0.634	0.715	0.798	0.812
200	0.713	0.729	0.729	0.802	0.860	0.870
250	0.789	0.801	0.801	0.861	0.903	0.909
300	0.849	0.858	0.858	0.904	0.933	0.937
350	0.898	0.903	0.904	0.937	0.956	0.959
400	0.938	0.941	0.941	0.963	0.974	0.976
450	0.971	0.973	0.973	0.983	0.988	0.989
n	τ = 0.95			τ = 0.99		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.395	0.587	0.671	0.416	0.658	0.798
100	0.615	0.773	0.822	0.641	0.830	0.904
150	0.746	0.857	0.888	0.770	0.899	0.943
200	0.828	0.904	0.925	0.848	0.934	0.963
250	0.882	0.934	0.949	0.897	0.956	0.975
300	0.920	0.956	0.965	0.931	0.970	0.983
350	0.948	0.971	0.977	0.956	0.981	0.989
400	0.969	0.983	0.987	0.974	0.989	0.994
450	0.986	0.992	0.994	0.988	0.995	0.997

Table 24. Values of the probability that the species will become absent in the lake, where n equals the sample size, τ equals the intrinsic extinction factor, and V equals the safe population size for the species. Note that N, the number of fish in the lake, equals 1,000, and X, the number of individuals of the species in the sample, equals 0.

n	τ = 0.10			τ = 0.20		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.056	0.056	0.056	0.063	0.063	0.063
100	0.111	0.111	0.111	0.123	0.123	0.123
150	0.165	0.165	0.165	0.182	0.182	0.182
200	0.218	0.218	0.218	0.239	0.239	0.239
250	0.271	0.271	0.271	0.295	0.295	0.295
300	0.323	0.323	0.323	0.350	0.350	0.350
350	0.375	0.375	0.375	0.403	0.403	0.403
400	0.426	0.426	0.426	0.455	0.455	0.455
450	0.477	0.477	0.477	0.506	0.506	0.506
500	0.527	0.527	0.527	0.556	0.556	0.556
n	τ = 0.50			τ = 0.60		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.097	0.097	0.097	0.116	0.118	0.118
100	0.183	0.183	0.183	0.216	0.219	0.219
150	0.262	0.262	0.262	0.304	0.308	0.308
200	0.334	0.335	0.335	0.383	0.386	0.386
250	0.401	0.401	0.401	0.453	0.456	0.456
300	0.462	0.463	0.463	0.515	0.518	0.518
350	0.519	0.519	0.519	0.572	0.575	0.575
400	0.572	0.572	0.572	0.623	0.626	0.626
450	0.621	0.621	0.621	0.670	0.672	0.672
500	0.667	0.667	0.667	0.713	0.715	0.715

Table 24. (Continued)

n	τ = 0.75			τ = 0.90		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.157	0.177	0.177	0.208	0.314	0.348
100	0.283	0.310	0.310	0.364	0.496	0.528
150	0.388	0.415	0.416	0.485	0.613	0.639
200	0.475	0.501	0.502	0.580	0.694	0.714
250	0.549	0.572	0.573	0.654	0.753	0.769
300	0.611	0.632	0.633	0.713	0.797	0.811
350	0.665	0.683	0.684	0.761	0.832	0.843
400	0.712	0.728	0.728	0.801	0.860	0.869
450	0.753	0.766	0.766	0.833	0.883	0.891
500	0.788	0.800	0.800	0.861	0.902	0.909
n	τ = 0.95			τ = 0.99		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.225	0.384	0.483	0.239	0.444	0.623
100	0.391	0.583	0.668	0.412	0.653	0.795
150	0.517	0.698	0.763	0.542	0.764	0.866
200	0.613	0.771	0.821	0.639	0.829	0.903
250	0.687	0.821	0.859	0.712	0.870	0.927
300	0.745	0.856	0.887	0.769	0.898	0.942
350	0.791	0.883	0.908	0.812	0.918	0.954
400	0.827	0.904	0.925	0.847	0.934	0.963
450	0.857	0.920	0.938	0.874	0.946	0.969
500	0.882	0.934	0.948	0.897	0.955	0.975

Table 25. Values of the probability that the species will become absent in the lake, where n equals the sample size, τ equals the intrinsic extinction factor, and V equals the safe population size for the species. Note that N, the number of fish in the lake, equals 2,500, and X, the number of individuals of the species in the sample, equals 0.

n	τ = 0.10			τ = 0.20		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.023	0.023	0.023	0.025	0.025	0.025
100	0.045	0.045	0.045	0.050	0.050	0.050
150	0.067	0.067	0.067	0.074	0.074	0.074
200	0.089	0.089	0.089	0.098	0.098	0.098
250	0.110	0.110	0.110	0.122	0.122	0.122
300	0.132	0.132	0.132	0.146	0.146	0.146
350	0.154	0.154	0.154	0.169	0.169	0.169
400	0.175	0.175	0.175	0.193	0.193	0.193
450	0.196	0.196	0.196	0.216	0.216	0.216
500	0.218	0.218	0.218	0.238	0.238	0.238
n	τ = 0.50			τ = 0.60		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.040	0.040	0.040	0.048	0.049	0.049
100	0.077	0.078	0.078	0.093	0.095	0.095
150	0.114	0.114	0.114	0.136	0.138	0.138
200	0.148	0.149	0.149	0.177	0.179	0.179
250	0.182	0.182	0.182	0.215	0.218	0.218
300	0.214	0.215	0.215	0.252	0.255	0.255
350	0.246	0.246	0.246	0.286	0.290	0.290
400	0.276	0.276	0.276	0.320	0.323	0.323
450	0.305	0.306	0.306	0.351	0.355	0.355
500	0.333	0.334	0.334	0.382	0.385	0.385

Table 25. (Continued)

n	τ = 0.75			τ = 0.90		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.067	0.077	0.077	0.090	0.147	0.170
100	0.127	0.144	0.144	0.169	0.262	0.294
150	0.183	0.204	0.205	0.240	0.354	0.389
200	0.234	0.259	0.259	0.304	0.431	0.464
250	0.282	0.308	0.309	0.362	0.494	0.525
300	0.326	0.353	0.354	0.415	0.547	0.576
350	0.367	0.395	0.395	0.462	0.592	0.618
400	0.405	0.433	0.433	0.504	0.630	0.655
450	0.441	0.468	0.468	0.543	0.664	0.686
500	0.474	0.500	0.500	0.578	0.693	0.713
n	τ = 0.95			τ = 0.99		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.098	0.185	0.257	0.104	0.219	0.357
100	0.183	0.324	0.418	0.195	0.376	0.553
150	0.260	0.430	0.528	0.275	0.494	0.669
200	0.328	0.514	0.607	0.347	0.583	0.743
250	0.389	0.581	0.665	0.410	0.651	0.793
300	0.444	0.634	0.710	0.467	0.704	0.828
350	0.493	0.678	0.746	0.517	0.745	0.854
400	0.537	0.714	0.775	0.562	0.778	0.874
450	0.576	0.744	0.799	0.602	0.805	0.890
500	0.612	0.770	0.820	0.637	0.828	0.903

Table 26. Values of the probability that the species will become absent in the lake, where n equals the sample size, τ equals the intrinsic extinction factor, and V equals the safe population size for the species. Note that N, the number of fish in the lake, equals 10,000, and X, the number of individuals of the species in the sample, equals 0.

n	τ = 0.10			τ = 0.20		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.006	0.006	0.006	0.006	0.006	0.006
100	0.011	0.011	0.011	0.013	0.013	0.013
150	0.017	0.017	0.017	0.019	0.019	0.019
200	0.022	0.022	0.022	0.025	0.025	0.025
250	0.028	0.028	0.028	0.031	0.031	0.031
300	0.033	0.033	0.033	0.037	0.037	0.037
350	0.039	0.039	0.039	0.043	0.043	0.043
400	0.044	0.044	0.044	0.050	0.050	0.050
450	0.050	0.050	0.050	0.056	0.056	0.056
500	0.055	0.055	0.055	0.062	0.062	0.062
n	τ = 0.50			τ = 0.60		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.010	0.010	0.010	0.012	0.013	0.013
100	0.020	0.020	0.020	0.024	0.025	0.025
150	0.030	0.030	0.030	0.036	0.037	0.037
200	0.039	0.039	0.039	0.048	0.049	0.049
250	0.049	0.049	0.049	0.059	0.060	0.060
300	0.058	0.058	0.058	0.070	0.072	0.072
350	0.068	0.068	0.068	0.082	0.083	0.083
400	0.077	0.077	0.077	0.093	0.095	0.095
450	0.086	0.086	0.086	0.104	0.106	0.106
500	0.095	0.095	0.095	0.114	0.116	0.116

Table 26. (Continued)

n	τ = 0.75			τ = 0.90		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.017	0.020	0.020	0.023	0.040	0.048
100	0.034	0.039	0.039	0.046	0.077	0.091
150	0.050	0.058	0.058	0.067	0.112	0.131
200	0.066	0.076	0.076	0.089	0.145	0.168
250	0.081	0.093	0.093	0.109	0.176	0.202
300	0.096	0.110	0.110	0.129	0.205	0.234
350	0.111	0.127	0.127	0.149	0.233	0.264
400	0.126	0.143	0.143	0.168	0.260	0.291
450	0.140	0.158	0.159	0.186	0.285	0.318
500	0.154	0.174	0.174	0.204	0.309	0.342
n	τ = 0.95			τ = 0.99		
	V=10	V=25	V=50	V=10	V=25	V=50
50	0.026	0.051	0.076	0.027	0.061	0.111
100	0.050	0.098	0.142	0.053	0.117	0.203
150	0.073	0.142	0.201	0.078	0.168	0.283
200	0.096	0.183	0.253	0.103	0.216	0.352
250	0.119	0.221	0.300	0.126	0.260	0.412
300	0.141	0.256	0.343	0.149	0.301	0.464
350	0.162	0.290	0.381	0.172	0.339	0.510
400	0.182	0.321	0.416	0.193	0.374	0.549
450	0.202	0.351	0.447	0.214	0.407	0.584
500	0.221	0.378	0.476	0.235	0.437	0.615

Table 27. Values of the probability that the species will become absent in the lake, where n equals the sample size, and X equals the number of individuals of the species in the sample. Note that N, the number of fish in the lake, equals 1,000; τ, the intrinsic extinction factor, equals 0.5; and V, the safe population size for the species, equals 25.

n	X				
	0	1	2	3	4
50	0.0970	0.0046	0.0002	0.0000	0.0000
100	0.1834	0.0167	0.0015	0.0001	0.0000
150	0.2623	0.0343	0.0045	0.0006	0.0001
200	0.3346	0.0558	0.0093	0.0015	0.0003
250	0.4012	0.0803	0.0160	0.0032	0.0006
300	0.4626	0.1068	0.0246	0.0057	0.0013
350	0.5194	0.1348	0.0349	0.0090	0.0023
400	0.5722	0.1636	0.0467	0.0133	0.0038
450	0.6214	0.1929	0.0599	0.0186	0.0058
500	0.6673	0.2225	0.0742	0.0247	0.0082
n	X				
	5	6	7	8	9
50	0.0000	0.0000	0.0000	0.0000	0.0000
100	0.0000	0.0000	0.0000	0.0000	0.0000
150	0.0000	0.0000	0.0000	0.0000	0.0000
200	0.0000	0.0000	0.0000	0.0000	0.0000
250	0.0001	0.0000	0.0000	0.0000	0.0000
300	0.0003	0.0001	0.0000	0.0000	0.0000
350	0.0006	0.0002	0.0000	0.0000	0.0000
400	0.0011	0.0003	0.0001	0.0000	0.0000
450	0.0018	0.0006	0.0002	0.0001	0.0000
500	0.0027	0.0009	0.0003	0.0001	0.0000

For the case x=0, i.e., absence of the species in the sample, let ω equal the value of p(0) that was calculated using the standardized set of parameters values. It is of interest to compare ω to ψ, the assessed probability that the species is absent in the lakegiven absence in the sample. Table 28 provides value of ω and ψ for N=1,000 and a range of values of n. The results in Table 28 show that the value of ψ consistently is smaller than the associated value of ω. Given absence of the species in the sample, it makes intuitive sense that the degree of belief that the species is absent in the lake is less than the degree of belief that the species will become absent in the lake. Recall that, in our definition of p(x), no limit is placed on the length of time within which the species will become absent in the lake.

Table 28. Values of ω, the probability that the species will become absent in the lake given that no individuals were observed in the sample, and ψ, the probability that the species is absent in the lake given absence in the sample, where N, the number of fish in the lake, equals 1,000.

	n									
	50	100	150	200	250	300	350	400	450	500
ω	0.097	0.183	0.262	0.335	0.401	0.463	0.519	0.572	0.621	0.667
ψ	0.051	0.101	0.151	0.201	0.251	0.301	0.351	0.401	0.451	0.500

5.3 Future Research

Three topics for future research will be discussed. First, suppose that unequal probability sampling was employed in a sampling design for assessing presence/absence of a species. A reasonable approach for this situation is to assume simple random

sampling in order to assess the probability of absence given that none were observed. One could then investigate the impact of this assumption on the probability assessment. Second, the estimators of the minimum value in a finite population that were based on the MLE performed adequately for a finite population selected from the Uniform distribution but not for a finite population selected from the Normal distribution. It would be worthwhile to investigate methodology for improving performance of those estimators for finite populations other than ones that can be approximated as a sample from the Uniform distribution. Third, for most of the finite populations investigated for the estimated cdf, the procedures that included a predictor variable overestimated the true distribution function in the lower tail, which resulted in confidence intervals that exceeded the nominal coverage. At the same time the estimated cdf for the predicted values usually was close to the true distribution function, which indicates that excessive error was being added to the predicted values. Therefore, one could investigate procedures that reduced the observed bias in the estimated cdf by decreasing the amount of error added to the predicted values via the Chambers and Dunstan protocol.

BIBLIOGRAPHY

Azuma, D.L., Baldwin, J.A., and Noon, B.R. 1990. Estimating the occupancy of spotted owl habitat areas by sampling and adjusting for bias. U.S. Department of Agriculture, Forest Service, Pacific Southwest Research Station, General Technical Report PSW-124.

Bolfarine, H. and Sandoval, M.C. 1993. Prediction of the finite population distribution function under Gaussian superpopulation models. Australian Journal of Statistics 35: 195-204.

Bolfarine, H. and Sandoval, M.C. 1994. On predicting the finite population distribution function. Statistics and Probability Letters 19: 339-347.

Bowman, K.O. and Shenton, L.R. 1988. *Properties of Estimates for the Gamma Distribution.* Marcel Dekker, New York.

Chambers, R.L. and Dunstan, R. 1986. Estimating distribution functions from survey data. Biometrika 73: 597-604.

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. 1993. Bias robust estimation in finite populations using nonparametric calibration. Journal of the American Statistical Association 88: 268-277.

Cleveland, W.S. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74: 829-836.

de Finetti, B. 1937. Foresight: its logical flaws, its subjective sources. Translated in H. Kyburg and H. Smokler, eds., Studies in Subjective Probability. Wiley, New York, 1964., pp. 93-158.

Dorfman, A.H. 1993. A comparison of design-based and model-based estimators of the finite population distribution function. Australian Journal of Statistics 35: 29-41.

Dorfman, A.H. and Hall, P. 1993. Estimators of the finite population distribution function using nonparametric regression. The Annals of Statistics 21: 1452-1475.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Fisher, R.A. 1930. Inverse probability. Proceedings of the Cambridge Philosophical Society 26: 528-535.

Fisher, R.A. and Tippett, L.H.C. 1928. Limiting forms of the frequency distributions of the largest or smallest member of a sample. Proceedings of the Cambridge Philosophical Society 24: 180-190.

Goel, N.S. and Richter-Dyn, N. 1974. *Stochastic Models in Biology.* Academic Press, New York.

Horvitz, D.G. and Thompson, D.J. 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47: 663-685.

Hosking, J.R.M., Wallis, J.R., and Wood, E.F. 1985. Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. Technometrics 27: 251-261.

Johnson, N.L. and Kotz, S. 1970. *Continuous Univariate Distributions - 1.* John Wiley & Sons, New York.

Kolmogorov, A.N. 1933. Grundbegriffe der Wahrscheinlichkeitrechnung. Translated in N. Morrison, Foundations of the Theory of Probability, 2nd edition. Chelsea, New York, 1956.

Konijn, H.S. 1973. *Statistical Theory of Sample Survey Design and Analysis.* North-Holland, Amsterdam.

Kuk, A.Y.C. 1988. Estimation of distribution functions and medians under sampling with unequal probabilities. Biometrika 75: 97-103.

Kuk, A.Y.C. 1993. A kernel method for estimating finite population distribution functions using auxiliary information. Biometrika 80: 385-392.

Kuo, L. 1988. Classical and prediction approaches to estimating distribution functions from survey data. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 280-285.

Linthurst, R.A., Landers, D.H., Eilers, J.M., Brakke, D.F., Overton, W.S., Meier, E.P., and Crowe, R.E. 1986. *Characteristics of Lakes in the Eastern United States. Volume 1: Population Descriptions and Physico-chemical Relationships.* EPA-600/4-86/007a. U.S. Environmental Protection Agency, Washington, D.C.

Messer, J.J., Ariss, C.W., Baker, J.R., Drouse, S.K., Eshleman, K.N., Kinney, A.J., Overton, W.S., Sale, M.J., and Schonbrod, R.D. 1988. Stream chemistry in the Southern Blue Ridge: feasibility of a regional synoptic sampling approach. Water Resources Bulletin 24: 821-829.

Overton, W.S. and Stehman, S.V. 1993. *Improvement of Performance of Variable Probability Sampling Strategies through Application of the Population Space and the Facsimile Population Bootstrap.* Technical Report No. 148. Department of Statistics, Oregon State University, Corvallis, Oregon.

Overton, W.S. and Stevens, D.L., and White, D. 1990. *The EMAP Design Perspective: A Prescription for Environmental Monitoring.* Technical Report No. 145. Department of Statistics, Oregon State University, Corvallis, Oregon.

Pyke, R. 1965. Spacings. Journal of the Royal Statistical Society, Series B 7: 395-449.

Rao, J.N.K., Kovar, J.G., and Mantel, H.J. 1990. On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika 77: 365-375.

Savage, L.J. 1954. *The Foundations of Statistics.* Wiley, New York.

Sedransk, N. and Sedransk, J. 1979. Distinguishing among distributions using data from complex sample designs. Journal of the American Statistical Society 74: 754-760.

Stehman, S.V. and Overton, W.S. 1994. Comparison of variance estimators of the Horvitz-Thompson estimator for randomized variable probability systematic sampling. Journal of the American Statistical Association 89: 30-43.

Wright, T. 1990. When zero defectives appear in a sample: upper bounds on confidence coefficients of upper bounds. The American Statistician 44: 40-41.

APPENDIX

A proof that the probability of absence of a species in a universe given absence of the species in a simple random sample is approximately equal to the sampling fraction follows. Let ψ equal the probability of absence. Then:

$$\psi = \left(\sum_{k=0}^{N-n} L(K=k \mid X=0) \right)^{-1}$$

$$= \left(1 + \sum_{k=1}^{N-n} \frac{\binom{k}{0}\binom{N-k}{n}}{\binom{N}{n}} \right)^{-1}$$

$$= \left(1 + \sum_{k=1}^{N-n} \frac{(N-k)!}{n! * (N-n-k)!} * \frac{n! * (N-n)!}{N!} \right)^{-1}$$

$$= \left(1 + \sum_{k=1}^{N-n} \prod_{j=0}^{k-1} \frac{N-n-j}{N-j} \right)^{-1}$$

As an approximation, replace the product in the summation with an approximate power function in N and n. Then:

$$\psi = \left(1 + \sum_{k=1}^{N-n} \prod_{j=0}^{k-1} \frac{N-n-j}{N-j} \right)^{-1}$$

$$\approx \left(\sum_{k=0}^{N-n} \left(\frac{N-n}{N} \right)^{k} \right)^{-1}$$

$$= \left(\frac{1 - \left(\frac{N-n}{N} \right)^{N-n+1}}{1 - \frac{N-n}{N}} \right)^{-1}$$

As a second approximation, replace the power term in the numerator with zero. Then:

$$\psi = \left(\frac{1 - \left(\frac{N-n}{N} \right)^{N-n+1}}{1 - \frac{N-n}{N}} \right)^{-1}$$

$$\approx \left(\frac{1}{1 - \frac{N-n}{N}} \right)^{-1}$$

$$= 1 - \frac{N-n}{N}$$

$$= \frac{n}{N}$$