Multidimensional Time Model for Probability Cumulative Function.

Michael Fundator, Sackler Colloquium of National Academy of Sciences.

abstract ID # **302990 Contributed Presentation at the 11th International Conference on Health Policy Statistics taking place October 7-9 in Providence, Rhode Island.**

Back in the Fall/97 I was doing my research in statistics and probability in the Columbia University Mathematical Library.

Abstract.

The question was how the explosion in the availability of health- and disease-related data from biological, biomedical, behavioral, social, environmental, and economical analyses could be addressed in view of the analysis of biomedical big data. And this was the challenge posed by complexity of data structures such as images, networks, and graphs, missing and sparse data, and complex dependence structures and interaction effects. The new method is based on changes of Cumulative Distribution Function in relation to time change in sampling patterns. Multidimensional Time Model for Probability Cumulative Function can be reduced to finite-dimensional time model, which can be characterized by Boolean algebra for operations over events and their probabilities and index set for reduction of infinite dimensional time model to finite number of dimensions of time model considering also the fractal-dimensional time arising from alike supersymmetrical properties of probability. The applicability of results are further extended to be used in innovative methodology for visualization, modeling, and analysis of biomedical big data to address the challenges posed by complex data structures such as images, networks, and graphs, missing and sparse data, and complex dependence structures and interaction effects such as various DNA analysis. The newly developed models, philosophically based on Erdos- Reney Law for the prediction are philosophically intended to reach high level of precision.

Keywords: finite-dimensional time model, Brownian motion, coefficient of concordance.

1. Introduction.

In contrast to the old and very ineffective notions of scan-statistics, Gundy-type inequalities, and selective rerandomization, which is mainly used to eliminate previous structures through business-like approach, mostly resembling radio or radar-like business analytics, which are tending to disperse attention of any researcher, applying any philosophy of any well developed subject orientation from concentrating on precision in analysis like in medicine, or in biomedical, or biological research, or develop a targeted treatment without further interest into unique characteristics of an individual such as genetic makeup, environmental factors, and/or lifestyle.

This new method is intended to address any challenges posed by previous visualization, modeling, and analysis of biomedical big data with previous complex data structures such as images, networks, and graphs, missing and sparse data with combination of complex dependence structures and interaction effects.
This new method is achieving the goal of precision medicine will require combining data across multiple formats that facilitate high-confidence predictions for individuals.

The last is very important in light of the recent applications of Biostatistics to various DNA analyses. It is also very important after the course of actions in the efforts fighting Ebola have developed sound perception that immune system patterns are following resembling manner of HIV.
The application of new method to immune system patterns could be a very good factor in the future cancer research, in contrast to controversial question of hypothesis whether long-term use of the drugs could prevent cancer.

The new method has very important methodological value, the methodology lies at the core of the research approach based on very valuable properties of Brownian Motion. One of the basic principles underlying the logical approach is to give some access to such mathematical and statistical concepts as ranks, ordinal numbers and higher dimensions, which are crucial to scientific mechanism, but otherwise not only inaccessible in the literature, but also ignored as in further quote of
S. Plouffe in the explanation of appearance of π^5 in the approximation for the proton-electron ratio, proposed by Richard P. Feynman in the 60's "and with powers of π if the dimension is higher. I could not find any evidence of such hypothesis."
The new systematic approach would lead to resolve the problems in the development of superstring theory, the first one remains a conjecture, which suggests that the five string theories might be different limits of a single underlying theory, called M-theory. The other is a dilemma of the lack of experimental evidence for superstring theory, which is based on supersymmetry. No supersymmetrical particles have been discovered and recent research at LHC and Tevatron has excluded some of the ranges.
The most obvious explanation lies in the complete ignoring by well developed mathematical theory of Mirror Symmetry and Algebraic Geometry the whole subject of study of ranks, ordinal numbers, and Mathematical Statistics, which is application of common parts of statistical analysis, that is based mostly on pure mathematical approach, including Brownian calculus, or sampling approach of Experimental Mathematics.

2. Reduction of the data

Assume that in related experiments we want to estimate some parameter, say θ. It can be mean or variance of the population.

Procedure decision δ is called dominating procedure decision δ' in the sense that

$R(\theta, \delta) \leq R(\theta, \delta')$, for all θ and (1)

$R(\theta, \delta) < R(\theta, \delta')$, for some θ. (2)

Where $R(\theta, \delta)$ is the risk function.

A class Ω of decision procedures is called complete, if for any δ' not in Ω there exists δ in Ω dominating it. A complete class is minimal, if it does not contain a complete subclass. Therefore, a minimal complete class is considered in such a sense as to provide the maximum possible reduction of a decision problem without loss of information.

The problem of reduction consists essentially in discarding that part of the data, which contains no information related to the decision problem, it can be, let say, unknown distribution P(θ). Therefore, this data has no value for any decision problem, concerning θ.

More general, by restricting attention to sufficient statistics T(x), which is defined as that, on which the conditional distribution Θ as a random variable of θ depends only given particular outcome X=x, one obtains a reduction of the data, and it is then desirable to carry this reduction as far as possible.

M and L-estimators are the examples of the reduction of the data.

Chebuchev's Theorem, which is studied in the very beginning of the introduction to the foundations of Mathematical Statistics, also suggests the reduction of the data. There are a lot of procedures that are examples of the reduction of the data.

However, when the Least Squares Estimators or weighted LSEs, as well as some other procedures are used for the decision problem, all of the observations of the population or available data usually can be used for the decision problem. This is actually one of the first accesses to the seemingly visual contradiction inside Mathematical Statistics and common parts of statistical analysis, which is based mostly on pure mathematical approach.

3. Emphasis on the importance of relation of logic and indexes to multi-dimensional approach.

"The Inverter and methodology" is a quote used by S. Plouffe, while explaining appearance of π^5 in the approximation for the proton-electron ratio, proposed by Richard P. Feynman in the 60's, S. Plouffe, mathematician, who discovered the formula, which permits the computation of the *n*th hexadecimal digit of π.

$$\pi = \sum_{k=0}^{\infty} \left[\frac{1}{16^k} \left(\frac{4}{8k+1} - \frac{2}{8k+4} - \frac{1}{8k+5} - \frac{1}{8k+6} \right) \right] \quad (3)$$

He naturally related it to multi-dimensional approach. But he gives a short account to his efforts: "Model #1, spheres or archimedean solid of n dimensions N - dimensional spheres of uniform matter is the first model considered. The volume of a n - dimensional sphere V(n) is $V_n = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$ (4), consequently the mass ratio should be rational in 3 dimensions and with powers of π if the dimension is higher." He could not find any evidence of such hypothesis, and his next step was" to consider semi - regular polytopes such as the archimedean solids. The volume of such polytopes is expressed in radicals", and he expected the mass ratios to be as well." However, his search using his database and programs with the other ratios failed to produce similar formulas in simple terms of π^k, exp(πk) with k being an integer or even rational. That idea of finding the best mathematical expression for dimensionless numbers of physics goes back to Sir Arthur Eddington in the 30's with Einstein at the beginning, later Feynman, Gellman and Dirac addressed or thought of that problem. There is an extensive literature on the subject.

4. Potential factor when all of the observed data is used and introduction of sampling theory.

But even in procedures, when all of the observed data is used for the decision problem, the amassing or collection of the data and the future division on subclasses for evaluation in the experiment or in statistical research is closely related to some saturating or exhausting of some potential factor, which in its

turn is related to some hedge strategy in the experiment or research process, that every analyst or researcher faces without a priory strategy or strict guidance.This could be seen from the number of throws in the experiments for the problem of Buffon's needle, or as in the nearly 100 years old experiment, in which a coin was thrown 20,000 times to find out, how many outcomes of Heads and Tails would be recorded.

Next analysis of the Theory of Sampling with large sampling leads to the following results:

1. random sequences of the sampling frame;

2. positive autocorrelation;

3. periodic fluctuations.

Systematic sampling would ensure more spread in the sample elements and will therefore be more precise than simple random sampling .

Therefore, though the observations are not equal, there are some resembling patterns in the sampling, that imply possible reduction of the data.

5. Emphasis on natural and social sciences and consecutive transfer to Biological, Biomedical, and Biostatistical sciences.

In science, we often deal with the issue of agreement or concordance in a relationship between phenomena under consideration. In several science disciplines, particularly in natural science, this can usually be measured by very precise and sophisticated methods.

The prerequisite for precise measurement of concordance is of course the availability and quality of data on the phenomena under consideration. However, in economics, and especially in social sciences and humanities, this prerequisite is often not met. Either, one does not have cardinal numerical data at his disposal, or the quality of such data is questionable. This would imply that there is some dispersion of the data in many areas, such as biostatistics or economics.

As to quote Jonathan M. Borwein in Barron's article for April/17 of the last year:"The higher the number of configurations tried, the greater is the probability that the backtest is overfit."
In the area of calculation of π–digits they introduced formulas for the fast convergence of sums, claimed results for hundreds of billions of digits, and discussed others results of up to 40 trillion of digits.

Now, one can ask a quite natural question, how these results could be used in innovative methodology for visualization, modeling, and analysis of biomedical big data to address the challenges posed by complex data structures such as images, networks, and graphs, missing and sparse data, and complex dependence structures and interaction effects such as various DNA analyses.

6. Introduction of a new idea through very often, but not yet resolved contradiction.

Though it looks like some kind of contradiction, some experiments with even not so large data can be considered as those, that have exhausted or saturated some potential factor, such as in Buffon's needle or

in coin tossing, while others as in Biostatistics or Economics, that even with much larger data look like lacking some data or possessing some dispersion factor or law.

However, both phenomena can be explained as related to the potential factor, which is subjected to some outside conditions. In such cases it is often convenient to turn over to the ordinal numerical data, or rank data, where the levels of analysis are indeed limited.

7. Introduction of the coefficient of concordance and consequently logic of Boolean algebra numbers for multi-dimensional approach.

Kendall and Smith (1939) provided a descriptive measure of agreement or concordance, which is called coefficient of concordance and is used for the test of consistency of more than two sets of rankings, such as the merit ordering of competitors by several judges in a sporting contest. If m judges each award ranks 1 to n independently to competitors, the sum s_i of the ranks awarded to competitor i has

mean $\frac{1}{2}(n+1)m$. (5)

The sum of the $S = \sum (s_i - \frac{1}{2}(n+1)m)^2$ (6)

with $W = \frac{12S}{m^2 n(n^2-1)}$ (7)

where $0 \leq W \leq 1$.

If all judges give identical rankings, than W = 1, with the case of W = 0 corresponding to compete disagreement.

In the case, which perhaps could little happen in sport competitions, but would rather resemble four judges rating 3 other, or the interpretation could be as in some talent contest, as in the table below, there is little agreement, it can be seen from the Table 1 below.

Table 1 S = 2 and W = 16.

	JUDGES			
GROUPS (COMPETITORS)	A	B	C	D
I	1	1	2	3
II	3	2	3	1
III	2	3	1	2

The rank is defined as the ordinal associated with an ordered observation.

In many cases it is possible to order observations according to some criterion without assigning exact measurements of objects. For example, it is possible to rank objects by length, without making measurements.

The other reference to a rank in pure and applied mathematics is the rank of a matrix, which is the maximum number of linearly independent columns or rows, or of a linear transformation, which is the dimension of its image.

Returning to the original problem of possible reduction of data with outcomes of the random variable X and some parameter θ, the problem was that of estimating the future value of this parameter θ.

8. Analysis in the system of Axioms of Probability of von Mises, and introduction to the logic of Boolean algebra.

Analysis, of how the integration over the events would work in the system of Axioms of Probability of von Mises, is leading to the decision, that it definitely would be an algebra with operations summation and multiplication.

It was to my great surprise, when after some 15-16 years I found something very resembling in von Mises book, it should not be surprising to the reader, as W.G. Cochran mentions in the Preface to "Sampling techniques": "This kind of assumption is not new-I noticed ***recently*** (the emphasis mine) that Laplace used it around 1800 in a sampling problem-but it clarifies the relation between sample survey theory and standard statistical theory."

Richard von Mises mentions "the four fundamental operations" on pages 38-58 of his book "Probability, Statistics and Truth":

1. Selection;

2. Mixing;

3. Partition; and

4. Combination,

with the last three defined so as to correspond to usual addition, division, and multiplication rules, and the 1st one is defined as the attributes unchanged and the sequence of elements reduced by place selection, which correspond to the unchanging of distribution, and can be viewed as identity element and related to the previous operations.

Proposition1. Analysis, of how the integration over the events would work in the system of axioms of probability of von Mises, is leading to the decision, that it would definitely be an algebra with operations summation and multiplication.

And we can inspect the logic of Boolean algebras related to the general model.

After it was established that the logic corresponding to the concept of algebra, Boolean algebra, algebra of polynomials of martingales, additivity, and additivity of probabilistic measures, the investigation should be made into the concept of product of algebras, indexes, and other logical operations bound by the definition and properties of Boolean algebra.

Consider that in most cases, any random process can be represented as sum of functions of normally distributed random variables, or so called Gaussian random variables (abbr. r.v.), or sum of functions of processes of Brownian motion, using Edgeworth's form, or that equivalent to the Gram-Charlier Type A series that use different types of the representation of the distribution function in terms of Chebyshev-Hermite polynomials. The constant r.v. are considered to be a particular case of $N(m, \sigma^2)$, that with $\sigma=0$.

9. Introduction of Brownian motion.

Brownian motion according to R. von Mises in "Probability, Statistics and Truth" page 186: "About a hundred years ago, the English botanist Brown observed under the microscope that certain organic liquids contain small particles moving to and fro in an incessantly agitated manner. It was discovered later that this so-called 'Brownian motion' is common to all sufficiently small particles suspended in a gas or in a liquid, and that it represents a mass phenomenon following the laws of probability calculus. Since we are only interested in the fundamental logical structure of this problem, we can simplify our conception by considering a two-dimensional scheme. We assume that the particles move in a zig-zag course in the horizontal plane, excluding any up or downward motion, or else, we may say that we consider only the projection of the three-dimensional motion onto a horizontal plane."

Today it is already 188 years since the discovery of the phenomenon of so-called 'Brownian motion', and many authors or almost all of them consider the definition of Brownian motion as starting from one-dimensional model, and further extending it to multy-dimensional case, instead of considering "only the projection of the three-dimensional motion onto a horizontal plane".

The physical theory of this motion was developed by Albert Einstein 110 years ago. It suggests that this motion is random, and has the following properties:

1 it has independent increments;
2 the increments are normally distributed random variables, or so called Gaussian random variables;
3 The motion is continuous.

10. Some of the properties of the Brownian Motion.

In order to make basis for operations with Brownian Calculus it is necessary to mention some of the properties of Brownian Motion and Brownian measure that would be defined later after short introduction of some concepts of Measure Theory, Probability Space, and Stochastic Processes:

(1) Time-homogeneity;

(2) Symmetry;

(3) Scaling;

(4) Time-inversion;

(5) That the probability transition function satisfies the heat equation:

There are some additional properties of Brownian motion, which are worth to mention:

(6) Brownian Bridge is a continuous-time stochastic process with probability distribution equal to conditional probability distribution of Brownian motion. It has very useful time transformation dependence with Brownian motion.

Consider Brownian bridge for $t \in [0,T]$. The expected value of the bridge is zero, with variance $t(T - t)$, implying time recurrence property and that the most uncertainty is in the middle of the bridge.

An attempt to analyze the question for the process of Brownian bridge B(t), which is defined as Brownian motion W(t), given the condition that B(T) = 0,

leads to the definite notion and concept of time recurrence. More precisely:

$B(t)=(W(t)|W(T)=0),\ t\in[0,T],$ (8)

And $B(T) =W(t)-tW(T),\ t\in[0,T].$ (9)

Since the concept of time recurrence is simple and not a complex concept, it can be expected that, recurrence is the axiomatically logical concept for the system of Axioms of Probability of von Mises.

Indeed, I found after some time in von Plato's book, that some few years after the very famous book by A.N. Kolmogorov "Foundations of the Theory of Probability", Alonso de Church introduced the concept of recurrence for the system of Axioms of Probability of von Mises in his paper "On the concept of a random sequence":

"Since the primitive recursive functions are effectively enumerable, sequences satisfying this criterion can be effectively constructed in accordance with Wald's Theorem"

Therefore, it can be expected that the number of dimensions in the general model is not infinite, but is bounded by some index number in the logical consideration of concepts, definitions and properties of Boolean algebra ideals and filters.

Some 15-16 years later I was surprised to find out, that resembling results in search for index numbers were reached in numbers of differentiable structures in many-dimensional manifolds. And though it was proved by J. Milner that for 7- dimensional sphere there are only 28 nonequivalent differentiable structures, 20 years later other mathematicians found that there is infinite number of ways to define velocities and accelerations over 4- dimensional manifolds.

11. Arithmetic and Number theory with methods of Linear Agebra.

After a year or so I found, that very resembling result was already proved in Arithmetic and Number theory simply by application of the methods of linear algebra.

A.Y. Khinchin in the very beginning of "Three Pearls of Number Theory" brings short, but a very interesting history of the Theorem of van der Waerden on Arithmetic progressions with the proof of M.A.

Lukomskaya of Minsk. He tells in his story, that in the Summer/28 upon his arrival to Göttingen for several weeks of the Summer semester, "nearly all mathematicians", whom he met told about the result, which "had just been obtained" in Göttingen "only few days before… arrival" "with enthusiasm".

His account is preceded by "A letter to the front (in lieu of a preface)", which was written in late March/45. As he recounts, the problem had very interesting history. One of the mathematicians there, whose name he forget, "most probably Baudet" "had come upon the following problem in the course of his scientific work: Imagine the set of all natural numbers to be divided in any manner whatsoever into two parts (e. g., into even and odd numbers, or into prime and composite numbers, or in any other way). Can one then assert that arithmetic progressions of arbitrary length can be found in at least one of these parts? (By the length of an arithmetic progression I mean here, and in what follows, simply the number of its terms.)....After several weeks of strenuous exertions, the problem finally yielded to the attack of a young man who had come to Gottingen to study, the Hollander van der Waerden. I made his acquaintance, and learned the solution of the problem from him personally. It was elementary, but not simply by any means. The problem turned out to be deep, the appearance of simplicity was deceptive."

Proposition2. Theorem Let k and l be arbitrary natural numbers. Then there exists a natural number n(k,l) such that, if an arbitrary segment, of length n(k,l) , of the sequence of natural numbers is divided in any manner into k classes (some of which may be empty), then an arithmetic progression of length l appears in at least one of these classes.(Khinchin "Three pearls of Number theory")

Consider Stone representation of Boolean algebra, which is represented by an algebra with known axioms for Boolean algebra and can be characterized by quadruplets **B** = <X, 0, *, ~>, where 0 is an element from a set X, and * is a binary operation and ~ is an unary operation, which would be a Boolean algebra with 1 as a unit on the operations $\wedge$, $\vee$, and ~. Besides that it has four unary operations, two of which are constant operations, another is the identity, and negation and besides the number of n-ary operations, the number of the dimensions that infinite-dimensional model can be reduced to through application of Boolean prime ideal theorem and Stone duality, can be indexed by an index set.

This can be done using the fact that filters in Boolean algebra are naturally in bijection with closed subsets of Stone spaces. Conversely, subalgebras of Boolean algebra correspond to quotients of Stone spaces. Further, any surjective map between compact Hausdorff spaces is a quotient map. Thus, if it is a bijection, it is automatically a homeomorphism.

Proposition3. Multidimensional Time Model for Probability Cumulative Function can be reduced to finite-dimensional time model, which can be characterized by Boolean algebra for operations over events and their probabilities and index set for reduction of infinite dimensional time model to finite number of dimensions of time model considering the fractal-dimensional time that is arising from alike supersymmetrical properties of probability,

One of the simplest examples is convolution, which is similar to cross-correlation in probability, statistics, computer vision, natural language processing, image and signal processing, electrical engineering, and differential equations.

12. Conjecture suggests that the five string theories might be different limits of a single underlying theory.

The development of a quantum field theory of a force invariably results in infinite (and therefore useless) probabilities. Physicists have developed mathematical techniques (renormalization) to eliminate these infinities which work for three of the four fundamental forces—electromagnetic, strong nuclear and weak nuclear forces—but not for gravity. The development of a quantum theory of gravity must therefore come about by different means than those used for the other forces.

Superstring theory is an attempt to explain all of the particles and fundamental forces of nature in one theory by modeling them as vibrations of tiny supersymmetric strings.

The notion of Superstring theory is a substitute for supersymmetric string theory, which is the version of string theory that combines fermions and supersymmetry.

'Superstring theory' is based on well developed mathematical apparatus based on Mirror Symmetry and Algebraic Geometry

However, there has been lack of experimental evidence for superstring theory, which is based on supersymmetry. No supersymmetric particles have been discovered and recent research at LHC and Tevatron has excluded some of the ranges. No report on suggesting large extra dimensions has been delivered from LHC. There have been no principles so far to limit the number of vacua in the concept of a landscape of vacua.

Our physical space is observed to have only three large dimensions and—taken together with duration as the fourth dimension—a physical theory must take this into account. However, nothing prevents a theory from including more than 4 dimensions

Theoretical physicists were troubled by the existence of five separate string theories. A possible solution for this dilemma was suggested at the beginning of what is called the second superstring revolution in the 1990s, which suggests that the five string theories might be different limits of a single underlying theory, called M-theory. This remains a conjecture.

Refe4rences:

1. J. von Plato “Probability Theory”.

2. S. Plouffe “A search for a mathematical expression for mass ratios using a large database”.

3. W. G. Cochran “Sampling techniques”. Wiley&sons

4. D. Nelson “Penguin Dictionary of mathematics”

5. Alonso de Church “On the concept of a random sequence”

6. W. Feller “Introduction to Probability Theory” v..1- 2

7. E. Calabi and H. S. Wilf "Note On the Sequential and Random Selection of Subspaces over a Finite Field" Journal of combinatorial theory (a) 22, 107-109 (1977)
8. J. MacQueen Some methods for classification and analysis of multivariate observations Proc. Fifth Berkeley Symp. on Math. Statist. and Prob., Vol. 1 (Univ. of Calif. Press, 1967)
9. Richard von Mises "Probability, Statistics, and Truth" Dover
10. Albert Einstein, "Investigations on the Theory of the Brownian Movement" Dover
11. D.Revuz , M. Yor "Continuous Martingales and Brownian Motion" Springer
12. Good, I.J. (56) Which comes first, probability or statistics? J.Inst. Actuaries 82 249-255.
13. P. Mörters, Y. Peres "Brownian Motion" CUP
14. A.Y. Khinchin "Three Pearls of Number Theory" Dover
15. D. Cox, S. Katz ,"Mirror Symmetry and Algebraic Geometry", AMS
16. C. Voisin "Mirror Symmetry" AMS
17. S.Blinnikov , R.Moessner Expansions for nearly Gaussian distributions Astrophysics
18. Halmos, Paul Lectures on Boolean Algebras. van Nostrand.