Coverage performance of confidence intervals for the empirical variance distribution approximation, to the quantile estimating function

John P. D. Martin

Friday, October 23, 2015

Executive Summary

This paper demonstrates the good coverage performance of several confidence interval estimators for sample quantiles based on the empirical variance distribution (Martin (1)). The empirical variance distribution uses a quadratic polynomial proxy, to calculate quantile variance estimates in the percentile scale, for the quantile estimating function (Koencker & Bassett (2)).

Analysis of the coverage of the confidence interval estimators is based on repeated sampling results for several distributions and samples sizes, and convergence comparison is conducted to existing quantile regression bootstrap results, for an "intercept only model" which directly calculates sample quantiles.

It has been found that three confidence interval estimators calculated in the percentile scale before back-transformation to the original measurement scale, (i) the binomial distribution, (ii) the empirical variance distribution and (iii) the total variance $E[Var(\theta|x)] + Var(E[\theta|x])$ confidence intervals (3), all produce nominal 95% coverage for quantiles between 0.1-0.9 similar to bootstrap results.

It is observed, for estimator (ii) the empirical variance distribution, that the coverage $\rightarrow 0$ as θ or $(1-\theta) \rightarrow 0$. This behaviour occurs because this estimator is a conditional variance estimator $E[Var(\theta|x)]$ which incorporates only type I errors. It assumes all observed samples, will be a representative sample of the distribution and this assumption is weakest for extreme quantiles and small samples.

Estimator (iii) $E[Var(\theta|x)] + Var(E[\theta|x])$, has the empirical variance distribution for the first term and the second term $Var(E[\theta|x])$ is taken to be equal to $\theta^2 Pr(\theta > 0)Pr(\theta = 0)$ for $\theta <= 0.5$ and equal to $(1-\theta)^2 Pr(\theta < 1)Pr(\theta = 1)$ for $\theta > 0.5$ using the variance of random indicators. The second variance term is largest for extreme quantiles near 0 & 1 and small samples, reflecting the impact of type II errors, when the sample size is not large enough to consistently collect a representative sample of rare events. The binomial distribution based confidence interval estimator is also a total variance estimator.

Importantly, for small samples (< 200) and extreme quantiles near 0 & 1, non-linear quantile regression extrapolation of the observed data is required by one estimated data point (at each end) for small samples to produce comparable coverage performance of empirical variance distribution based estimators to bootstrap results. Implicitly, the bootstrapping approach is also extrapolating beyond the observed data for extreme quantiles because of the symmetric confidence intervals that are derived for that method.

The use of the quadratic polynomial proxy approach to calculating confidence intervals (i) employs the quantile estimating function to perform the backtransformation from the percentile scale, (ii) only the sample size and chosen quantile value are needed for calculations in the percentile scale, and (iii) a density function has been derived which allows small samples point estimates to be calculated self-consistently (without CLT assumptions). For very small samples and extreme quantiles near 0 & 1, (cubic polynomial) quantile regression of bootstrap samples of size 400, is used to expolate data ranges to improve coverage performance.

Introduction

Quantiles (4) are an order statistic of a distribution defined by the equivalent probability amount contained under the cumulative distribution function up to the (ordered) value of the quantile point.

That is, x is a k-th q-quantile for a variable X if

Pr[X < x] $\leq$ k/q or, equivalently, Pr[X $\geq$ x] $\geq$ (1 - k/q)

So the 25th percentile point is the 25/100 (k/q) 100-quantile point where 25% of the probability under the cumulative density function has occurred.

An equivalent calculation of quantile points has been demonstrated (2) using least absolute deviation (LAD) regression of the following quantile estimating function

$$\min_{b\,\epsilon\mathbb{R}}\{\theta\,|x_t - b| + (1-\theta)\,|x_t - b|\} \tag{1}$$

where θ is the quantile of interest and x_t are the observed sample/population elements of the distribution X. As the absolute value functions in equation 1, create a piecewise linear function shape (convex polytope) to the estimating function, linear programming techniques are required to solve the minimsation problem. As such, closed form expressions for the standard error of the quantile estimates are not available from this approach.

In (1), it was demonstrated by transformation to the percentile scale of a cumulative density function for unweighted samples, that (i) the above quantile estimating function in the percentile scale (when normalised to have the minimum close to 1)

$$\min_{b\,\epsilon(0,1)}\left(\frac{\frac{|k/q-b|}{(1-\theta)} + \frac{|k/q-b|}{\theta}}{\frac{n(|\max(k/q)|+|\min(k/q)|)}{2}}\right) \in (1 - 1/2n, 1] \tag{2}$$

where k/q are the quantile values of the x_t data points, can be closely approximated by a (smoothing) quadratic polynomial proxy with analytic coefficients (in this reference frame),

$$\min_{b\,\epsilon(0,1)}\left(\frac{1}{\theta(1-\theta)}b^2 + \frac{-2}{(1-\theta)}b + \frac{1}{(1-\theta)}\right) \in [1, \text{inf}) \tag{3}$$

and (ii) comparable sample quantile standard error estimates (to bootstrap estimates) for equation (1) may be obtained by (numerical) backtransformation of the quadratic polynomial proxy variance results using quantile regression values for the evd points (with an intercept only model)

A integral part of the approach is to use the empirical cumulative distribution function with averaging and interpolation (ecdf_int), for k/q values. Using this quantile definition, in the percentile scale, the k/q data points have the quantile values

1/2n, 3/2n, 5/2n, ... , (2n-1)/2n

where the sample endpoints are essentially assigned with half weights since the population endpoints (estimated from the sample) have the highest uncertainty. In the percentile scale, for unweighted (or equally weighted) samples, the spacing between data points is equal. Given the quadratic polynomial proxy is an analytic (smoothing) fit to quantile regression in this scale, and was shown in (1) to have equivalent parametrisation to the normal distribution, only the number of data points and the estimated θ obtained from quantile regression (in original measurement scale) was required to derive the distribution density function in this frame

$$f_{q-proxy}(b,\theta,n) = \left(\frac{1}{\int_0^1 \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\}\,db}\right)\exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\} \tag{4}$$

where

$$\sigma_{CLT} = \sqrt{\frac{\theta(1-\theta)}{n}} \tag{5}$$

thus no sorting of the data is required using the quadratic polynomial proxy to calculate standard errors and confidence intervals. In the large n limit, the density function of the quadratic polynomial proxy for the quantile estimating function (in the percentile scale) converges to a (CLT) normal distribution form

$$f_{q-proxy}(b,\theta,n)_{CLT} \rightarrow \frac{1}{\sigma_{CLT}\sqrt{2\pi}} \exp\{\frac{-(b-\theta)^2}{2\sigma^2_{CLT}}\} \tag{6}$$

On backtransformation to the original measurement scale, the shape of the resulting empirical variance distribution (evd) using ecdf_int, shares the step function character of the cumulative distribution function and in general exhibits some asymmetry compared to the convention of outputting symmetric bootstrap standard errors.

In this paper,

(i) the coverage performance of several confidence interval (CI) estimators using the empirical variance distribution density function in the percentile scale are assessed and nominal 95% coverage is achieved with comparable results to quantile regression bootstrap estimates.

(ii) for small samples, the issue of lack of data for extreme quantile CIs is resolved by extrapolated quantile regression estimates resulting in comparable performance to bootstrap estimates.

Confidence interval estimation and the law of total variance

The confidence interval of an estimator, defines the range of uncertainty of an estimator under repeated sampling, such that the interval contains the true population value of the parameter (the estimator is estimating) for (1-α)% of the repeated samples. Typically, two sided 95% confidence intervals are used for estimation of population parameters. If specific analytical tests of predetermined accuracy are envisaged between subsamples or with other datasets, then in addition, sample sizes also achieving specific statistical power β may be required. The significance level (1-α) requirement establishes the sample size required for an acceptable false positive rate of descriptive statistical inferences derived about the sample/population (type I errors) based on a single sample. The statistical power requirement is an additional constraint on the sample size for an acceptable false negative rate of analytical statistical inferences derived about the sample/population (type II errors) comparing subsamples or with other data.

Relevant to this paper, while the concern about coverage performance appears to be a significance level (1-α) requirement for confidence levels, the statistical power requirement may also become important for extreme quantiles near 0 & 1, depending on (i) the estimator properties and (ii) whether the sample & estimation method is sufficient to define the endpoints of the sample/population. This complexity for bounded estimation examples has been previously investigated for estimates of proportions near 0 & 1 (4).

The direct way to include both type I & type II errors (in coverage performance) is to use confidence interval estimators which estimate selfconsistently the total variance (3).

$$Var(\theta)_x = E[Var(\theta|x)] + Var(E[\theta|x]) \tag{7}$$

In this paper, three confidence interval estimators based on the empirical variance distribution, in the percentile scale, are investigated and compared to quantile regression bootstrap results. The list of estimators considered are

(i) binomial variance distribution,

(ii) empirical variance distribution,
(iii) empirical variance distribution plus $Var(E[\theta|x])$ approximation and
(iv) quantile regression bootstrap

binomial CI estimator

The binomial variance distribution formula (5) is a total variance estimator, selfconsistently including the high probability of 0 (1) occurring for repeated sampling of bernoulli experiments with probabilities close to 0 (1). That is why it was described as the exact distribution for the estimate of proportions (4) from repeated sampling. It has an naturally skewed distribution reflecting the bounded distribution of the possible outcomes. For large n, CLT behaviour is observed.

In use for quantile variance estimation, in the percentile scale, the estimated quantile 2.5th & 97.5th points applicable to a 95% confidence interval are obtained by dividing the binomial cumulative distribution function by the sample size

$$F(k/n;n,\theta)_{binom} = \frac{1}{n}\sum_{i=0}^{\lfloor k \rfloor} \frac{n!}{k!(n-k)!}\theta^i(1-\theta)^{(n-i)} \tag{8}$$

where (i) θ is substituted for the proportion terms in the usual proportion formula as shown, and i are integers between 0 & n.

After obtaining the 2.5th & 97.5th quantile points for $F(k/n;n,\theta)_{binom}$ in the percentile scale, the quantile estimating function is then used, in the original measurement scale, to calculate the 95% confidence interval boundaries. Since the binomial distribution is discrete, in principle, this estimator may overestimate the 95% confidence interval since, for example the 2nd (98th) quantile points may be the same as the 2.5th (97.5th) quantile points, so the coverage performance may exceed 95%.

evd CI estimator

The empirical variance probability distribution, given in equation (4), is basically a normal distribution rescaled because of the boundary constraints applicable to quantiles. It is also only an estimator of the form $E[Var(\theta|x)]$ rather than the total variance equation (7). This is because the quantile estimating function minimisation, that the quadratic polynomial proxy mimics, is strictly a conditional variance of the observed sample, ie. if the observed quantile is 0 or 1, the calculated variance is zero.

In use for quantile variance estimation, the effect of rescaling in equation (4), in the percentile scale, means the estimated quantile 2.5th & 97.5th points applicable to a 95% confidence interval are obtained by using a normal cumulative distribution function bounded to the interval [0,1]

$$F(q)_{evd} = \frac{1}{2}[1 + erf(\frac{q-\theta}{\sigma_{CLT}\sqrt{2}})] \; \epsilon \, [0,1] \tag{9}$$

After obtaining the 2.5th & 97.5th quantile points for $F(q)_{evd}$ in the percentile scale, the quantile estimating function is then used, in the original measurement scale, to calculate the 95% confidence interval boundaries. Since the evd estimator is only $E[Var(\theta|x)]$, the coverage performance may be very poor for small samples or extreme quantiles.

Without additional adjustments such as adding a $Var(E[\theta|x])$ term, or using Wilson confidence interval or using an adjusted beta function approach (add one success/add one failure) of (3), this evd estimator will perform similarly to how the Wald confidence interval performs for the estimation of proportions near 0 & 1 (3).

evd plus $Var(E[\theta|x])$ approximation

Using the variance expression for indicator random variables, the $Var(E[\theta|x])$ term is approximated, in the percentile scale, by

$$Var(E[\theta|k/q]) = E(\theta)^2 Var(k/q)_\theta \tag{10}$$

$$\approx \begin{cases} \theta^2 Pr(\theta = 0) Pr(\theta > 0) \;\; for\, \theta \leq 0.5 \\ (1-\theta)^2 Pr(\theta = 1) Pr(\theta < 1) \;\; for\, \theta > 0.5 \end{cases} \tag{11}$$

where var(k/q) term is a population variance, in the percentile scale, and so is an unconditional variance contribution to the total variance.

In use for quantile variance estimation, the total variance estimator is obtained by

(i) calculating the 15.88th (the one standard error) quantile of the $E[Var(\theta|x)]$ evd term,

(ii) the probabilities $Pr(\theta = 0)$, $Pr(\theta = 1)$, $Pr(\theta > 0)$, $Pr(\theta < 1)$ are calculated using equation (9) (or more simply in r using pnorm() function),

(iii) the 15.88th (the one standard error) quantile of the $Var(E[\theta|x])$ term is then calculated using equation (11),

(iv) the total variance is calculated as the squared sum of the $E[Var(\theta|x)]$ and $Var(E[\theta|x])$ standard errors according to equation (7), and

(v) the 2.5th & 97.5th quantile points of an adjusted variance distribution

$$F(q)_{evd+Var(E[\theta|x])} = \frac{1}{2}[1 + erf(\frac{q-\theta}{\sigma_{total\,var}\sqrt{2}})] \; \epsilon \, [0,1] \tag{12}$$

are obtained and then backtransformed to the original measurement scale, using the quantile estimating function.

The advantage of the empirical variance distribution based 95% confidence interval estimators is that the quantile values of the intervals can be determined very simply in the percentile scale. Table 1 lists a comparison the intervals for the above estimators as a function of θ and sample size

Table 1: 95% evd based CI intervals, in the percentile scale

Sample size	quantile	binomial dist	evd dist	total var dist
50	0.025	(0,0.08)	(0.01,0.0659)	(0,0.0732)
100	0.025	(0.0,0.6)	(0.0050,0.0547)	(0.0,0.0590)
1000	0.025	(0.0160,0.0350)	(0.0152,0.0346)	(0.0153,0.0347)
50	0.1	(0.02,0.18)	(0.0199,0.1832)	(0.0131,0.1869)
100	0.1	(0.05,0.16)	(0.0415,0.1587)	(0.0409,0.1591)
1000	0.1	(0.0820,0.1190)	(0.0813,0.1185)	(0.0814,0,1186)
50	0.5	(0.36,0.64))	(0.3613,0.6385)	(0.3614,0.6386)
100	0.5	(0.4,0.6)	(0.4019,0.5979)	(0.4020,0.5980)
1000	0.5	(0.4690,0.5310)	(0.4689,0.5309)	(0.4690,0.5310)

Sample size	quantile	binomial dist	evd dist	total var dist
50	0.9	(0.82,0.98)	(0.8164,0.9734)	(0.8131,0.9869)
100	0.9	(0.84,0.95)	(0.8411,0.9584)	(0.8409,0.9591)
1000	0.9	(0.881,0.918)	(0.8813,0.9185)	(0.8814,0.9186)
50	0.975	(0.92,1.0)	(0.9290,0.9886)	(0.9268,1.0)
100	0.975	(0.94,1.0)	(0.9436,0.9931)	(0.9410,1.0)
1000	0.975	(0.965,0.984)	(0.9652,0.9846)	(0.9653,0.9847)

It can be seen from Table 1, that (i) the evd dist estimator does not include q=0.0 (1.0) in the 95% CI which will contribute to poor coverage for this estimator with small samples and (ii) the approximate total variance distribution estimator is similar to the binomial distribution values where the major difference is the rounding executed by the discrete nature of the binomial distribution whereas the total variance distribution estimator is a continuous distribution.

bootstrap replicate estimates

The "quantreg" r package for quantile regression uses as default a delete-d jackknife method described by Portnoy (6). The default number of bootstrap samples is 200 and the measured variance of the distribution of bootstrap samples is used to estimate a symmetric confidence interval and the associated 2.5th & 97.5th quantiles, directly in the original measurement scale.

A comparison of the four estimators coverage performance is given in a later section of this paper for sample sizes of 50, 100 & 1000.

continuity correction for extreme quantiles in finite samples

Before proceeding to the calculations for the above estimators, however, it is also necessary to consider the need for a continuity correction of extreme quantiles. Importantly, for small samples and extreme quantiles near 0 & 1, for example the 2.5th & 97.5th percentiles, depending on the sample size there may not be sufficient data points collected to accurately define the confidence interval of such extreme quantiles and hence poor coverage results, in spite of the estimator.

For the bootstrapping approach, extrapolation beyond the observed data for extreme quantiles already occurs because of the symmetric confidence intervals that are generated. In the case of empirical variance distribution based estimators, this most obviously occurs because the ecdf_int used for the cumulative distribution function is only defined between (1/2n,1-1/2n) but the 95% confidence intervals of extreme quantiles may extend beyond that interval, between [0,1/2n) and (1-1/2n,1].

For example, for a sample size of 50, the lowest known quantiles (using ecdf_int) are 0.01 (1/2/50), 0.03 (3/2/50), 0.05 (5/2/50). The expected 2.5th percentile value is 0.025 which lies between the two lowest possible data points and the binomial distribution based 95% confidence interval is (0,.08) from Table 1. Without adequate data, via some continuity correction it is difficult to estimate the 0th quantile.

To improve coverage performance for evd based CIs for extreme quantiles and to be consistent with quantile regression, the following extrapolation procedure has been trialled. For 2.5th and/or 97.5th percentiles of a 95% confidence interval lying in the intervals [0,1/n) or (1-1/n,1], a quantile regression extrapolation (using a cubic polynomial fit) of the 0th, 100th quantile is estimated from a bootstrapped sample size of 400. These extra data points, one at either end of the sample distribution are only used in the variance estimation. Figure 1 shows some examples of the model extrapolation for different samples, of n=50.

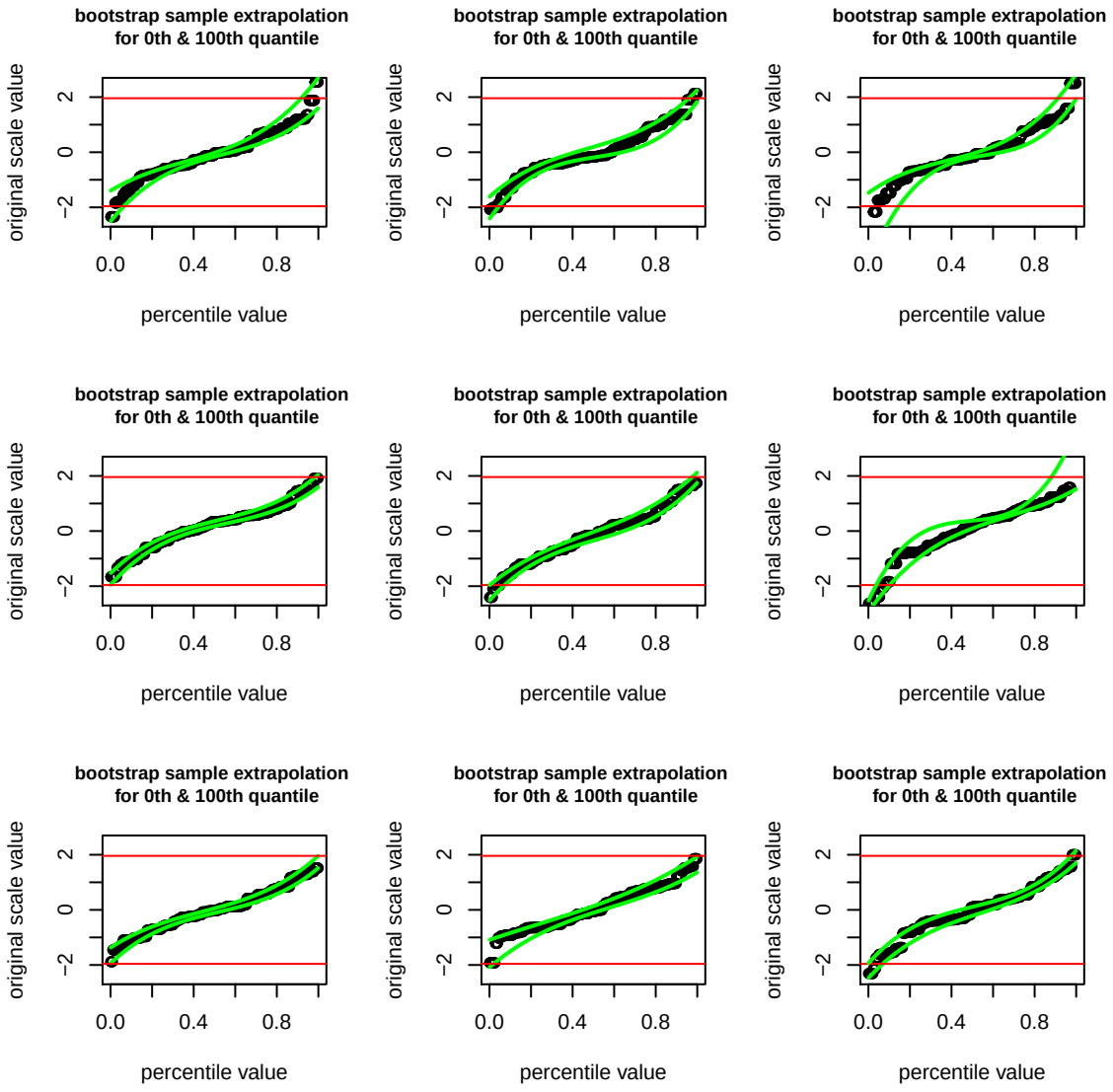


Figure 1: data extrapolation for variance estimation of extreme quantiles

where (i) the upper green line is the 100th quantile regression fit and the information obtained is the extrapolated value at q=1 and (ii) the lower green line is the 0th quantile regression fit and the information obtained is the extrapolated value at q=0. The choice of creating bootstrap samples of 400, was a compromise between (i) significantly decreasing the number of times under repeated sampling that the extrapolated values for 0th & 100th quantiles fail to exceed the known population values (eg. ± 1.96 for N(0,1), shown in red) for small samples and (ii) increasing the calculation time.

It is found, that the (sorted) repeated examples in the bootstrap sample of 400 usefully assists the quantile regression fit of the sample quantile distribution with a cubic polynomial, by creating finite scatter in the middle quantiles.

The cost of this extrapolation, is that the bootstrap dataset needs to be sorted to assign percentile values but this extrapolation is only necessary for small samples and extreme quantiles so is a reasonable burden. The alternative is bootstrap samples and performing 200 bootstraps samples imposes its on burden on calculation time.

Note that for small samples, this extrapolation does not guarantee the true population 2.5th & 97.5th

quantiles are reached or exceeded for each sample but brings the number of samples failing in this regard closer to the 5% expected with 95% confidence interval performance.

Because this estimation problem is one of order statistics rather than the relative contribution of two domains, this continuity correction for extreme quantiles is explicitly performed, rather than the Bayesian approach of nominally adding one success/one failure used by Agresti and Coull (3) for the similarly bounded case of variance estimates of proportions (near 0 & 1).

Numerical calculations of CIs for unweighted sample quantiles

Based on the above results, the algorithm for using the quadratic polynomial proxy as the quantile variance estimator (1) is now modified to the prescription

1. Use the quantile estimating function (or quasi-normalised version) in the original measurement scale to determine X distribution values for each given θ. In practice, this can be done using the "quantreg" r package (7) to perform quantile regression of the X distribution against an "intercept only" model.
2. Use the given sample size and quantile point θ to determine the quadratic polynomial proxy variance estimate or confidence interval in the percentile scale, using equations (8-12) for evd based CI estimators, according to choice and sample size.
3. Use the quantile estimating function (or quasi-normalised version) in the original measurement scale to determine X distribution values for each given θ value of the variance estimate or confidence interval obtained in step 2.
4. For extreme quantiles and small samples, to improve coverage, it may be required to create a bootstrap sample of 400 and perform quantile regression to obtain extrapolated estimates of the 0th & 100th quantiles. The ecdf_int percentile scale is assigned to a sorted bootstrap sample of 400 and a cubic polynomial quantile regression fit for the 0th & 100th quantiles respectively, is conducted. Using the two models, extrapolated values of the 0th and 100th quantiles are obtained and used in step 3 as approximate endpoints of the data distribution.

The bootstrap quantile variance estimates will be obtained by using the "quantreg" r package results in step 1 and calling the summary.boot option.

In Table 2 & 3, the coverage performance of backtransformed quadratic polynomial proxy (empirical) variance estimates of 95% confidence intervals of quantiles for unweighted samples are compared to bootstrap quantile confidence intervals for the N(0,1) and uniform(0,1) distributions for three sample sizes 50, 100, 1000. The number of repeated samples is 5000 to ensure convergence and the estimators whose coverage is compared are

(i) quantile regression bootstrap estimation,
(ii) binomial variance distribution (with extrapolation of 0th & 100th quantiles),
(iii) empirical variance distribution,
(iv) empirical variance distribution (with extrapolation of 0th & 100th quantiles)
(v) empirical variance distribution plus $Var(E[\theta|x])$ approximation (with extrapolation of 0th & 100th quantiles)

Table 2: CI coverage from 5000 repeated samples of unweighted N(0,1) samples

Sample size	quantile	bootstrap dist	binomial dist	evd dist	evd dist (extrapol)	total variance dist
50	0.025	0.8482	0.8422*	0.6894	0.7958*	0.8144*
100	0.025	0.9210	0.9490*	0.8874	0.9148*	0.9218*

Sample size	quantile	bootstrap dist	binomial dist	evd dist	evd dist (extrapol)	total variance dist
1000	0.025	0.9260	0.9386	0.9466	0.9466	0.9466
50	0.1	0.9174	0.9458*	0.9708	0.9724*	0.9732*
100	0.1	0.9206	0.9246	0.9394	0.9394	0.9394
1000	0.1	0.9402	0.9484	0.9478	0.9478	0.9478
50	0.5	0.9454	0.9538	0.9364	0.9364	0.9364
100	0.5	0.9488	0.9460	0.9568	0.9460	0.9460
1000	0.5	0.9484	0.9530	0.9536	0.9536	0.9486
50	0.9	0.9134	0.9438*	0.9438	0.9438*	0.9736*
100	0.9	0.9306	0.9392	0.9256	0.9392	0.9392
1000	0.9	0.9400	0.9426	0.9434	0.9434	0.9434
50	0.975	0.8410	0.8510*	0.6800	0.7940*	0.8216*
100	0.975	0.9108	0.8822*	0.9438	0.9068*	0.9162*
1000	0.975	0.9298	0.9454	0.9454	0.9454	0.9454

$*$ indicates extrapolation of data used near $\theta = 0, 1$

Table 3: CI coverage from 5000 repeated samples of unweighted uniform(0,1) samples

Sample size	quantile	bootstrap dist	binomial dist	evd dist	evd dist (extrapol)	total variance dist
50	0.025	0.9154	0.9032*	0.6732	0.8790*	0.9032*
100	0.025	0.9386	0.9372*	0.8844	0.9226*	0.9372*
1000	0.025	0.9300	0.9460	0.9460	0.9460	0.9460
50	0.1	0.9184	0.9316*	0.9676	0.9718*	0.9720*
100	0.1	0.9242	0.9344	0.9344	0.9344	0.9344
1000	0.1	0.9404	0.9540	0.9540	0.9540	0.9540
50	0.5	0.9236	0.9486	0.9316	0.9316	0.9316
100	0.5	0.9348	0.9512	0.9402	0.9402	0.9402
1000	0.5	0.9478	0.9554	0.9554	0.9554	0.9512
50	0.9	0.9360	0.9442*	0.9442	0.9442*	0.9724*
100	0.9	0.9344	0.9204	0.9342	0.9342	0.9342
1000	0.9	0.9406	0.9488	0.9510	0.9510	0.9510
50	0.975	0.9250	0.9362*	0.6822	0.8698*	0.9062*
100	0.975	0.9382	0.9604*	0.8814	0.9178*	0.9362*
1000	0.975	0.9386	0.9450	0.9544	0.9544	0.9544

$*$ indicates extrapolation of data used near $\theta = 0, 1$

It can seen that the performance of all five estimators is similar for sample quantiles in the range (0.1,0.9). Using the default quantile regression bootstrap settings for the two examples, the coverage performance in this range is ~0.92-0.95. Changing the default number of bootstrap samples (200) and/or using other bootstrap method choices may vary this performance. In the same range the evd based estimators have coverage 0.92-0.97, with the total variance estimator (v) having the most conservative coverage performance but they are all nominally performing at 95% coverage.

For the extreme quantiles near 0 & 1 (0.025 & 0.975 in the two tables), the coverage is higher for the uniform distribution than for the normal distribution. It can also be seen that the use of the quantile regression extrapolated data for the 0th & 100th quantiles has improved the evd based estimator coverage. In this range, the bootstrap estimates behaved well for samples of 100 & 1000 but starts to drop below nominal performance for n=50. The binomial variance distribution (with extrapolation) has the next best coverage, followed by the total variance CI estimator with the evd CI estimator (without extrapolation) having the weakest coverage.

In Appendix A, there are example histograms of the extrapolation produced by the (cubic polynomial) quantile regression fit of a bootstrap sample of 400 compared to (i) if no extrapolation of the observed data is conducted for the evd estimator of the 2.5th quantile and (ii) the standard quantile regression bootstrap estimates. The case given is for the uniform (0,1) distribution, where it is very obvious the evd estimator (without extrapolation) for sample size n=50, struggles to get the lower bound $\to 0$, whereas the standard quantile regression bootstrap CI estimator and the evd estimator with an extrapolated 0th quantile, have similar estimates that the lower bound ~0.

In Appendix B, the coverage performance of a finer range of quantile values near 0 is calculated and graphed, to investigate if the discreteness of the coverage for estimates of proportions (4) is replicated by the similarly bounded case of estimate of quantiles. From the results, the coverage performance for quantiles appear to be much smoother than for proportions (4).

Applying the evd method to the discrete Poisson distribution also results in nominal 95% coverage performance of the quantile confidence intervals in the quantile range (0.1,0.9).

Table 4: CI coverage from 5000 repeated samples of unweighted Poisson(600) samples

Sample size	quantile	bootstrap dist	binomial dist	evd dist	evd dist (extrapol)	total variance dist
50	0.025	0.8300	0.8176*	0.6924	0.7952*	0.8176*
50	0.1	0.9354	0.9434*	0.9752	0.9760*	0.9770*
50	0.5	0.9314	0.9626	0.9498	0.9498	0.9498
50	0.9	0.9162	0.9560*	0.9560	0.9560*	0.9786*
50	0.975	0.8406	0.8518*	0.7134	0.7968*	0.8198*

$*$ indicates extrapolation of data used near $\theta = 0, 1$

In Appendix A, the r code used for the coverage investigations is described and available on Github. To save run time, the example code has been written to only execute calculations for the 2.5th quantile of N(0,1) with a small number of repeated samples (500). There are options in the code to do additional quantiles, repeated samples of larger size and insert other distributions to be sampled.

Conclusions

Empirical variance distribution based confidence interval estimators, calculated in the percentile scale, have been found to exhibit nominal 95% coverage performance in the quantile range 0.1-0.9 similar to bootstrap estimates.

For extreme quantiles and small samples, coverage performance of the evd estimators are improved significantly, by quantile regression extrapolated estimates of the endpoints of sample distributions.

The good performance of empirical variance distribution based confidence interval estimators for the discrete Poisson distribution indicates that the approach may also be applicable to unequally weighted samples.

References

1. Martin J.P.D., 2015, http://figshare.com/articles/An_empirical_variance_distribution_approximation_using_backtransformation_from_the_percentile_scale_for_the_quantile_estimating_function_of_continuous_distributions/1566828
2. Koencker, R. W. & Bassett G., Econometrica, 1978, vol. 46, issue 1, pages 33-50
3. https://en.wikipedia.org/wiki/Law_of_total_variance
4. Agresti, A. and Coull, B. A. 1998 The American Statistician, vol. 52, p119-126. doi:10.2307/2685469. JSTOR 2685469
5. https://en.wikipedia.org/wiki/Binomial_distribution
6. Portnoy S., 2014 Journal Computational Statistics & Data Analysis, vol. 72, p273-281
7. Koencker, R. W., Portnoy S. et al, https://cran.r-project.org/web/packages/quantreg/quantreg.pdf

Appendix A - R code for coverage performance assessment

The r code contained in the Rmd version of this paper, calculates coverage performance for the five estimators in Tables 2-4. The current settings are set to calculate the value for one value, the 2.5th quantile and 500 repeated samples so that the program runs ~12 minutes on a dualcore 4GB windows based PC. However, the proper full run would be to do a loop over several quantile values and use 5000 repeated samples on a more powerful PC.

As shown below, the program outputs

(i) incremental coverage performance as the repeated samples grow
(ii) the percentile scale confidence intervals which only depend on the sample size and quantile point of interest,
(iii) the mean of the quantile regression estimate for the quantile point of the population,
(iv) the quantile point of the population,
(v) the lower and upper bounds of the confidence intervals (also as histograms for some of the estimators)

The code consists of two nested loops.

The outer loop sets up the sample distribution, the vector of quantiles to be investigated and initialises the vectors which will contain repeated sample coverage estimates for the CI estimators.

The inner loop does a repeat of the coverage calculations for each selected element of the vector of quantiles to be investigated.

The r code currently has six manual selection points indicated by comments.

1. The vector of quantiles of interest is defined. The sample size is entered, and "lambda" a variable used for standard error/mean of the distribution (normal, uniform / poisson) is given a value.
2. The inner for loop, is defined, to choose a subset or all of the vector of quantiles of interest.
3. The distribution being repeatedly sampled is defined, eg. rnorm(sampsize,0,lambda).

4. The quantile function for the above distribution is defined in order to extract the true population parameter that the repeated samples are attempting to estimate, eg. qnorm(taus,0,lambda).

5. The number of repeated samples is set in the inner "for loop"", currently 500 in the code below. The distribution of interest is defined a second time (inside the inner loop) to ensure new random samples are selected each increment of the inner loop. As mentioned 5000 repeated samples is a preferred conservative setting to ensure final convergence.

6. There is code currently "commented out", to save the datasets created to file.

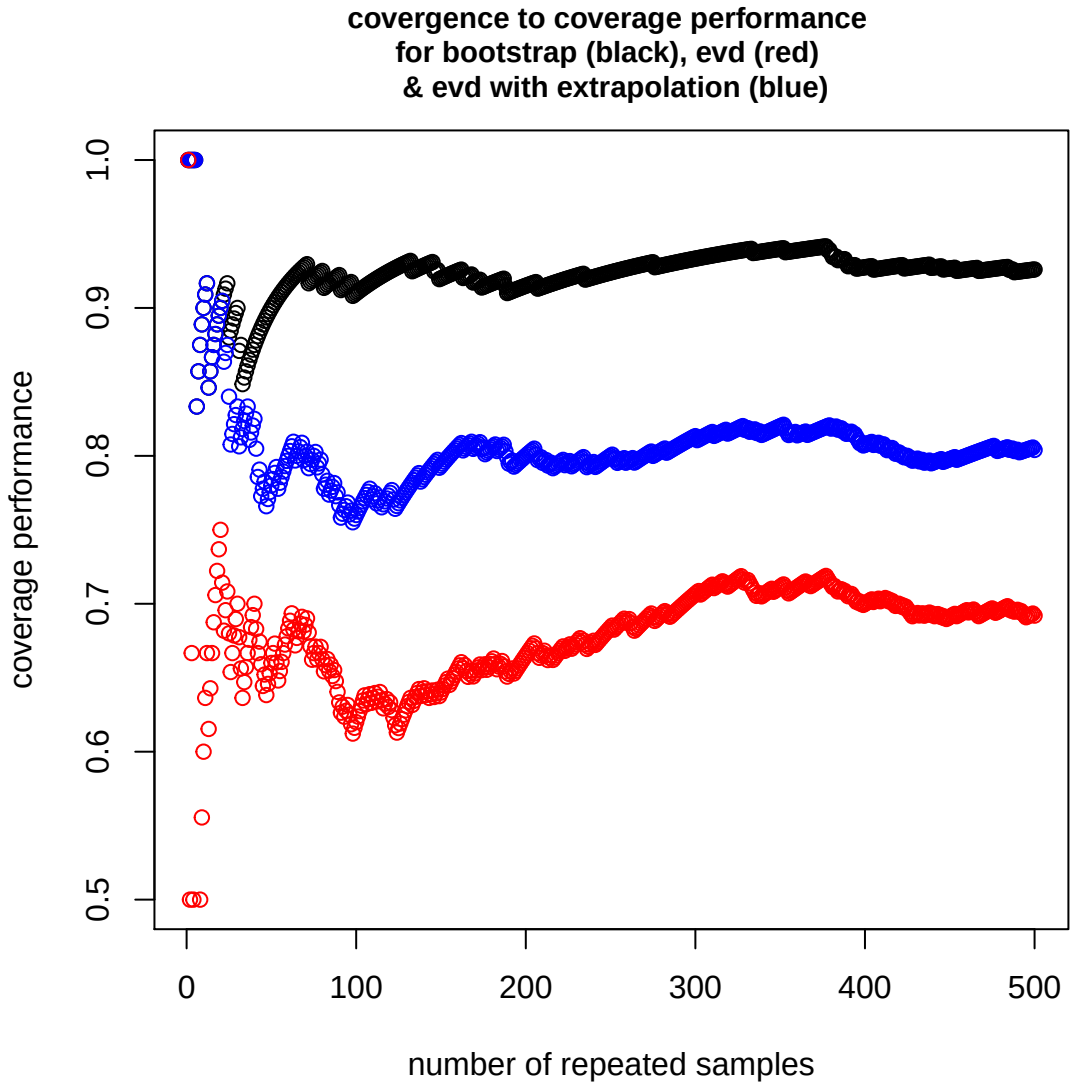
Important notes;

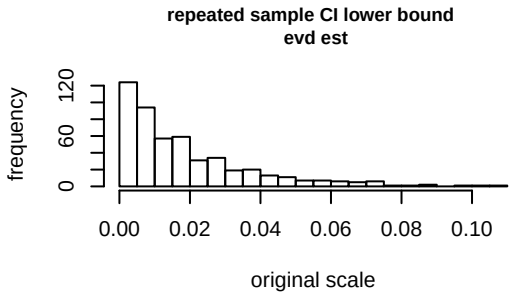
Two of the outputs printed to screen are the mean of the quantile regression estimates and the true population parameter. It has been observed that while quantile regression has non-unique solutions to the minimum, the "quantreg" package tends to regularly output the nearest non-unique solution to the median quantile. This behaviour means that the mean of the repeated estimates does not exactly equal the true population parameter, eg. -1.93 compared to -1.96 for the 2.5th quantile of the normal distribution. The difference in values with the true population parameter will be related to 1/2n where the length of non-unique solutions may be ~1/n. This difference is much less than the variance and CIs calculated by the quadratic polynomial proxy for a given sample size, which enables the quadratic polynomial proxy to be a good variance approximation of the quantile regression estimates.

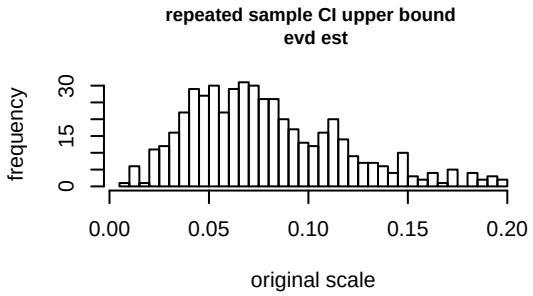
With the cubic polynomial quantile regression fit code used for extrapolated estimates of the 0th & 100th quantiles for extreme quantiles of small samples using "quantreg" package, the choice of solution algorithm must be "br". If "fn" is used for such small samples, the program will throw an error and stop.

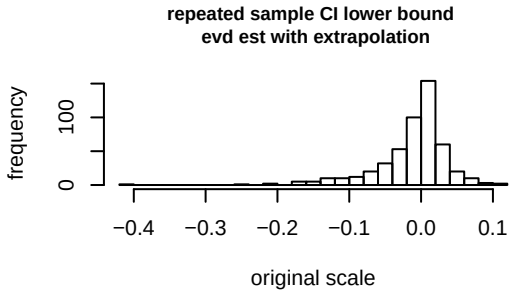
The percentile scale has been subdivided into 10000 points, regardless of the repeated sample size, in order to calculate the density function values. So this setting may need some adjustment if samples > 1000 are run with the code. Also this coarseness may be affecting the numerical precision of the existing results in tables 2-4.

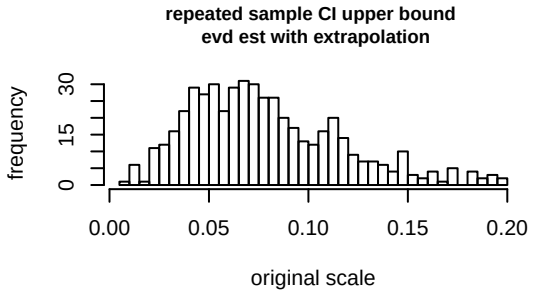
```
## [1] "binomial CI | evd CI | total var CI , in percentile scale"
## [1] "lower bounds"
## [1] "0 0.01 0"
## [1] "upper bounds"
## [1] "0.08 0.0658655865586559 0.0731778422964305"
## [1] ""
## [1] "coverage performance and mean quantile regression estimate"
## [1] "q pt | bootstrap | binomial | evd | evd with ext | total var | quantreg est | pop value"
## [1] "0.025 0.926 0.858 0.692 0.804 0.858 0.0386715354328044 0.025"
```

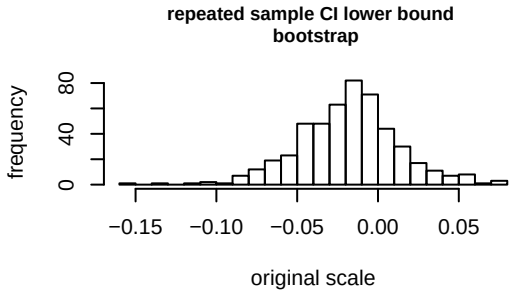

covergence to coverage performance
for bootstrap (black), evd (red)
& evd with extrapolation (blue)
coverage performance
0.5
0.6
0.7
0.8
0.9
1.0
0
100
200
300
400
500
number of repeated samples

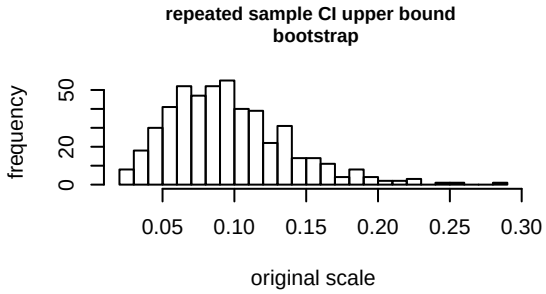

repeated sample CI lower bound
evd est
frequency
0
60
120
0.00
0.02
0.04
0.06
0.08
0.10
original scale


repeated sample CI upper bound
evd est
frequency
0
15
30
0.00
0.05
0.10
0.15
0.20
original scale


repeated sample CI lower bound
evd est with extrapolation
frequency
0
100
−0.4
−0.3
−0.2
−0.1
0.0
0.1
original scale


repeated sample CI upper bound
evd est with extrapolation
frequency
0
15
30
0.00
0.05
0.10
0.15
0.20
original scale


repeated sample CI lower bound
bootstrap
frequency
0
40
80
−0.15
−0.10
−0.05
0.00
0.05
original scale


repeated sample CI upper bound
bootstrap
frequency
0
20
50
0.05
0.10
0.15
0.20
0.25
0.30
original scale

Appendix B - Coverage performance plot for N(0,1)