BIG DATA AND THE CENTRAL LIMIT THEOREM: A STATISTICAL LEGEND

S. Allende-Alonso*, C. N. Bouza-Herrera*, S. E. H. Rizvi** and J. M. Sautto-Vallejo***
*Universidad de La Habana
**Agricultural Statistics, SKUAST- India
***Universidad Autónoma de Guerrero

ABSTRACT
Nowadays we deal with Big-Data commonly. The users of statistics rely on having a large sample size n for using the statistical methods based on normality. Usual inference methods are typically based on considering the Normal as the limit distributions of the sample mean for a large n. With large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems. This fact implies that we can use parametric procedures even when the data are not normally distributed. Al least a goodness-of-fit test must be performed for accepting whether normality is valid or not.
Monte Carlo (MC) techniques are used for selecting independent random samples of populations of means of three variables of importance in web network management. Different tests are performed to establish the acceptance of the normality. We did not find reliable results even for samples of size 10 000.

KEYWORDS: Big-Data, normality tests, asymptotic normality of means.

MSC: 62E99

RESUMEN
Nowadays we deal with Big-Data commonly. The users of statistics rely on having a large sample size n for using the statistical methods based on normality. Usual inference methods are typically based on considering the Normal as the limit distributions of the sample mean for a large n. With large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems. This fact implies that we can use parametric procedures even when the data are not normally distributed. Al least a goodness-of-fit test must be performed for accepting whether normality is valid or not.
Monte Carlo (MC) techniques are used for selecting independent random samples of populations of means of three variables of importance in web network management. Different tests are performed to establish the acceptance of the normality. We did not find reliable results even for samples of size 10 000.

PALABRAS CLAVE: grandes masas de datos, pruebas de normalidad, normalidad asintótica de medias

1. INTRODUCTION

Nowadays we deal with Big-Data commonly. The users of statistics rely on having a large sample size n for using the statistical methods based on normality. Fixing an appropriate sample size generally depends upon the size of the population studied. The usual idea is that collecting more data is better. It is supported by the idea that the statistical power is improved by increasing the sample size. To think in the power of computer for dealing with Big Data is a simplicity as dealing with an indefinitely large data is not intelligent as any research must consider the consumption of time and which is the point in which it becomes unproductive.

To establish the distribution of the variable of interest is one of the main issues in statistics as inferences are concerned with the validity of distributional assumptions.

We try to give evidence on the fact that is not so simple to confide in having a large n. Some theoretical research on the probabilistic model generating the data is needed. Actually is popular considering that having Big-Data and a Large-Computing power is enough for disregarding the analysis of the mathematical models. That attitude is foolish but is becoming a commonly supported by the increase of the brute force of computers capacity and speed. Take this paper as a counter example that having a large sample size normality is to be accepted without hesitation.

There are a lot of methods to test the goodness-of fit and some are more well-known and included in softwares . See the issues on SAS in Elliott-Woodward (2007) and Field (2009).

Pearson was the first one considering the problem of goodness of fit, see Pearson (1931). Its test statistic was a good approximation to the Maximum Likelihood one developed later. The use of the empirical distribution in goodness of fit tests was developed in the by Cramer (1928), von Mises (1931) and Kolmogorov (1933).

For testing normality some particular test was developed since the sixties, see Henry (2000) and D'agostino-Stephens (1986) for a large discussion. A more recent approach are tests that use methods coming from the theory of entropy.
Usual inference methods are typically based on considering the Normal as the limit distributions of the sample mean for a large n. With large enough sample sizes (> 30 or 40), the violation of the normality assumption should not cause major problems. This fact implies that we can use parametric procedures even when the data are not normally distributed. Asymptotic approximations are thumb rules. That n>30 is enough may be unreliable in most cases, and also even for larger values of n. In this paper we evaluate the behavior of goodness of fit tests for deciding if normality is to be accepted, as an adequate model, for the distribution of the sample mean. This fact has been analyzed in many papers . see for example Aishah (2011), Oztuna D, Elhan AH, Tuccar E. (2006), Royston P. (1991), Royston (1991), Steinskog (2007):
The goodness-of-fit tests are grouped considering their nature. Large sample sizes are taken for real life data coming from the evaluation of variables measuring the use of internet facilities in a network. Monte Carlo (MC) techniques are used for selecting independent random samples for the means of three variables of importance in web network management. The percent of acceptance of the normality is measured and compared qi with the expected 95%. In contrast with conventional wisdom of accepting that the mean distribution is described by a Gaussian, it seems to be not acceptable. We did not find reliable results even for samples of size 10 000.
Therefore, drawing inference on the means using normal theory is doubtfully universally correct. That poses a questioning of the solution of many statistical problems due to the question of how large n should be accepting limit theorems. The unquestioned question How large should the sample be for accepting that the Central Limit Theorem holds? Gay-Diehl (1992) recommended that the statistician should answer Large enough . More practical is to say take as large as possible
In section 2 the tests a to be compared are presented. Section 3 is concerned with the presentation of the data base used for the study and the estimation of the level of significance.

2. GOODNESS OF FIT TESTS FOR NORMALITY

The normal distribution is among the most useful distributions in statistical applications. Accordingly, testing for normality is of fundamental importance in many fields. Commonly the practitioners accept that the distribution of the sample mean is approximately normal, due to considering as an axiom the validity of the Central Limit Theorem. See for example Thode (2002). Commonly the experimenter considers that in experimental research n>30 is large enough, saying that samples sizes larger than 30 ensure the researcher that the Central Limit Theorem holds. Some conservative researchers insist that $a \leq n \leq 1000$, $a \in [50, 100]$, see Alreck-Settle (1995) . Among others Micceri (1989), Oztuna et al. (2006) pointed out how unfaithful is this theorem.
The normal probability density function (pdf) of a continuous cumulative probability distribution function F_0 is expressed by $f_0(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$. Statistical tests should be used for establishing if is valid the asseveration on the normality of a variable X.
Consider that we have a random sample of size n: $X_1, \dots, X_n$. The order statistics of the sample are denoted by $X_{(1)}, \dots, X_{(n)}, Z_{(i)} = \frac{X_{(i)}-\bar{X}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}}$. We may compute $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}, S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}, \breve{X} = \text{median}, S_{rob} = \sqrt{\frac{\pi}{2}} \frac{\sum_{i=1}^{n}|X_i-\breve{X}|}{n}, m_t = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^t}{n}$ $\sqrt{b_1} = \left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^3}{n}\right)\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}\right)^{-\frac{3}{2}}, b_2 = \left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^4}{n}\right)\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}\right)^{-2}$
These statistics are used by the tests described in the sequel. We are distinguishing some groups of normality tests:
We consider as classic tests those commonly appearing in text books and implemented in different softwares.

D ´Agostino tests (developed by D 'Agostino and followers)
Order statistic based tests (tests using explicit functions of order statistics)
Tests based on Entropy (using ideas coming from the theory of Entropy for developing test statistics).
A goodness-of-fit test (based on sample entropy) for normality was given by Vasicek. The test, however, can be applied only to the composite hypotheses. In this article an extended test of fit for normality is introduced based on. Kullback-Leibler information.

Other tests (tests not classifiable in one of the previous classes)

2.1. Classic tests

Karl Pearson was the first statistician who recognized in a paper published in 1900, the need of examining whether the observed data support the validity of considering that a certain pdf generates them. His paper introducing chi-squared test is one of the pillars of twentieth century statistics. He called it chi-square goodness of-fit test.

Afterwards of the most popular goodness of fit for normality were developed. Some of them are given in the sequel. We consider them as classical. These tests support the rejection of the null hypotheses when the corresponding test statistic is too large. The more popularly ones are presented in statistical courses are:

Chi squared, Pearson (1900)

Pearson's paper of 1900 introduced what subsequently became known as the chi-squared test. The set of n observation is divided into $K = \min\left\{m \in Z| \quad m \geq 2n^{\frac{2}{5}}\right\}$ disjoint classes and is computed

$$U = \sum_{i=1}^{K} \frac{(O_i - E_{0i})^2}{E_{0i}},$$

$$O_i = \text{number of observed variables falling in the ith class,}$$

E_{0i}=Expected number of observations falling in the ith class calculated using F_0.

The classes are constructed in such a way that $\forall i = 1, \dots, K, P\{X \in C_i\} = P$.

Large values of U supports that normality is not to be accepted.

A series tests are based on the empirical distribution. Some of them are:

Cramer-von Mises (1928, 1931)

They proposed using the test statistic

$$CM = \frac{1}{12n} + \sum_{i=1}^{n} \left(F_0\left(Z_{(i)}\right) - \frac{2i-1}{2n}\right)^2$$

Kolmogorov-Smirnov (1933)

They used the empirical distributing function evaluated in order statics. The test statistic proposed is

$$KS = \max\left\{\max_{1\leq j\leq n}\left[\frac{j}{n} - F_0(Z_{(j)})\right], m\ ax_{1\leq j\leq n}\left[F_0\left(Z_{(j)}\right) - \frac{j-1}{n}\right]\right\}$$

Lilliefors (1967) provided tables for testing whether KS is too large for accepting the normality of the variable. Applying this test copes with the difficulty of unknowing the parameters values and that it power is low.

See detailed discussion on this test in Steinskog D. J. (2007):

Anderson-Darling (1954)

They derived the statistics

$$AD = n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left(\log\left(F_0\left(Z_{(i)}\right)\right) + \log\left(1 - F_0\left(Z_{(n-i+1)}\right)\right)\right)$$

Watson (1961)

A function of CM was constructed by him. It is

$$A = (CM)^2 - n\sum_{i=1}^{n}\left(F_0\left(Z_{(i)}\right) - \frac{1}{2}\right)^2$$

2.2. D 'Agostino tests

A family of tests is defined by the principles fixed in the theory of goodness of fit based on the proposals of D 'Agostino in a series of paper. Examples are the papers of DOORNIK-Hansen (1994). Some of the most popular are given below

D 'Agostino (1970)

The first proposed was denominated omnibus test which uses

$$D_1 = \frac{\sum_{i=1}^{n}\left(i - \frac{n+1}{2}\right)X_{(i)}}{n^{\frac{3}{2}}S}$$

To reject normality is to be decided if this test statistic is considered whether extremely large or too small.

D ‘Agostino-Pearson (1973)
It considered the use of the sample kurtosis and/or skewness. Normality is,as with D_1, also rejected if is extremely large or small the test statistic

$$\sqrt{b_1} = \left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^3}{n}\right)\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}\right)^{-\frac{3}{2}}$$

And/or

$$b_2 = \left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^4}{n}\right)\left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}\right)^{-2}$$

Agostino- et al (1990)
This D ´Agostino´s family test is based on a transformation of the kurtosis and the skewness.
Take

$$c_t = \begin{cases} 6\dfrac{(n^2 - n + 2)\sqrt{6(n + 3)(n + 5)}}{(n + 7)(n + 9)\sqrt{n(n - 2)(n - 3)}} \text{ if } t = 1 \\ 6 + \dfrac{8}{2 + \sqrt{4 + c_1^2}} \text{ if } t = 2 \\ \dfrac{\left(b_2 - 3\dfrac{(n - 1)}{n + 1}\right)\sqrt{(n + 1)^2(n + 3)(n + 5)}}{\sqrt{24n(n - 2)(n - 3)}} \text{ if } t = 3 \\ \dfrac{\sqrt{b_1(n + 1)(n + 3)}}{\sqrt{6(n - 2)}} \text{ if } t = 4 \\ \dfrac{3(n^2 + 27n - 70)(n + 1)(n + 3)}{(n - 2)(n + 5)(n + 7)(n + 9)} \text{ if } t = 5 \\ \sqrt{2c_5 - 1} - 1 \text{ if } t = 6 \\ \sqrt{\dfrac{2}{c_6 - 1}} \text{ if } t = 7 \\ \dfrac{(n + 5)(n + 7)\left((n - 2)(n^2 + 27n - 70) + b_1\,(n - 7)(n^2 + 2n - 5)\right)}{6(n - 3)(n + 1)((n^2 + 15n - 4)} \text{ if } t = 8 \end{cases}$$

The normality is rejected for large values of

$$D_2 = \left(Z\ \ (\sqrt{b_1})\right)^2 + \left(Z\ \ (b_2)\right)^2$$

Where

$$Z\ \ (\sqrt{b_1}) = l\frac{\log\left(\frac{c_4}{c_7}\right) + \sqrt{\left(\frac{c_4}{c_7}\right)^2 + 1}}{\sqrt{\log(\sqrt{c_6})}}, Z\ \ (b_2) = 3\left[\left(1 - \frac{2}{9c_8}\right) - \sqrt[3]{\frac{1 - \frac{2}{c_8}}{1 + c_4\sqrt{\frac{2}{c_8 - 4}}}}\right]\sqrt{\frac{c_8}{2}}$$

Doornik-Hansen (1994)
They developed a transformation of D ‘Agostino’s statistic given by

$$D_3 = Z\ \ (\sqrt{b_1})^2 - DH^2$$

Where

$$DH = 3\left[\sqrt[3]{\frac{b_2 - 1 - b_1}{2c_8}}(2h) - 1 + \frac{1}{9}\right]\sqrt{2c_8}$$

$$h = \frac{(n + 5)(n + 7)(n^3 + 37n^2 + 11n - 313)}{12N - 3)(n + 1)(n^2 + 15n - 4)}$$

The normality is not accepted for large values of D_3.

2.3. Tests using order statistic

Order statistics are among the most important functions of a set of random variables that we study in probability and statistics. There is a well-developed theory of the order statistics of a fixed number n of observations from a fixed distribution, as also an asymptotic theory where n goes to infinity. Test of normality based on them have been developed. We will present some of them in this section
Denote the inverse of the standard normal distribution $M_{(i)} = \phi^{-1}(m_{(i)})$.
We do not accept normality if is obtained a large value of the statistic described below.

De Wet-Venter (1972)

$$W = \frac{\sum_{i=1}^{n}\left(X_{(i)} - \overline{X} - \phi^{-1}\left(\frac{i}{n+1}\right)\right)^2}{S^2}$$

Pettit (1977)

$$P = \sum_{i=1}^{n}\left(\frac{\phi\left(\frac{X_{(i)} - \overline{X}}{S}\right) - \left(\frac{i}{n+1}\right)}{\phi\left(\phi^{-1}\left(\frac{i}{n+1}\right)\right)}\right)^2$$

Del Barrio et al (1999)

$$B = 1 - \frac{\left(\sum_{k=1}^{n} X_{(k)} \int_{\frac{k-1}{n}}^{\frac{k}{n}} F_0^{-1}(t)dt\right)^2}{m_2}$$

Small values of the test static yields rejecting the normality in the following functions of order statistic.

Shapiro-Wilks (1962)
Take

$$E\begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix} = \begin{pmatrix} m_{(1)} \\ \vdots \\ m_{(n)} \end{pmatrix} = \vec{m}, V\begin{pmatrix} X_{(1)} \\ \vdots \\ X_{(n)} \end{pmatrix} = V_{n\times n} \text{ and } \left(a_{(1)}, \dots, a_{(n)}\right)^T = \frac{\vec{m}^T V_{n\times n}^{-1}}{\sqrt{\vec{m}^T V_{n\times n}^T V_{n\times n}^{-1}\vec{m}}}$$

The statistic for developing the normality test is

$$SW = \frac{\left(\sum_{i=1}^{\left\lfloor\frac{n}{2}\right\rfloor} a_{(n-i+1}\left(X_{(n-i+1)} - X_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(X_{(i)} - \overline{X}\right)^2}$$

Shapiro-Francia (1972)

In this case is to be determined $\left(b_{(1)}, \dots, b_{(n)}\right)^T = \frac{\vec{m}^T}{\sqrt{\vec{m}^T\vec{m}}}$. The test statistic to be use in the inferences is given by

$$SF = \frac{\left(\sum_{i=1}^{n} b_{(i)}\left(X_{(i)}\right)\right)^2}{\sum_{i=1}^{n}\left(X_{(i)} - \overline{X}\right)^2}$$

Filliben (1975)
Taking

$$m_{(i)} = \begin{cases} 1 - \sqrt[n]{0,5} \text{ if } i = 1 \\ \sqrt[n]{0,5} \text{ if } i = n \\ \dfrac{i - 0,3175}{n + 0,365} \text{ otherwise} \end{cases}$$

The test statistic is

$$F = \frac{\sum_{i=1}^{n} X_{(i)} M_{(i)}}{\sqrt{S \sum_{i=1}^{n} M_{(i)}^{2} \sqrt{n-1}}}$$

Chen-Shapiro (1995)

$$SC = \frac{1}{(n-1)S} \sum_{i=1}^{n-1} \frac{X_{(i+1)} - X_{(i)}}{\phi^{-1}\left(\frac{i+1-0,375}{n+0,25}\right) - \phi^{-1}\left(\frac{i-0,375}{n+0,25}\right)},$$

2.4. Tests based on Entropy.

The entropy of a random variable was introduced in the seminal paper of Shannon(1948).It measures information and uncertainty. Entropy is fundamental in information theory, communication, pattern recognition, statistical physics etc. See Choi (2008), Abbasnejad (2011), Abbasnejad et al.(2012), Alizadeh-Arghami (2012, 2013) for recent contributions in the theme. The sample entropy, the estimate of the entropy per observation, is commonly attributed to introduced by Vasicek (1976), but Dimitriev-Tarasenko (1973) were the first ones in introducing the use entropy for goodness of fit. The normality is rejected for small values of the following tests statistics

Dimitriev-Tarasenko (1973)
The test is based on the use of the Gaussian kernel based estimator of the density function

$$\hat{f}(X_i) = \frac{1}{1{,}06 n^{\frac{4}{5}} \sqrt{2S^2\pi}} \sum \exp\left(-\frac{1}{2} \left(\frac{X_i - X_j}{1{,}065 n^{\frac{1}{5}} S} \right)^2 \right)$$

For testing is used

$$V_{DT} = \frac{\exp\{HDT_{mn}\}}{S}, HDT_{mn} = -\int_{-\infty}^{\infty} \ln\left(\hat{f}(x)\right) \hat{f}(x) dx$$

Vasicek (1976)
He suggested a statistic which used order statistics, within an entropy functional. It is necessary that thm e positive integer satisfies the inequality m<n/2

$$V_P = \frac{\exp\{HV_{mn}\}}{S}, HV_{mn} = \sum_{i=1}^{n} \ln\left(\frac{n}{2m} \left(X_{(i+m)} - X_{(i-m)} \right) \right), i = \begin{cases} 1 \text{ when } i < 1 \\ n \text{ when } i > n \\ i \text{ in other case} \end{cases}$$

Van Es (1992)
The proposal was using

$$V_{ES} = \frac{\exp\{HES_{mn}\}}{S}$$

where

$$HES_{mn} = \ln(m) - \ln(n+1) + \frac{1}{n-m} \sum_{i=1}^{n-m} \ln\left(\frac{n+1}{m} \left(X_{(i+m)} - X_{(i)} \right) \right) + \sum_{k=m}^{n} \frac{1}{k},$$

Ebrahimi-Plufhoeft-Soofi (1994)
There test is derived from a transformation of V_P . Taking
$c_i = \left(1 - \frac{i-1}{m}\right) I_{[1,m]}(i) + 2I_{[m+1,n-m]}(i) + \left(1 + \frac{n-i}{m}\right) I_{[n-m+1,n]}(i)$
The test is performed using the statistic

$$EPS_P = \frac{\exp\{HE_{mn}\}}{S}, HE_{mn} = \frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{n}{c_i m} \left(X_{(i+m)} - X_{(i-m)} \right) \right), i = \begin{cases} 1 \text{ when } i < 1 \\ n \text{ when } i > n \\ i \text{ in other case} \end{cases}$$

Correa (1995)

The test is obtained by using

$$V_C = \frac{\exp\{HC_{mn}\}}{S},$$

Were defined

$$HC_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{\sum_{j=i-m}^{i+m} \quad \left(X_{(j)} - \breve{X}_{(i)}\right)(j-i))}{n\sum_{j=i-m}^{i+m} \quad \left(X_{(j)} - \breve{X}_{(i)}\right)^2}\right),$$

and

$$\breve{X}_{(i)} = \frac{1}{2m+1}\sum_{j=i-m}^{i+m} \quad X_{(j)}$$

Opposite to the decision rule of the previous tests, the non-parametric versions derived from the previous tests statistics, large values of the statistics indicates that normality must be rejected.

Park (2003)

Take $\xi_i = \frac{1}{2m}\sum_{j=0}^{2m-1} x_{(i-m+j)}$ and again

$HV_{mn} = \sum_{i=1}^{n} \ln\left(\frac{n}{2m}\left(X_{(i+m)} - X_{(i-m)}\right)\right), i = \begin{cases} 1 \text{ when } i < 1 \\ n \text{ when } i > n \\ i \text{ in other case} \end{cases}$

The proposed test statistic is

$$V_{NP} = \log\sqrt{2\pi}\left(\frac{\sum_{i=1}^{n} g_v\,(x) - \frac{1}{n}\sum_{i=1}^{n} g_v\,(x)}{n-1}\right)^{\frac{1}{2}} + 0{,}5 - HV_{mn},$$

where

$$g_v(x) = \begin{cases} 0 \text{ if } x < \xi_1 \\ \dfrac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)} \text{ if } \xi_i < \ x < \xi_{i+1} \\ 0 \text{ if } x > \xi_{n+1} \end{cases}$$

It is a non-parametric alternative for V_P.

The non-parametric version of EPS_P) is derived taking

$c_i = \left(1 - \frac{i-1}{m}\right) I_{[1,m]}(i) + 2I_{[m+1,n-m]}(i) + \left(1 + \frac{n-i}{m}\right) I_{[n-m+1,n]}(i)$

The test is performed using

$$EPS_{NP} = \log\sqrt{2\pi}\left(\frac{\sum_{i=1}^{n} g_e\,(x) - \frac{1}{n}\sum_{i=1}^{n} g_e\,(x)}{n-1}\right)^{\frac{1}{2}} + 0{,}5 - HE_{mn},$$

$$HE_{mn} = \sum_{i=1}^{n} \ln\left(\frac{n}{c_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right), i = \begin{cases} 1 \text{ when } i < 1 \\ n \text{ when } i > n \\ i \text{ in other case} \end{cases}$$

where

$$g_e(x) = \begin{cases} 0 \text{ if } x < \eta_1 \\ \dfrac{2m}{n\left(\eta_{(i+m)} - \eta_{(i-m)}\right)} \text{ if } \eta_i < \ x < \eta_{i+1}\,, \\ 0 \text{ if } x > \eta_{n+1} \end{cases}$$

$$\eta_i = \begin{cases} \eta_{m+1} - \dfrac{1}{m+k-1}\sum_{k=1}^{m}\left(X_{(k+m)} - X_{(1)}\right), 1 \le i \le m \\ \dfrac{1}{2m}\sum_{j=0}^{2m-1} x_{(i-m+j)} \text{ if } m+1 \le i \le n-m+1 \\ \eta_{n-m+1} + \dfrac{1}{m+k-1}\sum_{k=n-m+2}^{i}\left(X_{(n)} - X_{(k-m-1)}\right), n-m+2 \le i \le n+1 \end{cases}$$

Alizadeh-Nougal-Argami (2013) developed another non-parametric test derived from V_P. Their test statistic is

Alizadeh-Nougal-Argami (2013)

$$ANA_{NP} = \log\sqrt{2\pi}\left(\frac{\sum_{i=1}^{n} g_a(x) - \frac{1}{n}\sum_{i=1}^{n} g_a(x)}{n-1}\right)^{\frac{1}{2}} + 0{,}5 - HA_{mn},$$

$$HA_{mn} = \sum_{i=1}^{n}\ln\left(\frac{n}{a_i m}\left(X_{(i+m)} - X_{(i-m)}\right)\right), i = \begin{cases} 1 \text{ when } i < 1 \\ n \text{ when } i > n \\ i \text{ in other case} \end{cases}$$

$$g_a(x) = \begin{cases} 0 \text{ if } x < \eta_1 \\ \dfrac{1}{n\left(\eta_{(i+m)} - \eta_{(i-m)}\right)} \text{ if } \eta_i < x < \eta_{i+1}, \\ 0 \text{ if } x > \eta_{n+1} \end{cases}$$

$$\eta_i = \begin{cases} \eta_{m+1} - \dfrac{1}{m}\sum_{k=1}^{m}\left(X_{(k+m)} - X_{(1)}\right), 1 \le i \le m \\ \dfrac{1}{2m}\sum_{j=0}^{2m-1} x_{(i-m+j)} \text{ if } m+1 \le i \le n-m+1 \\ \eta_{n-m+1} - \dfrac{1}{m+}\sum_{k=n-m+2}^{i}\left(X_{(n)} - X_{(k-m-1)}\right), n-m+2 \le i \le n+1 \end{cases}$$

2.5. Other tests

Some tests suggest that normality is not to be accepted for large values of the statistics. We present some of them in the sequel.

Jacques-Bera (1980, 1987)
It is based on a function of the sample coefficients of skewness and kurtosis. Their proposal is using

$$JB = \frac{nb_1}{6} + \frac{n(b_2 - 3)^2}{24}$$

Its robustification was derived by Gel-Gastwirht (2008) yielding

$$JBG = \frac{n}{6}\left(\frac{m_3}{S_{rob}^3}\right)^2 + \frac{n}{64}\left(\frac{m_4}{S_{rob}^4} - 3\right)^2$$

Martínez-Iglewicz (1981)
It is based in the behavior of the observation with respect to their median. Take

$$\check{Z}_i = \begin{cases} \dfrac{X_i - \check{X}}{9M} \text{ if } \left|\dfrac{X_i - \check{X}}{9M}\right| < 1, \; M = \text{Median}\left\{\check{Z}_1, \dots, \check{Z}_n\right\} \\ 0 \text{ otherwise} \end{cases}$$

$$\check{S}^2 = n\frac{\sum_{|\check{Z}_i|<1} \left(X_i - \check{X}\right)^2 (1 - \check{Z}^2{}_i)^4}{\left(\sum_{|\check{Z}_i|<1} \left(1 - 5\check{Z}^2{}_i\right)^2 (1 - \check{Z}^2{}_i)\right)^2}$$

The proposed test statistic is

$$MI = \frac{\sum_{i=1}^{n} (X_i - M)^2}{(n-1)\check{S}^2}$$

Epps-Pulley (1983)

They used a function of the difference among the observations. Their proposal was using

$$EP = \frac{1}{\sqrt{3}} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \exp\left\{-\frac{\left(X_i - X_j\right)^2}{2m_2}\right\} - \frac{\sqrt{2}}{n}\sum_{j=1}^{n} \exp\left\{-\frac{\left(X_j - \overline{X}\right)^2}{4m_2}\right\}$$

Gel-Miao-Gastwirht (2007)

They simply considered a simple relation between dispersion measures. The tests static is given by

$$GMG = \frac{S}{S_{rob}}$$

Bonett-Seier (2002)

The absolute mean deviation is considered in its relation with the centered moment of order 2. A function of them is

$$BS = \left(13{,}29\left(\ln\sqrt{m_2} - \log\left(\frac{1}{n}\sum_{i=1}^{n}|X_i - \overline{X}|\right)\right) - 3\right)\frac{\sqrt{n+2}}{3{,}54}$$

The normality is not accepted if BS is whether to large or too small.

Commonly tables were developed for testing the significance of the goodness of fit tests. They are given in the referred papers. Some of them appear in specialized publications. The most popular appear in statistical softwares and the outputs provide also the p-value.

3. MONTE CARLO STUDY

Frequently more investigations are based on electronic inquires and is assumed that the sample size is to be large. The usually asked questioning of what size the sample should be used? is not made. It is a question pertinent to all investigations, but awkwardness arise in internet based electronic surveys.

We consider the fact that human interactions in the use of a network server. Due to the large quantity of data, web studies belong to the class of Big-Data problems. Web log data suggest using a series of measures for the evaluation of web´s performance. We decided studying

$$Y_t = \frac{\text{number of new connections}}{\text{number of disconnections}} \text{ in second t}$$

$$Z_t = \text{percent of active nodes in the network in second t}$$

$$W_t = \frac{\text{number of user sessions}}{\text{number of URL appearing in the processed log}} \text{ in second t}$$

In a year the data amounts 31´536.000 inputs of each variable. A very large sample was selected using simple random sampling and X=Y, Z, W were calculated. the respective sample means

$\overline{X}$ were computed for each sample. We took n=1.000, 10.000 and 100.000.

Monte Carlo experiments are commonly used for evaluating the behavior of goodness so fit tests. Ee for example Esteban et al. (2001). We performed Monte Carlo experiments for repeating the process of selecting independent sample for each n and the normality evaluated with each test described above with α=0,06. The experiment was performed B= 10.000 occasions. Take

$$P(Q:X) = \text{proportion of experiments in which the normality was accepted}$$

This proportion is an estimation of 1-α. Due to the large sample sizes and the properties of it as estimator of 1-α is expected that they must be close if the convergence of the distribution sample mean to a Gaussian is valid

Table 1. Proportion of experiments in which the normality was accepted. Classic tests with α=0,05

Statistic		Y			Z			W	
Tests	10^3	10^4	10^5	10^3	10^4	10^5	10^3	10^4	10^5
U	0,34	0,23	0,41	0,85	0,90	0,92	0,80	0,88	0,86
CM	0,48	0,27	0,58	0,81	0,93	0,93	0,80	0,86	0,86
KS	0,15	0,21	0,29	0,76	0,93	0,93	0,82	0,88	0,88
AD	0,18	0,18	0,43	0,82	0,94	0,94	0,64	0,65	0,70
A	0,48	0,27	0,58	0,79	0,94	0,95	0,86	0,88	0,87

See from Table 1 that classic tests tend to accept normality for Z with a close value to 0,52 for $n\geq1000000$. That is a really large sample size. For the other variables the sample mean could hardly be considered as normal

Table 2. Proportion of experiments in which the normality was accepted. **D ‘Agostino tests** with $\alpha=0,05$

Statistic		Y			Z			W	
Tests	10^3	10^4	10^5	10^3	10^4	10^5	10^3	10^4	10^5
D_1	0,61	0,75	0,83	0,80	0,85	0,88	0,41	0.52	0,58
D_2	0,55	0,64	0,85	0,83	0,88	0,83	0,45	0,65	0,67
D_3	0,51	0,73	0,85	0,79	0,83	0,87	0,44	0,62	0,70
$\sqrt{b_1}$	0,61	0,75	0,83	0,89	0,85	0,93	0,41	0,68	0,78
b_2	0,65	0,64	0,95	0,83	0,88	0,90	0,45	0,65	0,67

Table 2 suggests that Agostino´s tests have a not so stable behavior. The best was for Y using b_2 and is needed $n\geq100000$. For Z was $\sqrt{b_1}$ needing $n\geq1000000$. Both are very large sample sizes

Table 3. Proportion of experiments in which the normality was accepted. **Tests using order statistic** with $\alpha=0,05$

Statistic		Y			Z			W	
Tests	10^3	10^4	10^5	10^3	10^4	10^5	10^3	10^4	10^5
W	0,27	0,47	0,52	0,49	0,69	0,84	0,34	0,39	0,40
P	0,23	0,28	0,53	0,61	0,66	0,87	0,26	0,51	0,63
B	0,27	0,31	0,54	0,34	0,41	0,84	0,32	0,44	0,60
SW	0,27	0,27	0,52	0,59	0,66	0,85	0,33	0,43	0,60
SF	0,22	0,28	0,53	0,69	0,68	0,84	0,28	0,41	0,53
SC	0,27	0,30	0,44	0,64	0,69	0,90	0,22	0,44	0,61

Table 3 gives support to rejecting the normality of the sample mean as the proportion of acceptations are very low. Only Z obtained an acceptable level of acceptation if $n\geq1000000$ when using SC. This test statistic presented the larger proportions of not rejection for Z.

Table 4. Proportion of experiments in which the normality was accepted. **Tests based on Entropy** with $\alpha=0,05$

Statistic		Y			Z			W	
Tests	10^3	10^4	10^5	10^3	10^4	10^5	10^3	10^4	10^5
V_{DP}	0,27	0,30	0,32	0,87	0,90	0,94	0,67	0,70	0,72
V_{ES}	0,29	0,34	0,45	0,92	0,94	0,98	0,32	0,54	0,75
EPS_P	0,27	0,49	0,52	0,84	0,95	0,96	0,20	0,49	0,62
V_C	0,58	0,83	0,83	0,79	0,95	0,92	0,50	0,83	0,87
V_{NP}	0,27	0,30	0,62	0,92	0,96	0,94	0,62	0,93	0,92
EpS_{NP}	0,29	0,44	0,57	0,92	0,98	0,98	0,32	0,74	0,85
ANP_{NP}	0,20	0,59	0,62	0,84	0,95	0,94	0,27	0,79	0,82

Table 4 suggest that tests based on entropy performed well for Z but not with the other variables. In any case the needed values of n are very large.

Table 5. Proportion of experiments in which the normality was accepted. Other tests with $\alpha=0,05$

Statistic		Y			Z			W	
Tests	10^3	10^4	10^5	10^3	10^4	10^5	10^3	10^4	10^5

JBG	0,33	0,59	0,63	0,81	0,84	0,93	0,59	0,53	0,51
MI	0,68	0,77	0,78	0,77	0,85	0,81	0,77	0,78	0,77
EP	0,73	0,82	0,88	0,82	0,85	0,93	0,82	0,88	0,88
GMG	0,62	0,72	0,78	0,74	0,81	0,90	0,72	0,78	0,79
BS	0,53	0,59	0,63	0,81	0,84	0,93	0,59	0,55	0,58

From Table 5 we have again that that tests based on EP, GMG and BS performed well for Z and n≥1000000 but not with the other variables.

4. CONCLUSIONS

Accepting that n>30 is enough for accepting normality of the sample mean has not a real basis. The experiments gave evidence that for some variables its practically impossible fixing a reasonably large value of n for accepting that the hypothesis n is large is to be accepted.

ACKNOWLEDGMENTS: We are grateful to the support obtained from the projects CU2018JOt0044104 y CYTED 2015-P515RT01236-7. The authors thank the criteria of two anonymous referees which allowed improving the quality of the paper.

REFERENCES

[1] ABBASNEJAD, M (2011): Some Goodness of Fit Tests Based on Renyi Information. **Applied Mathematical Sciences**, 5, 1921 - 1934

[2] ABBASNEJAD, M., N. R. ARGHAMI, M. TAVAKOLI (2012): A Goodness of Fit Test For Exponentiality Based on Lin-Wong Information. **JIRSS** . 11, 191-202

[3] AISHAH AHAD, N., T. SIN YIN, A. OTHMAN and C. ROHANI YAACO (2011): Sensitivity of Normality Tests to Non-normal Data. **Sains Malaysiana**, 40, 637–641

[4] ALIZADEH NOUGHABI, H. and ARGHAMI, N.R. (2012): General Treatment of Goodness-of-FitTests Based on Kullback-Leibler Information. **Journal of Statistical Computation and Simulation**, 83, 1556-1569.

[5] ALIZADEH NOUGHABI, H. and ARGHAMI, N.R. (2013): Goodness-of-Fit Tests Based on Correcting Moments of Entropy Estimators. **Communications in Statistics - Simulation and Computation**, 42, 499-513.

[6] AHMAD, F. and SHERMAWI, R.A.K. (2015): Power comparison of various normality test. **Pk. Stat. Oper. Res.** 11, 331-345.

[7] ANDERSON, T.W. and DARLING, D.W. (1954): A test of goodness of fit. **J. Amer. Statist. Assoc**., 49, 765 – 769.

[8] CHOI, B. (2008): Improvement of Goodness of Fit Test for Normal Distribution Based onEntropy and Power Comparison. Journal of Statistical Computation and Simulation, 78,781-788.

[9] CORREA, J. C. (1995), A new estimators of entropy. **Commun. Statist. Theory Meth**., 24, 2439-2449.

[10] D'AGOSTINO, RALPH B. (1970): Transformation to normality of the null distribution of g_1. **Biometrika,** 57 , 679–681.

[11] D'AGOSTINO, R., PEARSON, E. S. (1973): Testing for departures from normality. **Biometrika**, 60, 613 - 622.

[12] D'AGOSTINO, RALPH B.; ALBERT BELANGER; RALPH B. D'AGOSTINO, Jr (1990): A suggestion for using powerful and informative tests of normality . **The American Statistician**. 44, 316–321.

[13] D'AGOSTINO, R. B. and STEPHENS, M. A. (1986): **Goodness-of-fit Techniques**. Marcel Dekker, New York.

[14] DOORNIK, J.A. and HANSEN, D. (1994): **An omnibus test for univariate and multivariate normality**. Working Paper, Nuffield College, Oxford.

[15] DMITRIEV, Y. G. and F. P. TARASENKO. (1973): On estimation of functionals of the probability density function and its derivatives. **Teor. Veroyatnost. i Primenen**., 18, 662–668.

[16] ELLIOTT AC, WOODWARD WA. (2007): **Statistical analysis quick reference guidebook with SPSS examples.** 1st ed.: Sage Publications, London

[17] EPPS, T.W. and PULLEY, L.B. (1983): A test for normality based on the empirical characteristic function. **Biometrik**a, 70, 723 -726.

[18] ESTEBAN, M.D., CASTELLANOS, M.E., MORALES, D. AND VAJDA I. (2001): Monte Carlo Comparison of Four Normality Tests Using Different Entropy Estimates. **Communications in Statistics Simulation and computation**, 30, 761-785.

[19] FIELD A. (2009): **Discovering statistics using SPSS**. 3 ed. SAGE publications Ltd., London.

[20] HENZE, N. (1990): An Approximation to the Limit Distribution of the Epps-Pulley Test Statistic for Normality. **Metrika**, 37, 7 – 18.

[21] HILL, R. (1998): WHAT SAMPLE SIZE is ENOUGH in INTERNET SURVEY RESEARCH. **Interpersonal Computing†and Technology An Electronic Journal for the21st Century**, 6, hillSamplesizeÆhtml.

[22] JARQUE, C.M. and BERA, A.K. (1987): A test for normality of observations and regression residuals. **Int. Stat. Rev**., 55, 163 – 172.

[23] JARQUE, W.M. and BERA, A.K. (1987): E±cient tests for normality, homoscedasticity and serial independence of regression residuals. **Economics Letters** 6, 255-259.

[24] MICCERI, T. (1989): The unicorn, the normal curve, and the other improbable

a. creatures. **Psychological Bulletin**, 105, 156-166.

[25] OZTUNA D, ELHAN AH, TUCCAR E. (2006): Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. **Turkish Journal of Medical Sciences**, 36, 171–176.

[26] PARK, S. and PARK, D. (2003), Correcting moments for goodness of fit tests based on two entropy estimates. **J. Statist. Comput. Simul**., 73, 685-694.

[27] PEARSON, EGON S. (1931): Note on tests for normality. **Biometrika,** 22, 423–424.

[28] ROYSTON P. (1991): Estimating departure from normality. **Stat Med**.10, 1283–93.

[29] STEINSKOG D. J. (2007): A cautionary note on the use of the Kolmogorov-Smirnov test for normality. **American Meteor Soc**.135:1151–1157

[30] SHAPIRO, S. S. and FRANCIA, R. S. (1972): An approximate analysis of variance test for normality. **J. Amer. Statist. Assoc**., 67, 215-216.

[31] THODE HJ. (2002): **Testing for normality**. Marcel Dekker, New York.

[32] VAN Es, B. (1992), Estimating functional related to a density by a lass of statistic based on spacings. **Scand. J. Statist**., 19, 61-72.

[33] VASICEK , O. (1976): Test for Normality Based on Sample Entropy . **Journal of the Royal Statistical Society. Series B.** 38, 54-59