Automation of the process of statistical analysis of the asymmetry of a quantitative variable

NIANGORAN Aristhophane K, ACHIEPO Odilon Yapo M., MENSAH Edoété Patrice

Abstract-- In the field of statistical analysis, the study of the asymmetry of the distribution of a quantitative variable plays a very important role in the implementation of many methods. For example, it determines the validity of the use of arithmetic mean to describe phenomena, it is one of the main elements for choosing statistical tests appropriate for real data, etc. However, to date, the techniques used to assess the asymmetry of a quantitative variable require human interpretation, whether algebraic or graphical. This need for human interpretation limits the verification of the asymmetry of quantitative variables by analysts because their exploitation requires that the analyst has a perfect command of the Exploratory Data Analysis approach as developed by John Turkey, or an advanced knowledge of the conditions of validity of the statistical methods used. As a result, many analyses are carried out implicitly assuming the symmetry of the distributions of quantitative variables, which generally leads to abusive or erroneous conclusions. This misuse of statistical methods has been exacerbated by the systematic use of calculation software and the computational aspect of new analytical branches such as Artificial Learning, Data Mining and Data Science. The purpose of this paper is to propose a method for studying the asymmetry of the distribution of a quantitative variable capable of automatically identifying the symmetry or otherwise of the said variable without human intervention. The developed method is tested with simulated data obeying theoretical laws of probability as well as with existing real data. All calculations are performed with the programming language R.

Keywords : quantitative variables, Exploratory Data Analysis, Asymmetry, Mustache Box, R Programming.

1. INTRODUCTION

When analysing a distribution, the nature of the asymmetry of a quantitative variable is very important for the quality of the conclusions [1]. A distribution is symmetrical if the observed values are evenly distributed around the three central values : mean, mode and median [2] [3]. The use of the tools such as the bar graph or the histogram, from classical statistics, makes it possible to realize the symmetrical or not of a distribution. The examination of the moustache box, a tool developed by statistician John Tukey, also gives an idea of this question depending on whether the box and whiskers are symmetrical or, on the contrary, of smaller amplitude on the left (asymmetry on the left) or on the right (asymmetry on the right) [4]. Indeed, the nature of the asymmetry of a distribution impacts the choice of analytical tools. Parametric and non-parametric indicators are used in the analysis when the distribution is symmetrical. Otherwise (non-symmetrical distribution) the use of parametric indicators is excluded; only non-parametric indicators can be used for the analysis.

In general, the techniques of EDA are all graphical [5]. The analysis of the box plot for the purpose of assessing the symmetrical or not nature of a distribution is very often subject to errors [6]. Because it requires a human interpretation of the graph but they are often misinterpreted by analysts. This subjective assessment of these tools is very often a source of error in the conclusions of the analyses. In our work, it is a question of determining the nature of the symmetry or not of a distribution of a quantitative variable automatically, without interpreting the box plot. In this paper, we propose the approach of detecting the symmetric nature of a quantitative variable distribution by software, without graph interpretation. Our method was developed using the theory of box plot. It has the advantage of providing fixed limit values to detect the symmetrical or not nature of the distribution of a quantitative variable. This detection is done without requiring a human appreciation in the form of interpretation of a graph as do the existing classical tools (density curve, histogram, etc.). Then we use Lagrange polynomial interpolation to calculate these bounds. To simulate

- *NINAGORAN Aristhophane Kerandel is currently pursuing PHD degree program in Computational Mathematics and Data Science in INP-HB Yamoussoukro, Côte d'Ivoire. E-mail:* kerandelniangoran@gmail.com
- *ACHIEPO Odilon Yapo Melaine is an Assistant Master in Artificial Intelligence in University of Korhogo, Côte d'Ivoire.* kingodilon@gmail.com
- *MENSAH Edoete Patrice is Professor in Theorical and Applied Mathematics in INP-HB Yamoussoukro, Côte d'Ivoire).* pemensah@hotmail.com

the work performed, we implement the model obtained in the R language.

2. THE MAIN ALGEBRAIC TOOLS FOR MEASURING ASYMMETRY

Several algebraic indicators are used to measure the degree of asymmetry of a quantitative distribution. Although producing numerical values, there are no limit values from which a distribution can be considered symmetrical. The main algebraic indicators used are as follows:

- **The Fisher asymmetry coefficient**

The Fisher coefficient (skweness) is the square root of the Pearson β_1 coefficient.

We have $\mu_2 = Var(x) = \sigma^2$ then the Fisher coefficient is as follows: $\gamma_1 = \frac{\mu_3}{\sigma^3}$

The interpretation of the Fisher coefficient is as follows :

- if γ_1 tends towards 0 then the distribution is symmetrical
- if γ_1 and moves away from 0 then the distribution is spread to the right;
- if γ_1 and moves away from 0 then the distribution is spread to the left.

The assessment of distance or proximity to 0 is subjective because it is left to the analyst's discretion.

- **The Yule coefficient**

The Yule coefficient is calculated from the position of quartiles Q_1, Q_2 and Q_3 and is written as follows :

$$S = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)}$$ becomes

$$S = \frac{Q_1 + Q_3 - 2Q_2}{Q_3 - Q_1}$$

- If S tends towards 0 then the distribution is symmetrical
- If $S > 0$ and moves further and further away from 0 then the distribution is spread to the right
- If $S < 0$ and moves further and further away from 0 then the distribution is spread to the left

The assessment of distance or proximity to 0 is subjective because it is left to the analyst's discretion.

- **Pearson's asymmetry coefficients**

There are two of them. The first is based on the average A and mode B. The second is defined from the centred moments of order 2 and 3.

Coefficient 1 : $S = \frac{\overline{\chi} - M_o}{\sigma}$. It is interpreted as the Yule coefficient.

Coefficient 2 : $\beta_1 = \frac{\mu_3^2}{\mu_2^3}$. It is the centered moment of order 2 squared divided by the moment of order 3 cube. μ_2 being the variance.

β_1 can only take positive or zero values.

- If β_1 tends towards 0 then the distribution is symmetrical.
- If β_1 moves further and further away from 0 and $\mu_3 > 0$ then the distribution is asymmetrically spread to the right.
- If β_1 moves further and further away from 0 and $\mu_3 < 0$ then the distribution is asymmetrically spread to the left.

The assessment of distance or proximity to 0 is subjective because it is left to the analyst's discretion.

3. THE MAIN GRAPHICAL TOOLS FOR ASSESSING ASYMMETRY

Several graphical tools are used to assess the asymmetry of a quantitative distribution. These graphical tools require a visual assessment of whether or not the distribution is symmetrical. The main algebraic indicators used are as follows :

- **The density plot**

kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. Kernel density estimate (KDE) with different bandwidths of a random sample of 100 points from a standard normal distribution. Grey: true density (standard normal). Red: KDE with $h = 0.05$. Green : KDE with $h > 0$. Black : KDE with $h = 0.0337$. Let $(x_{1,}, x_2, ..., x_n)$ be an sample drawn from some distribution with an unknown density f. We are interested in estimating the shape of this function f. Its kernel density estimator is $\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{n})$ where $K(\bullet)$ is the kernel — a symmetric but not necessarily positive function that integrates to one — and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as $Kh(x) = 1/h\,K(x/h)$. Intuitively one wants to choose as small as the data allows, however there is always a trade-off between the bias of the estimator and its variance. A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others. The Epanechnikov kernel is optimal in a minimum variance sense, though the loss of efficiency is small for the kernels listed previously, and due to its convenient mathematical properties, the normal kernel is often used $K(x) = \phi(x)$, where ϕ, is the standard normal density function.

If Gaussian basis functions are used to approximate univariate data, and the underlying density being estimated is Gaussian then it can be shown that the optimal choice for h is $h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\hat{\sigma}\, n^{-1/5}$

Where $\hat{\sigma}$ is the standard deviation of the samples.

- **The symmetry plot**

The symmetric plot was developed by John Turkey. Suppose we have a collection of values values $x_1, x_2, ...x_n$. We will say that the values are symmetrically distributed if their quantile function satisfies :

$Q(0.5) - Q(p) = Q(1 - p) - Q(0.5)$, for $0 < p < 5$.

This says that the *p*th quantile is the same distance below the median as the (1 - *p*)th quantile is above it. The obvious way to check the symmetry of a set of numbers is to plot the values

$Q(1 - p_1), ..., Q(1 - p_{n/2})$ against the values of $Q(p_1), ..., Q(p_{n/2})$. If the plotted points fall on the line $y = x$, then $x_1, ..., x_n$ are symmetrically distributed.

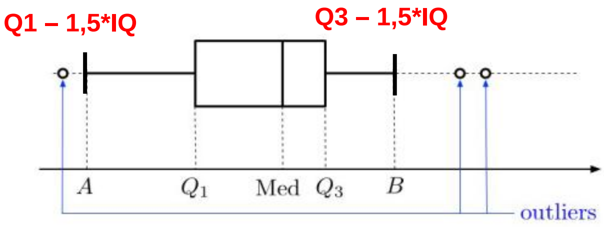
- **The box plot**

A boxplot, also known as a box-and-whisker plot, is a convenient way to graphically present numerical data. This plot is generated from the five-number summary of a distribution which consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation, written in order from smallest to largest. The boxplot was introduced by John Tukey in 1977. The center box (rectangle) of a boxplot contains the middle 50% of the ordered data. The two edges of the center box indicate the first and third quartiles. The range of the center box, which is also the difference between the third and first quartiles of the data, is usually known as the interquartile range (IQR). The only line inside the box marks the median. The line extending from the box out to the smallest observation contains the smallest 25% of observations and the line extended from the box out to the largest observation contains the largest 25% of observations. These two lines that extend from the box are known as whiskers. The far ends of the two whiskers indicate the range of the data. A symmetric distribution has two equal whiskers and a box separated into two equal parts by the median. When this is not the case, the distribution is considered to be skewed to the right or to the left. There is a provision for representing extreme values, which are determined using quartile and IQR values of the data.

- The interval quantile range (IQ) is calculated with the formula $IQ = Q3 - Q1$
- The upper extreme value limit is calculated with the formula $Q_3 + 1.5 * IQ$
- The lower extreme value limit is calculated with the formula $Q_1 - 1.5 * IQ$.

4. OPERATION OF THE BOXPLOT

Our method was developed using the theory of the moustache box.



The median: Med or Me

The quartiles : Q_1 , Q_3

$IQ = Q_3 - Q_1$: interval quantile range

By posing $\varepsilon^- = \dfrac{Q_1 - M_e}{Q_3 - Q_1}$ and $\varepsilon^+ = \dfrac{Q_3 - M_e}{Q_3 - Q_1}$ with $Q_3 \neq Q_1$ we have :

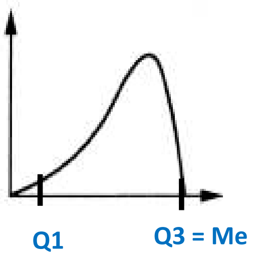
$$\Lambda = \varepsilon^- + \varepsilon^+ = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1}$$

As defined, we notice that: $\varepsilon^- \leq 0$ and $\varepsilon^+ \geq 0$.

5. THE DIFFERENT FORMS OF SYMMETRY NATURE

We use the Λ coefficient defined in a little above in each of the three cases of asymmetry.

Case 1 : Asymmetric distribution on the left



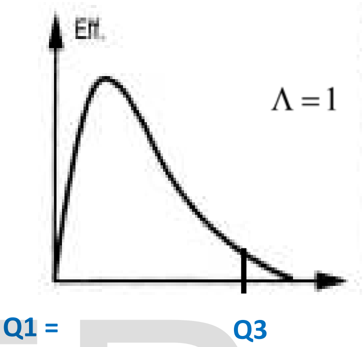
$\Lambda = -1$

In the case of a left-hand distribution, the median tends to approach the third quartile Q_3. The more the distribution is spread to the left, the more the median tends to merge with the third quartile. In the extreme case, the median coincides with the third quartile ($M_e = Q_3$). When we reach this limit, we have :

$$\Lambda = \frac{Q_1 + Q_3 - 2M_e}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Q_3}{Q_3 - Q_1} = \frac{Q_1 - Q_3}{Q_3 - Q_1} = -1$$

Indicator Λ has a minimum value of -1.

Case 2 : Asymmetric distribution on the right

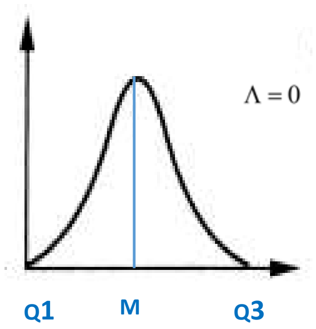


In the case of a distribution spread to the right, the median tends to approach the first quartile Q_1. The more the distribution is spread to the right, the more the median tends to merge with the first quartile. In the extreme case, the median coincides with the third quartile $(M_e = Q_1)$. When we reach this limit, we have :

$$\Lambda = \frac{Q_1 + Q_3 - 2M_2}{Q_3 - Q_1} = \frac{Q_1 + Q_3 - 2Q_1}{Q_3 - Q_1} = \frac{Q_3 - Q_1}{Q_3 - Q_1} = 1$$

Indicator Λ has a maximum value of 1.

Case 3 : Symmetrical distribution



In the case of a symmetric distribution, the median tends to position itself equidistant from the first and third quartiles. The more symmetrical the distribution, the more likely the median tends to merge with the mean value of Q_1 and Q_3. In the extreme case, the median coincides with the average of Q_1 and $Q_2(M_e = \frac{Q_1+Q_3}{2})$. When we reach this limit, we have:

$$\Lambda = \frac{Q_1+Q_3-2\frac{Q_1+Q_3}{2}}{Q_3-Q_1} = \frac{Q_1+Q_3-(Q_1+Q_3)}{Q_3-Q_1} = 0$$

The indicator has a central value of 0.

6. DETERMINATION OF THE FUNCTION FOR DETECTING THE NATURE OF ASYMMETRY

To establish our function for determining the asymmetric nature of a distribution, we used Lagrange's polynomial interpolation technique. More precisely, we wish to determine two functions Ψ^- and Ψ^+ such that :

$$\begin{cases} \Psi^-(Q_1 - 1.5IQ) = -1 \\ \Psi^-(Q_1) = -1/2 \\ \Psi^-(M_e) = 0 \end{cases}$$

$$\begin{cases} \Psi^+(Q_3 + 1.5IQ) = 1 \\ \Psi^+(Q_1) = 1/2 \\ \Psi^+(M_e) = 0 \end{cases}$$

Q being the the interval quantile range $(Q_3 - Q_1)$.

Let's pose :

$$\Psi(x) = \sum_{j=1}^{3} f_j(x) \text{with } f_j(x) = b_j \prod_{\substack{k=1 \\ k \neq j}}^{3} \frac{x - x_k}{x_j - x_k}$$

When we apply the polynomial interpolation technique, we obtain the following two equations :

$$\Psi^-(x) = \frac{(2x - Q_1 - Q_3)(x - 4Q_1 + 3Q_3)}{6(Q_1 - Q_3)^2} \quad (1)$$

$$\Psi^+(x) = \frac{(x + 3Q_1 - 4Q_3)(2x - Q_1 - Q_3)}{6(Q_1 - Q_3)^2} \quad (2)$$

The function Ψ is then given by :

$$\begin{cases} \Psi(x) = \Psi^-(x) \quad si \quad x \leq M_e(x) \\ \Psi(x) = \Psi^+(x) \quad si \quad x > M_e(x) \end{cases}$$

The following graph shows the appearance of the graphical representation of function Ψ obtained with a variable of 1000 observations according to a reduced centred normal distribution :

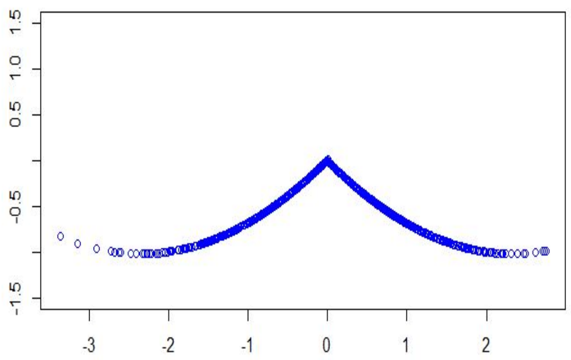


Figure 1 : Typical pace of an Ψ curve

The curve obtained with function Ψ developed for this article has a typical look of an inverted V for a symmetrical distribution.

Automatic determination of asymmetry

The automatic determination of asymmetry is based on three indicators using function $\Psi(x)$. These indicators are as follows :

- **Symmetrical centering index :**

$$iSym = [card\{x \mid \Psi(x) \in [-1/4;1/4]\}] / [card\{x \mid \Psi(x) \in [-1/2;1/2]\}]$$

- **Median connectivity index :**

$$ICC = \left\| \Psi^{+}(Me) \right| - \left| \Psi^{-}(Me) \right\|$$

- **Spread control index**
- $ETD^{-} = [1 + card\{x \mid \Psi(x) \le -1/2\}] / [card\{x \mid \Psi(x) \le -1/4\}]$
- $ETD^{+} = [1 + card\{x \mid \Psi(x) \ge 1/2\}] / [card\{x \mid \Psi(x) \ge 1/4\}]$
- $ETD = \left| (ETD^{+}) - (ETD^{-}) \right|$

The idea of our approach is based on the fact that a symmetrical distribution must have values $iSym$, ICC and ETD below simultaneously precise thresholds. To identify these thresholds, we generated two thousand (2000) series according to a reduced centred normal distribution. Thus, in principle of the law of large numbers, the limits were obtained by using the arithmetic means of the values of each parameter obtained with the simulated data. Then, these values were empirically corrected using many simulated distributions. Based on this approach, a variable is symmetric if and only if it meets the following three conditions : $iSym < 0.08$, $ECC < 0.008$ and $ETD < 0.4$.

The challenge of the proposed method is to be able to automatically detect if the distribution of a quantitative variable is symmetric or not. To do this, the approach presented above can be implemented using the following algorithm :

Algorithm 1 : Automatic Asymmetry Detection

```
1) Numerical variable input $X = (x_1, x_2, ...x_n)$
2) Calculate quartiles $Q_1(X)$, $M_e(X)$ and $Q_3(X)$
3) Calculate functions $\Psi^{-}(X)$ and $\Psi^{+}(X)$
4) Create empty sets « ES », « EN », « GMin », « GMax », « Dmin » and « Dmax »
5) Create a « Status » variable initialized with an empty character string
6) For each mode $x_i$ of $X$ :
       If $x_i < M_e(X)$ then :
           a) Calculate $\Psi^{-}(x_i)$
           b) If $\Psi^{-}(x_i) \ge -1/2$ then add $x_i$ to $ES$
           c) If $\Psi^{-}(x_i) \ge -1/4$ then add $x_i$ to $GMin$
           d) If $\Psi^{-}(x_i) \ge -1/4$ then add $x_i$ to $EN$
           e) If $\Psi^{-}(x_i) \le -1/4$ then add $x_i$ to $GMax$
       Else
           a) Calculate $\Psi^{+}(x_i)$
           b) If $\Psi^{+}(x_i) \le 1/2$ then add $x_i$ to $ES$
           c) If $\Psi^{+}(x_i) \ge 1/2$ then add $x_i$ to $DMin$
           d) If $\Psi^{+}(x_i) \le 1/4$ then $x_i$ to $EN$
           e) If $\Psi^{+}(x_i) \ge 1/4$ then $x_i$ to $DMax$
       End
   End
```

7. TESTS WITH SIMULATED DATA

Symmetrical distributions

To test the automatic detection capability of the symmetrical or non-symmetrical nature of the distributions, simulations were performed using the R programming language. The following figures show the results obtained with simulated symmetric distributions :

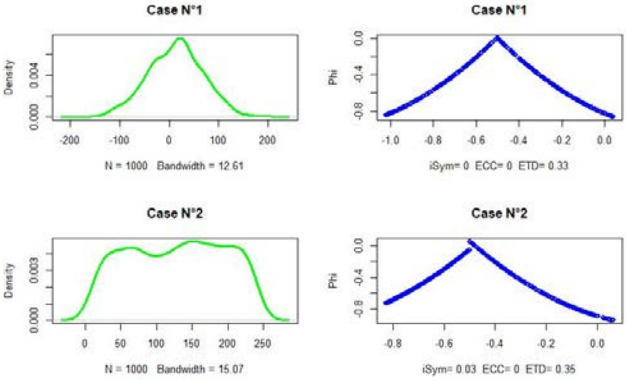


Figure 2 : Results on simulated symmetric distributions

The paces of the density curves clearly show that the simulated distributions (Case 1 and Case 2) are well symmetrical. After being subjected to the algorithm, we can see that the parameters iSym, ECC and ETD of case 1 (iSym=0, ECC=0, ETD=0, ETD=0.33) and case 2 (iSym=0.03, ECC=0, ETD=0.35) are well below the limit values (iSym*=0.08, ECC*=0.008, ETD*=0.4) ; which proves that the algorithm has detected the symmetric character of the simulated distributions.

Asymmetric distributions

To test the ability of the developed method to recognize non-symmetric distributions, two non-symmetric series were generated with R. The following figures show the results obtained with these distributions :

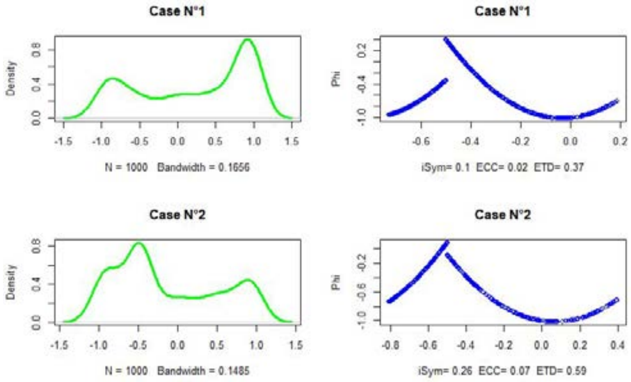


Figure 3 : Results on simulated non-symmetric distributions

The paces of the density curves clearly show that the simulated distributions (Case 1 and Case 2) are not symmetrical. After being subjected to the algorithm, we can see that the paces of the Ψ curves are not inverted V. In addition, the parameters iSym, ECC and ETD of case 1 iSym=0 and ECC=0.02 are higher than the limit values iSym*=0.08, ECC*=0.008.

The Symmetrical Illusion

Some distributions, such as contaminated distributions, may have a symmetrical appearance. However, these distributions are not a coherent phenomenon. One of the challenges of the method is to verify the ability of our approach to reject this type of symmetry. To do this, almost symmetrical contaminated distributions were generated with R. The following figures show the results obtained with such distributions :

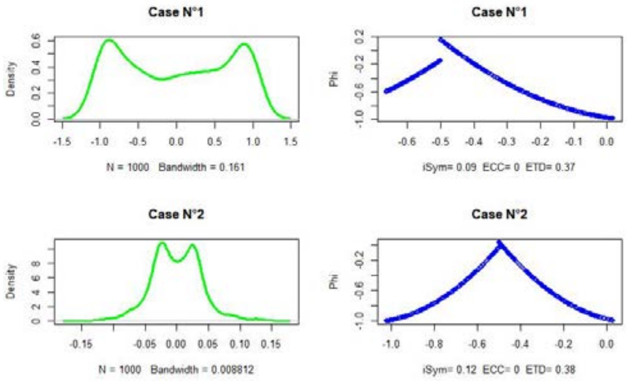


Figure 4 : Results on simulated contaminated distributions

The paces of the density curves clearly show that the simulated distributions (Case 1 and Case 2) are contaminated distributions. After being subjected to the algorithm, we find that the appearance of the A curves seems to suggest that these distributions are symmetrical. But this visual interpretation is illusory. Indeed, if we look closely at the iSym, ECC and ETD indicators, it is easy to see that these two distributions are considered as non-symmetrical by the algorithm despite their appearance. In Case 1, the iSym and ETD indicators are well above acceptable limits ; while in Case 2, it is the iSym and ETD indicators that exceed tolerable limits. It is therefore clear that the algorithm is able to detect these situations of symmetrical false distributions.

8. TESTS WITH REAL DATA

Sepal distributions

Tests on simulated data are supplemented by tests on real data. To do this, the iris database integrated into the R software was used. This database contains data on 150 iris flowers. For each iris, the length and width of the petals and the length and width of the petals were measured. The results of the sepal measurements are given in the following figure :

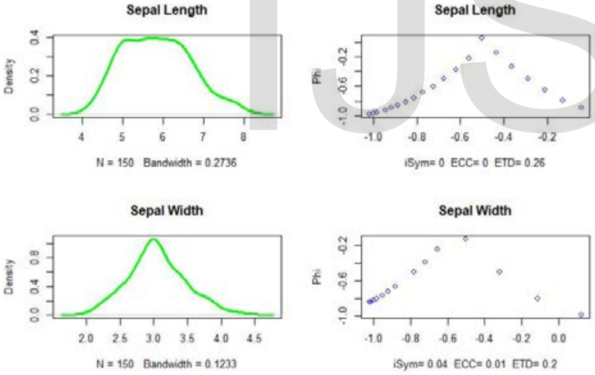


Figure 5 : Results on sepal measurements of iris data

The paces of the density curves clearly show that the lengths and widths of the petals have symmetrical distributions. The paces of the curves and the values of the iSym, ECC and EDT parameters show that the algorithm correctly identified these distributions as symmetric.

Petal distributions

The results on the petal measurements are given in the following figure :

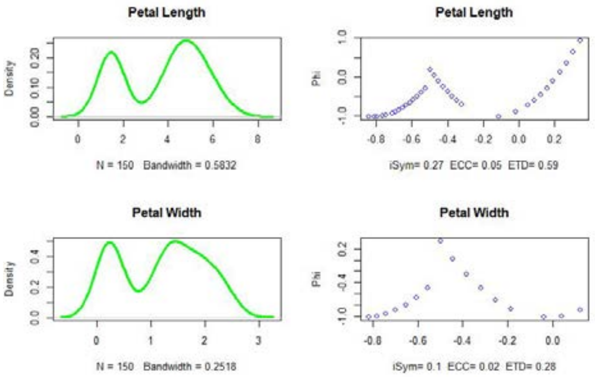


Figure 6 : Results on petal measurements of iris data

The paces of the density curves show that the lengths and widths of the sepals have non-symmetrical distributions.

CONCLUSION

The method developed in this article is based on the properties of the boxplot and Lagrange polynomial interpolation. For a symmetrical distribution, it provides a graph in the form of an inverted V. However, simulations have shown that this pattern is only characteristic of symmetric distributions in the case of uncontaminated distributions. The main objective of the method is to automate the detection of asymmetry in a distribution. Simulations prove that the method is not only able to distinguish symmetric distributions from those that are not, but also has the ability to recognize illusory symmetry situations, especially in the case of contaminated distributions. The advantage of such a method lies in the possibility of automatically testing the asymmetry of the distributions of quantitative variables, which will make it possible to automate correct data analysis methodologies in the methods developed in order to reduce analytical errors due to a lack of knowledge of the limitations and validity conditions of statistical data analysis methods. With the tool developed in this paper, future work will be able to focus on automating many analytical processes that are difficult for users to control, such as statistical testing.

REFERENCES

[1] B. Riou et P. Landais, « Principes des tests d'hypothèse en statistique: α, β et P », *Ann. Fr. Anesth. Réanimation*, vol. 17, n° 9, p. 1168-1180, oct. 1998.

[2] F. Mazerolle, *Statististique descriptive: séries statistiques à une et deux variables, séries chronologiques, indices*. Paris: Gualino, 2006.

[3] M. Carricano, F. Poujol, et L. Bertrandias, *Analyse de données avec SPSS*. Paris: Pearson Education France, 2008.

[4] R. L. Nuzzo, « The Box Plots Alternative for Visualizing Quantitative Data », *PM&R*, vol. 8, n° 3, p. 268-272, mars 2016.

[5] D. C. Hoaglin, « Exploratory Data Analysis: Univariate Methods », in *International Encyclopedia of the Social & Behavioral Sciences*, Elsevier, 2015, p. 600-604.

[6] S. Lem, P. Onghena, L. Verschaffel, et W. Van Dooren, « The heuristic interpretation of box plots », *Learn. Instr.*, vol. 26, p. 22-35, août 2013.