effect of sample size and repeated values for jackknife CI estimates of unweighted median

John P.D. Martin

Thursday, February 12, 2015

Executive Summary

A comparison of bootstrap variance and jacknife variance estimates for different sample sizes is presented, including examples of continuous and discrete distributions. The jackknife variance estimates are calculated in the percentile scale and then backtransformed to the original measurement scale.

Useful agreement between the jackknife and bootstrap confidence interval estimates is observed across the spectrum of sample sizes.

Introduction

For the median (pth-quantile) distribution, the ordered distribution in the percentile frame is linear. As such for continuous distributions, it has been shown (1) by numerical examples of symmetric and skewed distributions and algebraically for the asymptotic limit of the normal distribtuion; that conducting jackknife calculations in this reference frame, followed by backtransformation to the measurement scale using the (empirical) cumulative distribution function (CDF) produces variance estimates in close agreement with bootstrap variance estimates.

Each data point in the ordered percentile scale distribution of the median has the following cdf values, (using a basic median definition)

1/n , 2/n , 3/n , ... , n/n

In the drop one unit jackknife variance approach for such a distribution (ignoring continuity corrections), the (n-1) subsampled cdf values assigned to each ordered point are

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

The estimated first order jackknife variance in this reference frame is

$$(n-1)\cdot\frac{1}{n}\sum(jk\_est-jk\_mean)^2\approx\frac{1}{4n}\cdot\frac{n-1}{n} \qquad (1)$$

The contribution of the observed sample distribution is then handled, non-parametrically, by the backtransformation of the jackknife results to the original measurement scale using the empirical cdf.

For discrete distributions, there are repeated values in the observations. Therefore, to use the above transformed jackknife estimator approach for discrete distributions, requires each data point (in the ordered percentile scale and the algorithm program) to be assigned a unique ordered position number. The step function empirical cdf that is generated from the original observations is then used to perform the backtransformation of the jackknife results.

In this paper, the performance of the jackknife confidence interval estimate for unweighted samples, using this approach is demonstrated (i) across a range of samples sizes, (ii) the algorithm is extended to the case of discrete distributions and (iii) confidence interval corrections for bias (for n odd) and t distribution (for small sample sizes) are introduced.

Calculations and estimators

In the original paper (1), relatively large sample sizes were considered and the jackknife confidence interval was considered to follow the normal approximation. In this paper, (i) three versions of jackknife estimator are investigated

1. normal approximation

$$jk\_normal\_app = jk\_mean \pm 1.96 \cdot \sqrt{jk\_var} \tag{2}$$

2. bias corrected (n odd)

$$jk\_bias\_corr = 0.5 \pm 1.96 \cdot \sqrt{jk\_var} \tag{3}$$

3. bias corrected (n odd) and t distribution for 2.5th/97.5th percentile

$$jk\_bias\_tdist\_corr = 0.5 \pm tdist(0.95, df = (n-1)) \cdot \sqrt{jk\_var} \tag{4}$$

and (ii) the algorithm code is amended to deal with discrete distributions.

In particular, the algorithm for discrete distributions has the steps

1. the unweighted data observations are ordered by the sorting variable (univariate case)
2. ordered percentile values are assigned to each data point (this is the mapping of empirical cdf to observed values which is later used in step 6)
3. the sample median estimate is obtained as the ordered data point with quantile >= 0.5
4. In the drop one unit jackknife estimation loop, calculated in the ordered percentile frame and hence using the linear distribution of percentile values, the drop one unit percentile estimate of the sample median data point (the data point with the same order position as determined in step 3) is obtained and stored (following the Woodruff method (2))
5. calculations of the confidence interval, variance and mean of the jackknife estimates are conducted and scaled using standard formula
6. the empirical CDF (determined in step 2) is then used to backtransform the jackknife confidence interval values to the original measurement scale

Note that in the examples investigated, where Poisson distributions of mean 40, 60, 80 are used, the t distribution may be considered applicable as a confidence interval correction as the Poisson distributions approximates to normal distribution for such large means.

Results

For comparison, the improved jackknife confidence interval estimates are plotted with bootstrap interval and BCa estimates, for three cases

1. Normal distribution N(60,sqrt(60))

to examine the behaviour of the confidence intervals with sample size

2. Poisson distribution Pois(60)

to examine the effect of discrete distributions

3. bivariate Poisson distributions Pois(40)+Pois(80)

to examine the effect of skewness in the distribution

For completeness, the three cases are presented by graphs of (i) the large sample distribution, (ii) and example small sample distribution and (iii) the confidence interval estimates of the unweighted sample median by sample size.

The graphs are essentially convergence studies, as for a fixed random generating seed, the sample size is increased from 5 to 3600.

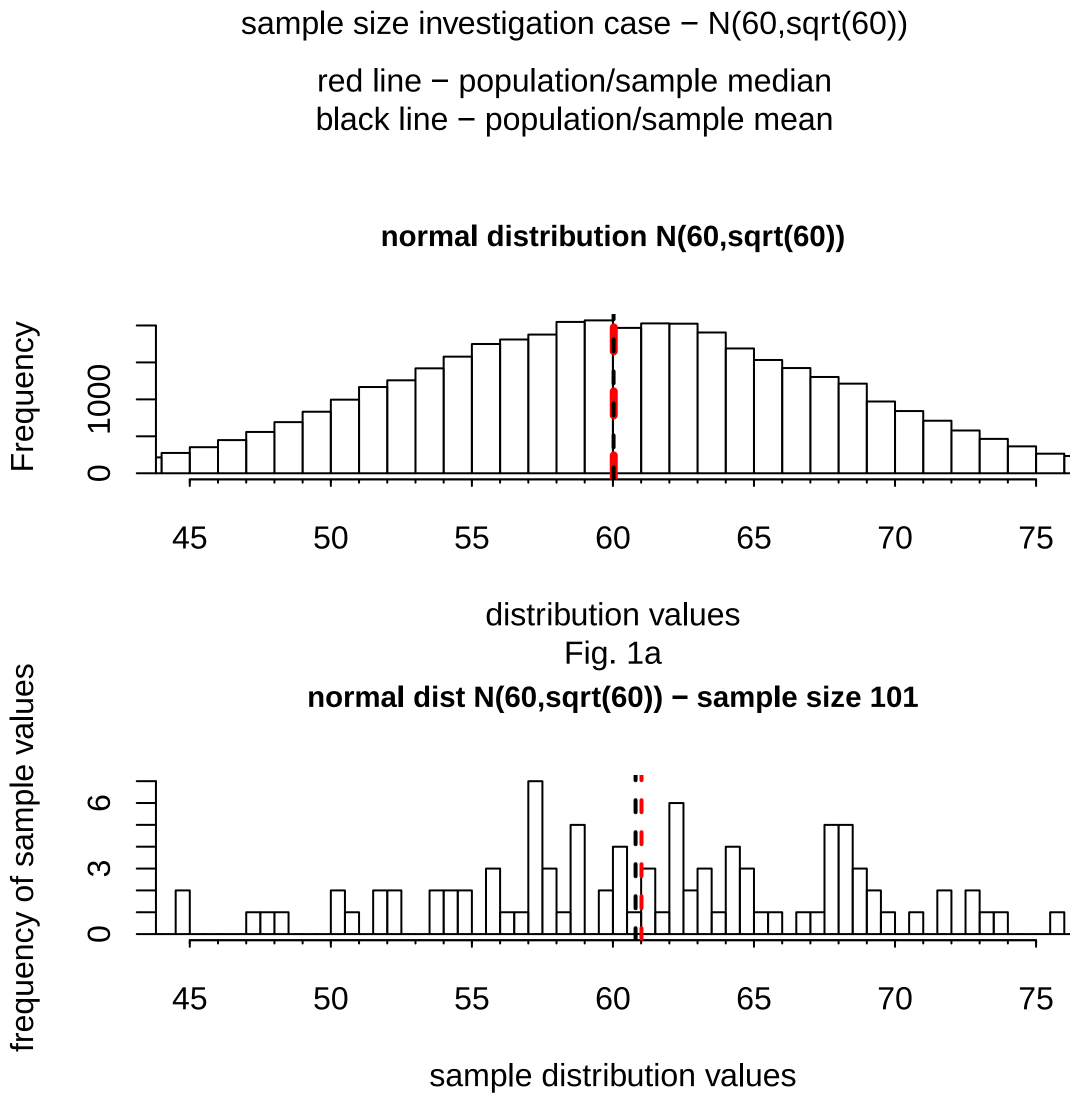

Fig. 1a

Fig. 1b

In Figure 2, for n > 20 there is good agreement between the jackknife and bootstrap confiodence interval estimators. For n < 20 there are more noticeable differences across the various estimators. The deviation of the bootstrap Bca estimates is caused by algorithm issues for the sample data causing default intervals to be output rather than optimal solution. For this continuous distribution it can be seen that the bootstrap mean and the sample median are in close agreement.

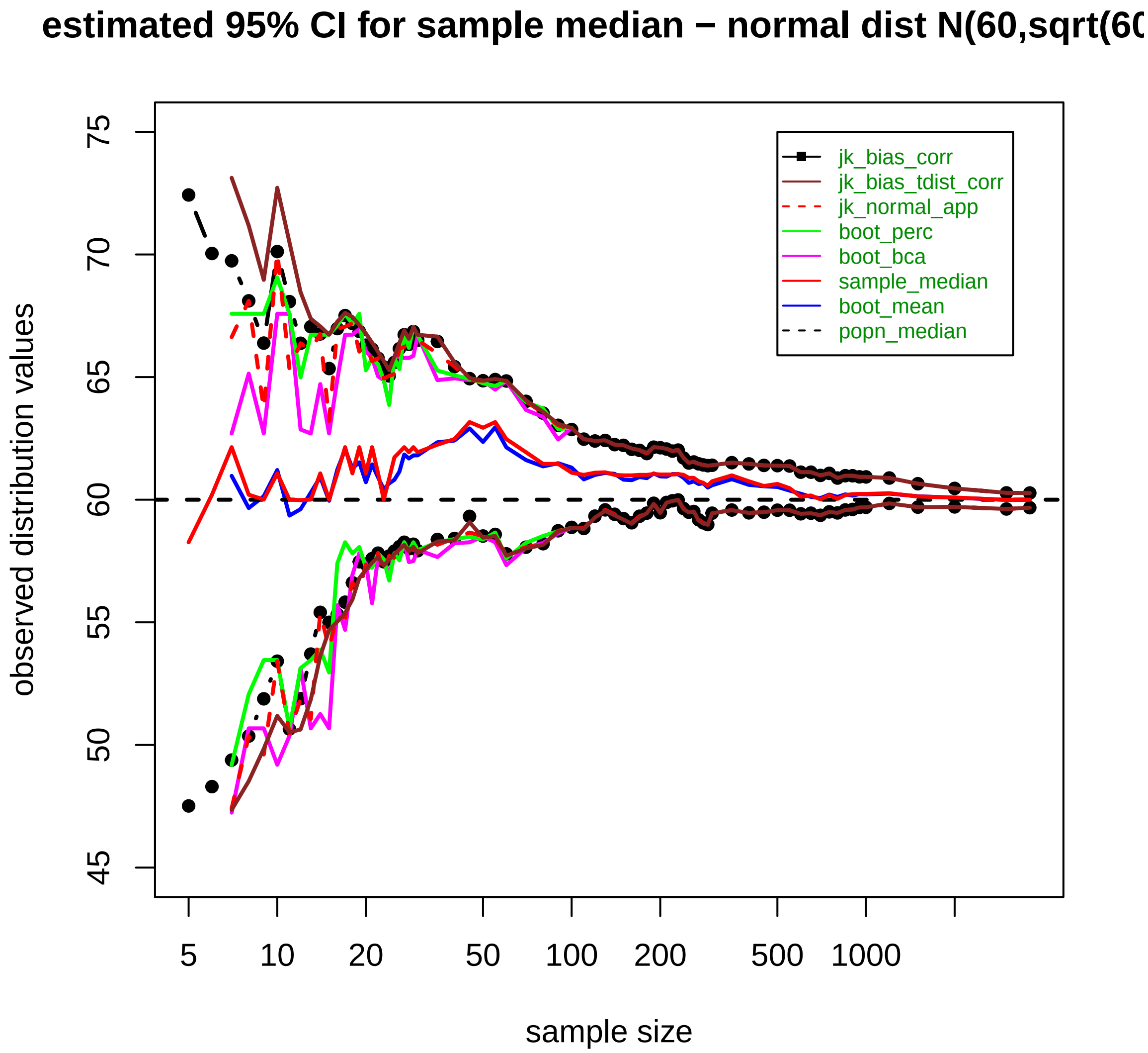
estimated 95% CI for sample median − normal dist N(60,sqrt(6(
observed distribution values
45
50
55
60
65
70
75
5
10
20
50
100
200
500
1000
sample size
jk_bias_corr
jk_bias_tdist_corr
jk_normal_app
boot_perc
boot_bca
sample_median
boot_mean
popn_median

Fig 2

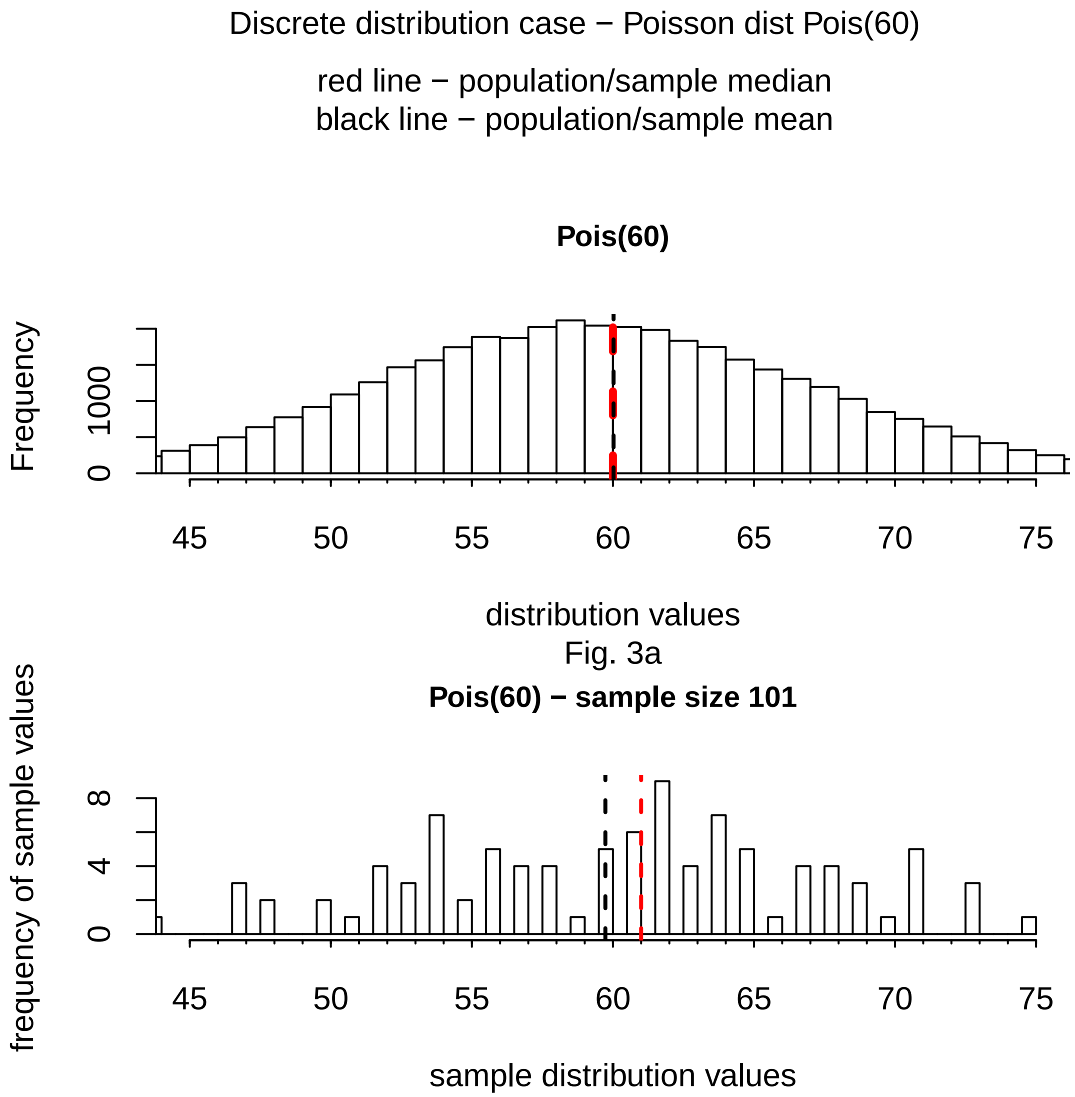

Fig. 3a

Fig. 3b

Looking at the case of a symmetric discrete distribution, in Figure 4, for n > 100 there is good agreement between the jackknife and bootstrap confiodence interval estimators. For, 20 < n < 100 there are some differences across the various estimators with the bootstrap estimator occassionally be smaller by one unit (sensitive to the discrete nature of the distribution) while a couple are issues with robust Bca solutions. Below n = 20, there are slightly wider differences across the estimators compared to the continuous case fig 2). For this discrete distribution example it can be seen that the bootstrap mean is similar but not unbiassed with respect to the sample median.

Although the population is a symmetric distribution, each repeated sampling experiment like figure 4 would be expected to display some asymmetry until convergence is achieved.

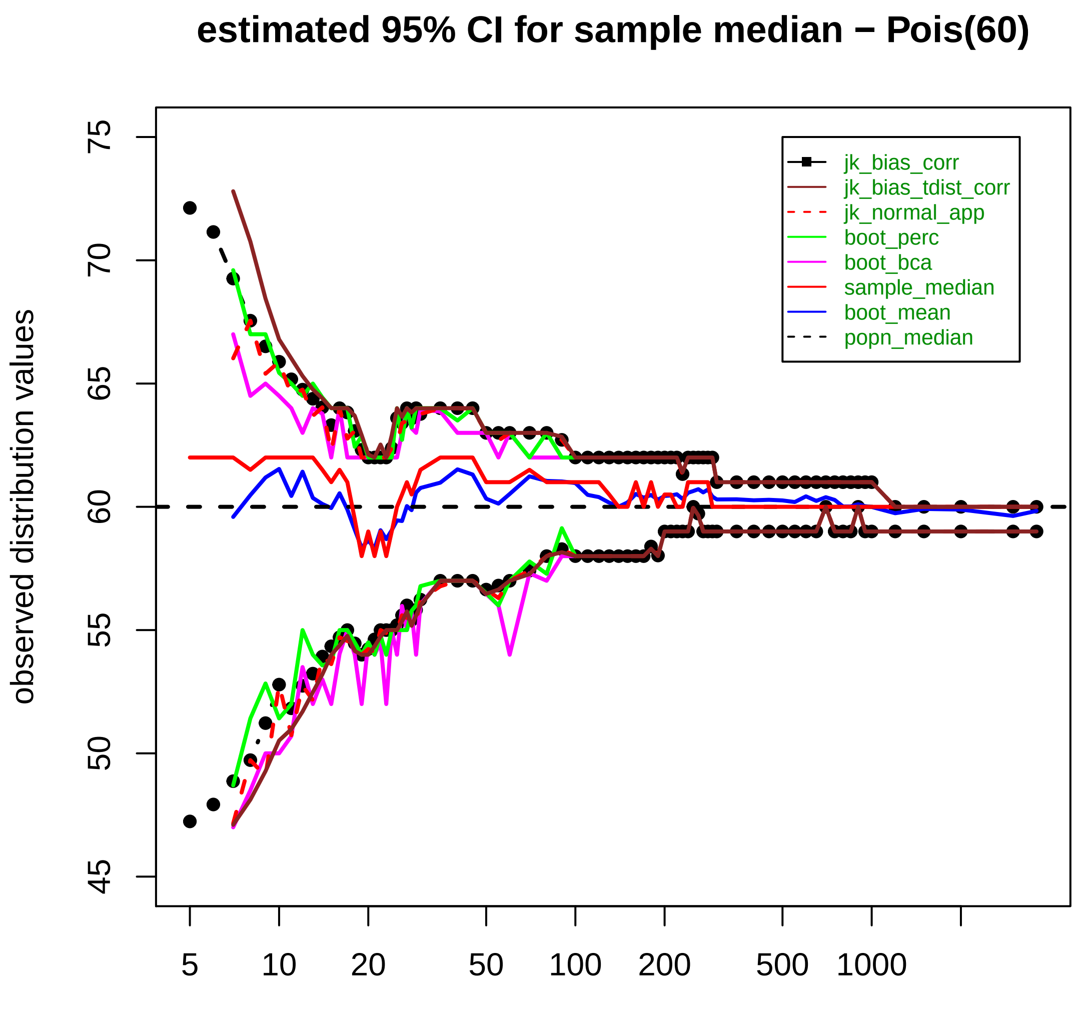
estimated 95% CI for sample median − Pois(60)
observed distribution values
45
50
55
60
65
70
75
5
10
20
50
100
200
500
1000
sample size
jk_bias_corr
jk_bias_tdist_corr
jk_normal_app
boot_perc
boot_bca
sample_median
boot_mean
popn_median

Fig 4

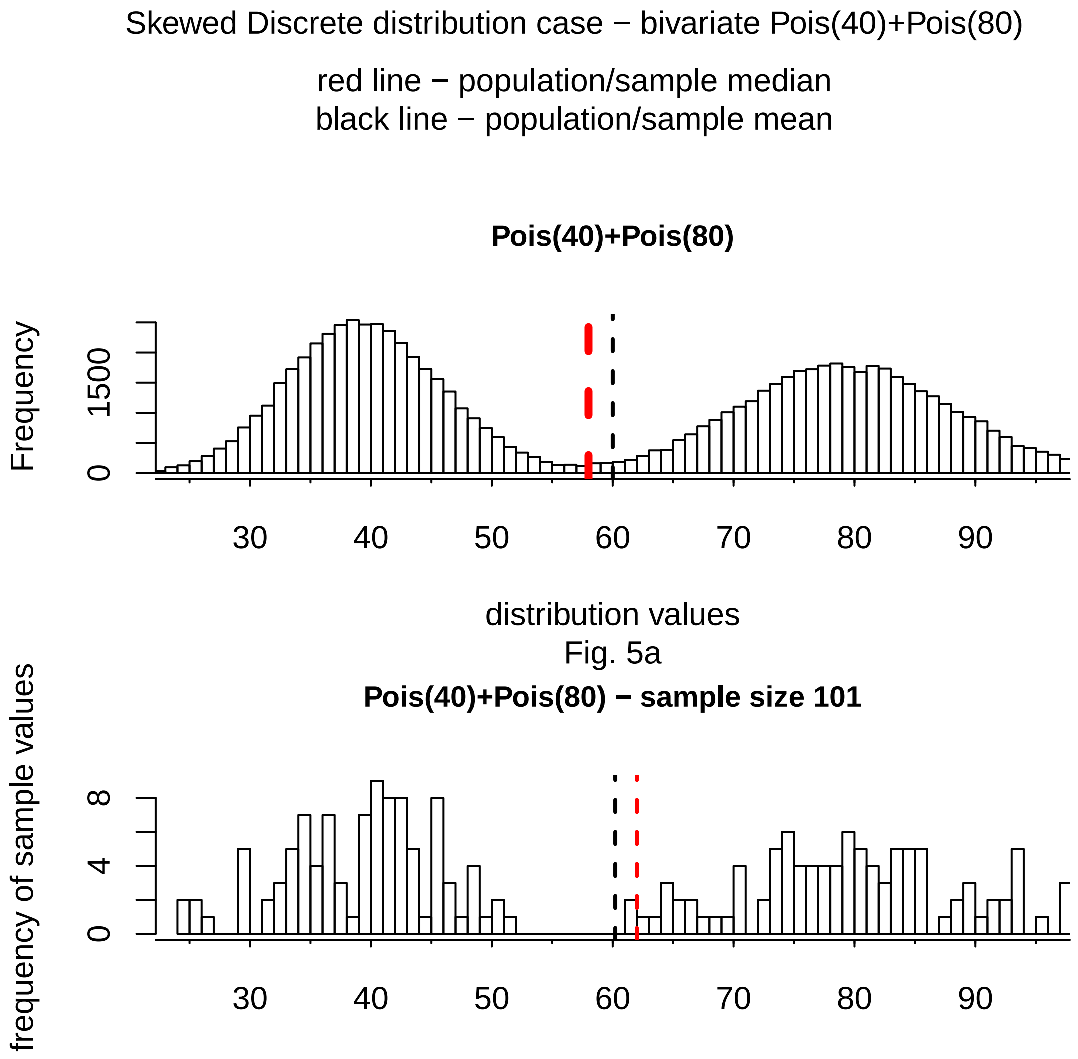

Fig. 5a

Fig. 5b

Looking at the case of a skewed distribution with discrete components, given in Figure 6 by a bivariate Poisson distribution, for n > 20 there is good agreement between the jackknife and bootstrap median confidence interval estimates. Below n = 20, there are slightly wider differences across the estimators.

For this skewed discrete distribution example it can be seen that the bootstrap mean converges more quickly to the population median (=57) than does the sample median.

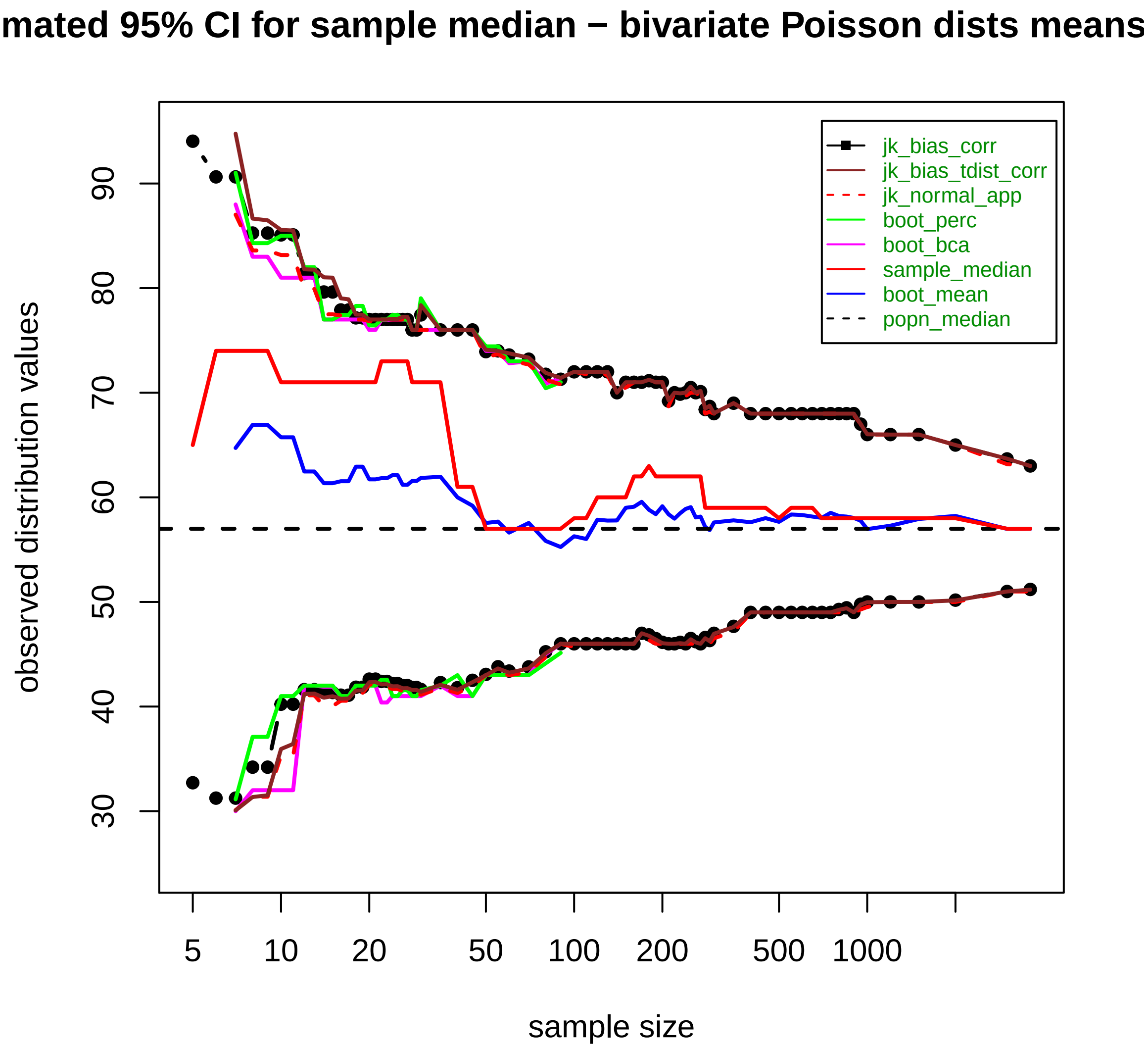

Fig 6

References

1. Martin J.P.D. (2015) https://github.com/johnpdmartin/sampling-investigations/blob/master/jackknife_for_unweighted_median_with_normal_dist_proof.pdf

2. John W. Rogers (2003), Estimating the variance of percentiles using replicate weights, 2003 Joint Statistical Meetings - Section on Survey Research Methods, p3525-3532, http://www.amstat.org/sections/SRMS/Proceedings/y2003/Files/JSM2003-000742.pdf