Une approche “ hétéro-statistique ” et graphique des masses de données d’enquête – le logiciel PointG

Stéphane Champely, Brice Lefevre, Julie Thomas, Sylvain Ferez

Une approche « hétéro-statistique » et graphique des masses de données d'enquête – le logiciel PointG

Article *in* Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique · November 2012

DOI: 10.1177/0759106312454633

CITATIONS
2

READS
1,182

4 authors:

Stephane Champely
University of Lyon
74 PUBLICATIONS **2,198** CITATIONS
SEE PROFILE

Brice Lefèvre
Claude Bernard University Lyon 1
52 PUBLICATIONS **264** CITATIONS
SEE PROFILE

Julie Thomas
Université Jean Monnet
37 PUBLICATIONS **92** CITATIONS
SEE PROFILE

S. Ferez
Université de Montpellier
197 PUBLICATIONS **466** CITATIONS
SEE PROFILE

Bulletin de Méthodologie Sociologique
116 25–43

DOI: 10.1177/0759106312454633
http://bms.sagepub.com

Une approche « hétéro-statistique » et graphique des masses de données d'enquête – le logiciel PointG

Stéphane Champely
Brice Lefèvre
(Université Lyon 1, EA 647)

Julie Thomas
Sylvain Ferez
(Université Montpellier 1, EA 4614)

Abstract

A "Hetero-statistical" and Graphical Approach for Massive Survey Data – The PointG Software. In sociology, one is often confronted with the problem of graphic presentations based on processing massive quantities of data from questionnaire surveys. Moreover, potential users can be of low statistical expertise, thus requiring relatively easy to use software such as PointG with drop-down menus. It is aimed at facing the difficulties due to the mass of collected quantitative data, both at the strategic, tactical and operational levels of statistical processing. It offers conventional ways of analysis going from univariate to multivariate data, but also automates treatments on groups of variables and calculates the size of overall and local effects. Apart from the graphic aspects, its distinctive "hetero-statistic" approach permits the user not to worry about the nature of the variables in bivariate analyses and during regression and factor analysis. PointG was developed within the powerful distribution free R programming environment.

Corresponding Author:
Stéphane Champely, UFR STAPS, Université Lyon 1, 27-29 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France
Email : champely@univ-lyon1.fr

Résumé
En sociologie, on est souvent confronté au problème de présentations graphiques basées sur le traitement des grandes quantités de données provenant des enquêtes par questionnaire. En outre, les utilisateurs potentiels peuvent être d'expertise statistique faible, ce qui nécessite un logiciel d'une utilisation relativement simple tel que PointG avec ses menus déroulants. Il est conçu pour faire face à des difficultés dues à la collecte massive de données quantitatives, tant aux niveaux stratégique et tactique qu'opérationnel du traitement statistique. Il propose de façon classique des analyses allant de l'univarié au multivarié, mais surtout permet d'automatiser des traitements sur des groupes de variables et de calculer des tailles d'effet globales comme locales. Il se distingue par une approche « hétéro-statistique » qui permet à l'utilisateur de ne pas se soucier de la nature des variables lors d'analyses bivariées, ainsi que lors de régressions et d'analyses factorielles. PointG a été développé dans l'environnement puissant et libre de distribution de R.

Keywords
Sociology, Survey Questionnaire Data, Statistics, R Software, Graphics

Mots clés
Sociologie, Données de questionnaire, Statistiques, Logiciel R, Graphiques

Introduction

Depuis plusieurs années certains d'entre nous, en tant que spécialistes des méthodologies quantitatives au sein de laboratoires de sciences sociales appliquées au sport, analysent statistiquement des données de questionnaire, pour notre propre compte ou celui de nos collègues. Nous enseignons également ces techniques et recevons régulièrement des demandes de consultation ou de formation à ce travail d'analyse. Si le logiciel est un outil indispensable à cette activité, nous allons ici défendre l'idée que la création d'un logiciel peut également être un vecteur de la diffusion de notre expérience, de notre conception de la statistique et probablement de la progression de notre propre expertise. En tant que sociologues, on ne s'étonnera pas de « l'influence des objets-outils » (Selz et Maillochon, 2009 : 27) dans ce processus d'analyse et notamment de l'importance de la connaissance de l'arsenal statistique moderne (Wheaton, 2003). En tant que concepteur de logiciel, on essaiera de « réguler » cette influence.

La conception d'un logiciel devrait prendre en compte primo la classe de problèmes statistiques considérée, en l'espèce l'analyse de données de questionnaire en sciences sociales, secundo notre conception de la statistique, et tertio les utilisateurs ciblés, leur(s) culture(s) scientifique(s), leurs objets et problématiques, leurs méthodes statistiques de base et leurs données.

Tous les questionnaires ne se ressemblent pas. Selon la discipline, on observe des variations importantes. Ainsi, la règle d'or de « faire au plus court » est rarement de mise en sociologie – comme nous le verrons sur l'exemple qui soutient notre présentation – et la structure de l'objet y est plus complexe que dans d'autres domaines. La formulation des

questions conduit généralement à des réponses catégorielles[1] plus que numériques. Les batteries d'échelles – que nous prendrons en considération car une part de notre activité de formation et de recherche est proche du « socio-marketing » – sont plus rares et moins « validées » comme elles peuvent traditionnellement l'être en psychologie ou en marketing. De même, la posture d'analyse quantitative est plutôt exploratoire par rapport aux disciplines précédentes, où le paradigme hypothético-déductif est prévalent et l'analyse en grande partie confirmatoire. Il ne s'agit toutefois pas d'une exploration « aveugle » comme peut l'être le *data mining*, mais d'un regard large orienté par la connaissance du terrain et la problématique sociologique (Blöss et Grossetti, 1999 ; Martin, 2005 ; Cibois, 2009). Dans ce cadre, les sociologues cherchent certes à vérifier que les réponses attendues aux questions posées sont obtenues (les hypothèses) ; mais s'efforcent d'entendre et de prendre en compte les réponses inattendues. Quant à la posture méthodologique que nous défendons en statistique, elle peut être résumée dans la recherche de « méthodes (essentiellement graphiques) plus orientées vers la découverte que les résultats quantifiables[2] » (Huber, 2011). Ainsi s'agit-il d'explorer, de découvrir, de décrire, en utilisant les méthodes inférentielles surtout comme garde-fou pour savoir si le hasard peut produire des découvertes similaires, et à l'occasion pour extrapoler les résultats de notre échantillon[3] à une population.

Enfin, le public d'utilisateurs que nous visons est plutôt celui des étudiants, doctorants ou sociologues avec une expertise statistique limitée. Que supposer des connaissances de cette audience en la matière ? Une bonne maîtrise du traitement du tableau croisé (les données catégorielles et leur relation étant la « base du métier ») ? Une connaissance de quelques tests bivariés, surtout celui du chi-carré d'indépendance de Pearson ? Et quelques expériences de lecture de plans factoriels ? En outre, pour se faire une idée du niveau anxiogène de la relation du sociologue à la statistique, on tapera dans le moteur de recherche du site Amazon® (anglais) les deux mots clés « Sociology Statistics »[4] et on répétera l'expérience avec « Psychology Statistics », « Linguistics statistics » ou « Biology Statistics ».

On concevra alors aisément que, pour ce public d'utilisateurs, les données de questionnaire semblent trop nombreuses et de structure trop complexe, et que le traitement de ces données pose à la fois des problèmes stratégiques, tactiques et opérationnels pouvant leur faire perdre leur « lucidité analytique » (Blöss et Grossetti, 1999). Comment les aider à acquérir les compétences qui leur manquent ? Notre réponse a longtemps été « un bon livre de statistiques et un logiciel »[5] mais serait à présent « un bon livre de statistique est un logiciel ». Entendre par là un logiciel adapté à la sociologie, au traitement de questionnaire, au type d'utilisateurs (selon leurs connaissances et leurs moyens financiers), afin qu'ils puissent se concentrer sur les problèmes stratégiques et interprétatifs plutôt que sur les problèmes tactiques ou opérationnels.

Après avoir présenté brièvement le jeu de données (inclus dans le logiciel) qui servira de support à notre présentation, nous examinerons les principes sous-jacents à la conception de ce logiciel. Nous suivrons une progression traditionnelle avec l'étude univariée, la classique étude bivariée, puis une spécificité de l'activité qui est la multiplication des traitements bivariés, que nous appellerons étude *mulBivariée* ; pour finir par l'approche réellement *mulTivariée*. Les activités fondamentales de prétraitement ne seront rapidement évoquées qu'au moment de la conclusion.

Le jeu de données

L'enjeu de l'enquête est d'étudier l'accès aux activités physiques et/ou sportives (APS) des personnes vivant avec le VIH (PVVIH)[6]. Etaient notamment étudiés l'effet du diagnostic ; l'effet de la socialisation au rôle/statut de PVVIH comme malade et plus précisément comme « malade chronique » : conséquence des parcours de soins (en lien avec les services hospitaliers, les médecins de ville, etc.), de la prise d'informations et de « l'éducation thérapeutique », de la fréquentation et de l'implication ou non dans les associations, etc. ; l'effet du rapport à son corps et au regard de l'autre, et inversement, l'effet du diagnostic sur cela.

Pour ce faire, hormis une cinquantaine d'entretiens non-directifs, nous disposons de 619 questionnaires. L'enjeu de l'échantillonnage a été de diversifier au maximum les profils de répondants (au regard de la pratique des APS, mais aussi au regard de l'ancienneté du diagnostic et des modes de contamination, des modes de prise en charge du VIH, des situations sanitaires, professionnelles et sociales). Nous avons cherché à montrer à la fois les effets communs du diagnostic de l'infection au VIH sur la problématique de l'accès aux APS, mais aussi l'hétérogénéité des ressources et des stratégies pour gérer cette problématique.

Le questionnaire comporte quatre parties : 1. Expérience en matière d'activités physiques et/ou de sport ; 2. Expérience du VIH, rapports au corps, traitements ; 3. Soin de soi et habitudes de vie ; 4. Informations générales (variables sociodémographiques, statut social, gestion globale de l'information sur le VIH auprès des proches, variables sociopolitiques et religieuses). La présentation du logiciel portera essentiellement sur cette dernière partie, largement commune à toutes les études.

Univarié - Saucisson et visualisation

Afin de s'adapter au type de public visé par ce logiciel, il a été choisi d'employer une interface graphique avec menus déroulants, format auquel il est habitué[7]. L'architecture de ces menus est simple – deux niveaux – et sa progression suit le déroulement théorique[8] des analyses de questionnaire : prétraitement, univarié dont signalétique[9], bivarié, mulBivarié et mulTivarié. L'objectif du logiciel est de donner un point de vue graphique sur les données de questionnaire – d'où son nom PointG. Aussi les sous-menus commencent-ils systématiquement par proposer les procédures graphiques.

Toutefois, ces graphiques et plus généralement les actions vont opérer sur deux types d'objets, avec, d'une part, les classiques variables d'intérêt et, d'autre part, des groupes de variables. Les variables d'intérêt sont des variables cruciales pour les problématiques sociologiques, souvent incluses dans les hypothèses. Elles servent à guider l'exploration. Les groupes de variables correspondent à des questions souvent placées côte-à-côte dans le questionnaire, d'où le nom de tranches de variables, et le recours à la stratégie dite du saucisson consistant à découper le jeu de données. Les tranches regroupent les questions portant sur un même concept, plus généralement un même thème, ou bien que l'on pense reliées sociologiquement à une variable d'intérêt. L'intérêt de cette construction est à la fois de permettre une grande vitesse d'exécution pour éviter des opérations répétitives, d'offrir une plus grande correspondance entre les objets sociologiques et les procédures statistiques, et d'inviter à l'exploration de multiples relations.

Bien que les variables catégorielles soient les plus nombreuses, les tranches regroupent souvent des variables de nature différente. Nous allons le voir sur un premier exemple qui réunit des éléments de signalétique : variables catégorielles (Sexe, Profession), ordinales (Diplôme, Revenu) et numériques (Age, Poids et Taille). Dans le logiciel PointG, les actions s'adaptent alors à ces objets composites comme dans la Figure 1, où le sous-menu : « Graphiques à plat » a été employé[10]. Les graphiques diffèrent en particulier selon la nature de la variable (et le nombre de valeurs prises). Ils sont de nature exploratoire et n'ont pas vocation à être utilisés pour communiquer les résultats[11]. Les choix de graphiques sont optimisés selon les conseils de Tufte (1983) et Cleveland (1993). Pour les variables numériques, un histogramme est réalisé, sauf si le nombre de valeurs prises est faible ($\leq$11, par exemple une échelle de Likert codée numériquement) : un graphe en bâtons est alors préféré. Le nombre d'individus étant important dans les questionnaires, les histogrammes sont stables et, dans notre exemple, ne montrent rien de surprenant sur l'âge (éligibilité à partir de 20 ans), la taille ni le poids. Pour les données ordinales, un graphe en barres respectant l'ordre est proposé. On repère, en ce qui concerne le revenu, un résultat étonnant (une surreprésentation des salaires supérieurs à 2.000 euros), et qui est lié au nombre d'homosexuels cadres parisiens dans l'échantillon. Concernant les données catégorielles, un camembert est tracé, qui vient troubler les repères habituels pour le sexe (transsexuel et intersexe ; une nouvelle variable Sexe2 marginalisera ces 13 sondés comme données manquantes). Si le nombre de catégories dépasse cinq comme pour la profession, un graphique en points est préféré où l'ordre n'est pas alphabétique (les agriculteurs ne sont premiers que dans les questionnaires) mais est réglé par l'importance des effectifs (les cadres).

La collection de graphiques invite immédiatement à se poser la question des relations entre ces informations.

Ensuite sont fournis deux versions de résumés numériques. L'une « Bref. » (voir Tableau 1) vise à définir le comportement majoritaire en proposant le mode pour les variables catégorielles (34 pour cent de cadres et 67 pour cent d'hommes) ou ordinales (32 pour cent d'interrogé-e-s dont le diplôme maximum est du secondaire hors bac), et la moyenne pour les variables numériques (âge typique de 45 ans). L'autre version plus complète, nommée « Statistiques à plat » (non présentée ici), permet de mieux saisir la diversité sociale des situations.

Parmi les tranches de variables de nature particulière, certaines sont homogènes (mêmes mesures pour toutes les variables), le plus souvent constituées d'échelles de Likert, dichotomiques ou résultant de questions à réponses multiples. Des procédures graphique et numérique spécifiques sont alors possibles.

L'étude de la signalétique est un exercice fondamental pour le sociologue. Ainsi un sous-menu dédié permet-il de réaliser une pyramide des âges, une cartographie à partir des codes postaux ou de départements (Figure 2, où l'on relève le déséquilibre spatial de l'échantillonnage), mais aussi de disposer dans un fichier spécifique des trois premiers niveaux de codage INSEE des PCS.

Enfin, destiné à l'enseignement, sachant la difficulté pour les étudiants de prendre le réflexe de se reporter à des données de cadrage, des liens directs vers des sites Internet comportant de tels renseignements (âge et sexe de la population française, PCS, consommation des ménages, sport) ont été intégrés.

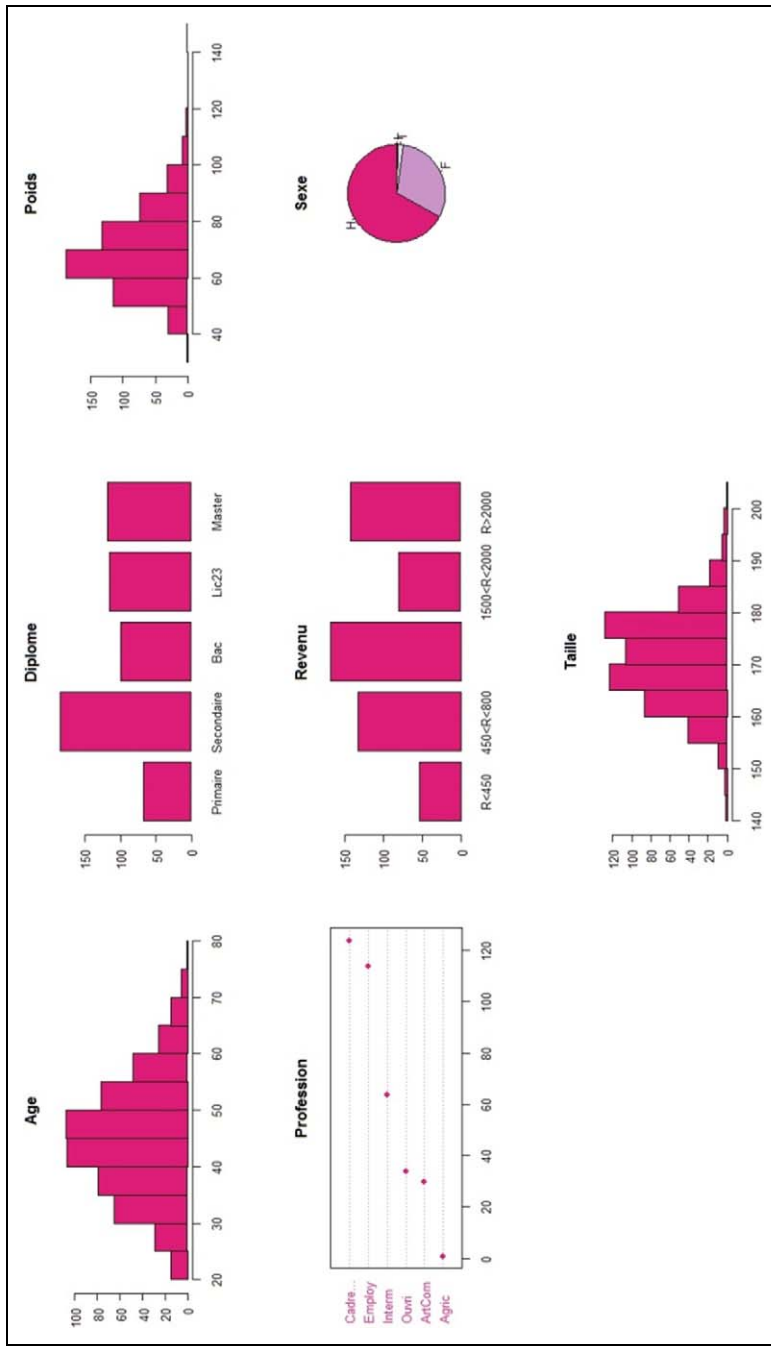

Figure 1. Résultat du sous-menu "Graphiques à plat" pour les variables de signalétique

Table 1. Résultat du résumé statistique opéré par le sous-menu "Bref" - Moyenne pour les variables numériques et catégorie à l'effectif le plus important pour les variables catégorielles et ordinales, accompagnée du pourcentage correspondant

Variable	Résumé statistique
Age	M: 45.5
Diplôme	Secondaire (%): 32
Poids	M: 70.4
Profession	Cadre... (%): 34
Revenu	800<R<1500 (%): 29
Sexe(2)	H (%): 69
Taille	M: 172.1

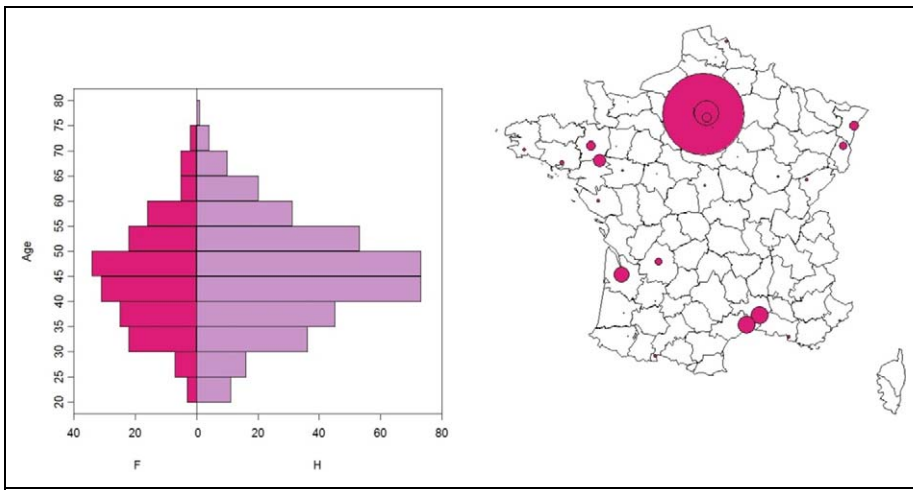

Figure 2. Pyramide des âges et cartographie des sondés

Bivarié - Dépendance vs. association et local vs. global

L'analyse d'une table croisée est bien connue du sociologue. Cela va nous permettre, dans un contexte familier, de présenter des notions parfois moins bien acquises. La première est de bien différencier, lors de l'étude de la relation entre les deux variables X et Y, s'il s'agit de se soucier de dépendance ou d'association. Prenons le cas de la relation entre le diplôme maximum obtenu par le sondé et son niveau de revenu. Le graphique typique de dépendance est le graphe en mosaïque du panneau (a) de la Figure 3. Il s'agit pour certaines valeurs du diplôme (variable X dite indépendante) de voir comment la distribution du revenu (variable Y dite dépendante) change. Pour le dire autrement, d'essayer de modéliser sous une forme $Y = f(X)$. Le graphique est donc la traduction visuelle du tableau des pourcentages en lignes (ou en colonnes). Le panneau (b) donne le graphique dit d'association[12] qui permet symétriquement de détecter comment les valeurs de X et Y sont reliées et traduit les résidus standardisés au modèle d'indépendance.

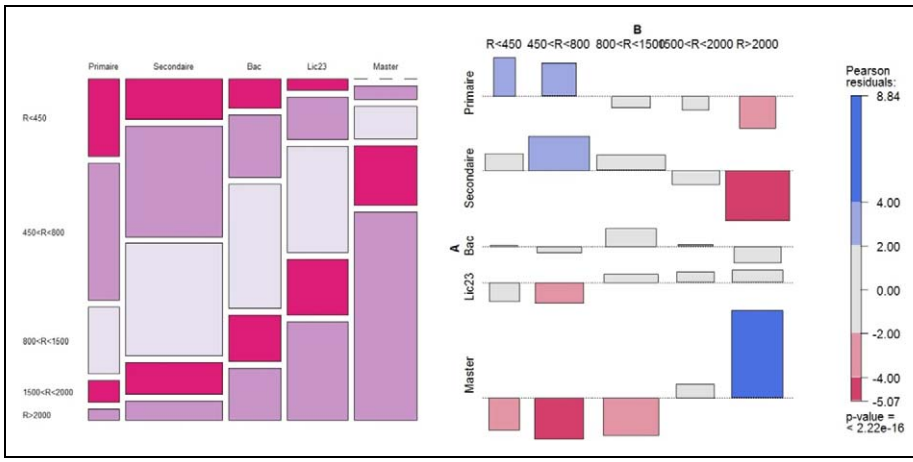

Figure 3. (a) Graphique en mosaïque et (b) graphique d'association de la relation diplôme-revenu

Les tables correspondantes des pourcentages en ligne (plutôt qu'en colonne, voir à ce sujet Ehrenberg, 1981) et des résidus de Pearson sont générées. La relation positive observée entre le niveau de revenu et le niveau de diplôme peut-elle être due au hasard ? Le test de significativité du chi-carré d'indépendance nous apprend clairement qu'il y a peu de chances que cela soit le cas, par le truchement de la p-value ($p < 0.001$).

Cependant, la valeur de la statistique X^2 (ou de p) ne nous renseigne pas directement sur l'intensité de la relation et par ailleurs est sensible à la taille de l'échantillon. C'est le rôle d'autres statistiques, dites de taille d'effet (*effect size*), qui vont devenir particulièrement intéressantes lorsque nous étudierons simultanément plusieurs relations bivariées. De nombreuses propositions existent, telles que $\phi = \sqrt{(X^2/n)}$ laquelle présente le défaut de n'être pas bornée par 1 dans le cas de tableaux croisés à plus de deux lignes et deux colonnes. Le V de Cramer corrige ce défaut mais est également critiqué sur le fait qu'atteindre 1 n'est possible qu'à certaines conditions sur les marges du tableau, ce qui conduit à d'autres méthodes dite phi/phimax ayant par ailleurs des inconvénients de robustesse (Davenport et El-Sanhurry, 1991) ; le PEM global de Cibois (1993) appartient à ce dernier type de méthodes.

Nous avons choisi d'employer ϕ dans PointG car il présente une interprétation simple en terme de pourcentage d'explication dans le cas des tableaux 2*2, et des valeurs – proposées par Cohen (1988) – aidant l'interprétation en rangeant les effets de petit ($\phi = 0.1$), moyen ($\phi = 0.3$) à grand ($\phi = 0.5$). Nous avons utilisé ces repères pour classer métaphoriquement les tailles (d'effet) en XS ($\phi \leq 0.05$), S ($0.05 < \phi \leq 0.20$), M ($0.20 < \phi \leq 0.40$), L ($0.40 < \phi \leq 0.75$) et XL ($\phi > 0.75$).

Les idées précédentes sont transposables aux situations où l'on étudie la relation :

- entre deux variables numériques, le graphique de référence étant le nuage de points, le test celui basé sur le coefficient de corrélation linéaire r de Pearson et la taille d'effet mesurée par le même r avec les mêmes classements que pour ϕ ;

- entre une variable numérique et une variable catégorielle, le graphique étant la collection de boîtes à moustaches, le test celui de l'analyse de la variance à un facteur, la taille d'effet pouvant être mesurée par le rapport de corrélation η^2 et les effets classés selon la mesure dérivée $f = \sqrt{(\eta^2/(1-\eta^2))}$ (XS si $f \leq 0.05$, S si $0.05 < f \leq 0.175$, M si $0.175 < f \leq 0.325$, L si $0.325 < f \leq 0.70$ et XL si $f > 0.70$ sur la base des repères de Cohen, 1988).

Dans le cas du croisement entre deux variables catégorielles, il est non seulement possible de s'intéresser à des tailles d'effet de la relation dans sa globalité mais plus localement à la mesure de l'attraction ou de la répulsion entre deux catégories particulières. La méthode consiste à créer un tableau 2*2 en conservant les deux modalités concernées, et en regroupant les effectifs des autres (voir par exemple Blöss et Grossetti, 1999 : 159). On peut alors mesurer la taille d'effet local par un coefficient d'association. Pour choisir ce dernier, on peut déjà utiliser les critères de Warrens (2008) demandant que cette taille d'effet soit nulle en situation d'indépendance, égale à 1 (ou -1), quelles que soient les marges, en cas de liaison parfaite. On peut aussi souhaiter une interprétation relativement intuitive, que le coefficient ne soit pas trop inconnu, et enfin qu'il soit possible de calculer des intervalles de confiance pour enrichir des représentations graphiques (employer un test du chi-carré permet de la même façon d'éviter toute surinterprétation bien que la multiplicité de ces tests ou intervalles rend le seuil théorique de significativité illusoire). Le coefficient Q de Yule que nous avons retenu satisfait à ces exigences[13].

Multivarié I - MulBivarié et taille d'effet

La première façon d'étudier les relations entre plusieurs variables simultanément est de multiplier les études bivariées. Il est possible de sélectionner une tranche de variables sur la base de leurs relations supposées avec une variable d'intérêt, sachant qu'on a alors plutôt affaire à un contexte de dépendance, ou bien d'explorer de façon systématique, et plus dans un contexte d'association, les relations internes à la tranche – une démarche dont Martin expose les faiblesses (2005 : 100).

En commençant par les approches purement graphiques, la Figure 4 montre comment la variable d'intérêt Sexe (à deux catégories) peut être reliée à une tranche de quatre variables, issues à nouveau de la signalétique. La variable d'intérêt joue ici le rôle (X) de la variable indépendante et on étudie comment les autres variables (Y) changent en fonction d'elle. A nouveau, selon la nature des variables en jeu, des graphiques adaptés sont produits. Pour les variables catégorielles, les graphiques en mosaïque indiquent que les hommes sont plus souvent titulaires d'un diplôme de master et d'un revenu supérieur à 2.000 euros. En ce qui concerne les variables numériques, les boîtes à moustaches rappellent les différences morphologiques entre les sexes : les hommes ont des poids et tailles plus élevés. Dans le cas où l'on indique la variable d'intérêt comme la variable dépendante, les graphiques subissent une rotation de 90°[14] et constituent une première étape pour une procédure de modélisation. Si seule une tranche est étudiée, le graphe par paires généralisé d'Emerson et Green (2011) est intéressant mais ne peut être employé efficacement en cas de nombre trop élevé de variables.

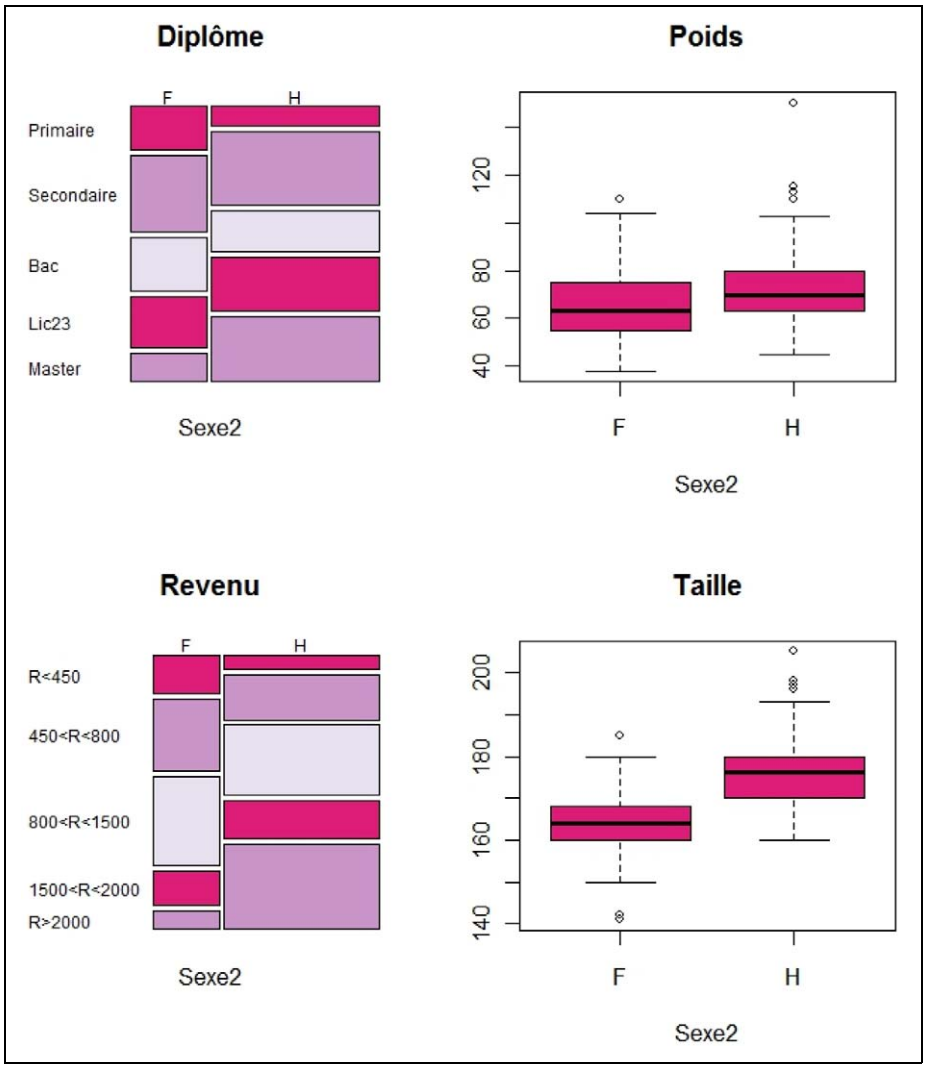

Figure 4. Graphiques croisés entre le sexe et une tranche composée du diplôme, du poids, du revenu et de la taille

Une autre approche basée sur les tests de significativité ou les tailles d'effet, dite du tamis (Blöss et Grossetti, 1999 : 159) permet de rechercher automatiquement et à plus grande échelle ce qui est intéressant (afin de l'explorer par la suite de façon graphique pour le mieux comprendre). Cette procédure peut-être globale, au niveau des variables, ou parfois au niveau de leurs catégories.

En ce qui concerne le niveau global, dans une approche avec une simple tranche, toutes les relations bivariées sont étudiées et un graphe de relations permet de signaler celles qui sont significatives (au seuil de 5 pour cent pour le test correspondant à la nature

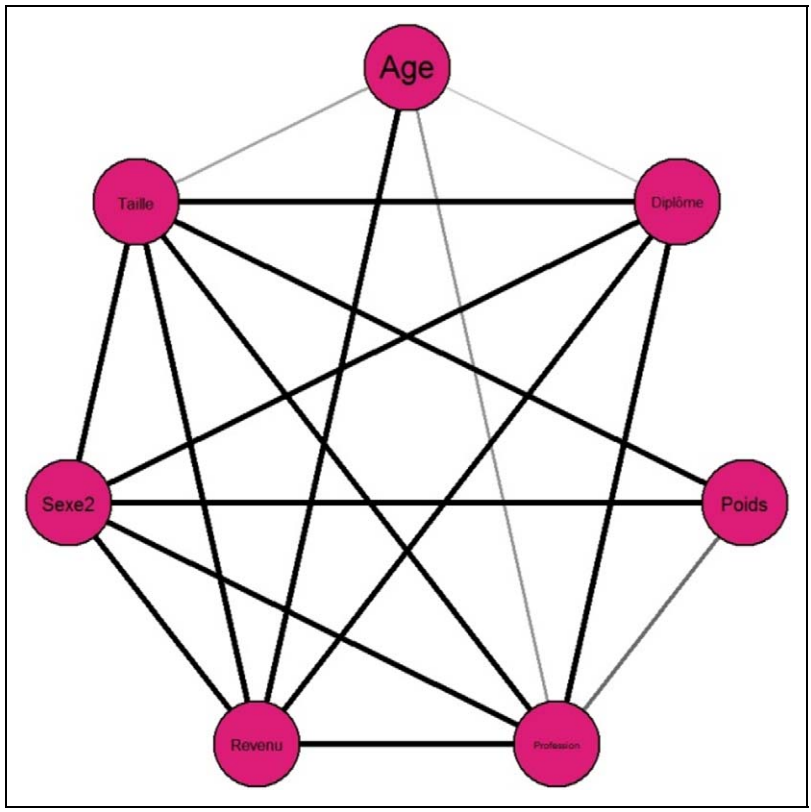

Figure 5. Graphe des relations entre les variables de signalétique. Un trait relie celles dont la relation est statistiquement significative au seuil de 5 pour cent

de la relation : test du X^2 d'indépendance, test F de l'ANOVA ou test r de corrélation linéaire) par un trait les reliant (Figure 5 avec les variables de signalétique). Le poids n'est ainsi par exemple relié qu'à la taille et au sexe (nous y reviendrons) et dans une moindre mesure à la profession (Detrez, 2002).

Dans le cas d'une étude couplant une variable d'intérêt avec une tranche, la même procédure est possible, et les tailles d'effet permettent d'établir un classement des variables les plus reliées. Sur le cas de la Figure 4, on constate alors que la taille du sondé (XL) est nettement plus reliée à son sexe que le diplôme (M), le revenu (M) ou le poids (M). Il est également possible de poursuivre l'étude au niveau local, en choisissant une catégorie (homme) de cette variable et en tentant de la relier avec les catégories des autres variables de signalétique (méthode dite du « profil de modalité », Cibois, 1993). Si certaines de ces variables sont numériques, elles sont automatiquement découpées en classes (deux par défaut : plus et moins). Les attractions non statistiquement significatives sont écartées (ainsi que toutes les répulsions[15]) et celles qui restent sont ordonnées graphiquement sur la base du Q de Yule (Figure 6). Les hommes sont caractérisés par leur plus grande taille, ils sont plus souvent Artisan-Commerçant ou

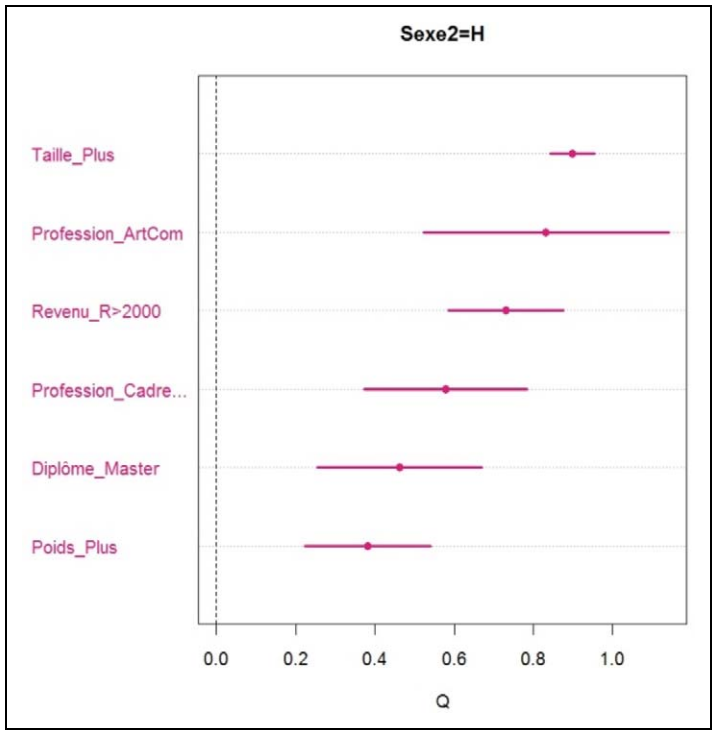

Figure 6. Graphique des Q de Yule positifs et statistiquement significatifs (et leurs intervalles de confiance) entre la catégorie Homme de la variable sexe et les catégories des autres variables de signalétique

Cadres, sont plus nombreux dans la tranche des hauts revenus, des hauts diplômés et des poids les plus élevés.

Multivarié II - MulTivarié et hétéro-statistique

L'analyse tabulaire multivariée est une approche séduisante du point de vue pédagogique, en étendant les règles familières de lecture d'un tableau croisé (et avec un vocabulaire bien spécifique à la sociologie, voir par exemple le chapitre 11 de Fox, 1999). Une traduction graphique peut en être donnée comme dans le package Vcd (Meyer et al., 2006), en cours d'insertion dans le logiciel PointG. Toutefois, le nombre de variables pouvant être ainsi traitées semble limité (en termes conceptuels comme en termes d'effectif raisonnable des cases), et nous leur préférerons des outils plus efficaces : modélisation linéaire généralisée pour la dépendance et analyse factorielle pour l'association (voire log-linéaire).

Pédagogiquement se pose une difficulté lorsqu'il s'agit d'aborder le modèle linéaire généralisé en sociologie : il est plus simple de le faire en présentant en premier lieu la régression multiple qui emploie une variable dépendante numérique. Or, ce type de variables n'est pas le plus courant (à part sous la forme un peu « dégénérée » d'une échelle de

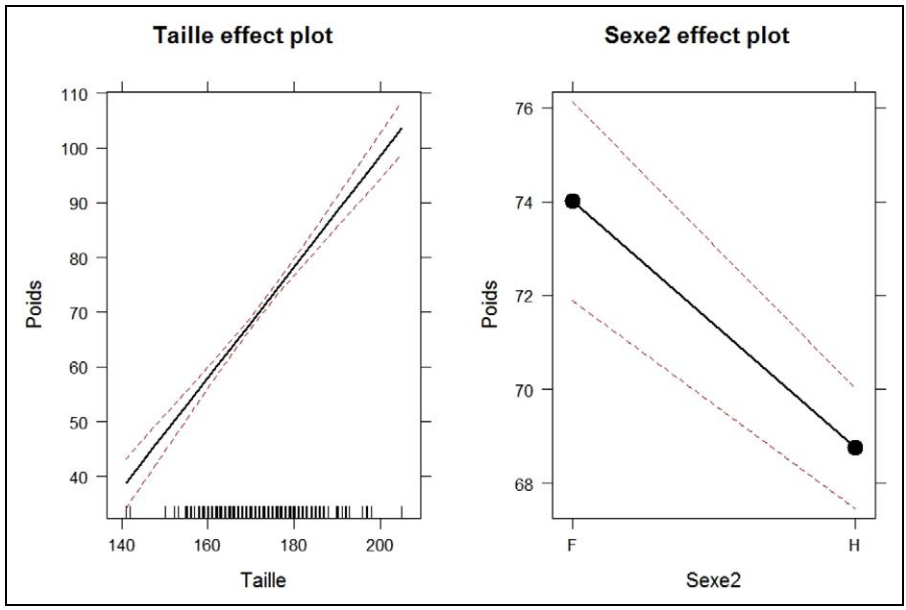

Figure 7. Graphique des effets pour la modélisation du poids par la taille et le sexe

Likert codée numériquement). Ceci dit, dans l'étude « Sport et VIH », ce problème ne se pose pas car nous pouvons étudier les effets de diverses variables – que tout étudiant connait bien – sur le poids des sondés.

Nous avons vu précédemment que le poids des individus était d'abord relié à leur taille, mais aussi à leur sexe, les hommes étant plus lourds que les femmes. La modélisation linéaire permet de décomposer ces deux effets. Dans le logiciel PointG, la procédure comprend deux étapes. Dans la première, on construit le modèle en indiquant la variable dépendante et les variables indépendantes (dont d'ailleurs l'une est numérique et l'autre catégorielle). Un tableau de la significativité statistique des effets est produit, en respectant les contraintes de marginalité (McCullagh et Nelder, 1989 : 89). On peut alors modifier le modèle (ce n'est pas la peine ici : $p(\text{Sexe}) = 0.0001$ et $p(\text{Taille}) < 0.0001$, les deux effets sont significatifs) ou passer à la seconde étape d'interprétation. Celle-ci produit la Figure 7, le graphe des effets de Fox (2003). Ce graphe donne la prédiction de la variable dépendante selon chaque variable indépendante, une fois les autres variables fixées à leur moyenne. On lit alors sur le panneau de droite que, pour un sondé de taille moyenne, les femmes sont plus lourdes que les hommes ! On voit donc que la prise en compte simultanée a conduit à une correction très importante de l'effet du sexe puisqu'il a été inversé. Ce phénomène porte le nom de paradoxe de Simpson (Huber, 2011 : 114). Il est également important de noter que ce raisonnement n'est pas totalement généralisable à toutes les situations. Par exemple, on ne peut pas vérifier si les femmes d'un mètre quatre-vingt dix, absentes de l'échantillon, sont effectivement plus lourdes que les hommes de même taille. Enfin, ce raisonnement toutes choses égales par ailleurs est parfois critiqué,

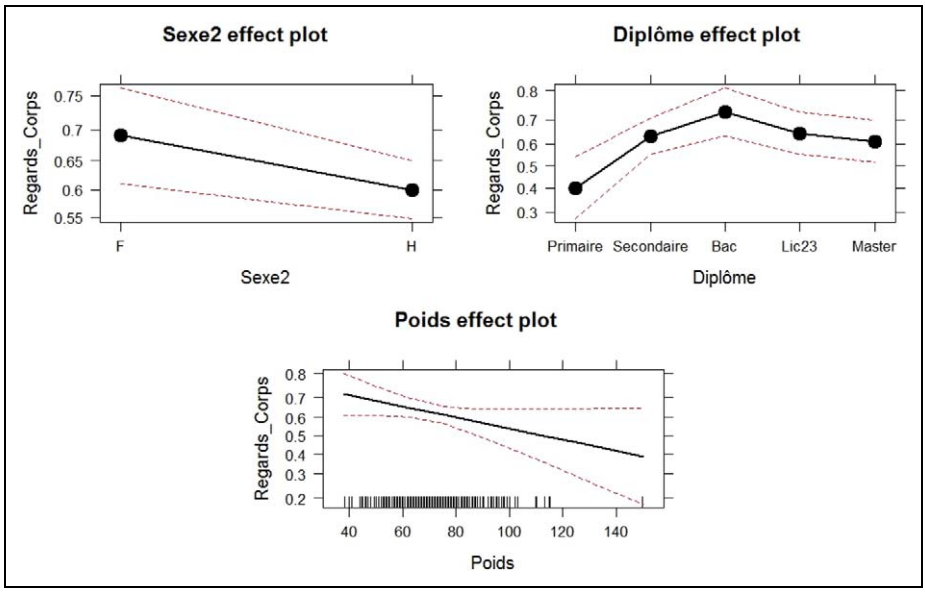

Figure 8. Graphique des effets pour la régression logistique du changement de regard sur son corps en fonction du sexe, du diplôme et du poids (en ordonné est présentée la probabilité de répondre « Oui »)

mais rien n'empêche de complexifier le modèle en y ajoutant des termes quadratiques et surtout d'interaction (Venables, 2000).

Dans le logiciel PointG, cette démarche de modélisation est étendue, quelle que soit la nature de la variable dépendante également. On peut, de façon transparente, employer une variable dichotomique, plus généralement catégorielle ou même ordonnée et un modèle linéaire généralisé correspondant sera ajusté (régression logistique binomiale, régression logistique-multinomiale, régression *proportional-odds*, voir McCullagh et Nelder, 1989, respectivement : chapitre 4 ; chapitre 6, page 151 ; et chapitre 6, page 159), une table de significativité adéquate sera produite (test du maximum de vraisemblance), ainsi qu'un graphe des effets (produisant cette fois des prédictions en termes de pourcentages). Ainsi, une modélisation du « changement de regard sur son propre corps » (Oui/Non) par le poids du sondé, sa profession et son sexe s'avère être en fait une régression logistique. On obtient la Figure 8 qui montre que les femmes ont plus changé leur regard sur leur propre corps que les hommes[16], que les sondés ayant un niveau d'étude primaire en ont bien moins changé par rapport aux autres catégories de diplôme, et que le changement de regard est d'autant plus faible que le poids du sondé est important[17].

On peut donc modéliser avec une variable dépendante et des variables indépendantes de tout type, et ce de façon invisible pour l'utilisateur et lui permettant de disposer d'outils cohérents et sophistiqués. Ici, il ne s'agit pas simplement d'adapter la procédure à la nature de l'objet ; c'est le modèle lui-même qui, par des procédures générales (maximum de vraisemblance, tests emboités, prédictions), permet de véritablement réaliser de l'hétéro-statistique.

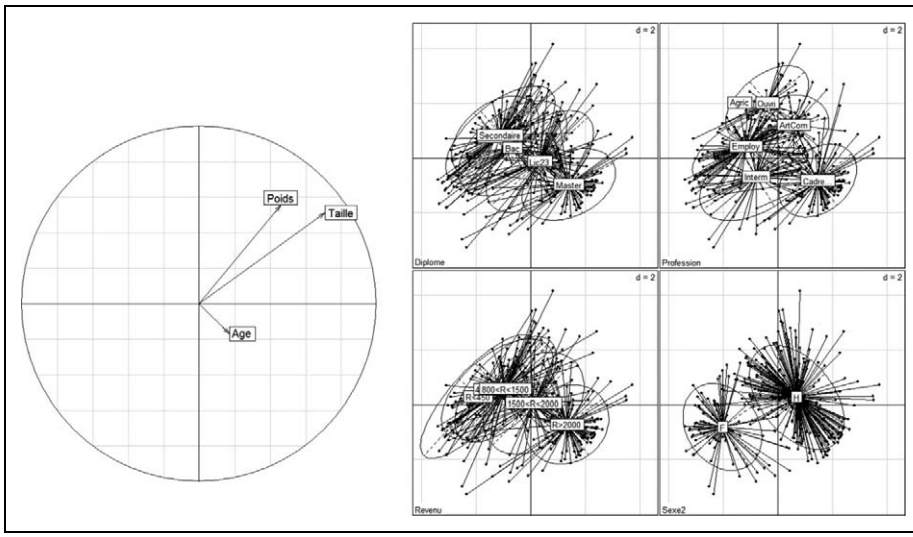

Figure 9. Analyse de Hill et Smith des variables de signalétique : (a) Cercle de corrélation pour les variables numériques ; (b) centres de gravité et ellipses de dispersion pour les modalités des variables catégorielles

Nous retrouvons cette idée dans le cas de l'analyse factorielle où, dans le logiciel PointG, une seule procédure[18] est employée : l'analyse de Hill et Smith (1976), qui permet d'utiliser conjointement variables numériques et variables catégorielles – alors qu'usuellement on choisissait, pour des raisons techniques, de recoder les variables (l'âge devenant des classes d'âge) pour les « faire entrer » dans une analyse des correspondances. L'analyse mixte revient à effectuer, si toutes les variables sont numériques, une Analyse en Composantes Principales standardisée avec pour production l'habituel cercle de corrélation ou, si elles sont toutes catégorielles, une Analyse des Correspondances Multiples entraînant une représentation visuelle de l'attraction des modalités. Dans le cas général associant les deux types, les deux éléments graphiques sont simultanément générés. Il est par ailleurs possible d'ajouter des variables supplémentaires, de toute nature également, pour enrichir l'analyse. En reprenant les variables de signalétique comme variables actives, l'analyse de Hill et Smith des données « Sport et VIH » construit deux axes intéressants et produit la Figure 9 en deux parties. On retrouve beaucoup des observations déjà faites : la relation entre sexe et morphologie, la forte relation entre diplôme et revenu – et, dans une moindre mesure, catégories socioprofessionnelles – l'isolement relatif de l'âge qui ne paraît pas lié avec ces autres variables.

Conclusion

Ce qui rend possible et évolutif le logiciel PointG est l'environnement de programmation R. Il permet de bénéficier d'algorithmes de qualité, de remarquables possibilités graphiques et d'une extraordinaire quantité de bibliothèques (*packages*) de programmes écrits par d'autres statisticiens[19], le tout disponible sur n'importe quel système d'exploitation.

La structure de construction d'un package permet aussi de générer très facilement des aides en ligne, et d'intégrer des documents divers dans le logiciel (comme un mode d'emploi dans PointG). Depuis la création de l'extension R-Commander, (Fox et al., 2011), il devient plus facile de créer des menus déroulants et par conséquent de rendre les choses plus conviviales pour l'utilisateur néophyte. Divers forums permettent une interactivité avec les utilisateurs ou les autres programmeurs, et ainsi de connaître les erreurs de programmation, les manques, et d'y remédier (souvent avec leur aide). Enfin, la libre distribution et l'esprit communautaire s'accordent avec la communication du travail universitaire.

Quelles sont les limites et les perspectives actuelles du logiciel PointG ? En premier lieu, nous en sommes à un stade expérimental (pour tester, télécharger RcmdrPlugin. pointG version 0.2.1 depuis http://CRAN.R-project.org/, Champely, 2012) où quelques options pour modifier les choix par défaut (de graphiques, de seuil de significativité) seront bienvenues. De façon nette, les opérations de codages ne sont pas satisfaisantes (ce sont pour l'heure celles assez générales du R-Commander). Or, d'une part, le logiciel repose en grande partie sur la correcte identification des variables en catégorielle, ordinale et numérique (C/O/N), voire sur le « jeu » dans le cadre de cette codification, pour les variables ordinales en particulier. D'autre part, le cours de l'analyse d'un questionnaire n'est pas aussi linéaire que notre présentation le laisse supposer. Il y a de fréquents retours en arrière pour corriger des erreurs mais surtout pour opérer des recodages (C/O/N), des transformations, des regroupements de catégories, pour traiter les données manquantes . . . Il convient donc de mener une réflexion à ce sujet et de l'envisager sous un angle interactif et dynamique.

De la même façon, les procédures factorielles, bien que puissantes, manquent nettement d'interactivité (retirer une variable de l'analyse, puis l'ajouter en supplémentaire, visualiser les variables les plus contributives, détecter celles qui sont reliées à une variable, voire une catégorie spécifique, identifier un point). On trouvera à ce sujet des sources d'inspiration dans Cook et Swayne (2007).

La difficile question de l'utilisation des pondérations reste également posée. Quelques améliorations pour les échelles de Likert en particulier (traitement des attentes, de la satisfaction de membres, d'usagers ou de clients) sont à l'ordre du jour, mais il ne s'agit pas d'aller jusqu'aux équations structurelles comme dans d'autres disciplines. Si bien des extensions sont en effet imaginables (modèle log-linéaire, positionnement multidimensionnel . . .), le propos de PointG est de conserver sa simplicité pour un utilisateur débutant. Lorsque celui-ci devient plus aguerri, il peut alors passer aux autres menus du R-Commander et par exemple calculer immédiatement des intervalles de confiance pour des *odds-ratio*, mieux tester l'adéquation de ses modélisations . . . Pour les plus attirés par les approches quantitatives, l'immersion dans l'environnement R sera alors envisageable, promettant une liberté solidement structurée et des possibilités incroyables il y a seulement une dizaine d'années, et dont l'évolution est difficilement prédictible. Du point de vue informatique et de R, l'utilisation de l'objet « tranche » (un sous-tableau finalement) et de la notion d'hétéro-statistique (un sous-tableau composite) mérite approfondissement avant d'envisager par exemple la combinaison de plusieurs tranches comme peut le faire par exemple une analyse factorielle multiple (Escoffier et Pagès, 1994, programmé et amélioré dans le package FactomineR, Husson

et al., 2011) ou des régressions emboîtées. In fine, il s'agira également de proposer la possibilité de réaliser des typologies (Nakache et Confais, 2004), notamment des classifications mixtes à articuler avec des régressions (Des Nétumières, 1997).

Notes

1. On n'abordera pas ici le problème des questions ouvertes et la statistique textuelle.
2. *statistical methods (mostly graphical) geared toward providing insight rather than quantifiable results*
3. Qui est parfois exhaustif comme il peut l'être dans une étude sur une (petite) organisation sportive, et parfois clairement non représentatif (voir notre exemple).
4. Pour obtenir le 21/12/2011 comme liste des dix premiers ouvrages : (1, 3, 8) *Statistics for People Who (Think They) Hate Statistics* (Salkind et Francis, 2011) ; (2) *How to Lie with Statistics* (Huff et Geis, 1993) ; (4) *The Foundations of Statistics* (Savage, 1972) ; (5) *The Tao of Statistics: A Path to Understanding (With No Math)* (Keller, 2005) ; (6) *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activists* (Best, 2001) ; (7) *Simple Statistics: Applications in Criminology and Criminal Justice* (Miethe, 2006) ; (9) *Reading & Understanding Multivariate Statistics* (Grimm et al., 1995) ; (10) *Statistics without Tears : A Primer for Non-mathematicians* (Rowntree, 2003).
5. Et plus précisément en version commerciale : Modalisa, Ethnos, Sphinx (Ganassali, 2007) et en version gratuite R (Barnier, 2009 ; Falissard, 2012) ou TriDeux (Cibois, http://cibois.pagesperso-orange.fr/Trideux.html).
6. Cette enquête nationale a été réalisée avec le soutien financier de Sidaction et de la région Île-de-France (2009-2011). Pour plus de détails sur ce travail, voir Ferez et Lomo (sous presses).
7. Alors que l'environnement **R** (R Development Core Team, 2011) dans lequel il est inclus est initialement un logiciel à langage de commandes ; c'est-à-dire, que des lignes de programmes sont tapées au clavier puis interprétées.
8. Nous reviendrons sur le déroulement pratique dans la discussion.
9. Un menu déroulant dédié à cette partie fondamentale pour l'analyse a été distingué (voir plus loin).
10. Les figures présentées sont les sorties graphiques directement issues du logiciel (sans retouches).
11. A cette fin, on trouvera dans le logiciel **R**, des programmes remarquables : *Grid* (Murrell, 2002), *Lattice* (Deepayan, 2008), et *Ggplot2* (Wickham, 2009).
12. Le sous-menu « Table croisée » produit également le plus classique graphe en barres juxtaposées et le premier plan factoriel de l'Analyse Factorielle des Correspondances (AFS).
13. On peut noter en passant que le PEM local de Cibois (1993) est exactement le coefficient phi/phimax de la table 2*2 associée.
14. Et ils sont modifiés en conséquence.
15. De façon optionnelle, les répulsions sont disponibles car les sous-représentations sont parfois sociologiquement instructives ; par exemple, sur le plan de l'analyse des consommations, des pratiques et (dé)goûts.
16. L'effet du genre est tout à fait sensible sur l'ensemble des questions portant sur le rapport au corps.

17. On sait que les traitements antirétroviraux ont tendance à faire fortement perdre du poids aux PVVIH. Or la maigreur est très souvent associée, dans les imaginaires, à la mort – les PVVIH ne faisant pas exception, bien au contraire.
18. Empruntée au package Ade4 (Chessel et al., 2004)
19. Ainsi *PointG* utilise des éléments de *Ade4* (Chessel et al., 2004), *Car* (Fox et Weisberg, 2011), *Effects* (Fox, 2003), *Hmisc* (Harell et al., 2011), *Lattice* (Deepayan, 2008), *MASS* (Venables et Ripley, 2002), *Maps* (Becker et al., 2011) , *Qgraph* (Epskamp et al., 2011), *RColorBrewer* (Neuwirth, 2011), *VIM* (Temple et al., 2011) et *YaleToolkit* (Emerson et Green, 2010).

Bibliographie

Barnier J (2009) *R pour les Sociologues et Assimilés*. Disponible à : http://cran.rproject.org/.

Becker RA, Wilks AR, Brownrigg R et Minka TP (2011) *Maps : Draw Geographical Maps, version 2.2-2*. Disponible à : http://CRAN.R-project.org/.

Blöss T et Grossetti M (1999) *Introduction aux méthodes statistiques en sociologie. Paris* : Presses Universitaires de France.

Champely S (2012) *RcmdrPlugin.pointG: Rcmdr Graphical Point of View for Questionnaire Data Plug-In, version 0.2.1*. Disponible à : http://CRAN.R-project.org/.

Chessel D, Dufour AB et Thioulouse J (2004) The Ade4 Package-I- One Table Methods. *R News* 4 : 5-10.

Cibois P (1993) Le PEM, pourcentage de l'écart maximum - Un indice de liaison entre modalités d'un tableau de contingence. *Bulletin de Méthodologie Sociologique* 40 : 43-63.

Cibois P (2009) *Les Méthodes d'Analyse d'Enquêtes*. Paris : PUF, Que sais-je, n. 3782 (épuisé), version corrigée disponible à http://cibois.pagesperso-orange.fr/.

Cleveland W (1993) *Visualizing Data*. Summit NJ : Hobart Press.

Chambers JM (2010) *Software for Data Analysis*. New-York: Springer.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (Second Edition). Hillsdale, NJ : Lawrence Erlbaum.

Cook D et Swayne DF (2007) *Interactive and Dynamics Graphics for Data Analysis*. New-York: Springer.

Davenport EC et El-Sanhurry NA (1991) Phi/phimax: Review and Synthesis. *Educational and Psychological Measurement* 51 : 821-28.

Deepayan S (2008) *Lattice: Multivariate Data Visualization with R*. New York : Springer.

Des Nétumières F (1997) Méthodes de régression et analyse factorielle. *Histoire et mesure* 12 : 271-97.

Detrez C (2002) *La construction sociale du corps*. Paris : Seuil.

Ehrenberg ASC (1981) Rudiments of Numeracy. *American Statistician* 35 : 67-71.

Emerson JW et Green WA (2010) *YaleToolkit: Data Exploration Tools from Yale University, version 3.2*. Disponible à : http://CRAN.R-project.org/.

Emerson JW et Green WA (2011) *Gpairs: The Generalized Pairs Plot, version 1.0*. Disponible à : http://CRAN.R-project.org.

Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD et Borsboom D (2011) *Qgraph: Network Representations of Relationships in Data, version 0.5.3*. Disponible à : http://CRAN.R-project.org/.

Escoffier B et Pagès J (1994) Multiple Factor Analysis (AFMULT package). *Computational Statistics and Data Analysis* 18 : 121-40.

Falissard B (2012) *Analysis of Questionnaire Data with R*. Boca Raton, FL : CRC Press.
Fox J (2003) Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software* 8(15) : 1-27.
Fox J et Weisberg S (2011) *An {R} Companion to Applied Regression, Second Edition*. Thousand Oaks CA: Sage.
Fox J et al. (2011) *Rcmdr: R Commander, version 1.7-0*. Disponible à : http://CRAN.R-project.org/.
Fox W (1999) *Statistiques Sociales*. Paris : DeBoeck Université.
Ferez S et Lomo A (eds) (sous presses). *Sport et VIH. Un corps sous contrainte médicale*. Paris : Téraèdre.
Ganassali S (2007) *Les enquêtes par questionnaire avec Sphinx*. Paris : Pearson Education France.
Harrell JR FE, et al. (2011) *Hmisc: Harrell Miscellaneous, version 3.9-0*. Disponible à http://CRAN.R-project.org/.
Hill MO et Smith AJE (1976) Principal Component Analysis of Taxonomic Data with Multi-state Discrete Characters. *Taxon* 25 : 249-55.
Huber PJ (2011) *Data Analysis. What Can be Learned from the Past 50 Years*. Hoboken, NJ: Wiley.
Husson F, Josse J, Le S et Mazet J (2011) *FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R, version 1.16*. Disponible à : http://CRAN.R-project.org/.
Martin O (2005) *L'analyse de données quantitatives (L'enquête et ses méthodes)*. Paris: Armand Colin.
McCullagh P et Nelder JA (1989) *Generalized Linear Models*. London : Chapman and Hall.
Meyer D, Zeileis D et Hornik K (2006) The Strucplot Framework: Visualizing Multi-way Contingency Tables with Vcd. *Journal of Statistical Software* 17(3) : 1-48.
Murrell P (2002) The Grid Graphics Package. *R News* 2 : 14-19.
Nakache JP et Confias J (2004) *Approche pragmatique de la classification - Arbres hiérarchiques et partitionnements*. Paris : Technip.
Neuwirth E (2011) *RColorBrewer: ColorBrewer Palettes, version 1.0-5*. Disponible à : http://CRAN.R-project.org/.
R Development Core Team (2011) *R - A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Disponible à http://www.R-project.org/.
Selz M et Maillochon F (2009) *Le raisonnement statistique en sociologie*. Paris : Presses Universitaires de France.
Temple M, Alfons A et Kowarik A (2011) *VIM - Visualization and Imputation of Missing Values, version 2.0.4*. Disponible à : http://CRAN.Rproject.org/.
Tufte ER (1983) *The Visual Display of Quantitative Information*. Cheshire, CT : Graphics Press.
Venables WN (2000) Exegeses on Linear Models. In *S-Plus User's Conference, Washington DC, USA, 8-9 October 1998*. Disponible à http://www.stats.ox.ac.uk/pub/MASS3/Exegeses.pdf.
Venables WN et Ripley BD (2002) *Modern Applied Statistics with S (Fourth Edition)*. New York : Springer.
Warrens MJ (2008) On Association Coefficients for 2*2 Tables and Properties that Do Not Depend on the Marginal Distributions. *Psychometrika* 73 : 777-89.
Wheaton B (2003) Quand les méthodes font toute la différence. *Sociologie et Société* 35 :19-48.
Wickham H (2009) *Ggplot2: Elegant Graphics for Data Analysis*. New York : Springer.