ALTERING THE DECISION BY DISTORTING THE PRESENTATION OF STATISTICAL INFORMATION

Viorel ȚARCĂ[1], Dumitru POPA[2], Elena ȚARCĂ[3]

[1]PhD, ″Apollonia″ University of Iaşi, Romania
[2]Prof. PhD, ″Apollonia″ University of Iaşi, Romania
[3]Lecturer PhD, "Grigore T. Popa" University of Medicine and Pharmacy of Iaşi, Romania
Corresponding author: Viorel Țarcă; e-mail: tarcaviorel@hotmail.com

Abstract

In any managerial process, the decision must first of all be well-founded scientifically, based on authentic information, properly synthesized by applying methods and techniques that lead to a result based primarily on the rigorous logic of the phenomenon under investigation and not statistically significant at any cost.Distortion of statistical information refers to the usually intentional distortion of the essence of a data set by the controlled extraction of most of the component units of a sample (violating the principles of random selection) - the so-called "cherry picking" and the distortion of the graphical representation used in the result dissemination process, in order to communicate a certain idea about the whole community under analysis in terms of the vision of a reality different from that existing in the statistical model of concrete data. Another way to apply the "cherry picking" process in combination with "data dredging" (looking for a potential correlation at any cost) is to search the statistical significance of the intensity of the link between two or more investigated variables.The aim of the paper is to present some common procedures in practice, through which the process of distorting statistical information is carried out in order to minimize the possibilities of occurring decisions that are fundamentally inappropriate.

Keywords: *distortion of statistical information, statistical error, cherry picking, data dredging statistically significant.*

1. INTRODUCTION

From the oldest times people tried to come up with various explanations specific to the natural phenomena, initially appealing to knowledge through mythology. The first steps of the long road which led to the scientific method that we currently refer to took place in ancient Greece and belonged to Aristotle, the father of logics, who presented the basic principles of reasoning, the conclusion being supported by the application of some rules of inference. The scientific processes are currently based on a generalisation of the results that researchers obtained from their experiments and they were turned into different theories, which objectively contribute to the sequential formation of knowledge (LAHOZ-BELTRA, 2021).

In the development and continuous perfection of the scientific method, an extremely important role was held by the concept of hypothesis, the scientist's development of a question inevitably generating the need for a preliminary experiment, which would represent the first step in starting the research necessary to obtain an expected answer. The data initially obtained following the experiment represent the basis for the proposal of a provisional explanation (statistical hypothesis), a stage in which the descriptive and exploratory statistics analysis techniques play a determined role. Subsequently, in the second research stage another experiment shall be projected, with the role of testing this assumption. The experimental information represents the basis for the assessment of the proposed scientific model, by confirming or infirming, according to each particular case, of the initial hypothesis. If the obtained statistical data lead to supporting the hypothesis, the researcher might generalize them, using the inductive reasoning as a fundament. Also, in order to validate the model and the inductive reasoning used within the scientific method, two requirements have to be met: the experiment has to be reproductible by any other independent researcher and the results obtained have to objectively have both the confirmation and the rejection of the initial labour hypothesis as their finality (WAYNE & CHAD, 2013).

In every statistical research we can identify the following four main stages (ȚARCĂ, 1998):

1. *Obtaining old and new statistical information*, necessary for the researched phenomena;
2. *Their multilateral processing* in order to determine the necessary synthetical indicators, characterised under different aspects of the analysed phenomena or processes;
3. *The analysis of the statistical indicators,* obtained through processing, in order to establish and to measure the connection between various phenomena, in order to get to their scientific explanation to the theoretical generalisations under the form of laws and statistical regularities;
4. *A phenomena prediction*, based on the results of the statistical analysis in order to use them in the decisional process.

In the current period most statistical research, due to a significant data volume which has to be recorded, is inevitably confronted with the differences between the recorded levels of some characteristics and their real value. Therefore, the information collected has to go through an ample verification and correction process in order to discover and eliminate most statistical errors.

2. ERRORS WHICH MAY INFLUENCE A STATISTICAL STUDY

Over the course of the development stages of statistical research one might encounter the following types of statistical errors (ȚIȚAN et al., 2021):

Observation errors, which in their turn can be:

a) *accidental*, with an unpremeditated character, because of a lack of attention or the appearance of some accidental causes; they take place in both directions in regard to the real values of a phenomenon and come under the action of the law of large numbers, compensating each other at the level of the whole, thus not influencing in an essential way the average results;

b) *systemic*, which take place intentionally and act in a well-determined direction, always distorting the final result; they determine a deviation of the real values in one direction and that is why they cannot be compensated at the level of a community. In different situations they can be triggered by malevolence or premeditated (the intention to distort reality, to exaggerate/diminish the proportions of a phenomenon or collective process from nature or society). They are usually caused by a low understanding level of the instructions of a statistical survey, by the convenience of the person responsible for data collection, bad faith, and sometimes by a misunderstanding of the purpose of the survey.

The representativity errors, specific to surveys, represent the difference between the value of a synthetic indicator – means, variance – obtained on the basis of the data from the sample and the dimension of the same indicator (an estimator of the desired parameter) determined on the basis of a total recording, provided that the recording error is the same in both cases. These types of errors appear in statistical research because of the fact that the structure of the general community can never be perfectly reproduced at the level of the sample taken from it. The estimator or "the statistics calculated at the sample level" is therefore an assumption of the true value of the parameter followed in the studied community, constituting without a doubt a statistical hypothesis. They can be of two types:

a) random (they occur during all statistical surveys by sampling, appearing even in the situation of very strict observance of the principles regarding the selection of each statistical unit in the sample);

b) systematic (they appear in the situation when the principles of random selection are violated in the constitution of the sample, in the situation when a statistical unit is introduced in a subjective, preferential way in the sample).

Modeling errors, which occur in the situation of an unjustified choice of a statistical calculation model, which results in obtaining indicators devoid of real content, hidden behind a mathematical calculation formula without any practical purpose. This type of error can be eliminated by properly using the verification procedures for the statistical hypothesis and the significance tests for the value of the indicators.

3. DISTORTING THE PRESENTATION OF INFORMATION – A MORE "SPECIAL" ERROR

Regardless of their type, errors might appear at any state of the complex collection, processing, analysis, synthesis process of statistical data, obviously influencing the decision-making process at the level of various managerial processes.

Unlike the error, distorting the presentation of statistical information refers to the usually intentional distortion of the essence of a data set by the controlled extraction of most of the component units of a sample (in violation of the principles of random selection) – the so-called "cherry picking" and to the deformation of the structure of the graphical representations used in the result dissemination process, with the purpose of communicating a certain idea referring to the whole community under analysis through the vision of a reality different from the one existing in the statistical model of concrete data. The distorted presentation of statistical information usually appears in the final part of the statistical research, respectively at the moment of the presentation of results. It is very interesting the fact that, although the data collection and processing stages took place accordingly, the final information is transposed in a deceiving manner, which can negatively influence subsequent decisions.

Therefore, the basic principle regarding the graphical presentation of statistical information is distorted, as well as the well-structured emphasis of the intimate essence of statistical data which allows clear, precise and efficient communication of complex ideas, allowing the beneficiary or the viewer to optimally absorb and assimilate a huge volume of information. In this context, the idea presented by Andy Field, professor at the University of Sussex, is highly suggestive. He says that "the way in which the graphical data is presented makes a huge difference when it comes to the message broadcasted to the public" (FIELD, 2009).

Statistical graphics offer valuable help in order to get inside the profound structure of the data logical model, "in order to offer a more attractive and precise presentation of information, to synthesize the existing relationships among variables, to help more clearly highlight existing trends and to clarify, from a visual point of view, a series of significant differences" (RUNYON, 1982).

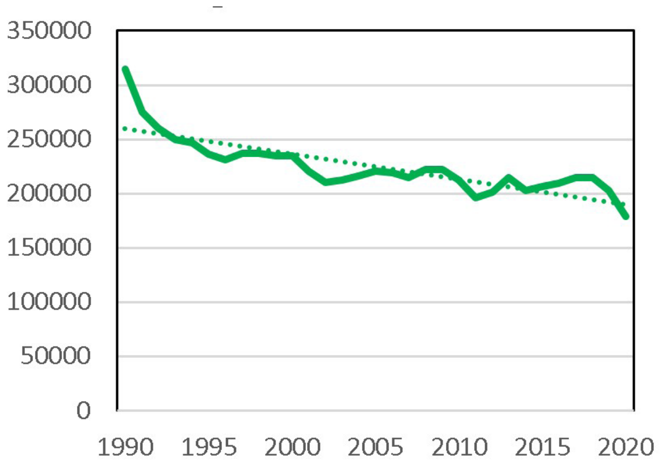
4. PRACTICAL WAYS OF DISTORTING THE PRESENTATION OF THE STATISTICAL DATA ESSENCE

4.1. GRAPHICAL DISTORSION

The improper graphical representation models of statistical data were developed a long time ago, "highlighting a wide variety of interesting and inventive schemes" (WAINER, 1984). Used very often, the "cherry picking" procedure refers to the voluntary selection of those levels of studied variables which support a certain statistical hypothesis preferred by the researcher to the detriment of other values that might lead to its rejection. This represents a deliberate action which discredits the entire statistical research, misleading the beneficiaries of the study by highlighting only a part intentionally cut from the phenomenon studied and hiding certain features that could highlight a general trend in a completely different direction.

We can exemplify a slightly simplified variant of this procedure, through the graphical representation of a statistical time series, which presents the evolution of the number of newly born in Romania, between 1990 and 2020, where one can notice a clear decreasing trend compared to the 2011-2018 period which, taken out of the context, seems to highlight a slightly ascending evolution (Figure 1).

A. The complete series: 1990 – 2020



B. The fractional series: 2011 - 2018

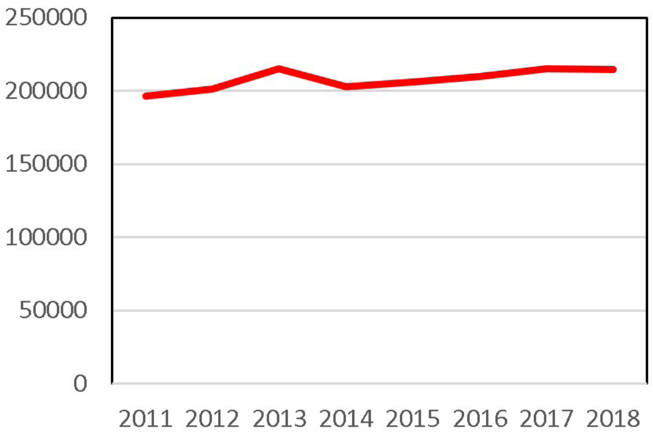
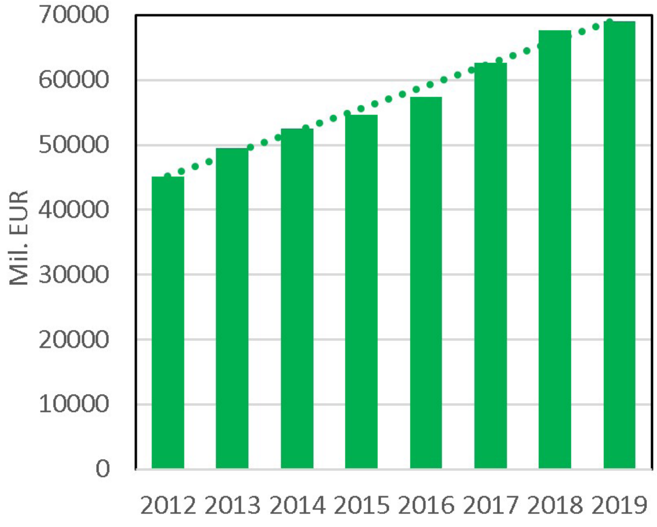


Fig. 1 The evolution of newly born in Romania (INSSE, n.d)

The distortion of the structure of the graphics with the purpose of exaggerating an evolutive tendency in a certain direction represents another side of the statistical distortion process which may directly influence the decisional process by creating the appearance of a phenomenon with a much higher magnitude than in reality. This procedure was firstly described by Darrell Huff in his book "How to Lie with Statistics" (HUFF, 1954), which, according to the former president of the American Institute of Statistical Mathematics, in 2010, was "the best-sold statistics book of the last 50 years" (STEELE, 2005).

A. The correct variant



B. The distorted variant

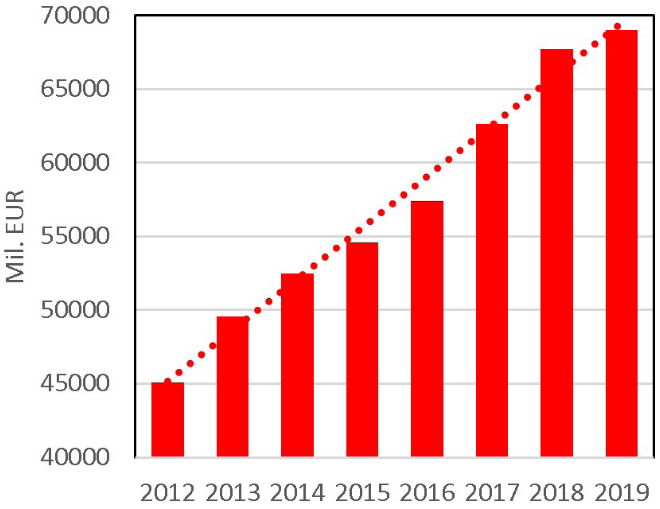
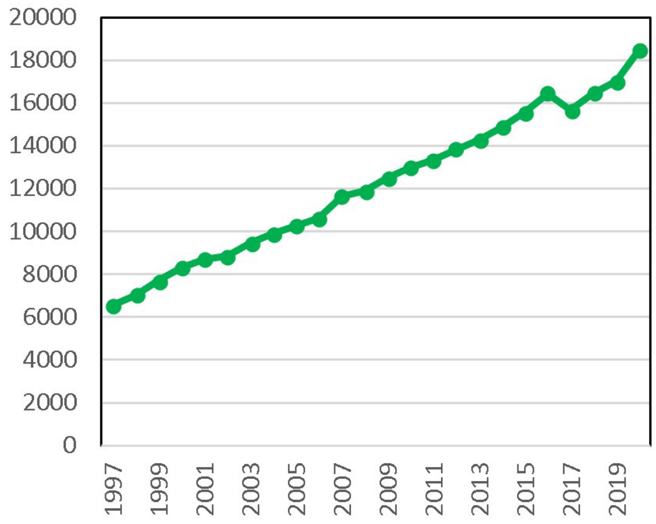
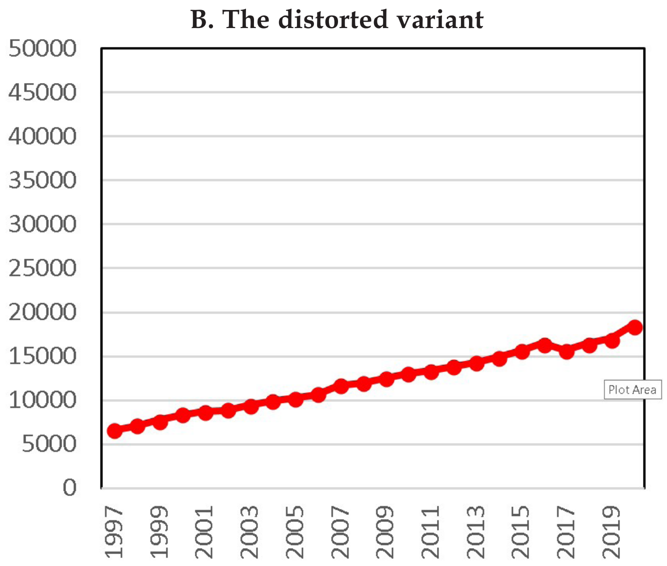


Fig. 2 The value of Romania's exports between 2012-2019 (INSSE, n.d)

As one can notice from Fig. 2, the same correctly represented data (variant A) show a slight increase in the value of the indicator, which the distorted variant (B) highlights, in an unreal manner, its highly significant increase. The explanation is very simple, in the case of our example, in which the vertical axis starts from 40000, the slightly low authentic differences among groups (the annual export values) appear as being huge. When a look, the graphic in figure 3, variant B, seems to incorrectly illustrate an increase in the value level of Romania's exports between 2012-2019, of almost 6 times, instead of a real growth of approximately 50%, as one can notice in the correct variant A.

A. The correct variant





B. The distorted variant

Fig. 3. The evolution of the total number of dentists in Romania between 1997-2020 (INSSE, n.d)

The distortion, from a graphical point of view, of the presentation of statistical data may also be achieved in the opposite direction, in the sense of minimising the tendencies recorded by the statistical data regarding a phenomenon or a group of phenomena, by using a distorted scale for the dependent variable (the vertical axis), in order to highlight a maximum level much higher than the real value of the model presented (Figure 3, variant B).

According to the information freely offered by the National Statistics Institute with the help of the TEMPO-Online database (INSSE, n.d), between 1997-2020, the total number of dentists, who work both in the public and in the private sector, increased by approximately 2.8 times, reaching a value of 18491 in 2020, an aspect which can be correctly observed in figure 3, variant A, but which is easily ridiculed in variant B of the same graph, the increase having an insignificant aspect.

4.2. "APPARENT" STATISTICALLY IGNIFICANT CORRELATIONS

Another way of applying the "cherry picking" procedure in combination with "data dredging" (the search for potential correlations at any cost) refers to the statistical significance of the intensity of the bond between two or more researched variables. According to the Biostatistics Encyclopaedia for medical professionals, published by the Indian professor and researcher Abhaya Indrayan, the founder and head of the Biostatistics and Medical Computer Science Department, from the University College of Medical Sciences in Delhi, India, the concept of "data dredging" refers to "the examination of the comparisons from a set of data that were not explicitly planned prior to the start of the study; also known as data fishing it is a form of data mining, in which large volumes of information are explored to discover any possible relationship between the variables. This technique is often described as an attempt to find more information in the datasets than they actually possess" (INDRAYAN & HOLT, 2017).

Using the p level of significance, which shows the likelihood of a particular event to take place by chance (in practice we usually use $p < 0.05$), one can discover various correlations among variables even if they appear due to hazard. In order to offer an example, we shall use a set of 12 variables, each of them being comprised of 40 observations, randomly generated (with values between 0 and 1) within a specialised computer science programme (Table 1).

Table 1. Randomly generated numerical statistical variables

A	B	C	D	E	F	G	H	I	J	K	L
0.394	0.058	0.801	0.184	0.719	0.337	0.487	0.135	0.588	0.846	0.480	0.486
0.418	0.991	0.386	0.073	0.373	0.702	0.001	0.539	0.370	0.021	0.765	0.265
0.468	0.546	0.958	0.484	0.384	0.592	0.802	0.972	0.052	0.251	0.517	0.002
0.331	0.347	0.829	0.557	0.643	0.668	0.357	0.246	0.547	0.988	0.086	0.332
0.595	0.978	0.075	0.986	0.329	0.505	0.648	0.042	0.427	0.002	0.215	0.454
0.715	0.524	0.229	0.208	0.303	0.128	0.593	0.377	0.827	0.763	0.986	0.945
0.177	0.512	0.457	0.921	0.697	0.872	0.525	0.439	0.709	0.303	0.961	0.495

0.495	0.504	0.175	0.223	0.549	0.365	0.755	0.004	0.918	0.016	0.736	0.144
0.076	0.141	0.225	0.027	0.664	0.310	0.738	0.203	0.957	0.817	0.600	0.691
0.286	0.457	0.890	0.656	0.506	0.156	0.636	0.303	0.361	0.001	0.626	0.909
0.842	0.735	0.216	0.037	0.129	0.671	0.948	0.215	0.564	0.875	0.179	0.579
0.054	0.327	0.631	0.569	0.335	0.046	0.736	0.962	0.912	0.333	0.609	0.025
0.189	0.673	0.946	0.047	0.610	0.935	0.566	0.920	0.536	0.373	0.614	0.075
0.364	0.453	0.252	0.079	0.432	0.988	0.595	0.127	0.867	0.934	0.866	0.809
0.886	0.728	0.372	0.563	0.817	0.566	0.121	0.775	0.504	0.502	0.210	0.746
0.851	0.800	0.597	0.945	0.605	0.133	0.742	0.645	0.713	0.312	0.888	0.220
0.303	0.088	0.119	0.462	0.309	0.908	0.447	0.386	0.460	0.032	0.043	0.404
0.573	0.126	0.928	0.971	0.514	0.798	0.925	0.012	0.716	0.637	0.609	0.177
0.999	0.842	0.398	0.275	0.069	0.420	0.471	0.398	0.923	0.750	0.028	0.517
0.413	0.648	0.209	0.326	0.070	0.959	0.142	0.336	0.329	0.267	0.782	0.989
0.809	0.721	0.475	0.574	0.335	0.894	0.653	0.896	0.503	0.574	0.055	0.071
0.327	0.467	0.998	0.971	0.084	0.411	0.533	0.924	0.536	0.209	0.283	0.297
0.901	0.224	0.072	0.753	0.920	0.473	0.165	0.464	0.913	0.106	0.994	0.229
0.054	0.345	0.551	0.351	0.242	0.507	0.704	0.246	0.885	0.850	0.200	0.888
0.946	0.898	0.020	0.724	0.612	0.052	0.006	0.128	0.842	0.629	0.876	0.939
0.428	0.696	0.842	0.142	0.537	0.082	0.423	0.285	0.967	0.479	0.867	0.487
0.409	0.978	0.032	0.525	0.844	0.149	0.109	0.291	0.841	0.005	0.827	0.459
0.386	0.290	0.947	0.639	0.075	0.004	0.537	0.289	0.175	0.143	0.871	0.880
0.423	0.542	0.669	0.103	0.676	0.781	0.481	0.843	0.597	0.464	0.947	0.131
0.892	0.812	0.334	0.193	0.933	0.238	0.388	0.235	0.281	0.527	0.385	0.176
0.477	0.674	0.950	0.398	0.260	0.034	0.546	0.235	0.580	0.932	0.489	0.134
0.528	0.847	0.302	0.839	0.972	0.348	0.820	0.871	0.062	0.144	0.220	0.508
0.182	0.393	0.725	0.619	0.297	0.363	0.796	0.270	0.237	0.956	0.249	0.174
0.142	0.788	0.364	0.957	0.279	0.117	0.627	0.764	0.549	0.762	0.427	0.425
0.005	0.040	0.035	0.564	0.406	0.488	0.480	0.952	0.261	0.211	0.381	0.250
0.582	0.143	0.248	0.730	0.893	0.795	0.352	0.141	0.112	0.288	0.151	0.497
0.476	0.655	0.112	0.556	0.202	0.151	0.843	0.130	0.375	0.485	0.037	0.605
0.216	0.030	0.783	0.035	0.993	0.424	0.463	0.469	0.485	0.856	0.239	0.288
0.890	0.746	0.919	0.262	0.682	0.947	0.302	0.660	0.147	0.729	0.506	0.064
0.674	0.968	0.892	0.034	0.564	0.192	0.031	0.639	0.442	0.858	0.082	0.581

Source: the 12 variables were generated in EXCEL, by using the =RAND() function, which returns a random, evenly distributed real number greater than or equal to 0 and less than 1

We shall have a number of (12 x 11) / 2 = 66 possible correlations, one for each pair of random variables, that we include in the matrix of coefficients, in table 2.

Table 2. The matrix of correlation coefficients

	B	C	D	E	F	G	H	I	J	K	L
A	**0.48**	-0.17	0.02	0.15	0.00	-0.26	-0.12	0.00	0.05	-0.03	0.02
B		-0.13	-0.01	-0.09	-0.16	-0.26	0.12	-0.03	-0.09	0.02	0.04
C			-0.07	-0.07	-0.04	0.15	0.25	-0.20	0.26	-0.03	-0.31
D				-0.04	-0.13	0.17	0.08	-0.16	-0.37	-0.03	-0.05
E					0.04	-0.29	0.02	-0.02	-0.06	0.14	-0.20
F						-0.05	0.11	-0.20	-0.02	-0.12	-0.18
G							-0.04	0.00	0.13	-0.17	-0.19
H								-0.26	-0.19	-0.06	**-0.41**
I									0.20	0.38	0.14
J										-0.26	0.09
K											0.11

By analysing table 2 we notice that the two values of the correlation coefficients between the variable pairs randomly generated (A and B) and (H and L) seem to be significant ($p < 0.01$) following the processing with the help of the statistical analysis programme R, an aspect that may seem impossible at first sight given the working hypotheses.

The chance to obtain an extreme correlation, as it is, for example, the one between A and B, where the value of the correlation coefficient is 0.48, presenting a positive connection, of medium intensity, between the two quantitative variables, is of (1 / 66) = 0.015 (a value situated much below the 5% limit), a fact which seems to prove that the analysed effect did not appear by chance, as it is a significant one from a statistical point of view.

If we look closely at Figure 4, which shows the distribution of the percentiles (parameters that structure the distribution of the 66 correlation coefficients in 100 equal parts), we notice that the aforementioned value (r = 0.48) actually appeared due to chance, being intentionally chosen by the person conducting the research, as that extreme level, corresponding to the 99th percentile, where only 1% of the elements of the series are located, which allows us to highlight a statistically significant pseudo-value ($p < 0.01$). Also, if we perform a statistical analysis using the nonparametric correlation methods, the distributions compared not being normally distributed, the correlation coefficient of the Spearman ranks reaches a level of 0.53 and that of Kendall 0.39, both being highlighted by any specialised software as having a high statistical significance $p < 0.001$, the relationship between the two variables (in our case A and B) being able to be regarded as a viable result from a scientific point of view.

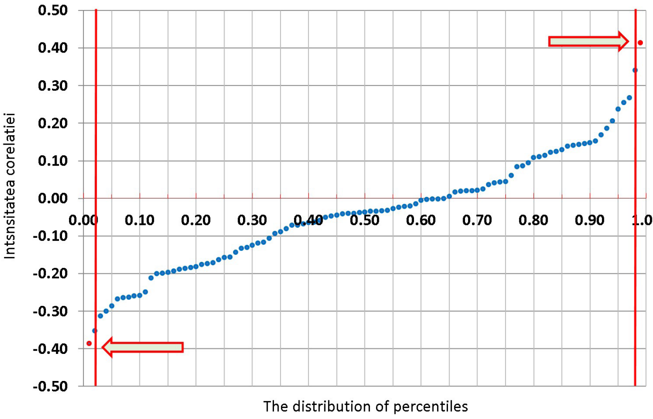


Fig. 4. The distribution of percentiles for the correlation coefficients

The search for a statistical correlation at any cost with the different variable sets analysed, goes, at most times, beyond the logic of the statistical phenomenon of explaining, first and foremost the reason for which that connection exists, intentionally giving way to arbitrary correlations. The traditional statistical methods on which any experiment is based have as a starting point the defining of the work hypothesis (H_1), followed by the analysis of the data collected, in an attempt to validate it. Completely opposed to the primordial significance of the traditional methods, the "data dredging" procedure represents the obtaining of some random correlations that best fit the desired hypothesis, most of the times without offering an analysis based on logic which may enter the profound intimacies of the connections between the analysed phenomena (INDRAYAN & HOLT, 2017).

In all practical situations, the value of the statistical significance will identify the chance as the result of a particular event which can or cannot be guaranteed with a certain level of trust. The statistical game between the null hypothesis H0 (the lack of a correlation in our case) and the alternative hypothesis H1 (the possible existence of a link) is of probabilistic essence, two types of errors planning on the final result of the research, respectively the erroneous acceptance of the work hypothesis when it is not true or its rejection when it is true. In this context, it is worth mentioning the leading article of Arthroscopy: The Journal of Arthroscopic and Related Surgery (4.772 impact factor), published in April 2021, in which it is clearly stated the fact that "the statistical significance dichotomizes the research results into significant versus insignificant ones, creating a false sense of certainty" (COTE, et al., 2021).

5. CONCLUSIONS

Summarizing, in any managerial process, the decision has to firstly be very well-founded scientifically, based on authentic information, properly synthesized by applying methods and techniques that lead to a result based primarily on the rigorous logic of the phenomenon under investigation and not statistically significant at any cost. Also, special attention, although at first sight it seems less important, has to be given to the correct graphical representation of statistical information, so that it can fulfil its fundamental role, that of supporting some viable managerial decisions.

In the decision-making process, managers issue hypothesis that they need to scientifically test regarding the parameter/parameters which synthetically characterize a community / study population or the distribution law that the different random valuables fallow within research.

Also, the research activity inevitably implies a series of ethical principles, such as seriousness, professional, moral and social responsibility, respect towards work and the topic analysed, sincerity and a fair cooperation among the members of the research team, in order to communicate the results of the research following their thorough verification, in order to finally generate a pure scientific value for that respective study.

References

COTE, M.P., LUBOWITZ, J.H., BRAND, J.C. & ROSSI, M.J. (2021) Misinterpretation of P Values and Statistical Power Creates a False Sense of Certainty: Statistical Significance, Lack of Significance, and the Uncertainty Challenge, Arthroscopy: The *Journal of Arthroscopic and Related Surgery*, 37(4), pp. 1057-1063.

FIELD, A. (2009) *Andy, Discovering Statistics using SPSS (Third Edition)*. London: SAGE Publications Ltd.

HUFF, D. (1954) *Darrell, How to Lie with Statistics*. London: Penguin Books.

INDRAYAN, A. & HOLT, M.P. (2017) Concise Encyclopedia of Biostatistics for Medical Professionals. Boca Raton:CRC Press, Taylor & Francis Group.

INSSE (n.d) The figures were processed according to the TEMPO-Online Database of the National Institute of Statistics. Available from:*http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table* [12 January 2022].

LAHOZ-BELTRA, R. (2021) *Behind the big data scenes. A statistical approach* [in Romanian]. Bucureşti: Litera Publishing House.

RUNYON, R. (1982) *Audrey Haber, Business Statistics*. Burr Ridge, IL: Richard Irving, Inc.

STEELE, J. M. (2005) Darrell Huff and Fifty Years of How to Lie with Statistics. *Statistical Science*, 20(3), pp. 205–209.

ȚARCĂ, M. (1998) *Treaty on Applied Statistics* [in Romanian]. Bucureşti: Didactic and Pedagogical Publishing House.

ȚIȚAN, E., GHIȚĂ, S. & TRANDAȘ, C. (2001) *The basics of statistics* [in Romanian]. Bucureşti: Meteora Press Publishing House.

WAINER, H. (1984) How to Display Data Badly. *The American Statistician*, 38(2), pp. 137-147.

WAYNE, WD & CHAD, C.L. (2013) *Biostatistics: A Foundation for Analysis in the Health Sciences (Tenth Edition)*. Hoboken, NJ: John Wiley & Sons, Inc.