A renewed approach to the foundations of SIA theory: generalizing SIA to incorporate multiple behavior hypotheses. Thoughts on the implicative intensity

Une approche renouvelée des fondements de l'ASI : Une généralisation de l'ASI afin d'introduire des hypothèses de pluralités des comportements; Réflexion sur l'intensité implicative.

THOMAS DELACROIX[1]

Abstract

The aim of this paper is to recall the foundations of SIA theory and modify these slightly in order to clarify and reinforce them, as well as make them more adaptable to various scientific models. Some issues regarding one of the founding blocks of SIA, the implicative intensity, are discussed and detailed. This paper is composed of two mostly separable sections. The first half of this paper focuses on a generalization of SIA so that known differences between individuals, which we shall refer to as multiple behaviors, can be taken into account in an analysis. The second half focuses on two issues related to the implicative intensity, the second of which being the well-known issue raised by large numbers of individuals in SIA data sets.
Keywords : *Statistical implicative analysis, Foundations, Multiple behaviors, Probability matrix, Distributions, Implicative intensity, Relative implicative intensity.*

Résumé

On souhaite ici rappeler les fondements de l'ASI et les modifier légèrement de manière à clarifier certains points et à les renforcer, ainsi qu'à les rendre plus adaptables à différentes modélisations scientifiques. Des questions soulevées concernant l'un des piliers de l'ASI, l'intensité d'implication, sont discutées et détaillées. L'article est articulé autour de deux parties relativement dissociables. La première partie concerne une généralisation de l'ASI permettant de prendre en compte dans les analyses des différences connues entre individus qui seront désignées par le terme de comportements multiples. La seconde partie concerne deux questions soulevées relatives à l'intensité d'implication, en particulier la question bien connue en ASI des difficultés d'analyse pour des données portant sur un grand nombre d'individus.
Mots-clés : *Analyse statistique implicative, Fondements, Comportements multiples, matrice de probabilités, Distributions, intensité d'implication, Intensité d'implication relative.*

We will consider in this paper only classical SIA theory, i.e. SIA on binary variables and a discrete set of individuals.

[1] The University of Auckland, Department of Mathematics, City-104.135, Auckland, New Zealand, maths.delacroix@gmail.com

The basic idea of SIA theory works as follows. Suppose we have a classical SIA data set. We have *n* individuals and a list of *p* different properties that are either true or false for each of these individuals. In SIA we first consider each set of two properties *a* and *b*. For these properties we can look at the following contingency table of the number of individuals having properties *a* or *b* :

	a	$\neg a$
b	$\left\|A \cap B\right\|$	$\left\|\overline{A} \cap B\right\|$
$\neg b$	$\left\|A \cap \overline{B}\right\|$	$\left\|\overline{A} \cap \overline{B}\right\|$

We then try to establish a probability distribution on the bottom left-hand box of such tables to which we will compare $\left|A \cap \overline{B}\right|$ by considering random variables *X* and *Y* corresponding to *A* and *B*. Firstly, we believe that this methodology might be a bit too hasty in the construction of the probability distribution that is used. This can be confusing regarding the hypotheses that justify the distribution. Hypotheses are often made on the behavior of individuals and the transition to groups of individuals is not necessarily that straightforward for an end user. By making the different steps in this construction more explicit, we render these hypotheses more apparent and more adaptable if they are to be modified. For this we will define a distribution at an anterior phase, from which we can infer a distribution on the contingency tables. Secondly, we look at the implicative intensity used to compare the empirical data with the constructed probability distribution. We define a new relative implicative intensity that we believe is more adequate than the regular implicative intensity. Then we outline the reasons for the limitations to the use of the implicative intensity in SIAs.We also chose to modify the notational conventions from regular SIA to render this renewed approach easier to express mathematically and possibly more adapted to straightforward programming of the formulas.

A distribution based on behaviors

We will no longer start with a probability distribution on the intersection of two sets, but infer this probability distribution from a probability distribution on the entire data set. By doing this, we make apparent the hypotheses and assumptions made about the data. It will make it easier to change them, so they are more adapted to a given situation. If we study

a group of individuals, it is often possible to model the behavior of these individuals. Such a model can easily be incorporated in an SIA using this more general approach. We will consider a certain number of individuals, finite or infinite, and a finite number p of properties. Each property is either true (1) or false (0) for a given individual[2].

Finite number of individuals

We first start looking at a probability distribution on the entire population of individuals, assuming this population is of finite cardinality n.

Let $\boldsymbol{M}_{p,n}$ be the set of $p \times n$ matrices, with all coordinates 0 or 1, and $\boldsymbol{D}$ be a probability distribution on $\boldsymbol{M}_{p,n}$. The random variables we will be considering are matrices $X \in M_{p,n}$. We define X_i the resulting random variable taken as the i-th row of X and $X_{i,j}$ the resulting random variable taken as the j-th coordinate of X_i.

Our main objective is to determine a probability distribution on contingency tables for any two properties (say 1 and 2) as is the case in regular SIA theory. This comes round to determining the value of

$$P(X_1 \cdot X_2 = m)$$

Note that we use different conventions from regular SIA theory for simplicity. The intersection is now seen as a dot product between two vectors. Also, it is of no importance yet that we look at the intersection of individuals that have property 1 and not property 2 rather than individuals that both share property 1 and property 2. We can infer the results on the prior from the latter, and doing otherwise would render reading more complicated.

Probability matrix

From any distribution $\boldsymbol{D}$, we can define a $p \times n$ probability matrix P as such:

$$\forall i, j, P_{i,j} = P(X_{i,j} = 1)$$

This matrix gives us the probability that a given property is true for a given individual, for all properties and individuals. In many cases, this can be determined by a model. We

[2]It is not because we have used the term individual here that these must be people in an applied context, this can be anything from people to objects to instances.

would therefore like to define $\boldsymbol{D}$ from P. Generally, P does not define the probability distribution. One probability matrix corresponds to a great number of different probability distributions. However, this is no longer true if we make adequate assumptions.

Generalized binomial distribution

The simplest assumption we can make is that of the independence of variables between all the $X_{i,j}$. In this case, we easily see that P defines a probability distribution $\boldsymbol{D}$. Indeed, let A be a matrix in $\boldsymbol{M}_{p,n}$, we have:

$$P(X = A) = \prod_{i,j}\left(P_{i,j}A_{i,j} + (1 - P_{i,j})(1 - A_{i,j})\right)$$

In this case, considering only properties 1 and 2, we have, for all m:

$$P(X_1 \cdot X_2 = m) = \sum_{\substack{I \subset [0,n] \\ |I| = m}} \prod_{i \in I} P_{1,i}P_{2,i} \prod_{j \notin I}(1 - P_{1,j}P_{2,j})$$

The name generalized binomial distribution is apparent in 1.1.4.

Generalized hypergeometric distribution

Another assumption we can make, is that of a conditional independence of variables between the $X_{i,j}$ relative to knowing all $X_i \cdot X_i$. This is a reasonable assumption in certain cases. For example, suppose we are looking at n students taking a list of p different competitive exams. As these are competitive exams (as opposed to regular exams), the number of candidates n_i that will pass a given exam i is predefined. This means $P(X_i \cdot X_i = n_i) = 1$or all i. If this is the only information we know about these exams, then conditional independence is a reasonable assumption.

In this case, considering only properties 1 and 2, we have, for all m:

$$P(X_1 \cdot X_2 = m) =$$

$$\sum_{\substack{I \amalg J \amalg K \amalg L = [0,n] \\ |I| = m \\ |I| + |J| = n_1 \\ |J| + |K| = n_2}} \prod_{i \in I} P_{1,i} P_{2,i} \prod_{j \in J} P_{1,j} \left(1 - P_{2,j}\right) \prod_{k \in K} \left(1 - P_{1,k}\right) P_{2,k} \prod_{l \in L} \left(1 - P_{1,l}\right)\left(1 - P_{2,l}\right)$$

The name generalized hypergeometric distribution is apparent in the following paragraph.

Regular SIA assumptions on *P*

Regular SIA theory does not look at each individual's particularities. All individuals are seen as the same. This means that, for all i, there exists a probability P_i, such that for all j, $P_{i,j}=P_i$. The value for P_i which is taken is $\frac{n_i}{n}$.

Using this matrix P under the assumptions for the generalized binomial distribution. We get:

$$P(X_1 \cdot X_2 = m) = \binom{n}{m} (P_1 P_2)^m (1 - P_1 P_2)^{n-m}$$

This gives us the same binomial distribution as the one used in regular SIA theory.

Now, using the same matrix P under the assumptions for the generalized hypergeometric distribution. We get:

$$P(X_1 \cdot X_2 = m) = \frac{\binom{n_1}{m}\binom{n - n_1}{n_2 - m}}{\binom{n}{n_2}}$$

This gives us the same hypergeometric distribution as the one used in (older) regular SIA theory.

We have no pretence here to say that these formulas have been found with better arguments than originally (see Models 1 and 2 from Lerman et al. (1981)). The reasoning here is actually very much the same as the one made for this 1981 paper. However, we believe it is important to keep this construction apparent, so as to allow more complex models.

Limitations

From a theoretical point of view, it is very easy to write down formulas for $P(X_1 \cdot X_2 = m)$. From a practical point of view, it can be very hard and even impossible to calculate them in a reasonable time. This is due to the extra complexity brought by considering different assumptions on individuals. If all assumptions on each individual are different, then the number of calculations that need to be done to calculate $P(X_1 \cdot X_2 = m)$ for the generalized binomial distribution will be of the order $\binom{n}{m}$. For a full SIA, the order can be up to n times the sum for all m less than n of all these orders, which is $n2^n$. This can easily become huge for greater values of n.

It is however possible to group individuals with common behavior together to make calculations less time consuming.

Grouping individuals by common behavior

We say that two individuals j, j' have same behavior if for all i, $P_{i,j}=P_{i,j'}$. We can group all individuals who have same behavior into q classes. We define $t=(t_1, ..., t_q)$ the vector such that t_k represents the number of individuals in each class. We have $1 \leq t_k \leq n$ and $t_1+...+t_q=n$. Furthermore, we define a $p \times q$ matrix $\tilde{P}$ such that the k-th column for $\tilde{P}$ is equal to any j-th column of P, where the individual j is in the behavior class k. Under the assumptions of the generalized binomial or hypergeometric distribution, the couple $(\tilde{P}, t)$ defines the probability distribution on random variables $X_1 \cdot X_2$.

We will only study the formula for the general binomial distribution[3]. This is:

$$P(X_1 \cdot X_2 = m) = \sum_{\substack{s_1+...+s_q=m \\ 0 \leq s_k \leq t_k}} \prod_{k=1}^{q} \binom{t_k}{s_k} \left(\tilde{P}_{1,k}\tilde{P}_{2,k}\right)^{s_k} \left(1-\tilde{P}_{1,k}\tilde{P}_{2,k}\right)^{t_k-s_k}$$

Now the number of calculations for a full SIA is of order:

[3]The formula for the general hypergeometric distribution can be written this way too, but is is quite complex and of little use. Indeed, as is true for the hypergeometric distribution, the general hypergeometric distribution gives way to symmetrical SIAs which in some sense waters down the whole idea behind SIA.

$$q\prod_{i=1}^{q}(t_i+1)$$

This is maximal when all classes have same size, so we can give an upper bound for this order, which is:

$$q\left(\frac{n}{q}+1\right)^q$$

This shows that there are still limitations even when grouping individuals in a limited number of behavior classes. With modern home computers, if it may be possible (meaning doable in a reasonable time) to go up to 10 classes for one hundred individuals for a full SIA, the number of classes would be limited to 5 for one thousand individuals and would fall down to 2 for one million individuals.

Infinite number of individuals

If we suppose we have an infinite number of individuals, we cannot define any matrix *P*. Even if we can define all $P_{i,j}$, it is still much less straightforward to look at different behaviors between each individual. For the time being we will take all assumptions on individuals to be the same. Therefore, even if we cannot consider the matrix *P*, we can define P_i as the probability that any individual has property *i*.

Even if the number of individuals is infinite, the data collected will always be on a finite number of individuals. If we can find a probability that the data collected is related to any given number of individuals, then we can adapt our previous distribution to this infinite case.

Indeed, let us consider that *X* is a random variable in $\boldsymbol{M}_{p,N}$ where *N* itself is a random variable. Then, under the assumptions for generalized binomial distribution, we have:

$$\begin{aligned}
P(X_1\cdot X_2=m) &= \sum_{k=0}^{\infty}P(N=k)\binom{k}{m}(P_1P_2)^m(1-P_1P_2)^{k-m}\\
P(X_1\cdot X_2=m) &= (P_1P_2)^m\sum_{k=m}^{\infty}P(N=k)\binom{k}{m}(1-P_1P_2)^{k-m}\\
P(X_1\cdot X_2=m) &= (P_1P_2)^m\sum_{k=0}^{\infty}P(N=k+m)\binom{k+m}{m}(1-P_1P_2)^{k}
\end{aligned}$$

Simple Poisson distribution

For many research studies in social sciences, it makes sense to model the probability that the collected data is related to a given number of individuals using a Poisson law. Indeed, the data is often collected over time and the fact that an individual participates in the study usually does not depend on the time since the last individual participated in the study. For example, the number of people agreeing to respond to a questionnaire on the street during a given time can easily be modeled through a Poisson distribution. And this is not only limited to social sciences.

In such a case, if n is the (empirical) number of collected data, we have:

$$P(N = k) = \frac{n^k}{k!}\mathrm{e}^{-n}$$

By substituting this in the previous formula, we see that:

$$\begin{aligned} P(X_1 \cdot X_2 = m) = & \quad (P_1P_2)^m \sum_{k=0}^{\infty} \mathrm{e}^{-n} \frac{n^{k+m}}{k+m!} \frac{(k+m)!}{m!k!} (1-P_1P_2)^k \\ P(X_1 \cdot X_2 = m) = & \quad \frac{(nP_1P_2)^m}{m!}\mathrm{e}^{-n}\mathrm{e}^{n(1-P_1P_2)} \\ P(X_1 \cdot X_2 = m) = & \quad \frac{(nP_1P_2)^m}{m!}\mathrm{e}^{-nP_1P_2} \end{aligned}$$

Therefore, $X_1 \cdot X_2$ follows a Poisson law of parameter nP_1P_2. Furthermore, if we take P_i to be equal to $\frac{n_i}{n}$ for all i as in regular SIA theory, then this parameter is equal to $\frac{n_1n_2}{n}$. This is the distribution that is used in most SIA.

This is of course no surprise. As previously, and even more so for this case, the construction of the formula is basically the same as in the original version (see Model 3 Lerman et al. (1981)). However, it is important to make apparent and adaptable the introduction of any new hypotheses in the model.

On a more practical note, it is good for users of SIA to know why they can (or cannot) use SIA for their studies. In most SIA literature, the reader is invited to read Lerman et al. (1981) or more recently Saporta (2006) for proof that the Poisson distribution discussed here is adapted to a certain drawing process. The hypotheses for the model are

not made easily apparent. It seems to us that the methodology presented in this article allows a better understanding of these hypotheses. These are the following:

1. The probability for each individual to have a given property is equal to the ratio of individuals having this property on the total number of individuals observed (i.e. $P_{i,j} = \frac{n_i}{n}$).

2. The fact that a given individual has a given property should not influence other such facts (i.e. independence of variables $X_{i,j}$).

3. The number of individuals on which data was collected can be modeled by a Poisson law (i.e. N follows a Poisson law).

In the most recent literature Gras et al. (2013), a list of three hypotheses h1, h2 and h3 are given for a potential SIA user. These hypotheses are based on the event [A and non B] and the characterization of a Poisson law in 2.4 Saporta (2006). We believe this is misleading as it is easy to forget that this implies that N is a random variable. In the next section of this article we show how this affects SIAs. Furthermore, the characterization for a Poisson law in Saporta (2006) is time based and though it is perfectly valid, we prefer to leave the possibility for a user to consider a Poisson law with space based hypotheses for example.

Multiple behaviors

We now come back to the general case where the population can be grouped into different behavior classes. As our population is infinite, we no longer consider a couple $\left(\tilde{P},t\right)$ where the coordinates of the vector t are integers, but a couple $\left(\tilde{P},r\right)$, where the k-th coordinate of r is the fraction of the population that belongs to the k-th behavior class.

This couple defines the distribution on variables $X_1 \cdot X_2$ as such:

$$P\left(X_1 \cdot X_2 = m\right) =$$

$$\sum_{k=0}^{\infty} P(N=k) \sum_{t_1+\ldots+t_q=k} \binom{k}{t} \prod_{i=1}^{q} r_i^{t_i} \sum_{\substack{s_1+\ldots+s_q=m \\ 0\le s_j\le t_j}} \prod_{j=1}^{q} \binom{t_j}{s_j} \left(\tilde{P}_{1,j}\tilde{P}_{2,j}\right)^{s_j} \left(1-\tilde{P}_{1,j}\tilde{P}_{2,j}\right)^{t_j-s_j}$$

Where $\binom{k}{t}$ is the classical multinomial coefficient equal to $\frac{k!}{t_1!...t_q!}$.

This formula corresponds to a very straightforward approach towards this problem as it has been defined. It is possible however to look at this differently and give a formula which is easier to work with, particularly so in the Poisson case. Indeed, let $y_{1,2,k}$ be the random variable corresponding to the number of individuals in a given behavior class k having properties 1 and 2. Then we have:

$$P(y_{1,2,k} = m) = \sum_{i=0}^{\infty} P(N = i)\binom{i}{m}\left(r_k \tilde{P}_{1,k}\tilde{P}_{2,k}\right)^m \left(1 - r_k \tilde{P}_{1,k}\tilde{P}_{2,k}\right)^{i-m}$$

Furthermore, as the different behavior classes are disjoint, we can write that:

$$P(X_1 \cdot X_2 = m) = \sum_{t_1+...+t_q=m} \prod_{k=1}^{q} P(y_{1,2,k} = t_k)$$

Generalized Poisson distribution

We now consider that N follows a Poisson law of parameter n as previously. Using the reasoning we have used for the simple Poisson distribution, we see that $y_{1,2,k}$ follows a Poisson law of parameter $\lambda_{1,2,k} = n r_k \tilde{P}_{1,k}\tilde{P}_{2,k}$. Therefore:

$$P(X_1 \cdot X_2 = m) = \sum_{t_1+...+t_q=m} \prod_{k=1}^{q} \frac{\lambda_{1,2,k}^{t_k}}{t_k!} e^{-\lambda_{1,2,k}} = \frac{\left(\sum_{k=1}^{q} \lambda_{1,2,k}\right)^m}{m!} e^{-\sum_{k=1}^{q} \lambda_{1,2,k}}$$

This last equality is given quite simply by the multinomial formula.

Thus, we see that in the case where there is a finite number of different behaviors, it is very simple to take these different behaviors into account while trying to determine the probability distribution on $X_1 \cdot X_2$ This probability distribution is in fact a Poisson distribution with parameter $\lambda_{1,2} = \sum_{k=1}^{q} \lambda_{1,2,k}$, where $\lambda_{1,2,k} = n r_k \tilde{P}_{1,k}\tilde{P}_{2,k}$ for all k.

Note that these simplifications are not surprising for an experienced probabilist as relations between multinomials and Poisson distributions are well known.

Overview of the approach

Generalized framework

This new approach defines a framework for performing an SIA on data while considering a model for the behavior of individuals. The model for the behavior of individuals is not provided by the framework and this is precisely the point of defining such a framework: the researcher can now consider which ever model for the behavior of individuals is best suited to his/her research.

It has been shown here that we can consider the model for the behavior of individuals used in classical SIA. In this case, the SIA obtained following the process defined by the multiple behavior framework would be the same as a classical SIA as shown in sections 1.1.4 and 1.2.1 of this paper. Hence, the new framework encloses classical SIA while allowing further possibilities.

As this paper focuses on the framework for considering different models for the behavior of individuals rather than these models, it does not present any alternative model to the one used in classical SIA. However, the development of this framework was motivated by the will to take into account a specific model for the behavior of individuals as explained in section 1.3.3 below.

Linearization

One of the other significant aspects of this new approach is the linearization of the problem: the variables considered are matrices rather than sets. This serves the mathematical clarity of the theory, which is crucial for defining a more complex framework than the classical one.

It also makes the transition from theory to computer programming much more straightforward. This aspect must not be underestimated in regard to a dissemination of SIA theory to a wider audience.

There does not seem to be any obstacles to an analogous linearization of the other current SIA approaches involving non-binary variables. Such a process would most certainly enable a grouping of all these different approaches into a single, more general framework and is called for by the author of this article.

Applications

As stated previously, the framework presented here was developed to incorporate in an SIA a specific model for the behavior of individuals. This model, constructed in Delacroix and Boubekki (2012), describes how students of different level manage questions of different difficulty. The computation of an SIA using this model via the multiple behavior SIA framework is presented in Delacroix and Boubekki (2013).

The use of this particular model in its given context helps, and the use of alternative models for the behavior of individuals to the classical SIA model in other contexts can help, reduce the number of irrelevant or parasitical quasi-implications in a SIA. This can be a great improvement to certain SIAs. Indeed, as it is recalled in section 2.2 of this present paper, SIA theory has to deal with the issue of excessive quasi-implications when considering large numbers of individuals. By enabling the researcher to reduce the number of irrelevant quasi-implications at an early stage, the multiple behavior SIA framework is part of the solution to this issue.

On the notion of implicative intensity

Comparing contingency tables

Implicative intensity

In the previous section, we have shown how a probability distribution for the values of what is usually denoted $\left|X \cap \bar{Y}\right|$ in the literature is defined. The next step in regular SIA is to look at the probability that $\left|X \cap \bar{Y}\right|$ is greater or equal to the empirical value $\left|A \cap \bar{B}\right|$ (this is usually denoted $\varphi(a,b)$ and called the implicative intensity). As we are looking at pseudo-implications $a \Rightarrow b$, the number of counter-examples to this rule is given by $\left|A \cap \bar{B}\right|$. Which is usually the justification for defining the implicative intensity as such:

$$\varphi(a,b) = P\left(\left|X \cap \bar{Y}\right| > \left|A \cap \bar{B}\right|\right)$$

Or with the conventions used in this article:

$$\varphi\left(a_1, a_{\bar{2}}\right) = P\left(X_1 \cdot X_{\bar{2}} > n_{1\wedge\bar{2}}\right)$$

Relative implicative intensity

The implicative intensity does not, however, take into account the fact that *N* can be a random variable. Indeed, as explained in the previous section (see 1.2), the number of individuals observed can be considered to follow a certain probability law. This is the case in most SIAs today, which correspond to the simple Poisson distribution case presented previously.

Let us consider the following examples for contingency tables:

	a	$\neg a$
b	1	1
$\neg b$	1	1

	a	$\neg a$
b	10	10
$\neg b$	2	10

Even though the number of counter-examples to the $a \Rightarrow b$ is greater for the second table, the ratio of counter-examples per reported data is 4 times higher in the first table than in the second.

If we analyse the data using a binomial distribution or a hypergeometric distribution (or even the generalised forms described previously), then this problem will not arise. In such a case, the number of individuals on which data was collected is not to be considered a random variable and the sum of all elements in each possible table will always be the same.

But, if we analyse the data using a Poisson distribution (or any distribution for which *N* is a random variable), this problem will arise. Actually, in the Poisson case, for any natural number *k*, the probability that the sum of the elements of a table is equal to *k* is non-zero. Therefore, it is important to take this into account. We suggest to define a new relative implicative intensity ψ by using the random variable *N* defined in 1.2 as such:

$$\psi(a_1, a_2) = P\left(\frac{X_1 \cdot X_{\bar{2}}}{N} > \frac{n_{1\wedge\bar{2}}}{n}\right)$$

This can be also written as:

$$\psi(a_1, a_2) = \sum_{k=0}^{\infty} P\left(X_1 \cdot X_{\bar{2}} > k\frac{n_{1\wedge\bar{2}}}{n} \middle| N = k\right) P(N = k)$$

Poisson case

If we look at how we have obtained a Poisson distribution on the value of $X_1 \cdot X_{\bar{2}}$ in 1.2.1, we can determine an expression for $\varphi(a_1, a_2)$ which can be used for the regular SIA Poisson case.

$$\psi(a_1, a_2) = \sum_{k=0}^{\infty} \left(\sum_{i>k\frac{n_{1\wedge\bar{2}}}{n}}^{k} \binom{k}{i} \left(\frac{n_1 n_{\bar{2}}}{n^2}\right)^i \left(1 - \frac{n_1 n_{\bar{2}}}{n^2}\right)^{k-i} \right) \frac{n^k}{k!} e^{-n}$$

This value is to be compared to the implicative intensity used in SIA today:

$$\varphi(a_1, a_2) = \sum_{j>n_{1\wedge\bar{2}}}^{\infty} \frac{1}{j!} \left(\frac{n_1 n_{\bar{2}}}{n}\right)^j e^{-\frac{n_1 n_{\bar{2}}}{n}}$$

These two formulas do not yield the same result. On the one hand, the formula for φ is much easier to compute. But on the other hand, we believe it does not give the most adequate information for an SIA under the Poisson hypotheses. If *n* is not too large, less than 100 000 for example, then modern computers can easily compute ψ. And if *n* is larger, then it is of little consequence to use one rather than the other, as we will explain in the following paragraph. Therefore, we recommend the use of the relative implicative intensity ψ for smaller samples and a different approach altogether for larger samples. We wish to reassure SIA users who have been using φ, that both the implicative intensity and the relative implicative intensity "behave" in a very similar fashion for the Poisson case.

Limitations of the (relative) implicative intensity

It is a known issue in the SIA community, that if *n* is too large (this problem is frequently observed with populations of thousands already), the implicative intensity tends to take values either 0 or 1. And using a relative implicative intensity does not help solve this problem in any way. This is the main motivation for introducing the entropic approach in SIA, for example.

We will not deal too much with such solutions to this problem here, although we will share a few remarks. Mainly, we will try to make apparent the mathematical reasons that

are behind this issue.

In most SIAs, the probabilities named P_i in this article, that any individual has property i, are determined by the ratio of the number of observed individuals that have property i on the number of individuals observed, which is $\frac{n_i}{n}$. When we do this, we usually implicitly consider this ratio to be a good approximation for the real value of this probability. If we consider a larger set of individuals, we do not expect this value to vary much, or this would mean that our approximation was not a good one. Therefore, when considering two properties 1 and 2, we expect $\beta_{1,2} = \frac{n_1 n_{\bar{2}}}{n^2}$ to be an invariant for our model. Furthermore, if we observe a system and we wish to generalize results from it to other systems, to larger systems, through an SIA, then it is necessary that the ratio of the number of individuals having property 1 and not 2 on the number of individuals is an invariant for such systems, as an SIA uses this value as a benchmark. This gives a second invariant $\alpha_{1,2} = \frac{n_{1\wedge\bar{2}}}{n}$.

We can therefore define, under whichever hypotheses we choose to work with, a sequence of implicative intensities $\varphi_n(a_1, a_2)$ entirely defined by these invariants. We show that, under the hypotheses used for SIAs:

$$\lim_{n\to\infty} \varphi_n(a_1, a_2) = 1 \quad if \quad \alpha_{1,2} < \beta_{1,2} \tag{1}$$

$$\text{and} \quad \lim_{n\to\infty} \varphi_n(a_1, a_2) = 1 \quad if \quad \alpha_{1,2} > \beta_{1,2} \tag{2}$$

In a certain sense, we show that the cumulative distribution function for an adequately defined sequence of probability distributions tends towards the cumulative distribution function of a Dirac measure of parameter $\beta_{1,2}$.

Binomial case

Let $0 < \alpha < \beta < 1$. We first show that :

$$\sum_{k=0}^{\alpha n} \binom{n}{k} \beta^k (1-\beta)^{n-k} \underset{n\to\infty}{\to} 0 \tag{3}$$

Indeed, we see that :

$$\sum_{k=0}^{\alpha n} \binom{n}{k} \beta^k (1-\beta)^{n-k} = (1-\beta)^n \sum_{k=0}^{\alpha n} u_k^{(n)} \quad \text{where} \quad u_k^{(n)} = \binom{n}{k} \left(\frac{\beta}{1-\beta}\right)^k$$

A quick study of the sequence $\left(u^{(n)}\right)_{k\in N}$ shows that :

$$u_{k+1}^{(n)} - u_k^{(n)} \geq 0 \Leftrightarrow k \leq \beta(n+1) - 1$$

Therefore, as $\alpha < \beta$, for any large enough *n* :

$$\sum_{k=0}^{\alpha n} \binom{n}{k} \beta^k (1-\beta)^{n-k} \leq (1-\beta)^n (\alpha n + 1) \binom{n}{\alpha n} \left(\frac{\beta}{1-\beta}\right)^{\alpha n}$$

Using Stirling's formula, we find that :

$$\binom{n}{\alpha n} \sim \frac{1}{\left(\alpha^\alpha (1-\alpha)^{1-\alpha}\right)^n \sqrt{2\pi\alpha(1-\alpha)n}}$$

Therefore :

$$(1-\beta)^n (\alpha n + 1) \binom{n}{\alpha n} \left(\frac{\beta}{1-\beta}\right)^{\alpha n} \sim \left(\left(\frac{\beta}{\alpha}\right)^\alpha \left(\frac{1-\beta}{1-\alpha}\right)^{1-\alpha}\right)^n \sqrt{\frac{\alpha n}{2\pi(1-\alpha)}}$$

It suffices to show that :

$$f_\beta(\alpha)^n \sqrt{\frac{\alpha n}{2\pi(1-\alpha)}} \underset{n\to\infty}{\to} 0 \quad \text{where} \quad f_\beta(\alpha) = \left(\frac{\beta}{\alpha}\right)^\alpha \left(\frac{1-\beta}{1-\alpha}\right)^{1-\alpha} \tag{4}$$

We study the function f_β defined on]0,1[using $\log(f_\beta)$. We find that f_β is strictly concave, going from $\lim_{\alpha\to 0^+} f_\beta(\alpha) = \beta$ to $\lim_{\alpha\to 1^-} f_\beta(\alpha) = 1-\beta$ with maximum reached only at $f_\beta(\beta) = 1$. Therefore, as $\alpha < \beta$:

$$0 < f_\beta(\alpha) < 1$$

From this comes the limit in (4) and the limit in (3) follows. This is the same as the limit given by (1), only taken in the binomial case.

Now, if $0 < \beta < \alpha < 1$, an entirely symmetrical proof gives :

$$\sum_{k=\alpha n}^{n} \binom{n}{k} \beta^k (1-\beta)^{n-k} \underset{n\to\infty}{\to} 0 \qquad (5)$$

This last result (5) is the same as (2) taken in the binomial case.

Poisson case

Let $0 < \alpha < \beta < 1$. The proof here is completely analogous to the binomial case. We will show now that :

$$\sum_{k=0}^{\alpha n} \frac{(\beta n)^k}{k!} e^{-\beta n} \underset{n\to\infty}{\to} 0 \qquad (6)$$

We see that :

$$\sum_{k=0}^{\alpha n} \frac{(\beta n)^k}{k!} e^{-\beta n} = e^{-\beta n} \sum_{k=0}^{\alpha n} v_k^{(n)} \quad \text{where} \quad v_k^{(n)} = \frac{(\beta n)^k}{k!}$$

And a quick study of the sequence $\left(v^{(n)}\right)_{k\in N}$ shows that :

$$v_{k+1}^{(n)} - v_k^{(n)} \geq 0 \quad \Leftrightarrow \quad k \leq \beta n - 1$$

Therefore, as $\alpha < \beta$, for any large enough *n* :

$$\sum_{k=0}^{\alpha n} \frac{(\beta n)^k}{k!} e^{-\beta n} \leq (\alpha n + 1) \frac{(\beta n)^{\alpha n}}{(\alpha n)!} e^{-\beta n}$$

Using Stirling's formula, we find that :

$$(\alpha n+1)\frac{(\beta n)^{\alpha n}}{(\alpha n)!}\mathrm{e}^{-\beta n} \sim \left(\left(\frac{\beta}{\alpha}\right)^{\alpha}\mathrm{e}^{\alpha-\beta}\right)^{n}\sqrt{\frac{\alpha n}{2\pi}}$$

It now suffices to show that :

$$g_\beta(\alpha)^n\sqrt{\frac{\alpha n}{2\pi}} \underset{n\to\infty}{\to} 0 \quad \text{where} \quad g_\beta(\alpha)=\left(\frac{\beta}{\alpha}\right)^{\alpha}\mathrm{e}^{\alpha-\beta} \tag{7}$$

The study of the function g_β defined on $]0,1[$ using $\log(g_\beta)$ gives that g_β is strictly concave, going from $\lim_{\alpha\to 0^+} g_\beta(\alpha)=\mathrm{e}^{-\beta}$ to $\lim_{\alpha\to 1^-} g_\beta(\alpha)=\beta\mathrm{e}^{1-\beta}<1$ with maximum reached only at $g_\beta(\beta)=1$. Therefore, as $\alpha<\beta$:

$$0<g_\beta(\alpha)<1$$

Which gives us (7), from which follows (4), which is the same as (1) in the Poisson case.

Now if $0<\beta<\alpha<1$, then we can show that:

$$\sum_{k=\alpha n}^{\infty}\frac{(\beta n)^k}{k!}\mathrm{e}^{-\beta n} \underset{n\to\infty}{\to} 0 \tag{8}$$

Indeed, if we separate this sum in two :

$$\sum_{k=\alpha n}^{\infty}\frac{(\beta n)^k}{k!}\mathrm{e}^{-\beta n}=\sum_{k=\alpha n}^{n^2-1}\frac{(\beta n)^k}{k!}\mathrm{e}^{-\beta n}+\sum_{k=n^2}^{\infty}\frac{(\beta n)^k}{k!}\mathrm{e}^{-\beta n}$$

Then, as by Taylor-Lagrange :

$$\sum_{k=n^2}^{\infty}\frac{(\beta n)^k}{k!}\le\frac{(\beta n)^{n^2}}{(n^2)!}\mathrm{e}^{\beta n}$$

And :

$$\sum_{k=\alpha n}^{n^2-1}\frac{(\beta n)^k}{k!}\le n^3\frac{(\beta n)^{\alpha n}}{(\alpha n)!}$$

We have :

$$\sum_{k=\alpha n}^{\infty} \frac{(\beta n)^k}{k!} e^{-\beta n} \leq n^3 \frac{(\beta n)^{\alpha n}}{(\alpha n)!} e^{-\beta n} + \frac{(\beta n)^{n^2}}{(n^2)!}$$

By Stirling's formula, we find that :

$$n^3 \frac{(\beta n)^{\alpha n}}{(\alpha n)!} e^{-\beta n} \sim \frac{n^3}{\sqrt{2\pi n}} \left(\frac{\beta}{\alpha} e^{\alpha - \beta} \right)^n$$

And :

$$\frac{(\beta n)^{n^2}}{(n^2)!} \sim \frac{1}{n\sqrt{2\pi}} \left(\frac{\beta}{n} e \right)^{n^2}$$

The first part of the sum goes to 0 by our previous result on $g_\beta(\alpha)$ and the second part clearly goes to 0. Therefore, we have shown (8) which is the same as (2) in the Poisson case.

Gaussian case

Even though we have not discussed this case in this paper, a normal distribution can also be used in SIA (see, for example, Gras et al. (2013)). In this case, the result comes spontaneously. Indeed, (1) and (2) are in this case the same as :

$$\frac{1}{\sqrt{2\pi}} \int_{h_n(\alpha,\beta)}^{\infty} e^{\frac{-t^2}{2}} dt \underset{n\to\infty}{\to} 1 \quad if \quad \alpha < \beta \text{ and } \frac{1}{\sqrt{2\pi}} \int_{h_n(\alpha,\beta)}^{\infty} e^{\frac{-t^2}{2}} dt \underset{n\to\infty}{\to} 0 \quad if \quad \alpha > \beta$$

Where $h_n(\alpha,\beta) = \frac{\alpha n - \beta n}{\sqrt{\beta n}}$.

And the result is straightforward.

Relative implicative intensity in the Poisson case

There is no need to look at the relative implicative intensity in the binomial case as in this case it is equal to the implicative intensity. We limit ourselves to the most used case in SIAs, i.e. the Poisson case.

Let $0 < \alpha < \beta < 1$. We want to show that :

$$\sum_{k=0}^{\infty}\left(\sum_{i=0}^{\alpha k}\binom{k}{i}\beta^{i}(1-\beta)^{k-i}\right)\frac{n^{k}}{k!}\mathrm{e}^{-n} \underset{n\to\infty}{\to} 0$$

This is actually quite straightforward from what we have done in 2.2.1. Indeed, let $f_{\beta}(\alpha) < \gamma < 1$, for *k* large enough (say $k > K$) we have :

$$\sum_{i=0}^{\alpha k}\binom{k}{i}\beta^{i}(1-\beta)^{k-i} \le \gamma^{k}$$

Therefore :

$$\sum_{k=0}^{\infty}\left(\sum_{i=0}^{\alpha k}\binom{k}{i}\beta^{i}(1-\beta)^{k-i}\right)\frac{n^{k}}{k!}\mathrm{e}^{-n} \le \sum_{k=0}^{K}\left(\sum_{i=0}^{\alpha k}\binom{k}{i}\beta^{i}(1-\beta)^{k-i}\right)\frac{n^{k}}{k!}\mathrm{e}^{-n} + \sum_{k=K+1}^{\infty}\frac{(\gamma n)^{k}}{k!}\mathrm{e}^{-n}$$

$$\sum_{k=0}^{\infty}\left(\sum_{i=0}^{\alpha k}\binom{k}{i}\beta^{i}(1-\beta)^{k-i}\right)\frac{n^{k}}{k!}\mathrm{e}^{-n} \le M\mathrm{e}^{-n} + \mathrm{e}^{(\gamma-1)n} \underset{n\to\infty}{\to} 0$$

If $0 < \beta < \alpha < 1$, an entirely symmetrical proof gives :

$$\sum_{k=0}^{\infty}\left(\sum_{i=\alpha k}^{k}\binom{k}{i}\beta^{i}(1-\beta)^{k-i}\right)\frac{n^{k}}{k!}\mathrm{e}^{-n} \underset{n\to\infty}{\to} 0$$

Therefore, we have shown that the relative implicative intensity will be of no more help than the implicative intensity for most practical applications when the number of individuals is too large.

Remarks

We have shown that the issues raised by large numbers of individuals in SIAs are inherent to the method itself, if it is applied to the type of study it was precisely designed for. Furthermore, the calculations show that the convergence of the implicative intensity

towards 0 or 1 is at least geometrical. Such a convergence is quite fast so it will occur whenever slightly bigger populations are considered. Thus, it is an issue that SIA theory must deal with. The current solution to this issue is the entropic approach. It is an interesting approach and we believe it should be developed. However, some matters of clarification are necessary. It is argued that the quasi-implicative model is more concerned about the rule $a \Rightarrow b$ than the rule $\neg b \Rightarrow \neg a$. This is false. In non-entropic SIAs, there is no difference between these two rules. Determining the implicative intensity of the rule $\neg b \Rightarrow \neg a$ gives exactly the same result as determining the rule $a \Rightarrow b$. And this is how it should be in any SIA. Therefore, the correction of this "issue" in the entropic approach seems a bit artificial. Indeed, a difference between these two rules is artificially created, so that these two "different" rules may be balanced in the new model. An entropic approach, without this construction, seems however entirely conceivable. One alternative approach is to consider that if *n* is large enough that the implicative intensity appears to be either 0 or 1 for all observed properties, then this simply shows that we have enough data to consider alternative implicative indices. And that we can use these indices to construct a hierarchical tree, rather than the implicative intensity. Even though implicative indices have been considered before, it seems that this approach has not been considered much and we believe it should not be overviewed. Another approach, which is currently investigated by the researcher, is to consider quasi-implications between crossed properties (e.g. $a \wedge b \wedge c \Rightarrow d \wedge e$). The implicative intensities for such rules are mechanically less than those for simple properties. If the computational complexity of a systematic review of all rules on crossed properties is exponential therefore ruling it out, algorithmically selecting a reduced number of relevant such rules is entirely conceivable.

References

DELACROIX T., (2012), É*tude d'un module "langage mathématique" en tant que module préparatoire à l'activité mathématique en algèbre linéaire de L1*, Masters thesis, Université Paris 7 Denis Diderot.

DELACROIX T., Boubekki A. (2012), *A regression analysis for taking students' levels into account in didactics studies*, preprint.

GRAS R., Bailleul M., et al. (2000), *Actes des journées sur : La fouille dans les données par la méthode d'analyse statistique implicative*, Eds Régis Gras et Marc Bailleul.

GRAS R., Kuntz P., Briand H. (2001), Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, **n° 154-155**, 9-29.

GRAS R., Suzuki E., Guillet F., Spagnolo F. (2008), *Statistical Implicative Analysis: theory and applications*, Vol. 127, Springer Verlag.

GRAS R., Régnier J.C., Marinica C., Guillet F. (2013), *L'analyse statistique implicative: Méthode exploratoire et confirmatoire à la recherche des causalités*, Ed. Cépaduès, Toulouse.

LERMAN I.C., Gras R., Rotsam H. (1981), Élaboration et évaluation d'un indice d'implication pour des données binaires, I et II, *Mathématiques et Sciences Humaines*, **n° 74**, 5-35, **n° 75**, 5-47.

REGNIER J.C., Bailleul M., Gras R., et al. (2012), *L'analyse statistique implicative: de l'exploratoire au confirmatoire*, Université de Caen.

SAPORTA G. (2006), *Probabilités, analyse de données et statistique*, Edition Technip.