Gradient Statistic: An option for conducting hypothesis testing in small sample size scenarios

Sébastien Lozano Forero, Vladimir Ballesteros Ballesteros, and Jorge Luis Nisperuza Toledo
Fundación Universitaria los Libertadores, Facultad de Ingeniería y Ciencias Básicas, Bogotá, Colombia.

Abstract: The gradient statistic, recently proposed in the literature, has gained attention on statistical practitioners due to the fact that it is a competitive alternative for the traditional statistics (likelihood ratio, Wald, Score) for performing hypothesis testing in parametric models with the special property of being easily computable. In this work, we present an exhaustive Monte Carlo simulation study with the four test statistics to assess its performance in a controlled scenario based on the exponential distribution. The obtained results suggest that the gradient test statistics is, indeed, a competitive option for the "Holy trinity" of statistical inference (Likelihood ratio, Score and Wald test statistics).

AMS Subject classification:
Keywords: hypothesis testing, gradient statistic, power of test and exponential distribution.

OVERVIEW

Initially, let us assume that, based in a parametric model, the following hypothesis testing should be performed:

\[ \begin{align*}
H_0 & : \theta = \theta_0, \\
H_1 & : \theta \neq \theta_0
\end{align*} \]

where \( \theta_0 \) is a specified vector and \( \theta \) is the parametric vector indexing the initial parametric model.

The tests based on large samples approximations are often used in statistics, due to the fact that exact tests are not always available. These tests are called "first-order asymptotic", that is, they are based on critical values obtained from a known null limit distribution. A natural problem that arises is whether the first order approximation is adequate for the null distribution of the test statistic under consideration. The best known asymptotic test statistics whose reference distributions is chi-square are: likelihood ratio, score and Wald (often referred as the "holy trinity" of statistical inference). The statistics of these three tests are equivalent in large sample sizes and, in regular problems, converge under the null hypothesis \( H_0 \), to the distribution \( \chi_q^2 \), where \( q \) is the number of restrictions imposed by \( H_0 \).

In small samples, the first-order approximation may not be satisfactory and may lead to quite distorted null hypothesis rejection rates. The statistics involved and which will be considered are: Likelihood ratio, Wald and Score. The main idea is to test the hypothesis \( H_0 \). According to the notation used in Lemonte (2016), the Likelihood ratio, Wald and Score test statistics are, respectively, defined as:

\[ \begin{align*}
S_{LR} &= 2[\ell(\hat{\theta}) - \ell(\theta_0)], \\
S_W &= (\hat{\theta} - \theta_0)^T K(\hat{\theta})(\hat{\theta} - \theta_0), \\
S_R &= U(\theta)^T K(\theta_0)^{-1} U(\theta_0)
\end{align*} \]

where \( \hat{\theta} \) is the maximum likelihood estimator of \( \theta \) which may be obtained from \( U(\hat{\theta}) = 0 \) where \( U(\theta) \) is the score vector defined as

\[ U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} \]

and \( \ell(\theta) \) is defined as the log-likelihood function, given by

\[ \ell(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta) \]

where \( f(. \mid \theta) \) is the probability function of the random sample \( (x_1, \cdots, x_n)^T \).
The Fischer information matrix is a way of measuring the amount of information a random variable contains over an unknown parameter vector and upon which the distribution depends. It is defined as

\[ K(\theta) = E[\mathbf{U}(\theta)\mathbf{U}(\theta)^T] = -E\left( \frac{\partial \mathbf{U}(\theta)}{\partial \theta^T} \right) \]

Is now important to have tools to better address the problem of small-sample test statistics. A first attempt is given by a modified version of the Wald statistic, which is defined as

\[ S_{WM} = (\hat{\theta} - \theta_0)^T K(\theta_0)(\hat{\theta} - \theta_0) \]

where a correction on the quantity that modifies the difference between the estimator and the parameter \( \theta_0 \), is made. This is, the only difference between \( S_W \) and \( S_{WM} \) is the object to be evaluated in \( K(\theta) \)

\[ S_W \rightarrow K(\hat{\theta}), \quad S_{WM} \rightarrow K(\theta_0) \]

The test statistic \( S_{LR}, S_W, S_R \) e \( S_{WM} \) have an approximate chi-square distribution (central) with \( p \) degrees of freedom \( (\chi^2_p) \) and, under null hypothesis \( H_0 : \theta = \theta_0 \), we reject \( H_0 \) if the observed value of the test statistic exceeds the quantile 100(1 - \( \alpha \))% of the distribution \( \chi^2_p \), where \( \alpha \) is the nominal level of the test.

The proposal of Terrell (2002) is widely known as Gradient Statistics can be obtained, as a result of the previous computations, as follows:

**Definition 1.** The Gradient Statistic, \( S_T \), to test the simple null hypothesis \( H_0 : \theta = \theta_0 \) against \( H_1 : \theta \neq \theta_0 \) has the form

\[ S_T = \mathbf{U}(\theta_0)^T (\hat{\theta} - \theta_0) \]

Note that this statistic is derived from the Rao statistic and modified Wald statistic (see Lemonte (2010)). The following are some important properties of this new statistic

- Under \( H_0 \), \( S_T \) has an approximate distribution \( \chi^2_p \).
- \( S_T \) is very simple to calculate, not involving estimation of the information matrix nor the calculation of its inverse.
- It is not evident that \( S_T \) is nonnegative, but it must be asymptotically nonnegative.

**COMPOSED NULL HYPOTHESIS**

Now, we take a quick look at the case where the null and compound hypothesis. First, partition the vector \( \theta = (\theta_1, \ldots, \theta_p)^T \) such that

\[ \theta = (\theta^T_1, \theta^T_2)^T \]

where \( \theta_1 = (\theta_1, \ldots, \theta_q)^T \) and \( \theta_2 = (\theta_{q+1}, \ldots, \theta_p)^T \). Now, consider to test \( H_0 : \theta_2 = \theta_{2,0} \) against \( H_0 : \theta_2 \neq \theta_{2,0} \) em que \( \theta_{2,0} \) is a vector of known constants. To test this hypothesis, one can use the statistics \( S_{LR}, S_W, S_R, S_T \) given by

\[ S_{LR} = 2[\ell(\theta) - \ell(\hat{\theta})], \quad S_W = (\hat{\theta} - \tilde{\theta})^T K(\tilde{\theta})(\hat{\theta} - \tilde{\theta}) \]
\[ S_R = \mathbf{U}(\tilde{\theta})^T K(\tilde{\theta})^{-1} \mathbf{U}(\hat{\theta}) \]
\[ S_T = \mathbf{U}(\tilde{\theta})^T (\hat{\theta} - \tilde{\theta}) \]

where \( \hat{\theta} \) and \( \tilde{\theta} \) are the unrestricted maximum likelihood estimators (under \( H_1 \)) and restricted (under \( H_0 \)) of \( \theta \) respectively. All the statistics \( S_{LR}, S_W, S_R \) and \( S_T \) follow approximately \( \chi^2_{p-q} \) distribution under \( H_0 \).

Now, there are natural questions regarding the performance of the statistics under study. The first fundamental fact to be solved is to establish when the gradient statistic
can be considered in relation to the classic test statistics. We have the following facts

- The simple form of $S_T$, which in practice can be the simplest to calculate, is an interesting feature.
- In complex problems, not having to calculate, estimate and invert a Fischer information matrix is at an even more positive point.
- For example, problems in survival analysis in which there is censorship, the $S_T$ statistic could be used without problems and thus would be an alternative to the likelihood ratio statistic.

There is still the natural question of the will to compare the proposed statistics. We want to know if $S_T$ is more, less or equally powerful than the other statistics. In order to answer, we will study the local power of the gradient test. According to Lemonte (2013), the following strategy was proposed:

1. to present the asymptotic expansion of the local power function (up to order $n^{-1/2}$) of the gradient test under the sequence of alternative hypotheses

$$H_{1n} : \theta_2 = \theta_{20} + \frac{\epsilon}{\sqrt{n}}$$

where $\epsilon = \sqrt{n}(\theta_2 - \theta_{20}) = (\epsilon_{q+1}, \ldots, \epsilon_p)^T$

2. Make a local power study of the gradient test by comparing it with the local power of the likelihood ratio tests, Wald and score.

Some math involved due to Lemonte (2013):

- Derivatives of the log-likelihood function

$$y_r = n^{-1/2} \frac{\partial \ell}{\partial \theta_r}, \quad y_{rs} = n^{-1} \frac{\partial^2 \ell}{\partial \theta_r \partial \theta_s},$$

$$y_{rst} = n^{-3/2} \frac{\partial^3 \ell}{\partial \theta_r \partial \theta_s \partial \theta_t},$$

with $r, s, t = 1, \ldots, p$

- The arrays

$$y = (y_1, \ldots, y_p)^T, \quad Y = ((y_{rs}), \quad Y_{rs} = ((y_{rst})).$$

- Cumulants

$$\kappa_{rs} = E(y_{rs}), \quad \kappa_{r,s} = E(y_{r}y_{s}), \quad \kappa_{rst} = \sqrt{n}E(y_{rst}),$$

$$\kappa_{r,s,t} = \sqrt{n}E(y_{r}y_{s}y_{t})$$

com $r, s, t = 1, \ldots, p$.

- The respective arrays

$$K = ((\kappa_{r,s})) = -((\kappa_{s,r})), \quad K_{rs} = ((\kappa_{rst})), \quad K_{r,s} = ((\kappa_{rst})).$$

- For quantities with three indexes, the adopted notation would be:

$$K_{rs} \circ a \circ b \circ c = \sum_{r,s,t=1}^{p} \kappa_{rst}a_r b_s c_t$$

where $M = ((m_{rs}))$ is a matrix $p \times p$ and $a, b, c$ are vectors $p-$dimensional.

- Define the matrices:

$$K = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}, \quad A = \begin{bmatrix} K_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$$

$$M = K^{-1} - A, \quad K_{21,1} = K_{22} - K_{21} K_{11}^{-1} K_{12}$$

- The vector

$$\epsilon^* = \begin{bmatrix} K_{11}^{-1} K_{12} \\ I_{p-q} \end{bmatrix} \epsilon$$

where $I_{p-q}$ is the identity matrix of order $p - q$.

**MOMENT GENERATING FUNCTION**

The moment generating function can be rewritten as

$$M(t) = (1 - 2t)^{1/2} \epsilon^T K_{22,1} t + O(n^{-1})$$

where $a_0 = -(a_1 + a_2 + a_3)$ and

$$a_1 = \frac{1}{4} \{K^\dagger \circ (K^{-1})^\dagger \circ (\epsilon^*)^\dagger - 2(K_{rs} + 2K_{rs})^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger$$

$$- (4K_{rs} + 3K_{rs})^\dagger \circ A^\dagger \circ (\epsilon^*)^\dagger$$

$$- 2(K_{rs} + 2K_{rs})^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger \}$$

$$a_2 = \frac{1}{4} \{K^\dagger \circ M^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger$$

$$- (4K_{rs} + 3K_{rs})^\dagger \circ A^\dagger \circ (\epsilon^*)^\dagger$$

$$- 2(K_{rs} + 2K_{rs})^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger \}$$

$$a_3 = \frac{1}{12} \{K^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger \circ (\epsilon^*)^\dagger$$

Referring the function $M(t)$, the following result is derived.
The main idea now is to be able to compare the podes of the tests that are being studied. The following facts held
1. Until the first order, the statistics $S_{LR}, S_{W}, S_R$ and $S_T$ have the same asymptotic properties under a null hypothesis $H_0$, or under an alternative local hypothesis.
2. Up to an error of order $n^{-1}$, the corresponding tests have the same size, however, their powers differ by the order term $n^{-1/2}$.
3. Thus the powers of the different tests can be compared on the basis of the expansions of their power functions ignoring lesser terms of order than $n^{1/2}$.

The local power function of the tests that use statistics is defined $S_{LR}, S_{W}, S_R \in S_T$ as
$$
P_i = 1 - P(S_i \leq x_γ) = P(S_i > x_γ), \quad i = LR, W, R, T$$
where $x_γ$ represents the proper quantile of the $\chi^2_{p-q}$ distribution for a choose nominal level $γ$ and
$$
P(S_i \geq x_γ) = G_{p-q,λ}(x_γ) + \frac{1}{\sqrt{n}} \sum_{k=0}^{3} a_k G_{p-q+2k,λ}(x_γ) + O(n^{-1})$$

**Simulation Results**

Until this point, the different test statistics have been presented and, as mentioned before, there may be differences in the powers of the four tests in small sample sizes, which is the reason the present simulation is made, to make a comparison between the powers of the tests that are generally used (Likelihood ratio, Score and Wald) and the statistical gradient for testing as well. Simulations are made via Monte Carlo for the exponential distribution function in the following form:

$$f_X(x \mid \theta) = \theta e^{-x\theta}$$

with $\theta > 0$ and $x > 0$. The corresponding log-likelihood function is given by:

$$\ell(\theta) = \log (\prod_{i=1}^{n} f_X(x_i \mid \theta)) = \log (\theta^n e^{\theta \sum_{i=1}^{n} x_i})$$

We have that

$$\ell(\theta) = -n \log (\theta) + \theta \sum_{i=1}^{n} x_i$$

To obtain the maximum likelihood estimator of $\theta$, we make

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = -\frac{n}{\theta} + \sum_{i=1}^{n} x_i = 0$$

this way, we obtain the following estimator

$$\hat{\theta} = \frac{1}{\bar{X}}$$

Since some of the test statistics depend on the Fisher information matrix, this corresponds to:

$$\frac{\bar{X} t}{t} - \sum_{i=0}^{n} x_i$$
where \( n \) is the sample size.

We fixed the number of Monte Carlo replicas in 10,000. In each Monte Carlo simulation, samples for different sample sizes (10, 30, 50 and 100) and different levels of significance

\[
(\alpha = 0.1, \alpha = 0.05, \alpha = 0.01)
\]

were generated in order to have the value of the parameter \( \theta = 1 \) as the value of the real parameter and for the generation of the samples.

To calculate the power of the test at each step the sample was generated with the actual value \( \theta \) and the hypothesis \( \theta = \theta_0 \) against \( \theta \neq 0 \) was tested. For different values of \( \theta \) at the end the percentage of times that was rejected test will be the measure of the power in each case.

RESULTS

For a sample size of 10 and for values lower than 1, the gradient statistic was found to be more potent, even than

\[
\text{Figure 1. Power of the considered test statistics with } n = 10 \text{ and } \alpha = 0.01
\]

the likelihood ratio test statistic. Although the differences in powers are not very large and all of them tend to have the same behavior from the same point. These behaviors were observed at the three levels of significance, these results can be seen in the table 13 and in the figures 1, 2, 3.

For a sample size of 30 and for all values of \( \theta_0 \), the statistics tend to have the same behavior, the three present values close to their power values and in this case the curves
are faster than in the case. The results of this study are shown in the table below and in the graphs in the following table: 14 and figures 4, 5, 6.

For sample sizes of 50 and 100, the power curves are much faster growing, that is to say that the tests are much more sensitive to small changes in the value of \( \theta \) real and \( \theta_0 \), besides we see that in these large sample sizes, the differences between the four tests are really minimal, so it makes no difference in how much power in the exponential distribution use any of the four tests. This can be observed in tables 15 and 16, and figures 7, 8, 9, 10, 11, 12.

In summary, we conclude, via Monte Carlo simulations, that for small sample sizes differing in the power of the test is influenced by the true value of the parameter and that on average the behavior of the four tests exposed is the same, in addition we see that for large sample sizes the
FIGURE 12. Power of the considered test statistics with \( n = 100 \) and \( \alpha = 0.1 \)

<table>
<thead>
<tr>
<th>Significance level</th>
<th>Value of ( \theta )</th>
<th>Wald</th>
<th>Score</th>
<th>RV</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>0.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>71.5</td>
<td>71.5</td>
<td>78.7</td>
<td>81.8</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>30.6</td>
<td>30.6</td>
<td>25.1</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>98.3</td>
<td>98.3</td>
<td>97.9</td>
<td>97.2</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

FIGURE 13. Behaviour of the considered test statistics with \( n = 10 \)

FIGURE 14. Behaviour of the considered test statistics with \( n = 30 \)

and by both asymptotically these tests are equivalent. As the gradient statistic is much simpler in its calculation it is an attractive alternative since it is equally powerful for the three tests that are commonly used and does not need a complex evaluation as to its calculation.

CONCLUSIONS

- All four considered test statistics are locally unbiased
- If \( K_0 = 0 \) the likelihood ratio, Wald and gradient tests have identical local power properties.
- If \( K_0 = 2K_0 \), the Score and gradient test statistics have identical local power properties.
<table>
<thead>
<tr>
<th>Significance level</th>
<th>Value of θ</th>
<th>Size 50</th>
<th>Value of θ</th>
<th>Size 100</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Wald</td>
<td>Score</td>
<td>RV</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gradient</td>
</tr>
<tr>
<td>10%</td>
<td>0.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>91.7</td>
<td>91.7</td>
<td>94.1</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>43.3</td>
<td>43.3</td>
<td>39.3</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>100</td>
<td>100</td>
<td>99.7</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>99.7</td>
<td>99.7</td>
<td>99.7</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>61.4</td>
<td>61.4</td>
<td>58.2</td>
</tr>
<tr>
<td></td>
<td>1.3</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1.6</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>1.9</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>2.2</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**FIGURE 15.** Behaviour of the considered test statistics with $n = 50$

- There is no uniform superiority of one test statistic over others.

**REFERENCES**


