J Stat Phys
DOI 10.1007/s10955-011-0252-5

Fundamental Limitation on Applicability of Statistical Methods to Study of Living Organisms and Other Complex Systems

Yitzhak Rabin

Received: 27 March 2011 / Accepted: 12 June 2011

Abstract A living organism is a complex system whose state is characterized by extremely large number of variables that far exceeds the number of individual organisms that can be experimentally studied. Since the relations between these variables and even their identities are largely unknown, the applicability of statistical methods of inference to the outcome of experiments in biomedical sciences is severely limited. Far from being a purely theoretical issue, this explains the recently proposed "Truth Wears Off" effect and sets a fundamental limitation on the applicability of machine-like approaches to the study of living organisms.

Keywords Living organism · Complex system · Hidden variable · Gaussian distribution · Statistical independence

In a recent article titled *The Truth Wears Off* published in the annals of science section of the New Yorker [1], the author asks whether something is wrong with the scientific method and discusses many instances in which experiments on living organisms in general and humans in particular, produced significant experimental findings that did not survive the test of time. We are all familiar both with the rather rare cases of scientific fraud and the much more prevalent cases of unintentional experimental biases or of rushing to publish insufficiently tested results (for a recent discussion of such biases, see e.g., [2]) but, as will be argued below, the issues raised in this article go beyond "poor science" and involve fundamental problems that face any experiments on complex systems, and in particular, on living organisms.

Y. Rabin (✉)
Department of Physics and Institute for Nanotechnology, Bar-Ilan University, Ramat Gan 52900, Israel
e-mail: rabinserious@gmail.com

Y. Rabin
Department of Biomedical Engineering, Northwestern University, Evanston, IL 60208, USA

Y. Rabin
James Franck Institute, University of Chicago, Chicago, IL 60637, USA

Published online: 21 June 2011

Lets assume that we have $N \gg 1$ "identical" systems (a group of mice, people, etc.), each of which possesses a measurable attribute (variable) X. This variable represents a complex property of the system such as the ability to recognize a face, the beneficial effect of a drug, etc. and therefore a measurement of X does not, in general, yield a yes or no (binary) result. In order to describe this property we will assume that X can take a large number ($M_x \gg 1$) of values in the range from $x_{\min}$ to $x_{\max}$. Notice that the assignment of a particular numerical value x to the result of such measurement is problematic by itself; however, we will ignore this difficulty in what follows.

We perform an experiment in which we measure the property X for each of the systems and obtain a set of N results $\{x_i\}_{i=1\ldots N}$. In order to generate a good statistical sample, the number of systems has to be much larger than the number of possible x values, i.e., $N \gg M_x$. We will assume that this indeed is the case. We can now construct a histogram of these results and obtain the normalized distribution $P(x_i)$.

Now, lets consider the two simplest limiting cases:

(a) The flat distribution $P(x_i) = const$.
(b) The peaked distribution such that $P(x_i = a) \simeq 1$ and $P(x_i \neq a) \simeq 0$.

If the measured distribution is flat, we will conclude that X is not an interesting/relevant property of these systems and forget about it. However, if the distribution is peaked, we will conclude that we discovered an important (i.e., worth of publication) property of our systems [3].

Since we are only interested in generic behavior, in the following we assume that X is a continuous variable that takes values in the interval $[x_{\min}, x_{\max}]$ and notice that both limits can be described by a Gaussian distribution of width σ, centered around a

$$P(x) = A \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right] \tag{1}$$

Here A is a constant that ensures that the probability distribution is properly normalized. This distribution is peaked (flat) when the width σ is much smaller (larger) than the range of values taken by the variable X. We would like to stress that the use of a Gaussian distribution is not essential for our argument and is given only as an illustrative example; any peaked distribution will be adequate for our discussion.

Now, lets assume that each of the systems is complex in the sense that it is not completely defined by the property X (actually, if it was not the case, observation of different values of x would imply that the systems are not even approximately identical) and that there is at least one additional "hidden" variable Y (in fact, one expects that there is a huge number of such hidden variables Z, W, etc.). In general, unless we are either lucky or have deep knowledge/intuition about the complex system, the variables X and Y will not be statistically independent. The crucial point is that the unknown variable Y can take M_y values where M_y can be an arbitrarily large number and, in particular, much larger than N. Each time we perform an experiment that measures a property X, we sample a particular subset $\{y_i\}_{i=1\ldots N}$ of all M_y possible values of Y. Some of these subsets will yield very broad distributions of X but it is only when a particular subset yields a peaked distribution,

$$\begin{aligned} P_{apparent}(X) &= P(X|\{y\}) \\ &= A \exp\left\{-\frac{[x - a(\{y\})]^2}{2\sigma^2(\{y\})}\right\} \end{aligned} \tag{2}$$

where $\sigma \ll x_{\max} - x_{\min}$ that we conclude that the experiment yielded a significant and therefore a publishable result (for convenience of notation, we have replaced the discrete set y_i by continuous y, in the above equation). The next time we perform an experiment on another group of "identical" individuals, we will sample another subset $\{y'\}$ of the set of possible values of the unknown variable Y and will obtain different values a' and σ' which, in general, will no longer correspond to a narrow distribution.[1] Notice that since we already decided that X is a "good" variable, we are no longer free to dismiss the new results and we will conclude that "the truth wears off" with time. The extent to which this happens will depend on the correlation between the measured X and the hidden Y variables: weak correlations will yield relatively robust results which will change only weakly in repeated studies. This would happen, for example, for logarithmic dependence of the width on Y. Strong correlations will lead to irreproducible results and loss of significance once the experiment is repeated (e.g., when the dependence of the width on Y is of power law or exponential type). The use of control groups will not help to overcome this problem since such groups consist of similarly complex individuals and therefore introduce a baseline that may shift with time. Although the time scale for the truth to wear off may be affected by the intrinsic dynamics of the system (e.g., due to mutations or environmental changes that affect the population), in many cases such changes are too slow to be observed and the relevant time scale is the response time of the scientific community in question (the time it takes new results to be disseminated through publication and seminars, and the time it takes to design and carry out a new study)—usually, several years.

How are such difficulties handled in the oldest natural science, physics? In equilibrium statistical physics we normally consider relatively simple fundamental elements for which the number of internal states is rather small (just 2 in the case of electron spin) and one can always generate a large enough ensemble of identical elements in order to obtain reliable statistics. It is only when one considers a large number of *interacting* elements that one observes complex phenomena such as phase transitions. The behavior of macroscopic systems with practically infinite number of elements (atoms and molecules) is the subject of thermodynamics. A student of thermodynamics is immediately faced with the question: how is it that the thermodynamic state of such systems can be completely characterized by specifying a small number of thermodynamic variables such as number of particles, volume, pressure, temperature, etc.? The answer is that in physics we have symmetries and conservation laws (conservation of mass, energy, momentum, electric charge, etc.) that allow us to identify the thermodynamic variables and deduce the relationships between them, without getting into the largely unknown microscopic details of the system under consideration (e.g., a steam engine). However, no such general prescription exists for identifying the "good" variables ("order parameters", in the language of statistical physics) that capture the behavior of complex physical systems far from equilibrium and therefore non-equilibrium thermodynamics remains an open problem. In some cases (hydrodynamics is perhaps the best example), we are sufficiently familiar with the system studied, to be able to identify the slowly changing variables and represent the cumulative effect of all the other (fast) degrees of freedom as dissipation and random noise. This is not possible in general, either because no separation of time scales exists for the systems under study or because we do not understand the laws that govern the relations between the different variables. Our very incomplete understanding of the physics of glasses is a case in point.

[1] While there is no fundamental law of nature that ensures that the narrow distribution will broaden upon further experimentation, we expect this to happen for the same reasons that lead to the 2nd law of thermodynamics (with time the system will go through its available states).

Clearly, the situation is much more complicated in studies of living systems, the subject of biology and medicine. Even on a single cell level, the identity of genomes in a clonal cell population does not ensure identical response to external stimuli, because of random variations in gene expression, different initial conditions, etc. These issues become even more problematic for multicellular organisms where time-dependent phenomena such as aging may be important. Our ability to isolate variables and to understand the inter-dependence of different variables is fundamentally limited by the fact that the organisms are viable only within a very limited range of external conditions. Biomedical research faces a formidable problem: the a priori probability that a statistically significant result obtained in response to a random question about the organism will survive the test of time, is vanishingly small. Of course, even though unlike in thermodynamics, there is no systematic way of identifying the relevant variables and questions, the odds are greatly improved by experience and familiarity with the organism studied, as one develops intuition about the system and begins to feel the structure of the underlying relationships that control its behavior. From the fact that a non-negligible fraction of published results appears not to "wear out" with time, one concludes that such intuitive understanding is not extremely rare. On occasion, some of us may have a glance at the book Einstein referred to, when he wrote about Newton: "Nature to him was an open book". Since "intuition" and "feeling" are essentially human traits which are usually associated with art rather than with computers, one expects major advances in our understanding of complex systems in the foreseeable future to be the result of human rather than machine efforts.

The above discussion has important implications for the organization and funding of research in life sciences and in other disciplines dealing with systems of comparable complexity. If, as is claimed here, even the best statistical analysis of experimental data on such systems can not fully ensure the scientific validity of the results, "truth" is more elusive than we normally tend to believe and, consequently, critical studies that attempt to test the validity of known results should be encouraged by the scientific community.

Acknowledgements Helpful discussions and correspondence with Jonathan and Ilana Rabin, Jonathan Widom, Mitchell Feigenbaum, David Biron, Igal Szleifer, Alexander Grosberg and Bruce Spencer are gratefully acknowledged.

References

1. Lehrer, J.: The truth wears off. The New Yorker, Annals of Science, December 13, 52 (2010)
2. Ioannidis, J.P.A.: Why most published research findings are false. PLoS Med. **2**, 696–701 (2005)
3. Schooler, J.: Nature **470**, 437 (2011)