Improved jackknife estimates for median variance in equally weighted samples using percentile scale based calculations

John P. D. Martin

Thursday, February 5, 2015

Executive Summary

This paper demonstrates the improvement in accuracy of jackknife estimates of the classical confidence intervals of medians from equally weighted (or unweighted) samples using calculations based on the percentile scale rather than the original scale of the observations. The method used is similar to the Woodruff approach for weighted samples (1) and the improved estimates are consistent with the expectation that suitably transforming the reference frame can improve variance estimation accuracy (2).

As shown in Table 1, the jackknife estimates of median confidence intervals for unweighted samples conducted in the percentile scale, and then backtransformed to the original scale are found to be highly consistent with bootstrap estimates for several common distributions; uniform, normal, log-normal and skewed bivariate normal.

Table 1: Estimated median 95% confidence intervals of unweighted samples for different re-sampling methods

Distribution	popn median	bootstrap	original scale jackknife	percentile scale jackknife
random uniform(0,1)	0.5	(0.4612,0.5218)	(0.4402,0.5513)	(0.4612,0.5218)
normal N(0,1)	0	(-0.0993,0.0576)	(-0.08749,.02576)	(-.09972,.05772)
systematic uniform(0,1)	0.5	(0.4532,0.5494)	(0.4534,0.5492)	(0.4545,0.5505)
log-normal exp(N(0.1,1))	1.09	(1.026,1.18)	(0.9434,1.236)	(1.026,1.181)
bivariate normal	76.65	(74.7,77.73)	(75.98,77.75)	(74.71,77.74)

To highlight the need for a locally linear (smooth) quality to the unweighted sample distribution (2,3,4) in the region of the median point, in order for jackknife median variance calculations to match the performance of bootstrap median variance calculations. The third example in Table 1, is the highly artifical case of evenly spaced observations. In this rare linear example in the original scale, as expected, the regular jackknife calculation approach also closely matches the bootstrap confidence interval estimate.

Various aspects of transforming the median variance estimation calculation to the percentile scale, have been in long use by statistical agencies such as the Australian Bureau of Statistics and Westat for weighted data. From Rogers 2003 (1), in particular,"... the Woodruff method calculates a confidence interval for the sample amount below the estimated percentile. That confidence interval is then transformed to the measurement scale using the inverse of the sample quantile function".

This is basically the same approach as what has been performed in this paper. There have been descriptions of the benefits of transformation for both bootstrap and jackknife variances estimates (2,3) but what report (i) describes this good performance of the jackknife approach for estimating the median variance for the special case of unweighted (or equally weighted) samples or (ii) uses this special case as a limit for convergence of bootstrap resampling calculations.

Introduction

In a number of introductory and intermediate online data science and statistics courses the selection weights of the samples and datasets are not explicitly considered. This is often because simple random sampling (with experimental design randomisation) is the selection method and other concepts of the subject matter were rightly considered more important to highlight than the rigorous issues of probability sampling.

However, in several of the courses, the historical problems with the accuracy of jackknife variance estimation in contrast to bootstrap estimation did get mentioned. With my working knowledge in clustered survey sampling and the Woodruff method I really wondered whether the issue was the same magnitude for unweighted samples.

With the John Hopkins University, Data Science and Biostatistics online courses. The drop one unit jackknife variance estimation for median (5) was presented with the following R code

```
n <- length(gmVol)

theta <- median(gmVol)

jk <- sapply(1 : n,

function(i) median(gmVol[-i]) )

thetaBar <- mean(jk)

biasEst <- (n - 1) * (thetaBar - theta)

seEst <- sqrt((n - 1) * mean((jk - thetaBar)^2))
```

where gmVol is the dataset under consideration and the loop "sapply" with argument median(gmVol[-i]) drops one unit at a time and recalculates the median estimate as per the jackknife prescription. The distrubtion of the jackknife replicate estimates are then collated to estimate the mean, bias and standard deviation of the sample.

The Woodruff method for percentile variance estimation uses (i) the whole sample to estimate the percentile boundaries and then (ii) calculates the uncertainty in the amount of sample within each percentile boundary by subsampling or other methods. The above example code for jackknife variance estimation, in the (original) scale can be converted to the woodruff method with only a few modifications (for continuous distributions)

```
n <- length(gmVol)

theta <- median(gmVol)

jk <- sapply(1 : n,

function(i) {

Fn <- ecdf(gmVol[-i]); # empirical cumulative density function

Fn(theta) # calculates the percent of sample < percentile boundary theta for each jackknife subsample

} )

thetaBarinv <- mean(jkinv) # using inv to indicate the jackknife calculations are in percentile space

jkexpinv <- (jkinv-thetaBarinv)*sqrt(n)+thetaBarinv # a scaled version of jackknife estimate close to 1 sd

jkexpandinv <- quantile(samp,jkexpinv)

biasEstinv <- (n - 1) * (thetaBarinv - theta)

seEstinv <- sqrt((n - 1) * mean((jkinv - thetaBarinv)^2))

jackknife_median <- quantile(gmVol,0.5) # the sample median back again

jackknife_median <- quantile(gmVol,0.5+1.96*seEstinv) # 2.5th percentile jackknife in original scale
```

jackknife_median <- quantile(gmVol,0.5+1.96*seEstinv) # 97.5th percentile jackknife in original scale

As will be presented in this paper, the performance of this jackknife variance estimation (using the Woodruff method approach) for unweighted samples very closely matches bootstrap calculations for several continuous distributions. The issue of the linearity of the percentile scale distribution, for unweighted samples, is highlighted as the reason for the good performance of this version of the jackknife method.

Median variance estimation for unweighted samples

The issue with the jackknife estimator as a linearisation of the bootstrap method, is that it relies on the distribution of sample estimates of mean, median etc being asymptopically linear and smooth.

For estimates of means, as the sample size grows, the central limit theorem is applicable and the distribution of sample means with jackknife subsampling becomes increasingly smooth and locally linear.

For sampling estimates of median variance, the distribution of resampled (bootstrap) or subsampled (jacknife) estimates of median is generally asymmetric in the original scale (2) of observed distributions and the central limit theorem does not apply.

However, in the case of unweighted samples, it is also generally true that the percentile distribution (empirical cumulative density distribution (ecdf)) about the median value 0.5, is linear. This is shown in figure 1b&d, for a N(0,1) unweighted sample of size 50, where each unweighted observation contributes the same amount 1/n to the ecdf.

Note that there are some more complex versions of cdfs where the ends of sample distributions are given less weighting via continuity corrections, eg. empirical distribution function with averaging and interpolation (used by Australian Bureau of Statistics). Such modified cdfs, value the sample distribution end points more weakly as the true end points of population distributions are harder to accurately sample.

Performing the jackknife calculations for estimates of median variance for unweighted samples in the percentile scale has the advantage that the percentile distribution of unweighted samples is linear while the jackknife is a linearised bootstrap estimator. Hence, the jackknife estimator could in principle approach the performance of bootstrap calculations.

The standard error estimates about the median point (0.5) for unweighted samples in the percentile scale (ecdf) using delete one unit jackknife estimation are

1/sqrt(n)

and the 2.5th/97.5th percentiles

0.5 +/- 1.96/sqrt(n)

under the normal approximation consistent with comparing to standard CI bootstrap calculations (shown in the R code given in the introduction) and the sample sizes 270, 415 & 1000 used in this paper.

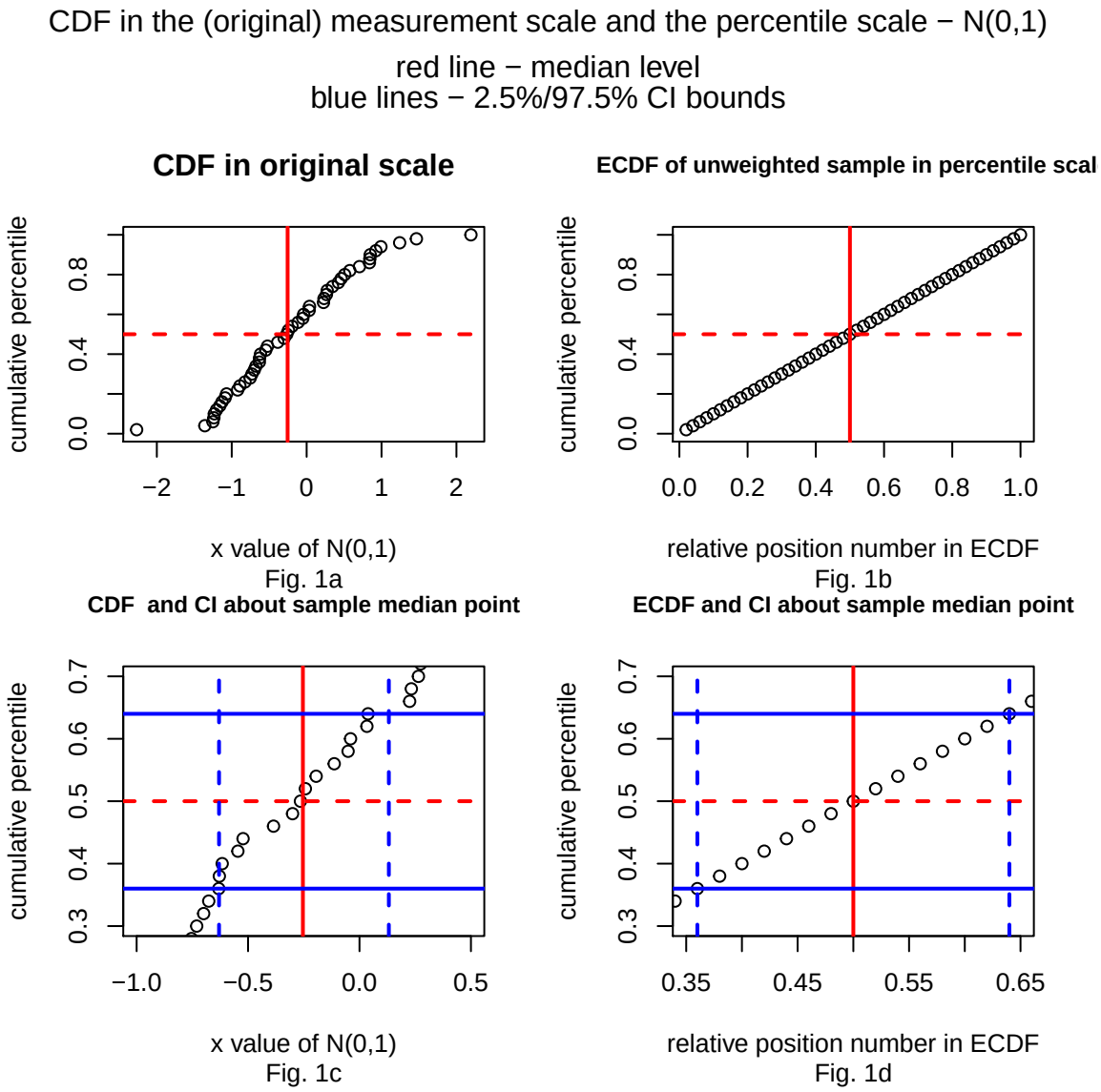


Figure 1

In Table 1 and the appendix, jackknife variance estimates of unweighted medians in the percentile scale are compared to standard jackknife calculations and classical bootstrap median confidence intervals (based on 10000 resamples) for five examples of continuous distributions.

Two common symmetric distributions

(i) uniform distribution unif(0,1) of sample 1000, and

(ii) the standard normal distribution N(0,1) of sample 1000

an artificial linear smooth symmetric distribution in the measurement scale

(iii) systematic unif(0,1) of random sample size 415 using start/skip selection

and two common skewed distributions

(iv) a log-normal distribution exp(N(0.1,1)) of sample 1000, and

(v) a continuous bivariate normal distribution 1/3(N(55,5.5))+2/3(N(80,6)) of sample 270. This example is a continuous distribution analogue of the old faithful geyser waiting duration data located as data(faithful) on the r library(datasets).

This range of datasets has different smoothness and nonlinearity of the cdf about the median point. As such, the confidence interval results for the regular jackknife calculation conducted in the (original) measurement scale, do not agree with the bootstrap calculation except for the artifical linear distribution example. Of course, for this artifical example (iii), the original scale cdf is linear and the jackknife calculation is expected to be accurate.

However, comparing the "jackknife variance estimates using the percentile scale"" to the bootstrap calculations there is good agreement in the confidence intervals of the unweighted medians across all five examples. The tranformation of the jackknife calculation to the linear ecdf scale allows the jackknife method to consider the full characteristic of the sample distribution and the backtransformation from the ecdf to the original measurement scale handles the nonlinearity in the observed data distribution.

Table 1: Estimated median 95% confidence intervals of unweighted samples for different resampling methods

Distribution	popn median	bootstrap	original scale jackknife	percentile scale jackknife
random uniform(0,1)	0.5	(0.4612,0.5218)	(0.4402,0.5513)	(0.4612,0.5218)
normal N(0,1)	0	(-0.0993,0.0576)	(-0.08749,.02576)	(-.09972,.05772)
systematic uniform(0,1)	0.5	(0.4532,0.5494)	(0.4534,0.5492)	(0.4545,0.5505)
log-normal exp(N(0.1,1))	1.09	(1.026,1.18)	(0.9434,1.236)	(1.026,1.181)
bivariate normal	76.65	(74.7,77.73)	(75.98,77.75)	(74.71,77.74)

Conclusions

The transformation of the jackknife calculation to the percentile scale for the special case of median estimates of unweighted samples, allows jackknife variance estimates to be consistent with median confidence interval estimates using first-order bootstrap calculations.

This special case provides (i) a limiting value for bootstrap calculations as the number of resamples approaches infinity, (ii) supports jackknife estimation as a linearisation of classical bootstrap estimation and (iii) supports the paradigm that SRSWR bootstrap resampling is not somehow modifying the information given by the collected distribution since the jackknife estimation is using subsampling which is more clearly thought not to modify the information given by the sample.

The issue of the accuracy of the Woodruff method for unequally weighted data (complex surveys) using replicate weighting and jackknife in the percentile scale, should be assessed by investigation of the linearity and smoothness of the cdf about the estimation point (mean, median etc).

References

1. 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
2. Efron, B. (1979) Bootstrap Methods: Another look at the Jackknife, The Annals of Statistics 7, pp1-26
3. Efron, B. (2003) Second thoughts on the bootstrap, Statistical Science 18, pp135-140
4. Miller, R.G. (1974) The Jackknife–a review, Biometrika 61, pp1-15

5. Caffo, B. (2007) Mathematical Biostatistics Boot Camp Lecture 12 on bootstrapping and resampling, http://www.biostat.jhsph.edu/~bcaffo/651/files/lecture12.pdf

Appendix - Results and calculations of bootstrap and jackknife resampling for several continuous distributions Using continuous distribution examples, allows the empirical cumulative density function (ecdf) to be as smooth as feasible allowing for a clear demonstration of percentile based jacknkife calculations.

With drop one unit jackknife median estimation it is well known that the median estimates of jackknife subsamples have only two values. This is because the full sample median point will only move one ordered place as the number of data points in the subsamples is (n-1) compared to the full sample n. Hence in the graphs of jackknife median estimates shown below, you will only see two peaks compared to the multiple discrete distribution of bootstrap median resample estimates.

Figure 2

```
## Warning: bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.4667,  0.5277 )   ( 0.4697,  0.5303 )
##
## Level     Percentile            BCa
## 95%   ( 0.4612,  0.5218 )   ( 0.4605,  0.5214 )
## Calculations and Intervals on Original Scale

## [1] "mean of bootstrap samples"

## [1] 0.4943

## [1] "standard error of bootstrap samples"

## [1] 0.01556
```

Symmetric distribution case 1 − unif(0,1)

red line − population median, black line − sample median
blue lines − 2.5%/97.5% percentile bounds

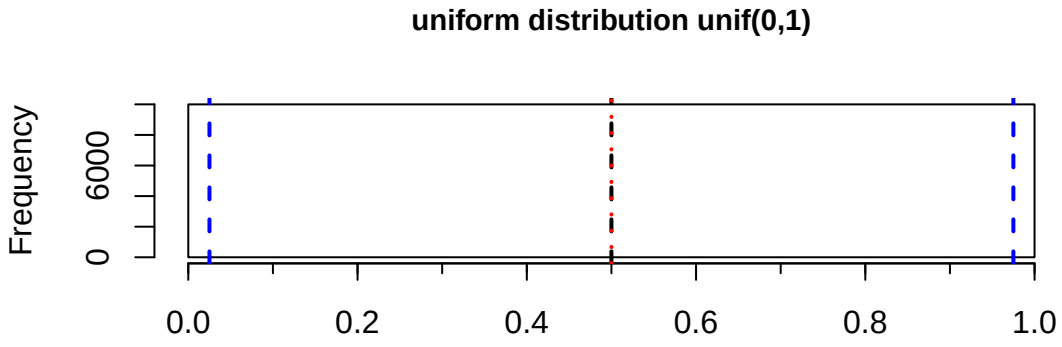


Fig. 2a

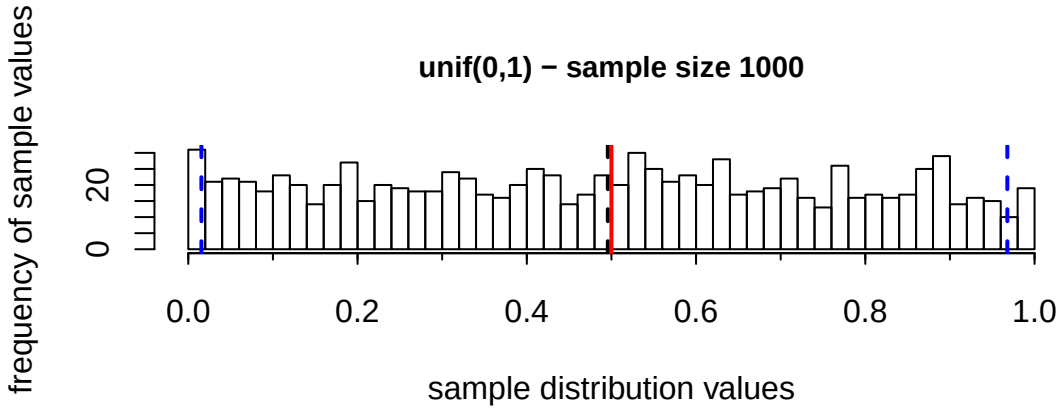


Fig. 2b

```
## [1] "jackknife median estimate using calculations in original scale"

## [1] 0.4957

## [1] "jackknife median CI lower bound using calculations in original scale"

## [1] 0.4402

## [1] "jackknife median CI upper bound using calculations in original scale"

## [1] 0.5513

## [1] ""

## [1] "jackknife median estimate using calculations in percentile scale"
```

```
## [1] "and then backtransformed to original scale"

##    50%
## 0.4957

## [1] "jackknife median estimate lower bound using calculations in "

## [1] "percentile scale and then backtransformed to original scale"

##  46.9%
## 0.4612

## [1] "jackknife median estimate upper bound using calculations in "

## [1] "percentile scale and then backtransformed to original scale"

##  53.1%
## 0.5218
```

Median variance and 95% Confidence Interval estimates
symmetric distribution case 1 − uniform − unif(0,1)

red − pop'n value, black − sample median, blue − 2.5%/97.5% CI bounds

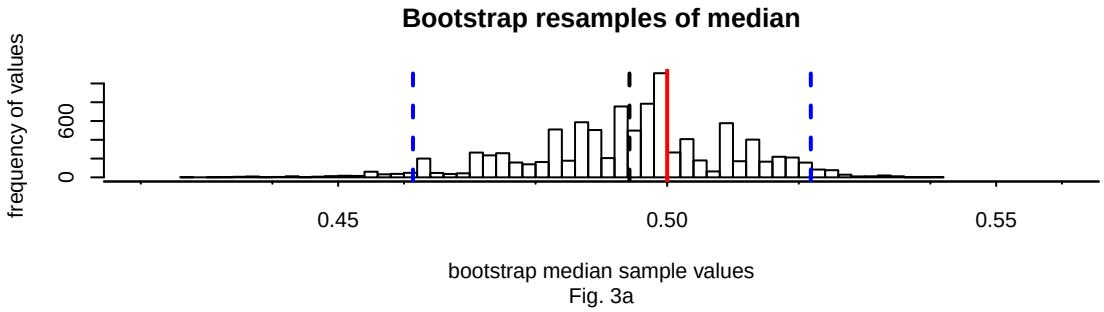


Fig. 3a

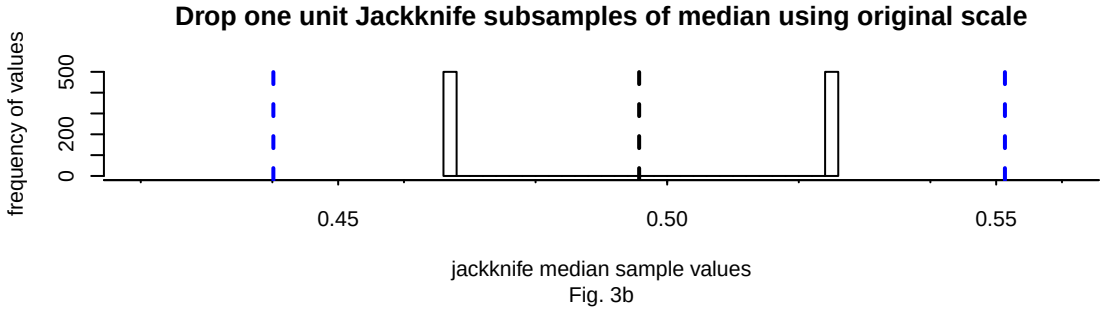


Fig. 3b

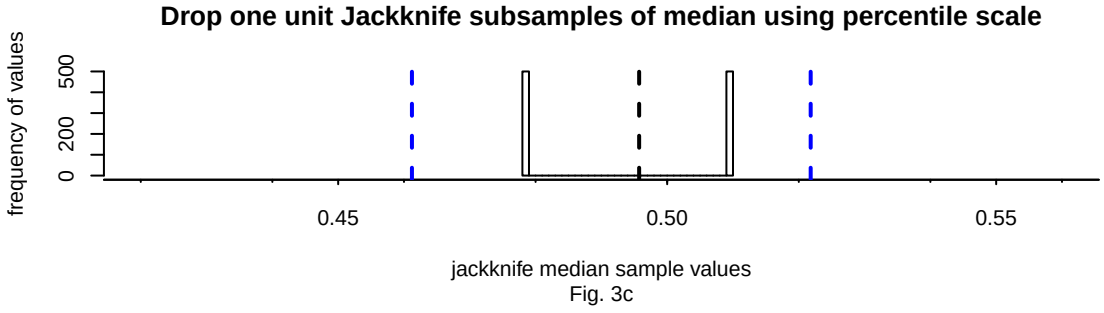


Fig. 3c

Symmetric distribution case 2 − normal − N(0,1)

red line − population median, black line − sample median
blue lines − 2.5%/97.5% percentile bounds

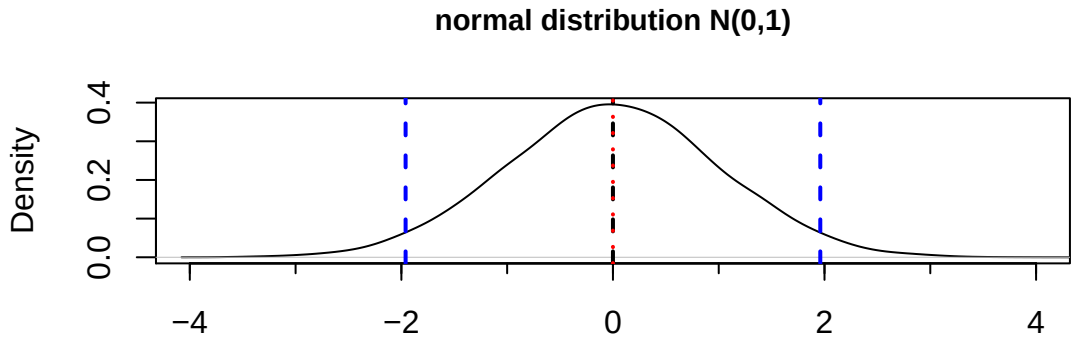


Fig. 4a

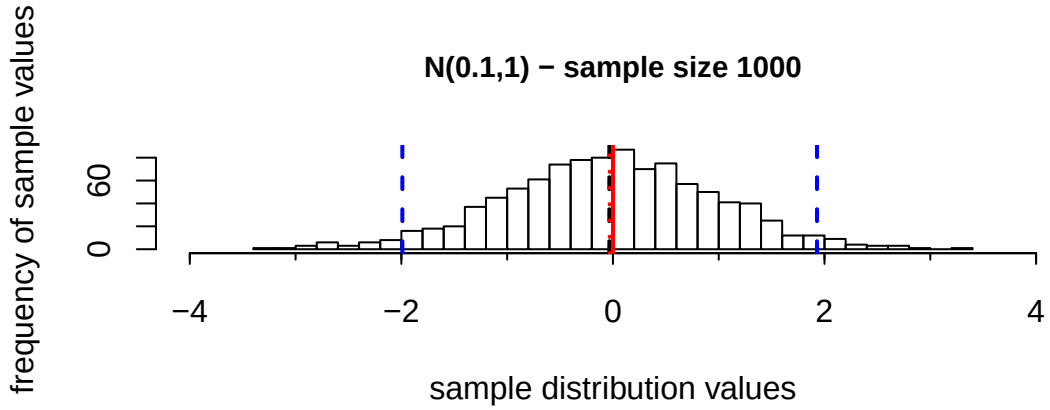


Fig. 4b

Figure 4

Figure 5

```
## Warning: bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-0.1192,  0.0455 )   (-0.1193,  0.0376 )
##
## Level     Percentile            BCa
## 95%   (-0.0993,  0.0576 )   (-0.1025,  0.0563 )
## Calculations and Intervals on Original Scale
```

```
## [1] -0.02484

## [1] 0.04202

## [1] "jackknife estimate based on original scale"

## [1] -0.03087

## [1] -0.08749

## [1] 0.02576

## [1] "jackknife estimates based on percentile scale"

##      50%
## -0.03087

##    46.9%
## -0.09972

##   53.1%
## 0.05772
```

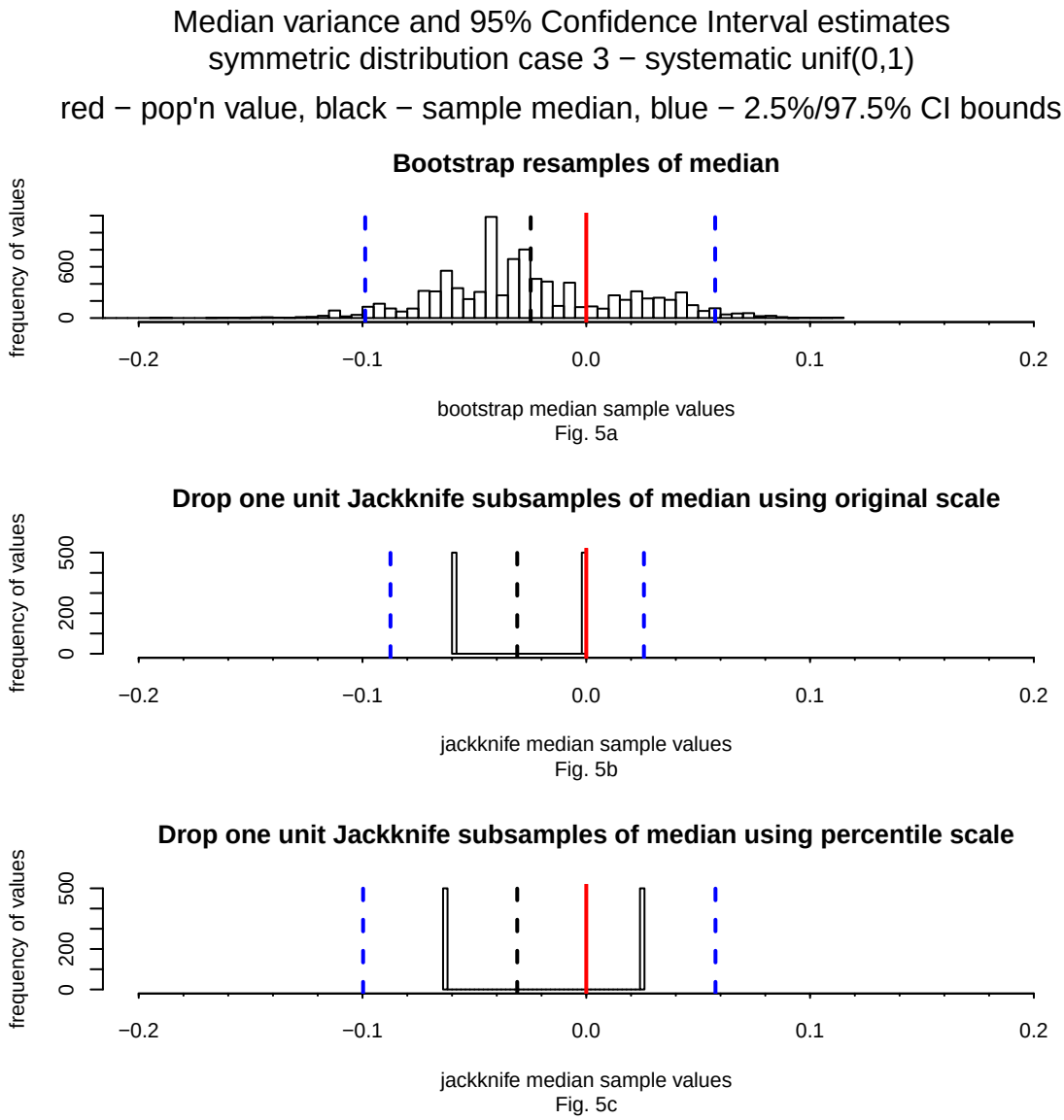


Figure 6

```
## Warning: bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 0.4533,  0.5497 )   ( 0.4532,  0.5494 )
##
## Level     Percentile            BCa
## 95%   ( 0.4532,  0.5494 )   ( 0.4532,  0.5494 )
## Calculations and Intervals on Original Scale
```

```
## [1] "mean of bootstrap samples"

## [1] 0.5011

## [1] "standard error of bootstrap samples"

## [1] 0.0246
```

Symmetric distribution case 3 − systematic unif(0,1)

red line − population median, black line − sample median
blue lines − 2.5%/97.5% percentile bounds

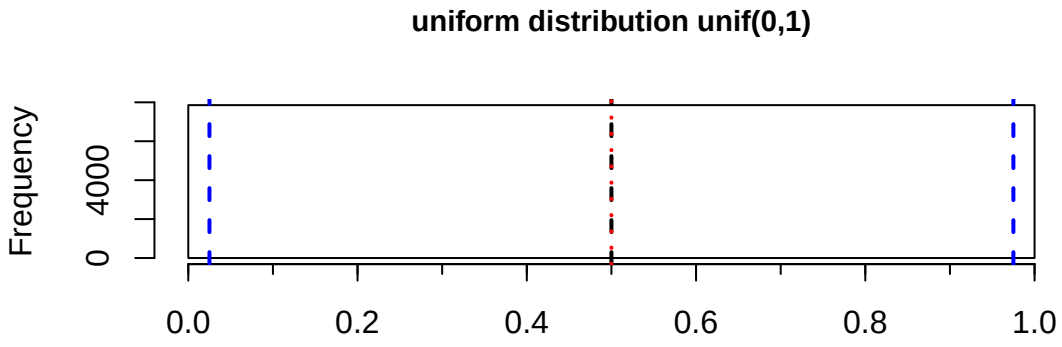


Fig. 6a

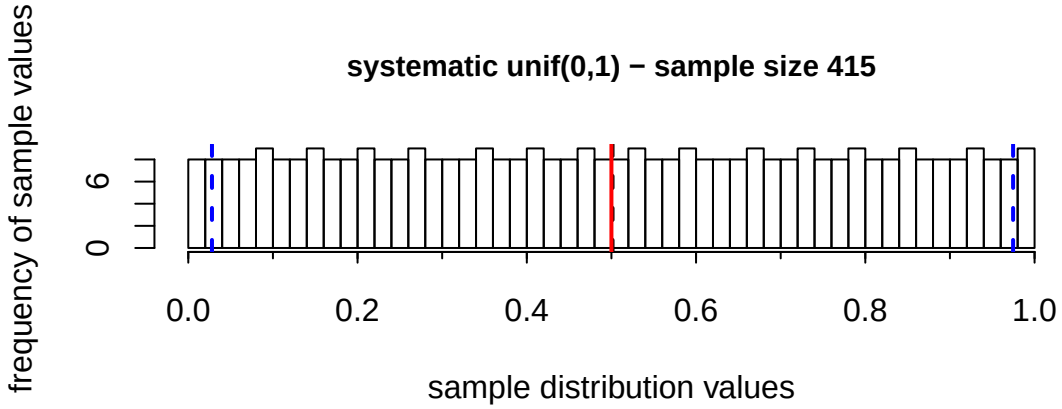


Fig. 6b

```
## [1] "jackknife median estimate using calculations in original scale"

## [1] 0.5013

## [1] "jackknife median CI lower bound using calculations in original scale"

## [1] 0.4534
```

```
## [1] "jackknife median CI upper bound using calculations in original scale"

## [1] 0.5492

## [1] ""

## [1] "jackknife median estimate using calculations in percentile scale "

## [1] "and then backtransformed to original scale"

## 50.12%
## 0.5025

## [1] "jackknife median estimate lower bound using calculations in "

## [1] "percentile scale and then backtransformed to original scale"

##  45.3%
## 0.4545

## [1] "jackknife median estimate upper bound using calculations in "

## [1] "percentile scale and then backtransformed to original scale"

## 54.94%
## 0.5505
```

Median variance and 95% Confidence Interval estimates
symmetric distribution case 3 − systematic uniform − evenly spaced points (0,1)
red − pop'n value, black − sample median, blue − 2.5%/97.5% CI bounds

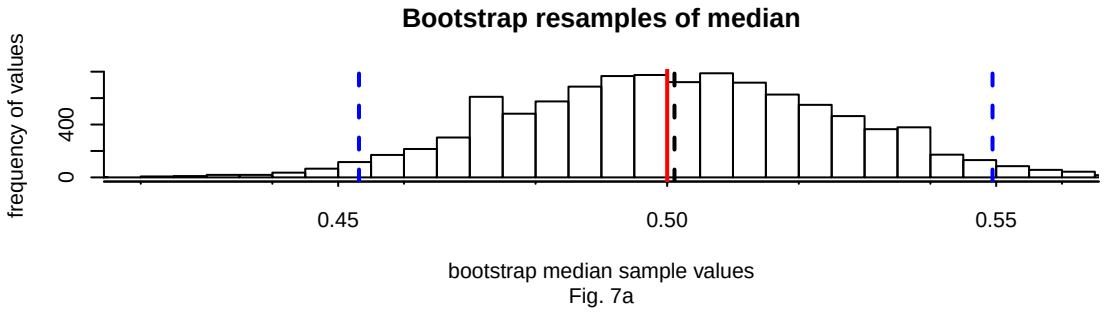


Fig. 7a

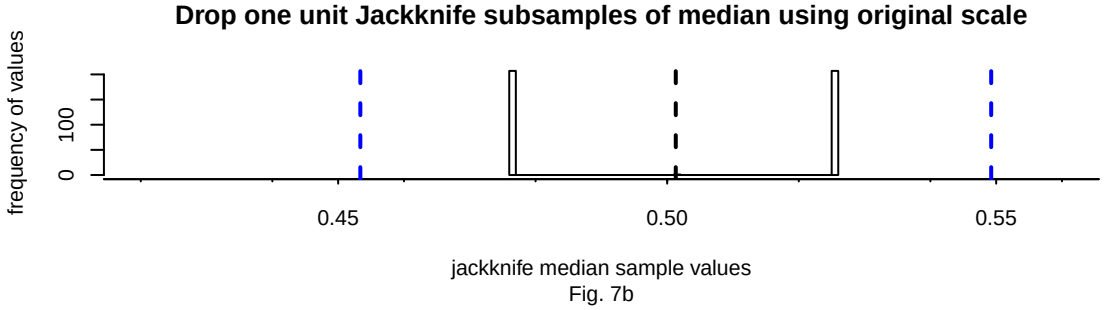


Fig. 7b

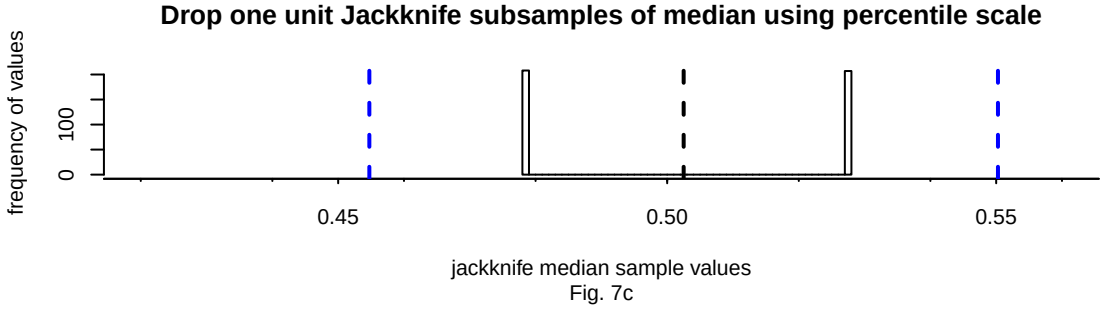


Fig. 7c

A simulated log-normal dataset is created to have a continuous right skewed distribution. This allows examination of the asymmetric nature of the median confidence interval in the original scale.

Skewed distribution case 1 − log−normal exp(N(0.1,1))

red line − population/sample median
black line − population/sample mean

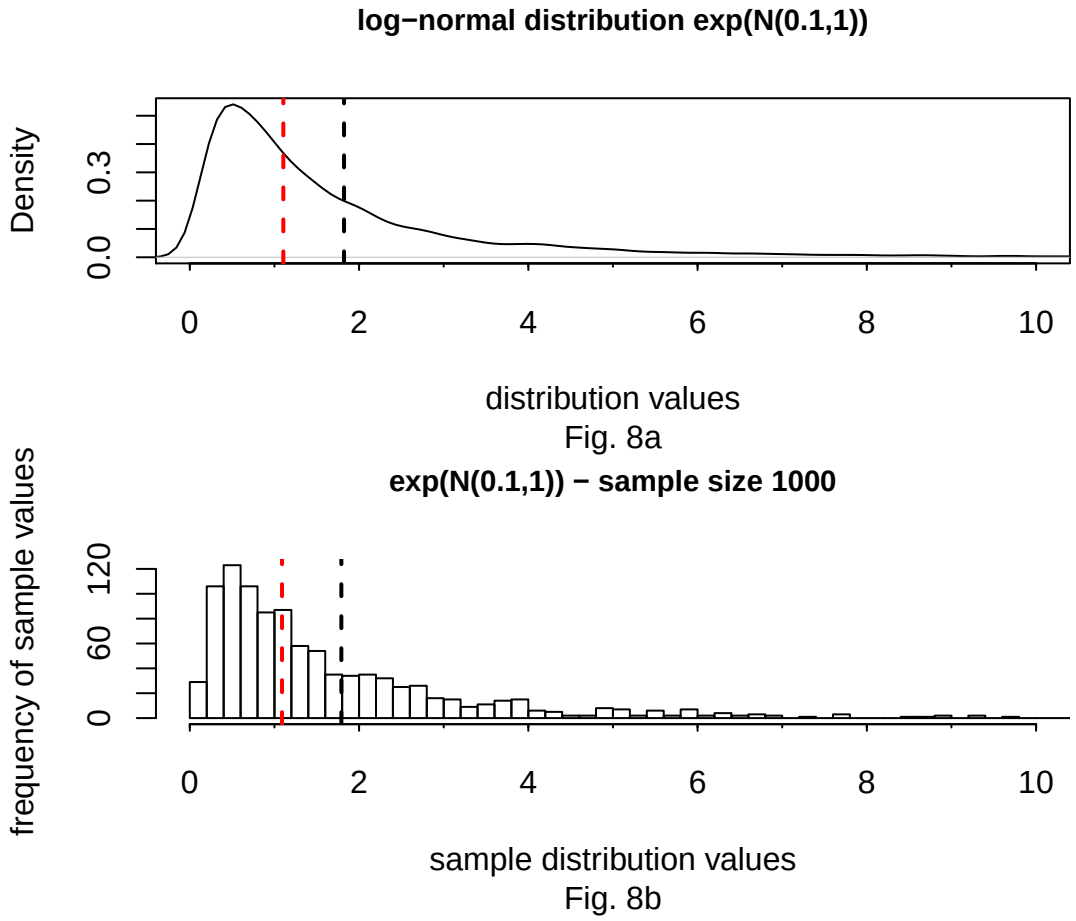
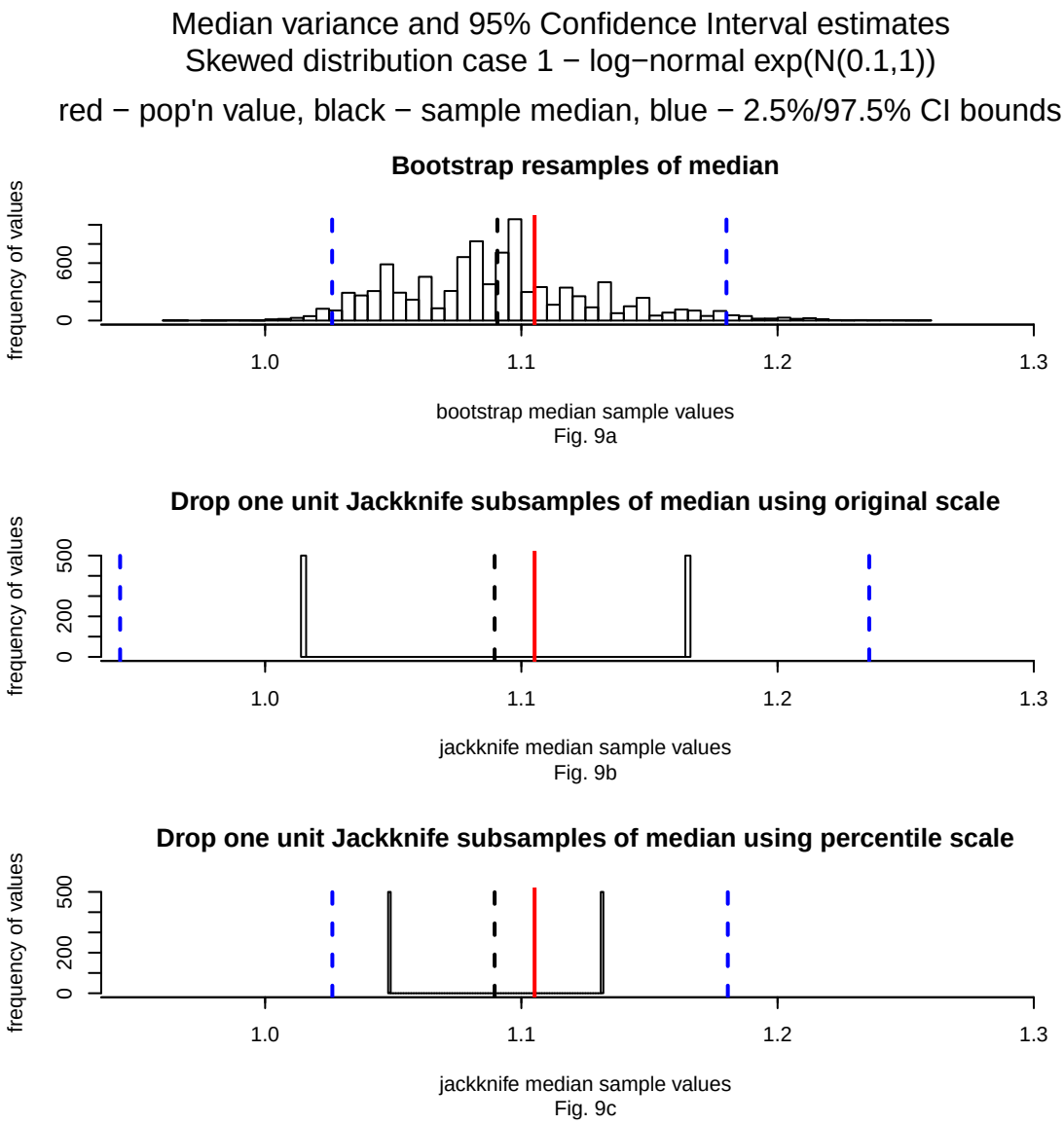


Figure 8

Figure 9

```
## Warning: bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 1.011,  1.166 )   ( 0.999,  1.153 )
##
## Level     Percentile            BCa
## 95%   ( 1.026,  1.180 )   ( 1.026,  1.179 )
## Calculations and Intervals on Original Scale
```

```
## [1] 1.091

## [1] 0.03959

##   2.5% 15.87%    50% 68.27%  97.5%
##  1.026  1.049  1.090  1.100  1.180

## [1] "Jackknife estimates using original scale"

## [1] 1.09

## [1] 0.9434

## [1] 1.236

## [1] "Jackknife estimates using percentile scale and backtransformed"

##  50%
## 1.09

## 46.9%
## 1.026

## 53.1%
## 1.181
```



This final example is a continuous distribution analogue of the waiting duration data for the "old faithful geyser" on r library(datasets) as data(waiting). The sample size of the simulation (270) and the asymmetry of the bivariate distribution is closely matched to the real dataset.

Skewed distribution case 2 − weighted bivariate normal
approximation of old faithful geyser wait duration
red line − population/sample median
black line − population/sample mean

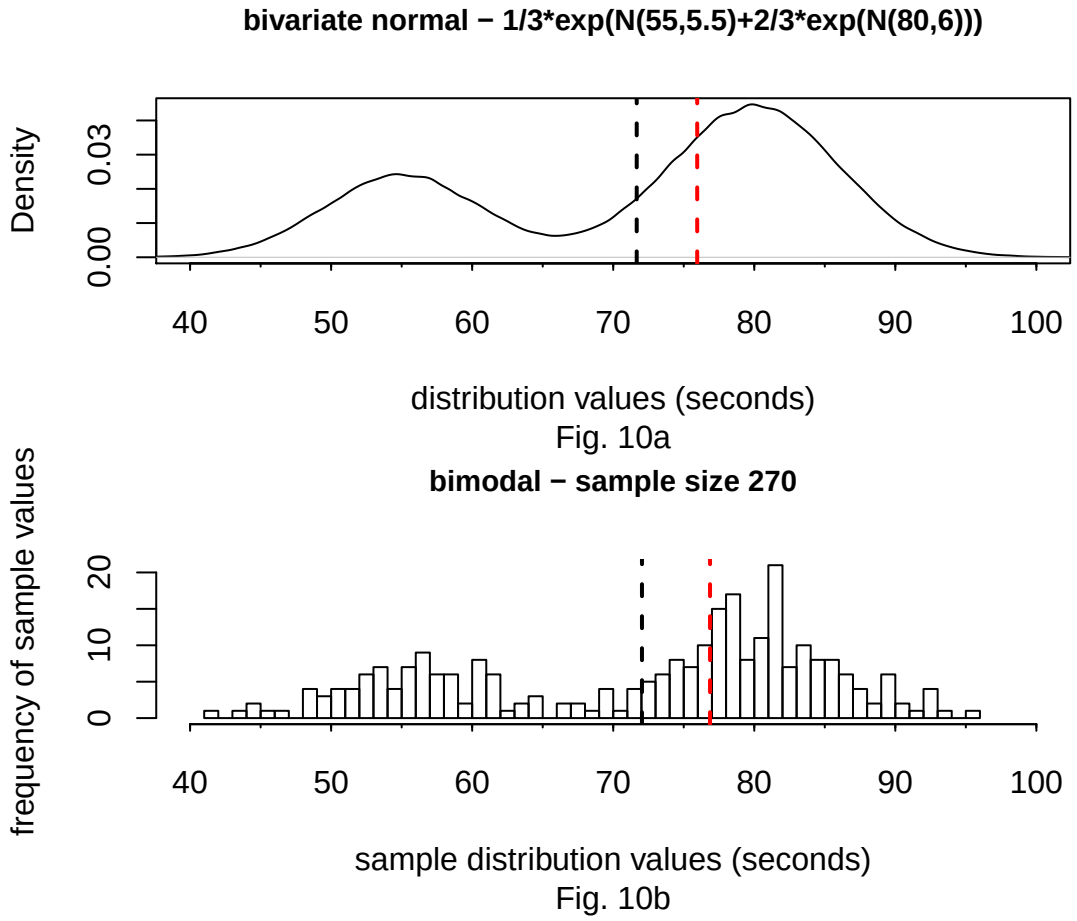


Figure 10

Figure 11

```
## Warning: bootstrap variances needed for studentized intervals

## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 40000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot.out)
##
## Intervals :
## Level      Normal              Basic
## 95%   (75.56, 78.60 )   (75.99, 79.02 )
##
## Level     Percentile            BCa
## 95%   (74.70, 77.73 )   (74.70, 77.73 )
## Calculations and Intervals on Original Scale
```

```
## [1] 76.65
```

```
## [1] 0.7763
```

```
##   2.5% 15.87%    50% 68.27%  97.5%
##  74.70  75.91  76.86  77.11  77.73
```

```
## [1] "Jackknife estimates using the original scale"
```

```
## [1] 76.86
```

```
## [1] 75.98
```

```
## [1] 77.75
```

```
## [1] "Jackknife estimates using percentile scale and backtransformed"
```

```
##   50%
## 76.86
```

```
## 44.02%
##  74.71
```

```
## 55.98%
##  77.74
```

Median variance and 95% Confidence Interval estimates
skewed distribution case 2 − weighted bivariate normal

red − pop'n value, black − sample median, blue − 2.5%/97.5% CI bounds

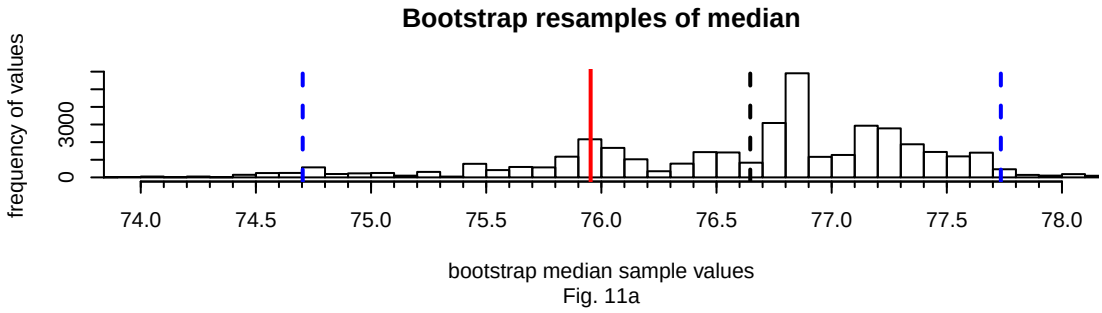


Fig. 11a

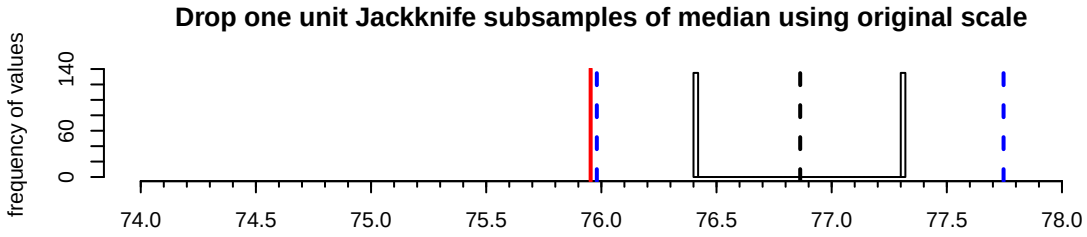


Fig. 11b

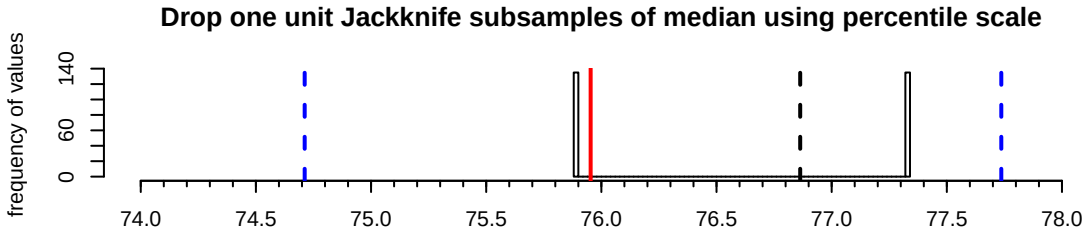


Fig. 11c