Model averaging quantiles for censored data

R. Nysen[1], M. Aerts[1], C. Faes[1]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Universiteit Hasselt, Belgium

E-mail for correspondence: ruth.nysen@uhasselt.be

Abstract: Quantiles are of interest in food safety data dealing with a limit of detection. The limit of detection introduces a lot of uncertainty in the left tail of the underlying distribution, making quantile estimation for this part of the distribution difficult. Therefore we fit a model to the data and derive the model-based estimate for the quantile. Since the true distribution is unknown, model averaging is used to combine information from a set of models. In this paper we discuss two approaches to use model averaging for quantiles. The methods are applied to a data example and compared in a simulation study. The effect of an increasing percentage of censoring on the estimates is explored.

Keywords: Censoring; Model averaging; Quantiles.

1 Introduction

We are interested in the lower quantiles of the distribution: e.g. which concentration is the cut-off for the 10% lowest concentrations? The most common estimate is the nonparametric or empirical quantile. However, censoring occurs in particular in the left tail of the distribution and introduces a lot of uncertainty about the quantiles. Therefore we fit a model to the data and derive the model-based estimate for the quantile. First we fit several parametric models to the data. All models are related to the log-normal distribution, a distribution that is regularly used in food safety data. Next we consider a semi-nonparametric family of distributions that consists of extensions of the log-normal distribution. Gallant and Nychka (1987) and Fenton and Gallant (1996) studied this family of distributions. Nysen, Aerts and Faes (2012) based a goodness-of-fit test for censored data on this semi-nonparametric family of distributions.
We can select the best model from the set of parametric ($\mathcal{M}_P$) and semi-nonparametric models ($\mathcal{M}_S$), based on a model selection criterion, e.g. Akaike's information criterion (AIC). This model is considered as the best approximating model and the cumulative distribution is used to estimate the quantile. However, if we would have a second and similar data set, it is possible that a different model is selected by the model selection criterion. This might result in a quite different estimate of the quantile. By

selecting one single model, we ignore the model uncertainty. The idea of model averaging is to start from a large set of plausible models and combine information from all models. In Section 2 we discuss two approaches to apply model averaging for quantiles. The approaches are applied to a data example in Section 3 and compared in a simulation study in Section 4. Although we focus on left censored data in this paper, the estimation can deal with other types of censoring, like right and interval censoring.

2 Model averaging of quantiles: two approaches

Let $M_i, i = 1 \ldots, K$ be a rich set of candidate models, with F_i the corresponding cumulative distribution function. The natural parameters θ_i are estimated by maximum likelihood theory and their variance-covariance matrix is denoted by $\text{Var}(\hat{\theta}_i)$. Suppose we want to estimate the p-quantile of the data ξ_p, where p is a fixed number between 0 and 1. There are several approaches in model averaging to obtain an estimate for the quantiles.
In the first approach, the quantile is estimated for each candidate model and the model averaged estimate is a weighted average of the K estimates. Based on model M_i, the quantile is obtained by $\xi_{p,i} = F_i^{-1}(p; \theta_i)$. The variance of the quantile can be approximated by the delta method:

$$\text{Var}(\hat{\xi}_{p,i}) \approx \nabla F_i^{-1}(p; \theta_i)^T \text{Var}(\hat{\theta}_i) \nabla F_i^{-1}(p; \theta_i). \tag{1}$$

In (1) the gradient $\nabla F_i^{-1}(p; \theta_i)$ is with respect to θ_i and can be estimated by $\nabla F_i^{-1}(p; \hat{\theta}_i)$.
Burnham and Anderson (1998) calculate the weight for each model, based on the difference of its AIC with the smallest AIC of all candidate models:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2}\Delta_j\right)},$$

where $\Delta_i = \text{AIC}_i - \text{AIC}_{\min}$.
The model averaged value of the quantile $\xi_{p,MA1}$ is given by the weighted average of the estimates of all candidate models:

$$\hat{\xi}_{p,MA1} = \sum_{i=1}^{K} w_i \hat{\xi}_{p,i} \tag{2}$$

with estimated variance

$$\widehat{\text{Var}}(\hat{\xi}_{p,MA1}) = \left[\sum_{i=1}^{K} w_i \sqrt{\widehat{\text{Var}}(\hat{\xi}_i) + \left(\hat{\xi}_i - \hat{\xi}_{p,MA1}\right)^2} \right]^2.$$

The variance estimator is the sum of two components. The first component is the conditional variance, given model M_i. The second component reflects the variation in the estimates across the K models.

A second approach applies model averaging at a different level. The distribution of each candidate model is estimated and combined in a model averaged cumulative distribution function. The quantile of the combined distribution is the model averaged quantile estimate. Let x be a real number in the domain of the candidate distribution function. The averaged distribution function is given by

$$\hat{F}_{MA}(x;\theta) = \sum_{i=1}^{K} w_i \hat{F}_i(x;\theta_i)$$

with θ the vector of the natural parameters of all candidate models. The estimated variance of $\hat{F}_{MA}(x;\theta)$ is

$$\widehat{\mathrm{Var}}(\hat{F}_{MA}(x)) = \left[\sum_{i=1}^{K} w_i \sqrt{\widehat{\mathrm{Var}}(\hat{F}_i(x;\theta_i)) + \left(\hat{F}_i(x;\theta_i) - \hat{F}_{MA}(x;\theta)\right)^2}\right]^2. \tag{3}$$

The quantile can be estimated by

$$\hat{\xi}_{p,MA2} = F^{-1}_{MA}(p;\theta). \tag{4}$$

Based on the implicit function theorem, the variance of $\hat{\xi}_{p,MA2}$ can be approximated by

$$\mathrm{Var}(\hat{\xi}_{p,MA2}) = \frac{\mathrm{Var}(\hat{F}_{MA}(\hat{\xi}_{p,MA2}))}{\hat{f}^2_{MA}(\xi)},$$

where $\mathrm{Var}(\hat{F}_{MA}(\hat{\xi}_{p,MA2}))$ can be estimated by (3). Because the true value ξ is unknown, we need to estimate $\hat{f}_{MA}(\xi)$ by $\hat{f}_{MA}(\hat{\xi}_{p,MA2})$.
In general, the estimates (2) and (4) result in different estimates, because quantile estimation is not a linear functional. When estimating for instance the cumulative distribution function, the two approaches would be equivalent. We compare the performance of the two approaches in a simulation study, but first we illustrate the approaches in a real data analysis. We will consider a set of parametric and semi-nonparametric models $\mathcal{M}_P \cup \mathcal{M}_S$.

3 Data analysis

The motivating data for this study, is a sample of measurements of cadmium level. The data set consists of almost 100 observations, but 37% of the measurements are censored by the limit of detection (LOD). The LODs are small and lie in between 0.001 and 0.01. A visual representation of the data is given in Figure 1, where a kernel density of the logarithm of the concentrations is shown.

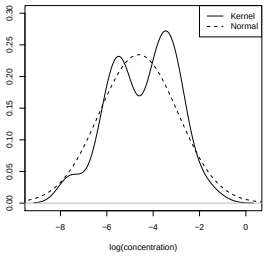

FIGURE 1. Cadmium data. Kernel density function of the concentrations where LODs are replaced by LOD/2 (solid line), transformed to the log scale. Normal fit (dashed line) based on likelihood for censored data.

Figure 1 shows that the (log-)normal distribution is a reasonable fit, but other distribution might fit better. Table 1 lists the AIC for several parametric models and for the semiparametric extensions of the log-normal distribution. The best global fit is given by the log-skew-t distribution and the distribution with two extra parameters from the semi-nonparametric family of distributions. For each model, the 10% and 25% quantiles are given in Table 1. For comparison, we also computed the nonparametric quantile estimates, where we substituted the LOD in the data with LOD/2 to cope with the left censoring. We see that the parametric models have smaller estimates for the quantiles. This is due to the left censoring, because the parametric models put weight on the left of the distribution, while the raw data only provide information on the LOD.
The quantile estimates based on the two model averaging approaches from Section 2 are close together. If we average the distribution function before computing the quantile (approach 2), the 10% quantile is slightly smaller. The estimated standard error for the estimate of the second approach is also smaller.

4 Simulation study

The performance of both methods is demonstrated in a simulation study. Data are simulated from 3 kinds of distributions, i.e. the log-normal distribution, the gamma distribution with same mean and variance as the log-normal distribution, and a mixture of 2 log-normal distributions. For each distribution, 500 samples are drawn. The censoring scheme is defined by limits of detection that correspond to five quantiles (1%, 5%, 10%, 20%, 25%) of the original log-normal distribution, resulting in 12% censoring on average when sampling from the log-normal distribution.
On each sample we fit 13 models: seven parametric (log-normal, log-skew-normal, log-t, log-skew-t, gamma, Weibull), denoted by $\mathcal{M}_P$, and seven

TABLE 1. Cadmium data. AIC and model averaging weights. For each parametric and semi-nonparametric model, the quantile was estimated (standard error).

	AIC	weight	$\hat{\xi}_{0.1}$	$\hat{\xi}_{0.25}$
Non-parametric			0.00250	0.00500
Log-normal	-135.7	0.000	0.00110 (0.00036)	0.00310 (0.00076)
Log-skew-n	-149.3	0.023	0.00045 (0.00026)	0.00226 (0.00086)
Log-t	-133.4	0.000	0.00111 (0.00036)	0.00314 (0.00078)
Log-skew-t	-154.2	0.262	0.00000 (0.00000)	0.00109 (0.00122)
Weibull	-149.1	0.021	0.00061 (0.00019)	0.00291 (0.00058)
Gamma	-151.8	0.081	0.00029 (0.00021)	0.00222 (0.00093)
GenGam	-150.0	0.033	0.00023 (0.00021)	0.00208 (0.00097)
SemiNP1	-151.5	0.068	0.00060 (0.00013)	0.00128 (0.00028)
SemiNP2	-154.3	0.275	0.00019 (0.00006)	0.00137 (0.00133)
SemiNP3	-152.6	0.120	0.00022 (0.00054)	0.00149 (0.00118)
SemiNP4	-150.7	0.047	0.00023 (0.00111)	0.00156 (0.00109)
SemiNP5	-148.8	0.018	0.00021 (0.00372)	0.00160 (0.00105)
SemiNP6	-150.3	0.038	0.00025 (0.00293)	0.00152 (0.00122)
SemiNP7	-148.3	0.014	0.00015 (0.00581)	0.00154 (0.00119)
$\mathcal{M}_P \cup \mathcal{M}_S$ 1 *			0.00020 (0.00052)	0.00147 (0.00120)
$\mathcal{M}_P \cup \mathcal{M}_S$ 2 **			0.00015 (0.00050)	0.00148 (0.00107)

* quantiles are estimated for each distribution and then averaged (MA1)
** cumulative distribution function is averaged (MA2)

extensions of the log-normal defined by the semi-nonparametric family of distributions, denoted by $\mathcal{M}_S$. The generalized gamma distribution was not included in the simulation study due to convergence issues. We provide here the preliminary results for data simulated from the log-normal distribution, estimating the 1% quantile. Table 2 shows the variance and the squared bias, together with the sign of the bias. The mean squared error (mse) is obtained by adding these two quantities.

A first observation is that the mean squared error is larger when the data are censored, which is to be expected because censoring introduces more uncertainty in the data and the estimation process. From the family of parametric models, the log-t and the true log-normal distribution provide the best fits. In the semi-nonparametric family of distributions, the models perform worse, because more parameters are added. Indeed, since we simulate from the log-normal distribution, the extra parameters are redundant to describe the data. If the data are not censored, there is no difference between the two modeling approaches, regarding the mean squared error. On the contrary, the bias is smaller for the second approach, while the variance is smaller for the first approach. In the censoring case, the first approach does result in a smaller mean squared error. The variance is higher in the

model averaging compared to the single-model inference, because the model selection uncertainty is now incorporated.
In future research we will compare the model averaging approaches for the other distributions, sample sizes and different quantiles. We will also consider a different family of distributions, e.g. restricted to the parametric models $\mathcal{M}_P$ or to the semi-nonparametric family of distributions $\mathcal{M}_S$.

References

Burnham, K.P. and Anderson, R.A. (1998). *Model selection and inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Fenton, V.M. and Gallant, A.R. (1996). Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics*, **74**, 77 – 118.

Gallant, A.R. and Nychka, D.W. (1987) Semi-nonparametric maximum likelihood estimation. *Econometrica*, **55**(2), 363 – 390.

Nysen, R., Aerts, M. and Faes, C. (2012), Testing goodness of fit of parametric models for censored data. *Statistics in Medicine*, **31**, 2374 – 2385.

TABLE 2. Simulation study. Data simulated from log-normal distribution with sample size 100. ($\times 10^{-5}$)

Censoring	No			Yes		
	bias^2 (sign)	var	mse	bias^2 (sign)	var	mse
Log-normal	0.258 (+)	3.700	3.957	0.283 (+)	4.381	4.664
Log-skew-n	0.338 (+)	7.184	7.522	2.266 (+)	10.909	13.175
Log-t	0.000 (+)	3.467	3.467	0.051 (−)	4.482	4.533
Log-skew-t	0.006 (+)	7.179	7.185	0.357 (−)	15.422	15.780
Weibull	40.546 (−)	0.252	40.799	43.228 (−)	0.211	43.439
Gamma	39.419 (−)	2.136	41.555	45.539 (−)	1.064	46.603
SemiNP1	0.494 (+)	6.076	6.569	0.020 (−)	12.088	12.108
SemiNP2	1.208 (+)	6.999	8.207	0.100 (+)	16.869	16.969
SemiNP3	1.467 (+)	8.543	10.010	0.601 (+)	19.938	20.540
SemiNP4	1.641 (+)	8.811	10.452	0.944 (+)	22.621	23.565
SemiNP5	1.673 (+)	9.383	11.056	1.394 (+)	23.612	25.006
SemiNP6	1.880 (+)	9.594	11.474	2.222 (+)	24.477	26.699
SemiNP7	2.053 (+)	10.417	12.471	3.280 (+)	25.546	28.826
$\mathcal{M}_P \cup \mathcal{M}_S$ 1 *	0.351 (+)	6.018	6.369	0.056 (+)	10.916	10.973
$\mathcal{M}_P \cup \mathcal{M}_S$ 2 **	0.248 (+)	6.121	6.369	0.199 (−)	13.016	13.215

* quantiles are estimated for each distribution and then averaged (MA1)
** cumulative distribution function is averaged (MA2)