Comments on "Detecting Outliers in Gamma Distribution" by M. Jabbari Nooghabi et al. (2010)

M. Magdalena Lucini* Alejandro C. Frery†

September 14, 2015

This note shows that the results presented by Jabbari Nooghabi et al. (2010) do not hold in all expected cases. With this, the technique proposed by Kumar and Lalhita (2012) for detecting upper outliers in Gamma samples is also not valid. Specifically, this note shows that the probability density functions (pdf) under the null hypothesis of the test statistics therein proposed are not always valid.

In the aformentioned works the authors propose test statistics to detect outliers in Gamma samples using a *test of discordancy for outliers* framework as defined in Barnett and Lewis (1994).

Following the approach of Barnett and Lewis (1994), the null hypothesis (H_0) of a test for discordancy is a statment of an initial probability model that explains the data generating process. For instance, in the case here considered, H_0 states that data are generated as independent observations from a common distribution F. If F is a Gamma distribution, as in Jabbari Nooghabi et al. (2010) and Kumar and Lalhita (2012), then $H_0\colon X_1, X_2, \ldots, X_n$ are n independent random variables, each following a Gamma distribution with shape parameter $m > 0$ and scale parameter $\sigma > 0$, denoted by $\Gamma(m, \sigma)$, whose probability density function (pdf) is given by

$$f(x; m, \sigma) = \frac{1}{\Gamma(m)\sigma^m} x^{m-1} \exp\left(-\frac{x}{\sigma}\right), x > 0.$$

As σ is a scale parameter, without loosing generality, it will be assumed from now on that these random variables are distributed according to a $\Gamma(m, 1)$ law, that is, with pdf given by

$$f(x; m) = \frac{1}{\Gamma(m)} x^{m-1} \exp(-x), x > 0.$$

The alternative hypothesis used in Jabbari Nooghabi et al. (2010) and Kumar and Lalhita (2012) is the *slippage* alternative.

We are interested in detecting $1 \leq k < n$ upper outliers using Z_k, the statistic proposed by Kumar and Lalhita (2012). This statistic, after some computations, can be written as

$$Z_k = \frac{\sum_{j=n-k+1}^{n} (n-j+1) Y_j}{\sum_{j=2}^{n} (n-j+1) Y_j}, \tag{1}$$

where

$$Y_j = X_{(j)} - X_{(j-1)}, \tag{2}$$

*Facultad de Ciencias Exactas, Naturales y Agrimensura, Universidad Nacional del Nordeste, Av. Libertad 5460, 3400 Corrientes, and CONICET, Argentina, magdalenalucini@gmail.com

†Alejandro C. Frery is with the LaCCAN, Universidade Federal de Alagoas, Av. Lourival Melo Mota, s/n, 57072-900 Maceió – AL, Brazil, acfrery@gmail.com

$X_{(j)}$ denotes the j-th order statistics of the ordered sample from $(X_i)_{1\leq i\leq n}$ in nondecreasing order, that is, $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$, and k is the number of observations suspected to be upper outliers.

As in any statistical test, once the test statistic is proposed we need to determine rejection criteria related to a previously specified significance level. To do that, and to compute the p-value associated to a sample, the distribution of the test statistic under the null hypothesis must be known.

In Kumar and Lalhita (2012) the distribution of Z_k under the null hypothesis was obtained based, mainly, on the distribution of differences of subsequent order statistics from Gamma random variables, i.e., the distribution of the Y_j given in Eq. (2). However, when performing simulations we observed that the empirical pdf of Z_k under the null hypothesis given by Kumar and Lalhita (2012) gave a proper adjustment only for $m = 1$, that is, when the random variables $(X_i)_{1\leq i\leq n}$ follows an exponential law.

Jabbari Nooghabi et al. (2010) also used the random variables Y_j to find the pdf of the test statistic they proposed under the alternative (Theorem 3.1) and null (Corollary 3.1) hypotheses. Kumar and Lalhita (2012), followed the very same reasoning and methodology used in Theorem 3.1 of Jabbari Nooghabi et al. (2010) to derive the pdf of Z_k under the null hypothesis.

A strong assumption made in both works is that, under the null hypothesis, each Y_j follows a $\Gamma(m, (n-j+1)^{-1})$ distribution. This is not true when $m \neq 1$, as we show in what follows.

Recall that under the null hypothesis of a test for discordancy, $X_1, \ldots, X_n$ are independent identically distributed Gamma random variables. In general, if $X_1, \ldots, X_n$ are independent identically distributed random variables the pdf of $Y_{sr} = X_{(s)} - X_{(r)}$ can be found by solving the following integral (David and Nagaraja, 2003):

$$\begin{aligned} f_{Y_{sr}}(y) =& \frac{n!}{(r-1)!(s-r-1)!(n-s)!} \\ & \int_{-\infty}^{\infty} F^{r-1}(x) f(x) [F(x+y) - F(x)]^{s-r-1} f(x+y) [1 - F(x+y)]^{n-s} dx, \end{aligned} \tag{3}$$

where F and f are the cumulative distribution function and the pdf, respectively, of any of the X_i (without sorting).

Replacing s by j and r by $j-1$ in Eq. (3), the pdf of $Y_j = X_{(j)} - X_{(j-1)}$ can be found by solving the following integral

$$f_{Y_j}(y) = \frac{n!}{(j-2)!(n-j)!} \int_{-\infty}^{+\infty} F^{j-2}(x) f(x) f(x+y) [1 - F(x+y)]^{n-j} dx. \tag{4}$$

Let us suppose that the sample is only composed by two random variables X_1 and X_2, each $\Gamma(m, 1)$ distributed with shape parameter $m \in \mathbb{N}$. Then $n = 2$, and we just have to compute $Y_2 = X_{(2)} - X_{(1)}$. Making $n = 2$, and $j = 2$ in Eq. (4), and having in mind than $m \in \mathbb{N}$, after some computations (see Appendix) the pdf of Y_2 can be written as

$$f_{Y_2}(y) = \frac{\exp(-y)}{\Gamma^2(m)} \sum_{i=0}^{m-1} \binom{m-1}{i} \frac{\Gamma(2m-i-1)}{2^{2(m-1)-i}} y^i, y > 0. \tag{5}$$

As already mentioned, a strong assumption made by Jabbari Nooghabi et al. (2010) and by Kumar and Lalhita (2012) is that if X_1, X_2 are random variables distributed according to a $\Gamma(m, 1)$ law then $Y_2 \sim \Gamma(m, 1)$. But, if for instance $m = 2$ and using Eq. (5),

the pdf f_{Y_2} can be expressed as

$$f_{Y_2}(y) = \frac{1}{2}\big(\exp(-y) + y\exp(-y)\big), y > 0. \tag{6}$$

This is a composition of a $\Gamma(1,1)$ and a $\Gamma(2,1)$ distributions with same probability, and not a $\Gamma(2,1)$ distribution as claimed by both Jabbari Nooghabi et al. (2010) and by Kumar and Lalhita (2012). The discrepancy is notorious, as will be shown henceforth.

Algorithm 1 presents the pseudocode used for the discussion. We implemented it in the `R` programming language R Core Team (2014), and run it with $R = 10000$ replications for each case of $m \in \{1, 3, 8\}$.

Algorithm 1: Pseudocode for the analysis of Y_2.

```
Data: Read m, R, and the pseudorandom number generator seed.
Initialize Z of length R;
Initialize r = 1;
for 1 ≤ r ≤ R do
    Obtain X = (X_1, X_2) from the Γ(m, 1) distribution;
    Sort X and obtain X_(1) ≤ X_(2);
    Compute Y_2 = X_(2) − X_(1);
    Store Z(r) = Y_2;
    Update r = r + 1;
Analyze Z;
```

Figure 1 presents the results obtained with this simulation with $R = 10^4$: histograms of Y_2 and the densities proposed by Jabbari Nooghabi et al. (2010) and Kumar and Lalhita (2012) (dashed lines), and the one we obtained and presented in Eq. (5) (solid lines).

(a) $X_i \sim \Gamma(1,1)$ (b) $X_i \sim \Gamma(3,1)$ (c) $X_i \sim \Gamma(8,1)$

Figure 1: The pdf of Y_2 assumed by Jabbari Nooghabi et al. (2010) and by Kumar and Lalhita (2012) in dashed lines, and in solid lines the pdf given in Eq. (5)

Both densities coincide in the case $m = 1$, i.e., when X_1, X_2 follow unitary mean Exponential distributions; cf. Fig. 1(a). Figures 1(b) and 1(c) show the discrepancy between the observed data and the model claimed by Kumar and Lalhita (2012). The data is well fit by the distribution we obtained, though.

Conclusions

In this work we have shown that if X_1, X_2 are independent random variables, each $\Gamma(m,1)$ distributed, with $m \in \mathbb{N}_{\geq 2}$, then the pdf of $Y_2 = X_{(2)} - X_{(1)}$ is a composition of m Gamma distributions, and not a $\Gamma(m,1)$ law as claimed by Jabbari Nooghabi et al. (2010) and then assumed by Kumar and Lalhita (2012). Therefore, with this counterexample we conclude that if $m \in \mathbb{N}_{\geq 2}$ then Y_j, as in Eq. (2), does not follow a Gamma distribution. This implies that most computations presented by Jabbari Nooghabi et al. (2010) and by Kumar and Lalhita (2012) are not valid, including the pdf of Z_k given by Kumar and Lalhita (2012).

Appendix

From Eq. (4)

$$\begin{aligned}
f_{Y_2}(y) &= \frac{2!}{(2-2)!(2-2)!}\int_0^{+\infty} F^{2-2}(x)f(x)f(x+y)[1-F(x+y)]^{2-2}dx \\
&= 2\int_0^{+\infty} f(x)f(x+y)dx \\
&= \frac{2}{\Gamma^2(m)}\int_0^{+\infty} \exp(-x)x^{m-1}\exp(-(x+y))(x+y)^{m-1}dx \\
&= \frac{2\exp(-y)}{\Gamma^2(m)}\int_0^{+\infty} \exp(-2x)(x^2+xy)^{m-1}dx.
\end{aligned}$$

Having in mind that $m \in \mathbb{N}$, expanding the binomial $(x^2+xy)^{m-1}$ and using that $\int_0^{+\infty} x^n e^{-ax}dx = a^{-(n+1)}\Gamma(n+1)$, follows that

$$\begin{aligned}
f_{Y_2}(y) &= \frac{2\exp(-y)}{\Gamma^2(m)}\left\{\int_0^{+\infty} \exp(-2x)\left[\sum_{i=0}^{m-1}\binom{m-1}{i}y^i x^{(2(m-1)-j)}\right]dx\right\} \\
&= \frac{2\exp(-y)}{\Gamma^2(m)}\left\{\sum_{i=0}^{m-1}\binom{m-1}{i}y^i\left[\int_0^{+\infty} x^{(2(m-1)-j)}\exp(-2x)dx\right]\right\} \\
&= \frac{\exp(-y)}{\Gamma^2(m)}\left[\sum_{i=0}^{m-1}\binom{m-1}{i}\frac{\Gamma(2m-i-1)}{2^{2(m-1)-i}}y^i\right].
\end{aligned}$$

Incidentally, the expression given for the Dixon's D_k statistic by both the articles commented in this work are wrong. They state that $D_k = (X_{(n)} - X_{(n-k)})/X_{(n)}$ when, in fact, it is

$$D_k = \frac{X_{(n)} - X_{(n-k)}}{X_{(n)} - X_{(1)}},$$

the ratio of the gap to the range of the data.

References

Barnett, V. and T. Lewis (1994), *Outliers in Statistical Data*, third edition. John Wiley and Sons.

David, H.A. and H.N. Nagaraja (2003), *Order Statistics*, third edition. John Wiley and Sons.

Jabbari Nooghabi, M., H. Jabbari Nooghabi, and P. Nasiri (2010), "Detecting outliers in gamma distribution." *Communications in Statistics-Theory and Methods*, 39, 698–706.

Kumar, N. and S. Lalhita (2012), "Testing for upper outliers in gamma sample." *Communications in Statistics-Theory and Methods*, 41, 820–828.

R Core Team (2015), *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.