**Homework 2. Bootstrapping**

**Student: Hongxing NIU; Matricule ULB: 000342366**

1. Simulate a sample from two independent Gamma distributions X and Y with

different parameters α and λ. .

Sample size *n*=10 for both X and Y. We want to estimate and construct confidence interval. Define and let *t* be the statistic for the simulated sample.

1. Simulate the true distribution of *T* (using 100,000 simulated samples *T*). Use the kernel density estimation to plot the true distribution

We define the parameter values as follows,

=4.3, =1 ;

=4.7, =0.8 ;

So, the true value of theta .

To simulate the true distribution of the statistic of *T*, we randomly take samples, and estimate the two variables by the equations below,

(1)

(2)

(3)

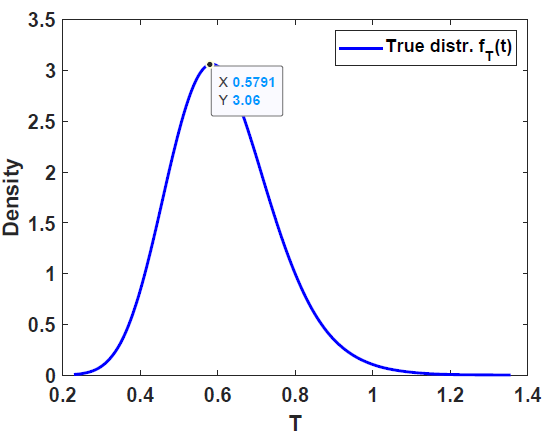
where “” is a function which generates (pseudo) random numbers according to the gamma function.

Afterwards, we compute the Kernel Density Estimation for the data series using the *Cosine Kernel* by the “” function, where the bandwidth is chosen to be

(4)

and the sample size is 1001 (the range of values is split by 1000 equally).

As a result, the true distribution of is simulated by the figure below,



1. For the given sample *t*, find the parametric and non-parametric empirical distributions, i.e., the distributions of *T*\*, based on parametric and non-parametric resampling.

moment estimators of parameters

* The parametric empirical distribution relies on parametric estimators for α and λ. Use the moment estimators:

Since and ,

we obtain and

Similarly, and

For the parametric resampling, use the estimated values ( ) and () to generates B= 10,000 (parametric bootstrapping) pseudo new random samples according to the gamma functions:

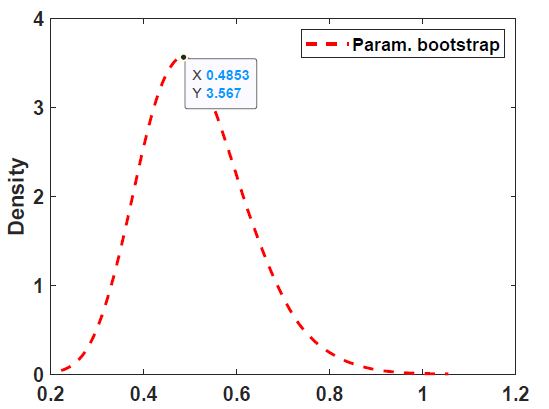
( matrix) (5)

( matrix) (6)

Then, is generated by

(B=10,000 vector) (7)

As a result, the parametric empirical distribution of is shown below,



By comparison, the non-parametric bootstrapping is done by the following equations,

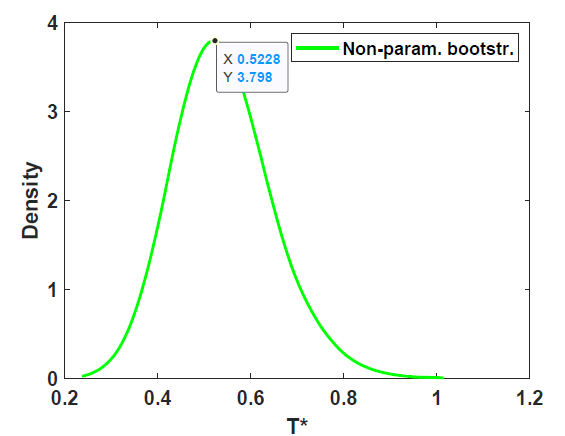
( matrix) (8)

( matrix) (9)

Then, is generated by

(B=10,000 vector) (10)

As a result, the non-parametric empirical distribution of is shown below,



1. Find/estimate the exact bias of *T* from the simulated values of *T*. Compare with exact expression. Compare with the bootstrap estimated bias. Compare also with the exact and estimated variance of T. Repeat the calculations for larger sample size(s) *n*. Any conclusions?

* Find/estimate the exact bias of T from the simulated values of T and compare with exact expression.

Refer to the separate file, the bias of *T* using “exact expression” is calculated by

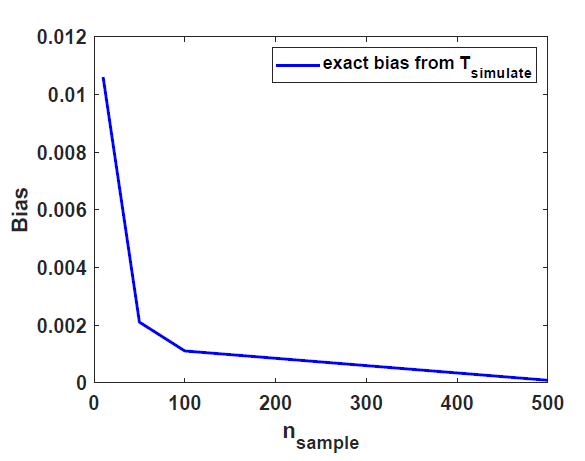
(11)

While the “exact” bias of T from simulated values of T is obtained by

(12)

When the sample size *n*=10, the bias is0.0108 and 0.0106 for (11) and (12), respectively.

Obviously, the value of *β* decreases with the increase of sample size *n* and it is experimentally proved as well that decreases with the increase of sample size *n* , as shown below



* Compare with the bootstrap estimated bias.

(13)

The bias values are 0.0155 and 0.0129 for non-parametric and parametric bootstrapping, respectively.

* Compare also with the exact and estimated variance of T.

In this context, "exact" refers here to the simulated T and the variance of T is in matlab

0.0157 (14)

While "estimated" variance of T refers to a parametric or non-parametric bootstrap estimation, respectively,

0.0189 (15)

0.0193 (16)

* What can we expect from the dependence of the quality of the bootstrap estimator on *n* (sample size)?

By a series of experiments (sample size *n*=10,50,200,500), it was found that all estimated bias values became smaller (the smallest was 8.4485e-05) with the increase of sample size.

As a matter of fact, if we increase the sample size of original data, of course we get smaller bias thanks to more information from the original statistic, which does not mean the quality of T as estimator increase with the increase of sample size *n*. The bootstrap distribution is centered at the observed statistic, not the population parameter, for example, at x¯, not μ. We cannot use the bootstrap to improve on x¯; no matter how many bootstrap samples. In this regard the bootstrap is like formula methods that use the data twice—once to compute an estimate, and again to compute a standard error for the estimate.

1. We now use the fact . We want to construct a percentile confidence interval, but for that we need that the parameter of interest is the median of the estimator. If , then define the parameters and . We will construct a CI for and correct it with *δ*.

We set the sample size *n=10* in question (d).

* Find numerical values for *τ* and *δ*, using simulations of T

= 0.6018 (17)

= -0.0017 (18)

* Find the parametric and non-parametric bootstrap estimators *τ\** and *δ\**.

For the parametric bootstrap estimator,

= 0.4926 () (19)

= -6.0297e-04 (20)

For the non-parametric bootstrap estimator,

= 0.4921 () (21)

= 1.7467e-04 (22)

* Use the parametric and non-parametric bootstrap simulations of *T\** to construct a parametric and nonparametric percentile bootstrap confidence interval for *τ*, and from there a confidence interval for *θ*.

Let α=0.05. For the parametric bootstrapping, the percentile CI for is:

(23)

The result is [0.3614 0.6828]

For the nonparametric bootstrapping, the percentile CI for is :

(24)

The interval is [0.3510 0.6782]

Afterwards, we add *δ* to these intervals to account for a “median bias” (very small correction),

= [0.3614 0.6828] (25)

[0.3510 0.6782] (26)

1. Construct the non-parametric basic bootstrap confidence interval.

Let α=0.05. The non-parametric basic (1- α)×100% confidence interval can be found using

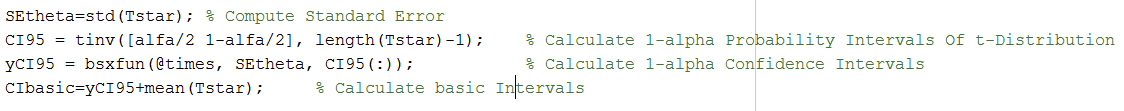
(27)

where is the standard error for the statistic , obtained using the bootstrap. Usually, the CI in above equation can be used when the distribution for is normally distributed or the normality assumption is plausible.

Bootstrap estimate of standard error is done by

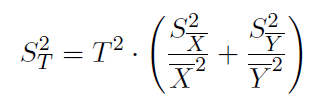
where is the sample mean of the B bootstrap replicates.

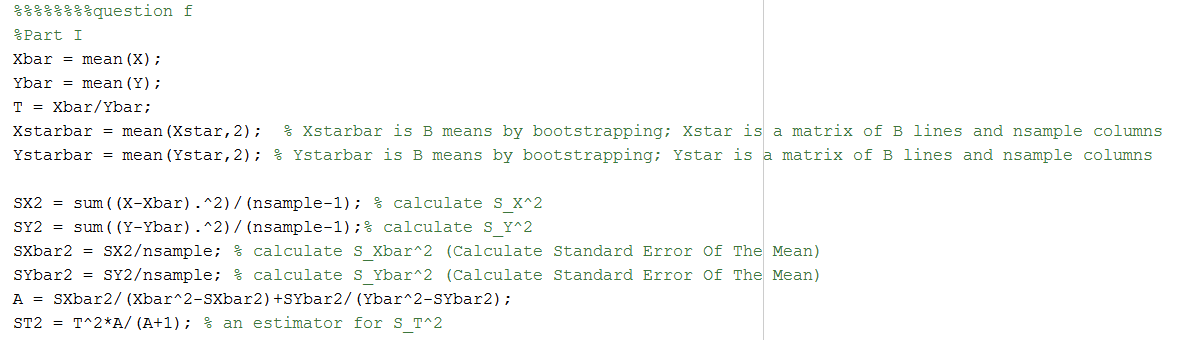
In Matlab, we use SEtheta=std(T\*) to get the value=0.0840.

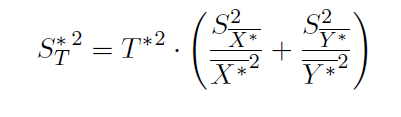


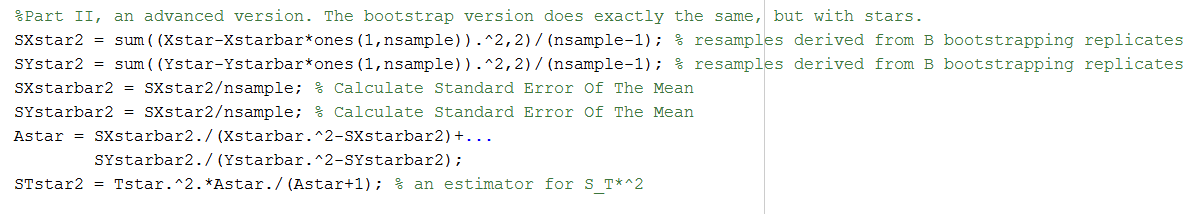
At the end, = [0.3369 0.6661]

1. Construct the non-parametric bootstrap-*t* confidence interval

* Using the approximation , the Matlab code is shown below,



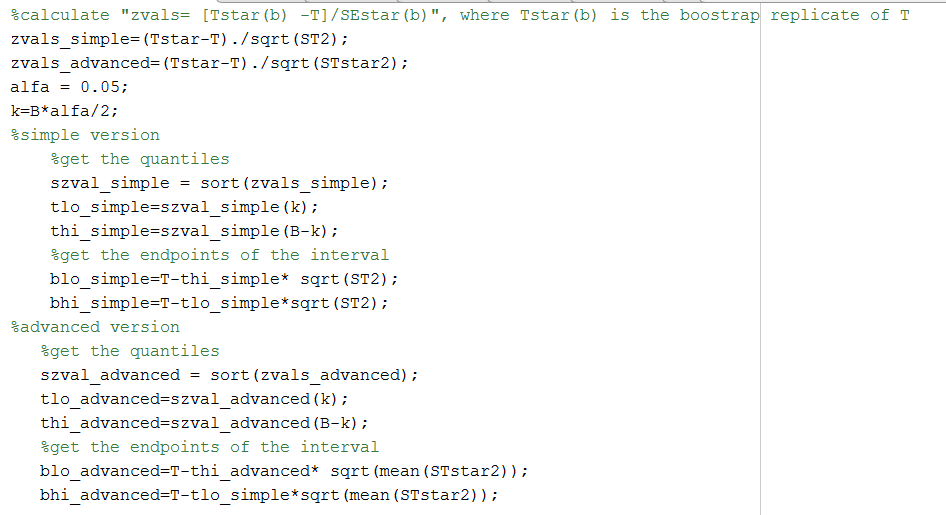
* An advanced version using  is coded in Matlab below,



After we calculate the statistic , we can get the quantiles that we need for the intervals given in below equation and calculate the confidence intervals,

.

They are implemented in Matlab,



When α=0.05, the non-parametric bootstrap-t intervals are [0.3420 0.7722] and [0.3894 0.7029] for the simple and advanced versions, respectively. The bootstrap-t intervals in this study are slightly broader than the bootstrap percentile confidence intervals and bootstrap basic confidence intervals. The bootstrap-t interval estimates the standardized version of T from the data, avoiding the normality assumptions used in the basic interval estimation.