Assignment 3 – Semi- and Nonparametric methods

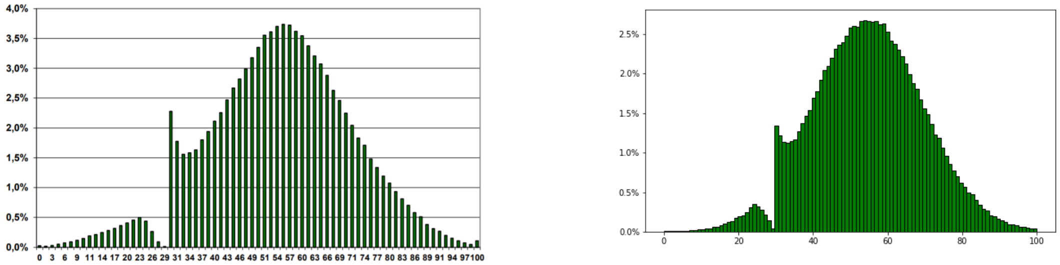
Advanced Econometrics 2

Stepan Svoboda, Jan Hynek

January 26, 2018

1 Report

We have chosen to study an interesting data set that we have stumbled upon the internet. The results of Polish high school graduation exam have a very peculiar shape, visible in figure 1a).



(a) Original

(b) Simulation

Figure 1: Comparison of original distribution with our simulation – both with 330000 observations

Unfortunately this data set is not available online so we have simulated this data set to the best of our abilities so it captures the main behavior of the original. This can be seen in Figure 1b).

1.1 Research question

We want to estimate the underlying distribution of this data set. The easiest way to describe our data is a slightly right-skewed normal distribution of grades between 0 and 100 with a pit just before the passing grade and a jump just at and above the passing grade.

An alternative description would be to think of this distribution as a bimodal one which we can split into two normal-looking right-skewed distributions with heavier tails.

Our goal is to estimate the density of this data set and deal with the two issues that inevitably face us. The first one is common for both ways of thinking about this issue – capturing the split within of our data and the second one is exclusive to the second way of thinking about our data – the end-point problem we encounter if we divide our data into two separate densities and estimate those separately.

1.2 Methodology

In this section we wish to describe how we approached this issue and how have we gotten the results. We simulated our data because the original data set was not available readily online. Luckily this distribution shape is quite straighforward to replicate.

After we created our data we moved on to the density estimation. There were several issues we encountered including some very general and not data set-specific.

The first step was a straightforward kernel density estimation method,

$$\hat{f}(x) = \frac{1}{hN} \sum_{i=1}^{N} K\left(\frac{x_i - x}{h}\right). \tag{1}$$

To calculate this estimator we must first determine our kernel function and our bandwidth h, the rest is given by our data. We chose the Gaussian kernel as the implementation is not difficult and the impact of kernel choice should be only marginal in case of many observations, which is our case. We focused more on the choice of bandwidth. We have chosen the plug-in estimator of the bandwidth for the Gaussian kernel of the form:

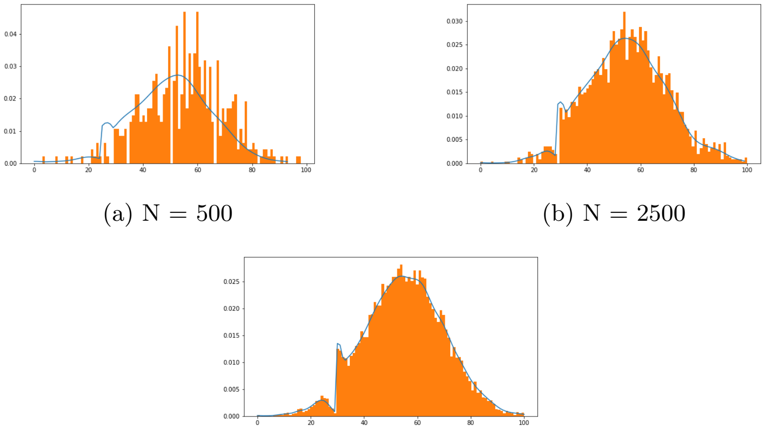
$$h = 1.059\, sd(x)\, N^{-\frac{1}{5}}. \tag{2}$$

This kind of estimator unfortunately cannot estimate the bimodality present in our data and thus we had to apply the described kernel density estimator to two subsets of our data. Since we can see the low share of students who just failed and high share of students who just passed we estimated two densities. The density of the distribution of the failed and of the successful students. We simply applied this estimator to both parts and then in the final density estimate reweighted the densities. There are a few pitfalls such as using variance of the subset in the plugin estimator that lead to bad outcomes but these are avoidable.

The last problem that came up was the end-point bias. There is only place where it truly matters that our kernels lose mass at the end of the distribution and that is the left end of the density of passing students. We applied there the $reflection$ boundary correction technique by reinstating this mass using $K_h(x-x_i)+K_h(x+x_i)$ instead of just $K_h(x-x_i)$. This in practice means we took the part of the kernel that stretched beyond our distribution and replaced this missing mass with the equivalent from the actual distribution.

1.3 Results

We can observe in Fig. 2 that our estimated density provides us with feasible estimates from cca N = 2500.



(a) N = 500

(b) N = 2500

(c) N = 10000

Figure 2: Samples with different amount of observations along with their density estimations

The case of 500 observations is obviously problematic since we have 101 different possible realizations and only 500 observations, which necessarily leads to a model that can at best somewhat approximate the true distribution. In our case it did detect the soft discontinuity around the point 30, which is probably all we can ask from our model with this amount of observations.

When we had 2500 observations it functioned much better and our approximation captured nearly all the information present in the data. If we compare the estimates from 2500 and 10000 observations the main difference is that with more observations our model is able to capture the sharpness of the drop and the subsequent jump at the passing grade much better. Second difference, less visible one, is in the tails. When we have much more observations our density estimate, expectedly, behaves better in the tails.

1.4 Conclusion

We have done several things in our assignment. Firstly we found an interesting real-life data set that described a peculiar phenomenon. Then we took our density and estimated it as a combination of two different densities, i.e. captured the bimodality. Then we tackled the end-point bias problem which gave us the final model that we presented in the previous section.