Student Number ....

Exam of Statistical Computing II — 24-06-2019

EcoStat PhD Program, Università degli Studi di Milano Bicocca

Task 1

We observed the following sample of iid data extracted from the distribution F:

$$2, 2, 3, 3, 3, 4, 5$$

Our aim is to estimate the median of F using the plugin estimator.

a) Give a general definition of the plugin estimator, and use it to compute the sample median.

b) Compute the 6 Jacknife replications of the median, and give an estimate of the standard error of the plugin estimator.

c) Is the estimate of the standard error reliable? Comment on the result, and explain if it is possible to provide a better estimate.

Task 2

We observe a sample of size $n = 20$ from an unknown continuous distribution. We denote by F the unknown cdf. Our aim is to give a confidence interval for the ratio between the first and the third quartile of the distribution:

$$\theta = \frac{Q1}{Q3}$$

where $Q1$ and $Q3$ are the first and third quartiles, respectively: $Q1 = F^{-1}(0.25)$; $Q2 = F^{-1}(0.75)$. A bootstrap is performed on the data set, using $B = 10000$ replications. Figure 1 shows the obtained histogram and QQ-plot.

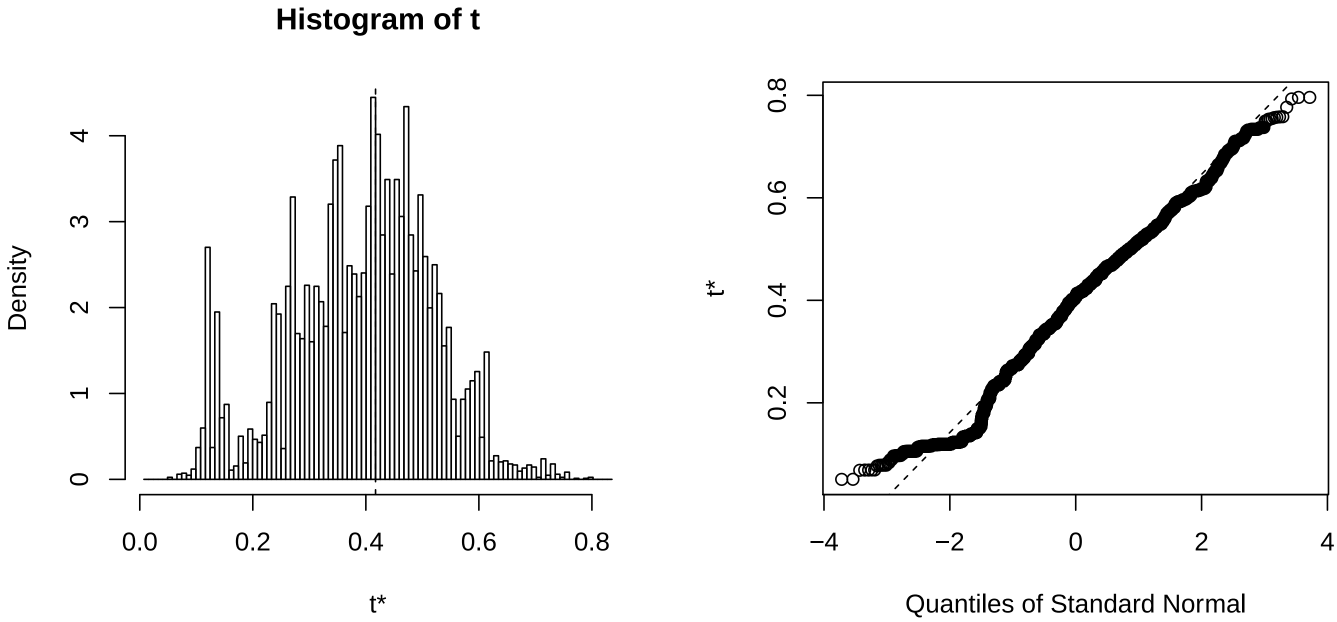

Figure 1: Histogram and QQ-plot of 10000 Bootstrap replications of θ.

a) What do the histogram and QQplot in Figure 1 represent? Briefly explain how they are obtained.

b) The computation of 95% Bootstrap confidence intervals gives the following result:

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.score, conf = 0.95)

Intervals :
Level      Normal              Basic
95%   ( 0.1936,  0.6871 )   ( 0.2200,  0.7123 )

Level     Percentile            BCa
95%   ( 0.1221,  0.6143 )   ( 0.1230,  0.6174 )
Calculations and Intervals on Original Scale
```

Explain the differences between the obtained intervals, both in terms of their theoretical definition, and in terms of the numerical result. Are the differences observed consistent with theory?

c) The Bootstrap-t confidence interval is not reported. Discuss (just theoretically) how it is defined, and if you would expect different results.

d) Choose the confidence interval that has better properties in this case, motivating your answer.

Task 3

We observe a data set of size 40 from a 2-dimensional random variable. The observed data points are plotted in Figure 2. Our aim is to cluster the data into 3 groups.

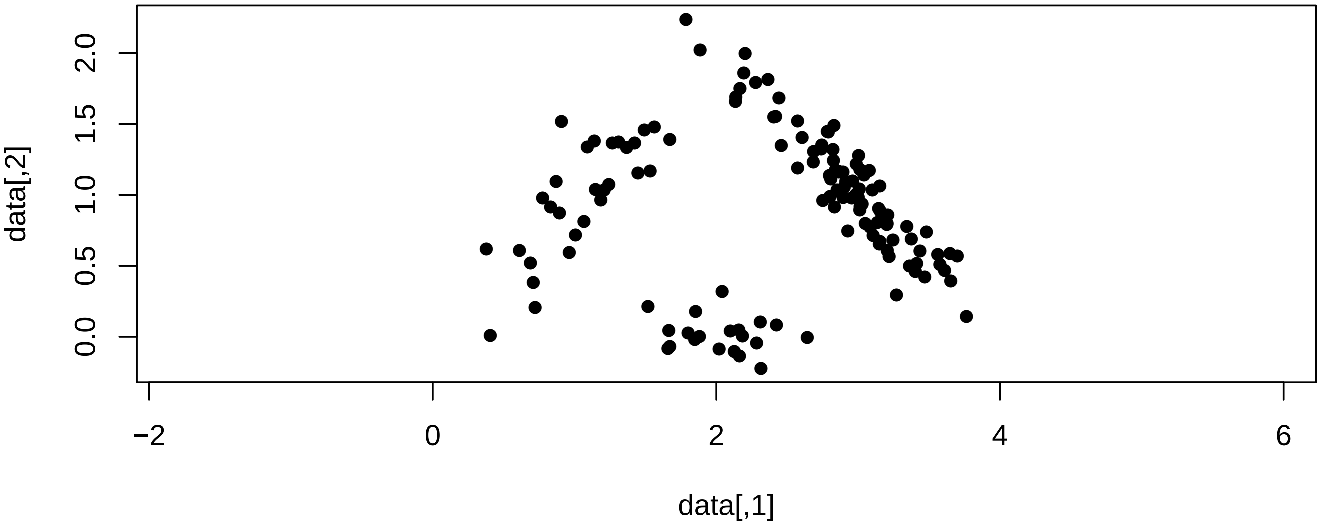

Figure 2: Scatterplot of data points.

a) Describe the k-mean algorithm.

b) Describe the EM algorithm in the setting of Gaussian mixtures, and point out the differences between the EM and the k-means.

c) You expect similar results of the two methods on the observed data? If yes, discuss why, if not, discuss which of the two methods you would rather use.