EXTENSION DE L'ANALYSE STATISTIQUE IMPLICATIVE AU CAS DES VARIABLES CONTINUES QUELCONQUES

Régis GRAS[1], **Jean-Claude REGNIER**[2]

EXTENSION OF THE IMPLICATIVE STATISTICAL ANALYSIS TO ANY CONTINUOUS VARIABLE

RESUME

Après avoir généralisé l'Analyse Statistique Implicative au cas où l'espace des sujets est continu, nous étendons son champ d'application au cas où cette fois les espaces des variables sont continus sur [0; 1]. Ainsi, les variables seront observées sur des intervalles munis d'une loi de répartition continue. Nous procédons, tout d'abord, à l'extension à partir du traitement connu en ASI des variables-intervalles. Puis, nous envisageons un cas particulier où les distributions sur les espaces des variables suivent une même loi uniforme. Enfin, nous traitons le cas général de l'extension aux espaces de variables munis de lois différentes et quelconques.

***Mots-clés** : variables, variable-intervalle, variable continue, densité, intensité d'implication, propension.*

ABSTRACT

After having generalized the implicative statistical analysis to the case in which the subjects' space is continuous, we extend its field of application to the case in which variables' spaces are continuous on [0 ; 1]. Thus, the variables will be observed as interval-variables. We present a specific case where the distributions of the variables' spaces follow the same uniform law. Finally, we treat the general case of the extension to variable spaces with various and undefined laws.

***Keywords:** variables, interval-variable, continuous variable, density, intensity of application, propensity.*

1 Introduction

Une matrice de données numériques croisant classiquement sujets et variables étant donnée, l'Analyse Statistique Implicative (ASI ou SIA en anglais) attribue une mesure de qualité à des énoncés du type « si un sujet satisfait la variable a alors, généralement, il satisfait la variable b ». Nous avons traité (Gras et Régnier, 2012) le cas où, en ASI, l'espace des sujets E est continu muni d'une loi de répartition donnée. Nous abordons ici la situation où, cette fois, les variables actives elles-mêmes sont continues. Quel est le sens de cette continuité ? Elle exprime que, dans la population-mère de sujets, les observations de la variable suivent une loi continue, par exemple une loi gaussienne ou

[1] Ecole Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, Site de la Chantrerie, rue C.Pauc, BP 44306, Nantes cedex 3, e-mail : regisgra@club-internet.fr,

[2] Laboratoire UMR 5191 ICAR – Université Lumière de Lyon –Lyon2, 86 rue Pasteur, 69635 LYON Cedex 07, e-mail : jean-claude.regnier@univ-lyon2.fr

une loi uniforme. La population E de la matrice des données est censée en être une fidèle réalisation.

Précisons. Le type de variable générique que nous étudierons est une variable numérique dont la distribution est continue, de densité donnée et dont les valeurs finies sont prises sur IR ou sur un intervalle de IR. Si sa nature initiale est qualitative, nous nous limiterons au cas où cette « qualité » est quantifiable et ordonnée. Citons, par exemple, des variables qui expriment des saveurs, des sensations, des phénomènes climatiques,… On rencontre en effet ce type de situation dans le cas de données sensorielles. Mais aussi, comme dans (Gras et al, 2001), à la faveur du recueil des notes attribuées à des élèves dans des disciplines variées et conduisant à des intervalles de IR.

L'ensemble V des variables a, b, c… de l'étude peut être constitué de telles variables continues de distributions souvent différentes, de densités de mesure : f_a, f_b, f_c,… et présenter des espaces de valeurs également différents que l'on peut tous ramener par une homothétie convenable, à l'intervalle [0 ; 1]. Si X est la variable aléatoire représentant la valeur de a inférieure à α, alors Prob [$X<\alpha$] = $\int_0^{\alpha} f_a(t)dt$ et de même : Prob [$X\geq\alpha$] = $\int_{\alpha}^{1} f_a(t)dt$ représente la probabilité pour que la valeur de a soit supérieure ou égale à α. Si a(s) (resp. b(s)) sont les valeurs observées chez le sujet s selon a (resp . b), nous évaluerons dans quelle mesure on observe « rarement » dans E : a(s) > b(s).

Pour chaque sujet s de E, espace des sujets supposé ici discret et fini, nous disposons de la valeur numérique réelle prise selon chacune des variables soit : a(s), b(s), c(s),…Ces valeurs traduisent, dans la plupart des cas sémantiques, un degré d'intensité d'adhésion du sujet s à la variable ou de satisfaction de celle-ci. Elles sont, rappelons-le, manifestement ordonnées sur IR ou un intervalle réel.

Dans cet article, nous revenons d'abord sur la situation traitée dans le cadre des variables-intervalles où chaque intervalle est pondéré par la fréquence de ses instanciations. Le découpage de l'ensemble des valeurs de chaque variable par une partition optimale permet de définir les implications d'intervalles. Dans la section suivante 2, nous étudions le cas où les variables sont continues et uniformément distribuées sur l'ensemble de leurs valeurs. Nous définissons alors un indice de propension entre ces variables. Dans les deux dernières sections 3 et 4, nous étendons l'étude au cas où les variables ont une distribution quelconque mais, possiblement, différentes entre elles. Dans un premier temps, nous traitons le problème par la méthode générale de l'implication statistique. Ensuite, nous envisageons la même situation mais par l'approche propensive comme dans la section 2.

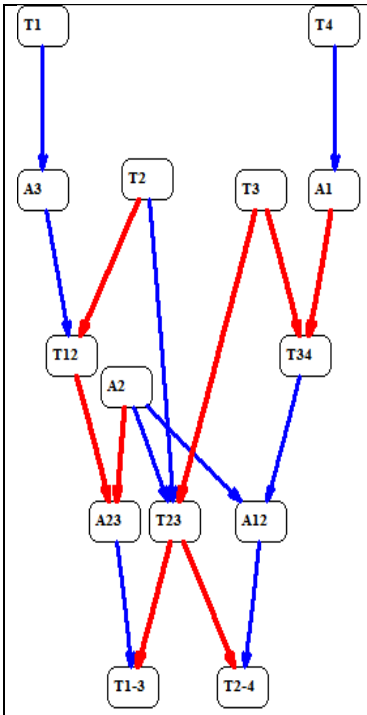
2 Première approche

Pour l'ensemble des variables a, b, c,…, nous examinons, sur leur espace de réalisation ramené à [0, 1], l'ensemble des valeurs prises par les sujets de E selon leurs densités respectives f_a, f_b, $f_{c,\ldots}$non explicitement définies. Notre objectif est de décomposer l'intervalle des valeurs de chaque variable sous forme de sous-intervalles réels. Pour ce faire, sur la base de la répartition des valeurs prises par les n sujets, nous

recherchons la meilleure partition de ces sous-intervalles en un nombre k pour la variable a (l pour b, m pour c, …) maximisant la variance interclasse pour un nombre fixé de sous-intervalles. Cette partition est effectuée selon la méthode inspirée et élargie à partir de celle des « Nuées dynamiques » de E. Diday (1972) et que nous avons utilisée précédemment pour définir l'intensité d'implication entre les variables-intervalles (Gras et al, 2001, 2013). Supposons la partition optimale obtenue constituée des sous-intervalles distincts A_1, A_2,…, A_k (resp. B_1, B_2,…B_l et C_1, C_2, …C_m), eux-mêmes avatars de sous-variables a_1, a_2, …, a_k (resp. b_1, b_2, …,b_l et c_1, c_2, …,c_m.) L'ensemble des valeurs prises par l'ensemble des sujets sur ces sous-intervalles est fini et est représentable par un intervalle de IR, réunion des A_i, que nous pouvons ramener à [0 ; 1] par une homothétie bijective. Nous convenons qu'à A_1 (resp. à A_k), par exemple, s'agrègeront les valeurs non observées « en bout » respectivement à gauche et à droite. Le tableau de données original est remplacé par un nouveau tableau à valeurs binaires. Sur la ligne du sujet x qui prend sa valeur en a selon la modalité a_i (ou de façon équivalente dans A_i) par exemple, on note la nouvelle valeur $1(A_i)$. (mesure de A_i). Ainsi, pour tout x qui satisfait a_i, alors $a_i(x) = 1$ et $a_j(x) = 0$ pour les autres sous-intervalles. Les différentes intensités d'implication des a_i sur les modalités b_j et vice-versa, sont calculées comme pour des variables binaires sous-intervalles par sous-intervalles. Un algorithme permet également de conjoindre, des sous-intervalles contigus d'une variable et d'évaluer les implications optimales d'une variable segmentée vers une autre. Puis les règles sont représentées par un graphe implicatif.

Exemple

On cherche à valider l'hypothèse (certes très vraisemblable) que, généralement, « être de grande taille » en 2013 implique « être jeune ». On dispose pour cela d'un échantillon, pris au hasard dans la population de français, de 77 sujets dont les âges varient de 19 ans à 60 ans et les tailles de 168 cm à 198 cm.



On obtient par C.H.I.C., logiciel de traitement de l'ASI, le graphe implicatif ci-contre lorsque l'on demande que les données numériques de la variable A (âges) soient partitionnées en 3 sous-intervalles optimaux et celles de la variable T (tailles) en 4 sous-intervalles optimaux :

Partitions optimales des données de la variable A :

A1 de 19 à 31 (soit moins de 31)

A2 de 32 à 43

A3 de 46 à 60 (soit plus de 46)

Partitions optimales des données de la variable T:

T1 de 168 à 172 (soit moins de 172 cm)

T2 de 173 à 176

T3 de 178 à 184

T4 de 185 à 198 (soit plus de 185 cm)

On observe que les grandes tailles impliquent (intensité d'implication 0.95) plutôt les jeunes : T4 => A1 (ce qui ne surprendra pas le lecteur) et que ces mêmes jeunes impliquent les tailles de 178 à 198 : A1 => T34

Remarques

1° La prise en compte des mesures des sous-intervalles permet de limiter l'intérêt aux implications à la hauteur de leur importance pondérale.

2° Par la finesse du découpage (k sous-intervalles) certes d'une part on multiplie le temps de calcul, mais d'autre part, on améliore la représentation des nuances de la distribution des instanciations.

3° Une autre approche du problème des variables continues, comparable à l'extension de l'espace des sujets au cas continu, sera à l'étude dans les trois sections suivantes.

4° Ainsi, la distribution a posteriori définie pour chaque variable telle que a à partir de la partition de A en sous-intervalles est explicitement donnée sous forme d'histogramme sur l'intervalle [0 ; 1]. Elle est une des réalisations contingentes de la distribution a priori de a soit f_a. De ce fait, l'implication entre intervalles pourrait être désignée comme implication entre histogrammes.

3 Cas où la variable est uniformément distribuée sur [0 ; 1].

Les extensions de traitements de données dans une approche ASI, développées à partir des premiers travaux de R. Gras, par M. Bailleul (1994) et J.B. Lagrange (1998),

considèrent les variables à valeurs sur [0 ; 1] comme des variables discrètes. Ce que nous souhaitons introduire maintenant, est la prise en considération de variables continues à valeurs sur l'intervalle [0 ; 1] ou s'y ramenant. Ici, nous allons, dans un premier temps, étudier le cas des variables continues sur [0 ; 1] dont la distribution de probabilité est uniforme sur cet intervalle et identique pour toutes les variables.

3.1 Propriétés de la variable Z, produit de deux variables aléatoires continues de loi uniforme sur [0 ; 1]

La problématique de l'étude présente s'exprime ainsi :

Exprimer par un nombre compris entre 0 et 1, dans quelle mesure l'observation, sur un ensemble de sujets E, des valeurs (numériques ou numérisées) prises par une variable continue a s'accompagne généralement de l'observation de valeurs plus grandes prises par une variable continue b.

Nous satisfaisons bien ainsi la philosophie de l'Analyse Statistique Implicative puisque nous voulons quantifier la tendance de la variable a à prendre des valeurs sur E inférieures aux valeurs prises selon la variable b. Cette quantification porte le nom de **propension**. Nous suivons la même procédure que celle suivie par J.-B. Lagrange à propos des variables modales discrètes. Soient X et Y variables aléatoires continues à valeurs dans [0 ; 1] de loi uniforme continue. Elles représentent les valeurs aléatoires des variables a et b qui se réaliseront dans l'observation. Comme il est coutumier en Analyse Statistique Implicative, les deux variables X et Y sont supposées indépendantes. Nous savons alors que la densité de probabilité de X ou de Y est donnée par la fonction $f(u)=1_{[0\,;\,1]}(u)$ de laquelle nous pouvons calculer la fonction de répartition :

$$F(u) = \begin{cases} 1 & u > 1 \\ u & 0 \le u \le 1 \\ 0 & u < 0 \end{cases}$$

Par ailleurs sans difficulté on montre que les espérances respectives de X et Y valent $E(X) = E(Y) = \frac{1}{2}$ et les variances $V(X) = V(Y) = \frac{1}{12}$.

De manière évidente, la variable $\overline{Y} = 1 - Y$ est encore une variable aléatoire à valeur dans [0 ; 1] de loi uniforme continue dont l'espérance et la variance sont celles de Y.

Poursuivant le chemin réalisé dans la construction de l'indice de propension établi avec des variables modales, nous posons Z = X(1-Y). Z est encore une variable aléatoire continue à valeurs dans [0 ; 1] mais qui n'admet plus une densité uniforme. On montre que la densité de probabilité de Z est $g(u) = -\ln(u)\, 1_{]0\,;\,1]}(u)$ ou encore que sa fonction de répartition est :

$$G(u) = \begin{cases} 1 & u > 1 \\ -u\ln(u) + u & 0 < u \leq 1 \\ 0 & u \leq 0 \end{cases}$$

À partir de ces informations, nous pouvons calculer l'espérance puis la variance de la variable aléatoire continue à valeurs sur [0; 1] de la manière suivante:

$E(Z) = \int_{0+}^{1} uf(u)du = \left[-\frac{1}{2}u^2 \ln(u) + \frac{1}{4}u^2 \right]_{0+}^{1} = \frac{1}{4}$ mais comme Z est le produit de deux variables indépendantes, nous pouvons obtenir directement : $E(Z) = E(X)\,E(1-Y) = \frac{1}{4}$

Pour ce qui est de la variance, nous savons que :

$V(Z) = E[(Z-E(Z))^2] = E(Z^2)-[E(Z)]^2$

Or $E(Z^2) = \int_{0+}^{1} u^2 f(u)du = \left[-\frac{1}{3}u^3 \ln(u) + \frac{1}{9}u^3 \right]_{0+}^{1} = \frac{1}{9}$

donc la variance de Z vaut $V(Z) = \left(\frac{\sqrt{7}}{12} \right)^2$

3.2 Indice de propension et intensité de propension pour des variables aléatoires continues de loi uniforme sur [0 ; 1]

Pour suivre la procédure de construction qui mène à la formule de l'indice de propension, nous considérons le n-uplet de variables iid (Z_1, Z_2,…, Z_n).

Posons $T_n = \frac{1}{n}\sum_{k=1}^{k=n} Z_k$. Il en résulte que $E(T_n) = \frac{1}{n}\sum_{k=1}^{k=n} E(Z_k) = \frac{1}{4}$ ainsi que $V(T_n) = \left(\frac{\sqrt{7}}{12\sqrt{n}} \right)^2$ puisque les variables Z_k sont indépendantes deux à deux. Nous sommes en mesure de fournir l'expression algébrique exacte de l'indice de propension entre les deux variables X et Y.

$$Q_u(X, \overline{Y}) = \frac{T_n - \frac{1}{4}}{\frac{\sqrt{7}}{12\sqrt{n}}} = \sqrt{n}\,\frac{T_n - \frac{1}{4}}{\frac{\sqrt{7}}{12}} = \sqrt{n}\,\frac{\left[\frac{1}{n}\sum_{k=1}^{k=n} X_k(1-Y_k)\right] - \frac{1}{4}}{\frac{\sqrt{7}}{12}}$$

On peut considérer que la variable aléatoire $Q_u(X, \overline{Y})$ indice de propension entre deux variables continues de loi uniforme sur [0 ; 1] suit approximativement une loi de Laplace-Gauss centrée réduite dans la mesure où sa construction même nous place dans les conditions d'un des théorèmes central-limite.

L'intensité d'implication se calcule donc de la façon suivante :

$$\varphi_u(X,Y) = 1 - \Pr ob\left[Q_u(X,\bar{Y}) \le q_u(X,\bar{Y})\right] = \frac{1}{\sqrt{2\pi}} \int_{q_u(X,\bar{Y})}^{\infty} e^{-\frac{t^2}{2}} dt$$

4 Cas général traité par la méthode ASI classique

4.1 Problématisation

Pour chaque variable observée, nous avons jusqu'alors examiné, sur son espace de réalisation, l'ensemble des valeurs prises sur l'ensemble des sujets E. Nos travaux sur les variables numériques et modales ont porté, comme dans la section 2, sur la définition d'une mesure de propension dont nous venons de rappeler une modélisation. A l'instar de J.B. Lagrange (1998) et de S. Guillaume (2000), nous utilisons encore l'expression : **propension** (ou tendance) de a vers b, dès lors que l'on rencontre généralement dans E peu de transactions s∈ E dans lesquelles a(s) > b(s)) pour la relation d'ordre sur [0 ; 1] où a(s) et b(s) sont respectivement les valeurs de a et b observées en s. Notons que si a(s) et b(s) sont des valeurs de rang ou d'intensité préférentielle, on peut exprimer l'inégalité a(s) > b(s) par : « a est préférée à b en s »

Notre exploration implicative, notée encore a=>b, dans le cas des variables continues est ici plus générale et s'exprime de la façon suivante :

Dans quelle mesure peut-on considérer que, pour x∈ E « ***si a(x) ≥ α, alors on peut affirmer que b(x) ≥ β*** » (F1) où α et β sont des réels de l'intervalle [0 ; 1] choisis dans l'analyse. Autrement dit, observe-t-on généralement que : « *si une réalisation pour x selon a dépasse la valeur α alors sa réalisation selon b dépasse aussi la valeur β* ? » (F2). Cette fois, contrairement à ce qui est envisagé dans la section 2, ce sont toutes les valeurs de deux intervalles, pondérés par des densités, qui sont mobilisées. De plus, les bornes de ces intervalles peuvent être indépendantes et de sémantique différentes. La réponse confère à la mesure associée une valeur prédictive.

Notons que si $\alpha = \beta$ égalité à laquelle on peut toujours ramener les deux ensembles de valeurs des deux variables, nous retrouvons la formulation de l'ensemble des contre-exemples, en tout point comparable à celle de la propension conceptualisée précédemment pour les variables numériques (Gras et al, 2009, 2013). En effet, un contre-exemple aux énoncés des règles (F1) et (F2) apparaît dès lors que a(s) > b(s). Comparable certes, mais pas identique épistémologiquement car le caractère continu demeure par la prise en compte des intervalles. Or, rappelons que dans le cadre des variables numériques a et b (Gras et Régnier, 2009, 2013), pour tout i∈ E, nous notons $\bar{b_i}$ le complément à 1 de bi : $\bar{b_i}$ = 1-bi, valeur de la variable b qui récuse b_i. On choisit comme pour l'implication entre variables binaires, l'indice $\sum_{i \in E} a_i \bar{b_i}$ - qui prend la valeur $n_{a \wedge \bar{b}}$ dans le cas binaire - comme indice de non-propension (ou de non-tendance) de a vers b. Ainsi, intuitivement et grossièrement, plus cet indice sera petit, plus on pourra s'attendre à une propension de a vers b (Gras et Régnier, 2009). Toutefois, cet indice ne tient plus dans le cas qui nous intéresse maintenant où les variables sont continues.

4.2 Formalisation de l'intensité d'implication entre variables continues

La formalisation va s'inspirer de celle retenue dans le cas où l'espace des sujets est continu (Gras et Régnier, 2013, p. 165-175). Considérons les sous-ensembles de sujets A et B de E admettant les valeurs respectives selon les variables a et b et satisfaisant les inégalités suivantes : $A=\{s\in E\ /a(s) \geq \alpha\}$ et $B=\{s\in E\ /b(s) \geq \beta\}$ où α et β sont deux valeurs quelconques appartenant à l'intervalle [0 ; 1]. Les cardinaux de A et B sont respectivement n_a et n_b (avec $n_a \leq n_b$). Par conséquent, les sujets s du sous-ensemble $A \cap \overline{B}$ vérifient $a(s) \geq \alpha$ et $b(s) < \beta$ ou encore $1-\beta \leq 1-b(s)<1$

Comme nous le faisons dans le cas binaire, nous comparons le nombre de contre-exemples à l'implication a => b (au sens de la relation F1 indiquée dans la section 3.1) observés dans la contingence et celui que l'on obtiendrait si les variables a et b étaient indépendantes. Pour cela, choisissons respectivement au hasard, de façon indépendante, deux parties $\hat{X}$ et $\hat{Y}$ de E de mêmes cardinaux respectifs que A et B et vérifiant les mêmes inégalités que ces deux sous-ensembles. Cependant la probabilité associée au tirage de $\hat{X}$ et $\hat{Y}$ parmi les parties satisfaisant les propriétés cardinales doit être affectée de celle qui traduit les contraintes pesant sur les valeurs prises selon a et b par les sujets des sous-ensembles $\hat{X}$ et $\hat{Y}$. Cette seconde probabilité est calculée à partir de la distribution produit entre les lois respectives de a et de b. Or, sous l'hypothèse d'indépendance *a priori* de a et b, la loi produit est égale au produit des lois respectives de a et b de densités f_a et f_b nulles à l'extérieur de [0 ; 1]. Les variables a et $\overline{b}=1-b$ sont également indépendantes et $f_{\overline{b}}=1-f_b$ est la densité de probabilité de $\overline{b}$. Les fonctions de répartition associées sont respectivement F_a, F_b et $F_{\overline{b}}$.

Ainsi la probabilité qu'un sujet pris au hasard ait un comportement vis-à-vis des modalités continues de a et de celles de b qui soit un contre-exemple à a => b (c'est à dire que $a(s) \geq \alpha$ tandis que $0 \leq b(s) < \beta$ ou $1-\beta \leq \overline{b}(s) <1$) est

$$\left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{0}^{\beta} f_b(u)du\right]=\left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{1-\beta}^{1} f_{\overline{b}}(u)du\right]=\left[1-F_a(\alpha)\right]\left[1-F_{\overline{b}}(1-\beta)\right]$$

A chaque sujet s de $\hat{X}\cap\hat{\overline{Y}}$, nous associons une variable de Bernoulli de paramètre p, qui est la probabilité qu'il soit à la fois dans l'ensemble des contre-exemples et qu'il vérifie la double inégalité $a(s)\geq \alpha$ et $b(s)<\beta$. Ces quatre événements étant indépendants par hypothèse, la probabilité p s'écrit comme produit de leurs probabilités, à savoir :

$$p=\frac{n_a n_{\overline{b}}}{n^2}\left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{0}^{\beta} f_b(u)du\right]$$

De là, Card($\hat{X}\cap\hat{\overline{Y}}$), somme des variables de Bernoulli indépendantes associées à chaque sujet de $\hat{X}\cap\hat{\overline{Y}}$ peut être considérée comme une variable binomiale de paramètres n et p.

$$\Pr\left[\text{Card}(\hat{X}\cap\hat{\overline{Y}})=k\right]=C_n^k p^k (1-p)^{n-k}$$

Espérance et variance de cette variable aléatoire sont donc :

E[Card ($\hat{X}\cap\hat{\overline{Y}}$] = np et Var [Card ($\hat{X}\cap\hat{\overline{Y}}$)]= np(1-p)

Par suite,

$$\Pr[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) \leq \mathrm{Card}(A \cap \overline{B})] = \sum_{k=0}^{k=\mathrm{Card}(A \cap \overline{B})} C_n^k p^k (1-p)^{n-k}$$

Comme dans le cas de variables discrètes, nous définissons l'intensité d'implication de a sur b par :

$$\varphi(a,b) = 1 - \Pr\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) \leq \mathrm{Card}(A \cap \overline{B})\right]$$

Remarque 1

Cette approche a-t-elle un sens dans sa restriction au cas de variable binaires ? La réponse est oui. En effet si les variables a et b sont binaires, la relation d'implication a => b n'est pas satisfaite lorsque pour $s \in \hat{X} \cap \hat{\overline{Y}}$, a(s) = 1 et b(s) = 0 ou $\overline{b}(s) = 1$. Les densités de probabilité sont dégénérées. Pour que l'on ait Pr[{$s \in X$ | a(s)=1]=1, il est suffisant que $\alpha = 0$. De même, pour que ait Pr[{$s \in \overline{Y}$ | b(s)=0]=1, il est également suffisant que $\beta=1$. Les deux intégrales définissant p sont alors égales à 1 et p devient égale à $\frac{n_a n_{\overline{b}}}{n^2}$, expression effectivement conforme à ce que nous avons dans la théorie de l'ASI dans le cas binaire.

Remarque 2

Comme nous le montrons dans (Gras et al, 2009, 2013), nous pouvons modéliser le cardinal aléatoire de $X \cap \overline{Y}$ par une variable de Poisson de paramètre

$$\lambda = \frac{n_a n_{\overline{b}}}{n} \left[\int_{\alpha}^{1} f_a(t)dt\right]\left[\int_{0}^{\beta} f_b(u)du\right]$$

Par suite $\forall s \in \{0,1,2,...,n\}$ $\Pr[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) = s] = \frac{\lambda^s}{s!} e^{-\lambda}$

et $\varphi(a,b) = 1 - \Pr\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) \leq \mathrm{Card}(A \cap \overline{B})\right] = 1 - \sum_{s=0}^{s=\mathrm{Card}(A \cap \overline{B})} \frac{\lambda^s}{s!} e^{-\lambda}$

Pour $\lambda \geq 5$, la variable « indice d'implication empirique », notée :

$$Q(a,\overline{b}) = \frac{\mathrm{Card}(X \cap \overline{Y}) - E\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}})\right]}{\mathrm{Var}\left[\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}})\right]} = \frac{\mathrm{Card}(\hat{X} \cap \hat{\overline{Y}}) - \lambda}{\sqrt{\lambda}}$$

qui résulte du centrage-réduction de la variable de Poisson, $\mathrm{Card}(\hat{\overline{X}} \cap \hat{\overline{Y}})$, peut être approchée par la variable gaussienne centrée réduite N(0 ;1).

Si nous considérons la valeur empirique $q(a,\overline{b}) = \frac{n_{a \wedge \overline{b}} - \lambda}{\sqrt{\lambda}}$, alors l'intensité d'implication estimée de la quasi-règle $a \Rightarrow b$, est approximativement :

$$\varphi(a,b) = 1 - \Pr\left[Q(a,\bar{b}) \le q(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Insistons sur le sens de cette intégrale. Elle représente la probabilité gaussienne pour que le nombre de transactions observées satisfaisant la quasi-règle $a \Rightarrow b$, soit supérieur à celui qui serait observable sous l'hypothèse d'indépendance de a et b. Autrement dit, $\Pr[Q(a,\bar{b}) \le q(a,\bar{b})]$ est la p-value du test visant à réfuter l'hypothèse de l'indépendance de a et b au profit d'une relation de type quasi-implication.

5 Cas général traité par la méthode ASI propensive

5.1 Formalisation de la relation de propension entre variables modales

J.B. Lagrange (1998) a construit dans le cas des variables modales, un indice de **propension** entre variables modales qui généralise l'indice d'implication entre variables binaires. En posant les conditions suivantes :

- si $a(x)$ et $\bar{b}(x)$ sont les valeurs prises en x par les variables modales a et $\bar{b}$, où $\bar{b}(x) = 1 - b(x)$

- si s_a^2 et $s_{\bar{b}}^2$ sont les variances empiriques des variables a et $\bar{b}$

Définition 1: L'indice de propension de variables modales est :

$$\tilde{q}(a,\bar{b}) = \frac{\sum_{x \in E} a(x)\bar{b}(x) - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{(n^2 s_a^2 + n_a^2)(n^2 s_{\bar{b}}^2 + n_{\bar{b}}^2)}{n^3}}}$$

Cette solution apportée au cas modal est aussi applicable au cas des *variables fréquentielles,* voire *des variables numériques positives,* à condition d'avoir normalisé les valeurs observées sur les variables, telles que a et b, la normalisation dans [0 ; 1] étant faite à partir du maximum de la valeur prise respectivement par a et b sur l'ensemble E.

Tout en suivant une modélisation comparable à celle de la section 2.1 où la loi de chaque variable est celle de la variable continue uniforme sur [0 ; 1], nous envisageons le cas où les variable sont continues, mais de lois données différentes et à valeurs ramenées par normalisation sur l'intervalle [0 ; 1]. Par ailleurs, notre recherche porte, comme dans la section 3, sur la relation F 1 :

Dans quelle mesure peut-on considérer que, pour $x \in E$ « ***si a(x) ≥ α, alors on peut affirmer que b(x) ≥ β*** » ?

Reprenons les notations de la section 2 : X (resp. Y, resp . $\overline{Y}$) est une variable aléatoire de [0 ; 1], ensemble normalisé des valeurs a_i (resp. b_j et 1- b_j) de la variable a

(resp. b, resp. $\bar{b}$) mais de loi de densité de probabilité f_a (resp. f_b et f_{1-b}) quelconque. De cette manière nous avons Prob $[X \geq \alpha] = \left[\int_{\alpha}^{1} f_a(t)dt\right]$ et Prob$[Y < \beta] = \left[\int_{0}^{\beta} f_b(u)du\right]$.

Les variables aléatoires X et Y (donc également $\bar{Y}$) sont, par hypothèse commune en ASI, indépendantes. Leurs paramètres, espérance et variance, sont calculables et donnés par les formules habituelles.

Introduisons la variable aléatoire $Z = X\bar{Y} = X.(1-Y)$. Elle représente la conjonction d'une réalisation des variables a et 1-b. La loi de Z est le produit des lois de X et de 1-Y et sa densité de probabilité est le produit des celles des lois de probabilité de X et 1-Y. L'espérance de Z est alors connue par : $E(Z) = E(X)E(1-Y) = M_1$ et sa variance est : $V(Z) = E(Z^2) - [E(Z)]^2 = M_2 - M_1^2$

Nous considérons le n-uplet $(Z_1, Z_2, \ldots, Z_n)$ composé de n variables aléatoires indépendantes de même loi que Z et dont la réalisation est celle des n sujets de E. Nous définissons alors comme dans la section 2, la variable somme : $T_n = \frac{1}{n}\sum_{k=1}^{k=n} Z_k$. Il en résulte que l'espérance $E(T_n) = \frac{1}{n}\sum_{k=1}^{k=n} E(Z_k) = M_1$ ainsi que la variance $V(T_n) = \frac{1}{n}\sum_{k=1}^{k=n} V(Z_k) = \frac{1}{n}(M_2 - M_1^2)$ puisque les variables Z_k sont mutuellement indépendantes.

En utilisant le théorème de la limite centrale dit de Lindeberg-Lévy puisque l'indice de propension est une moyenne de variables aléatoires indépendantes identiquement distribuées, on démontre que T_n suit approximativement, pour n grand, la loi de Laplace-Gauss d'espérance $E(Z) = M_1$ et de variance $\frac{1}{n}(M_2 - M_1^2)$

Autrement dit, la loi de la variable « indice de propension empirique »

$$\tilde{Q}(a, \bar{b}) = \frac{T_n - M_1}{\sqrt{\frac{1}{n}(M_2 - M_1^2)}}$$

est approximativement la loi gaussienne centrée réduite N(0 ; 1).

Dans une mise en pratique, la calcul d'une réalisation de la variable « indice de propension empirique », nécessite d'estimer l'espérance et la variance de la variable T_n. Pour ce faire, nous utilisons m_a la moyenne des observations a_i et m_b la moyenne des observations b_i ; qui nous permet d'obtenir une estimation de la moyenne M_1 avec $m_a(1-m_b)$. Ensuite nous réalisons une estimation de la variance de T_n à partir de la variance v_a des a_i et v_b celles des b_i, une estimation du moment M_2 d'ordre 2 est obtenue à partir des observations a_i et b_i par $(v_a + m_a^2)(v_b + m_{\bar{b}}^2)$ puisque les observations b_i et $\bar{b}_i$ ont la même variance.

L'indice empirique d'implication devient :

$$\tilde{q}(a,\bar{b}) = \frac{\frac{1}{n}\sum_{i\in E} a_i \bar{b}_i - m_a m_{\bar{b}}}{\sqrt{\frac{v_a(v_b + m_{\bar{b}}^2) + m_a^2 v_b}{n}}}$$

Quant à l'estimation de l'intensité de propension, elle est encore obtenue par :

$$\varphi(a,b) = 1 - \Pr\left[\tilde{Q}(a,\bar{b}) \le \tilde{q}(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{\tilde{q}(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Nous avons démontré (Gras, Régnier, 2013) que la réduction de ces formules exprimées dans le cas continu à l'instar des variables numériques au cas binaire était parfaitement valide.

6 Conclusion

Nous avons examiné différentes situations relatives aux variables principales continues dans le cadre de l'Analyse Statistique implicative de l'A.S.I. Partant de la situation la plus élémentaire où ces variables sont uniformément distribuées, nous avons considéré les variables continues à distribution quelconque sur [0 ; 1]. Après un retour simplificateur par les variables continues par intervalle, nous avons établi des lois permettant de calculer le critère fondamental en A.S.I., l'intensité d'implication qui évalue la qualité implicative d'une variable sur une autre, leurs distributions pouvant être différentes. Nous pouvons retenir la méthodologie adaptée pour traiter le cas continu dans d'autres types d'analyse de données. De plus, nous retenons la capacité dont nous bénéficierons dans des applications où il s'agit de nuancer spécifiquement la connaissance des variables à prendre en compte dans l'analyse, comme par exemple nuancer l'information sur la complexité *a priori* de ces variables pour une population de sujets donnée.

Références

[1] Bailleul, M., (1994). *Analyse statistique implicative : variables modales et contribution des sujets. Application à la modélisation de l'enseignant dans le système didactique*. Thèse Université de Rennes 1

[2] Diday E, (1972), *Nouvelles méthodes et nouveaux concepts en classification automatique et reconnaissance des formes*, Thèse d'Etat, Université de Paris VI.

[3] Gras R., Régnier J.C. et Guillet F. (2009). *L'Analyse Statistique Implicative. Une méthode d'analyse de données pour la recherche de causalités*, RNTI-E-16, Cépaduès Editions

[4] Gras R., Diday E., Kuntz P et Couturier R. (2001), Variables sur intervalles et variables-intervalles en analyse statistique implicative, *Actes du 8ème Congrès de la Société Francophone de Classification, Université des Antilles-Guyane, 17-21 décembre 2000,*166-173

[5] Gras R. et Régnier J.-C. (2013). Extension de l'A.S.I. aux variables non binaires in *L'Analyse Statistique Implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, Cépaduès Editions, 70-77

[6] Lagrange J.-B., (1998), Analyse implicative d'un ensemble de variables numériques ; application au traitement d'un questionnaire à modalités modales ordonnées, *Revue de Statistique Appliquée*, XLVI, 71-93.

[7] Régnier, J.-C., & Gras, R. (2005) Statistique de rangs et Analyse Statistique Implicative. *Revue de Statistique Appliquée*. 53(1) p.5-38

Ouvrages de référence

[1] *L'implication statistique. Nouvelle méthode exploratoire de donnée*, sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble (1996)

[2] *Mesures de Qualité pour la Fouille de Données*, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

[3] *Quality Measures in Data Mining* , F.Guillet et H.Hamilton eds, Springer, 2007,

[4] *Statistical Implicative Analysis, Theory and Applications*, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

[5] *Analyse Statistique implicative. Une méthode d'analyse de données pour la recherche de causalités*, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, 2009.

[6] *Teoria y Aplicaciones del Analisis Estadistico Implicativo*, Eds : P.Orus, L.Zamora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, 2009..

[7] *L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire*. Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012

[8] *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités*, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, 2013.