ANOVA = bad science

Can bad science practices be fixed?

Ulisse Di Corpo[1]

Abstract

The symposium on reproducibility and reliability of medical research, held at the Wellcome Trust in London on 1-2 April 2015, touched one of the most sensitive issues in science today: the fact that something has gone fundamentally wrong with one of the greatest endeavors of Western civilization. Participants expressed the concern that more than 50% of what is published on the main scientific journals is simply untrue. Scientific journals are afflicted by conflicts of interest and by a scientific liturgy which has turned dogmatic. They publish only results that have been obtained using ANOVA (Analysis of Variance). ANOVA requires that effects can be added, that data is quantitative and normally distributed, and groups are initially similar and are from the same population. Requirements which cannot be met in life sciences and lead to results that are inconsistent, unstable and often incorrect. Nevertheless ANOVA has become a requirement of all the scientific journals and only results obtained using ANOVA are published. Consequently a lot of what is published in the main scientific journals is simply incorrect. The endemicity of bad science has become alarming and using ANOVA scientists too often sculpt data and results to fit their expectations.

Can bad science practices be fixed?

[1] ulisse.dicorpo@syntropy.org

Introduction

A study published on JAMA (Journal of the American Medical Association), which revisited the results produced using ANOVA and published in the period from 1990 to 2003 in 3 major scientific journals and cited at least 1,000 times, found that a study out of three was refuted by other experimental works. This finding raises serious doubts about ANOVA, when used in life sciences.[2] In May 2011 Arrosmith published in the Journal Nature[3] a study which shows that the ability to reproduce the results from phase 1 to phase 2 decreased in the period 2008-2010 from 28% to 18%, despite results were statistically robust in phase 1 (using ANOVA). Gautam Naik in the article "Scientists' Elusive Goal: Reproducing Study Results" published on the Wall Street Journal on December 2, 2011 points out that one of the secrets of medical research is that the majority of results, including those published in major scientific journals, cannot be reproduced. Reproducibility is at the foundations of making science and when results are not reproduced the consequences can be devastating for the biomedical industry, which only in the U.S. invests each year more than 100 billion dollars in research. Naik suggests the hypothesis that researches, particularly those carried out in universities, are often biased by the need to find positive results, in order to publish and receive funding and because of increased competition.

In the December 23, 2010 Jonah Lehrer[4] writes of a meeting of neuroscientists, held in Brussels on September 18, 2007, and in which the reducing effect of the second-generation antipsychotic drugs was discussed. During this conference it was suggested that the decline of the effect of today's best sellers drugs, such as Abilify, Zyprexa and Serequel, is due to the fact that the environment becomes accustomed to their effects, similarly to what happens with antibiotics. The use of antibiotics leads to select and enhance microorganisms which become in this way immune and "get used" to the antibiotic. However, the attempt to extend this explanation to psychiatric drugs is inconsistent as it is known that there are no microorganisms which cause schizophrenia. In the January 3, 2011 article entitled "More Thoughts on the Decline Effect," Jonah Lehrer answers readers' letters and notes that the reduction effect occurs in biology, medicine and psychology (i.e. in life sciences).

[2] Ioannidis J.P.A. (2005), *Contradicted and Initially Stronger Effects in Highly Cited Clinical Research*, JAMA 2005; 294: 218-228.

[3] Arrosmith J. (2011), *Trial watch: Phase II failures: 2008-2010*, Nature, May 2011, 328-329.

[4] "The Truth Wears Off," he New Yorker.

Lehrer quotes a passage of a letter from a university professor, now an employee of a biotechnology industry:

> "*When I worked in a university lab, we'd find all sorts of ways to get a significant result. We'd adjust the sample size after the fact, perhaps because some of the mice were outliers*[5] *or maybe they were handled incorrectly, etc. This wasn't considered misconduct. It was just the way things were done. Of course, once these animals were thrown out [of the data] the effect of the intervention was publishable.*"

Leher continues:

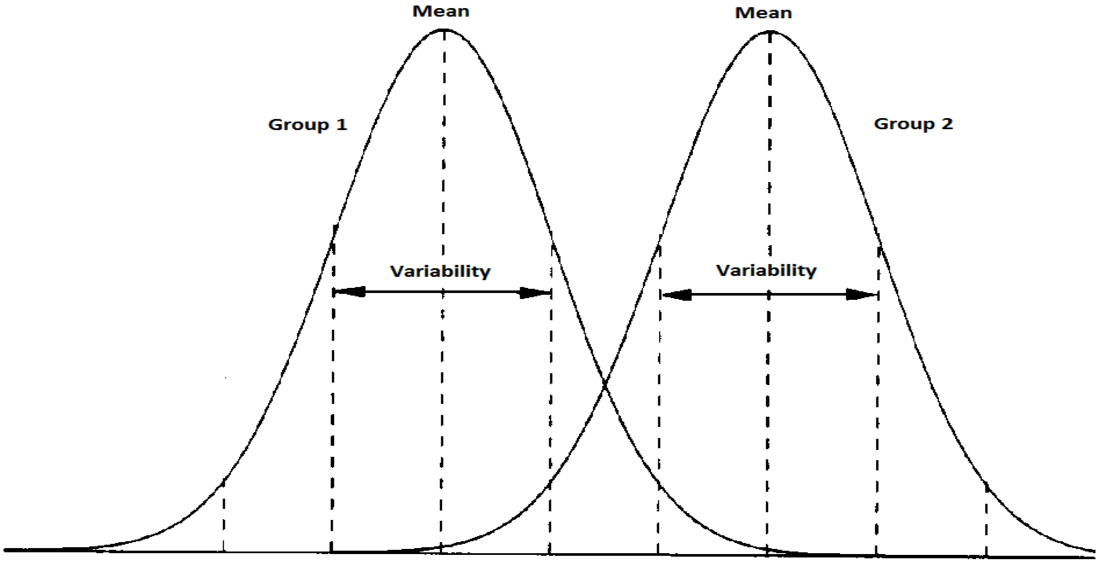
> "*Of course, once that basic research enters clinical trials, there's plenty of evidence that the massive financial incentives often start warping the data, leading to the suppression of negative results and the misinterpretation of positive ones. This helps explain, at least in part, why such a large percentage of randomized clinical trials cannot be replicated.*"

ANOVA and bad science

Analysis of variance (ANOVA) is a collection of statistical models in which the observed variance is partitioned into components: *treatment variability* (between groups) and *error variability* (within groups). The ratio of the treatment variability and the error variability produces a value, F, of which the statistical distribution is known and from which the statistical significance of the effect is obtained.

ANOVA assesses statistical significance by comparing the variance between groups with the variance within groups.

[5] In statistics, an outlier is an observation that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error. Consequently, it is commonly accepted that researchers can freely include or exclude outliers from the data set, changing in this way the outcome of the results.



Comparison of variability of two groups

Initial similarity between groups is a fundamental requirement, without which it is impossible to state that the difference observed between the experimental and the control group is a consequence of the treatment.

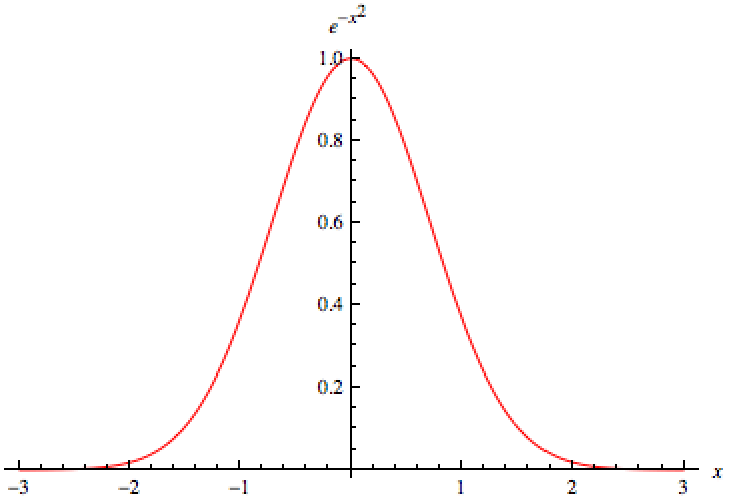
In order to satisfy this requirement, randomization is generally used. Randomization tends to distribute all the intervening variables in a similar way, thereby making groups similar. For example, randomization will distribute subjects of different ages in a similar proportions between the experimental and the control group. The same will happen to all the other variables. But, generally speaking, no controls are performed in order to verify if the condition of similarity is satisfied and often the experimental and control groups are different ever since the beginning of the experiment. A single person with extreme values can produce differences which are not due to the cause (i.e. treatment), but are due to the initial dissimilarity of the control and experimental groups.

Increasing the samples size is also used, since it allows small differences to reach statistical significance. But, in clinical trials the variability of subjects can be so great that even increasing the sample size does not lead to statistical significant results. When this is the case laboratory animals are used. Laboratory animals are all very similar and are used in order to decrease the variability of the sample, allowing in this way small differences to become statistically significant. But, there is now mounting evidence that animal experimentation constitutes an artifact.[6] The reason is very

[6] In experimental science, the expression ‘artifact’ is used to refer to experimental results which are not manifestations of the natural phenomena under investigation, but are due to the particular experimental arrangement, and hence indirectly to human agency.

simple. Statistical significance is stronger when the variability is smaller. Consequently, when the effect size is small, the way to obtain results is to reduce the variability of the sample. When using animals, which are all very similar, the variability of the sample tends to be null, and consequently also insignificant differences become statistically significant. In other words, animals are too similar and differences that have no actual value become significant. Furthermore, one of the fundamental rules in science is to use samples that are representative of the population to which results will be generalized. It is obvious that laboratory animals are not representative of humans and that the effects observed using laboratory animals are difficult to generalize to humans.

Finally, the methodology of differences uses parametric statistical techniques, which require data distributed according to the Gaussian curve. This condition is usually not met.



In the 1960s Simon Shnoll and coworkers were probably the first scientists to show that the assumption of the Gaussian distribution is only mathematical, and that in life sciences and also in physics it is false. In a review of studies performed over more than forty years, Shnoll[7] shows the non-randomness of the fine structure of the distributions, starting from biological objects and moving into the purely physical domain. The implication is huge: tests based on the assumption of Gaussian random distributions (i.e. ANOVA) are fundamentally biased and produce results which are often incorrect.

[7] Shnoll SE, Kolombet VA, Pozharskii EV, Zenchenko TA, Zvereva IM and AA Konradov, Realization of discrete states during fluctuations in macroscopic processes, Physics – Uspekhi 162(10), 1998, pp.1129–1140. http://ufn.ioc.ac.ru/abstracts/abst98/abst9810.html#d

Can bad science practices be fixed?

Science (from Latin *scientia*, meaning knowledge) is a systematic enterprise that builds and organizes knowledge in the form of testable explanations and predictions. An explanation is a set of statements which clarify the relations among causes, context, and consequences of facts. Explanations may establish rules or laws which allow to formulate predictions. Relations among causes, context and consequences are at the basis of explanations and predictions and, when relations are studied in a replicable way, it is possible to talk about science.

In 1843, John Stuart Mill stated that causal relations can be studied using:[8]

1. The *methodology of differences*: "*If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.*"
2. The *methodology of concomitant variations*: "*Whatever phenomenon varies in any manner whenever another phenomenon varies in some particular manner, is either a cause or an effect of that phenomenon, or is connected with it through some fact of causation.*"

From a statistical point of view the methodology of differences is embodied in parametric statistical techniques which compare mean and variance values, such as Student's t and the analysis of variance (ANOVA). The following conditions are needed:

1. In order to study differences between groups it is necessary that the effects can be added among the experimental subjects. For example, if a drug increases in some subjects the reaction times, whereas in others subjects it reduces the reaction times, when adding these opposite effects a null effect is obtained. The effect exists, but it is invisible to ANOVA.
2. Differences can be calculated only when using quantitative data, i.e. data which can be added together. For this reason, experiments are conducted using laboratory measurements. On the contrary, qualitative data cannot be added and it is unsuitable when using ANOVA.
3. All possible sources of variability must be controlled. It is important that nothing, besides the treatment, the cause that we administer, can influence the variability of groups. For this reason a controlled environment, which allows to keep similar all the possible sources of variability and in which each subject is treated exactly in

[8] Mill J.S. (1843), *A System of Logic*, University of Toronto Press, 1843.

the same way, is needed. Controlled environments require laboratory settings, which are very different from the natural context. The need for controlled settings excludes the big picture and limits ANOVA to analytical knowledge, detached from the context and from complexity.

Common mistakes:

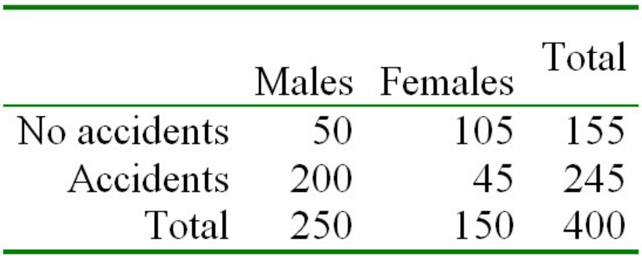
1. Differences can be caused by single extreme values. Just one single outlier can cause statistical significant results and lead to assert effects that do not exist. Outliers are often kept or removed in order to manipulate results.
2. Data transformation refers to the application of a deterministic mathematical function to each point in a data set which is replaced with the transformed value. A common example are logarithmic transformations. In theory, any mathematical function can be used to transform the data set. Operating in this way, it is often possible to obtain differences between the two data sets, when there are no effects.

Let us see now if good science is possible.

In 1992 physicists at LEP (Large Electron-Positron Collider in operation at CERN in Geneva) could not explain some annoying fluctuations in the beams of electrons and positrons. Although very small, these fluctuations created serious problems when the energy of the rays must be measured with great precision. ANOVA did not provide any clue and in order to solve the dilemma the methodology of concomitant variations was used in order to test different hypotheses. Results showed the concomitant fluctuation in the energy of the particle beams of LEP and the tidal force exerted by the Moon. A more detailed analysis showed that the gravitational attraction of the Moon distorts very slightly the vast stretch of land where the circular tunnel of LEP is recessed. This tiny change in the size of the accelerator caused fluctuations of about 10 million electron volts in the energy rays.

The methodology of concomitant variations uses double entry tables of dichotomous variables.

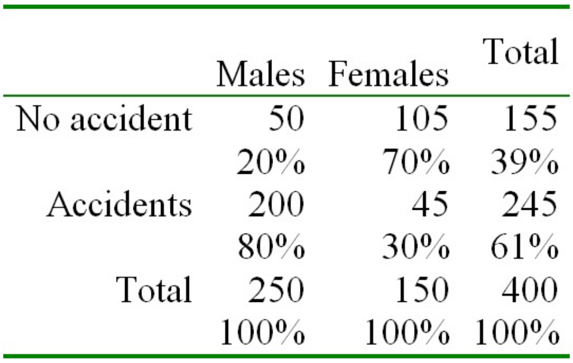
For example:



	Males	Females	Total
No accidents	50	105	155
Accidents	200	45	245
Total	250	150	400

Concomitances between sex and car accidents (data invented for this example)

This table shows the concomitance of the variable sex and car accidents. But, it is difficult to assess the existence of relations, since the total values of each column differ. When the absolute frequency values are converted into column percentage values, as shown in the next table, it becomes possible to compare the dichotomous variables "*Males*" and "*Females.*"



	Males	Females	Total
No accident	50	105	155
	20%	70%	39%
Accidents	200	45	245
	80%	30%	61%
Total	250	150	400
	100%	100%	100%

Concomitances between sex and car accidents (columns percentages)

We now see a strong relation (concomitance) between "*Males*" and "*Accidents*" (80%) and between "*Females*" and "*No accidents*" (70%). Concomitances are assessed according to the differences between observed frequencies (column percentages of the dichotomous variable) and expected frequencies (column percentages of the total column). For example, the expected percentage for "*no accidents*" (total column) is 39%, whereas in the cell "*females*" we have 70%.

Since being male is determined before car accidents take place, we easily fall in the trap of stating that being male is the cause of car accidents. However, the methodology of concomitant variations allows to check for intervening variables by splitting in two the previous table.

In the next table, the previous table is split between those who drive little and for those who drive a lot:

	Drive little		*Drive a lot*	
	Males	Females	Males	Females
No accidents	70%	70%	20%	20%
Accidents	30%	30%	80%	80%
Total	100%	100%	100%	100%

Concomitances between sex, km driven and car accidents

It turns out that the concomitance between sex and accidents vanishes, since there is no difference between males and females in the group of those who drive little and in the group of those who drive a lot. The relation "*males are involved in more accidents*" is therefore mediated by the variable number of kilometers driven, which is therefore an intervening variable. Consequently the relation becomes "*males drive a lot and consequently are involved in more accidents.*"

Crossing three variable at a time allows to identify intervening variables and to study the context within which relations are valid. For example, when a concomitance is found between drug and healing it is possible to study if it is valid always or only at certain conditions, such as specific age groups, sex, habits and other conditions.

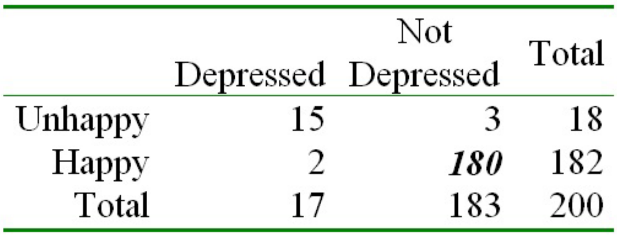
The advantages of the methodology of concomitant variations are:

1. It allows the study of many variables at the same time, thereby it can take into account the complexity of the phenomena, whereas ANOVA can study only two or a very limited number of variables at a time, producing knowledge which is detached from the context and complexity of natural phenomena.
2. It transforms quantitative, qualitative, objective and subjective information into one or more dichotomous variables. In this way quantitative and qualitative, objective and subjective can be studied together.
3. It performs controls for intervening and spurious variables, and this is done afterward and not before. Therefore it does not require controlled environments such as laboratory conditions and it is possible to perform studies in natural contexts.
4. Contrary to the ANOVA, results are impossible to manipulate.
5. When using subjective variables people often respond using masks. For example, even when we feel unhappy, lonely, depressed, usually we try to give an image of ourselves (a mask) which is positive. With ANOVA masks constitute a problem which is insurmountable and which is solved removing qualitative and subjective

variables from the analyses. On the contrary, the methodology of concomitant variations can handle correctly responses which are masked.

Let us discuss this last point. A property of masks is that they are used not only on one variable, but on all those that express the trait that we are trying to mask. For example, if a person responds by saying no to "*I feel depressed,*" when he is depressed, he will also say no to "*I feel unhappy,*" when he is unhappy. The relation between depression and unhappiness remains unchanged, because both responses have moved in the same direction and continue to remain concomitant. For this reason, the methodology of concomitant variations allows for direct questions such as: "*do you feel depressed*?"

An example:

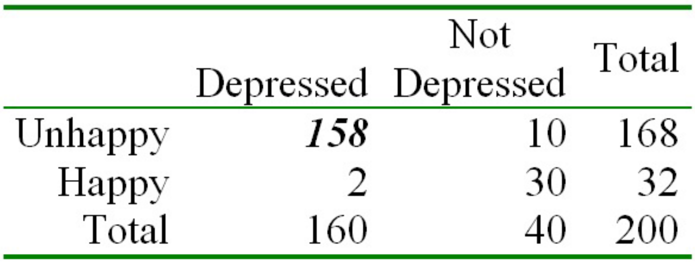


	Depressed	Not Depressed	Total
Unhappy	15	3	18
Happy	2	***180***	182
Total	17	183	200

Concomitances between masked answers

The two variables, "*I feel happy*" and "*I do not feel depressed*", although masked, turn to be related.

When using psychological tests, which produce "objective" measurements of depression and happiness which are not distorted by the effect of masks, answers shift from the positive to the negative side. But the relation remains practically the same.



	Depressed	Not Depressed	Total
Unhappy	***158***	10	168
Happy	2	30	32
Total	160	40	200

Concomitances obtained when using "objective" information

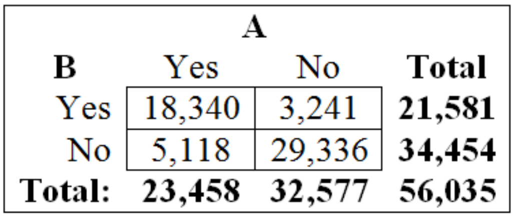
Since relations are studied as concomitances, results continue to show the relation between the variables depression and unhappiness. *Relation* and *correlation* are similar concepts. The term *correlation* is usually used when handling quantitative data, whereas when dealing with dichotomous variables the term *relation* is more appropriate.

The previous example shows that if a relation exists it will emerge also when responses are masked, since masks are applied in a coherent way to all those variables which are related. This is a fundamental issue, as the problem of the mask is a ubiquitous problem in psychological, social and economic sciences. The methodology of concomitant variations solves this problem and allows in this way to widen science to subjective and qualitative data.

Statistics

In the context of the methodology of concomitant variations studies are carried out using nonparametric statistics, among which the Chi Square test (χ^2) is today one of the most widely used statistical indexes. The χ^2 test calculates the differences between observed frequencies and expected frequencies. In the absence of correlation χ^2 is equal to 0, whereas in the case of maximum correlation it is equal to the size of the sample. The comparison with the χ^2 probability distributions allows to know the statistical significance of the correlation. Statistical significance indicates the risk which is accepted when we state the existence of the correlation. Conventionally correlations are considered as valid when the risk is below 5% or 1%. With dichotomous variables χ^2 values have a risk lower than 1% with values greater or equal to 6.635 and with values greater than 3.841 the risk is lower than 5%.

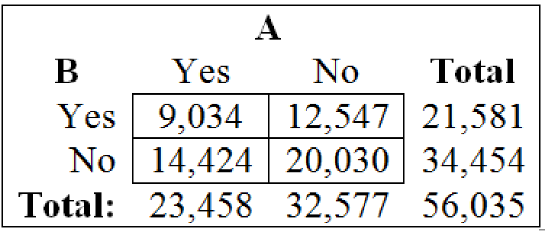
As already mentioned when using the methodology of concomitant variations all variables are translated into the dichotomous form, Yes/No. Crossing two dichotomous variables produces a 2x2 table.



	A		
B	Yes	No	**Total**
Yes	18,340	3,241	**21,581**
No	5,118	29,336	**34,454**
Total:	**23,458**	**32,577**	**56,035**

In this example the total number of cases is 56,035. The Chi Square value is obtained by comparing observed and expected frequencies. Expected frequencies are calculated by dividing the product of the total values of row and column by the general total. For the first cell (Yes / Yes) we have: 21,581 x 23,458/56,035 = 9,034.

Following this procedure we get the table of expected frequencies:



	A		
B	Yes	No	**Total**
Yes	9,034	12,547	21,581
No	14,424	20,030	34,454
Total:	23,458	32,577	56,035

The Chi Square formula is the following:

$$Chi\ Square = \sum \frac{(f_o - f_e)^2}{f_e}$$

where f_o indicates observed frequencies and f_e expected frequencies

In other words, for each cell we calculate the square of the differences between observed frequencies and expected frequencies divided by expected frequencies and we sum the results together. In this example the Chi Square value is 26,813, well above the value 6.635 from which the statistical significance of 1% starts.

Since the maximum value of χ^2 varies depending on the number of cases, it is useful to standardize it, making it vary between 0 and 1. This transformation is known as the rPhi test and is obtained as the square root of the value of χ^2 divided by the sample size. RPhi values obtained from quantitative variables behave similarly to the classical correlation index (Pearson's r). Correlations can be of two types: direct or inverse. If the correlation is directed the two dichotomous variables are concomitantly true or false, whereas if the correlation is inverse one variable is true when the other is false. Inverse correlations have negative sign (-) while direct correlations are shown without sign. High correlation values, that is from 0.35 onward, typically identify trivial correlations that are known without resorting to statistical analyses. Lower values, below 0.35 and in particular around 0.200, identify correlations which are not trivial. In order to study non-trivial correlations it is necessary that the sample exceeds 100 subjects.

Software

Statistical software was developed by the author in order to use the methodology of concomitant variations. A complete description is available in the book "The Methodology of Concomitant Variations"[9]. The first version of this software dates back to 1982, it was distributed with the name *DataStat*, and extensively used in the Department of Statistics of the University of Rome. It is now named Sintropia-DS and merges database and statistical analyses (this is the reason of the extension DS: database and statistics).

Some characteristics of Sintropia-DS are:

1. *Online coding of data*. Statistical analyses require data which has been translated in a numeric form. Online coding makes data-entry easy, more efficient, and allows to check constantly the quality of data, reducing in this way errors.
2. *Unity of structures*. Commercial data-bases are organized in sub-archives which are related together. This architecture conflicts with the statistical unit requirement. Sintropia-DS records are united in one archive, one structure, which allows to perform easily the analysis of concomitant variations.
3. *Easy editing of forms*. It is possible to use forms of any level of complexity. Editing a Sintropia-DS form is easy. The same file used to print the form with a word processor can be used (with minor changes) to edit the structure of the Sintropia-DS database and data entry form. Extensive diagnostics guarantees that the final form is suitable for statistical analyses.

Other characteristics:

1. Integration of database and statistical analyses optimizes data-entry for statistical analyses. The grid which translates data into the dichotomous form is produce automatically, reducing in this way errors and fatigue. Automatic checks during data-entry drastically increase the quality of data, and reduce data-entry time.
2. Only few statistical techniques, coherent with the methodology of concomitant variations, are provided. Users with no background in statistics, can produce robust and correct statistical analyses.
3. The integration of qualitative and quantitative data allows for the complexity of natural phenomena.
4. Instantaneous analyses, independent from the dimension of the archive, allow immediate visualization of the most complex results.

[9] http://www.amazon.com/dp/B00MOBIGWC

Conclusion

The methodology of concomitant variations and non-parametric statistics provide a path that shows that it is possible to go back to good-science.

In 1989, the American National Academies of Science (NAS) published a booklet entitled *On Being a Scientist,* in 1995 it added the sub-title *A Guide to Responsible Conduct in Research*. In the same period, the National Institutes of Health (NIH) established an Office of Research Integrity[10], which all too often reports penalties enacted on researchers who have been found dishonest. On the first of October 2012, The Guardian published the article "*Tenfold increase in scientific research papers retracted for fraud. Study of 2,047 papers on PubMed finds that two-thirds of retracted papers were down to scientific misconduct, not error.*"[11] A study, published on the Proceedings of the National Academy of Sciences (PNAS)[12], found that papers are retracted mainly because of fraud. In the 5 October 2012 editorial of the New York Times "*Fraud in the scientific literature*"[13] it is suggested that researchers are competing for inadequate available resources[14] and have become grant-seekers, who continuously need to publish. This situation is leading researchers towards deliberate fraud and dishonesty, which is now considered to be endemic within science.[15,16]

Publish or perish is a phrase coined to describe the pressure to rapidly and continuously publish scientific works. Frequent publication is one of few methods at disposal to demonstrate scientific talent. Successful publications bring attention and sponsoring institutions, and facilitate funding. Scientists who publish infrequently, or who focus on activities that do not result in publications, find themselves out of the funding tracks. It is now widely recognized that the pressure to publish is one of the main causes of poor research and fraud in science.

Scientific fraud is usually perpetrated using ANOVA, at the moment of data analysis, which allows for easy manipulation. ANOVA provides a path that, by assessing differences can be manipulated by keeping or removing outliers.

[10] http://ori.hhs.gov/

[11] www.theguardian.com/science/2012/oct/01/tenfold-increase-science-paper-retracted-fraud

[12] www.pnas.org/content/109/42/17028

[13] www.nytimes.com/2012/10/06/opinion/fraud-in-the-scientific-literature.html?_r=0

[14] Freeland Judson H. (2004), *The Great Betrayal: Fraud In Science*; Etchells P. and Gage S. (2012), *Scientific fraud is rife: it's time to stand up for good science. The way we fund and publish science encourages fraud*, The Guardian, 2 November 2012.

[15] Broad W. and Wade N. (1982), *Betrayers of the Truth: Fraud and Deceit in the Halls of Science*, Simon & Schuster, 1982.

[16] Bauer H. (2014), *The Science Bubble*, EdgeScience #17, February 2014, http://www.scientificexploration.org/edgescience/

A widespread chorus of scientists is calling for a change towards a new way of doing science, which will comprise qualitative and quantitative information, objective and subjective, and take into account the context and complexity.

We here suggest that this change coincides with the transition from the methodology of differences to the methodology of concomitant variations.

A more detailed analysis is available in the book "*The methodology of concomitant variations*": http://www.amazon.com/dp/B00MOBIGWC