4-28-2005

A Comparison of Parametric and Coarsened Bayesian Interval Estimation in the Presence of a Known Mean-Variance Relationship

Kent Koprowicz
University of Washington, kentk@u.washington.edu

Scott S. Emerson
University of Washington, semerson@u.washington.edu

Peter Hoff
University of Washington, hoff@stat.washington.edu

1. INTRODUCTION

Many scientific questions are addressed statistically by making inference about some population parameter. In a clinical trial, for example, treatment effect is most often measured by comparing some summary measure, θ, of the distribution of responses, $\vec{y}$, across populations. One might consider a difference or ratio of means, a difference or ratio of medians, an odds ratio, a hazard ratio or a number of other possibilities. The standard procedure is to use data from the samples in order to make inference about the true value of the treatment effect θ in the populations.

When making such inference, there are two major schools of statistical thought: frequentist and Bayesian. Both schools employ parametric and nonparametric methods but the operating characteristics considered and interpretations differ. A key difference between the frequentist and Bayesian approaches concerns the quantities that are conditioned upon in statistical models: the frequentist approach considers the distribution of the data given a parameter, $p(\vec{y}|\theta)$, whereas the Bayesian approach considers $p(\theta|\vec{y})$. Conceptually, both of these models can be derived from the joint distribution $p(\theta, \vec{y})$ but standard practice shows the frequentist and Bayesian approaches diverging in the probability model. In particular, there is a greater tendency in the frequentist approach to interpret results nonparametrically and it is this difference we most want to address.

We regard the frequentist and Bayesian approaches as complementary, each providing assistance in answering important scientific questions. We also regard that nonparametric methods, for reasons of robustness, are to be preferred to parametric models. However, nonparametric approaches should be broad enough to include any reasonable parametric model that might have instead been chosen to represent the data. Thus our goal in this paper is to explore a nonparametric Bayesian method of analysis that exists in the same probability space as the most common nonparametric frequentist approach. We view this as a first step in working toward a more general method for evaluating both Bayesian and frequentist operating characteristics of statistical procedures for clinical trial design.

In the next section we provide a more detailed account of the background and motivation for the current research. In § 3 we outline a proposed gen-

eral approach to the problem. In § 4 we present the results of simulation studies for inference about a population mean and median. In § 5 we apply the method to a hospital charge dataset, obtaining confidence intervals for measures of location for a possibly Lognormal population and compare our results to those obtained by previous researchers. We follow with a discussion in § 6.

2. BACKGROUND AND MOTIVATION

2.1 *Parametric and nonparametric models*

We desire to avoid, as far as possible, parametric assumptions. Aside from being *a priori* unverifiable, assuming a specific parametric model often seems unreasonable given otherwise limited information. Often such models require more detailed assumptions than the experiment we are analyzing was designed to address. In a two-sample clinical trial, for example, the scientific question to be addressed is usually one of central tendency: Does the treatment tend to result in higher (lower) values for the outcome of interest? If we choose to make inference about the mean outcome, the use of a parametric model is equivalent to admitting ignorance about the effect of treatment on the first moment of the outcome distribution but then assuming we know the effect of treatment on all of the higher moments of the distribution including the variance, skewness, kurtosis and so on. Nonparametric methods allow us to relax these overly detailed assumptions. However, to provide the greatest utility, our nonparametric model should be be broad enough to include the particular parametric model that others would have chosen.

Attempting to circumvent the issues raised by use of parametric models by relying on model checking is problematic. Reviewers of earlier drafts of this manuscript, in fact, suggested that much of our statistical and scientific concerns could be addressed by simple model checking. We feel, however, that a reliance on model checking is misguided in a number of respects. First, it may be that no parametric modeling is required; e.g. when the use of a simple statistical functional meets the goals of the analysis. Second, the application may be one such that the statistical model must be specified in advance; this is often the case when dealing with governmental regulatory agencies, as in seeking approval for a novel treatment. In the Bayesian paradigm we have a philosophical difficulty with model checking in that it seems reasonable to expect that any uncertainty in the model should have already been incorpo-

rated in the prior distribution. A final difficulty with model checking we'd like to point out is that typical methods of testing for lack of model fit are lacking in power at smaller sample sizes. Later in this paper, for example, we will consider data simulated from a class of mixtures of Exponential and Lognormal distributions: Figure 1 shows some typical distributions from which such data might be sampled. One can see that in practice it may be difficult to distinguish which distribution gave rise to a particular dataset and, in the case of a mixture, especially difficult to determine the exact composition. In the situation we are considering in this paper, where the model is posited in advance yet there is concern over possible unanticipated departures, the use of an omnibus test such as the Kolmogorov-Smirnov would seem most appropriate. Simulations show, however, that the Kolmogorov-Smirnov test has poor power discriminating between the Exponential and Lognormal in smaller sample sizes. Even for a sample size of 40 the Kolmogorov-Smirnov statistic showed only about 25% power to reject the $\alpha = 0.05$ level hypothesis test that the data were from a Lognormal distribution when they were in fact from the Exponential and about 12% power to reject the hypothesis that the data were Exponential when they were Lognormal.

2.2 *How parametric models influence analysis*

In order to have nonparametric models which encompass common parametric models, it is useful to consider the ways in which parametric models may drive a statistical analysis and to examine whether that role is central to answering the scientific question. Parametric assumptions influence:

1. *The choice of parameter to be compared.* For a location problem, for instance, should we consider the median or the mean? If we assume our data arises from a Normal distribution, for example, then we usually focus on the mean. If, on the other hand, we assume the data come from a Double-Exponential distribution our focus might be the median. That is, a fully parametric approach often chooses parameters corresponding to sufficient statistics. A nonparametric statistician might believe that a scientifically more appropriate criterion would be based on some meaningful loss function. For instance, we may desire to downweight the influence of outliers and so might choose to make inference about the geometric mean. Such an approach is common in health services cost data where interest often lies in using a mean to

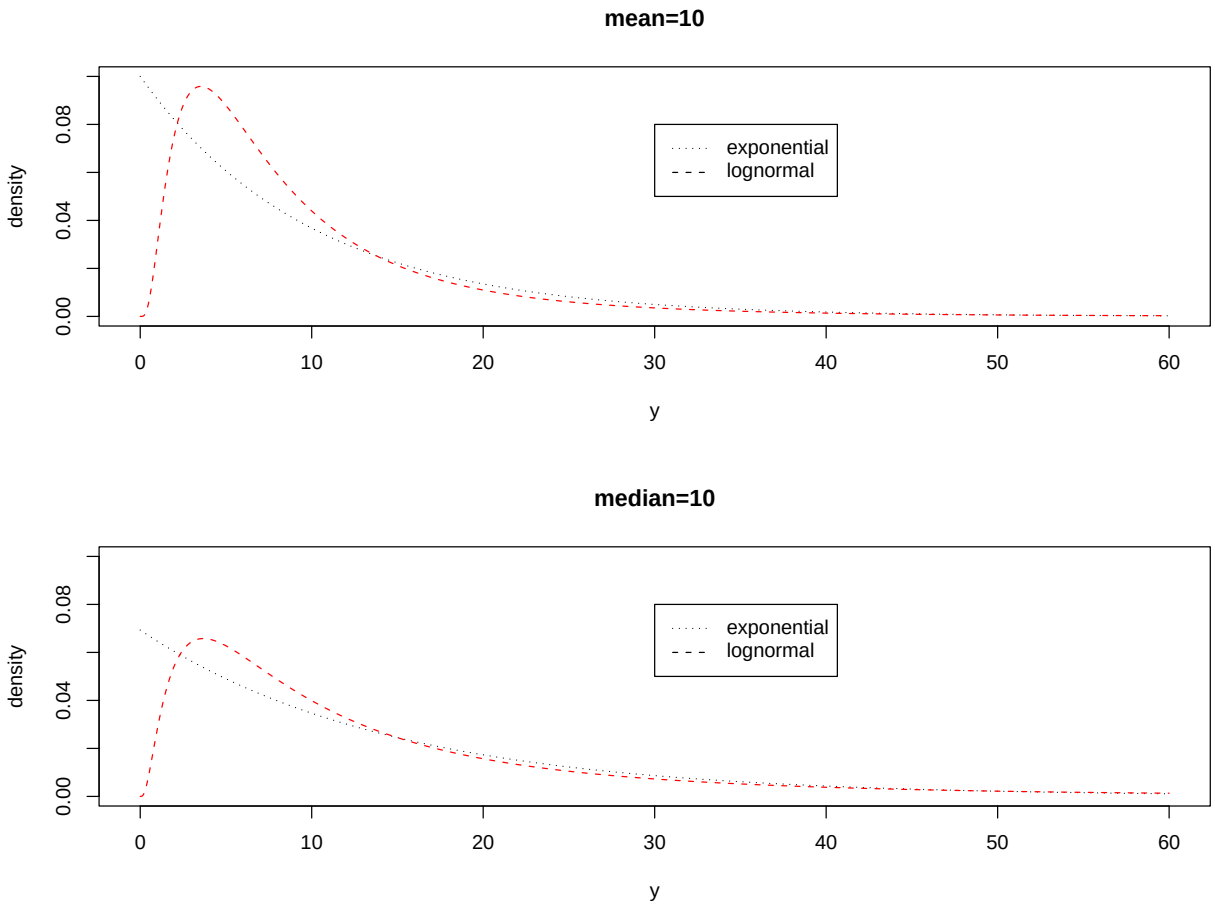

Figure 1. Top: Typical Exponential and Lognormal distributions for mean inference.
Bottom: Typical Exponential and Lognormal distributions for median inference

estimate total costs in a population, regardless of the "natural parameter" from some presumed probability model.

2. *The way the parameter is computed.* For example, if we wanted to use the mean for scientific reasons, in estimating the population mean under an Exponential model we might consider the sample arithmetic mean while for a Lognormal model the formula would be a function of the sample geometric mean. It would be appropriate, however, to use the nonparametric estimator in either case. To the extent that science dictates the mean but *not* the parametric model, the robustness of a nonparametric estimator may be preferred.

3. *The hypothesized mean-variance relationship of our selected statistic.* This is a key idea which we will discuss in more detail later in this section. For an example, in the case of both the Exponential and Lognormal models the variance of the sample mean is proportional to the square of the population mean. For the Poisson model, on the other hand, the variance is assumed equal to the mean. In the standard nonparametric use of the t-test we generally construct confidence intervals and the like assuming a constant variance across alternatives. The nonparametric approach to the mean-variance relationship that we consider in this paper encompasses specific classes of parametric models. Hence we might, for example, look at nonparametric models with mean-variance relationships that include all Exponential and certain Lognormal models as a subclass.

4. *The shape of the distribution of the statistic under different parameter values.* For many test statistics the distribution of the nonparametric estimator converges to a Normal distribution. Thus in large samples, a parametric model has little impact on this aspect of parametric analyses.

5. *The hypothesized shape of the distribution.* This level of detail is necessary only if we wish to make single value predictions. Here we are

addressing inference about a population treatment effect.

The above argues that we might expect nonparametric models to capture much of the information contained in parametric models provided the mean-variance relationship is adequately addressed. In fact, much frequentist inference may be interpreted nonparametrically, even when the choice of statistic is derived under parametric models. For instance, Lumley et al. Lumley et al. (2002) show that although the t-test comparing means is derived under an assumption of Normally distributed data, it is valid to an excellent approximation even given the extreme departures from normality found, for instance, in medical cost data. Lumley et al. do not address the issue of the mean-variance relationship, however, and so their approach would not have corresponded well with parametric estimation based on, for example, a Lognormal model.

2.3 *The mean-variance relationship in frequentist statistics*

The major difficulty posed by a mean-variance relationship comes down to this: In general, when performing frequentist hypothesis testing, only the sampling distribution of the estimator $\hat{\theta}$ under the null hypothesis $\theta = \theta_0$ need be known. Often, such an estimator is asymptotically normally distributed, and so frequentist hypothesis testing amounts to estimating the expectation and the variance of the sampling distribution when $\theta = \theta_0$. For accurate confidence intervals, however, these first two moments of the statistic must be known under all parameter values. Meeting this further requirement is difficult and hence many methods of interval estimation, parametric as well as nonparametric, simply ignore the presence of a mean-variance relationship.

That the mean-variance relationship matters in nonparametric frequentist inference can be seen by considering the most common method of obtaining an interval estimate for a parameter: the "Wald" interval. That is

$$\hat{\theta} \ \pm \ Z_{1-\alpha/2} \times \frac{\hat{\sigma}}{n}$$

where $\hat{\sigma}$ is estimated under $\hat{\theta}$. Under an Exponential model, for example, one might take $\hat{\sigma} = \hat{\theta}$. Such intervals enjoy many nice properties with simplicity and ease of construction perhaps chief among them. It is recognized, however, that under various circumstances such intervals possess severe failings

as well. The Wald interval is based on asymptotic results and is valid only for sufficiently large samples. An exact confidence interval would be preferred, of course, but is not always available. Recall that $C_\alpha(\vec{y})$ is a $100(1-\alpha)\%$ confidence interval for θ if $Pr(\theta \in C_\alpha(\vec{y})|\theta) = 1-\alpha$. In general, confidence intervals are related to hypothesis tests of the form $H_0 : \theta = \theta_0$ versus a two-sided alternative $H_1 : \theta \neq \theta_0$. Specifically, $C_\alpha(\vec{y})$ should contain all of those values of θ that would *not* be rejected under H_0. Thus a Wald-based confidence interval corresponds to the inversion of a Wald-based hypothesis test. An important point about Wald intervals is that the variance does not change with θ.

Somewhat less commonly encountered are "Score" intervals (θ_L, θ_H) where

$$\theta_L = \hat{\theta} - Z_{1-\alpha/2} \times \sqrt{\frac{V(\theta_L)}{n}}$$
$$\theta_H = \hat{\theta} + Z_{1-\alpha/2} \times \sqrt{\frac{V(\theta_H)}{n}}.$$

Such score intervals are based on inversion of the Rao score test. Here, under an Exponential model, one might take $V(\theta_L) = \theta_L$. The score test, in turn, is based on the score function

$$\ell'_n(\theta_0) = \sum_{i=1}^{n} \frac{f'_{\theta_0}(y_i)}{f_{\theta_0}(y_i)}.$$

Under many circumstances the score interval has more accurate coverage than the corresponding Wald interval. The key difference between the Wald and Score intervals is that the former is based on the asymptotic distribution of the single observed $\hat{\theta}$ while the latter is based on the distribution of the score statistic $\ell'(\theta)$.

Table 1 shows the results of a simple simulation study for inference about an Exponential mean. For the Exponential model, $V(\theta) = \theta^2$. Nominal 95% exact, Wald and Score-based confidence intervals were constructed from datasets simulated over a variety of sample sizes. Here we arbitrarily (as this is a scale family) chose mean, θ, equal to 2. In terms of frequentist coverage probability we see that the Wald interval performs most poorly, attaining only approximately 90% coverage for the smaller sample sizes but improving steadily. The Score interval, in contrast, provides reasonably good

Table 1

Summary of results of simulation study for inference on an Exponential mean. Based on 10,000 simulated datasets for each sample size (n). Monte Carlo error for proportions: $\sqrt{(0.95)(0.05)/10000} \approx 0.002$.

		coverage probability			mean interval width		
		exact	Wald	Score	exact	Wald	Score
$\theta = 2$	n= 10	0.9486	0.9046	0.9536	2.9899	2.4825	4.0311
	n= 25	0.9524	0.9313	0.9563	1.6776	1.5617	1.8453
	n= 40	0.9480	0.9383	0.9512	1.2978	1.2425	1.3746
	n=100	0.9540	0.9461	0.9546	0.7965	0.7840	0.8153
	n=250	0.9497	0.9473	0.9502	0.4988	0.4964	0.5041

coverage probabilities but at the cost of requiring notably larger intervals for the smaller sample sizes. In either case we are using a Normal approximation to the distribution of $\overline{y}$. In the Exponential model, the variance of the data, and hence of many statistics of interest, is a deterministic function of the mean. The Wald interval fails mainly because it ignores this functional relationship. It does not perform well until n is large enough to allow for nearly constant variance $V(\theta)$ over the range of precision. The Score interval, however, explicitly accounts for the mean-variance relationship and performs well even for small n. So we see by this example that parametric frequentist methods still require particular attention to the mean-variance relationship and note that nonparametric frequentist methods tend not to address the issue in practice.

2.4 *The mean-variance relationship in Bayesian statistics*

The impact of the mean-variance relationship on the Bayesian approach to inference may seem moot. This is because parametric Bayesian methods account for the presence of a mean-variance relationship in an accurate and natural way provided the posited sampling distribution for the data is correct under all alternatives. Unfortunately these parametric methods are not usually robust to model misspecification. On the other hand, standard nonparametric Bayesian methods may be able to account for a mean-variance relationship by placing a prior on the space of all distributions Ferguson (1974). These so-called Dirichlet process prior models are relatively com-

plicated, however, and it is not clear how to place mass on specific mean-variance relationships. Also, the probability space of such models is more cumbersome than that of nonparametric frequentist models. For instance: If the prior is subjective, can the investigator state how much mass is placed, for example, on bimodal distributions?

One goal of this paper is to explore a nonparametric Bayesian method of analysis which is comparable to the probability models used in the nonparametric frequentist setting but which can accommodate a mean-variance relationship. We consider a Bayesian approach based on treating the frequentist nonparametric estimator, $\hat{\theta} = t(\vec{y})$, as a "coarsening" of the data. Our approximate likelihood, $\hat{p}(\hat{\theta}|\theta)$, is based on the asymptotic distribution of a sample statistic, i.e. the same likelihood used in nonparametric frequentist analyses. The use of an approximate distribution for $\hat{\theta}$ is robust in that it should hold under a wider variety of sampling distributions. We believe such an approach enjoys the characteristic strengths of both schools of statistical thought, robustness and validity of interval estimates, without the accompanying weaknesses. We compare this coarsened Bayesian approach to the fully parametric approach via simulation study. As a first step we wish to examine the accuracy of such an approach and consider its efficiency as measured by the width of posterior credible intervals. We do this in the context of a known mean-variance relationship, with and without the presence of a nuisance parameter, where interest lies in making inference for a population location parameter.

3. Methods

Standard parametric Bayesian inference for a population parameter, θ, requires the specification of a prior distribution $p(\theta)$ of the parameter and a sampling distribution $p(\vec{y}|\theta)$ for the data given θ. All inference is then based on the posterior distribution

$$\begin{aligned} p(\theta|\vec{y}) =& \frac{p(\vec{y}|\theta)p(\theta)}{\int p(\vec{y}|\theta)p(\theta)d\theta} \\ \propto& L(\theta|\vec{y})p(\theta) \end{aligned} \tag{1}$$

Here we assume satisfaction with $p(\theta)$. Equation 1 makes clear that much is required even for a simple Bayesian hypothesis test. For example, calculating the posterior probability of the null hypothesis $\theta = \theta_0$ requires integrating

the sampling distribution $p(\vec{y}|\theta)$ over all possible values of θ. As previously noted a frequentist hypothesis test requires knowledge only of what attains under the null hypothesis of $\theta = \theta_0$.

Our particular concern is that in the nonparametric frequentist world we may not know or even have a likelihood, $L(\theta|\vec{y})$. What then should one "plug-in" for the likelihood in equation 1? Ideally we should use the "true" sampling distribution. It is not clear when (or how or why) this true distribution is known and thus we might like to consider alternatives robust to model misspecification. There are many possibilities. Pratt et al. suggest using the approximate Normal distribution for the sample mean while treating the estimated variance as known Pratt et al. (1965). Boos & Monahan consider this large sample approximation as well as an approximation based on a bootstrap estimate of the sampling distribution of an estimator Monahan and Boos (1992) Boos and Monahan (1986). Lazar Lazar (2003) examines the use of empirical likelihood and the frequentist properties of resultant posterior intervals. None of these authors explicitly consider a mean-variance relationship and, in particular, we note that in the general setting bootstrapping cannot reproduce or discover the mean-variance relationship.

Thus we consider basing inference on the approximate posterior distribution

$$\hat{p}(\theta|t(\vec{y})) = \frac{\hat{p}(t(\vec{y})|\theta)p(\theta)}{\int \hat{p}(t(\vec{y})|\theta)p(\theta)d\theta}$$

where $t(\vec{y})$ is a statistic (or set of statistics), such as the sample mean or median, and $\hat{p}(t(\vec{y})|\theta)$ is an approximate sampling distribution that might be employed in a typical nonparametric frequentist setting. A Bayesian analysis using the information in $t(\vec{y})$ rather than $\vec{y}$ would lead naturally to computation of the posterior distribution $p(\theta|t(\vec{y}))$. Since $p(t(\vec{y})|\theta)$ is not necessarily known, for robustness to model misspecification we compute instead $\hat{p}(t(\vec{y})|\theta)$ - an approximation that is valid under a variety of $p(\vec{y})$. For inference about the population mean $\theta = E(y)$, for example, we might take the nonparametric estimate $t(\vec{y}) = \hat{\theta} = \frac{1}{n}\sum y$ and $\hat{p}(t(\vec{y})|\theta) = \text{Normal}(\theta, \text{var}(y|\theta)/n)$. For inference about the population median $\theta = F_y^{-1}(\frac{1}{2})$ we might take $t(\vec{y}) = \hat{\theta} = \hat{F}_n^{-1}(\frac{1}{2})$ and use its asymptotic distribution $\hat{p}(t(\vec{y})|\theta) = \text{Normal}(\theta, 1/4np_y^2(\theta))$. We note that in each case, under standard regularity conditions, the approximation becomes more accurate as sample size increases. We also note that

the variance functions of these asymptotic distributions may involve a nuisance parameter which would need to be estimated as well.

The Bayesian counterpart to the frequentist idea of a confidence interval is usually referred to as a "(posterior) credible interval" and corresponds to $100(1-\alpha)\%$ of the posterior probability $p(\theta|\vec{y})$. Commonly used are central posterior intervals and regions of highest posterior density. In contrast to frequentist confidence intervals, Bayesian credible intervals possess individual coverage probability. For example, a single 95% Bayesian credible interval for a parameter is interpreted as having 95% probability of containing the true parameter value. A single frequentist confidence interval, on the other hand, either contains the parameter value or not - there is no probability statement to be made. Frequentist coverage probabilities arise from the (possibly hypothetical) replication of a procedure and taking the ratio of favorable outcomes. Later, when we discuss the results of simulation studies, we look largely at frequentist probabilities but also consider the Bayesian interpretation. It should be noted that Bayesian analysis using a noninformative prior often leads to procedures with approximate frequentist validity.

3.1 *Example*

We illustrate the proposed method by considering inference about, in turn, a population mean and population median. The mean is of interest due to its scientific relevance and great familiarity. The median is of interest as it illustrates the difference between the parametric and nonparametric estimates and introduces the idea of nuisance parameters and how to account for such.

First, suppose we were to assume that $p(y|\theta)$ is the Exponential distribution with mean θ. Standard parametric Bayesian inference for θ given a sample of size n is based upon

$$\begin{aligned} p(\theta|\vec{y}) &\propto p(\theta) \quad \prod \frac{1}{\theta}\, e^{-y_i/\theta} \\ &\equiv p(\theta) \quad \left(\frac{1}{\theta}\right)^n \; e^{-\sum y_i/\theta} \end{aligned}$$

where the sample mean is a sufficient statistic for the parameter of interest. However, having no certain knowledge of the parametric model we would propose instead to base inference on the approximate posterior obtained by

using the asymptotic distribution for the sample mean, $t(\vec{y}) = \hat{\theta} = \frac{1}{n}\sum y_i$

$$\hat{p}(\theta|\hat{\theta}) \propto p(\theta) \;\; \phi\left(\frac{\hat{\theta}-\theta}{\theta/\sqrt{n}}\right)$$

In general, if $t(\vec{y})$ is a sufficient statistic then we are simply approximating the likelihood with a robust alternative. If $t(\vec{y})$ is not sufficient then the method should still prove robust for inference about the mean, though with some possible loss of efficiency (as compared to the unknown true model). Most importantly, this approximation will be valid for any family of distributions with the given mean-variance relationship.

For a slightly more interesting example consider the mixture distribution

$$p(y|\theta,\lambda) = \lambda \text{ Exponential } + (1-\lambda) \text{ Lognormal}$$

where both the Exponential and Lognormal distributions are parameterized with median θ and variance $c\,\theta^2$. That is

$$p(y|\theta,\lambda=1) = \frac{\log 2}{\theta}\exp\{\frac{-y\log 2}{\theta}\} \qquad \text{completely Exponential sample}$$

$$p(y|\theta,\lambda=0) = \frac{1}{\sqrt{2\pi}}\exp\{-(\log y - \log\theta)^2/2\}/y \qquad \text{completely Lognormal sample}$$

Suppose interest is in making inference about the population median. Again, standard inference would be based on the full parametric posterior

$$p(\theta,\lambda|\vec{y}) \propto p(\theta,\lambda) \;\; \prod p(y|\theta,\lambda)$$

which is easily manipulated to yield the marginal posterior for θ. Again, lacking certainty about the parametric model we propose basing inference on

$$\hat{p}(\theta|\hat{\theta}) \propto p(\theta,\lambda) \;\; \phi\left(\frac{\hat{\theta}-\theta}{c_\lambda\theta/\sqrt{n}}\right)$$

where c_λ is a nuisance parameter, independent of θ, that needs to be considered (the formula for c_λ comes from the asymptotic distribution for sample quantiles as noted earlier for the median). There is a hierarchy of approaches that are typically taken to address c_λ:

1. in parametric analyses one often implicitly assumes such nuisance parameters to be known;

2. in a more honest Bayesian analysis one might place a joint prior distribution on the nuisance parameter and the parameter of interest;

3. in a rather simple-minded approach (the common frequentist approach based on asymptotics with consistent estimators) one might just calculate some plug-in estimate of c_λ and then treat it as known.

In the next section we examine the performance of the approximate posterior distribution for θ where we take these various approaches for dealing with c_λ.

4. Simulations and results

Using the mixture distribution example from the previous section we consider, in turn, inference for the mean and inference for the median using the proposed approximate posterior distributions. Specifically, we generate data from

$$y|\theta, \lambda \sim \lambda p_1(y|\theta) + (1-\lambda)p_2(y|\theta)$$

where $p_1(y|\theta)$ is an Exponential distribution and $p_2(y|\theta)$ is from a class of Lognormal distributions. When interest is in inference for the mean we consider parameterizations of p_1 and p_2 such that $E(y|\theta) = \theta$ and $\text{var}(y|\theta) = \theta^2$. We also consider the distribution of λ as a point mass at 1, 0, and 1/2, corresponding to samples that are entirely Exponential, entirely Lognormal, or a 50/50 mixture of the two, respectively. When we turn our attention to inference about the median, we reparameterize p_1 and p_2 so that $p(\theta)$ is explicitly a prior for θ as the median. We also examine the case when $\lambda \sim \text{Uniform}(0, 1)$, that is, the samples are a random mixture of Exponential and Lognormal random variables.

For each sample size (n=10, 25, 40, 100, 250) and sampling scheme (Exponential, Lognormal, mixture) combination, we sampled 10,000 values of θ from a Normal(10, 1) distribution and simulated a corresponding dataset. Using Markov Chain Monte Carlo we estimated the posterior distribution based on each of the parametric sampling distributions and the proposed approximations. We calculated 95% posterior credible intervals, (θ_L, θ_H) for

each sampled θ by setting θ_L equal to the 2.5% and θ_H to the 97.5% MCMC sample quantiles (alternatively one might consider the highest posterior density (HPD) intervals). We also calculated the width of each interval and determined whether or not it contained θ.

Note that although the data in these simulations are generated from linear combinations of two particular parametric distributions the estimation approach is applicable to a much larger class. Here we require only that the mean-variance relationship of $E(\hat{\theta}) = \theta$ and $\text{var}(\hat{\theta}) = c\theta^2$ be approximately correct. We summarize the results of the simulations in the next two sections.

4.1 *Inference about a population mean*

It is a simple matter to obtain, analytically, the asymptotic relative efficiencies (ARE's) of the coarsened estimator and that of the parametric estimator for the Exponential and Lognormal models. For inference about the mean the coarsened approach and the Exponential model rely on the same statistic, $\overline{y}$, and thus the coarsened approach is asymptotically efficient. The sufficient statistic for the Lognormal model, $\overline{\log y}$, has asymptotic variance $\theta^2 \log 2$ while the asymptotic variance for $\overline{y}$ is θ^2. Thus the ARE of the coarsened estimate relative to the parametric Lognormal is $\log 2 \approx 0.693$.

We focus now on small sample properties. Table 2 displays the results of a simulation study comparing correct model parametric inference, incorrect model parametric inference, and the nonparametric coarsened approach. Inference based on the coarsened approach relies on the approximate Normal distribution for the sample mean with the assumed mean-variance relationship.

We see that inference based on the correct parametric model attains, as expected, approximately 95% frequentist coverage probabilities and that coverage improves with increasing sample size. Inference based on the proposed coarsened approach attains nearly the same coverage probabilities with no appreciable loss of efficiency in the larger sample sizes when the correct parametric model is Exponential and some small loss of efficiency when the data are Lognormal.

When inference is based on the incorrect parametric model, however, things can go terribly wrong. Inference based on the Exponential model

Table 2

Summary of results of simulation study for inference on a population mean. Based on 10,000 simulated datasets for each $p(y|\theta)$ and sample size (n) combination. Monte Carlo error for proportions: $\sqrt{(0.95)(0.05)/10000} \approx 0.002$.

true $p(y\|\theta)$		coverage probability assumed $p(y\|\theta)$			mean interval width assumed $p(y\|\theta)$		
		exp	lnorm	coarse	exp	lnorm	coarse
Exponential	n= 10	0.947	0.907	0.944	3.725	3.698	3.705
	n= 25	0.948	0.817	0.944	3.490	3.353	3.480
	n= 40	0.953	0.727	0.952	3.295	3.063	3.291
	n=100	0.952	0.409	0.950	2.755	2.320	2.760
	n=250	0.949	0.081	0.947	2.084	1.574	2.090
Lognormal	n= 10	0.946	0.945	0.938	3.723	3.656	3.709
	n= 25	0.950	0.953	0.946	3.491	3.339	3.482
	n= 40	0.949	0.947	0.947	3.296	3.092	3.296
	n=100	0.950	0.947	0.946	2.756	2.481	2.763
	n=250	0.948	0.946	0.947	2.083	1.807	2.090
50/50 mixture	n= 10	0.952	0.933	0.947	3.722	3.676	3.707
	n= 25	0.950	0.904	0.945	3.490	3.348	3.482
	n= 40	0.947	0.874	0.942	3.297	3.081	3.293
	n=100	0.953	0.746	0.952	2.757	2.402	2.760
	n=250	0.949	0.489	0.950	2.086	1.690	2.090

when the data are truly Lognormal poses no difficulties since the likelihood from the Exponential model is similar to that used in the coarsened approach. The reverse, however, is not true because we are inappropriately estimating a function of the geometric mean rather than the arithmetic mean. This inappropriate estimate becomes more precise and hence more incorrect as sample size increases. For instance, when a Lognormal model is assumed but the data come from an Exponential distribution we see that our coverage probability is only 91% for a sample size of 10 and decreases to about 8% for a sample size of 250. We note again that poor coverage occurs even at small sample sizes where the Kolmogorov-Smirnov statistic shows low power to differentiate the two distributions (e.g. about 12 and 25% when $n = 40$ when the data are truly Lognormal and Exponential, respectively), hence model checking will not solve the problem.

4.2 *Inference about a population median*

We again consider a sample $\vec{y}$ from a mixture of Exponential and Lognormal distributions but, as stated, reparameterized so that $p(\theta)$ is explicitly a prior distribution for θ as the median. For the coarsened approach we use the sample median, $\hat{\theta} = F_n^{-1}(\frac{1}{2})$, which can be shown (e.g. Ferguson (1996)) to have an asymptotic Normal distribution with $E(\hat{\theta}) = \theta$ and $\text{var}(\hat{\theta}) = 1/[4np_y^2(\theta)]$. In the present case the variance formula may be written $\text{var}(\hat{\theta}) = c_\lambda\ \theta^2/n$, where $c_1 = 1/(\log^2 2)$ for the Exponential distribution and $c_0 = \pi/2$ for the particular class of Lognormal distributions considered here. As mentioned earlier, there are different ways of dealing with this nuisance parameter (including the common approach of just ignoring it and the mean-variance relationship). We obtain a simple plug-in estimate for the mean-variance relationship in the following manner. For each simulated data set we took the variance stabilizing logarithmic transformation of the data (failure to do so produced poor results) and bootstrapped (Efron and Tibshirani (1993)) the resulting sample median to obtain a variance estimate, $\hat{V}(\widehat{\log\theta})$. Doing this and applying the δ-method allowed us to set $\hat{c_\lambda} = n\hat{V}(\widehat{\log\theta})$. For each run of the simulation we used 200 bootstrap samples to obtain this estimate. This is an admittedly ad-hoc approach but our simulations show reasonable performance and the approach is better than basing confidence intervals on a single plug-in estimate of the variance as is the case with a Wald interval (see Table I). Aside from this additional estimation of c_λ for the coarsened approach we proceeded in a similar manner

as in the case of the sample mean simulation.

Just as in the case of the mean, the ARE's of the coarsened and parametric approaches for inference about the median are simple to obtain analytically. For inference using the Exponential model when the data are Exponential the asymptotic variance is $\frac{\theta^2}{\log^2 2}$ while that of the sample median is simply θ^2 leading to an ARE of $\log^2 2 \approx 0.480$. For the Lognormal estimator the asymptotic variance is $\frac{2\theta^2}{\pi}$ and so the ARE of the coarsened estimate relative to the parametric Lognormal is $\frac{2}{\pi} \approx 0.637$. We now consider some small sample results.

Table 3 summarizes the results for inference about the median when λ is fixed at 0, 1 and 1/2. As in the case of inference about the mean, we see that the coarsened inference is nearly as accurate as the correct parametric inference. However, as expected, there is a loss of efficiency when using the sample median since it is not a sufficient statistic for either parametric distribution. The parametric models tend toward smaller confidence intervals but with incorrect coverage probabilities when the model is wrong. Using an incorrect model results in increasingly poor results as sample size increases - similar to the results found for inference about the mean. Assuming the Lognormal model when the correct model is Exponential leads to coverage probabilities ranging from 93% down to 21% for sample sizes of 10 to 250, respectively.

Furthermore, and in contrast to the results obtained for inference about the mean, the assumption of the Exponential model when the data actually come from the Lognormal distribution also results in very poor coverage probabilities: For a sample size of 10 the coverage probability is approximately only 93% and worsens to an approximate coverage of 56% for a sample size of 250. We also note that the accuracy of the coarsened approach is not as good as one might hope: the nominal 95% credible intervals only cover the true θ approximately 93% of the time even in the larger sample sizes.

The failure of the coarsened approach to attain 95% coverage is at least in part due to the way we estimated the nuisance parameter c_λ and then treated it as known. We explore this further by considering a simulation study of median estimation when the data come from a random λ-mixture of

Table 3

Summary of results of simulation study for inference on a population median. Based on 10,000 simulated datasets for each $p(y|\theta)$ and sample size (n) combination. Monte Carlo error for proportions: $\sqrt{(0.95)(0.05)/10000} \approx 0.002$.

true $p(y\|\theta)$		coverage probability assumed $p(y\|\theta)$			mean interval width assumed $p(y\|\theta)$		
		exp	lnorm	coarse	exp	lnorm	coarse
Exponential	n= 10	0.947	0.932	0.923	3.723	3.763	3.720
	n= 25	0.951	0.896	0.931	3.488	3.522	3.618
	n= 40	0.946	0.838	0.930	3.296	3.298	3.501
	n=100	0.947	0.611	0.935	2.756	2.647	3.157
	n=250	0.950	0.212	0.934	2.085	1.874	2.619
Lognormal	n= 10	0.933	0.947	0.925	3.690	3.733	3.689
	n= 25	0.920	0.951	0.935	3.443	3.488	3.556
	n= 40	0.892	0.947	0.932	3.248	3.290	3.404
	n=100	0.797	0.948	0.937	2.754	2.743	3.002
	n=250	0.559	0.947	0.937	2.159	2.070	2.413
50/50 mixture	n= 10	0.939	0.936	0.919	3.710	3.747	3.703
	n= 25	0.937	0.929	0.933	3.463	3.504	3.588
	n= 40	0.927	0.918	0.934	3.270	3.294	3.457
	n=100	0.897	0.847	0.935	2.756	2.700	3.084
	n=250	0.813	0.660	0.934	2.125	1.980	2.515

Exponential and Lognormal distributions. We consider a hierarchy of model assumptions where the inference is based on increasing levels of approximation for the posterior distribution:

1. $p(\vec{y}|\theta, c_\lambda)p(\theta, c_\lambda)$ - a full parametric model for θ and λ;

2. $\hat{p}(t(\vec{y})|\theta, c_\lambda)p(\theta, c_\lambda)$ - a halfway coarsened approach using the approximate Normal distribution for the sample median but estimating the nuisance parameter c_λ using the distribution of λ obtained from the full parametric analysis (this allows us to gain some insight into our bootstrap estimation of the nuisance parameter);

3. $\hat{p}(t(\vec{y})|\theta, \hat{c}_\lambda)p(\theta)$ - the fully coarsened nonparametric approach where the nuisance parameter is estimated via the bootstrap and then treated as known.

Table 4 displays results of this last simulation study. In terms of frequentist criteria our average coverage probabilities are maintained at a reasonably high level and are clearly improving with increasing sample size. As for efficiency, there is some loss - mainly due to the insufficiency of the sample median on which our inference is based. We again note that the loss in accuracy is due in part to our plug-in estimate of the nuisance parameter in the fully coarsened approach. We believe that the lower average coverage probability is due to the failure to account for the uncertainty in the estimation of the variance. Finally, we note that there is no apparent additional efficiency loss related to the estimation of the nuisance parameter.

We also turn to Table 4 as we consider a Bayesian interpretation of our results. If one considers the results from the full parametric analysis as "truth" then, by definition, each individual 95% posterior credible interval we construct contains exactly 95% of the posterior probability. As we remove parametric assumptions and consider the halfway and fully coarsened approaches, however, our individual credible intervals are not necessarily in agreement. As compared to the fully parametric model sometimes an individual coarse credible interval contains less, and sometimes more, of the posterior probability. That is, in the coarsened approach we may "on average" have 95% coverage probability but not necessarily always have 95% coverage. The last column of the table provides some information regarding the distribution of posterior coverage probabilities under the assumption

Table 4

Summary of results of simulation study for inference about a median from a mixture distribution. Based on 10,000 simulated datasets for each sample size (n). Monte Carlo error for proportions: $\sqrt{(0.95)(0.05)/10000} \approx 0.002$. *The last column gives information regarding the distribution of the posterior coverage probabilities as compared to the full parametric model.*

	method	coverage probability	mean interval width	distribution of posterior coverage probability (10%, 90%)
n=10	full parametric model	0.947	3.736	-
	coarsened θ, parametric λ	0.940	3.752	(0.916, 0.965)
	coarsened θ, bootstrap c_λ	0.925	3.702	(0.891, 0.964)
n=25	full parametric model	0.951	3.519	-
	coarsened θ, parametric λ	0.944	3.626	(0.913, 0.968)
	coarsened θ, bootstrap c_λ	0.934	3.591	(0.894, 0.968)
n=40	full parametric model	0.945	3.338	-
	coarsened θ, parametric λ	0.940	3.510	(0.911, 0.971)
	coarsened θ, bootstrap c_λ	0.924	3.456	(0.884, 0.971)
n=100	full parametric model	0.951	2.820	-
	coarsened θ, parametric λ	0.948	3.127	(0.903, 0.978)
	coarsened θ, bootstrap c_λ	0.934	3.083	(0.876, 0.980)
n=250	full parametric model	0.947	2.155	-
	coarsened θ, parametric λ	0.946	2.539	(0.891, 0.984)
	coarsened θ, bootstrap c_λ	0.940	2.515	(0.863, 0.987)

that the full parametric model is correct. When n=10, for example, we see that 80% of the intervals computed using the halfway coarsened approach cover between 92% and 97% of the posterior probability as measured by the full parametric model. For the fully coarsened approach 80% of the intervals contain between 89% and 96% of the same posterior probability. Thus we see that, for specific posterior credible intervals, there is not necessarily agreement among the approaches.

5. A real data example

Zhou and Gao Zhou and Gao (1997), compare four methods of constructing confidence intervals for the mean of skewed data having an approximately Lognormal distribution. The four methods considered are:

1. naive method - the exponentiation of a Wald-type interval based on log-transformed data;
2. Angus' conservative method - based on an approximate pivotal statistic;
3. parametric bootstrap - a bootstrap interval based on the approximate pivotal statistic used in Angus' method;
4. Cox's method - based on the UMVU estimators for $\log \theta$ and its variance under the assumed Lognormal distribution.

Simulation studies revealed that the naive method fails - the coverage error increases with σ^2 *and* with increases in sample size. For smaller sample sizes the bootstrap method has the smallest coverage error. Angus' method always gives too wide intervals but shows improvement as sample size increases. The Cox method has smallest error for moderate sample size (around n=50) and is comparable to the bootstrap for small samples when σ^2 is large. Also, the Cox method has the smallest intervals among the three appropriate methods. It is not clear how robust the Cox and bootstrap methods are.

The authors also apply the four methods to a real data set consisting of 355 measures of hospital charges following knee replacement procedures. The sample mean is \$9620.8, the sample median is \$8917.9 and the sample standard deviation \$3455.3. The data are noticeably right skewed (see Figure 2).

The reported 90% confidence intervals for the mean are:

method	CI	width
naive	(8839.6, 9363.7)	524.1
conservative	(9325.4, 9924.9)	599.5
bootstrap	(9330.1, 9897.1)	567.0
Cox	(9326.4, 9893.2)	566.8
coarse	(9336.9, 9896.3)	559.4

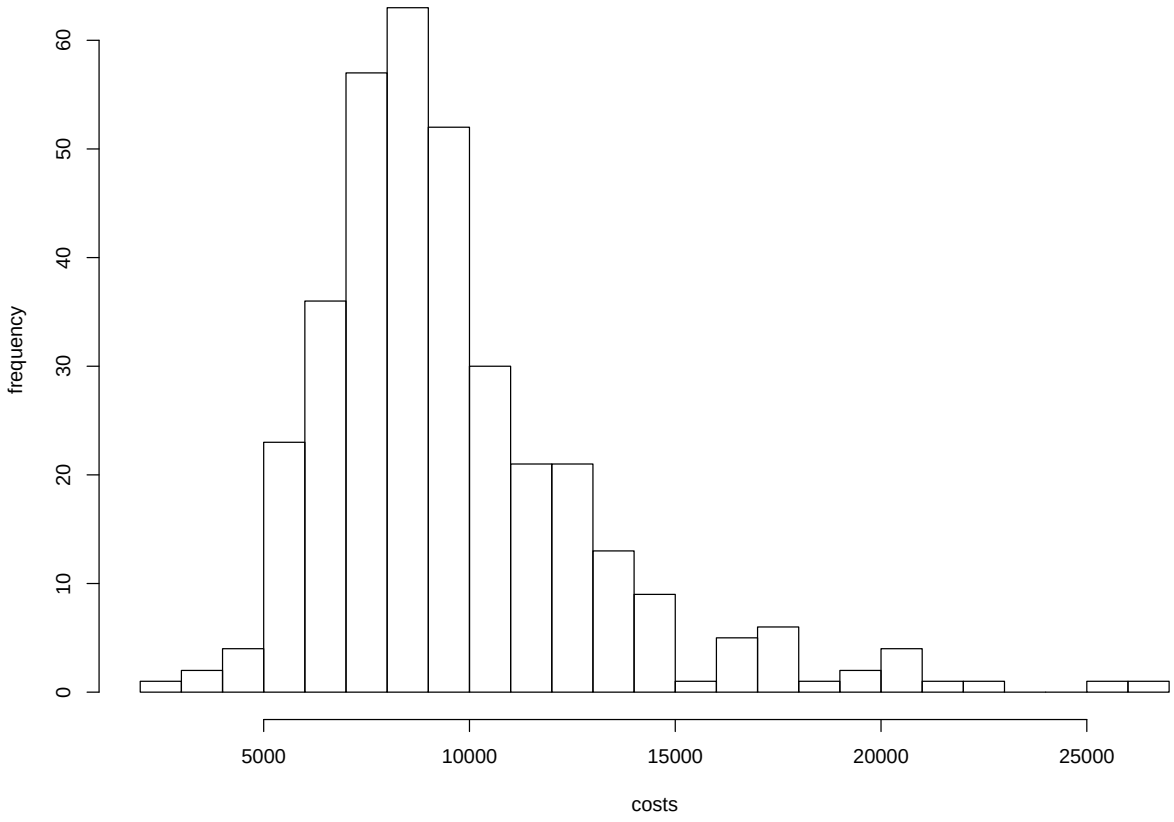

Figure 2. Histogram of the hospital charges following knee surgery from the paper by Zhou and Gao, 1997.

Based on the proposed coarsened approach using a noninformative prior we estimate a 90% confidence interval of (9336.9, 9896.3) with a width of 559.4 - comparable to both the bootstrap and Cox method.

For this particular health cost study inference about the mean was of primary interest. At other times and in the presence of such skewed data sociological questions might be better addressed by median inference. With the proposed coarsened approach it is a simple matter to obtain a confidence interval for the median. Using a noninformative prior we obtain a 90% confidence interval of (8658.0, 9183.8). It is also possible, if desired, to place an informative prior specifically on the mean or median rather than the parameters μ and σ of the standard parameterization of the Lognormal distribution.

6. Discussion

Nonparametric inference may be problematic in the presence of a mean-variance relationship. Parametric Bayesian methods account for a mean-variance relationship in an accurate and natural manner if the proposed parametric model is correct, but such models are not usually robust to misspecification. Standard nonparametric Bayesian methods can be more robust but are relatively complicated and the probability space is cumbersome as compared to that of standard frequentist models. As a simple and robust alternative, we propose replacing $L(\theta|\vec{y})$ in the standard Bayesian parametric analysis by an approximate sampling distribution that might be used in a nonparametric frequentist procedure. Our use of Bayesian methodology with the proposed replacement for the parametric likelihood allows us to account for a mean-variance relationship while attaining robustness to model misspecification. The method is intuitive and relatively simple to implement. An additional advantage of this approach is that by basing nonparametric Bayesian and frequentist inference on the same probability models, Bayesian analyses are made more accessible and communicable to the larger scientific community. Among applications, in the setting of group sequential clinical trials, the use of the coarsened approach facilitates the evaluation of clinical trial designs and the reporting of results in both frequentist and Bayesian contexts. In order to address the population of priors represented in the medical community, Bayesian inference can be represented graphically via a sensitivity analysis using the coarsened approach and a range of Normal priors Emerson et al. (2003).

A few key issues regarding the coarsened approach need to be addressed. We began by assuming the mean-variance relationship to be known and then introduced a nuisance parameter. We explored our ability to handle a nuisance parameter by crude methods as well as by use of formal prior distributions. If the method makes use of a formal prior can it still be considered nonparametric? We believe the answer is: Yes. In the real world the mean-variance relationship is nonidentifiable in general, and any formal prior placed on it is in essence a sensitivity analysis. Perhaps, then, a sensitivity analysis to a prior should be considered. Broadly, the nexus of the mean-variance relationship, prior distributions and model assumptions needs to be formalized.

In more practical terms we are currently researching extensions to more general classes of unknown mean-variance relationships given more than a single sample. Also of interest is a reasonable method of comparison to the standard nonparametric Bayesian approach that somehow limits priors to particular classes of mean-variance relationships.

acknowledgements
This research was supported in part by NHLBI grant R01 HL69719-01.

References

Boos, D. D. and Monahan, J. F. (1986). Bootstrap methods using prior information. *Biometrika* **73**, 77–83.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, London.

Emerson, S. S., Kittelson, J. M. and Gillen, D. L. (2003). Evaluation of group sequential clinical trial designs. *to be submitted* .

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2**, 615–629.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman and Hall, London.

Lazar, N. A. (2003). Bayesian empirical likelihood. *Biometrika* **90**, 319–326.

Lumley, T., Diehr, P., Emerson, S. and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health* **23**, 151–169.

Monahan, J. F. and Boos, D. (1992). Proper likelihoods for bayesian analysis. *Biometrika* **79**, 271–278.

Pratt, J. W., Raiffa, H. and Schlaifer, R. (1965). *Introduction to Statistical Decision Theory.* McGraw-Hill, New York.

Zhou, X. and Gao, S. (1997). Confidence intervals for the log-normal mean. *Statistics in Medicine* **16**, 783–790.