INVESTIGATION ON THE GUASSIANITY AND INTERPOLABILITY OF THE HOLY QURAN -APPLICATION OF AUTHOR DISCRIMINATION

H. Sayoud

http://sayoud.net

Abstract

It is well known that every physical phenomenon in the universe respects some specific rules, as we can scientifically observe and measure. For instance, the Gaussianity rule characterizes every physical phenomenon respecting the "Large Numbers" condition. On the other hand, the Interpolability can be noticed for almost every discrete curve representing a natural physical phenomenon. Those two rules are respected by a wide variety of data present in the Universe or let us say simply in our daily life.

In the case of textual data, the two rules should theoretically be respected, and indeed, they have been verified during some experiments conducted on a set of several books, which have been analyzed during this investigation.

However, in the case of the holy Quran, neither the Gaussianity nor the interpolability of the data curve is respected. Moreover, the evolution continuity of the data (second derivative of the curve) is not well respected and seems quite strange with regards to conventional or usual mathematical curves (e.g. Gaussian, exponential, linear, etc.).

Furthermore, we notice an inexplicable and strange statistical structure in the holy book, without any (prior) scientific interpretation. As a consequence, this investigation strongly confirms that the two books: Quran and Hadith should belong to two different Authors.

I. INTRODUCTION

Most of existing natural data obey a certain set of physical rules that seem to be quite simple, mathematically speaking. For instance, the gravity equation is basically simple (P=m.g); the energy equation is also simple ($E=mc^2$) and the electric voltage equation is further simpler (U=RI). This mathematical simplicity in most of the natural or physical data is almost always verified.

On the other hand, a famous theorem in statistics, called "Central Limit Theorem" or sometimes "Theorem of Large Numbers", stipulates (after a rigorous demonstration) that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed (i.e. Gaussian distribution), regardless of the underlying distribution [Siegrist, 2016] [Rice, 1995]. That is, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (i.e. Gaussian distribution) [Contributors, 2015].

The central limit theorem has an interesting history. The first version of this theorem was postulated by the French mathematician De Moivre who, in a remarkable article published in 1733, used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This finding was far ahead of its time, and was nearly forgotten until the French mathematician Laplace rescued it from obscurity in his work called “Théorie analytique des probabilities”, which was published in 1812. Laplace expanded De Moivre's finding by approximating the binomial distribution with the normal distribution. But as with De Moivre, Laplace's finding received little attention in his own time. It was until the end of nineteenth century that the importance of the central limit theorem was discerned, when, in 1901, the Russian mathematician Aleksandr Lyapunov defined it in general terms and proved precisely how it worked mathematically. Nowadays, the central limit theorem is considered to be the sovereign of probability theory [Contributors, 2015].

Sir Francis Galton described the Central Limit Theorem as [Galton, 1889]: <<…It reigns with serenity and in complete self-effacement, amidst the wildest confusion. The larger the mob, and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of Unreason. Whenever a large sample of chaotic elements are taken in hand and marshaled in the order of their magnitude, an unsuspected and most beautiful form of regularity proves to have been latent all along>> [Contributors, 2015].

Another important aspect is the natural continuity of every physical or natural phenomenon. In other words, the graphical representation of a measured physical data should present a certain continuity and regularity (i.e. the curve shape respects some well-known graphical models). Again, taking discrete samples from that measured physical data, will lead to a discrete curve from which it should be possible to interpolate and fit with usual interpolation or fitting functions.

Hence, it could be seen why Michael Whiteman stated [Whiteman, 1967]: “To speak of running through ‘all the points’ of a curve would therefore be inadmissible. Nevertheless, there are exact concepts of continuity and gradient, which are applicable to conceptually defined curves, and thence are applicable approximately to physical. curves. Thus, in any particular case a physical curve may be tested for continuity, and an approximate measure of its gradient may be found. Likewise a physical trajectory does not consist of isolated events. Nevertheless, by selecting points, the exact concept of velocity may be applied so as to obtain an approximate measure of velocity at any physical point of the trajectory” [Whiteman, 1967].

Now, by considering some natural/physical curves present in the universe, we observe continuity in the graphical representation of the measured dimension (continuity of the dimension). Moreover, we should even find some continuity in its first derivative (continuity of the variation). For instance, let us observe the temperature curve of the weather. Not only, the temperature should vary continuously and smoothly, but also its derivative does so with an extreme respect of the physical and natural well-known laws.

Again, by observing the natural curves present in real life, one remark that the form of the curves is quite identifiable by a simple visual observation (as experts do), and the curves are easily fitted by usual mathematical functions (e.g. Gaussian, Linear, Polynomial, Sinusoidal, Exponential, etc.).

For concreteness, if we take the text data for instance, we may remark that it is composed of several characters (i.e. A, B, C, … Z, for the English), several words (eg. So, You, Yes, …, for the English), several numbers (e.g. 1, 2, 3 etc.) and so on.

That is, if a large amount of textual data is analyzed by computing the frequency of some features, we should usually retrieve a Gaussian distribution of the data, when the features are represented from the most frequent to the least frequent.

As it can be seen in the next section, this particularity has been checked with 7 different books by taking the following feature: “Word Length Frequency”.

Contrariwise, for the holy Quran those mathematical rules do not seem to be respected, without any possible interpretation. Moreover, by analyzing another feature (i.e. Number citation frequency), we strangely noticed that this last feature does not respect the previous laws either, while for the case of the Hadith book, the mathematical laws are well respected for all those features.

This manuscript is organized as follows: We first present some important definitions of fitting and interpolation, then, two types of features are analyzed: Word length frequency and Number citation frequency. At the end, a discussion and conclusion are provided with some interesting bibliographical references.

II. FITTING AND INTERPOLATION DEFINITIONS

Given a set of data that results from an experiment or from a physical scenario, we show (in Mathematics) that there is some function that passes through the data points and optimally represents the area of interest at all present and absent points.
With *interpolation* [Milne, 2012], we look for a function that allows us to approximate the values between the original data points [William Edmund Milne]. The interpolating function should pass exactly through the original data points. In our experiments, we chose two different techniques: the Piecewise cubic Hermite interpolation (PCHIP), which preserves the monotonicity and shape of the data and the Bezier interpolation technique, which preserves the curve shape in a graphical way.

On the other hand, with curve fitting [Milne, 2012], we simply want a function that represents a *good fit* to the original data points with a minimum estimation error. With curve fitting the approximating function does not have to pass through the original data points. It should respect the overall data with the best possible fitting, and respect the chosen mathematical expression type as well.
For instance, in table 1, we have represented four columns : the first one represents the sample order, the second one represents the original data corresponding the real samples values, the third column represents the interpolation based values and the fourth one represents the fitting based values.

Fig. 1. Example of Interpolation and Fitting.

Sample	Original Samples	Interpolation	Fitting
1	1.00	1.00	1.07
2	absent	2.01	2.04
3	3.00	3.00	2.99
4	absent	3.97	3.98
5	5.00	5.00	5.03

III. INVESTIGATION ON THE WORD-LENGTH FREQUENCY

III.1 Definition of the Word Length Frequency (WLF)

Since the first part of our experiment concerns the Word Length Frequency (WLF) [Sayoud, 2012], we think that it could be useful to define some technical terms employed in this work:

-The "Word Length" represents the number of letters composing the word.
-The "Word Length Frequency" F(n) for a specific length ‘n’, represents the number (in percent) of words composed of n letters each, present in the text.

III.2 Graphical representation of the Word Length Frequency

In this section we will graphically represent the WLF of the two books : holy Quran and Hadith. Furthermore, we will represent the WLF of 6 other books written by 6 different authors to make a general comparison between their features.

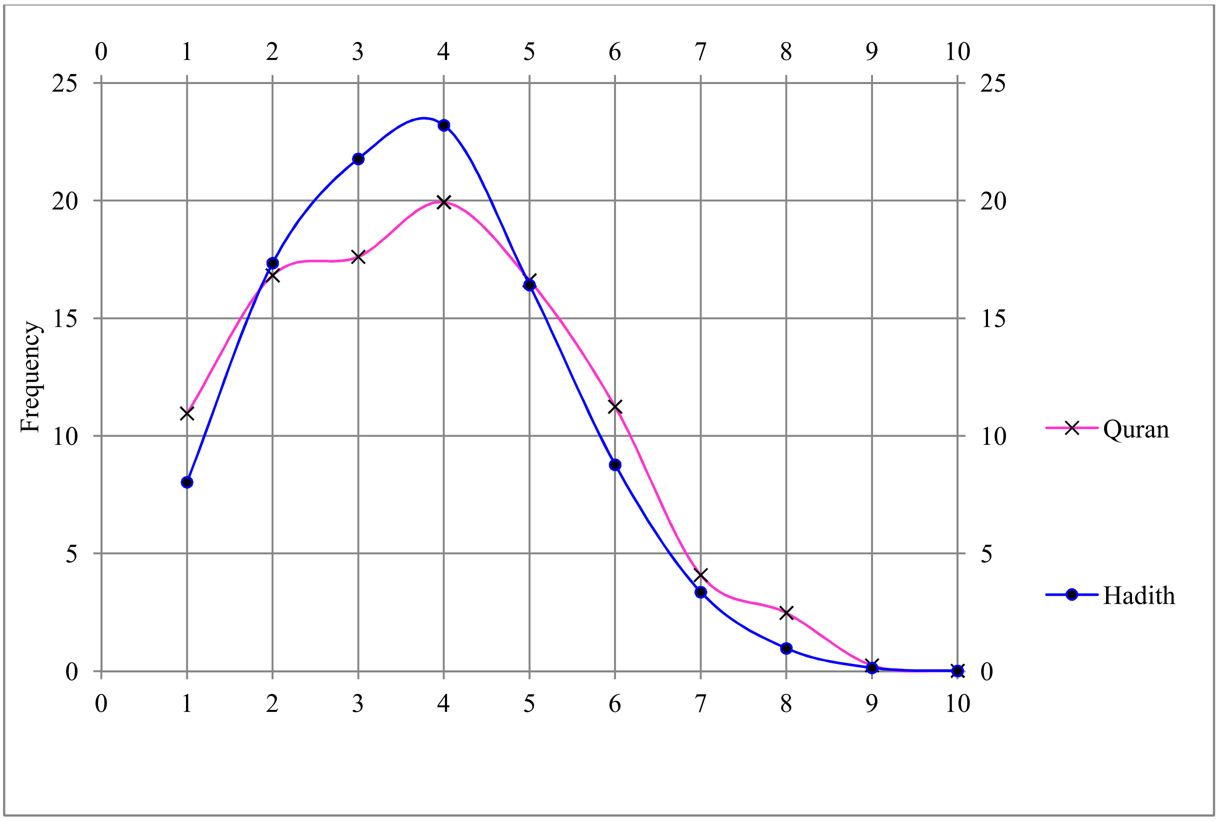


Fig. 1: Word length frequency of the Quran (in red) and Hadith (in blue), obtained with Bezier interpolation. We can notice that only the Hadith curve presents a Gaussian (or log-normal) shape.

As we can see from figure 1, in a visual point of view, the Hadith curve (in blue) respect the Guassianity, whereas the Quran curve (in red) does not seem to respect any Guassianity or at least any semi-Gaussianity (i.e. Gaussianity in one side).

By the term Gaussianity, we mean a smooth bell curve, which more or less resemble to a mono-Gaussian form in one or both sides.

We also visually notice that the ideal curve that could, maybe, ensure a certain Guassianity for the Quran is disconnected from the real Quran curve at abscises 3 and 7.

Due to that strange observation, we have decided to test if the Guassianity is also respected with other long Arabic texts (eg. testing other books/authors) or not. So, we have basically drawn the WLFs of 6 other books written by 6 different authors to see if there is any possible Gussianity or at least a log-normal shape.

Hence, several experiments of Word length frequency have been conducted on the holy Quran and the books of 7 other religious Authors, namely: the Prophet (PBUH), Dr Abd AlKafy, Dr Amro Khaled, Dr Al Ghazali, Dr Al Arifi, Dr Al Qarqdqwi and Dr Hassan. The 6 last ones represent contemporary authors from the 20th and 21th century, for which the main topic is also religion. Results are represented in figure 2.

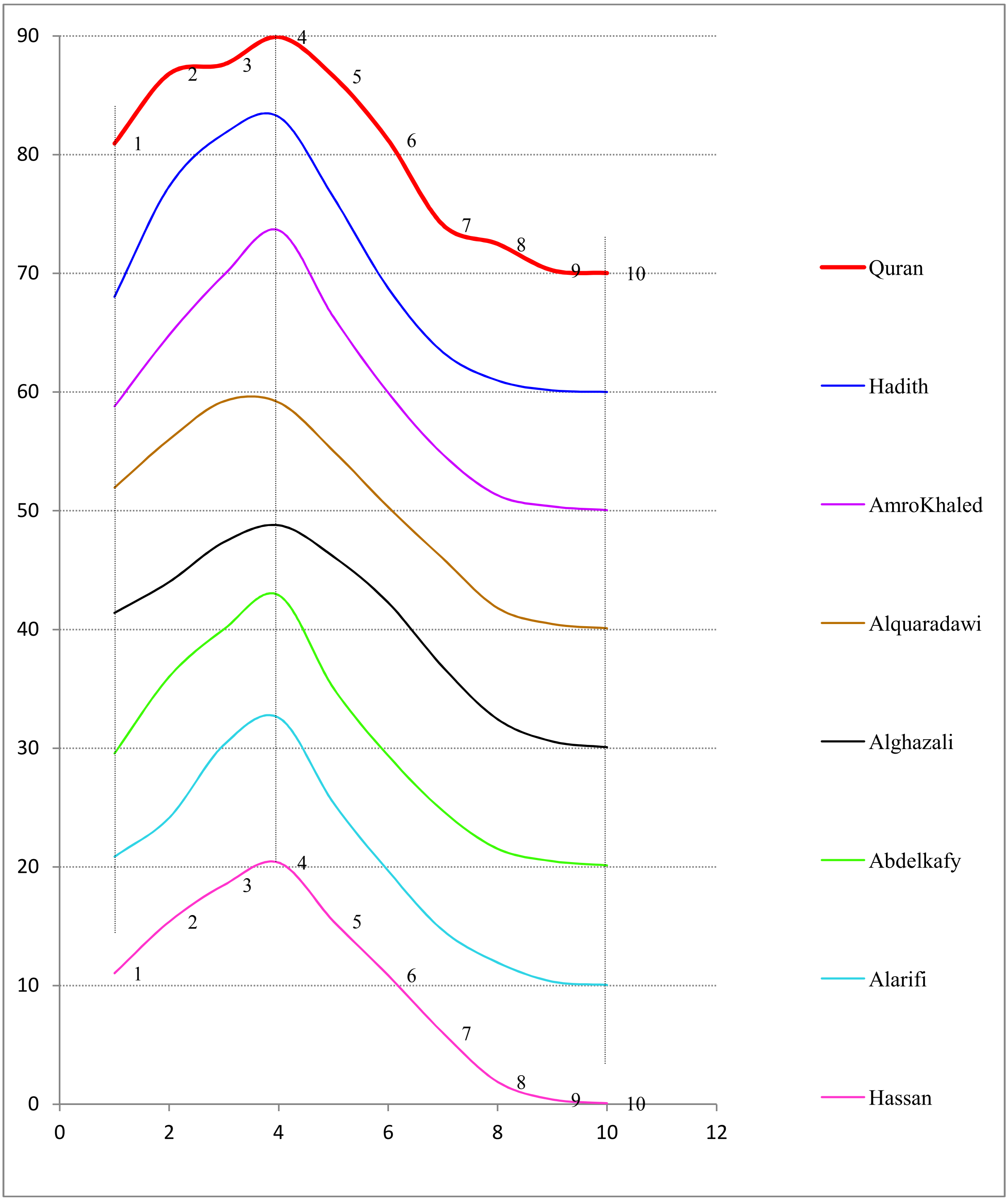


Fig. 2: The word length frequency of 8 different books: the Quran (in red), Hadith (in blue) and 6 other books written by 6 different contemporary authors, obtained with Bezier smoothing. We notice that all the curves, except the Quran, present a Gaussian or log-normal shape, at least for one of the two sides.

By observing figure 2, we notice that all the seven WLF curves present a certain gaussianity (for at least one side), except the Quran one, which has a strange graphical shape that does not seem to respect any form of gaussianity.

In a separate representation, we can see the Quran WLF with more details (see figure 3), where it clearly appears that it does not respect any Gaussianity or Interpolability. Moreover, we notice that the general form of the curve is quite strange (not familiar).

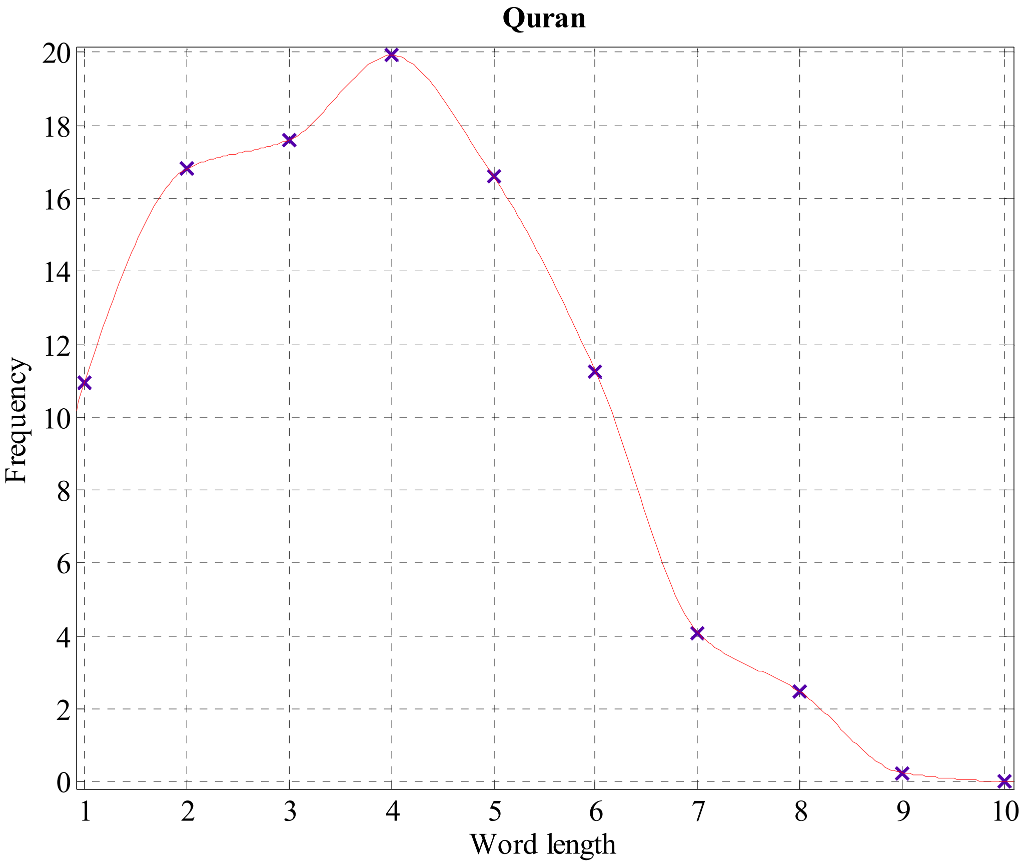


Fig. 3: The word length frequency of the holy Quran, obtained with PCHIP interpolation: we notice that the curve shape is not Gaussian and not log-normal either, in neither the left nor the right side. We also remark that it is not interpolable with conventional interpolation functions. In fact, two exceptions, precisely in 3 and 7, make the Gaussianity and Interpolability not respected.

In the previous figures (1 and 3), a strange form is noticed for the Quran WLF and a question could be asked then: is it the result of a mixture of two (or more) styles? In other terms, does the holy Quran result from a mixture of several authors? That question could be statistically stated as: is the Quran representation multi-gaussian?

To answer that question, we simulated several text mixtures and computed the corresponding approximated WLFs.

In the first experiment, we simulated the text mixture of the two authors: Dr Hassan and Dr Amro-Khaled. See figure 3.a.

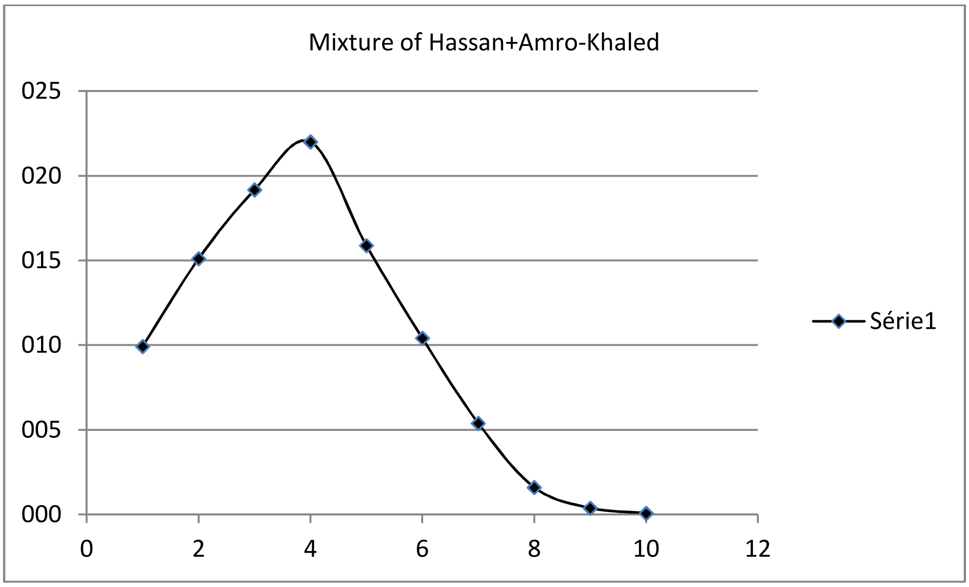


Fig. 3.a: WLF of a simulated mixture between the texts of two different authors: Dr Hassan and Dr Amro-Khaled.

In the second experiment, we simulated the text mixture of the two authors: Dr Abd-AlKafy and Dr Amro-Khaled. See figure 3.b.

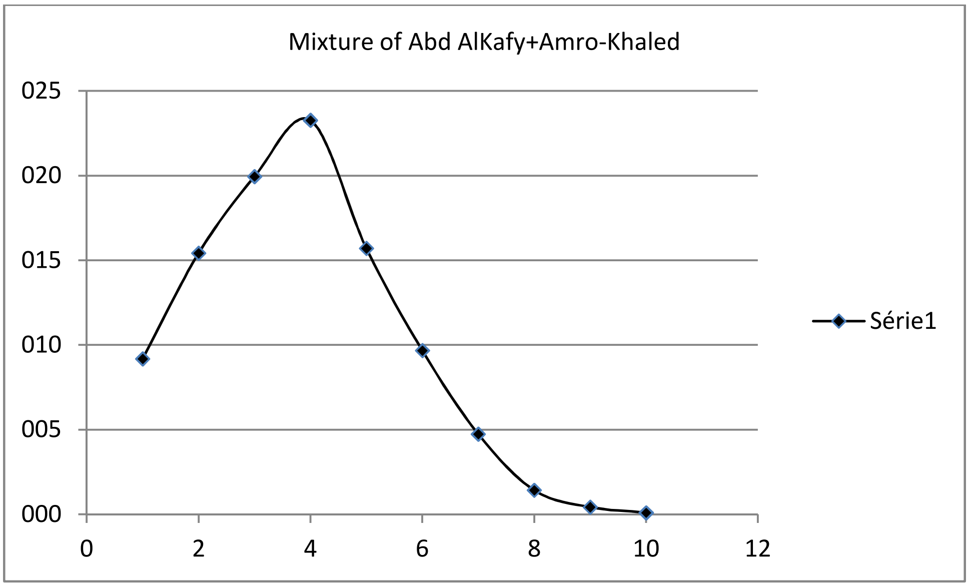


Fig. 3.b: WLF of a simulated mixture between the texts of two different authors: Dr Abd-AlKafy and Dr Amro-Khaled.

In the third experiment, we simulated the text mixture of the two authors: Dr Al-Ghazali and Dr Al-Qaradawi. See figure 3.c.

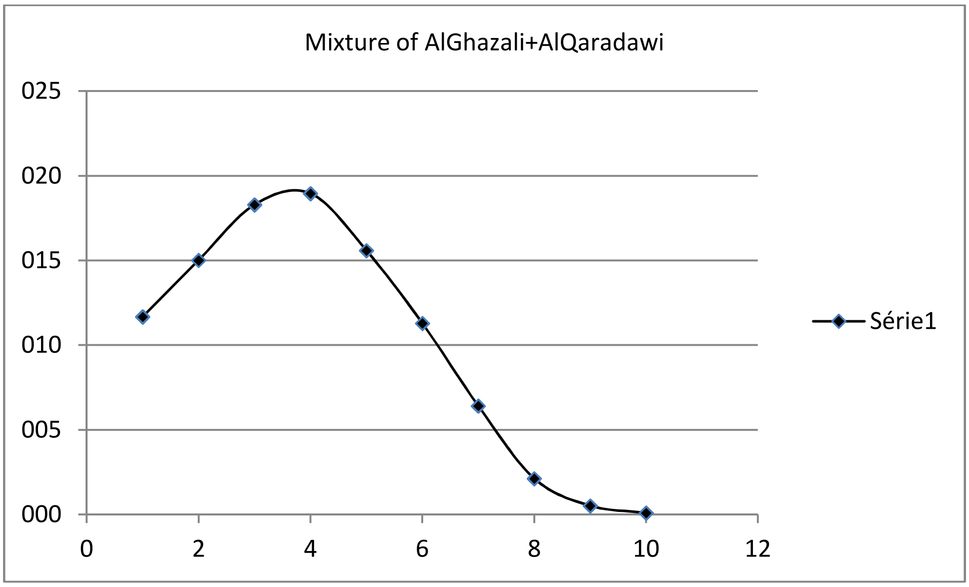


Fig. 3.c: WLF of a simulated mixture between the texts of two different authors: Dr Al-Ghazali and Dr Al-Qaradawi

In the fourth experiment, we simulated the text mixture of the two authors: Dr Al-Arifi and the Prophet (Pbuh). See figure 3.d.

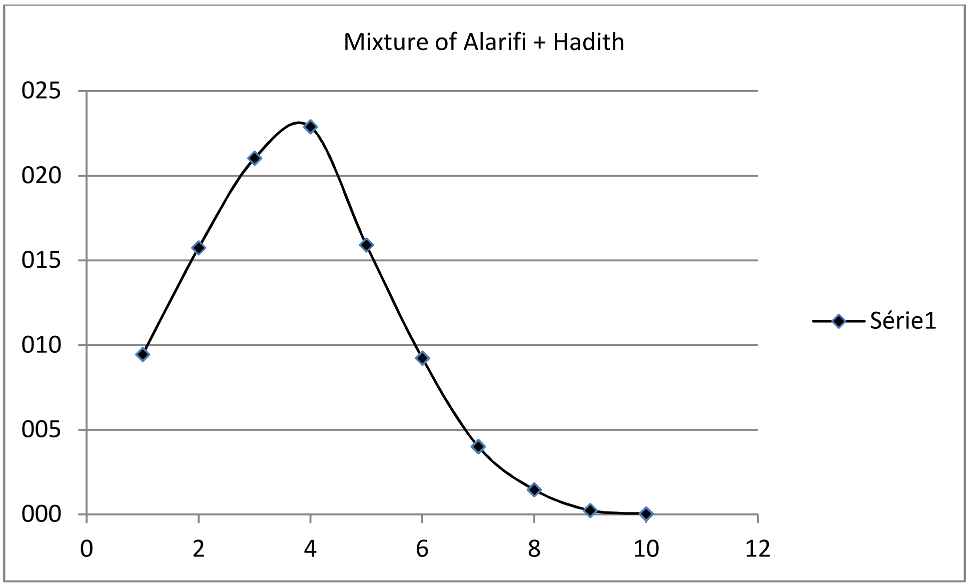


Fig. 3.d: WLF of a simulated mixture between the texts of two different authors: Dr Al-Arifi and the Prophet (Pbuh).

In the fifth experiment, we simulated the text mixture of 7 different authors: Dr Hassan, Dr Alarifi, Dr Alkarny, Dr Abdelkafy, Dr Alghazali, Dr Alquaradawi and Dr AmroKhaled. See figure 3.e.

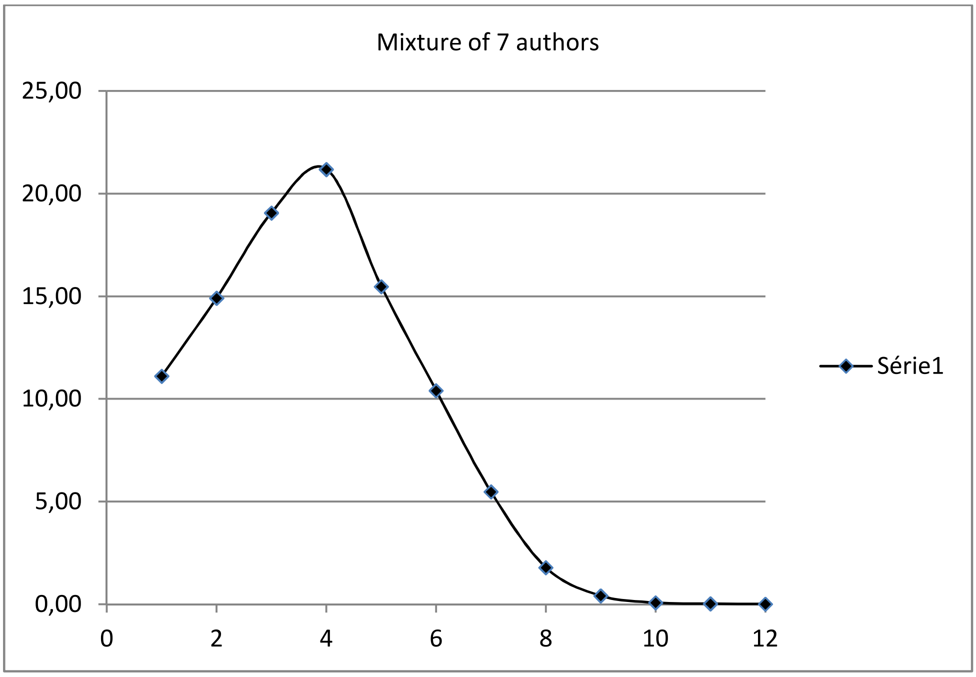


Fig. 3.e: WLF of a simulated mixture between the texts of 7 different authors: Dr Hassan, Dr Alarifi, Dr Alkarny, Dr Abdelkafy, Dr Alghazali, Dr Alquaradawi and Dr Amro-Khaled.

As we can see in all the simulated texts/authors mixture, we got always a Gaussian (on at least one side), and we do not observe any division of the curve into two or multiple Gaussian as one could expect.

This fact shows that the hypothesis of multiple styles in the holy Quran is excluded, since that phenomenon has not been noticed in the previous simulations.

So, what could be then the reason of that strange unexplained form?
In the opinion of the author, there is no classic explanation for that fact, except that the Quran should be the work of the Creator who made his holy scripture above the classic rules of statistics and mathematics.

III.3 Hadith model interpolated with Gaussian fitting

From the previous results, showing that the Hadith should respect a certain Gaussianity and Interpolability, we performed a computation of a Gaussian curve in a form given by equation 1, and optimized it to get the lowest error possible (i.e. optimal coefficients for the best fitting).

$$f(x) = a1*\exp(-((x-b1)/c1)^\wedge 2) \qquad (1)$$

The obtained results are given below:

Parameters
a1 = 23.61
b1 = 3.512
c1 = 2.497

Goodness of fit:
SSE: 2.263
R-square: 0.9969

RMSE: 0.5686

The resulting fitted curve is represented in figure 4.

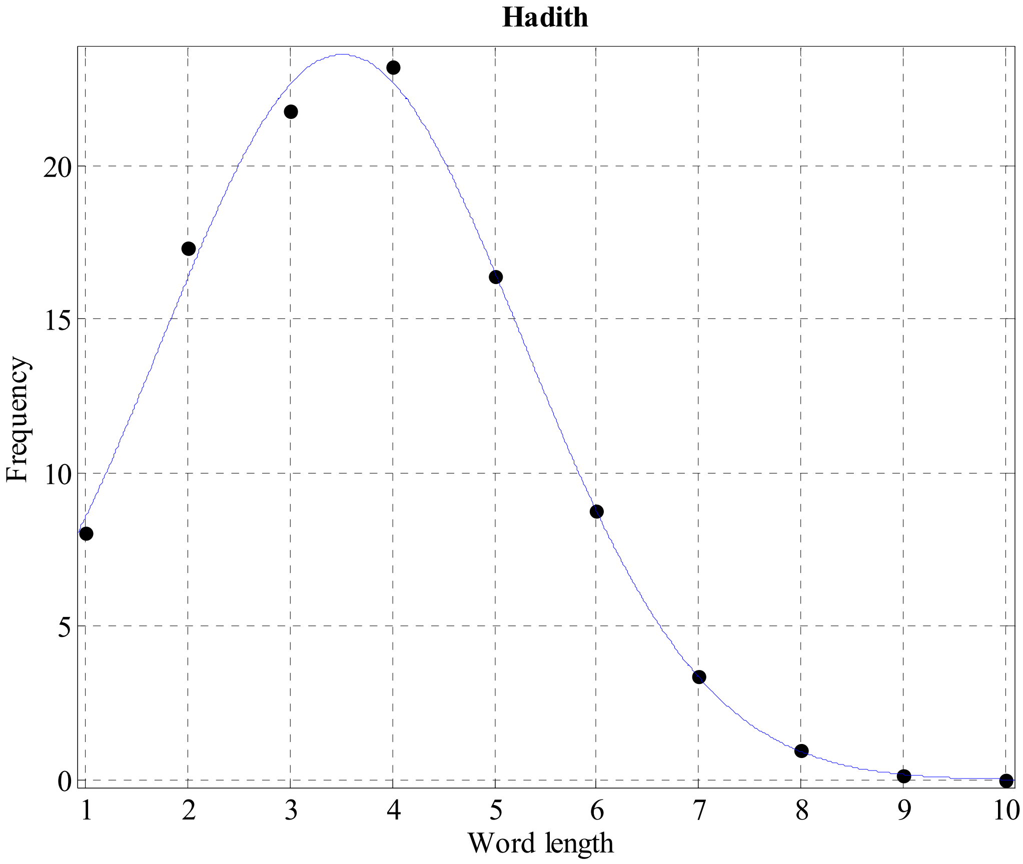


Fig. 4: The word length frequency of the Hadith , obtained with a Gaussian fitting: we notice that the Gaussian curve fits the data with a quite good precision.

Observation

From the previous results, we notice that the Hadith and all the cited books obey to the law of Gaussianity and Interpolability, except the holy Quran, which does not respect any of these classical laws. That is, we do not understand why this particular exception is noticed for the holy scripture. Again, the only interpretation, one can give, is that the concerned book should have a mysterious origin.

IV. INVESTIGATION ON NUMBERS CITATION FREQUENCY

In this second investigation, we consider the citation of numbers in the text, such as one, two, three, etc. However, those numbers are sorted from the most frequent to the least one. For concreteness, if the numbers 1, 2 and 3 have the following frequencies 10%, 15% and 12% respectively, then they will be sorted into the following sequence: 2 (*1st number*), 3 (*second number*) and 1 (*3rd number*). That scheme makes the representation curve monotone (*decreasing*) and easier to interpolate.

On the other hand, only numbers that are cited at least more than 5 times are considered, for a purpose of consistency. Consequently and in practice, only the 6 or 7 most frequent numbers are kept in the graphical representation.

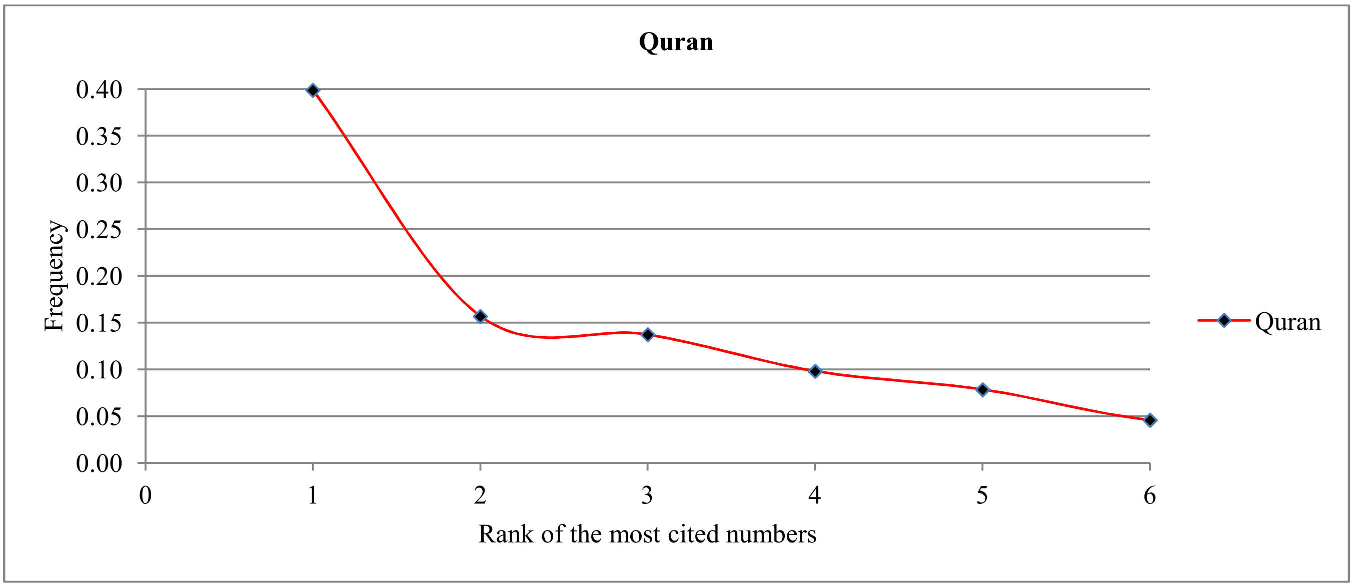


Fig. 9: Number citation in the holy Quran (*sorted from the most frequent to the least frequent*). The curve is obtained by Bezier interpolation.

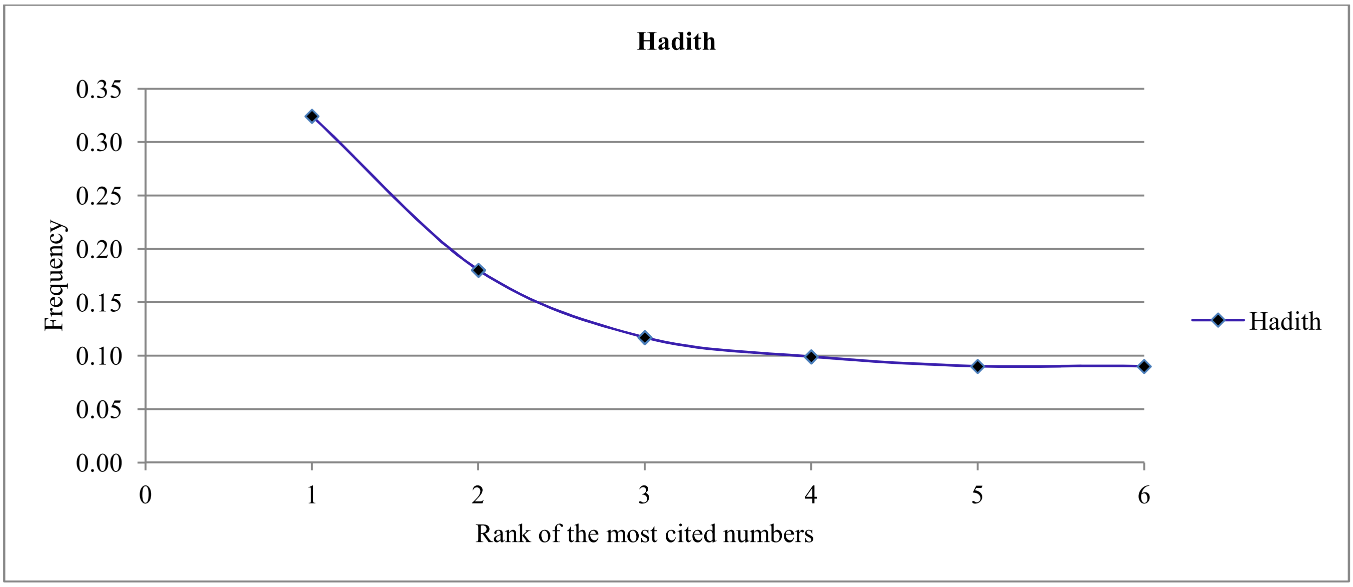


Fig. 10: Number citation in the Hadith (*sorted from the most frequent to the least frequent*). The curve is obtained by Bezier interpolation.

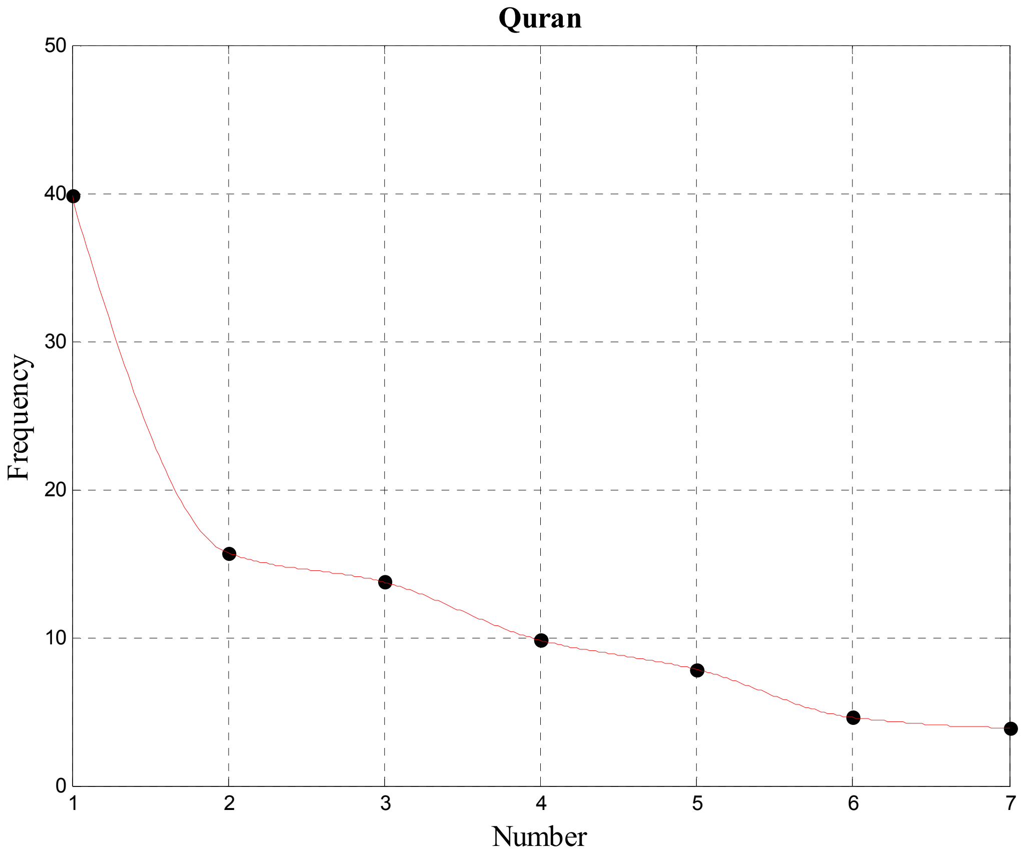


Fig. 11: Quran number frequency: there is no Gaussianity and no Interpolability either. Once again, the curve doesn't resemble to any known mathematical or physical shape and the curve shape is strange, mathematically speaking. The curve is obtained by PCHIP interpolation.

By observing figure 11, we remark that the curve presents different slopes at each segment located between two successive points. The general curve is quite strange and unfamiliar, mathematically speaking. Once again, there is no Gaussianity and no Interpolability either. Furthermore the curve does not resemble to any known mathematical or physical shape. Also, we notice that there exists a pseudo-horizontal segment between the 2nd and 3rd numbers and between the 6th and 7th ones.

Hadith model interpolated with Exponential fitting f(x) for the Number frequency

As in the previous investigation, and due to the fact that the Hadith curve appears to respect a certain Gaussianity and Interpolability (*i.e. visually*), we performed a computation of an exponential curve in a form given by equation 3, and optimized it to get the lowest error possible.

$$f(x) = a*\exp(b*x) + c*\exp(d*x) \qquad (3)$$

The obtained results are given below:

Parameters

a = 61.85
b = -0.9579
c = 8.81

d = -0.007993

Goodness of fit:
SSE: 0.472
R-square: 0.999
RMSE: 0.3967

The resulting fitted curve is represented in figure 12.

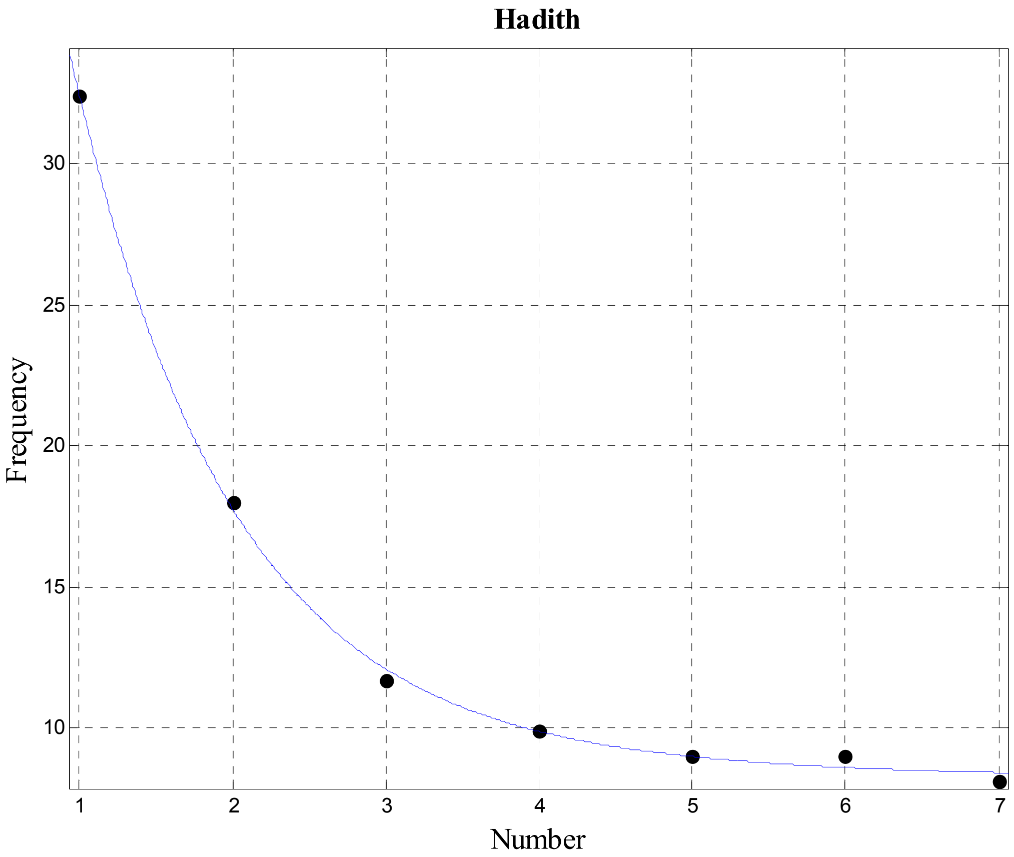


Fig. 12: Hadith number frequency: Once again, the overall curve seems to follow a partial Gaussianity shape and the Interpolability is possible for every point, by an exponential polynomial, as we can see in the corresponding fitting equation. The curve is obtained with exponential fitting.

Observation

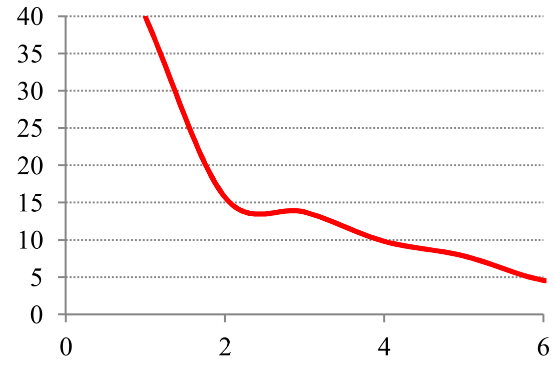
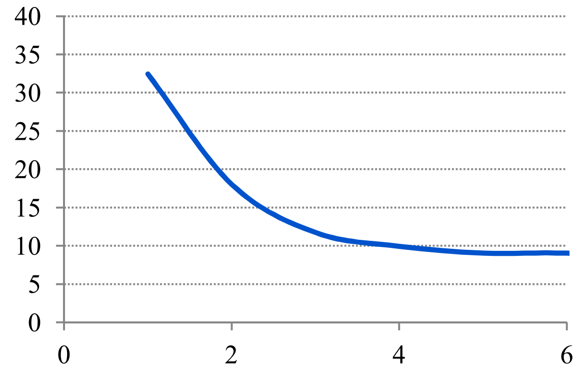
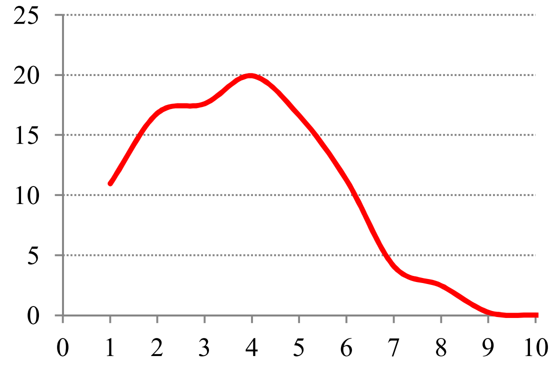
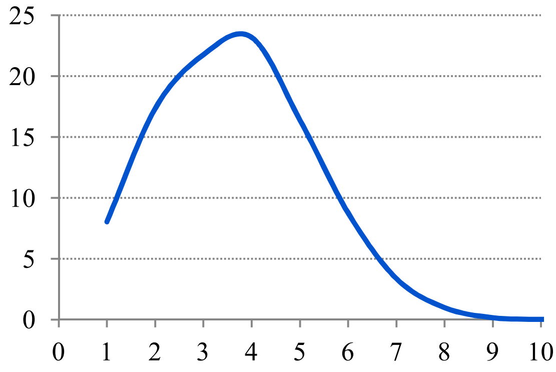
As in the previous investigation, we notice again that the Hadith obeys to the law of Guassianity and Interpolability by presenting a nice exponential shape for the sorted number citation. Contrariwise, the Quran's number citation presents a complex curve with no Guassianity or Interpolability either. This fact, once again, proposes that the holy scripture should have a mysterious origin.

V. Conclusion and Discussion

An investigation of Gaussianity and Interpolability has been conducted on the Holy Quran, in order to see if it respects, as the other human books, the physical properties of Gaussianity and Interpolability.

A summary of the different curves is given in the following table (table 2).

Table 2. Comparison between the Quran and Hadith curves (Bezier interpolation is used in both books).

	Holy Quran	Hadith
Sorted Number citation frequency		
Word length frequency		

As a comparison with other books, we positively verified that the Hadith book does obey to the Gaussianity and Interpolability rules for both Word length frequency and Number citation. Similarly, 6 other books written by different human authors have been analyzed and experimented in the same manner. Once again, those 6 different books appear to obey perfectly to the Gaussianity and Interpolability rules for the word length frequency.

On the other hand, and contrariwise, we strangely noticed that the holy Quran does not respect those rules for the word length frequency and for the number citation frequency either.

Theoretically, Gaussianity is a rule to which obey every physical phenomenon respecting the "Large Numbers" condition (*at least to the knowledge of the author*). However in the case of the holy Quran, neither the Gaussianity nor the continuity of the curve evolution (second derivative) is respected. This fact proposes that the holy Quran could not be a human invention but probably the work of a Superior Non-Human Intelligence who is beyond the prescribed rules and who does not respect any of the well known physical properties. Consequently, we may deduce two important facts:

- Firstly, the two investigated books: Quran and Hadith are quite different in terms of textual structure statistics, which leads to the conclusion that the two corresponding Authors should be different;
- Secondly and more strangely, we do not see any possible human origin for the holy Scripture. Hence, the hypothesis of a Divine origin, for the holy Quran, is widely supported by the result of this investigation.

Finally, the present paper is a pure Statistical/ Computational-Linguistic investigation regardless of the religious aspect of the studied books. It tries only to bring a new scientific discovery, which was hidden and unknown before, to the scientific community.

REFERENCES

[Siegrist, 2016] K. Siegrist, Random: Probability, Mathematical Statistics, Stochastic Processes. Department of Mathematical Sciences, University of Alabama in Huntsville. http://www.math.uah.edu/stat/sample/CLT.html. Last consultation in Fubruary, 18, 2016.

[Rice, 1995] J. *Rice (1995), Mathematical Statistics and Data Analysis (Second ed.), Duxbury Press, 1995. ISBN 0-534-20934-3).*

[Contributors, 2015] Contributors of Wikipedia. Central limit theorem. Consultation on August, 27, 2015. https://en.wikipedia.org/wiki/Central_limit_theorem.

[Galton, 1889] F. Galton (1889), Book : *Natural Inheritance*, 1889. p. 66. http://galton.org/cgi-bin/searchImages/galton/search/books/natural-inheritance/pages/natural-inheritance_0073.htm

[Sayoud, 2012] H. Sayoud. Author discrimination between the Holy Quran and Prophet's statements. Literary and Linguistic Computing 2012; doi: 10.1093/llc/fqs014, pp 427-444. Literary and Linguistic Computing, Vol. 27, No. 4, 2012.

[Whiteman, 1967] M. Whiteman. Philosophy of Space and Time and the Inner Constitution of Nature: A Phenomenological StudyRelié– 1967. New York Humanities Press.

[Milne, 2012] W. E. Milne. Numerical Calculus - Approximations, Interpolation, Finite Differences, Numerical Integration, and Curve Fitting. 2012, Cook Press, ISBN-13: 9781447457640.