Journal of Statistical Theory and Applications, Vol. 12, No. 2 (July 2013), 185-190

Testing the mean of an exponential distribution in the presence of outliers

Abbas Mahdavi

Department of Statistics, Faculty of Mathematical Sciences
Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran
E-mail address: a.mahdavi@vru.ac.ir

Received 17 March 2012
Accepted 17 March 2013

In this study, we propose a simple robust test for the mean of an exponential distribution by using the simplified version of "Forward Search" (FS) method. The FS method is a powerful general method for identifying outliers and their effects on inferences about the hypothesized model. The simulation studies indicate robustness of the testing method and the ability of the procedure to capture the structure of data. Results are presented through the plots which are powerful in revealing the structure of the data.

Keywords: Forward Search; Robust approach; Outlier; Least Median of Squares.

1. Introduction

The exponential distribution has an essential role in a variety of applications in reliability engineering and life testing problems. The exponential hazard rate is constant and the estimation and test theory can easily be detailed for the exponential model, therefore, the mean of this distribution is an important characteristic that is often of interest to an experimenter.

We should pay attention to outliers because a small departure from the assumed model can have negative effects on the efficiency of classical estimators. The Forward Search (FS) approach is a powerful general method that provides diagnostic plots for finding outliers and discovering their underlying effects on models fitted to the data and for assessing the adequacy of the model. Atkinson and Riani ([1], [2], [3]) developed the FS procedure for regression modeling and multivariate analysis frameworks. The FS method starts from a small, robustly chosen subset of the data. The method increases the subset size by using some measure of closeness to the fitted model until finally all the data are fitted. The outliers enter the model in the last steps and the entrance point of the outliers can be revealed by monitoring some statistics of interest during the process. Recently the FS method is implemented in wide applications, e.g. ANOVA framework [5] and testing normality [6]. For further results see [4].

The purpose of this article is to adopt the simplified version of FS method in testing the mean of an exponential distribution. The most popular test for the mean of exponential distribution is based on a Chi-square distribution, but presence of outliers influences this test strongly. In this paper we try to determine how many and which observations agree with the null hypothesis about the mean of an exponential distribution.

The paper is organized as follows. In Section 2 we briefly introduce testing the mean of an exponential distribution. Section 3 presents the proposed forward search algorithm in testing the mean of an exponential distribution. In Section 4, the performance of the method is illustrated with simulated data and the behavior of our procedure is analyzed. Finally concluding remarks are provided in Section 5.

2. Testing the Mean of an Exponential Distribution

Historically, the exponential distribution was the first widely discussed lifetime distribution. The probability density function (pdf) of a random variable $X \sim Exp(\mu)$, which represents the lifetime variable of interest, is given by

$$f\left(t \mid \mu\right) = \mu^{-1} \exp\left(-t/\mu\right), \quad t > 0. \tag{2.1}$$

The positive parameter μ is the mean lifetime and $\lambda = \mu^{-1}$ is the hazard rate.

Let $X_1, X_2, \ldots, X_n$ be a random sample of size n taken from the exponential distribution given in (2.1). The MLE of μ is given by

$$\hat{\mu} = \bar{X}, \tag{2.2}$$

where $\bar{X}$ denotes the sample mean. The pivotal quantity $Q = 2n\bar{X}/\mu$ which follows a Chi-square distribution with $df = 2n$ can be used for testing the null hypothesis $\mu = \mu_0$ against the alternative hypothesis $\mu \neq \mu_0$. The null hypothesis is rejected with a test of size α if $Q_0 = 2n\bar{X}/\mu_0 < \chi^2_{(2n,\alpha/2)}$ or $Q_0 = 2n\bar{X}/\mu_0 > \chi^2_{(2n,1-\alpha/2)}$. Here $\chi^2_{(\nu,p)}$ denotes the *p*-th quantile of a Chi-square distribution with ν degrees of freedom. The sample mean is not a robust statistic, hence testing the mean of exponential distribution is strongly affected by presence of outliers.

3. Forward Search in Testing the Mean of an Exponential Distribution

The FS method is useful not only to detect and investigate observations that differ from the bulk of data, but also to analyse the effect of outliers on the estimation of parameters and other inferences about the model of interest. The FS method has three steps: the first step is choosing an outlier free subset of all observations, the second step presents the plan for progressing in FS and the last step is monitoring statistics during the search. In this paper we are inspired by quantile-quantile (QQ) plot to choose the initial subset and also add observations during the search according to their closeness to appropriate quantiles of standard exponential distribution. In the following subsections we address these three points separately.

3.1. Step 1: Choice of the initial subset

Starting point of the FS procedure is choosing an outlier free subset of observations robustly. A QQ plot is a common and basic technique used for finding a suitable model to data. When comparing observed data to a hypothesized distribution, the plot of the ordered observations versus the appropriate quantiles of assumed distribution, should look approximately linear. For more details about QQ plot see [7].

Let $\mathbf{x}_{(.)} = \left(x_{(1)}, x_{(2)}, \ldots, x_{(n)}\right)$ be the vector of ordered observations from an exponential distribution with mean μ. Then its *p*-quantile, defined by $P\left(X \leq x_p\right) = p$, is

$$x_p = -\mu ln(1-p) \tag{3.1}$$

Thus x_p is a linear function of $\alpha_i = -\ln(1-p_i)$. Also for appropriate $p_i = (i-0.5)/n$ one can view the *i*-th ordered sample $x_{(i)}$ as a good approximation for x_{p_i}. Therefore, the plot of $x_{(i)}$ against α_i should look approximately linear (a line without intercept). Thus we can estimate the unknown parameters μ by writing the following regression model

$$x_{(i)} = \mu\alpha_i + \varepsilon_i \tag{3.2}$$

The unknown parameter μ of model (3.2) can be estimated using robust regression estimation, for example Least Median of Squares (LMS), proposed by Rousseeuw [8]. Here we only discuss the LMS regression briefly. If $\boldsymbol{\beta} \in \mathrm{R}^p$ denotes the vector of parameters in the classical linear regression model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, i = 1, 2, \ldots, n, \tag{3.3}$$

where $y_i \in \mathrm{R}, \mathbf{x}_i \in \mathrm{R}^p$ and ε_i is the error term, then the LMS estimator for $\boldsymbol{\beta}$ is defined as

$$\hat{\boldsymbol{\beta}}_{\text{LMS}} = \min_{\hat{\boldsymbol{\beta}}} med\ e_i^2, \tag{3.4}$$

where e_i denotes the *i*-th residual

$$e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}, \quad i = 1,2,\ldots,n. \tag{3.5}$$

The resulting estimator has a 50% breakdown point. For an exhaustive account about this estimator see [9].

After estimating the parameter of model (3.2) by LMS estimation method, the estimated expected value for $x_{(i)}$ is of the form

$$\hat{x}_{(i)} = \hat{\mu}_{LMS}\alpha_i\ , \quad i = 1,2,\ldots,n. \tag{3.6}$$

Let $r_i = \left|x_{(i)} - \hat{x}_{(i)}\right|$, the *i*-th absolute residual resulting from (3.6). The elements of $\mathbf{x}_{(.)} = (x_{(1)}, x_{(2)}, \ldots, x_{(n)})$ are reordered based on the values of r_i, and this new vector of reordered observations is denoted by $\mathbf{x}_{(\text{LMS})}$.

To start the FS approach, the size of initial subset must be specified. The breakdown point of (3.4) is 50%, hence we start the process with the first $[(n+1)/2]$ observations of $\mathbf{x}_{(\text{LMS})}$. Denote this subset by $S^{(*)}$ that involved $[(n+1)/2]$ observations x_i that correspond to the smallest of $r_i = \left|x_{(i)} - \hat{x}_{(i)}\right|;\ i = 1,2,\ldots,n$.

The error terms of (3.2) are not independently and identically distributed. However no inferences are presented for the model or the parameter of (3.2), but the LMS estimator maybe have not a good performance for small data size. Hence, based on the asymptotic theory it is better we use the proposed method for large enough data sizes.

3.2. Step 2: Adding observations during the FS

At each step, the procedure adds to the subset the observation that is closest to the previously fitted model. Since we use a robust method for estimating the parameter of (3.2), it is not necessary to refit the model and reorder the observations $\mathbf{x}_{(\text{LMS})}$ at each step of the search. It means we just estimate the parameter of (3.2) based on the all observations and this parameter would not change during the search, hence we call this method as simplified version of FS method. Therefore in the $n - [(n+1)/2]$ remaining steps we add the next observation of $\mathbf{x}_{(\text{LMS})}$ to the previously chosen subset. Let $S^{(m)}$ be the subset of the first *m* observations of $\mathbf{x}_{(\text{LMS})}$. Thus $S^{(m)}$ involved m observations x_i that correspond to the smallest of $r_i = \left|x_{(i)} - \hat{x}_{(i)}\right|;\ i = 1,2,\ldots,n$.

3.3. Step 3: Monitoring the search

To guide the researcher in outlier detection and in the analysis of their effect on model inference, some statistics of interest must be monitored during the search. According to the Section 2, for testing the null hypothesis $\mu = \mu_0$ we must obtain Q statistic under the null hypothesis $Q_0 = 2n\bar{X}/\mu_0$. In the forward search version of this test, in each step of the search we obtain this statistic for each subset of the search. The forward search version of the test for the mean of an exponential distribution, $\mathbf{Q}_{FS}$, is defined as a collections of Q statistics computed for the subset $S^{(m)}$ during the FS procedure under the null hypothesis as

$$\mathbf{Q}_{FS} = \left(Q_{S^{(*)}}, \ldots, Q_{S^{(m)}}, \ldots, Q_{S^{(n)}}\right) \tag{3.7}$$

where $Q_{S^{(m)}} = 2m\bar{X}_{S^{(m)}} / \mu_0$ and $\bar{X}_{S^{(m)}}$ is the sample mean of the subset $S^{(m)}$.

It is obvious that the size of (3.7) is dependent to the data size n. The quantiles of the test statistic (3.7) can be estimated by simulation in all steps of the procedure by generating numerous samples in size n from a standard exponential distribution. By simulating 10000 samples from the null hypothesis (or without lack of generality we can suppose $\mu_0 = 1$) and then by applying the FS method, we can obtain (3.7) for each sample. Now we have 10000 values for $\mathbf{Q}_{FS}$ under the null hypothesis, therefore the α-th quantile of $Q_{S^{(m)}}$ is $(10000 \times \alpha)$-th value of sort $Q_{S^{(m)}}$. We denote the empirical α-th quantile of Q for the subset $S^{(m)}$ by $\hat{q}_{(m,\alpha)}$. In any search steps the acceptance region lies between the chosen estimated quantiles, for example $\hat{q}_{(m,\alpha=0.025)}$ and $\hat{q}_{(m,\alpha=0.975)}$. Hence it is possible to determine from which observation onwards the null hypothesis is rejected.

4. Simulation Study

To evaluate the proposed statistic (3.7), we conduct simulation studies that aim to consider the behavior of this statistic in the presence of outliers and ability of FS to detect them. Table 1 reports the 2.5% and 97.5% empirical quantiles of Q at each step of the search estimated by generating 10000 samples from a standard exponential distribution with size $n = 100$.

Table 1. Empirical quantiles of Q statistics at each step of the search for $n = 100$.

m	$\hat{q}_{(m,\alpha=0.025)}$	$\hat{q}_{(m,\alpha=0.975)}$	m	$\hat{q}_{(m,\alpha=0.025)}$	$\hat{q}_{(m,\alpha=0.975)}$	m	$\hat{q}_{(m,\alpha=0.025)}$	$\hat{q}_{(m,\alpha=0.975)}$
50	56.85	183.55	67	76.47	176.13	84	108.58	184.45
51	57.01	182.23	68	77.86	176.33	85	111.00	186.10
52	57.70	182.30	69	79.45	176.11	86	113.17	187.73
53	58.52	181.58	70	80.83	176.03	87	116.07	189.33
54	59.43	181.23	71	82.37	176.25	88	118.39	191.57
55	60.54	180.42	72	84.30	176.22	89	121.53	193.70
56	61.35	179.53	73	86.03	177.10	90	124.23	196.30
57	62.54	179.00	74	88.01	177.63	91	127.26	199.00
58	63.77	178.74	75	90.07	177.95	92	130.20	201.62
59	65.03	178.19	76	92.20	178.60	93	133.17	204.71
60	66.15	177.65	77	94.10	179.08	94	136.63	208.07
61	67.49	176.92	78	96.06	179.65	95	139.92	211.93
62	69.03	176.68	79	97.94	180.16	96	144.08	215.81
63	70.39	176.31	80	99.70	181.15	97	148.13	220.47
64	71.76	176.49	81	102.15	181.82	98	152.33	225.70
65	73.51	176.33	82	104.27	182.97	99	157.18	231.43
66	75.03	176.31	83	106.34	183.66	100	163.57	240.11

Now consider four samples which are generated in the following way:

- Sample A: 100 observations are generated from an $Exp(\mu = 1)$.
- Sample B: 95 observations are generated from an $Exp(\mu = 1)$ and for contamination 5 observations are generated from a $N(\mu = 10, \sigma^2 = 1)$.
- Sample C: 90 observations are generated from an $Exp(\mu = 1)$ and for contamination 10 observations are generated from a $Uniform(a = 5, b = 6)$.
- Sample D: 90 observations are generated from an $Exp(\mu = 1)$ and for contamination 10 observations are generated from an $Exp(\mu = 5)$.

For each sample, the proposal is to test the null hypothesis $\mu = 1$ against the alternative hypothesis $\mu \neq 1$. In Fig.1, values of $\mathbf{Q}_{FS}$ during the search are plotted for samples A-D and compared with corresponding 2.5% and 97.5% quantiles (dashed lines) of its distribution obtained from the simulation study with clean data. The null hypothesis is accepted in each step of the search for clean sample A, but it is rejected after entrance of outliers in the last steps for contaminated samples B, C. In case D the null hypothesis is rejected just from step 96 onwards, may be this is due to the similarity between the clean data distribution and contaminated data distribution.

4.1. Empirical power of $\mathbf{Q}_{FS}$

In this subsection, our interest is to evaluate the empirical power of our approach. Consider testing the null hypothesis $\mu = 1$ against the alternative hypothesis $\mu \neq 1$. Fig.2 shows the empirical power of $\mathbf{Q}_{FS}$ against following alternative hypotheses by generating 10000 samples of size 100.

(a) $Exp(\mu = 0.7)$
(b) $Exp(\mu = 0.5)$
(c) $Exp(\mu = 1.3)$
(d) $Exp(\mu = 1.5)$

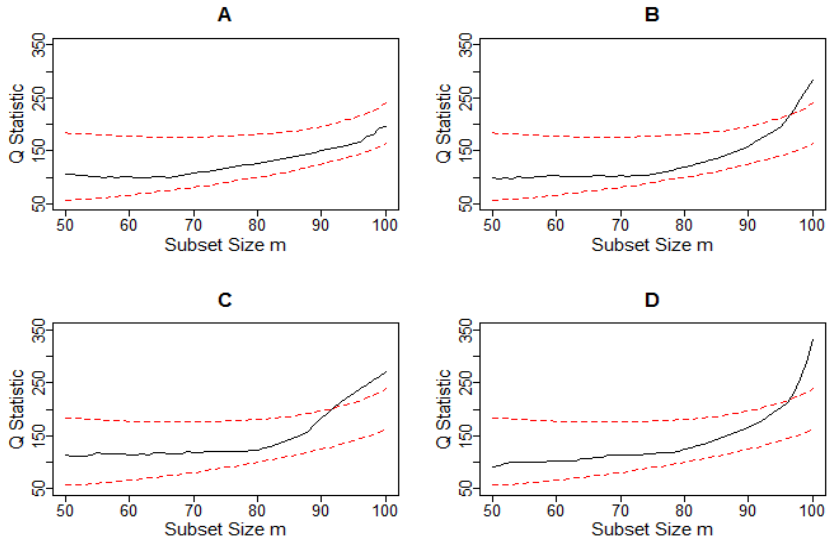

Fig.1. Forward plots of $\mathbf{Q}_{FS}$ during the search for samples A-D.

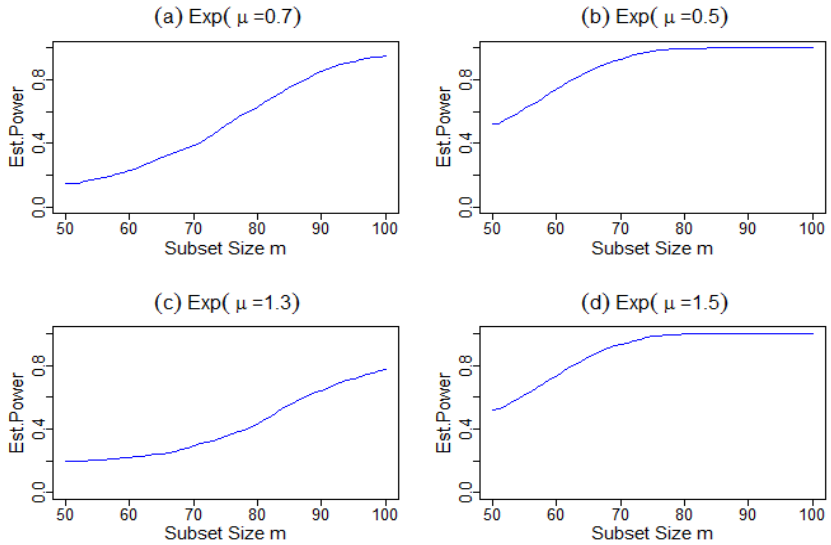

Fig.2. Empirical power of $\mathbf{Q}_{FS}$ versus alternative distributions (a-d).

Due to our aim is to find the largest subset of observations that can be distributed as the null hypothesis, the power of our proposed procedure in the first steps of the search is low. It means in the first step, procure choose the best subset of observations that can be generated from the null hypothesis although all dataset is generated from an alternative distribution. Thus the minimum power is always in the first step and it is increasing as the subset size increase. Therefore, the larger sample size provides safer procedure to detect and investigate the effect of outliers.

5. Concluding Remarks

In this paper, a new robust method for testing the mean of an exponential distribution has been presented. The approach gives information about the mean of majority of the data and the percentage of contamination. At every step of the FS, the proposed statistic is computed and with a graphical approach a cut-off point divides the group of outliers from the other observations. In order to illustrate the application and the advantage of the FS approach we used some artificial examples.

References

1. Atkinson, A. C., Riani, M., *Robust Diagnostic Regression Analysis*. Springer, New York, 2000.
2. Atkinson, A. C., Riani, M., Forward search added-variable t-tests and the effect of masked outliers on model selection, *Biometrika*. **89**(4) (2002), 939–946.
3. Atkinson, A. C., Riani, M., The forward search and data visualization, *Comp. Stat*. **19** (2004), 29-54.
4. Atkinson, A. C., Riani, M. and Cerioli, A., The forward search: theory and data analysis, *J. Korean. Stat. Soc*. **39** (2010), 117–134.
5. Bertaccini, B., Varriale, R., Robust analysis of variance: An approach based on the forward search, *Comp. Statist. Data. Anal*. **51** (2007), 5172-5183.
6. Coin, D., Testing normality in the presence of outliers, *Statist. Meth. Appl*. **17** (2008), 3-12.
7. D'Agostino, R. B., Stephens, M. A. (eds.), *Goodness of Fit Techniques*. Marcel Dekker, New York, 1986.
8. Rousseeuw, P. J., Least median of squares regression, *J. Am. Statist. Ass*. **79** (1984), 871–880.
9. Rousseeuw, P. J., Leroy, A. M., *Robust regression and outlier detection*. Wiley, New York, 1987.