PROCEDURE FOR ASSESSING NORMALITY IN SMALL SIZE SAMPLES OF PHYSIOLOGICAL DATA

Ileana BARAN[1], Cristina SĂVĂSTRU[2]

This paper reports an experimental study for physiological signals acquisition, using artery applanation tonometry method. The records acquired from a small size sample (20 subjects) were measured with Omron HEM 9000 AI device. Results are analysed at univariate level in order to identify, estimate the statistical model of data and validate the parameters inferred for each of the proposed theoretical distribution laws.

Keywords: probability plots, small size samples, goodness of fit test, applanation tonometry, Omron HEM 9000 AI

1. Introduction

Cardiovascular diseases are known as the main cause of mortality, influenced by stress, age, heredity, cholesterol, physical inactivity and diabetes. To prevent the development of arterial stiffness, stroke or myocardial infarctions, is important to monitor the parameters (central systolic pressure, systolic - diastolic pressure, pulse pressure etc.) that triggers such diseases [1]. Among modern methods for determining these parameters is the arterial applanation tonometry conducted with specific instruments for arterial pulse measurements (SphygmoCor, Omron HEM 9000 AI, Watchpat, etc.).

Inspired by intraocular tonometry, the arterial applanation tonometry is useful to analyze and capture the pulse wave in a noninvasive manner. Although most of the measurements are taken from the peripheral radial artery, the brachial–radial–ulnar artery system is also an eligible area because there is a rigid structure (the bones) near the blood vessel that facilitate the uniform compression and occlusion [2].

In this study, records of vital physiological parameters (blood pressure, augmentation index - AI, heart rate, etc.) where made with Omron HEM 9000AI device, for an experimental group of 20 subjects, aged between 21-58 years. Statistical analysis of acquired data is performed using a procedure tailored for small size samples.

[1] Faculty of Power Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: leana.baran@gmail.com

[2] Faculty of Electrical Engineering, Doctoral School of Electrical Engineering, University POLITEHNICA of Bucharest, Romania, e-mail: savastru_cristina@yahoo.com

2. Physiological signals acquisition

The achieved data set, gives information about systolic pressure (SYS), diastolic pressure (DIA), pulse pressure (PP), central systolic blood pressure (cSBP), augmentation index (AI) and also about normal pulse values (PULSE).

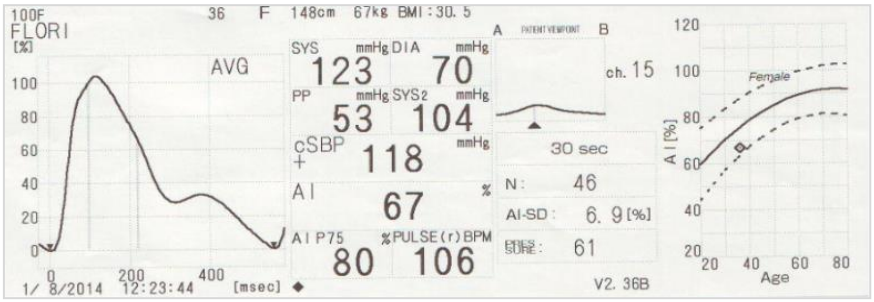


Fig. 1. Results generated by HEM Omron AI 9000

The right side of Fig. 1 is a diagram that evaluates if the recorded results are within normal limits for the patient age.

The first step for signal acquisition is adjusting the patient posture. Placing the tonometer on the wrist can sometimes be problematic. An angle of 30 degrees between the arm and the device, the existence of a rigid holder makes it easier for sensing the pulse [3].

One of the important advantages offered by HEM Omron 9000 AI is the simultaneous recording of the physiological variables listed above, opening therefore the way to multivariate techniques for the statistical analysis of the acquired data.

3. Structure of the acquired data

The experimental study targeted the population of healthy subjects without diagnosed cardiovascular diseases. The units included in the sample have been selected using a non-probability sampling technique namely the *convenience sampling*, made up of people who were easy to reach in the time frame of the experiment [4]. The main limitation of such a technique consists in the risk to obtain a sampling bias (i.e. the units selected from the population for inclusion in the sample does not reflect the population, and therefore, the sample could be unrepresentative of the population). The characteristics of the population that are of interest to this study are three anthropometric variables: the age (AGE), the height (H) and the weight (G) of the units.

The physiological variables recorded using HEM Omron AI 9000 {SYS, DIA, cSBP, PULS, AI} have been associated to the anthropometric variables of each unit. Both physiological and anthropometric variables are direct observable

variables (variables that can be observed and measured) [5] and together they form a random vector with 8 entries. The sample under study consists of 20 realizations of the observed random vector, as a single record has been performed for each of the 20 subjects of the experiment. Consequently, the data form a sample of $n = 20$ points in a space with $p = 8$ dimensions.

Table 1

Sample size structure

dimensions / *observations*	1	2	3	4	5	6	7	8
	AGE	H	G	cSBP	SYS	DIA	AI	PULS
1	y_{11}	y_{12}	y_{13}	y_{14}	y_{15}	y_{16}	y_{17}	y_{18}
…	…	…	…	…	…	…	…	…
k	y_{k1}	y_{k2}	y_{k3}	y_{k4}	y_{k5}	y_{k6}	y_{k7}	y_{k8}
…	…	…	…	…	…	…	…	…
n	y_{n1}	y_{n2}	y_{n3}	y_{n4}	y_{n5}	y_{n6}	y_{n7}	y_{n8}

The structure of the sample is made explicit in the table above. All the random variables are numerical and continuous, which implies the continuity of the random vector. All the entries of the random vector taken together have a multivariate distribution described by the joint probability density function. When isolated, each of the entries of the random vector has a univariate probability distribution that can be described by its own probability density function. This is called marginal probability density function, in order to distinguish it from the joint probability density function associated to the random vector. A more formal definition follows.

Let $Y_1,\ldots, Y_p$ be p continuous random variables forming a "p x 1" random vector. Then, the probability density function for each of the Y_i random variables with $i = 1,\ldots, p$ denoted by $f_{Y_i}(y)$ is called *marginal probability density function* and can be obtained from the joint probability density function by integrating with respect to all variables except Y_i

$$f_{Y_i}(y) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} f_Y\left(y_1,\ldots y_{i+1}, y_{i-1},\ldots y_p\right) dy_p \ldots dy_{i+1}\, dy_{i-1} \ldots dy_1 \qquad (1)$$

Many of the statistical procedures [6] performed on multivariate data are based on the assumption that the data follow a multivariate normal distribution (MVN) but, assessing multivariate normality is difficult in high dimension. A rational approach consists in analyzing the univariate marginal probability density functions followed by a bivariate analysis because if $Y \sim$ MVN, all the marginal and conditionals density functions are normal (MVN stands for Multivariate Normal Distribution). Consequently, the first step in any statistical analysis of multivariate data is to check if the marginal probability density functions defined in (1) are normal for each entry of the random vector, and if not, which deviations from normality can be identified. If for large enough sample sizes (> 30 or 40) the

violation of the normality assumption should not cause major problems (i.e. parametric procedures can be however used), for small samples such as the present one, verification of the normality assumption should be mandatory otherwise it would be difficult to draw accurate and reliable conclusions about reality [7], [8].

There are several approaches to check for normality but, for the purposes of this paper, two approaches have been selected: a graphic-analytical method based on the empirical cumulative distribution function and the use of goodness of fit tests.

4. Fitting a Univariate Distribution Using QQ Plots

When sample sizes are small, simply creating a histogram from the available data cannot be qualified as an objective method to judge the assumption of normality, because highlighting the shape of the theoretical distribution function using the sample histogram is difficult, mainly because the shape of the histogram can change significantly by simply changing the width of the class intervals [9]. For small sample sizes the quantile-quantile plots (QQ plots) may be used to assess more objectively whether data comes from a normal distribution.

For the univariate random variable $Y_{k\ (k\ =\ 1,\ 2\ ,\ldots p)}$ the available sample is formed by n observations $\{y_{k1},\ldots,y_{kj},\ldots,\ y_{kn}\}$ which rearranged in ascending order will produce the raw of sample order statistics $y_{k(1)} <\ldots< y_{k(i)} <\ldots< y_{k(n)}$, where $y_{k(i)}$ is called the i^{th} order statistic [10]. Special cases includes the minimum $y_{k(1)} = \min \{y_{kj}\}_{j\ =\ 1\ldots n}$ and the maximum $y_{k(n)} = \max \{y_{kj}\}_{\ j\ =\ 1\ldots n}$. When the sample quantiles are distinct (which, in general will be true for a continuous variable), exactly i observations will be smaller than or equal to $y_{k(i)}$.

The sample quantiles are plotted as a function of the corresponding normal order statistic medians which are defined as:

$$x_i = N^{-1}(U_i) \tag{2}$$

where U_i are the *uniform order statistic medians*, defined in [11] and N^{-1} is the percent point function of the normal distribution i.e. the inverse of the normal distribution function. It should be mentioned that when the sample quantiles are represented as a function of theoretical quantiles computed using a given distribution (normal distribution in this case), the QQ plot becomes a Probability Plot (PP plot), as it will be referred hereafter.

The uniform order statistic medians can be approximated by:

$$\begin{cases} U_i = 1 - U_n & \text{for } i = 1 \\ U_i = (i - 0.3175)/(n + 0.365) & \text{for } i = 2,3,\ldots,(n-1) \\ U_i = 0.5^{1/n} & \text{for } i = n \end{cases} \tag{3}$$

If the hypothesis of normality holds, the points $\{x_i, y_{k(i)}\}$ in the PP plot will fall along a straight line, because a PP plot compares the actual positions of the observed quantiles to their corresponding positions in the theoretical population. Departures from this straight line indicate departures from the specified statistical model in this case the normal distribution.

The advantages of working with PP plots are:

a) a quick check of the agreement between the proposed theoretical model (the normal distribution in this case) and the sample distribution,
b) it allows an easy detection of outliers and extreme values (sample values which are not within the normal behavior of the analyzed variable),
c) in case of lack of fit with the theoretical model, the PP plot highlights the nature of the deviations (for example different skewness, shorter or longer than expected tails).

A straight line can be fit to the points and added as a reference line. The intercept and slope of the fitted line are in fact estimators for the location and scale parameters of the normal distribution in a least square approach. The correlation coefficient associated with the linear fit to the data in the probability plot PP plot, i.e. the *Probability Plot Correlation Coefficient* (PPCC) [12] can be considered as a measure of the goodness of fit with the statistical model. Reference [11] offers a table of critical values that can be used as a formal test of the hypothesis that the sample comes from a normal distribution.

To find which statistical model provides the best fit for the data, different theoretical distribution functions can be used to generate PP plots; the probability plot with the highest correlation coefficient will be the best choice since it generates the straightest probability plot.

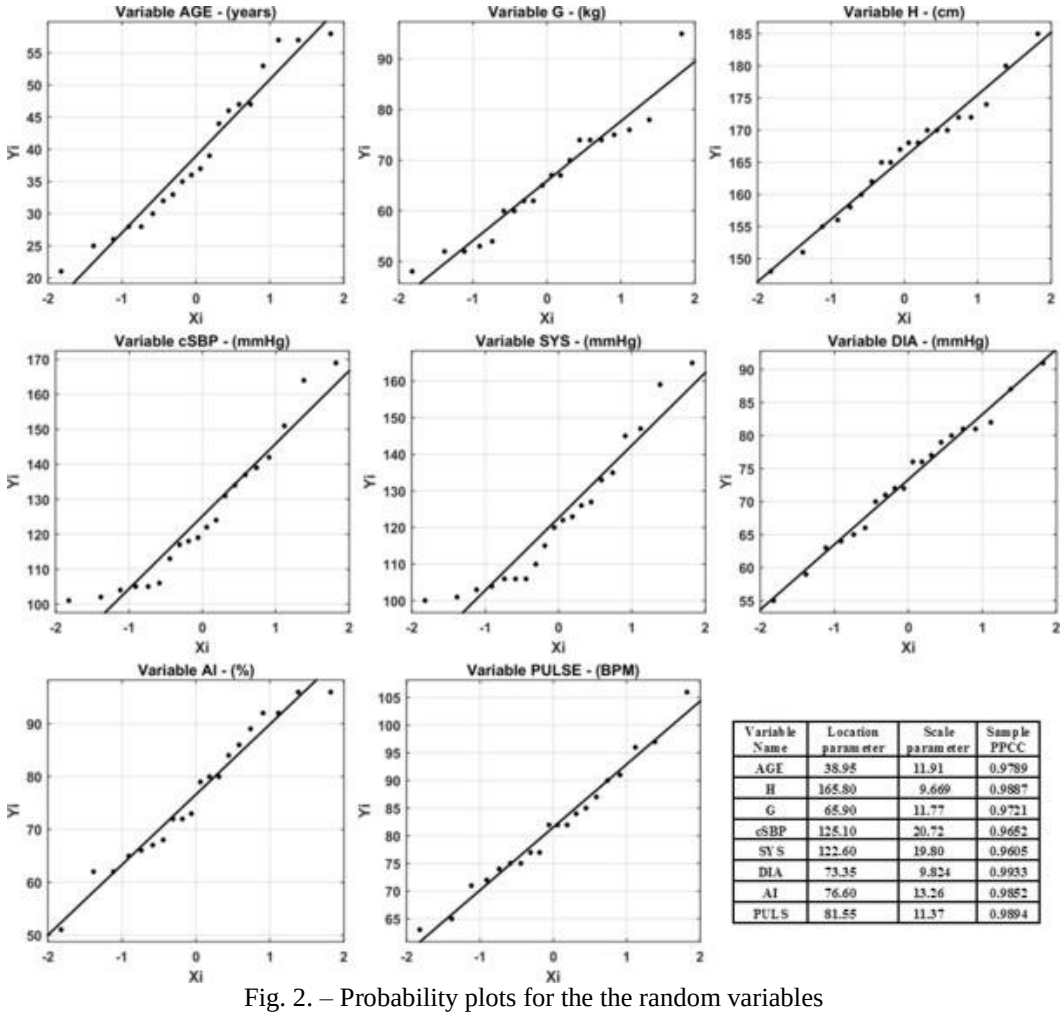
When applying the least square method, the independent variable or the regressor is the theoretical quantile from (2) while the dependent variable is the sample order statistic $y_{k(i)}$. The regression line is the model function:

$$y_{k(i)} = f(x_i; \mathbf{b}) = b_0 + b_1 \cdot x_i + r_i \tag{4}$$

where $\mathbf{b} = (b_1, b_0)$ is the vector of parameters, and r_i is a random variable named *residual* which captures all other factors influencing the dependent variable $y_{k(i)}$ other than the regressor x_i. The parameters estimated from the available data applying the least square method (LSM) contains all the information needed to identify the parameters of the normal distribution function used as statistical model. The correspondence between the vector of parameters **b** in (4) and the

parameters of the normal distribution is the following: b_1 or the slope of the regression line represents in this case the scale parameter σ (or standard deviation) of the normal distribution while b_0 the intercept of the regression line represents the location parameter (or theoretical mean) μ of the normal distribution.

PP plots have been drawn for all the random variables under study. The results are summarized in fig. 2. The figure includes also a table with the least squares estimates of the normal model's parameters (location and scale) together with the value of the sample PPCC. The visual analysis of the PP plots has not revealed the presence of extreme values or outliers, consequently, all available observations were kept. The point with G > 90 kg which seems to be an outlier was accepted after being checked using Grubb's test for outliers [14].



Variable Name	Location parameter	Scale parameter	Sample PPCC
AGE	38.95	11.91	0.9789
H	165.80	9.669	0.9887
G	65.90	11.77	0.9721
cSBP	125.10	20.72	0.9652
SYS	122.60	19.80	0.9605
DIA	73.35	9.824	0.9933
AI	76.60	13.26	0.9852
PULS	81.55	11.37	0.9894

Fig. 2. – Probability plots for the the random variables

The stronger sample linear correlation corresponds to the variable DIA, which also exhibits the best agreement with the normal distribution in the PP plot. The variables with a weaker correlation such as SYS and cSBP have also PP plots, which suggest that the data distribution has a longer left tail than would be expected under the theoretical distribution (i. e. normal) being considered. In conclusion, the interpretation of the PP plots shows that the variables have not outliers or extreme values and the null hypothesis that the data came from populations with a normal distribution cannot be rejected, even for the variables cSBP and SYS without an objective measure of the departure from normality.

4. Goodness of fit tests

As already was stated, an objective approach to check the normal distribution of data is to apply a goodness-of-fit test which**,** in general, refers to measuring how well do the observed data correspond to the fitted (assumed) model. Given a sample $y_1, y_2, \ldots, y_n$ of observations on a random variable, a goodness of fit test operates with the following hypothesis:

$H_0 : Y \in F(y)$, the null hypothesis - the model $F(y)$ fits

versus

$H_A : Y \in F(y)$, the alternative hypothesis - the model $F(y)$ does not fit

where $F(y)$ is the assumed model.

4.1 Shapiro – Wilk (SW) test for normality

The test, introduced in [15] for small samples, is based on the distribution of the following statistic:

$$W = \left[\sum_{i=1}^{n} a_{in}\, y_{k(i)}\right]^2 \Bigg/ \left[\sum_{i=1}^{n}\left(y_{k(i)} - \bar{y}_k\right)\right]^2 \quad , \quad \bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_{k(i)} \tag{5}$$

where $y_{k(i)}$, $i = 1\ldots n$ are the order statistics associated to the sample y_{kj} , $j = 1\ldots n$ from the random variable Y_k , a_{in} are suitably chosen constant coefficients, and $\bar{y}_k$ is the mean of the sample. It may be noted that if one is indeed sampling from a normal population then the numerator and the denominator of W are both, up to a constant, estimating the same quantity, namely σ^2, i. e. the variance of the population. On the contrary, for non-normal populations, these quantities would not in general estimate the same thing. The W statistic can also be interpreted as the square of the correlation coefficient between the coefficients a_{in} and the terms $y_{k(i)}$ having the same rank in the order statistics raw; consequently, W will take values between 0 and 1. For small values of the W statistic the null hypothesis is

rejected while for values approaching 1 the null hypothesis is accepted. The outcome of the SW test performed for the variables under study is given in table 2. The second and the third column display the sample mean and the sample standard deviation which are the maximum likelihood estimators for the mean and scale parameter of the normal distribution function $F(y)$ assumed as statistical model.

As it can be seen from the results summarized in table 2, the null hypothesis is accepted with a significance level going up to $\alpha = 0.10$ except for two variables, namely systolic (SYS) and central systolic (cSBP) pressure. For this two variables, the sample value of the statistic W is placed in the critical region of the test, so that the validation of the normal model requires further analysis, especially as the shapes of the PP plots and the PPCC values closer to the critical value, have already signalled possible deviations from normality.

Table 2

Shapiro Wilk test - results

Variable	Mean	SD	W	$W_{20;0.01}$ **= 0.868**	$W_{20;0.05}$ **= 0.905**	$W_{20;0.10}$ **= 0.920**
AGE	38.95	11.546	0.940	H_0 accepted	H_0 accepted	H_0 accepted
H	165.80	9.283	0.979	H_0 accepted	H_0 accepted	H_0 accepted
G	65.90	11.493	0.941	H_0 accepted	H_0 accepted	H_0 accepted
SYS	122.65	19.562	**0.910**	H_0 accepted	H_0 accepted	**H_0 rejected**
cSBP	125.15	20.376	**0.919**	H_0 accepted	H_0 accepted	**H_0 rejected**
DIA	73.35	9.388	0.984	H_0 accepted	H_0 accepted	H_0 accepted
AI	76.60	12.779	0.960	H_0 accepted	H_0 accepted	H_0 accepted
PULS	81.55	10.904	0.979	H_0 accepted	H_0 accepted	H_0 accepted

4.2 The Filliben test of normality

To refine the analysis concerning the validity of the normal model, a second goodness of fit test has been applied, the PPCC test for normality introduced in [11] and referred as the Filliben test for normality. The approach make use of a new test statistic for the composite hypothesis of normality (i. e. location and scale of the distribution both unspecified and therefore replaced by estimates of this two parameters), namely the *probability plot correlation coefficient* PPCC, which had been already introduced in relation with the PP plots. The test statistics is computed using relation (6), where $y_{k(i)}$ is the i-th order statistic for the sample on the random variable. Y_k and x_i, computed with (2) is a measure of location $\mathrm{loc}(X_i)$ of the i-th order statistic from a standardized normal distribution. The Filliben test uses as a measure of location the order statistic medians for reasons which are argued in [11]:

$$r = \frac{\sum_{i=1}^{n}\left(y_{k(i)} - \bar{y}\right)\cdot\left(x_i - \bar{x}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_{k(i)} - \bar{y}\right)^2 \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}, \quad \bar{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_{k(i)}, \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (6)$$

If the hypothesis H_0 is true, the theoretical distribution function of the r statistic (i. e. the probability to have $r < r_p$) depends upon the sample size n. This particular distribution function has been taken over from [11] and given in table 3.

Table 3

Theoretical distribution function of the r statistic for n = 20 observations

$p = \Pr(r < r_p)$

r_p	0.452	0.912	0.925	0.939	0.950	0.960	0.972
p	0.000	0.005	0.010	0.025	0.050	0.100	0.250
r_p	0.981	0.987	0.991	0.992	0.994	0.995	0.995
p	0.500	0.750	0.900	0.950	0.975	0.990	0.995

The test results for the Omron data set are summarized in table 4. The sample mean value reported in the second column of the table and the sample standard deviation (SD) reported in the third column are point estimates of the location and scale parameters of the populations under study.

Table 4

Results for the Filliben Test

Variable	Mean	SD	***Sample PPCC***	$r_{n=20;\alpha=0.05}$	$r_{n=20;\alpha=0.10}$	Observation
AGE	38.95	11.721	0.9789	0.950	0.960	
H	165.80	9.530	0.9887	0.950	0.960	
G	65.90	11.601	0.9721	0.950	0.960	
DIA	73.35	9.679	0.9933	0.950	0.960	$PPCC > r_{n,\alpha}$
AI	76.60	13.061	0.9852	0.950	0.960	H_0 is accepted
PULS	81.55	11.201	0.9894	0.950	0.960	
SYS	122.65	19.562	**0.9605**	0.950	0.960	
cSBP	125.15	20.376	**0.9652**	0.950	0.960	

The decision is to accept the hypothesis H_0 at a significance level $\alpha = 0.10$ for all the random variables tested, even for the variables SYS and cSBP which failed to pass the SW test for the $\alpha = 0.10$ significance level. The acceptance of the hypothesis H_0 is justified, because for alternative asymmetrical function to the normal distribution, characterized by sample asymmetry coefficients between 0.6

and 1.0, the power of the Filliben test is higher than the power of SW test, as has been demonstrated in reference [11]. For the variables SYS and cSBP, the Pearson asymmetry coefficients are 0.725 respectively 0.689, and consequently, the decision to accept H_0 is covered by the power of the Filliben test.

5. Conclusion

Subjects who participated in this experiment form a relatively representative group for the population of interest in terms of anthropometric characteristics (age, weight, height). Next to the anthropometric variables, five physiological variables (cSBP, SYS, DIA, AI and PULSE), considered as cardiovascular indicators, have been observed and simultaneously recorded using an Omron HEM 9000AI device. The advantage of simultaneity offered by the device, creates the opportunity to apply to the sample formed by the records, different multivariate analysis techniques.

The main contribution of the paper consists in the solutions proposed for performing the univariate analysis of data (the first necessary step of any multivariate approach) based on the empirical cumulative distribution function and not on the empirical density function. Although frequently used in other areas, techniques based on the empirical cumulative function are not common in biostatistics even though they present a number of advantages outlined in the paper. This approach allows the assessment, with a good confidence level, of the normality of small correlated data samples, a goal difficult to achieve by other means.

The normality of the marginal univariate distributions was checked using probability plots (PP) combined with the Filliben goodness of fit test (based on the probability plot correlation coefficient, PPCC), as an alternative to the more commonly used Shapiro-Wilk test. The proposed procedure can be used for small size samples and can discriminate between the normal model and other alternative asymmetric models. The use of PP plots facilitates the identification of extreme values and outliers, if any, and reveals the presence and causes of deviations from the normal distribution.

Further research will perform the bivariate analysis of the same data, to see if partial correlations can be considered as normal, followed by a factor analysis or a principal components analysis. These techniques are directed towards clarifying the nature and the structure of the correlations between the random variables under study (i.e. the entries of the random vector formed by the anthropometric and physiological variables put together).

Acknowledgement

The authors acknowledge the support granted by Dr. Mihai Chetan. The work was conducted at MED.CO S.R.L Company.

R E F E R E N C E S

[1]. S. *Mendis, P. Puska, B. Norrving*, Global Atlas on Cardiovascular Disease Prevention and Control, World Health Organization, Geneva 2011, ISBN: 978 92 4 156437 3.

[2]. *C. Savastru, A. Morega, M. Morega*, Sensing the Hemodynamic of the Brachial-Ulnar-Radial Arterial System, Proc. of the 8th Intl. Conference and Exposition on Electrical and Power Engineering, Iași, Romania, October 2014, ISBN: 978-1-4799-5848-1.

[3]. *** HEM 9000AI Omron, Non-Invasive blood pressure monitor with augmentation index, Instruction manual, 2006.

[4]. *F. J. Gravetter., L. A. Forzano*, Research Methods for the Behavioral Sciences, Australia: Cengage Learning, 5th ed, Jan 1, 2015, p. 147.

[5]. *L. Spinei, O. Lozan, V. Badan*, Biostatistica (Biostatistics), Univ. de Stat de Medicină şi Farmacie „Nicolae Testemiţanu", Şcoala de Management în Sănătate Publică, Chișinau 2009 (Î.S. F.E.-P. „Tipografia Centrală"), p 186, ISBN 978-9975-78-743-7.

[6]. R. Y. Jiang, Study on probability plot correlation coefficient of the log - Weibull distribution, Springer, **Vol. 20**, Issue 3, June 2015, pp 298–301, doi:10.1007/s12204-015-1625-8.

[7]. K. Hajian-Tilaki, Sample size estimation in diagnostic test studies of biomedical informatics. J Biomed Inform., **Vol. 48**, Apr. 2014, pp. 193-204, doi:10.1016/j.jbi.2014.02.013.

[8]. J. C. Oliver-Rodríguez, X. T Wang, Non-parametric three-way mixed ANOVA with aligned rank tests, British Journal of Mathematical and Statistical Psychology, **Vol. 68**, 2015, pp. 23–42, doi:10.1111/bmsp.12031.

[9]. W. Chantarangsi, W. Liu, F. Bretz, S. Kiatsupaibul, A. Hayter, and F.Wan, Normal probability plots with confidence, Biometrical Journal, **Vol. 57**, 2015, pp 52–63. doi: 10.1002/bimj.201300244.

[10]. *R. V. Hogg, A. T. Craig*, Introduction to Mathematical Statistics, 3rd ed. New York: Macmillan, 1970, p. 146.

[11]. *J. J. Filliben*, The probability Plot Correlation Coefficient Test for normality, Technometrics, **Vol 17**, No 1(Feb., 1975), pp. 111-117.

[12]. *V. Clocotici*, Dicționar de statistică (Statistics glossary), 2009, available at http://profs.info.uaic.ro/~val/statistica/StatGloss.htm.

[13]. *** Statistica, User Guide, 2016, available at http://www.statsoft.com/Textbook.

[14]. *F. E. Grubbs*, Procedures for Detecting Outlying Observations in Samples, Technometrics, No. 1, **Vol. 11**, Feb. 1969, pp. 1-21.

[15]. *S. S Shapiro, M. B Wilk*, An Analysis of Variance Test for Normality (Complete Samples), Biometrika, **Vol. 52**, No. 3/4. (Dec., 1965), pp. 591-611, available at sdittami.altervista.org/shapirotest/ShapiroTest.html.

[16]. *C. Dragomirescu., J.W. Drane.*, Biostatistică pentru începători (Biostatistics for beginners), **Vol I**, Biostatistică descriptivă, Ediția a 6°, Editura CREDIS, București 2009, p 207, ISBN 978-973-734-461-8.