A Randomized Bootstrap Approach to Overcoming Model Selection Uncertainty

Fredrik Thunarf

Master's Thesis in Mathematical Statistics

Supervisors

Leif Nilsson Patrik Rydén

Umeå University

Department of Mathematics
and Mathematical Statistics

June 2009

Abstract

Statistical inference is traditionally based on the assumption that one single model is the true model, whereas in fact several models could fit the data equally well. Following this common practice means that model selection uncertainty is ignored, with the result of biased estimates and too narrow intervals.

In this thesis we study criteria for model selection and introduce a bootstrap approach where a model is selected at random, based on measures of fit of the contending models. This method is tested in a simulation study, where we estimate the 10th percentile, and is compared with an available method developed by Buckland *et al.* (1997). We also consider the situation where the true model is known and no model selection procedure is applied, with the purpose of evaluating the effect of model selection.

We find that model selection yields broader intervals with good coverage and that the parametric two-step variant of the proposed randomized bootstrap approach is an applicable method for overcoming model selection uncertainty.

Keywords: bootstrap; confidence intervals; model averaging; model selection uncertainty; percentile

En slumpbaserad bootstrap-metod
för att övervinna osäkerhet vid modellval

Statistisk inferens baseras traditionellt på antagandet att en enskild modell är den sanna, medan det i själva verket kan finnas flera likvärdiga modeller. Att följa detta gängse bruk innebär att den osäkerhet som är associerad med modellval ignoreras, med icke väntevärdesriktiga skattningar och för smala intervall som följd.

I det här examensarbetet studerar vi kriterier för modellval och presenterar en bootstrap-metod där en modell väljs slumpmässigt, baserat på de antagliga modellernas anpassningsgrad. Metoden testas i en simuleringsstudie, där vi skattar den 10:e percentilen, och jämförs med en befintlig metod, utvecklad av Buckland *et al.* (1997). Vi betraktar även fallet när den sanna modellen är känd och ingen modellvalsprocedur tillämpas, i avsikt att evaluera vilken påverkan modellval har.

Vi konstaterar att modellval ger bredare intervall med bra täckningsgrad och att den parametriska tvåstegsvarianten av den föreslagna slumpmetoden är en användbar metod för att övervinna modellvalsosäkerhet.

Nyckelord: bootstrap; konfidensintervall; model averaging; modellvalsosäkerhet; percentil

Preface

Acknowledgments

First, I would like to thank my supervisors, Leif Nilsson and Patrik Rydén, for their support, encouragement and guidance. Thanks also to Benny, Mattias and Daniel for convincing me to learn LaTeX.

Finally, since this master thesis concludes my university studies, I would like to thank all the people at the Department of Mathematics and Mathematical Statistics at Umeå University for their ever kind treatment and their readiness to answer questions.

Umeå, May 2009

Fredrik Thunarf

Contents

List of Tables

List of Figures

List of Algorithms

CHAPTER 1

Introduction

Given a random sample, we are interested in constructing a confidence interval for some parameter θ, e.g. a small percentile. If the distribution of the sample is known, we will be able to make inferences based on the truth. This is mostly true only in statistical textbooks; however, such inferences will be considered in this thesis, for comparative purposes.

In inferential statistics it is not uncommon to fit a number of models to the sample and evaluate them, e.g. by performing some kind of goodness-of-fit test. Once the best fitting model has been identified, it is assumed to be the true model and inferences are made conditional thereon. This may lead to biases in parameter estimates, underestimation of variances (since the model selection uncertainty component of the variance is omitted), and confidence and prediction intervals being too narrow and having a coverage below the desired level of confidence (Augustin *et al.*, 2005; Buchholz *et al.*, 2008; Buckland *et al.*, 1997; Burnham & Anderson, 2002).

Burnham & Anderson (2002, pp. 153–155) have identified three approaches to evaluating the uncertainty originating from model selection: (*a*) theoretical studies (Monte Carlo simulation), (*b*) bootstrapping, and (*c*) model averaging based on Akaike's information criterion.

Buchholz *et al.* (2008) compare Bayesian model averaging (see e.g. Hoeting *et al.*, 1999; Raftery *et al.*, 1997) with a two-step bootstrap model averaging (bootstrapMA) method of Augustin *et al.* (2005). They recommend the latter, which is in fact a modification of the bootstrapMA method proposed by Buckland *et al.*, 1997, but remark that the choice of methods might rather be a philosophical question.

1.1 Disposition

In chapter 2 we give an account of the methods used. The chapter is divided into two sections, each concluded by a simulation study where the methods are tested and evaluated. First, we cover the topics of distribution fitting and

principles for model selection. Then, we describe the model selection procedures and confidence intervals. In the third chapter we present the results of the two simulation studies, and finally the conclusions are presented in chapter 4.

1.2 Notation

Throughout this thesis, $\log(x)$ is used to denote the natural logarithm of x. If there is no ambiguity as to what the argument of the function is, the parentheses will be omitted. The exponential function, e^x, will be written as $\exp\{x\}$.

The capitalized greek letter gamma denotes the gamma function, defined as

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, \mathrm{d}t, \quad \alpha > 0.$$

If α is a positive integer, then $\Gamma(\alpha) = (\alpha - 1)!$ (Casella & Berger, 2002, p. 99). We write $\Gamma^n(k)$ to denote $\Gamma(k)$ raised to the nth power.

Subscripts enclosed in parentheses are used to indicate order statistics, e.g.

$$x_{(1)} = \min\{x_1, \ldots, x_n\}.$$

CHAPTER 2

Methods

2.1 Distribution fitting and model selection principles

2.1.1 Models

Given a set of observations from an unknown distribution, the first task would be to identify a set of K plausible models, according to some principle. This topic will not be covered in this thesis, however. Instead, we will consider three different models: the Weibull, gamma and lognormal distributions. Some properties of these distributions are described in Table 2.1.

Simulation of data

In R (R Development Core Team, 2007), random variables may be generated with built-in functions. We will thus use the functions `rweibull()`, `rgamma()` and `rlnorm()` to simulate data from the Weibull, gamma and lognormal distributions.

Weibull random variables are generated by means of the inverse transform method (see Ross, 1997, pp. 62ff.). Methods suggested by Ahrens & Dieter (1974, 1982) are used to simulate from the gamma distribution. Random variables from the lognormal distribution are obtained from normal deviates which are generated using an implementation of an algorithm described by Wichura (1988) and then transformed using the relation between the two distributions: if X is normally distributed, then $exp\{X\}$ has a lognormal distribution.

2.1.2 Distribution fitting

Fitting a distribution to a set of data means to find the set of parameters for which the theoretical distribution achieves maximum resemblance with the distribution of the data. There are several methods for estimating the parameters,

Table 2.1: *Some properties (probability density function (pdf), mean and variance) of the distributions used in this thesis.*

Weibull(k, λ)	
pdf	$f_X(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left\{-(x/\lambda)^k\right\},$ $x \geq 0, \quad k, \lambda > 0$
mean	$EX = \lambda\Gamma\left(1+\frac{1}{k}\right)$
variance	$\operatorname{var} X = \lambda^2\left(\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right)$
Gamma(k, λ)	
pdf	$f_X(x;k,\lambda) = \frac{\lambda^k}{\Gamma(k)} x^{k-1} \exp\{-\lambda x\},$ $x \geq 0, \quad k, \lambda > 0$
mean	$EX = \frac{k}{\lambda}$
variance	$\operatorname{var} X = \frac{k}{\lambda^2}$
Lognormal(μ, σ)	
pdf	$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\},$ $x \geq 0, \quad -\infty < \mu < \infty, \quad \sigma > 0$
mean	$EX = \exp\left\{\mu + \sigma^2/2\right\}$
variance	$\operatorname{var} X = \left(\exp\{\sigma^2\} - 1\right)\exp\{2\mu + \sigma^2\}$

e.g. the method of maximum likelihood and the method of moments. The first of these is the most commonly used and will be used in this thesis.

The method of maximum likelihood was introduced by R. A. Fisher in 1922 (Aldrich, 1997). Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a sample of independent and identically-distributed (i.i.d.) observations from a distribution with probability density function $f(x|\boldsymbol{\varphi})$, where $\boldsymbol{\varphi} = \varphi_1, \ldots \varphi_k$ are the distribution parameters. Then, the likelihood function is defined as

$$\mathcal{L}(\boldsymbol{\varphi}|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\varphi}).$$

As the name of the method suggests, estimators are found by maximizing the likelihood function. Because of the monotonicity of the logarithm function on $(0, \infty)$, this may also be done by maximizing the log likelihood function,

$$\mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}) = \log \mathcal{L}(\boldsymbol{\varphi}|\mathbf{x}), \tag{2.1}$$

which is easier to differentiate. By solving the system of simultaneous equations

$$\begin{cases} \frac{\partial}{\partial\varphi_1}\mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}) &= 0 \\ \qquad\qquad\vdots & \\ \frac{\partial}{\partial\varphi_k}\mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}) &= 0 \end{cases} \tag{2.2}$$

for $\varphi_1, \ldots, \varphi_k$, the maximum likelihood estimator (MLE) is obtained (Casella & Berger, 2002, pp. 315–317). It can be expressed as

$$\widehat{\boldsymbol{\varphi}}_{\text{MLE}} = \arg\max_{\boldsymbol{\varphi}} \mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}).$$

In R, maximum likelihood estimates can be computed using the function `fitdistr()` of the package `MASS` (Venables & Ripley, 2002).

Weibull

The likelihood and log likelihood functions for the Weibull distribution are given by

$$\mathcal{L}(\boldsymbol{\varphi}|\mathbf{x}) = k^n\lambda^{-nk}\prod_{i=1}^{n} x_i^{k-1}\exp\left\{-\lambda^{-k}\sum_{i=1}^{n}x_i^k\right\}$$

$$\mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}) = n\log k - nk\log\lambda + (k-1)\sum_{i=1}^{n}\log x_i - \lambda^{-k}\sum_{i=1}^{n}x_i^k.$$

The maximum likelihood estimators of the parameters of the Weibull distribution cannot be written as a closed-form expression; hence, a numerical approach is needed. Using the function `fitdistr()` in R does yield estimates numerically, but sometimes the computations fail, and the function does not seem optimal for estimating parameters of the Weibull distribution.

To avoid this, another method was used instead, *the grid approach*. This is an iterative method which may be described as placing a grid over the log likelihood function around the initial values of the parameters. Then, a new grid of smaller meshwidth is placed over the function around the vertex with the greatest function value. This procedure is then repeated until a desired tolerance requirement is met. A summary of this method is given in Algorithm 1.

We will use a square grid of size 7×7 and compute the starting values of the parameters in the same way as the R function `fitdistr()` does, i.e.

$$k_0 = \frac{1.2}{\sqrt{v}}, \qquad \lambda_0 = \exp\left\{m + \frac{0.572}{k_0}\right\},$$

where m and v are obtained by logarithmizing the sample and computing the corresponding sample mean and variance, respectively. The initial meshwidth will be

$$w = \frac{2\left(\min(k_0, \lambda_0) - \epsilon\right)}{n-1}, \tag{2.3}$$

where $\epsilon > 0$ is a constant introduced in order to avoid values of zero (where the natural logarithm function is not defined). We will use a tolerance level of 10^{-5} for the iterations.

Algorithm 1 The grid approach to maximum likelihood estimation

1. Set grid size (n), initial mesh-width (w), initial parameter values (k_0, λ_0), level of tolerance (δ) for the iterations and a counter $r := 1$.

2. Compute parameter vectors of length n, centered around the initial values:
$\mathbf{k} = \left(k_0 - (\frac{n-1}{2})w, k_0 - (\frac{n-3}{2})w, \ldots, k_0, \ldots, k_0 + (\frac{n-1}{2})w\right)$,
$\boldsymbol{\lambda} = \left(\lambda_0 - (\frac{n-1}{2})w, \lambda_0 - (\frac{n-3}{2})w, \ldots, \lambda_0, \ldots, \lambda_0 + (\frac{n-1}{2})w\right)$

3. Create an $n \times n$ matrix (*the grid*) of log likelihoods.
$\mathbf{L} = [L_{ij}], \quad L_{ij} = \mathcal{L}^*(k_i, \lambda_j | \mathbf{x}), \quad i, j = 1, \ldots, n$

4. Locate the maximum of $\mathbf{L}$ ($L_r := \max \mathbf{L}$) and record the corresponding parameter values, (k^*, λ^*).

5. Compute the modulus of the difference of the two latest maxima.
$D = |L_r - L_{r-1}|$

6. If $D > \delta$, then set $k_0 := k^*$, $\lambda_0 := \lambda^*$, $w := \dfrac{2w}{n-1}$, $r := r + 1$ and goto step 2.

7. Return the maximum likelihood estimator $\widehat{\boldsymbol{\varphi}}_{\text{MLE}} = (k^*, \lambda^*)$.

Gamma

The likelihood and log likelihood functions for the gamma distribution are given by

$$\mathcal{L}(\boldsymbol{\varphi}|\mathbf{x}) = \frac{\lambda^{nk}}{\Gamma^n(k)} \prod_{i=1}^{n} x_i^{k-1} \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\}$$

$$\mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}) = nk \log \lambda - n \log \Gamma(k) + (k-1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i.$$

Solving (2.2) for k and λ yields no closed-form expressions of the estimators; hence, a numerical approach is required. In R, the function `fitdistr()` optimizes the log likelihood for the gamma distribution using the Nelder-Mead method (Venables & Ripley, 2002). For smaller samples (e.g., 20 observations), the computations sometimes fail and we will for this reason use the grid method (see Algorithm 1) when fitting a gamma distribution to a small sample.

We will use the same gridsize and tolerance level as in the Weibull case. The starting values of the parameters will be computed in the same way as by the R function `fitdistr()`:

$$k_0 = \frac{m^2}{v}, \qquad \lambda_0 = \frac{m}{v},$$

where m and v are the sample mean and variance, and the meshwidth will be computed using equation (2.3).

Lognormal

The likelihood and log likelihood functions for the lognormal distribution are given by

$$\mathcal{L}(\boldsymbol{\varphi}|\mathbf{x}) = \sigma^{-n}(2\pi)^{-n/2}\prod_{i=1}^{n}\frac{1}{x_i}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu)^2\right\}$$
$$\mathcal{L}^*(\boldsymbol{\varphi}|\mathbf{x}) = -n\log\sigma - \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n}\log x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu)^2.$$

Solving (2.2) for μ and σ yields the following maximum likelihood estimators for the parameters of a lognormal distribution:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\log X_i$$
$$\hat{\sigma}_{\text{MLE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log X_i - \hat{\mu}_{\text{MLE}})^2}.$$

In R, the function `fitdistr()` uses these closed-form expressions to compute the parameters (Venables & Ripley, 2002).

2.1.3 Model selection principles

Goodness-of-fit tests

Goodness-of-fit is a measure of how well a model fits the data. There are several different ways to test the goodness-of-fit, e.g. the Kolmogorov test and the chi-squared test, which are considered in this study and described below. The null hypothesis, H_0, is that a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ from an unknown distribution $F_X(x)$ follows a stated distribution $F_0(x)$ – the null distribution. The parameters of $F_0(x)$ may be unknown, in which case only the mathematical form of the distribution is specified (a composite hypothesis). The most common alternative hypothesis, H_A, for a goodness-of-fit test is simply that the null hypothesis is false. The hypotheses may be written as

$$\begin{aligned} H_0: &\quad F_X(x) = F_0(x) \\ H_A: &\quad F_X(x) \neq F_0(x). \end{aligned}$$

When performing goodness-of-fit tests, not rejecting the null hypothesis is more interesting than rejecting it, in contrast to other statistical test situations where rejection of the null hypothesis in most cases is the desired outcome of the test (D'Agostino & Stephens, 1986, p. 1).

The Kolmogorov goodness-of-fit test is based upon the difference between the empirical cumulative distribution function of the sample and the cumulative distribution function of the null distribution. The test statistic is given by

$$D = \max(D^+, D^-),$$

where

$$D^+ = \max_{1 \le i \le n} \left(\frac{i}{n} - z_i \right), \qquad D^- = \max_{1 \le i \le n} \left(z_i - \frac{i-1}{n} \right),$$

where

$$z_i = F_0(x_{(i)} | \varphi_1, \ldots, \varphi_k)$$

for $i = 1, \ldots, n$ (Stephens, 1970, 1974). The p-value of the test is given by $P_0\{D \ge d\}$, where d is the obtained value of D and P_0 denotes the use of the distribution of the null hypothesis when computing the probability. Furthermore, $P_0\{D \ge d\}$ is the same for any continuous distribution (Ross, 2006, pp. 222–225).

In R, the p-value is computed using a method described by Marsaglia *et al.* (2003) when using the built-in function `ks.test()` to perform a Kolmogorov goodness-of-fit test. However, exact p-values cannot be computed for samples containing ties (R Development Core Team, 2007).

The Kolmogorov test is not very well adapted for situations where the parameters need to be estimated from the data, although it is possible to compute a test statistic in the same way as above, but with $\{z_i\}$ based on an estimate $\widehat{\boldsymbol{\varphi}}$ of $\boldsymbol{\varphi}$ instead of predefined parameter values:

$$z_i = F_0(x_{(i)} | \widehat{\varphi}_1, \ldots, \widehat{\varphi}_k)$$

for $i = 1, \ldots, n$ (Durbin, 1973, pp. 47–48). In this case, the p-value can be roughly approximated by $P_U\{D \ge d\}$, where P_U denotes the use of the standard uniform distribution, i.e. the uniform distribution on the interval $(0, 1)$. This will give an overestimation of the p-value and if the obtained estimate is small, the p-value should be estimated more accurately by simulation (Ross, 1997, pp. 201–202).

In Pearson's chi-square test (the chi-squared goodness-of-fit test) the outcome space is divided into k disjoint classes. The test is based on the discrepancies between the number of observations, o_i, and the expected frequency (under H_0), e_i, for each class. The expected frequency for a class is the product of the sample size and the probability (under H_0) for an observation to belong to that class. The test statistic is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

and is asymptotically chi-square distributed with $(k-1-s)$ degrees of freedom, where s is the number of estimated parameters, if the null hypothesis is true (Snedecor & Cochran, 1980, pp. 75–78). The p-value of the test is given by $P(\chi^2 \ge x^2_{\text{obs}})$, where x^2_{obs} denotes the observed value of the test statistic.

Using equiprobable classes, i.e. where $P(\text{obs.} \in i\text{th class}) = 1/k$ for all k classes, as suggested by Mann & Wald (1942), yields an unbiased test (Cohen & Sackrowitz, 1975). Furthermore, the χ^2-distribution is a more accurate approximation to the distribution of the test statistic for equiprobable classes than for

unequiprobable classes (Moore, 1986, pp. 69–70). Hence, we will only consider equiprobable classes.

There are several recommendations on the choice of the number of classes. A commonly accepted rule of thumb for the choice of the number of classes is that no class should have an expected frequency below 1 and that at least 80 percent of the classes should have an expected frequency of at least 5 (Cochran, 1954; Moore, 1986, pp. 69–70). This means that the expected frequency for equiprobable classes should not fall below 5.

However, others dismiss this rule of thumb as being too conservative, e.g. Roscoe & Byars (1971) whose recommendation for $\alpha = 0.05$ is to choose equiprobable cells with an average expected frequency of at least 1.

One recommendation for the choice of the number of classes is

$$k = 4\left(\frac{2(n-1)^2}{c(\alpha)^2}\right)^{1/5}, \tag{2.4}$$

where $c(\alpha) = \Phi^{-1}(1-\alpha)$, i.e. the $1-\alpha$ quantile of the standard normal distribution (Mann & Wald, 1942).

However, this value is greater than the optimal number of classes (Schorr, 1974), and halving it yields no greater loss of power (Rayner & Best, 1989, pp. 23ff.; Williams, 1950). Moore (1986, pp. 69–70) takes this into account and recommend choosing the number of classes so that it falls between the value of (2.4) for $\alpha = 0.05$ and half that value. The 'convenient choice'

$$k = 2n^{2/5} \tag{2.5}$$

is suggested.

Finally, we will consider a recommendation mentioned by Greenwood & Nikulin (1996, pp. 39–40):

$$k \leq \min\left(\frac{1}{\alpha}, \log n\right). \tag{2.6}$$

Since the the number of classes affects the degrees of freedom, which must be at least 1, an absolute lower bound for the number of classes is given by

$$k \geq s + 2.$$

This means that with two estimated parameters the number of classes must be at least 4. If we apply the rule of thumb, this is in fact the only possible choice for samples of size 20, whereas on the other hand we may have up to 20 classes for samples of size 100.

The three recommendations described above are evaluated in Table 2.2, using samples of size 100. We can see that the first recommendation is not compliant with the rule of thumb mentioned earlier. The third recommendation results in very few classes and the second yields a number of classes between those of the other two.

Since no study on the subject has been established as norm, we have simply decided arbitrarily to follow the recommendation of Moore (1986, pp. 69–70), see equation (2.5); hence, we will use 13 equiprobable classes for samples of size 100.

Table 2.2: *The number of equiprobable classes (k), expected frequencies per class (e_i) and the corresponding number of degrees of freedom (df) for a chi-squared goodness-of-fit test at significance level 0.05 with samples of size 100.*

Formula	k	e_i	df
Mann & Wald (2.4)	24	4.17	21
Moore (2.5)	13	7.69	10
Greenwood & Nikulin (2.6)	4	25.0	1

Information criteria

Buckland *et al.* (1997) remark that 'there are problems in developing a general approach based on hypothesis tests' and suggest using likelihood-based information criteria instead. Such an information criterion can be written as

$$I = -2\mathcal{L}^*(\widehat{\boldsymbol{\varphi}}|\mathbf{x}) + q,$$

where $\mathcal{L}^*$ is the log likelihood function (2.1) and q is a penalty coefficient based on the number of parameters of the model and/or the number of observations in the sample.

The information criteria most commonly used for model selection are, according to Reschenhofer (1996), Akaike's information criterion (AIC) and the Bayesian information criterion (BIC):

$$AIC = -2\mathcal{L}^*(\widehat{\boldsymbol{\varphi}}|\mathbf{x}) + 2s$$

$$BIC = -2\mathcal{L}^*(\widehat{\boldsymbol{\varphi}}|\mathbf{x}) + s\log n,$$

where s is the number of parameters and n the number of observations.

Burnham & Anderson (2004) deal thoroughly with these two information criteria and their differences, and stress that they were designed to answer different questions. We will use the Akaike criterion in this thesis.

By normalizing the information criteria, we obtain weights:

$$w_k = \frac{\exp\{-I_k/2\}}{\sum_{i=1}^{K}\exp\{-I_i/2\}}, \qquad k = 1, \dots K, \tag{2.7}$$

where I_k is an information criterion for model k. When AIC is used, these weights are called Akaike weights (Burnham & Anderson, 2004). If all models have the same penalty coefficient q, then (2.7) can be simplified as

$$w_k = \frac{\mathcal{L}_k}{\sum_{j=1}^{K}\mathcal{L}_j}.$$

Let θ denote some parameter that is defined for all contending models, e.g. a percentile. A point estimate of θ can be obtained by fitting each of the contending models to the original data and computing the corresponding parametric estimates $\{\hat{\theta}_k\}$ of θ. We define the model averaging estimate of θ as the weighted mean

$$\hat{\theta}_{\mathrm{MA}} = \sum_{k=1}^{K} w_k\hat{\theta}_k, \tag{2.8}$$

where w_k is the Akaike weight for model k, given by (2.7), and $\hat{\theta}_k$ is the parametric estimate of θ for model k (Buckland *et al.*, 1997).

2.1.4 Simulation study I

In order to decide whether to base the model selection procedure on hypothesis tests or information criteria, a simple simulation study will be carried out. We will use data from known distributions, make assumptions about the true distribution and compute the different measures of goodness-of-fit. This yields nine possible combinations of true and assumed distributions. For each sample and test, we will make the assumption that the best fitting model (i.e. the model yielding the highest p-value or lowest AIC value) is the true model. Two sample sizes will be considered: 100 and 20, and in each case we will simulate 1000 samples from each distribution.

For comparative purposes, parameters will be chosen so that all distributions have the same mean and the same variance. The parameters of the Weibull distribution cannot be expressed explicitly in terms of the mean and variance whereas this is possible for the gamma and lognormal distributions. Consequently, the parameters for the Weibull distribution need to be chosen before the parameters for other distributions can be computed.

We will use the following distributions:

- Weibull$(3, 2)$
- Gamma$(7.570403, 4.238845)$
- Lognormal$(0.5179213, 0.3522335)$,

each of which has mean 1.785959 and variance 0.4213315. Their probability density functions and cumulative density functions are shown in Figure 2.1. As can be seen, the Weibull distribution is somewhat less similar to the gamma and lognormal distributions than these are to each other.

This simulation study will be carried out as described in Algorithm 2.

Algorithm 2 Simulation study I: Model selection principles

1. Generate N n-sized samples.
2. Repeat for each sample:
   - 2.1 Fit (all three) models to the sample.
   - 2.2 Perform a Kolmogorov test for each model.
   - 2.3 Perform a chi-squared test for each model.
   - 2.4 Compute AIC for each model.
3. For each test, compute the number of times each model was considered to be the 'best' model.
4. Repeat steps 1–3 for all three distributions.

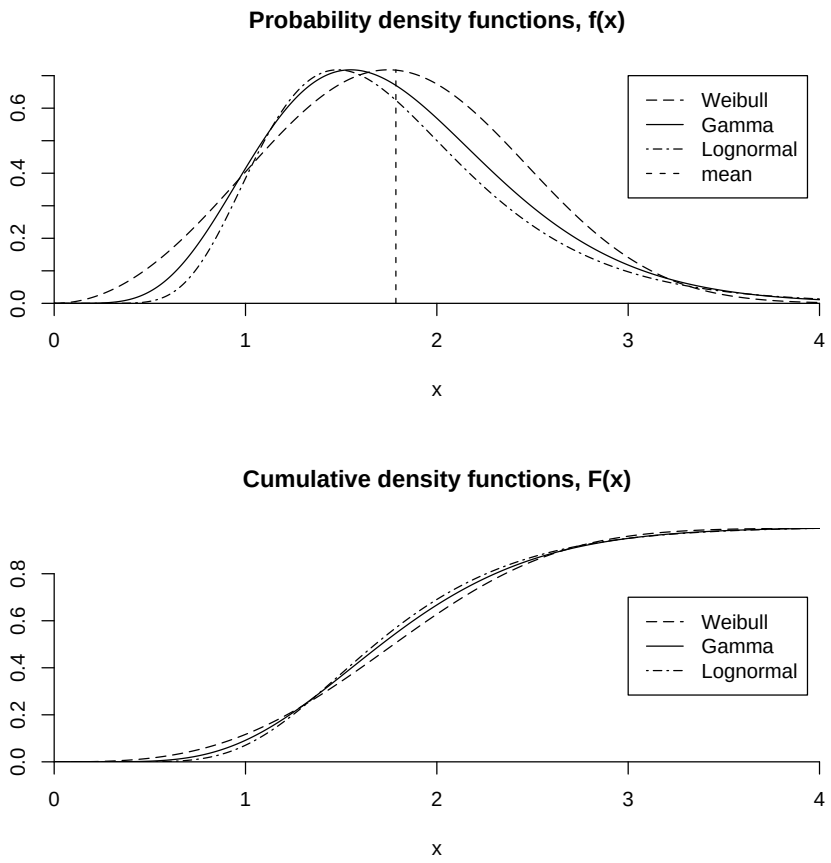

Figure 2.1: *A plot of the probability and cumulative density functions for the distributions used in the model selection principles simulation study.*

2.2 Model selection methods and confidence intervals

2.2.1 The bootstrap

The non-parametric bootstrap

Given a random sample $\mathbf{x} = (x_1, \ldots, x_n)$ from an unknown distribution $\mathcal{F}$, we can generate B equisized bootstrap samples by drawing with replacement from the original sample. We write $\mathbf{x} \rightarrow \mathbf{x}_b^* = (x_{b1}^*, \ldots, x_{bn}^*)$ to denote that $\mathbf{x}_b^*$ is the bth bootstrap sample based on $\mathbf{x}$.

The estimator for some parameter θ is given by

$$\hat{\theta}^* = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^*, \tag{2.9}$$

where $\hat{\theta}_b^*$ is the bth bootstrap replication of θ, i.e. an estimate of θ based on the bth bootstrap sample. The sample standard deviation of the bootstrap replications can be used to estimate the standard error of the θ estimate (Efron & Tibshirani, 1993, pp. 45–49).

The parametric bootstrap

Instead of resampling from the original sample, we may assume that the data has a certain probability distribution with estimated parameters and generate B samples from this distribution instead. This is called parametric bootstrapping. The samples, once generated, are treated as in the non-parametric case and we may compute estimates and their standard errors in the same way, see (2.9) (Efron & Tibshirani, 1993, pp. 53–56).

2.2.2 Model selection methods

The Buckland method

Buckland *et al.* (1997) suggest a non-parametric bootstrap approach where the model selection procedure is applied independently to each bootstrap sample. Akaike's information criterion is computed for all K contending models and the model with the best (i.e. smallest) value is selected. The method will be referred to as *the Buckland method* in this thesis and is described in Algorithm 3. This method has been developed further by Augustin *et al.* (2005) who suggest a two-step approach for situations with a larger number of contending models. The first step has the purpose of reducing the number of models before the rest of the model selection procedure is carried out.

Randomized Model Selection

What we have chosen to call *The method of randomized model selection* (RMS) is a bootstrap approach in which the model is selected at random from the set of possible models. We will consider both non-parametric and parametric bootstrap samples. The two corresponding variants of RMS are described below.

Algorithm 3 The Buckland method

1. Bootstrapping. Repeat this step B times.
   - 1.1 Generate a non-parametric bootstrap sample.
     $\mathbf{x} \to \mathbf{x}_b^*$
   - 1.2 Fit models to the bootstrap sample.
     $\mathbf{x}_b^* \to \{\hat{F}_{bk}^*\}$
   - 1.3 Compute AIC.
     $\{\hat{F}_{bk}^*\} \to \{AIC_{bk}^*\}$
   - 1.4 Model selection. Select the model with the smallest AIC value.
     $\min\{AIC_{bk}^*\} \to \hat{F}_b^*$
   - 1.5 Estimate θ.
     $\hat{\theta}_b^* = g(\hat{F}_b^*)$
2. Construct confidence interval from $\{\hat{\theta}_b^*\}$.

In the one-step (or non-parametric) method of randomized model selection, which will be denoted by RMS_1, the bootstrap samples are generated non-parametrically from the original sample.

For each bootstrap sample, we begin by fitting each of the K models and compute the values of AIC, from which Akaike weights are computed, see (2.7). These weights have the property that $\sum_k w_k = 1$, and regarding them as probabilities we may select a model at random by means of the discrete inverse transform method (see Ross, 1997, pp. 62ff.). Finally, θ is estimated parametrically from the selected model.

This method is summarized in Algorithm 4.

Algorithm 4 One-step randomized model selection

1. Bootstrapping. Repeat this step B times.
   - 1.1 Generate a non-parametric bootstrap sample.
     $\mathbf{x} \to \mathbf{x}_b^*$
   - 1.2 Fit models to the bootstrap sample.
     $\mathbf{x}_b^* \to \{\hat{F}_{bk}^*\}$
   - 1.3 Compute AIC and Akaike weights.
     $\{\hat{F}_{bk}^*\} \to \{AIC_{bk}^*\} \to \{w_{bk}^*\}$
   - 1.4 Model selection. Select a model at random.
     $\{w_{bk}^*\} \to \hat{F}_b^*$
   - 1.5 Estimate θ.
     $\hat{\theta}_b^* = g(\hat{F}_b^*)$
2. Construct confidence interval from $\{\hat{\theta}_b^*\}$.

The two-step method of randomized model selection (RMS_2) is a parametric

variant of RMS_1. As the name suggests, the model selection procedure is carried out in two steps.

In the first step, a model is selected at random from the set of possible models, in the same way as in RMS_1. Since no bootstrap sample is available at the time, the selecting probabilities are based on the original sample. The selected model is used to generate a parametric bootstrap sample; thus, we will call it the *generating model* and denote its distribution function by $\hat{G}_b^*$.

In the second step, the generating distribution is refitted to the bootstrap sample. This gives us another distribution (of the same mathematical form but with different parameters), which will be referred to as the *selected model* and is used to estimate θ parametrically.

This method is summarized in Algorithm 5.

Algorithm 5 Two-step randomized model selection

1. Fit models to the original sample.
   $\mathbf{x} \to \{\hat{F}_k\}$
2. Compute AIC and Akaike weights.
   $\{\hat{F}_k\} \to \{AIC_k\} \to \{w_k\}$
3. Bootstrapping. Repeat this step B times.
   - 3.1 Model selection I. Select a (generating) model at random.
     $\{w_k\} \to \hat{G}_b^*$
   - 3.2 Generate a parametric bootstrap sample.
     $\hat{G}_b^* \to \mathbf{x}_b^*$
   - 3.3 Model selection II. Refit the selected model.
     $\mathbf{x}_b^* \to \hat{F}_b^*$
   - 3.4 Estimate θ.
     $\hat{\theta}_b^* = g(\hat{F}_b^*)$
4. Construct confidence interval from $\{\hat{\theta}_b^*\}$.

2.2.3 Confidence intervals

Let $\hat{\theta}_L$ and $\hat{\theta}_U$ denote the lower and upper bounds of a confidence interval for the estimate of some parameter θ. Then, the coverage probability of a confidence interval is given by

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U).$$

When the true value of θ lies outside of the interval, we will refer to the event as a *miss*. Furthermore, a miss is the union of the two disjoint events *miss left*, defined as 'the confidence interval lies completely to the left of the true value', and *miss right*, defined analogously (Boos & Hughes-Oliver, 2000). Figure 2.2 graphically illustrates these events. If the corresponding *miss probabilities*

$$P(\theta < \hat{\theta}_L) \quad \text{and} \quad P(\theta > \hat{\theta}_U)$$

are equal, the confidence interval is called equal-tailed. This is the most common case (Efron & Tibshirani, 1993, pp. 155–156).

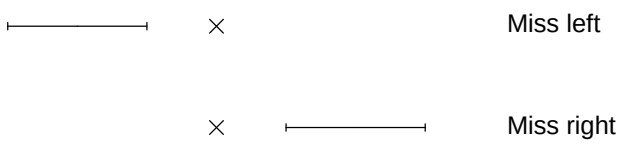

Figure 2.2: *The possible cases when a confidence interval misses (i.e. does not cover) the true value (×) of the parameter of interest.*

The length of a confidence interval is defined as the distance between its bounds:

$$\hat{\theta}_U - \hat{\theta}_L.$$

Standard normal confidence intervals

The standard normal confidence interval is the simplest confidence interval, obtained by using the standard normal distribution as an approximation to the distribution of the parameter of interest. The bounds for a $100 \cdot (1 - 2\alpha)\,\%$ approximate standard normal confidence interval are given by

$$\hat{\theta}_L = \hat{\theta} - z_\alpha \cdot \widehat{se}(\hat{\theta}), \qquad \hat{\theta}_U = \hat{\theta} + z_\alpha \cdot \widehat{se}(\hat{\theta}),$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$ is the $(1 - \alpha)$th quantile of the standard normal distribution (e.g. $z_{.025} \approx 1.96$). This interval has the drawback of always being symmetric around $\hat{\theta}$.

Percentile intervals

An improved interval is the percentile interval, which is based on the empirical percentiles of the bootstrap replicates (Efron & Tibshirani, 1993, pp. 170ff.). Given B bootstrap samples, we order the bootstrap replicates from smallest to largest and choose the $B\alpha$th and $B(1 - \alpha)$th replicates as bounds for the confidence interval. In other words, the bounds for a $100 \cdot (1 - 2\alpha)\,\%$ percentile interval are given by

$$\hat{\theta}_L = \hat{\theta}^*_{(B\alpha)}, \qquad \hat{\theta}_U = \hat{\theta}^*_{(B(1-\alpha))}.$$

Blom & Holmquist (1998, p. 280) mention a slightly different way of constructing percentile confidence intervals, using the percentiles of the distribution of $\hat{\theta}^* - \hat{\theta}$ as an approximation of those of the distribution of $\hat{\theta} - \theta$. The bounds for such a $100 \cdot (1 - 2\alpha)\,\%$ percentile interval are given by

$$\hat{\theta}_L = 2\hat{\theta} - \hat{\theta}^*_{(B(1-\alpha))}, \qquad \hat{\theta}_U = 2\hat{\theta} - \hat{\theta}^*_{(B\alpha)},$$

where we will use the model averaging estimate (2.8) for $\hat{\theta}$.

In this thesis, we will refer to these intervals as PCTL_1 and PCTL_2 intervals.

Improved percentile intervals

Efron & Tibshirani (1993, pp. 178–188) remark on the unsatisfactory coverage properties of percentile intervals and present an improved version called $\mathrm{BC_a}$, *bias-corrected and accelerated* intervals. Buckland *et al.* (1997) present another approach to improving the percentile interval: the *weighted percentile method*. As in the regular percentile method, the bounds of the confidence interval are based on the empirical percentiles of the bootstrap replicates.

In the $\mathrm{BC_a}$ intervals, the empirical percentiles of the bootstrap replicates are modified by two quantities: bias-correction and acceleration. The first is a measure of the discrepancy between the median of the bootstrap replicates and the estimate of θ based on the original sample $\mathbf{x}$, and the latter is a measure of the rate of change of the standard error of the estimate $\hat{\theta}$ with respect to the true value of θ.

The bias-correction, $\hat{z}_0$, is given by the proportion of bootstrap replicates less than the estimate $\hat{\theta}$ of θ based on $\mathbf{x}$, expressed in normal units:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}_b^* < \hat{\theta}\}}{B}\right),$$

where Φ^{-1} is the quantile function for the standard normal distribution, and we will use the model averaging estimate $\hat{\theta}_{\mathrm{MA}}$ (2.8) for $\hat{\theta}$.

The acceleration can be computed in the following way: Let $\mathbf{x}_{*i} = (x_1, \ldots, x_{i-1}, x_{i+1} \ldots, x_n)$ denote the ith jackknife sample of $\mathbf{x}$, $\hat{\theta}_{*i}$ the estimate of θ based on $\mathbf{x}_{*i}$ and define $\hat{\theta}_* = \frac{1}{n}\sum_{i=1}^n \hat{\theta}_{*i}$. Then, the acceleration is given by

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_* - \hat{\theta}_{*i})^3}{6\left(\sum_{i=1}^n (\hat{\theta}_* - \hat{\theta}_{*i})^2\right)^{3/2}}.$$

The bounds of the $\mathrm{BC_a}$ confidence interval are given by

$$\hat{\theta}_L = \hat{\theta}^*_{(B\alpha_1)}, \qquad \hat{\theta}_U = \hat{\theta}^*_{(B\alpha_2)},$$

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 - z_\alpha}{1 - \hat{a}(\hat{z}_0 - z_\alpha)}\right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_\alpha}{1 - \hat{a}(\hat{z}_0 + z_\alpha)}\right).$$

In the weighted percentile method, weights are assigned to the bootstrap samples before the confidence interval is constructed. The weight for a bootstrap sample depends on the selected model and is defined as the ratio of the theoretical proportion of bootstrap samples in which that model *would* be selected to the proportion of bootstrap samples in which it actually was selected. This means that the bootstrap samples for which a model *selected less often than expected* is selected are given more importance in the construction of the interval.

We begin by computing the Akaike weights w_k for the models, see (2.7). Then, we let $B_k, k = 1, \ldots, K$, denote the number of bootstrap samples in

which model k was selected and define the weight for the bth bootstrap sample as

$$v_b = \frac{w_k}{B/B_k},$$

where k indicates the selected model.

Using these weights, we can obtain a weighted percentile confidence interval based on the ordered bootstrap replicates. The bounds for such an interval with coverage probability $100 \cdot (1 - 2\alpha)\,\%$ are given by

$$\hat{\theta}_L = \hat{\theta}^*_{(r)}, \qquad \hat{\theta}_U = \hat{\theta}^*_{(s)},$$

where

$$r = \max\left\{t \leq B \in \mathbb{Z}_+ : \textstyle\sum_{b=1}^{t} v_b \leq (B+1)\alpha\right\},$$

$$s = \min\left\{t \leq B \in \mathbb{Z}_+ : \textstyle\sum_{b=1}^{t} v_b \geq (B+1)(1-\alpha)\right\}.$$

2.2.4 Simulation study II

To compare the model selection methods and the different kinds of confidence intervals, we will carry out a simulation study and compute interval estimates for the 10th percentile. The following distributions will be used:

- Weibull(3,2)
- Gamma(4,7)
- Lognormal(0.2,0.3),

the parameters of which have been chosen arbitrarily. Their probability density functions and cumulative density functions are shown in Figure 2.3. For each of these distributions we will consider two different sample sizes (100 and 20), which means that we will have six different cases.

First, we will generate a set of 1000 independent samples from the distribution. Then, for each sample we will generate 1000 bootstrap samples, nonparametrically. By applying the Buckland method and RMS_1 – separately – to each of the bootstrap samples[1] we will obtain two different sets of 1000 bootstrap replicates of the desired percentile. We will also apply RMS_2 to the original sample and obtain yet another set of bootstrap replicates; thus, we will have a total of three such sets for each original sample.

Next, we will use each set of bootstrap replicates to construct PCTL_1, PCTL_2, weighted and BC_a intervals at a nominal confidence level of 95 %. This means that we will obtain 1000 confidence intervals for each of the 12 combinations of methods and interval types.

Finally, considering one combination at the time, we will estimate the coverage probability by the proportion of intervals that contains the true value. Likewise, we will estimate the miss left and miss right probabilities. Furthermore, the average length of the intervals will be computed.

(See also Algorithm 6 for a summary.)

[1] It should be noted that both methods will be applied to the *same* bootstrap samples. This means that we will perform step 1.1 in Algorithms 3 and 4 (with repetition: 1000 times) before moving on to step 1.

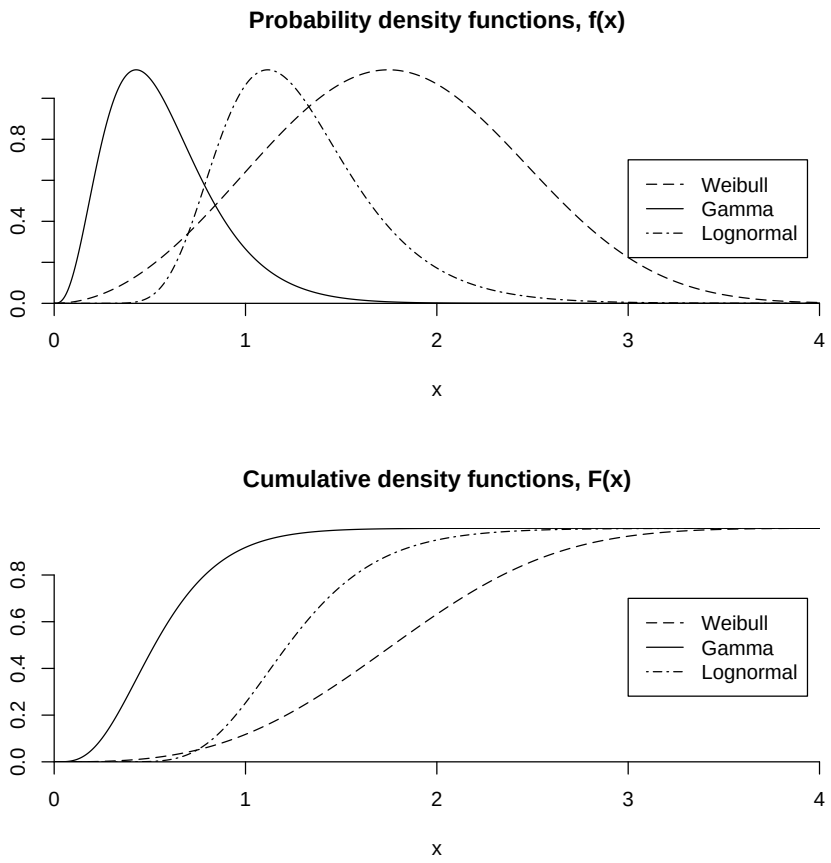

Figure 2.3: *A plot of the probability and cumulative density functions for the distributions used in the model selection methods simulation study.*

Algorithm 6 Simulation study II: Model selection methods

1. Generate 1000 samples of size 100 or 20 from the true distribution.
2. For each sample:
   - 2.1 Generate 1000 bootstrap samples non-parametrically
   - 2.2 Apply the Buckland method to the bootstrap samples.
   - 2.3 Apply RMS_1 to the bootstrap samples.
   - 2.4 Apply RMS_2 to the original sample
   - 2.5 For each set of BS replicates:
     - i. Construct a $PCTL_1$ interval.
     - ii. Construct a $PCTL_2$ interval.
     - iii. Construct a weighted percentile interval.
     - iv. Construct a BC_a interval.
3. For each combination of methods and interval types:
   - 3.1 Compute the average interval length.
   - 3.2 Compute the miss left and miss right probabilities.
   - 3.3 Compute the coverage probability.

In order to evaluate the effect of applying model selection methods, we will also carry out a regular parametric bootstrap where the model is known (i.e. the the mathematical form of the true distribution). Using 1000 samples from the Gamma(4,7) distribution, we will for each sample (*a*) fit parameters, (*b*) generate 1000 bootstrap samples, (*c*) refit the model to each bootstrap sample and estimate the percentile parametrically, and (*d*) construct confidence intervals.

CHAPTER 3

Results

3.1 Simulation study I

Detailed results (tables and figures) can be found in Appendices A–B.

For the samples of size 100 (see Tables A.1–A.3), the Akaike information criterion yielded greater numbers of correct highest ranks (for the true distribution being the best fit) than did the goodness-of-fit tests, of which the Kolmogorov test yielded better results. It can be seen that it did occur only a few times that the lognormal distribution was the best fit for a Weibull sample (and vice versa); the gamma distribution was more likely to be (erroneously) considered as being the best fit. Also, the error ratio was higher for the samples having gamma as the true distribution.

Equally, the information criterion yielded the best result also for the smaller sample size $n = 20$ (see Tables A.4–A.6), followed again by the Kolmogorov test. Compared to the larger samples, the number of incorrect assumptions is notably higher – in particular for the gamma distribution, where erroneous assumptions were made in 3 out of 4 samples, irrespective of method. Furthermore, there was less difference between the Weibull and lognormal distributions.

If we look at the estimated probability density functions of the test statistics (see Figures B.1–B.5), we can see that the curves have lower peaks and heavier right tails in the cases where the assumptions about the true distribution were wrong – except for the chi-squared test and small samples, where it is difficult to see any differences between the graphs. Similarly, the AIC graphs (see Figures B.3–B.6) show little difference.

The density estimates of the Kolmogorov test p-values (see Figures B.7–B.9) show that we do not tend to reject the hypothesis in the correct cases. The willingness to reject the hypothesis in the incorrect cases depends on the sample size. The graphs for the p-values of the chi-squared test (see Figures B.8–B.10) are more interesting. Whereas the curves are flat for the larger samples and almost look like the probability density function of the Uniform(0,1) distribution – with the exception of the Weibull–lognormal cases (see the second paragraph

of this section) –, they go up and down and show a much greater tendency to rejection of the null hypotheses for $n = 20$.

3.2 Simulation study II

Detailed results can be found in the tables in Appendix C. The confidence intervals (based on the average values of the interval bounds) are also illustrated graphically[1], see Appendix D.

We begin by looking at the two Weibull cases (see Tables C.1 and C.5). The RMS_2 intervals are slightly shorter than those of the other methods. The difference in coverage between the methods are small when $n = 100$. For the small samples, the randomized methods yield higher coverage rates than does the Buckland method; the highest coverage is achieved by RMS_2. This order among the methods can also be seen in the gamma and lognormal cases (see Tables C.2–C.3 and C.6–C.7), irrespective of sample size, not only with reference to the coverage but also the interval length. Some estimated values of the coverage probability exceed the nominal confidence level of 95 %, in particular for RMS_2 in the lognormal cases.

Considering the case in which we knew the true distribution (see Table C.4), we see that the intervals are shorter compared with when any of the RMS methods was applied and that the coverage probabilities are lower. However, compared with the Buckland method the intervals are longer, but the differences in coverage are small. Interestingly, the weighted percentile interval has much less coverage, despite being remarkably wider, in the case of no model selection method applied.

Finally, we will make some comments on the results for the different types of confidence intervals. The two regular percentile intervals are of equal length in all cases, and generally shorter than the improved interval types. For the small samples from gamma or lognormal distributions, BC_a and $PCTL_2$ intervals yield greater coverage than the other two intervals. Looking at the miss probabilities for $PCTL_1$ and weighted intervals we see that the miss right probability is much greater than that of miss left, which means that these intervals are more inclined to overestimating the parameter of interest. Contrariwise, BC_a and $PCTL_2$ intervals tend to underestimate the value (exception: Weibull) although the difference between the miss probabilities is smaller.

Such tendencies can also be seen in the interval plots (see Figures D.1–D.7).

[1] The intervals are plotted using the R function `plotCI()` from the `plotrix` package, see Lemon *et al.* (2008).

CHAPTER 4

Discussion

It could be argued that it would be possible to improve on the performance of the chi-squared test by choosing the classes differently. In any case, optimizing Pearson's test is not an aim of this thesis, and we will therefore disregard the possibility of improvement. As was established in section 3.1, the best results were obtained using AIC, and we thus advocate the use thereof as model selection principle.

The high number of incorrect assumptions that occurred for the samples of size 20 can probably be accounted for by the sample size. Naturally, fitting a distribution is more difficult with fewer observations. It is also reasonable to believe that the small sample size accounts for the appearance of the density estimates of the chi-squared test p-values.

The use of model selection procedures has an evident effect on the confidence intervals. On the one hand, the lengths of the intervals increase, but on the other hand we get higher coverage probabilities. That the Buckland method resulted in a shorter interval without any notable difference in coverage disagrees with this conclusion; however, this is in all probability an accidental occurence, since *a priori* knowledge of the true model should yield a better result. It would be reasonable to believe that the odd behaviour of the weighted percentile intervals in the regular parametric bootstrap case can be explained by the fact that Buckland *et al.* (1997) suggest using it combined with model selection.

We find it natural that the two-step RMS performs better than the one-step version, since the first step (selecting a generating model for the bootstrap samples) provides additional information that is used when selecting a model in the second step.

What kind of confidence interval we use seem to be of little importance. However, if a situation required that the parameter of interest would not be overestimated, $\mathrm{BC_a}$ and $\mathrm{PCTL_2}$ intervals might be preferred over $\mathrm{PCTL_1}$ and weighted intervals – and conversely if underestimating should be avoided. While it may be true that the first two interval types are better suited for small samples than are the last two (cf. section 3.2), we may not say so for sure.

Of course, the sample size affects the confidence intervals: the fewer the observations the longer the interval needs to be in order to achieve the same coverage probability. Throughout this study, the intervals of the larger samples had somewhat better coverage than those of the smaller samples.

It should be stressed that we have limited ourselves to studying a case of three plausible models. Whereas this might seem like a small number, in order to deal with more models it would have been necessary to introduce some criterion for determining the plausibility of a model, but that is beyond the scope of this thesis. Nevertheless, it would be interesting to extend the methods we have introduced for use in situations with a larger number of contending models.

Moreover, we have evaluated the methods for only one set of parameters for each distribution. How they would perform on distributions with other parameters can not be answered here, but is of course an interesting question. Equally important, we have focused only on estimating the 10th percentile. Using these methods to estimate the mean value or the median would probably yield good results, whereas it is difficult to say whether they maintain their level of performance for more extreme percentiles – or extreme quantiles.

Bibliography

Ahrens, J. H., & Dieter, U. 1974. Computer methods for sampling from gamma, beta, Poisson, and binomial distributions. *Computing*, **12**, 223–246.

Ahrens, J. H., & Dieter, U. 1982. Generating gamma variates by a modified rejection technique. *Communications of the ACM*, **25**(1), 47–54.

Aldrich, J. 1997. R. A. Fisher and the making of maximum likelihood 1912–1922. *Statistical Science*, **12**(3), 162–176.

Augustin, N., Sauerbrei, W., & Schumacher, M. 2005. The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modeling*, **5**(2), 95–118.

Blom, G., & Holmquist, B. 1998. *Statistikteori med tillämpningar*. 3. [revised and extended] edn. Lund: Studentlitteratur.

Boos, D. D., & Hughes-Oliver, J. M. 2000. How large does n have to be for Z and t intervals? *The American Statistician*, **54**(2), 121–128.

Buchholz, A., Hollander, N., & Sauerbrei, W. 2008. On properties of predictors derived with a two-step bootstrap model averaging approach – a simulation study in the linear regression model. *Computational Statistics & Data Analysis*, **52**(5), 2778–2793.

Buckland, S. T., Burnham, K. P., & Augustin, N. H. 1997. Model selection: An integral part of inference. *Biometrics*, **53**(2), 603–618.

Burnham, K. P., & Anderson, D. R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. 2. edn. New York: Springer.

Burnham, K. P., & Anderson, D. R. 2004. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, **33**(2), 261–304.

Casella, G., & Berger, R. L. 2002. *Statistical inference*. 2. edn. Pacific Grove, CA: Duxbury.

Cochran, W. G. 1954. Some methods for strengthening the common χ^2 tests. *Biometrics*, **10**(4), 417–451.

Cohen, A., & Sackrowitz, H. B. 1975. Unbiasedness of the chi-square, likelihood ratio, and other goodness of fit tests for the equal cell case. *The Annals of Statistics*, **3**(4), 959–964.

D'Agostino, R. B., & Stephens, M. A. (eds). 1986. *Goodness-of-fit techniques*. New York: Marcel Dekker.

Durbin, J. 1973. *Distribution theory for tests based on the sample distribution function*. Philadelphia: SIAM.

Efron, B., & Tibshirani, R. 1993. *An introduction to the bootstrap*. New York: Chapman & Hall.

Greenwood, P. E., & Nikulin, M. S. 1996. *A guide to chi-squared testing*. New York: Wiley.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. 1999. Bayesian model averaging: A tutorial. *Statistical Science*, **14**(4), 382–401.

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., & Kane, J. 2008. *plotrix: Various plotting functions*. R package version 2.4-3.

Mann, H. B., & Wald, A. 1942. On the choice of the number of class intervals in the application of the chi square test. *The Annals of Mathematical Statistics*, **13**(3), 306–317.

Marsaglia, G., Tsang, W. W., & Wang, J. 2003. Evaluating Kolmogorov's distribution. *Journal of Statistical Software*, **8**(18), 1–4.

Moore, D. S. 1986. Tests of chi-square type. *In:* D'Agostino, R. B., & Stephens, M. A. (eds), *Goodness-of-fit techniques*. New York: Marcel Dekker.

R Development Core Team. 2007. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Raftery, A. E., Madigan, D., & Hoeting, J. A. 1997. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, **92**(437), 179–191.

Rayner, J. C. W., & Best, D. J. 1989. *Smooth tests of goodness of fit*. New York: Oxford University Press.

Reschenhofer, E. 1996. Prediction with vague prior knowledge. *Communications in Statistics – Theory and Methods*, **25**(3), 601–608.

Roscoe, J. T., & Byars, J. A. 1971. An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic. *Journal of the American Statistical Association*, **66**(336), 755–759.

Ross, S. M. 1997. *Simulation*. 2. edn. San Diego: Academic Press.

Ross, S. M. 2006. *Simulation*. 4. edn. Amsterdam: Elsevier / Academic Press.

Schorr, B. 1974. On the choice of the class intervals in the application of the chi-square test. *Math. Operations Forsch. u. Statist.*, **5**, 357–377.

Snedecor, G. W., & Cochran, W. G. 1980. *Statistical methods*. 7. edn. Ames, IA: Iowa State U.P.

Stephens, M. A. 1970. Use of the Kolmogorov–Smirnov, Cramer–von Mises and related statistics without extensive tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **32**(1), 115–122.

Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, **69**(347), 730–737.

Venables, W. N., & Ripley, B. D. 2002. *Modern applied statistics with S*. 4. edn. New York: Springer.

Wichura, M. J. 1988. Algorithm AS 241: The percentage points of the normal distribution. *Applied Statistics*, **37**(3), 477–484.

Williams, C. A., Jr. 1950. On the choice of the number and width of classes for the chi-square test of goodness of fit. *Journal of the American Statistical Association*, **45**(249), 77–86.

APPENDIX A

Tables: Simulation study I

Table A.1: *Frequencies of highest ranks for the Kolmogorov goodness-of-fit test for samples of size 100.*

	Assumed model		
True model	Weibull	Gamma	Lognormal
Weibull	817	169	14
Gamma	164	556	280
Lognormal	29	274	697

Table A.2: *Frequencies of highest ranks for the chi-squared goodness-of-fit test (13 equiprobable classes, 10 degrees of freedom) for samples of size 100.*

	Assumed model		
True model	Weibull	Gamma	Lognormal
Weibull	708	249	43
Gamma	202	441	357
Lognormal	53	384	563

Table A.3: *Frequencies of highest ranks for Akaike's information criterion for samples of size 100.*

	Assumed model		
True model	Weibull	Gamma	Lognormal
Weibull	890	107	3
Gamma	111	645	244
Lognormal	0	271	729

Table A.4: *Frequencies of highest ranks for the Kolmogorov goodness-of-fit test for samples of size 20.*

	Assumed model		
True model	Weibull	Gamma	Lognormal
Weibull	611	232	157
Gamma	341	240	419
Lognormal	201	264	535

Table A.5: *Frequencies of highest ranks for the chi-squared goodness-of-fit test (13 equiprobable classes, 10 degrees of freedom) for samples of size 20.*

	Assumed model		
True model	Weibull	Gamma	Lognormal
Weibull	455	275	270
Gamma	368	289	343
Lognormal	290	320	390

Table A.6: *Frequencies of highest ranks for Akaike's information criterion for samples of size 20.*

	Assumed model		
True model	Weibull	Gamma	Lognormal
Weibull	673	203	124
Gamma	321	252	427
Lognormal	172	234	594

APPENDIX B

Figures: Simulation study I

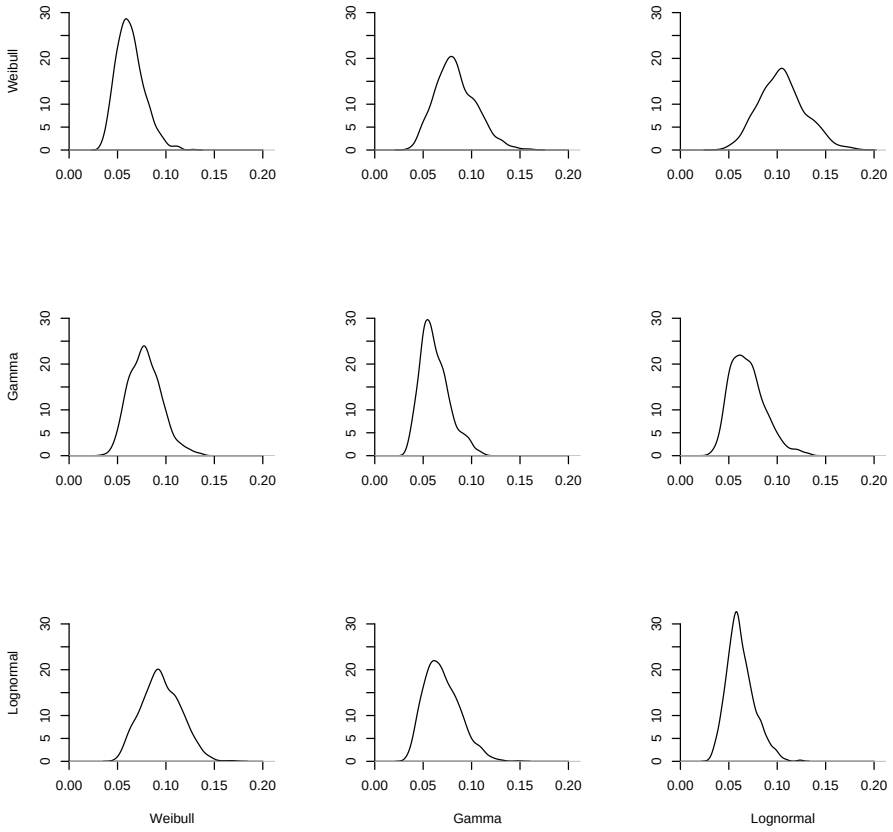

Figure B.1: *Density estimates for the test statistics of the Kolmogorov goodness-of-fit tests. True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 100.*

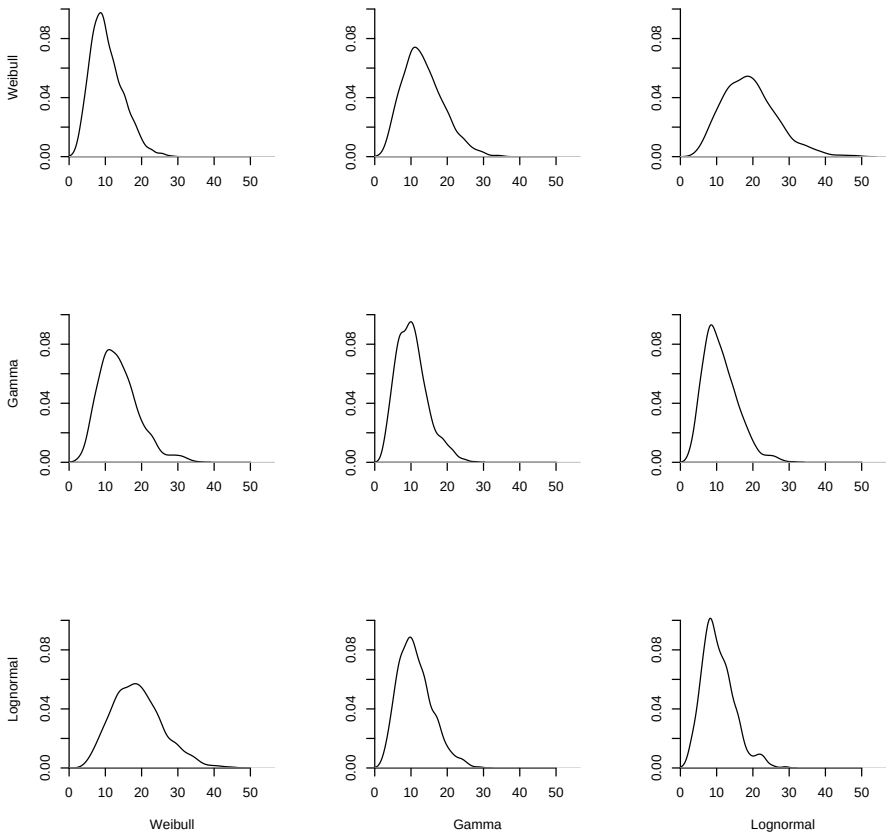

Figure B.2: *Density estimates for the test statistics of the chi-squared goodness-of-fit tests (13 equiprobable classes, 10 degrees of freedom). True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 100.*

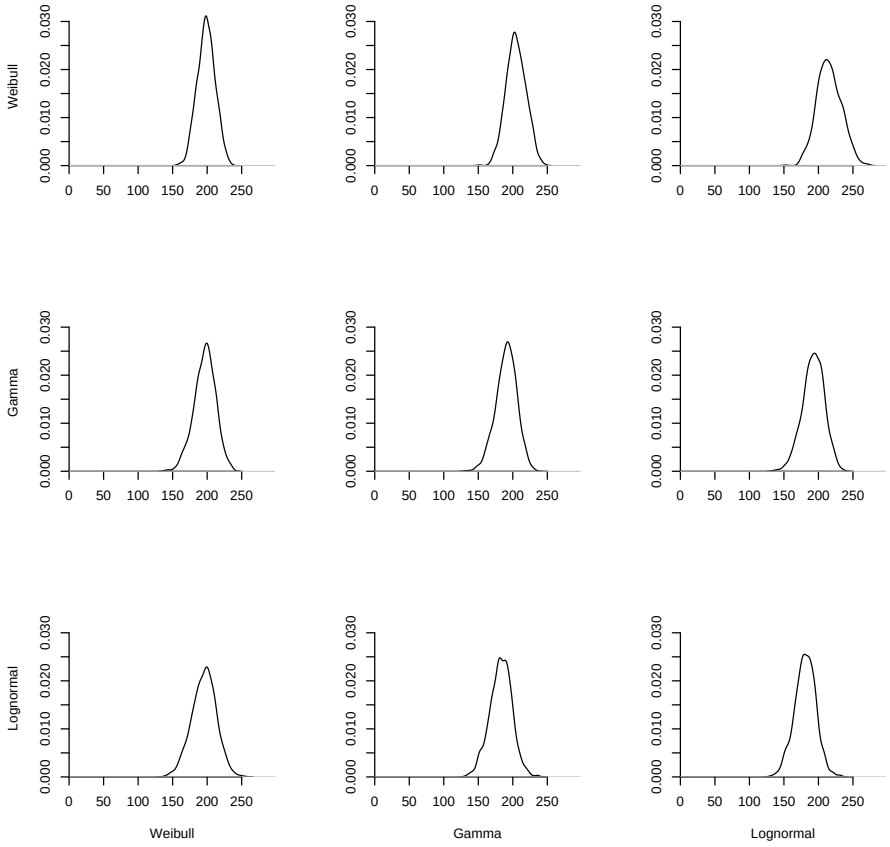

Figure B.3: *Density estimates for the values of the Akaike information criterion. True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 100.*

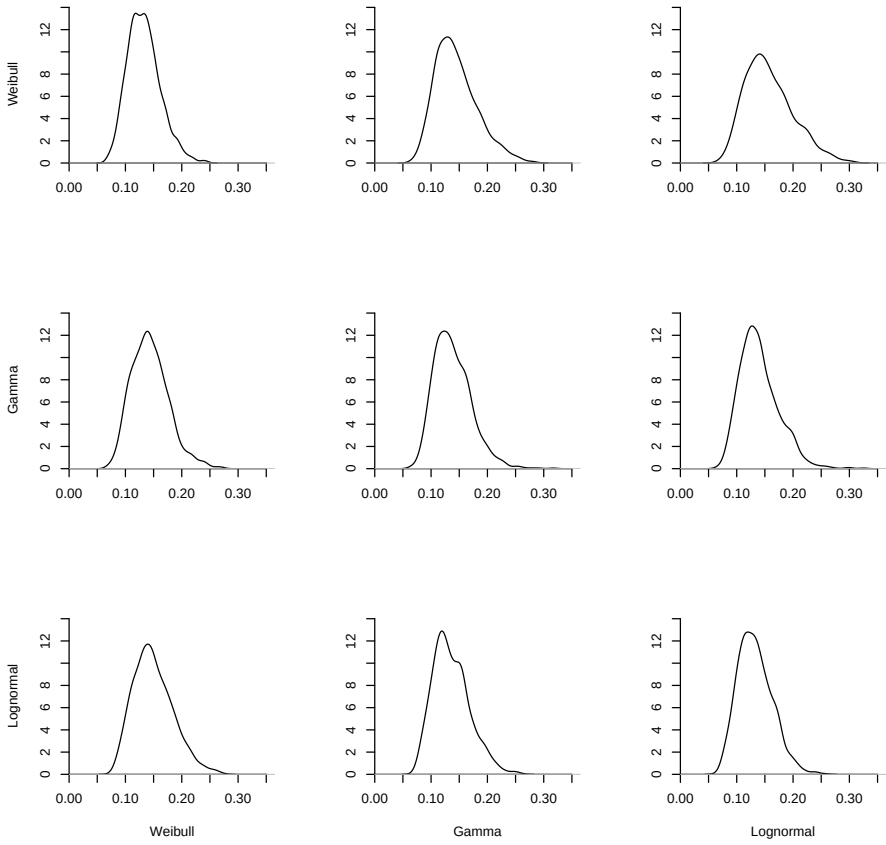

Figure B.4: *Density estimates for the test statistics of the Kolmogorov goodness-of-fit tests. True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 20.*

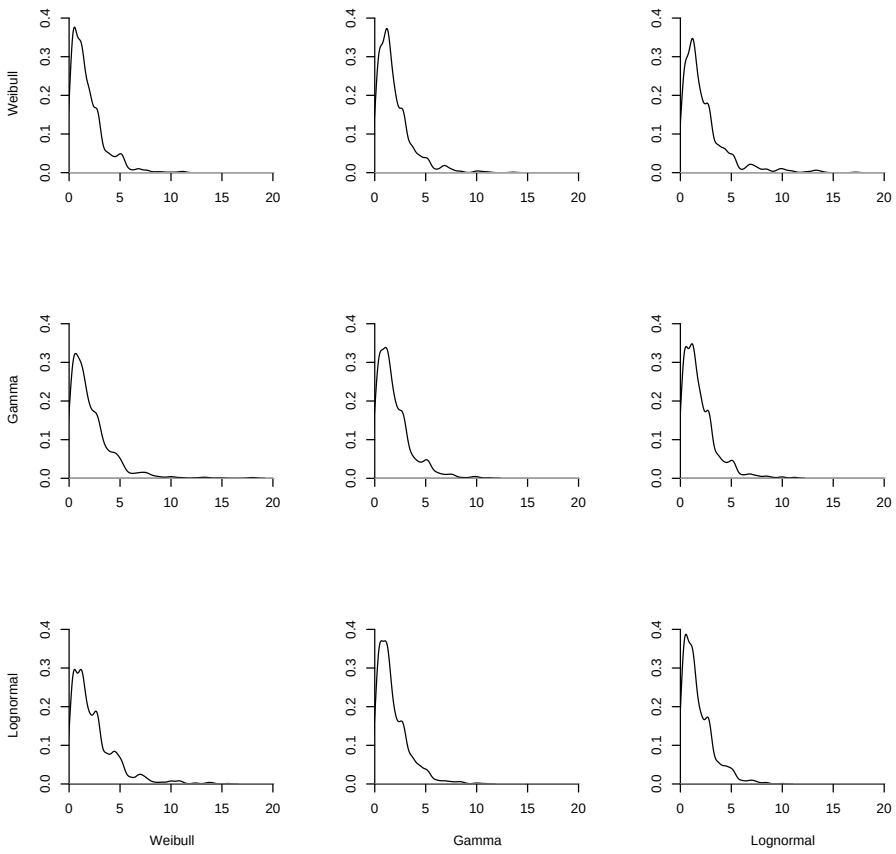

Figure B.5: *Density estimates for the test statistics of the chi-squared goodness-of-fit tests (4 equiprobable classes, 1 degree of freedom). True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 20.*

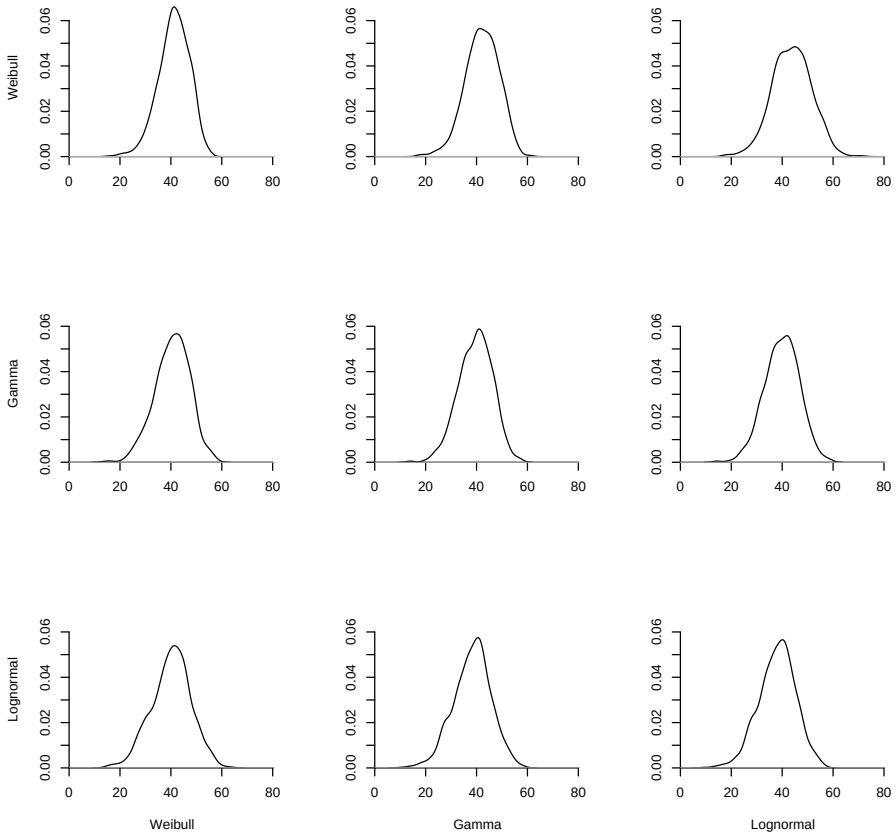

Figure B.6: *Density estimates for the values of the Akaike information criterion. True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 20.*

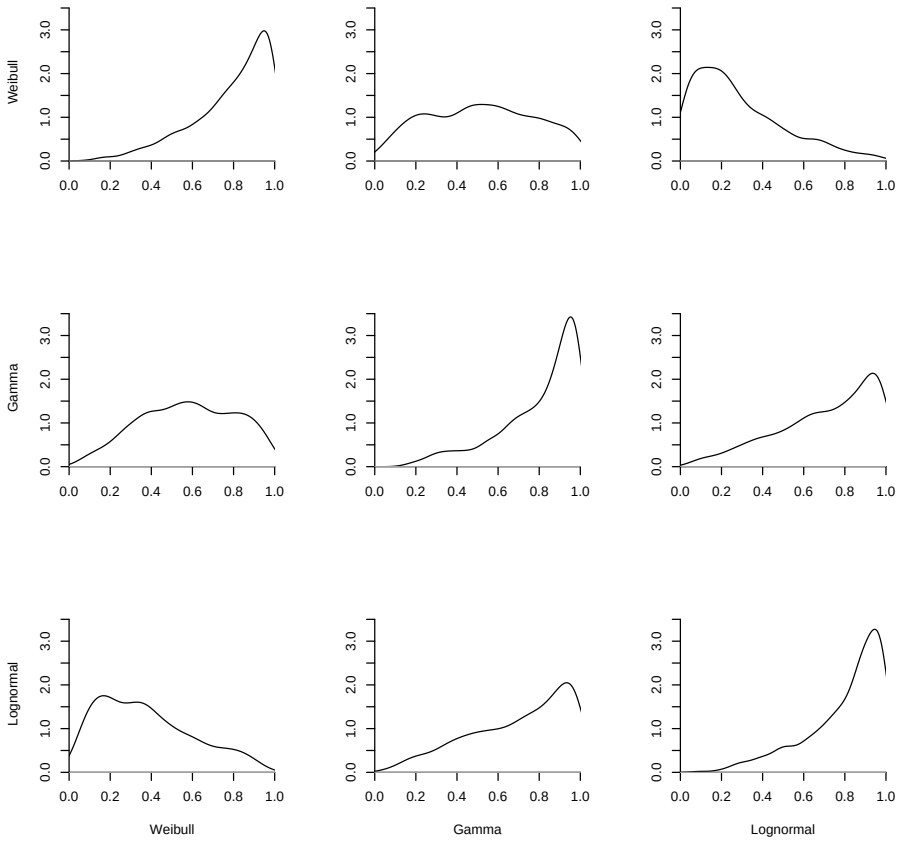

Figure B.7: *Density estimates for the p-values of the Kolmogorov goodness-of-fit tests. True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 100.*

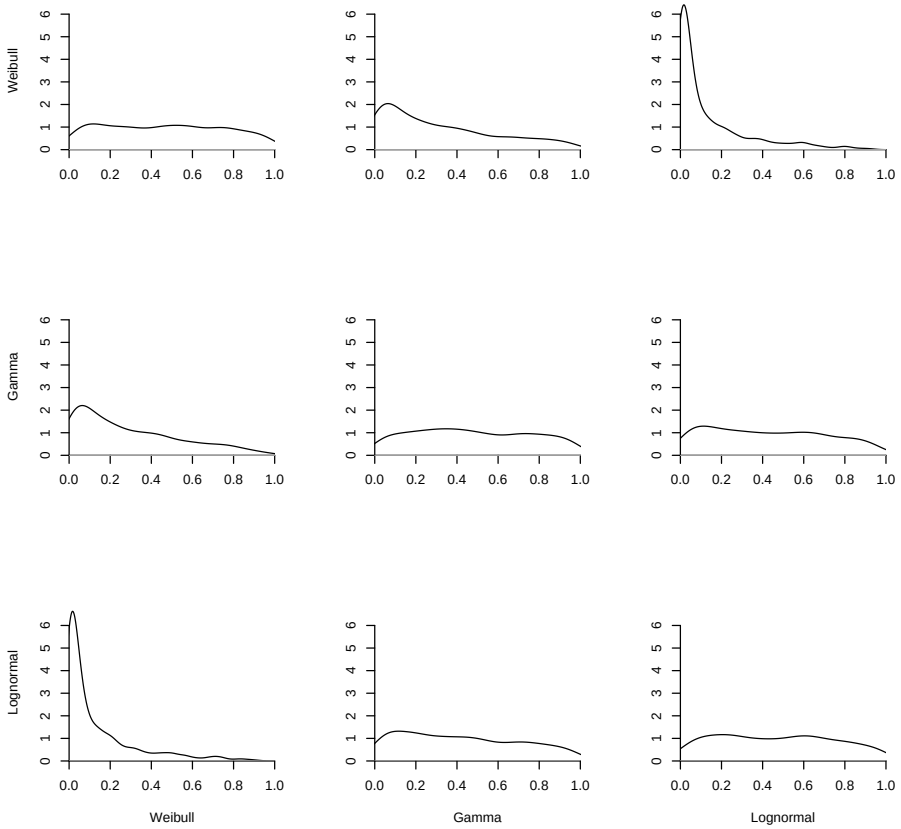

Figure B.8: *Density estimates for the p-values of the chi-squared goodness-of-fit tests (13 equiprobable classes, 10 degrees of freedom). True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 100.*

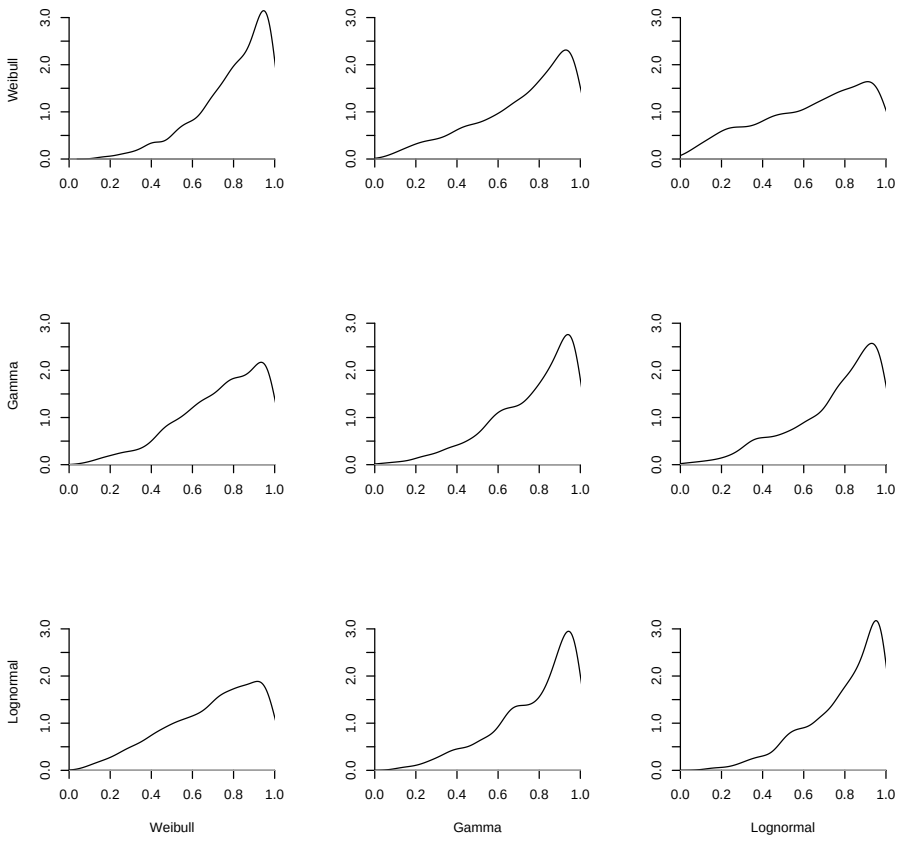

Figure B.9: *Density estimates for the p-values of the Kolmogorov goodness-of-fit tests. True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 20.*

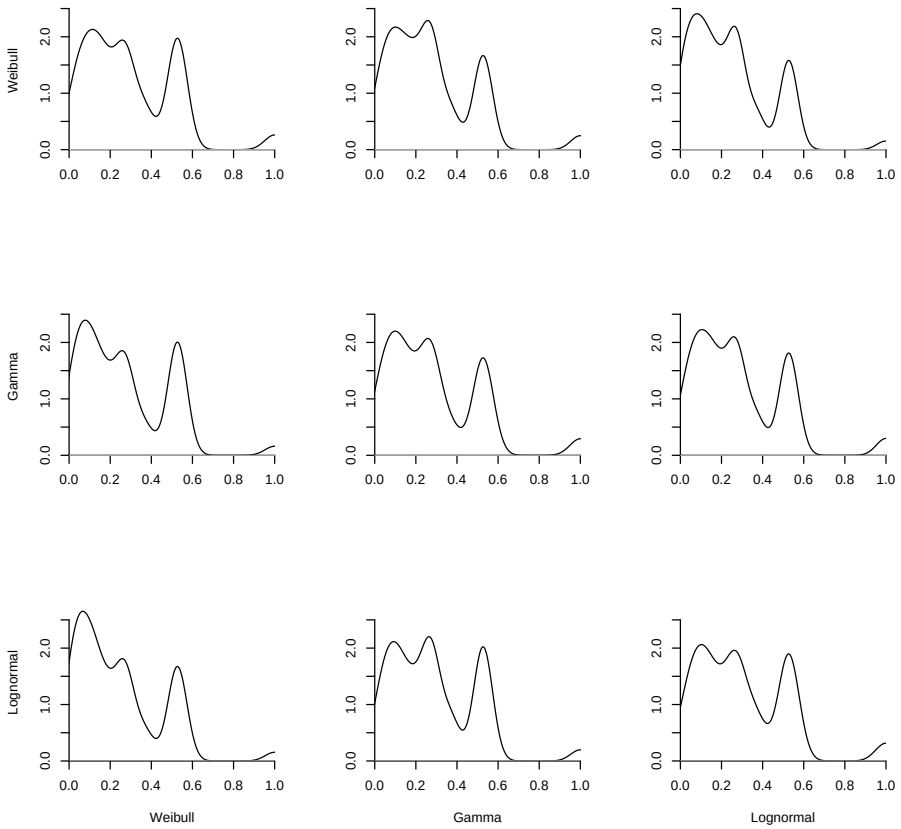

Figure B.10: *Density estimates for the p-values of the chi-squared goodness-of-fit tests (4 equiprobable classes, 1 degree of freedom). True distributions row-wise (in order from the top: Weibull, Gamma and Lognormal); assumed distributions column-wise (from left to right, in the same order). Sample size: 20.*

APPENDIX C

Tables: Simulation study II

Table C.1: *Results for large samples ($n = 100$) from the Weibull(3,2) distribution when model selection was applied. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.9446.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
Buckland	BC_a	0.7995	1.1015	0.3019	0.027	0.036	93.7
Buckland	$PCTL_1$	0.8321	1.1319	0.2998	0.007	0.074	91.9
Buckland	$PCTL_2$	0.8004	1.1002	0.2998	0.029	0.047	92.4
Buckland	Weighted	0.8316	1.1343	0.3027	0.007	0.069	92.4
RMS_1	BC_a	0.7977	1.0999	0.3022	0.026	0.034	94.0
RMS_1	$PCTL_1$	0.8329	1.1329	0.3000	0.007	0.069	92.4
RMS_1	$PCTL_2$	0.7994	1.0994	0.3000	0.029	0.047	92.4
RMS_1	Weighted	0.8323	1.1350	0.3027	0.007	0.068	92.5
RMS_2	BC_a	0.7976	1.0912	0.2936	0.032	0.028	94.0
RMS_2	$PCTL_1$	0.8347	1.1228	0.2881	0.011	0.076	91.3
RMS_2	$PCTL_2$	0.8095	1.0976	0.2881	0.027	0.047	92.6
RMS_2	Weighted	0.8343	1.1249	0.2907	0.009	0.076	91.5

Table C.2: *Results for large samples ($n = 100$) from the gamma(4,7) distribution when model selection was applied. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.2493.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
Buckland	BC_a	0.2031	0.2891	0.0860	0.039	0.029	93.2
Buckland	$PCTL_1$	0.2146	0.3007	0.0861	0.006	0.069	92.5
Buckland	$PCTL_2$	0.2018	0.2879	0.0861	0.045	0.026	92.9
Buckland	Weighted	0.2144	0.3014	0.0869	0.005	0.070	92.5
RMS_1	BC_a	0.2010	0.2907	0.0897	0.036	0.024	94.0
RMS_1	$PCTL_1$	0.2116	0.3004	0.0888	0.005	0.045	95.0
RMS_1	$PCTL_2$	0.2021	0.2909	0.0888	0.040	0.026	93.4
RMS_1	Weighted	0.2114	0.3010	0.0896	0.006	0.046	94.8
RMS_2	BC_a	0.1971	0.2894	0.0923	0.035	0.017	94.8
RMS_2	$PCTL_1$	0.2094	0.2987	0.0893	0.008	0.038	95.4
RMS_2	$PCTL_2$	0.2038	0.2931	0.0893	0.033	0.026	94.1
RMS_2	Weighted	0.2092	0.2993	0.0900	0.008	0.036	95.6

Table C.3: *Results for large samples ($n = 100$) from the lognormal(0.2,0.3) distribution when model selection was applied. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.8315.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
Buckland	BC_a	0.7501	0.8838	0.1338	0.046	0.021	93.3
Buckland	$PCTL_1$	0.7672	0.8997	0.1325	0.017	0.046	93.7
Buckland	$PCTL_2$	0.7521	0.8846	0.1325	0.045	0.023	93.2
Buckland	Weighted	0.7670	0.9009	0.1340	0.014	0.043	94.3
RMS_1	BC_a	0.7478	0.8879	0.1401	0.035	0.020	94.5
RMS_1	$PCTL_1$	0.7605	0.8982	0.1376	0.020	0.032	94.8
RMS_1	$PCTL_2$	0.7536	0.8913	0.1376	0.029	0.023	94.8
RMS_1	Weighted	0.7602	0.8991	0.1388	0.018	0.031	95.1
RMS_2	BC_a	0.7385	0.8896	0.1511	0.036	0.008	95.6
RMS_2	$PCTL_1$	0.7542	0.9007	0.1465	0.012	0.021	96.7
RMS_2	$PCTL_2$	0.7511	0.8976	0.1465	0.020	0.019	96.1
RMS_2	Weighted	0.7539	0.9017	0.1477	0.012	0.021	96.7

Table C.4: *Results for large samples ($n = 100$) from the gamma(4,7) distribution when parametric bootstrap with knowledge of the true distribution was done. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.2493.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
noMS	BC_a	0.2031	0.2900	0.0869	0.045	0.028	92.7
noMS	$PCTL_1$	0.2132	0.3005	0.0873	0.007	0.054	93.9
noMS	$PCTL_2$	0.2020	0.2893	0.0873	0.048	0.023	92.9
noMS	Weighted	0.2210	0.3340	0.1130	0.000	0.103	89.7

Table C.5: *Results for large samples ($n = 20$) from the Weibull(3,2) distribution when model selection was applied. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.9446.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
Buckland	BC_a	0.6921	1.2813	0.5891	0.020	0.101	87.9
Buckland	$PCTL_1$	0.7761	1.3847	0.6086	0.002	0.183	81.5
Buckland	$PCTL_2$	0.6371	1.2457	0.6086	0.038	0.064	89.8
Buckland	Weighted	0.7755	1.3904	0.6149	0.002	0.182	81.6
RMS_1	BC_a	0.6741	1.2881	0.6140	0.019	0.068	91.3
RMS_1	$PCTL_1$	0.7618	1.3813	0.6195	0.003	0.158	83.9
RMS_1	$PCTL_2$	0.6405	1.2600	0.6195	0.033	0.066	90.1
RMS_1	Weighted	0.7607	1.3862	0.6255	0.001	0.157	84.2
RMS_2	BC_a	0.6467	1.2500	0.6034	0.027	0.034	93.9
RMS_2	$PCTL_1$	0.7625	1.3445	0.5820	0.004	0.140	85.6
RMS_2	$PCTL_2$	0.6774	1.2594	0.5820	0.033	0.073	89.4
RMS_2	Weighted	0.7613	1.3486	0.5872	0.004	0.137	85.9

Table C.6: *Results for large samples ($n = 20$) from the gamma(4,7) distribution when model selection was applied. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.2493.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
Buckland	BC_a	0.1766	0.3416	0.1650	0.034	0.087	87.9
Buckland	$PCTL_1$	0.2014	0.3824	0.1809	0.003	0.163	83.4
Buckland	$PCTL_2$	0.1442	0.3251	0.1809	0.078	0.044	87.8
Buckland	Weighted	0.2013	0.3837	0.1825	0.003	0.162	83.5
RMS_1	BC_a	0.1653	0.3478	0.1825	0.019	0.053	92.8
RMS_1	$PCTL_1$	0.1921	0.3802	0.1881	0.004	0.113	88.3
RMS_1	$PCTL_2$	0.1464	0.3344	0.1881	0.050	0.045	90.5
RMS_1	Weighted	0.1918	0.3818	0.1900	0.004	0.112	88.4
RMS_2	BC_a	0.1434	0.3446	0.2013	0.019	0.010	97.1
RMS_2	$PCTL_1$	0.1808	0.3783	0.1974	0.002	0.078	92.0
RMS_2	$PCTL_2$	0.1483	0.3457	0.1974	0.032	0.036	93.2
RMS_2	Weighted	0.1805	0.3798	0.1993	0.002	0.078	92.0

Table C.7: *Results for large samples ($n = 20$) from the lognormal(0.2,0.3) distribution when model selection was applied. The interval bounds and lengths are average values. The true value of the 10th percentile is 0.8315.*

		Bounds			Miss prob.		
Method	Interval	Lower	Upper	Length	Left	Right	Coverage (%)
Buckland	BC_a	0.6786	0.9489	0.2703	0.079	0.050	87.1
Buckland	$PCTL_1$	0.7289	1.0139	0.2850	0.009	0.107	88.4
Buckland	$PCTL_2$	0.6490	0.9340	0.2850	0.110	0.029	86.1
Buckland	Weighted	0.7283	1.0162	0.2879	0.010	0.109	88.1
RMS_1	BC_a	0.6441	0.9605	0.3163	0.049	0.020	93.1
RMS_1	$PCTL_1$	0.7063	1.0098	0.3036	0.011	0.068	92.1
RMS_1	$PCTL_2$	0.6530	0.9566	0.3036	0.058	0.030	91.2
RMS_1	Weighted	0.7056	1.0122	0.3066	0.010	0.068	92.2
RMS_2	BC_a	0.5844	0.9668	0.3824	0.026	0.011	96.3
RMS_2	$PCTL_1$	0.6734	1.0179	0.3444	0.002	0.035	96.3
RMS_2	$PCTL_2$	0.6450	0.9894	0.3444	0.011	0.025	96.4
RMS_2	Weighted	0.6727	1.0202	0.3474	0.002	0.035	96.3

APPENDIX D

Figures: Simulation study II

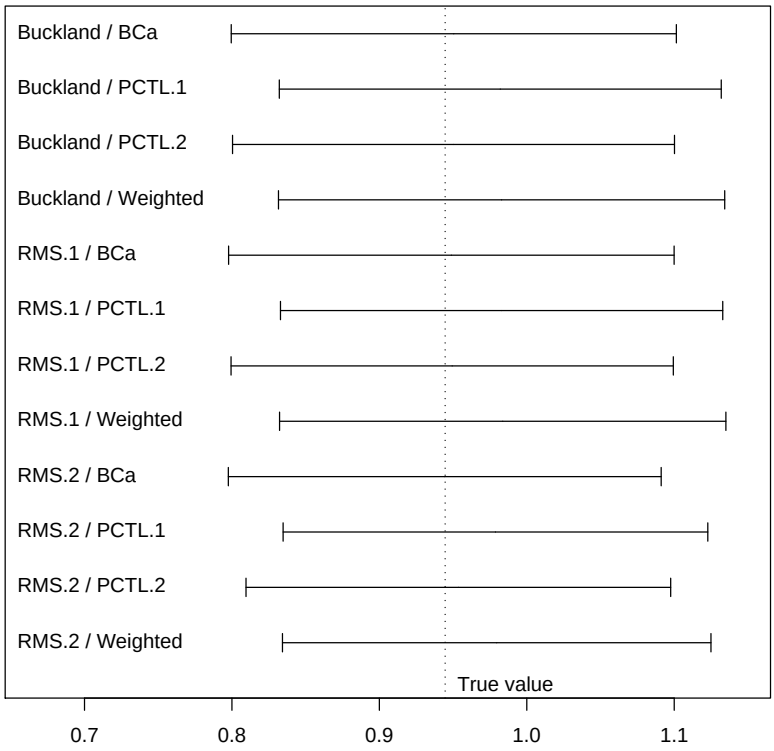

Figure D.1: *A plot of the average confidence intervals when model selection was applied to samples of size 100 from the Weibull(3,2) distribution. The dotted vertical line indicates the true value of the 10th percentile: 0.9446.*

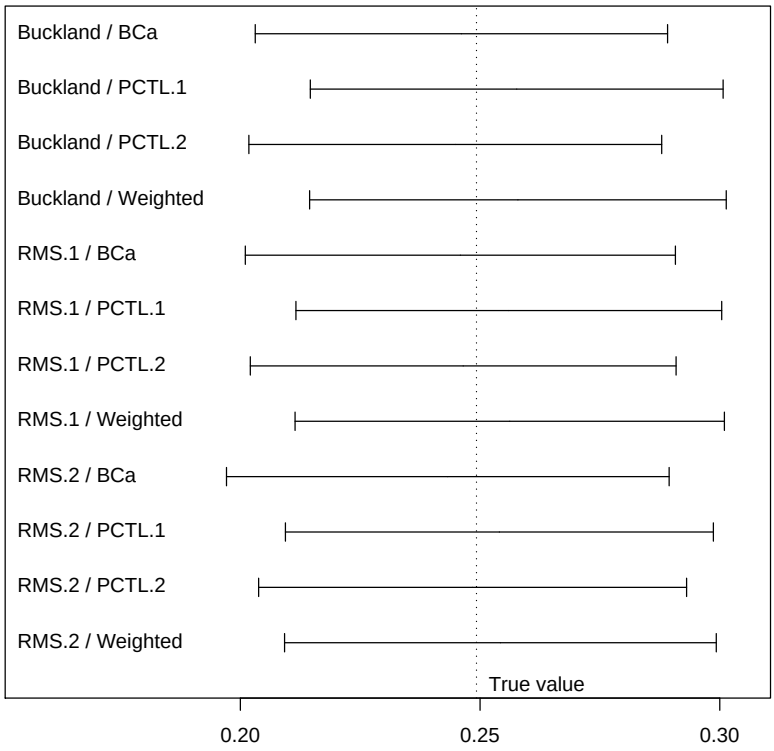

Figure D.2: *A plot of the average confidence intervals when model selection was applied to samples of size 100 from the gamma(4,7) distribution. The dotted vertical line indicates the true value of the 10th percentile: 0.2493.*

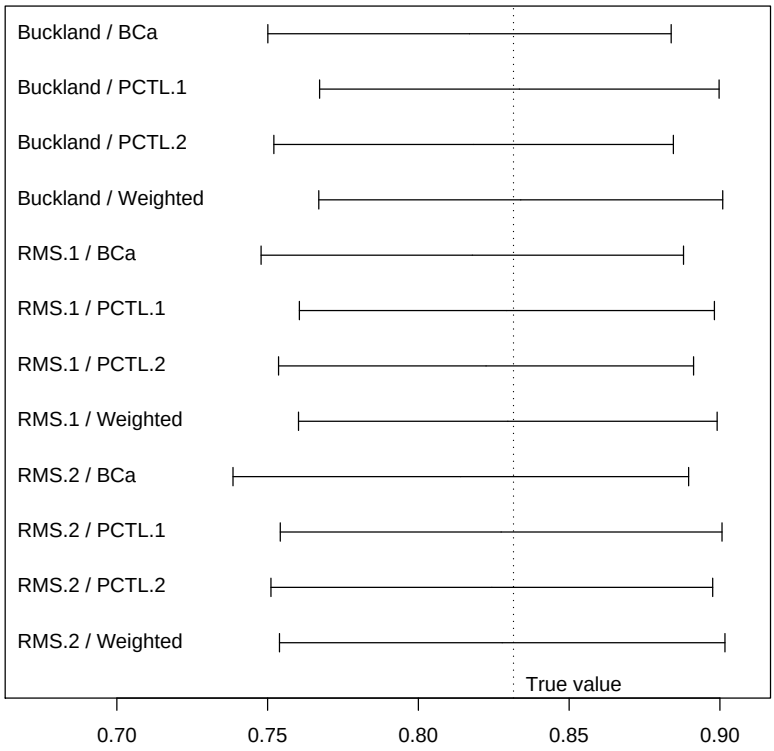

Figure D.3: *A plot of the average confidence intervals when model selection was applied to samples of size 100 from the lognormal(0.2,0.3) distribution. The dotted vertical line indicates the true value of the 10th percentile: 0.8315.*

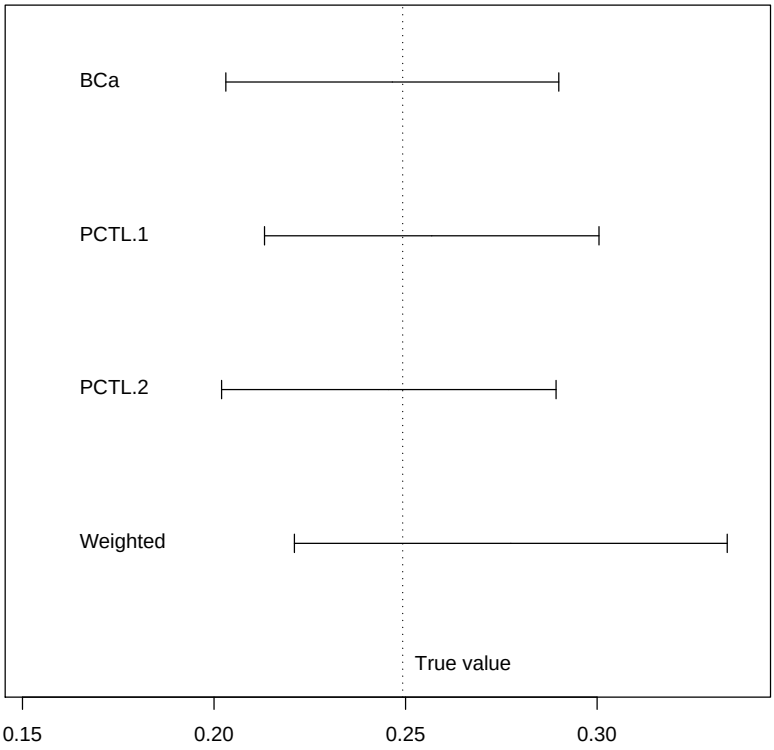

Figure D.4: *A plot of the average confidence intervals for samples of size 100 from the gamma(4,7) distribution on which parametric bootstrap with knowledge of the true distribution was carried out. The dotted vertical line indicates the true value of the 10th percentile: 0.2493.*

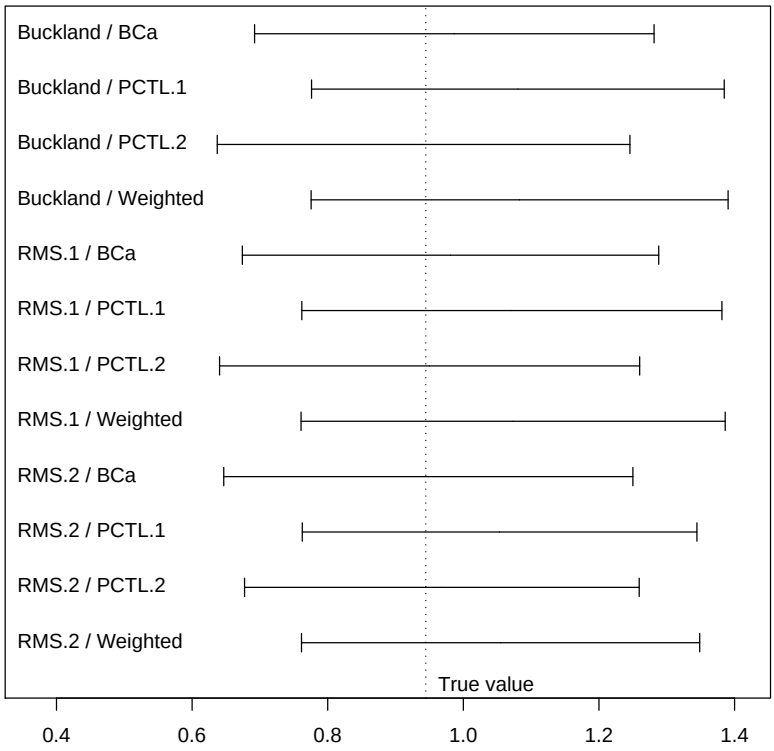

Figure D.5: *A plot of the average confidence intervals when model selection was applied to samples of size 20 from the Weibull(3,2) distribution. The dotted vertical line indicates the true value of the 10th percentile: 0.9446.*

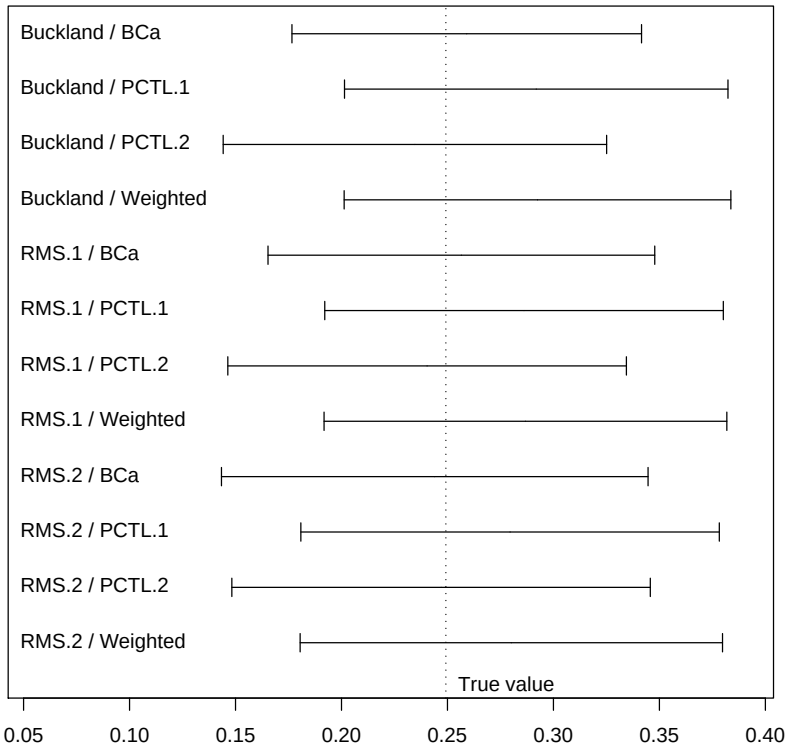

Figure D.6: *A plot of the average confidence intervals when model selection was applied to samples of size 20 from the gamma(4,7) distribution. The dotted vertical line indicates the true value of the 10th percentile: 0.2493.*

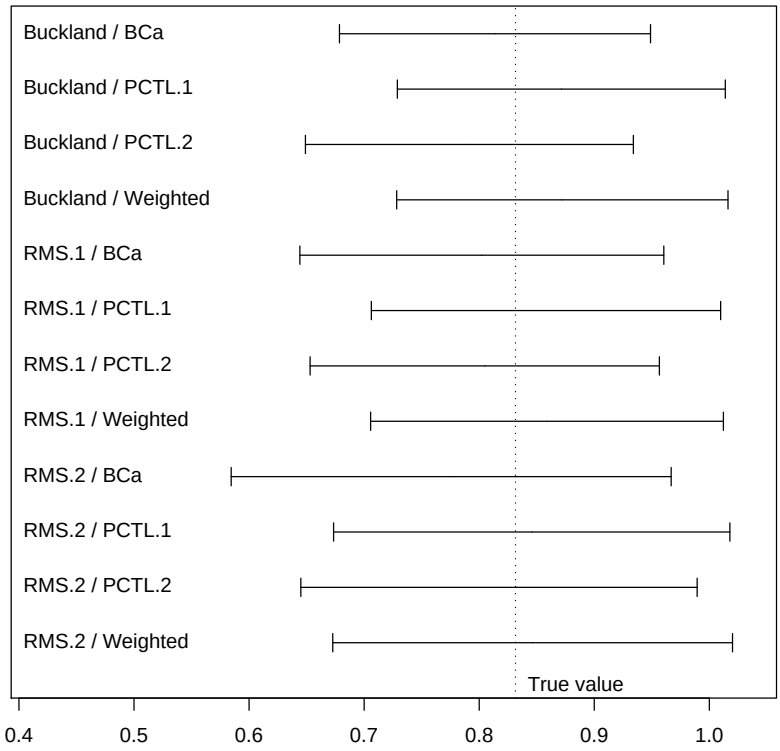

Figure D.7: *A plot of the average confidence intervals when model selection was applied to samples of size 20 from the lognormal(0.2,0.3) distribution. The dotted vertical line indicates the true value of the 10th percentile: 0.8315.*