Journal of Statistics: Advances in Theory and Applications
Volume 11, Number 2, 2014, Pages 109-115

RESAMPLING DIFFERENCES OF ORDER STATISTICS (RDOS)

ANDRÉ ADLER

Department of Applied Mathematics
Illinois Institute of Technology
Chicago
Illinois, 60616
USA
e-mail: adler@iit.edu

Abstract

Using an interesting limit theorem for differences of order statistics, we establish a new way to estimate the parameter from an exponential distribution. From a small set of observations, we can create many data sets using these pairs of order statistics. Each data set will produce one estimator of λ. We will then average all of these estimators to obtain our final estimator of our parameter. This technique does provide an alternative to the classic way of estimation while using both theory and application.

1. Introduction

Using the strong laws of large numbers for differences of pairs of adjacent order statistics, we will produce a new estimator for the parameter in the exponential distribution. The conclusion is both interesting for its theoretical value and also its application. The theory is nice and clean and the simulations work as well. This same technique

2010 Mathematics Subject Classification: Primary 62F10; Secondary 62F40.
Keywords and phrases: order statistics, exponential distribution, resampling procedures, strong law of large numbers.
Received June 14, 2014

performed quite well when we observed ratio of order statistics, RROS, from the Pareto distribution, see [1].

The resampling plan is to take the original N observations and randomly place them into an n by m matrix. From that one matrix, we will obtain one statistic. We then will randomly place all N data points into a similar matrix and obtain another statistic. Next, we will resample until the average of these statistics converge. We did this dozens of times with various N's and various λ's and even though we resampled 10,000 times, it turns out that the average converged every time around 4,000 resamples. Using a basic PC this only took two to three seconds. We can also remove one observation at a time, and obtain all matrices associated with these $N-1$ observations. Next, we can toss out two of the original observations or three and obtain many estimators of our parameter. Averaging the statistics obtained from each of these matrices will give us an excellent estimator of our parameter. One can clearly see that our original data can be reassigned many ways.

This estimator is a reasonable competitor to the classic MLE estimator. It wins about half the time. It not only predicts λ, but it also detects if the underlying distribution is not exponential or even continuous. If the differences of the data is not spread out sufficiently, then our statistic will be near zero, which means that the underlying distribution is degenerate, since $\tilde{\lambda}$ is near zero. On the other hand, if one uses the MLE, i.e., the sample mean, it will just average the data and estimate λ as this average. Hence, it will not detect any aberration from the continuous exponential distribution like our RDOS estimator does.

2. Exponential Distribution

The underlying density is $f(x) = (1/\lambda)e^{-x/\lambda}I(x \geq 0)$, where $\lambda > 0$. Our goal is to estimate the parameter λ by using pairs of order statistics and resampling the original data to make our estimator robust.

The joint density of the original i.i.d. random variables $X_1, \ldots, X_m$ is

$$f(x_1, \ldots, x_m) = \frac{1}{\lambda^m} e^{-\sum_{i=1}^m x_i/\lambda} I(x_1 \geq 0) \cdots I(x_m \geq 0).$$

Hence the density of the corresponding order statistics $X_{(1)}, \ldots, X_{(m)}$ is

$$f(x_{(1)}, \ldots, x_{(m)}) = \frac{m!}{\lambda^m} e^{-\sum_{i=1}^m x_{(i)}/\lambda} I(0 \leq x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(m)}).$$

Next we obtain the joint density of $X_{(1)}, D_1, \ldots, D_{m-1}$, where $D_i = X_{(i+1)} - X_{(i)}$ and we let $W = X_{(1)}$. In order to do that we need the inverse transformations, which are

$$X_{(1)} = W,$$

$$X_{(2)} = W + D_1,$$

$$X_{(3)} = W + D_1 + D_2,$$

through

$$X_{(m)} = W + \sum_{i=1}^{m-1} D_i.$$

So, in order to obtain this density we need the Jacobian, which is the determinant of the matrix

$$\begin{pmatrix} \frac{\partial x_{(1)}}{\partial w} & \frac{\partial x_{(1)}}{\partial d_1} & \frac{\partial x_{(1)}}{\partial d_2} & \cdots \frac{\partial x_{(1)}}{\partial d_{m-1}} \\ \frac{\partial x_{(2)}}{\partial w} & \frac{\partial x_{(2)}}{\partial d_1} & \frac{\partial x_{(2)}}{\partial d_2} & \cdots \frac{\partial x_{(2)}}{\partial d_{m-1}} \\ \frac{\partial x_{(3)}}{\partial w} & \frac{\partial x_{(3)}}{\partial d_1} & \frac{\partial x_{(3)}}{\partial d_2} & \cdots \frac{\partial x_{(3)}}{\partial d_{m-1}} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial x_{(m)}}{\partial w} & \frac{\partial x_{(m)}}{\partial d_1} & \frac{\partial x_{(m)}}{\partial d_2} & \cdots \frac{\partial x_{(m)}}{\partial d_{m-1}} \end{pmatrix},$$

which is the lower triangular matrix

$$\begin{pmatrix} 1 & 0 & 0 & \dots 0 \\ 1 & 1 & 0 & \cdots 0 \\ 1 & 1 & 1 & \cdots 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \cdots 1 \end{pmatrix}.$$

Thus the Jacobian is 1. So, the joint density of $X_{(1)}, D_1, \dots, D_{m-1}$ is

$$f(x_{(1)}, d_1, \dots, d_{m-1}) = \frac{m!}{\lambda^m} e^{-[mx_{(1)}+(m-1)d_1+(m-2)d_2+(m-3)d_3+\cdots+d_{m-1}]/\lambda},$$

where each of these random variables have support on $[0, \infty)$. This shows that the random variables $X_{(1)}, D_1, \dots, D_{m-1}$ are independent and that the density of D_i is

$$f_{D_i}(d_i) = \frac{(m-i)}{\lambda} e^{-(m-i)d_i/\lambda} I(d_i \geq 0).$$

Hence these random variables are independent exponentials with mean $\lambda/(m-i)$. This means that a natural estimator of λ is $D_i(m-i)$. Next, we will average over $i = 1, \dots, m-1$ to obtain a better estimate of our parameter. Finally, by resampling the original data repeatedly, our final statistic is a very robust estimate of λ.

3. Application

As we did with the Pareto random variables, see [1], we take a sample of size N and fill an n by m matrix with these observations from our exponential distribution. We can pick any n and m as long as $nm \leq N$ and we can do that by tossing aside a few observations for each matrix we generate. With this n by m matrix of i.i.d. exponential random

variables, we first order the elements within each row. That creates the order statistics, then we take the differences between columns. Thus, we have the difference of adjacent order statistics within each row. That creates a new n by $m-1$ matrix, with element D_{ji}. Since $E(D_{ji}) = \lambda/(m-i)$, by summing over both rows and columns of our matrix of differences of adjacent order statistics, we have as a natural estimator of λ

$$\hat{\lambda} = \frac{\sum_{j=1}^{n}\sum_{i=1}^{m-1} D_{ji}(m-i)}{n(m-1)},$$

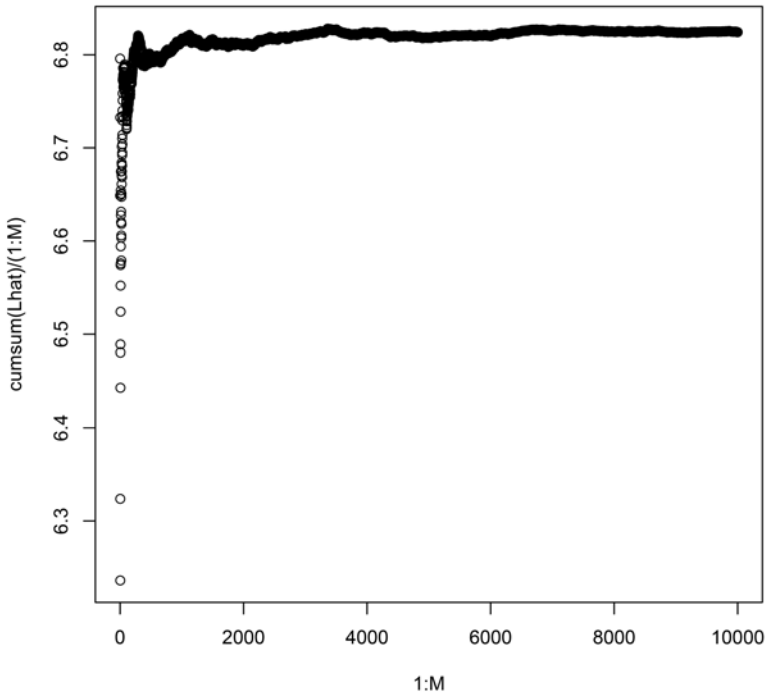
for this one matrix. Where, once again D_{ji} is the i-th difference of adjacent order statistics within the j-th row. Next, we repeat this over and over again by taking the original N random variables and obtaining a new $\hat{\lambda}$ for each resampled matrix.

We continue in this fashion by resampling $N-1$ exponential random variables by tossing out one of the original observations. Each time we place these $N-1$ values in a matrix, we obtain another estimator of λ. We can then toss out two of our N observations. This creates a huge collection of possible estimators of λ. Next, we average over all of these $\hat{\lambda}$'s to obtain our final estimator, $\tilde{\lambda}$. The strong law of large numbers forces this to be an excellent and of course consistent estimator of λ.

As with the RROS estimator in [1], our RDOS estimator wins about half the time. Running simulations using R with various N's and various λ's, we noticed that after about 4,000 resamples the RDOS estimator stabilized every single time. These simulations only take about two to three seconds on a basic PC. This proves that this is an estimator who's time has come. With the use of any sophisticated machine, these type of estimators will be immediate.

In the table we show what happened, when we set $\lambda = 7$. We generated $N = 30$ i.i.d. exponential random variables using R. The sample mean was 6.679467 and $\tilde{\lambda}$ was 6.824357. What one can see is that $\tilde{\lambda}$ does converge very quickly. In that setting, we had 6 columns and 5 rows for our $N = 30$ observations. Using these 6 by 5 matrices, our statistics converged very quickly, so we did not need to toss out any of the original 30 values and resample any further. However, we can toss out any 2 of the 30 original random variables and form all 4 by 7 and then all 7 by 4 matrices. With $\binom{30}{2}$ equal to 435, we see that we can create 435 new data sets with this type of resampling. And we can toss out any 3 of the 30, which has 4060 ways of being done. Then with those 27 observations, we can form all possible 3 by 9 and all possible 9 by 3 matrices. The possibilities are endless. In the end, we will use the strong law of large numbers and average over all of these statistics to obtain our $\tilde{\lambda}$.

We only need a few lines of programming in R to obtain $\tilde{\lambda}$. We start by letting R generate N uniform random variables, U, and by setting $Y = -\lambda \ln U$, we have our exponential random variables with mean λ. Since the program in R was only a few lines long, it would be very easy to incorporate these types of estimators into existing software. Some day these RDOS and RROS estimators should be included in all of these computer packages that statisticians use to estimate parameters. They are an interesting alternative to our classic estimators. What's quite nice about them, is that they combine theory and applications. The math is necessary to derive these various estimators and today's computers can now quickly compute them. This is a nice marriage of theory and application.



Acknowledgement

I would like to thank Allen Favell for helping me run these simulations in R.

Reference

[1] A. Adler and G. Skountrianos, Resampling ratios of order statistics, Communications in Statistics 41 (2012), 1891-1894.

■