General statistical framework:
draft for the proposition of a possible theory

Sébastien REBECCHI

Member of the LIRIS laboratory, INSA of Lyon, France
Research report for the MIVIA Laboratory, University of Salerno, Italy

November 2007

Abstract

In this report, I propose my vision of what could be a general constraintless statistical framework for the probabilistic characterization of potentially all sets, finite or infinite, discrete or continous, symbolic or numerical, and which can be sumarized as follows : a statistical process is represented by a *theoretical function* $\overline{f}$ that assigns a positive *theoretical absolute weight* for each *theoretical element* of the *theoretical set* $\overline{S}$. First, one has to *realise* the set S, *real* existence of $\overline{S}$, in which one wants to do statistics, thus also realising what are the elements of S. Then, one can go to the realisation, with respect to S, of each possible *theoretical object* asking by $\overline{f}$ in order to obtain the *realised function* f, real existence of $\overline{f}$. Finally, one has made it possible to deal with probabilities in S, producing the *realised relative weight* of each element x of S (also known as *mass* or *density* of probability), thanks to its now *realised absolute weight* $f(x)$, and by the way of an adequate *normalization* process. I detail these steps, provide some examples, discussions, and propositions, more particularly centered on the string case.

Keywords

Function, object, constraint, weigth, absolute, relative, mass, density, probability, element, set, theory, realisation, existence, discrete, continous, numerical, symbolic, finite, infinite, normal, uniform, distance, standard deviation, univariate, multivariate, distribution, mathematics, stochastic, statistics, philosophy, natural, artificial, complex structure, string.

Contents

1 Introduction

Statistics is related to the measurement of the influence, or *weigth*, of each element, or cluster of elements, in the shape and behaviour of a set of such elements. In order to deal with probabilistic issues, the classical framework imposes that the *function* responsible for this measurement has a total weigth of 1 over all the elements of the set. This results in a *different funtion* for each possible set in which one would translate the *same underlying process*, *i.e.* a process that has the same *general* behaviour, in which all elements would have the same amount of *relative* influence in the behaviour and shape of the set.

I think that this constraint makes classical statistics only a special and restrictive case, in the sense that it forbids the most part of positive mathematical functions to express themselves, their underlying shape and behaviour, in a probabilistic framework. I believe in a more general statistical framework in which all of these functions could potentially be considered, by cancelling the stochastic constraint on them, and reporting it in an independant probabilistic part, by the way of an adequate *a posteriori* normalization phase of the absolute contributions of each element of the set, to obtain their relative ones, their amount (mass or density) of probability. I thus believe in the separation of a pure mathematical part, reponsible for the absolute definitions of the weigth of each element in a set, and a probabilistic second part, that permits to take stochastic constraint into account in order to do probabilistic studies of the evolution of a general process, when realised in a particular set within its particular associated objects. Moreover, I think that there is the same undelying funtion for the same general stochastic behaviour (uniform, normal...), regardless to its possible future realisation in a specific set to be considered. To accept this vision, one first have to accept the notion of *theoretical existence* of a mathematical object, like a function, *i.e.* the not mathematically complete expression of it, with a possible theoretical part.

2 The framework

2.1 Theory

The *theoretical function* $\overline{f} : \overline{S} \longrightarrow \mathbb{R}+$ assigns to each *theoretical element* $\overline{x}$ of the *theoretical set* $\overline{S}$, a value $\overline{f}(\overline{x})$ called the *theoretical absolute weight* of $\overline{x}$. $\overline{f}$ is composed of a fixed mathematical part, and a possible realisation-dependant theoretical one, composed of what we call *theoretical objects*.

Example

- $\overline{f}(\overline{x}) = 3$, with no theoretical part, that assigns the same constant theoretical absolute weight to each element of $\overline{S}$
- $\overline{f}(\overline{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\overline{d}(\overline{x},\overline{\mu})^2}{2}\right)$, in which the theoretical part has two theoretical objects, namely the function $\overline{d}$ and the constant $\overline{\mu}$.

There is no constraint on $\overline{f}$. But theoretical objects could be constrained by $\overline{f}$ or other theoretical objects in $\overline{f}$ to be valid. In the last example above, $\overline{d}$ must give a numerical value and $\overline{\mu}$ must take its value in the domain of definition of the second parameter of $\overline{d}$.

Remind not to make this mistake: thoretical objects are not mere *parameters*. In fact, they could be considered as *second-order* parameters, *i.e.* parameters that affect the *nature* of a realisation of the theoretical function, whereas *first-order* parameters are those that affect the *value* of this realised function.

But for the moment, $\overline{S}$ and $\overline{f}$ have only *theoretical* existences, and one have to decide what one wants as their *real* exitences. This is the *realisation* step.

2.2 Realisation

In this phase, one chooses the set in which one wants to make statistics given the theoretical function.

The choice of the set S gives a real existence to $\overline{S}$, and thus to each element of $\overline{S}$. Once the set has been realised, the only step remaining is the realisation of the theoretical objects, with respect to their respective theoretical constraints and the possible new specific ones created by the set realisation, finally obtaining the real existence f of $\overline{f}$.

Example

- $\overline{f}(\overline{x}) = 3$. Let $S = [5;6]$. Then $f(x \in S) = 3$.
- $\overline{f}(\overline{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\overline{d}(\overline{x},\overline{\mu})^2}{2}\right)$. Let $S = \mathbb{N}$, $d(x \in S, y \in S) = \sqrt{(x-y)^2}$, and $\mu = 0$. The constraint on μ is respected ($0 \in S$), and so is that on d (numerical value). Then $f(x \in S) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$.

Now that we don't have any more theoretical existences, we are able to compute any abolute weigth in S to any element of this set. But a more interesting task is that of the probabilistic characterization of S by the distribution inherent to the underlying statistical process driven in S by the realisation of $\overline{f}$ as f. Let us talk about it now.

2.3 Probabilities

A very simple well-known issue to obtain a distribution from a set of alternatives is the *normalization* process. Here the set of alternatives is driven by the absolute weight $f(x)$ from f of each element $x \in S$. Normalizating $f(x)$ by the *total* weigth of S is the only right way to obtain the realisation in S of the distribution that is perfectly faithful to the underlying process represented by $\overline{f}$, and driven in S by f after the realisation of the theoretical objects of $\overline{f}$. What I want to say is that the representation of this underlying statistics is also realisation-dependant in S, because of the choices of the theoretical objects that compose the theoretical part of $\overline{f}$; but once them chosen, this normalization step is the true one to translate, independantly of all possible realisation issue, the stochastic constraint on S.

There is two well-known operators that permit to compute the total value of a function in a set S, depending on the nature of S: if S is discrete, this operator is the sum over S, and controversery if S is continous this is the integral one.

So computing the relative weigth of $x \in S$ is given by the equation:

$$rf(x) = \frac{f(x)}{\sum_{y \in S} f(y)}$$

in the discrete case, and:

$$rf(x) = \frac{f(x)}{\int_{y \in S} f(y)dy}$$

in the continous one.

Passing from $f(x)$ to $rf(x)$ is no more than taking stochastic constraint into account and thus realising a distribution in S. Now we can call $rf(x)$ the *mass* of probability of x in S by f in the discrete case, and its *density* of probability in the continous one.

Example

- $f(x \in [5;6]) = 3$. Then:

$$rf(x) = \frac{3}{\int_{y \in [5;6]} f(y)dy} = \frac{3}{\int_{y \in [5;6]} 3\ dy} = \frac{3}{3} = 1$$

- $f(x \in \mathbb{N}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. Then:

$$rf(x) = \frac{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)}{T}$$

with:

$$T = \sum_{y \in \mathbb{N}} \left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)\right)$$

The only difficulty is the analytical computation of the normalization denominator so has to be able to use rf as a distribution for infinite sets, which is a difficult task in general. Moreover this could lead to divisions by a null or positive infinity value, and so the strictly rejection of the possibility for $\overline{f}$ to have a reprentative distribution in S with respect to its realised objects. It could not be the same negative fact in the same set when other ojects are chosen. But we are talking about it again in 4.2 and 5.1.

2.4 Discussion

We can say that this general framework is divided into two independant parts: the mathematical and probabilistic ones. Moreover it is as itself constraintless, in the sense that the only constraints that one may uncountered are not added by this framework but are due to the mathematical one: one has to be rigourous when confronted to realisation tasks, so as to obtain a valid mathematical expression of a positive computable function.

The probabilistic step also doesn't exhibit any constraint, neither in itself nor retroactively in $\overline{f}$ or its objects realisations, but only potential computational difficulties that are also merely mathematical ones, inherent in the way that mathematics has historically been defined to (superbly [Pen89]) fit and explain the real physical world around us.

In summary and conclusion, this framework potentially allows to statistically characterize all sets with all positive functions, by throwing away the stochastic constraint from the function to an independant normalization process. We are now seeing how the classical numerical statistical framework is subsumed by our general one, thus being a special case.

3 Where is the constraintful classical statistical framework?

By classical statistical framework, you have probably guessed that I am talking about the well-known *numerical* representation of the Kolmogorov axiomatical probabilistic theory [Kol33] in some subsets of the Euclidian space, *i.e.* a subset of $\mathbb{R}^n$ within all the well-known powerful objects of the Euclidian geometry (distance...). I believe that the motivation for using this formalism is the power of numericity added with the seemingly great accuracy ([Pen89]) of the Euclidian geometry to represent the physical space of the world in which we live.

But the Kolmogorov axiomatics and the Euclidian space could now be considered as unsufficient and thus the classical framework could itself be considered as too restrictive. The scientist has always tried to understand the physics that we undergoes from *natural* processes, so as to explain it, and possibly control and reproduce it as accurately as possible. I think that the classical framework has been developed with respect to that desire, especially for the stochastic part of such processes. But now that we are dealing with *artificial* processes, who are we to decide that the underlying laws governing these artificial statistics could still be driven by the Euclidian theory? The question is of important relevance because of the profound *symbolic* nature of the interactions between the human and its created pseudo-physics. Who are we to consider that the human creations are faith enough to the natural ones, and so decide that we can still use the classical framework by the way of applying pseudo-adequate discretisation or symbolisation processes to laws for which we have experimentaly proven that they could be the ones for which we could have confidence enough to use it as natural representations?

That is why I think that it could be interesting to be free from this context and try to have a general vision of the statistics field in order to be able to realise some (not necessary) well-known processes for the caracterization of sets of complex symbolic structures that one is likely to handle thanks to our artificial machines creations. We would not like to see this symbolic system as a new special one, totally independant from the classical numerical framework, but it could be more interesting to have an overview of what could be the features of a statistical process, to be able to reproduce it in all environments, within the variables we have to deal with in some particular context. Then we could study the possible properties of symbolic distributions with respect to this general framework, and thus make it possible for us to use the maximum of the power of the symbolic system, like this has been done in the numerical case.

Finally, I don't need some mathematical proof and I am quick by merely saying that the classical framework is effectively represented, in the general one, by the realisation in each subset S of R^n of each theoretical function $\overline{f}$ for which all theoretical objects have been chosen as the ones of the Euclidian geometry, and for which the such realised function f gives to S a total weight of 1, thus respecting the stochastic contraint of the Kolmogorov axiomatics, and thus for which the normalization process changes nothing. This classical framework is a little part of the general one, avoiding all functions that don't possess such native features in some considered sets, thus avoiding most part of positive functions to be expressed within a considered set in a probabilistic fashion. This is not the case of the general framework that could be, I think, an overall representation of probabilistic mathematics. The key is the separation of the mathematical and probabilistic parts, thus cancelling the stochastic constraint on f, that makes classical framework a constraintfull one: this constraint in f is not due to mathematical limitations but added by the framework itself because of the desire for respecting the Kolmogorov axiomatics.

4 Theoretical functions related to the uniform and normal processes

We are now trying to embed more particularly these two well-known distributions in our framework, this for three possible sets, which are, respectively, finite, infinite discrete, and (infinite) continous.

4.1 The uniform law

In fact, you have probably guessed that the first one of the two running examples given above is representative of a uniform process. More generally, any theoretical function that, after realisation, always gives the same positive abolute weight to any element of any realised set, is representative of a uniform process. The only restriction is that this constant must neither be 0, nor $+\infty$, because of the mathematical constraint that is imposed to have a valid normalization process.

The running example given above belongs to the most simple class of such theoretical functions: those that have no theoretical part, and thus for which the realisation part only consists of the set one. Here is another one example:

Example
$\overline{f}(\overline{x}) = \overline{V}$. There is a constraint of numerical value on the theoretical constant $\overline{V}$.

- For the finite case, let $S = \{a, b\}$, $V = 12$. The constraint on V is respected and $rf(x) = 12/24 = 1/2$.
- For the infinite case, let $S = \mathbb{N}$, $V = 12$. The constraint on V is respected and $rf(x) = 12/+\infty$, which violates the constraint on the normalization denominator, and confirm the impossibility to realise an uniform process in infinite discrete sets.
- For the continous case, let $S = [1; 4]$, $V = 12$. The constraint on V is respected and $rf(x) = 12/36 = 1/3$.

4.2 The normal law

Once again, you have probably guessed that the second case of the two running examples given above could be related to a normal process. The normal law is known as well representing the distribution of errors around a true value. In contradiction with what we proposed in [RJ08], directly derivated from ideas exposed in [Jol03], I really think now that this error is mesured in terms of numerical *distance* (or even maybe only *dissimilarity*) from the true value, and more generally that the standard deviation is an universally defined concept, independant of all realisation issues, as the expectation of the distance between an element and the expected element of a distribution.

My vision is that the normal process is represented by those classes of theoretical functions:

$$\overline{f_\sigma}(\overline{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\overline{d[\overline{S}]}(\overline{x}, \overline{\mu})^2}{\sigma^2}\right)$$

with $\overline{d[\overline{S}]}$ belonging to the class of bi-parametered theoretical functions that take their parameters in $\overline{S}$, and fulfill the three mathematical constraints of distance in $\overline{S}$ (or maybe only dissimilarity if one accepts the violation of the triangular inequality). Thus $\overline{d[\overline{S}]}$ adds a mathematical constraint on $\overline{\mu}$, as being member of $\overline{S}$.

Let us see some realisation examples of one of these functions, namely $\overline{f_5}$:

Example
$\overline{f_5}(\overline{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\overline{d}(\overline{x},\overline{\mu})^2}{25}\right)$.

- For the finite case, let $S = \{a, b\}$, $\mu = a$, $d[S](a, a) = d[S](b, b) = 0$, $d[S](a, b) = d[S](b, a) = 10$. Mathematical constraints are fulfilled, and $rf(a) \simeq 0.88, rf(b) \simeq 0.12$.
- For the infinite case, let $S = \mathbb{N}$, $\mu = 0$, $d[S](x \in S, y \in S) = \sqrt{(x-y)^2}$. Mathematical constraints are fulfilled.
- For the continous case, let $S = [1; 4]$, $\mu = 1$, $d[S](x \in S, y \in S) = \sqrt{(x-y)^2}$. Mathematical constraints are fulfilled.

Like observing in 2.3, the difficulty is situated in the computation of the denominator of the normalization process, so as to obtain the distribution rf from f, with a possibility of violation of the strict mathematical constraint that imposes this value as neither being 0 nor $+\infty$. This fact especially arises in infinite discrete cases, when the sum of the values of a function in one of such sets is $+\infty$. I think this is because of the sum operator that is not as powerful as the integral one, that can deal in a elegant way with the continous case in order to compute the total value of a function in an infinite set.

Two solutions are offered to us, in order to make probabilistic studies in infinite discrete sets:

1. Propose to mathematicians to discover a new total operator that can be used especially for infinite discrete sets, thus letting the sum one especially for finite ones. This operator would have to be valid when using within probabilistic issues, and could have a power between the two philosophically opposite other total operators that are the sum and the integral.
2. One could believe that the first solution is totally inconsistant, incoherent, and irrelevant with the profound nature of mathematics. So one could finally prefer to take into consideration only the functions for wich the value is not null for only a finite number of elements of the set.

The second solution is that of restricting the probabilistic influence in the set for only a finite number of elements, and thus acting like in the finite case. Remind that this solution still doesn't tackle the problem of *computing* this total value, if one hasn't any analytical solution for doing it in finite time, *i.e.* a solution that is proven to take all elements into account in finite time, *i.e.* a way of ignoring the null valued elements of the set, which is a general difficult problem if one hasn't any powerful information or property related to the *adequation* between the set and the function to be considered (for example $\sum_{e \in \mathbb{N}} = \frac{(-1)^e}{2e+1} = \frac{\pi}{4}$). This computation process is not a problem in the finite case, which has the advantage of being the bounded special one of the discrete one.

This general difficulty is thus potentially inherited by all infinite discrete sets, more particularly in symbolic structural cases, for which I am talking about now, and proposing to apply a special already known solution, that takes into account the structural specificity. We are only seeing the string case in order to have a first simple viewpoint of the task.

5 How to make statistics for strings?

Before trying to answer to this question, let us just recall some necessary definitions and notations. We say that a *string* is a *concatenation*, or *succession*, of *letters* that are taking their values in a set called *alphabet*. In the rest of this paper, we are identifying this alphabet as A, the set of all finite length strings on A as A^*, and the empty string of length 0 as λ, where the length of a string X is obviously given as the number of letters that are to be concatenated in order to obtain X, and where we are saying that a letter is no more than a string of length 1, except, for simplicity, λ, that can even represent the empty letter.

5.1 The univariate general difficulty

The problem that has arised to us in 2.3, and that we have more particularly developed in 4.2, is to be considered for all discrete infinite sets where one wants to deal with probabilities. Thus, even if A is taken as finite, we would have general difficulties to define some distributions in the infinite set A^*.

We can not enumerate all strings of A^* in order to compute the total value of a function f to be considered, because of an infinite time asking by this procedure. So we can apply the two possible solutions given in 4.2, or take into account the structural specificity of a string, *i.e.* see a distribution of strings as merely being a succession of distributions of letters, *i.e.* possibly benefit of a *multivariate* paradigm to statistically handle sets of strings.

5.2 The multivariate possible solution

This solution could possibly be considered for all types of complex structure, *i.e.* seeing a complex structure distribution as a multivariate distribution of more simple primitives that are composing a structure. This primitives could even be expressed as complex structures, and thus we could apply this paradigm in a recursive way, until have reached a simple enough computational point, namely P.

Notice that this solution could only be totally applied in the cases where the primitives that are underlying of P are not even expressed within a discrete infinite primitive alphabet. In this case, we could only apply the two general solutions in P, if we are necessary to deal with probabilities in the first-order most complex set A^*. Notice also that this solution avoids a great part of the potential distributions that one would like to discover in A^*, but the multivariate solution is the only one that I currently see in order to tackle a part of this difficult problem, when one would like to potentially handle the entire set A^* (waiting for this new powerful mathematical operator (*cf.* 4.2)...).

Obviously, this solution is not a discovery for us, only a confirmation. I especially think of this powerful machine, namely *probabilistic grammar* [VTdlH+05a, VTdlH+05b] (with numerous less powerful derivatives: automata, HMM, n-gram...), that is a simple structure representing a potentially infinite set of probabilistic strings in a multivariate fashion, and has been initially devoted to the representation of the uncertainty in a formal language.

In [RJ08], we have proposed the formalisation of what could be realisations of some uniform and normal laws of strings, with respect to the multivariate paradigm, and a mutual independancy assumption. This work has lead to encouraging results for the control and estimation of these string distributions, but has suffered from a lack of theoretical justification, that I hope doesn't exist any more now, thanks to the framework I am proposing to you. These definitions could easily be adapted to this framework, thus having a real theoretical value, but I think that the most interesting work to do could not be to try to *translate* what could be some or others well-known properties of the numerical paradigm to the symbolic one, but much more to *discover* the inherent ones that could arise in the symbolic case. We could thus discover that there is effectively a part of *universality* in statistics, but this must be, I think, done in a *general* manner, and not in a *discretisative* or *symbolicative* one, even if this could be a difficult challenge...

6 Conclusion

In this paper, I propose my ideas for the development of a new general formalism that could lead to an universal framework for the potential caracterization of each set by each positive function taking its parameters in the set, and representing a special realisation of a statistical process. The two keys are, first, the abstraction of some mathematics that are realisation-dependant for the statistical process, and, second, the rejection of the stochastic constraint from the function to an independant probabilistic part, by the way of an adequate normalization step, that depends on the nature of the set.

Combining these two issues leads to a more general framework than the classical one, and I have shown that the later is effectively subsumed by the new one. I have also proposed some discussions, and raises some difficulties that could arise while trying to handle some distributions in infinite sets of complex discrete structures. Some partial solutions are given to us, but that still remains a great challenging task.

Finally, I am working to understand why this general paradigm could be considered as a relevant one with respect to some possible phylosophical issues that are hidden in the nature of statistics, and also how it could effectively bring us some new knowledge, and thus being useful in practice, not being regarded as a mere mathematical game...

Acknowledgment

I would like to thank Pasquale Foggia for his in-depth criticisms and suggestions, and for having motivated me to make my ideas in paper, even if, afterward, they could be regarded as irrelevant with respect to the profound nature of statistics. I hope it won't be the case.

My visit in the MIVIA laboratory has been supported by a grant from the Exploradoc program of the *Région Rhône-Alpes*, France. I am grateful to Mario Vento for having accepted to receive me, and to all the members of the MIVIA laboratory for their welcoming.

References

[Jol03] J.-M. Jolion. The deviation of a set of strings. *Pattern Analysis and Applications*, 6(3):224–231, 2003.

[Kol33] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Julius Springer, 1933.

[Pen89] R. Penrose. *The Emperor's New Mind.* Oxford University Press, 1989.

[RJ08] S. Rebecchi and J.-M. Jolion. Lois uniformes et normales de chaînes discrètes. In *RFIA*, Amiens, France, January 2008. To appear.

[VTdlH+05a] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines - part I. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1013–1025, 2005.

[VTdlH+05b] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta, and R. C. Carrasco. Probabilistic finite-state machines - part II. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1026–1039, 2005.