VERS UNE MODÉLISATION DU RAISONNEMENT QUALITATIF EN STATISTIQUES

Amel BORGI
e-mail : borgi@laforia.ibp.fr

Résumé :

Nous tentons dans cet rapport de mieux refléter le raisonnement qualitatif humain lors de l'analyse de données statistiques, et ce à partir d'un exemple. Nous présentons pour cela des éléments d'un modèle logique et cognitif du raisonnement qualitatif dans le domaine de la statistique descriptive élémentaire, le modèle concerne les experts et les novices en statistiques. Pour modéliser la description qualitative de données ou de paramètres statistiques nous introduisons la notion d'*appréciation qualitative*, pour leur comparaison, nous définissons la notion d'*écart* entre deux données. Ces deux points se basent sur la gradualité et la logique multi-valente.

Le modèle proposé aborde également la notion de résumé quantitatif d'une série statistique et les procédures de calcul associées. Nous présentons, dans ce cadre, une méthode de calcul approché de la moyenne arithmétique par les novices.

Mots clés :

Raisonnement qualitatif, modélisation logique et cognitive, statistique descriptive élémentaire, gradualité, calcul approximatif, experts et novices en statistiques.

Abstract :

This paper is devoted to the description, through an exemple, of a model for human qualitative reasoning in statistical data analysis. Some elements of a logical and cognitive model of qualitative reasoning in the field of elementary descriptive statitics are presented. This model concerns the experts and the novices in statistics. In order to modelize the qualitative description of statistical data or parameters, the notion of *qualitative appreciation* is introduced. In order to compare them, the notion of *gap* between two data is defined. These two notions are based on graduality and many-valued logic.

The model thus proposed allows us to tackle the notion of quantitative abstract of a statistical serie and the associated computation processes. Within this framework, a method for the approximate calculation of the arithmetic average by the novices is given.

Keywords :

Qualitative reasoning, logical and cognitive modelization, elementary decriptive statistic, graduality, approximate calculus, experts and novices in statistics.

VERS UNE MODÉLISATION DU RAISONNEMENT QUALITATIF EN STATISTIQUES

TABLE DES MATIÈRES

1. Introduction 3
2. Différentes approches du raisonnement qualitatif 4
3. Les bases du modèle logique et cognitif 5
    3.1. Les données statistiques 5
    3.2. Discrétisation du domaine de la variable statistique 6
    3.3. La notion d'écart entre deux données statistiques 7
4. Appréciation qualitative d'une variable note 9
    4.1. Le raisonnement des novices en statistiques 10
    4.2. Le raisonnement des experts en statistiques 10
5. Comparaison qualitative de données statistiques 11
6. Traitement de la moyenne arithmétique par les novices en statistiques 12
    6.1. Transformation des données statistiques 12
    6.2. Calcul approché de la moyenne 12
    6.3. Appréciation et comparaison qualitatives de la moyenne 14
7. Conclusion 15
Bibliographie 17

1. INTRODUCTION

Un des objectifs des sciences cognitives, et de l'intelligence artificielle en particulier, est de décrire, comprendre, modéliser et simuler les processus intelligents et cognitifs. Le raisonnement qualitatif est un exemple de ces processus et constitue un élément incontournable de l'activité cognitive humaine : l'homme appréhende et réagit à de multiples situations de façon très performante, non pas en résolvant des équations différentielles, mais en raisonnant de manière qualitative [Dormoy et al. 89].

De nombreux travaux ont étudié ce mode de raisonnement, en général dans le cadre de la description et de la simulation de phénomènes physiques [Haton et al.91], mais également dans l'objectif de formaliser et de modéliser le raisonnement qualitatif [Dormoy 89]. Notre travail se situe dans cette dernière optique et concerne le domaine particulier des statistiques.

Le raisonnement qualitatif est souvent utilisé en statistiques surtout par les novices mais également par les experts. Ce domaine se prête bien à ce type de raisonnement, car la notion d'ordre de grandeur y est importante et les analyses des séries statistiques et de leurs paramètres relèvent souvent du qualitatif. Nous nous sommes limités dans cet article à la statistique descriptive élémentaire. Rappelons qu'elle correspond à la présentation et au résumé des données et n'a aucune visée inductive.

Le travail réalisé a un double objectif. Le premier est de modéliser le raisonnement statistique naïf : les enfants ou les adultes sans formation scientifique peuvent commenter des données statistiques, les analyser et aboutir à des conclusions satisfaisantes sans pour autant maîtriser les outils statistiques ou faire des calculs exacts [Corroyer 92]. Ils raisonnent sur des *modèles mentaux de nature qualitative*. Le second objectif est d'étudier le raisonnement des experts en statistiques : eux aussi recourent à des méthodes qualitatives pour l'appréciation des paramètres statistiques ou l'analyse des données. Cependant ils n'utilisent en général ce raisonnement que dans un deuxième temps : après avoir fait le calcul exact des paramètres à considérer et utiliser les différentes méthodes statistiques dont ils disposent.

Nous avons utilisé comme support de la modélisation les données et résultats d'une étude expérimentale en psychologie cognitive menée par D. Corroyer [Corroyer 92] : on présente un ensemble de données numériques à analyser, à des individus n'ayant jamais suivi d'enseignement de statistiques. Cette analyse ne requiert aucune notion de probabilité, il s'agit de la description et de la comparaison de deux séries numériques.

Les bases du modèle proposé sont construites à partir de ces données. Nous commençons par discrétiser le domaine de la variable statistique étudiée en un nombre fini d'intervalles. Ce découpage permettra de définir la notion de *même ordre de grandeur* et d'*apprécier qualitativement* les données ou paramètres statistiques. Ce dernier point est basé sur la gradualité et la logique multi-valente [Akdag et al. 92], [Burattini et al. 92]. Pour la

comparaison des protocoles numériques, nous définissons la notion d'*écart* et établissons un lien entre cette notion et l'implication en logique multi-valente.

Ces concepts permettent de modéliser la description et la comparaison qualitatives de données statistiques ou de paramètres de position (la moyenne arithmétique, le mode et la médiane).

Pour modéliser le calcul de la moyenne par les novices, nous présentons une méthode de calcul approché de ce paramètre. Cette méthode est inspirée des travaux décrits dans [Missier et al. 89] relatifs à la structure d'Algèbre Qualitative et au calcul qualitatif.

Nous montrerons, en conclusion, l'interdisciplinarité du travail réalisé, ses applications et quelques évolutions possibles.

2. DIFFÉRENTES APPROCHES DU RAISONNEMENT QUALITATIF

Le raisonnement qualitatif, encore appelé physique qualitative, est né des préoccupations de Pat Hayes [Hayes 79], qui considère qu'une machine intelligente doit posséder la notion de "sens commun" et un modèle du monde qui nous entoure. Depuis, la physique qualitative s'est élargie et permet aujourd'hui d'étudier des systèmes physiques, de prédire leurs états futurs et d'expliquer leurs comportements en se fondant sur leur description qualitative [Fouché et al. 89]. L'intérêt de ce mode de raisonnement est de modéliser de tels systèmes en l'absence de données quantitatives mais également de réaliser des simulations : une simulation qualitative utilisant une faible quantité de données peut fournir de bons résultats rapidement et peut "remplacer" une simulation numérique parfois impossible à cause de la complexité des équations ou d'une connaissance non exacte des phénomènes étudiés [Haton 91], [Williams et al. 91].

Depuis les années quatre vingt, un nouvel intérêt pour les techniques et méthodes qualitatives est apparu dans de nombreux domaines tels que l'automatique, la chimie ou l'intelligence artificielle. De nouveaux cadres ont donc été développés pour prendre en compte d'autres idées intuitives de ce qui peut être rangé sous le terme qualitatif, comme par exemple les *ordres de grandeur*. Ainsi de nombreux travaux étudient l'aspect formel du raisonnement qualitatif. Les différentes approches utilisent une représentation discrète de l'espace continu des valeurs possibles que peut prendre une quantité ou une grandeur. L'un des premiers modèles étudié est celui basé sur les signes, il est utilisé par les économistes depuis plus de quarante ans [Dormoy 89]. D'autres modèles comme celui basé sur les ordres de grandeur [Raiman 91] ou les algèbres d'intervalles [Kuipers 86] ont ensuite été proposés. Ces différents exemples de modèles d'inspiration algébrique permettent de réaliser des calculs qualitatifs grâce à une règle de résolution.

Une direction de recherche future semble être l'unification de ces différents modèles en une structure axiomatique qui capture la notion intuitive de ce qu'est "le qualitatif". Des travaux réalisés en ce sens sont décrits dans [Missier et al. 89] et [Piera et al. 92].

Notre travail s'inscrit dans le cadre de la modélisation du raisonnement qualitatif et tente de mieux refléter le raisonnement humain lors de la description qualitative de données statistiques. Nous nous sommes inspirés pour cela des notions d'Algèbres Qualitatives et d'ordre de grandeur [Missier et al. 89], [Raiman 91], ainsi que de la gradualité et de la logique multi-valente [Akdag et al. 92], [Burattini et al. 92].

3. LES BASES DU MODÈLE LOGIQUE ET COGNITIF

Pour modéliser le raisonnement qualitatif des novices en statistiques, nous nous sommes basés sur une étude expérimentale en psychologie cognitive décrite dans [Corroyer 92]. Denis Corroyer a étudié une population d'élèves de classes de Sixième, de Seconde et des adultes. Les sujets avaient à comparer deux ensembles de 10 nombres : ces nombres étaient écrits sur de petits cartons de deux couleurs différentes selon l'ensemble auquel ils appartenaient et étaient présentés en vrac de telle manière que les sujets pouvaient les disposer comme ils le souhaitaient. Trois contextes ont été définis : le "*contexte des nombres*" dans lequel les sujets ont été amenés depuis l'école primaire à effectuer des opérations sans qu'elles aient nécessairement de signification, le "*contexte de la biologie*" où les valeurs représentent des résultats d'observations, enfin le "*contexte des notes*". Nous nous sommes intéressés au cas où ces nombres correspondent à des notes (variant de 0 à 20). La consigne données aux sujets étaient : "*Voici les notes des élèves de deux classes qui ont eu le même examen. Je voudrais que tu compares les notes de ces deux classes, que tu me dises ce que tu en penses. Je voudrais aussi que tu me dises tout ce que tu fais pour les comparer*".

La manière dont les sujets disposaient les nombres et toutes leur verbalisations étaient enregistrées. Les bases du modèle proposé sont construites à partir de ces données et de ces verbalisations.

3.1. Les données statistiques

Nous notons les données dont disposent les sujets de la manière suivante : série 1 *brute* : $\{x_1, x_2, ..., x_n\}$, série 2 *brute* : $\{x'_1, x'_2, ..., x'_n\}$, où n correspond à l'effectif total. Chacune de ces deux séries correspond donc aux notes des élèves de chacune des deux classes. Dans notre exemple, n=10. Nous rappelons qu'une série brute est constituée des valeurs prises par une variable statistique sur l'ensemble des individus étudiés [Droesbeke 92].

D'après les tests, les sujets passent en général par la première étape d'une analyse de données qui consiste à présenter ces dernières sous une forme commode, pour en déduire certaines impressions générales [Corroyer 92]. Ils commencent par ordonner chacune des deux séries de notes, cela est possible car la variable statistique considérée ici est une variable quantitative. Ensuite, dans la majorité des cas (54% dans le contexte des notes), les sujets présentent les deux séries sous forme de distribution observée [Droesbeke 92] : ils ordonnent les notes et regroupent celles qui sont identiques, pour chacune des deux séries.

Une distribution observée est définie par les valeurs distinctes qui apparaissent dans la série brute, et par l'effectif associé à chacune d'elles.

Ainsi, la distribution observée de la série 1, que l'on notera D.O. est définie par l'ensemble des couples suivants **{(X_j, n_j), j=1,2,...,J}**. X_1, X_2,..., X_J, sont les valeurs distinctes qui apparaissent dans la série brute, J est le nombre de ces valeurs distinctes, et n_j est l' effectif associé à X_j. Pour la série 2, on obtient de même une distribution observée notée D.O.' définie par l'ensemble des couples suivant : **{(X'_j, n'_j), j=1,2,...,J'}**.

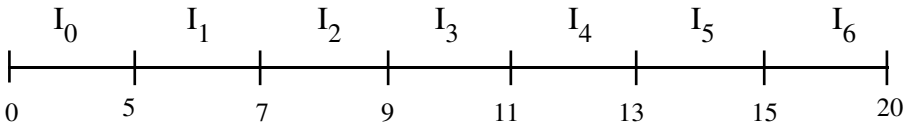
On travaillera désormais sur les distributions observées, cela est justifié par le comportement cognitif des sujets qui manipulent les données initiales et, en général, transforment les séries brutes en distributions observées ou du moins ordonnées. Pour l'étude du raisonnement qualitatif des experts en statistiques nous considérerons également les séries observées.

Après avoir présenté les données statistiques sous une forme commode, les sujets les analysent en deux étapes. La première est celle de la description des données : les sujets émettent des appréciations pour chacune des deux séries numériques. La deuxième étape est celle de la comparaison des deux groupes de notes. Pour cela, les sujets utilisent des paramètres de position et de dispersion sans forcément maîtriser ces notions ou même connaître le terme consacré au paramètre étudié [Corroyer 92]. Nous nous limiterons dans cette étude, aussi bien pour le cas des novices que des experts en statistiques, aux paramètres de position.

3.2. Discrétisation du domaine de la variable statistique

Les novices et les experts en statistiques déduisent des impressions générales à partir des données et font des analyses qualitatives en se basant sur les ordres de grandeur des valeurs observées. C'est la raison pour laquelle nous découpons le domaine de la variable statistique considérée en intervalles significatifs pour le problème.

Dans notre exemple, le domaine est, pour chacune des deux séries, l'intervalle : [0,20]. Par ailleurs, la variable considérée, qui est la même pour les deux séries, correspond à une note, d'où le choix suivant d'intervalles qui correspondent aux *appréciations des notes* :



I_0 = [0,5] correspond aux *très mauvaises* notes,
I_1 =]5,7] aux *mauvaises* notes,
I_2 =]7,9] aux notes *médiocres*,
I_3 =]9,11] aux notes *moyennes*,
I_4 =]11,13] aux notes *correctes*,
I_5 =]13,15] aux *bonnes* notes, et
I_6 =]15,20] aux *excellentes* notes.

On note ***I*** l'ensemble de ces intervalles et **c** son cardinal. On munit ***I*** d'une relation d'ordre total : pour tout i, $j \in \{0,1,...,c-1\}$, $I_i \leq I_j$ si et seulement si $i \leq j$. ***I*** muni des opérateurs min et max a une structure de treillis.

Selon le domaine, le sujet d'étude et la précision requise, on choisira un nombre plus ou moins élevé d'intervalles. Le choix des bornes des intervalles dépend d'une part du domaine considéré, afin d'obtenir un découpage significatif de ce domaine, et d'autre part du raisonnement des élèves et des experts. Un axe de recherche serait d'étudier de près les observations, réactions et analyses de ceux-ci afin d'en extraire les intervalles à considérer et leurs bornes.

Nous insistons sur le fait que le travail que nous exposons par la suite reste pertinent quel que soit le nombre d'intervalles et le choix de leurs bornes. La seule condition requise pour le bon fonctionnement du modèle est que l'ensemble ***I*** de ces intervalles soit muni d'une *relation d'ordre total.* Le chevauchement des intervalles reste en particulier possible, il permettrait une plus grande souplesse dans l'interprétation des notes en termes qualitatifs. L'ordre total est alors défini par : $[a,b] \leq [c,d]$ si et seulement si $a \leq c$ et $b \leq d$.
Les intervalles considérés peuvent être fermés, ouverts ou semi-ouverts.

Dans notre cas, nous avons considéré 7 intervalles (échelle de 0 à 6) ; cela est assez fin pour le domaine considéré [0,20] et pour la signification des valeurs observées : une répartition des notes sur ces 7 intervalles reflète correctement le niveau de l'élève (de façon qualitative bien sûr). Ce choix reste cependant subjectif.

La notion de même ordre de grandeur

Deux valeurs de même unité que les valeurs observées seront du même ordre de grandeur si et seulement si elles appartiennent au même intervalle I_k, $k \in \{0,1,...,c-1\}$, cette définition découle naturellement du découpage du domaine de la variable statistique en un nombre fini d'intervalles. L'hypothèse cognitive sous-jacente est que les sujets apprécient de la même manière deux notes appartenant au même intervalle I_k.

$x, y \in [0,20]$ $\quad x \in I_j \quad j \in \{0,1,...,c-1\}$

$\quad y \in I_k \quad k \in \{0,1,...,c-1\}$

x MOG y ssi j = k

x MOG y : x et y sont de *Même Ordre de Grandeur*.

3.3. La notion d'écart entre deux données statistiques

Nous désirons modéliser la comparaison qualitative de deux observations ou de deux paramètres de position s'exprimant dans la même unité qu'une observation (i.e. en *note*) comme par exemple les moyennes des deux séries numériques.

Notons x et y les deux valeurs à comparer, on a alors x et y dans [0,20], donc il existe j et k dans {0,1,...,c-1} tels que : $x \in I_j$ et $y \in I_k$

On s'intéresse à la *position relative* des deux paramètres, c'est pour cela que nous introduisons la notion d'**écart** définie et notée comme suit :

$$\mathbf{[x - y] = [I_j - I_k] = Card(\boldsymbol{I}) - 1 - sup(j,k) + inf(j,k)}$$
$$= c - 1 - sup(j,k) + inf(j,k)$$

Cette définition a un double fondement : logique et cognitif.

a) ***Fondement logique :***

Ce fondement étant en liaison avec la logique multi-valente, nous commencerons par présenter brièvement cette dernière.

La logique multi-valente

La logique multi-valente définie par H. Akdag, M. De Glas et D. Pacholczyk dans [Akdag et al. 92] généralise la logique classique en introduisant des degrés de vérité intermédiaires entre le *Vrai* et le *Faux*. Elle se fonde sur la théorie des multi-ensembles, proposée par De Glas, qui elle même généralise la théorie des ensembles en définissant la notion de degrés d'appartenance.

La logique multi-valente s'applique à la représentation des connaissances [Burattini et al. 92], elle permet en effet de formaliser la gradualité et de représenter des propositions du type "Jean est très grand". Un adjectif graduable, par exemple "grand", ne peut, de part sa nature, être satisfait qu'à un certain degré. Ce degré est verbalisé par un adverbe scalaire (ici "très") qui est formalisé par un degré de vérité.

Dans le cas d'un nombre fini M de degrés de vérité, l'ensemble L_M de ces degrés est noté $L_M=\{\tau_0,..., \tau_i,..., \tau_{M-1}\}$. L_M est muni d'un ordre total ($\tau_i \leq \tau_j$ si et seulement si $i \leq j$) et des opérateurs $\wedge$ (min, qui correspond à la conjonction), $\vee$ (max, qui correspond à la disjonction) et $\neg$ (avec $\neg \tau_i=\tau_{M-1-i}$, $\neg$ traduit la négation). L_M a une structure de treillis de De Morgan. Une implication multi-valente (qui est celle de Lukasiewiscz) est définie sur L_M par :

$$\tau_i \rightarrow \tau_j = \tau_{M-1} \qquad \text{si } i \leq j$$
$$= \tau_{M-(i-j)-1} \qquad \text{si } j < i$$

Le treillis $\boldsymbol{I}$ des intervalles, totalement ordonné, est isomorphe à L_7, (Card($\boldsymbol{I}$)=7). On peut donc considérer que les intervalles I_k correspondent à des degrés de vérité, on voit alors que :

Si $k \geq j$ alors $[I_j - I_k] = Card(\boldsymbol{I}) - 1 - k + j = I_k \rightarrow I_j$
sinon $[I_j - I_k] = Card(\boldsymbol{I}) - 1 - j + k = I_j \rightarrow I_k$

D'où : $\mathbf{[I_j - I_k] = I_{sup(j,k)} \rightarrow I_{inf(j,k)}}$

La position relative ainsi définie correspond donc à une implication en logique multi-valente. L'aspect ordinal est prédominant; si les données dont nous disposons sont purement qualitatives, les opérations et manipulations définies (écart, appréciation et comparaison qualitatives de notes) restent valables. Le modèle que nous proposons permet en effet de réaliser des manipulations symboliques sur les données, il n'est donc pas indispensable que ces dernières soient numériques. Il est par contre nécessaire que l'ensemble $\boldsymbol{I}$ des intervalles soit totalement ordonné comme l'est l'ensemble L_M des degrés.

b) ***Fondement cognitif :***

Pour comparer deux notes, les sujets utilisent les ordres de grandeur : ils font une comparaison "relative". L'*écart* que nous avons défini est bien une position relative des deux paramètres considérés et non pas une distance.

L'écart est symétrique par rapport aux variables x et y, on peut donc supposer que k≥j, on a alors : $[x - y] = [I_j - I_k] = c - 1 - k + j$. Les différentes valeurs possibles prises par [x - y] sont :

6 : *même ordre de grandeur* (x MOG y)
5 : valeurs *proches*
4 : valeurs *assez proches*
3 : valeurs *éloignées*
2 : valeurs *très éloignées*
1 : valeurs *incomparables*
0 : valeurs *extrêmes*.

Ces valeurs, totalement ordonnées, correspondent à une échelle de comparaison et peuvent traduire le degré auquel l'écart des deux variables satisfait un adjectif graduable de comparaison. Un adjectif graduable possible est "comparable" :

"x et y sont comparables"
↑
prédicat linguistique graduable, donc satisfiable à un certain degré

Les adverbes scalaires suivant : totalement, très, assez, moyennement, peu, très peu, pas du tout, verbalisent le degré auquel les variables x et y satisfont le prédicat "comparables" [Akdag et al. 92], [Burattini et al. 92]. Ils traduisent l'échelle de comparaison ordinale :

$$\text{"x et y sont} \begin{pmatrix} 6 : \text{totalement} \\ 5 : \text{très} \\ 4 : \text{assez} \\ 3 : \text{moyennement} \\ 2 : \text{peu} \\ 1 : \text{très peu} \\ 0 : \text{pas du tout} \end{pmatrix} \text{comparables"}$$

4. APPRÉCIATION QUALITATIVE D'UNE VARIABLE NOTE

Les intervalles $I_0, I_1, ..., I_6$ correspondent à une échelle totalement ordonnée de 0 à 6, ces valeurs peuvent traduire le degré auquel une note satisfait un adjectif graduable. En effet, la notion d'ordre qui ressort de cette échelle suggère un lien avec la gradualité. Un adjectif graduable possible est "élevé" : "x est élevée", x étant une valeur observée (une note) ou un paramètre de même unité que la variable étudiée (par exemple le mode).

On a alors 7 adverbes scalaires, correspondant aux 7 intervalles de ***I***, qui verbalisent le degré (de 0 à 6, indices respectifs des intervalles I_0 à I_6) auquel la variable considérée satisfait le prédicat linguistique "élevé", cela permet d'émettre une *appréciation qualitative* sur une variable note x (valeur observée, moyenne arithmétique, mode, ...).

$$\text{"x est} \begin{pmatrix} 0 : \text{pas du tout} \\ 1 : \text{très peu} \\ 2 : \text{assez peu} \\ 3 : \text{assez} \\ 4 : \text{correctement} \\ 5 : \text{très} \\ 6 : \text{totalement} \end{pmatrix} \text{élevé"}$$

4.1. Le raisonnement des novices en statistiques

Le lien établi avec la gradualité est intéressant pour la *représentation des connaissances* et l'interprétation des verbalisations faites par les sujets étudiés par Denis Corroyer[Corroyer 92] . En effet, les tests révèlent que, dans la plupart des cas, les sujets utilisent des adjectifs graduables, cela leur permet d'exprimer un degré et une intensité avec laquelle la variable vérifie plus moins l'adjectif. Leurs *appréciations qualitatives* du type "Le mode est (adverbe) élevée", pourront être modélisées en déterminant l'intervalle I_k dont l'indice k correspond à l'adverbe employé.

Inversement, la *représentation qualitative* d'une note *donnée* est faite par l'intervalle I_k auquel elle appartient. L'adverbe associé à cet intervalle I_k permet d'apprécier qualitativement la note en question.

4.2. Le raisonnement des experts en statistiques

Les experts en statistiques peuvent émettre des appréciations de nature qualitative sur des données ou des paramètres statistiques. L'interprétation qualitative intervient après le calcul exact des paramètres de position (la moyenne arithmétique, le mode et la médiane).

Le traitement des appréciations qualitatives émises par les experts est identique au cas des novices en statistiques, décrit dans le paragraphe précédent, et ce aussi bien pour la description des données que pour les paramètres de position. En effet une propriété commune aux paramètres de position considérés est qu'ils s'expriment dans les mêmes unités que les observations, cela est intéressant dans la mesure où nous pouvons utiliser la même description qualitative (basée sur les intervalles de ***I***) que pour les valeurs observées. L'étude de paramètres d'unité différente que celle des valeurs observées, comme par exemple la variance, nécessiterait l'introduction d'une autre échelle d'appréciation et de comparaison.

Prenons l'exemple du mode. Nous notons X_m le mode de la série 1 (sous l'hypothèse que ce mode existe et est unique). $X_m$est dans [0,20], il existe alors m* dans {0,1,...,c-1} tel que le mode X_m soit dans I_{m*}. Cela permet une interprétation de la valeur du mode en terme qualitatif : X_m sera médiocre, mauvais, bon, excellent, ..., en fonction de l'appréciation associée à l'intervalle I_{m*}.

5. COMPARAISON QUALITATIVE DE DONNÉES STATISTIQUES

Pour modéliser la comparaison qualitative de deux notes a et b ou de deux paramètres d'unité *note*, il suffit de calculer leur position relative [a-b]. Cela est valable aussi bien pour le raisonnement des novices en statistiques que pour celui des experts.

Si $a \in I_i$ et $b \in I_j$ $i, j \in \{0, 1, ...,c-1\}$, on a : $[a-b] = [I_i - I_j] = c-1 -\sup(i,j)+\inf(i,j)$.

On obtient ainsi un degré de satisfiabilité de l'adjectif "comparable" sur l'échelle des écarts, l'adverbe scalaire associé à ce degré permet alors de *comparer qualitativement* les deux notes considérées.

Si, par contre, nous disposons d'une verbalisation relative à la comparaison qualitative de deux notes ou de deux paramètres de position, comme par exemple "Les moyennes des deux séries de notes sont *très peu* comparables", nous pouvons représenter cette information. En effet, une phrase de ce type peut être traduite en terme d'écart entre les deux moyennes, c'est à dire en terme de "position" sur l'échelle des écarts. Il suffit pour cela de déterminer le correspondant ordinal (ici : 1) de l'adverbe scalaire utilisé (ici "très peu") dans l'échelle ordinale de comparaison. On obtient ainsi la position relative des deux notes considérées.

Remarque 1 :

Les propriétés de la théorie des multi-ensembles et de la logique multi-valente pourront être exploitées pour nos cas de gradualité, et notamment les possibilités de composer les degrés de vérité (conjonction, disjonction, négation, implication, ...), [Akdag et al. 92], [Burattini et al. 92].

Remarque 2 :

Ce modèle qui décompose l'intervalle réel [0,20] en un nombre fini de sous-intervalles totalement ordonnés, apporte deux réponses intéressantes de par la prise en compte de la notion de gradualité [Akdag et al. 92], [Burattini et al. 92] :

- il y a une inadéquation entre la variabilité potentiellement infinie des degrés auxquels un prédicat graduable peut être satisfait et le nombre fini de mots pour l'exprimer. Le langage discrétise par les mots, et cela peut se formaliser par la discrétisation de l'intervalle [0,20] en un nombre fini d'intervalles,

- plus le nombre de degrés est élevé, plus la connaissance relative à la satisfaction d'un prédicat par un objet est fine. Ainsi, la connaissance augmente avec la croissance du nombre de degrés.

Le découpage du domaine est donc très important, le nombre d'intervalles à considérer et leurs bornes dépendent du sujet d'étude et de la précision recherchée.

Remarque 3 :

Il est possible de modéliser l'appréciation qualitative et la comparaison d'effectifs [Borgi 94]. On procède de la même manière que pour les valeurs observées : on découpe l'intervalle [0,n], domaine des effectifs, en un nombre fini d'intervalles significatifs pour le problème et on définit l'écart entre deux effectifs comme celui défini entre deux valeurs observées. Pour l'interprétation des verbalisations des sujets relatives aux effectifs (comme par exemple : "Peu de notes sont élevées") ou pour la représentation qualitative d'un effectif donné, on procède également de la même manière que pour le traitement d'une note.

6. TRAITEMENT DE LA MOYENNE ARITHMÉTIQUE PAR LES NOVICES EN STATISTIQUES

Les novices en statistiques ont une approche plus grossière que les experts, ils font des approximations dans le calcul des paramètres caractérisant les séries statistiques et obtiennent donc des résultats moins précis et une analyse moins fine. Cependant leurs conclusions n'en sont pas forcément moins significatives, du moins si la précision requise n'est pas très élevée.

6.1. Transformation des données statistiques

En général les novices en statistiques ne font pas le calcul exact de la moyenne arithmétique, mais en donne une estimation pas toujours précise. La moyenne sera par exemple approchée par un intervalle (verbalisation du type : "*La moyenne est comprise entre 12 et 14* "). Ils ne distinguent pas les notes de même ordre de grandeur, cette approche est modélisée en regroupant les notes appartenant au même intervalle I_k, (k=0,1, ...,c-1).

Nous obtenons alors deux séries *approximatives* à partir des deux séries observées initiales dont nous disposons.

Pour la série 1, nous avons D.O. = $\{(X_j,n_j), j=1, 2, ..., J\}$

$\forall\, j \in \{1, 2, ..., J\}, \exists\, k \in \{0, 1, ..., c-1\}$ tel que : $X_j \in I_k$

On regroupe les X_j, j variant de 1 à J, appartenant au même intervalle, d'où une série *réorganisée* que nous noterons ***A*** définie par : $\boldsymbol{A = \{(I_k, sn_k), k=0,...,c-1\}}$

avec pour tout $k \in \{0,1,...,c-1\}$, $\quad sn_k = \sum_j n_j \quad j \in \{1,2,\ldots,J\} / \; X_j \in I_k$

sn_k est l'*effectif cumulé* sur l'intervalle I_k ($k \in \{0,1,...,c-1\}$).

S'il existe un intervalle I_k, (k=0,...,c-1), ne contenant aucune valeur observée X_j, (j=1,...,J), on lui associe l'effectif nul : $sn_k=0$.

Pour la série 2, D.O.', on obtient de même une série *approximative* ***A***' :

$$\boldsymbol{A}' = \{(I_k, sn'_k), k=0,...,c-1\},$$

6.2. Calcul approché de la moyenne

Le calcul de la moyenne se fera à partir des séries ***A*** et ***A'***. Considérons par exemple la série ***A***, on aura pour "valeur approchée" de la moyenne de D.O., que l'on notera $\sim \bar{x}$:

$$\sim \overline{x} = (\sum_{k=0}^{c-1}) \frac{sn_k}{n} \otimes I_k$$

Cette valeur approchée est en fait un intervalle de [0,20].

$\oplus$ et $\otimes$ sont une addition et une multiplication qu'il faut définir. La première méthode de calcul possible consiste à faire une première approximation lors du calcul de chaque terme $sn_k/n \otimes I_k$, pour k variant de 0 à c-1, et une deuxième approximation lors du calcul de la somme de ces termes [Borgi 94]. Cette méthode de calcul qualitatif s'avère très grossière et conduit à une trop grande perte d'information. La deuxième méthode de calcul que nous proposons est plus précise car une seule approximation y est faite.

Les approximations utilisées dans les deux méthodes sont inspirées des travaux de Antoine Missier, Nuria Piera et Louise Travé-Massuyès décrits dans [Missier et al. 89]. Leur étude est développée autour de l'axiomatique de l'égalité qualitative et de la définition d'une structure algébrique appelée Algèbre Qualitative. Une classe particulière d'Algèbres Qualitatives bien adaptée au calcul qualitatif peut se construire à partir de partitions de l'axe réel.

Dans notre cas, il ne s'agit pas d'une partition de IR mais du découpage du domaine [0,20] de la variable statistique étudiée en un nombre fini de sous-intervalles.

On considère l'ensemble S, univers de description, qui est l'union de ***I*** et de tout intervalle pouvant s'exprimer comme l'union d'éléments adjacents de ***I*** [Missier et al. 89], [Piera et al. 92]. On définit sur S un ordre partiel et une égalité qualitative identiques à ceux définis dans [Missier et al. 89] :

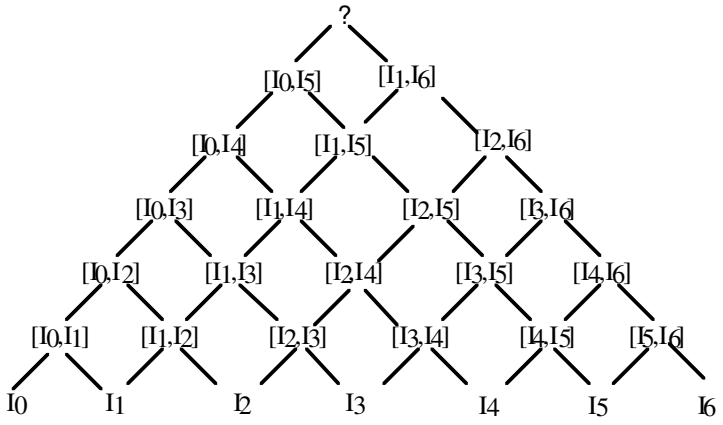
- ***ordre partiel*** : pour x, y dans S, $\mathbf{x} \leq \mathbf{y}$ si et seulement si $x \subset y$ (au sens ensembliste),
- ***égalité qualitative*** (qui traduit la possibilité d'être égal) : pour x, y dans S, $\mathbf{x} \approx \mathbf{y}$ si et seulement si il existe $a \in S$ tel que $a \leq x$ et $a \leq y$.

Pour toute partie X de [0,20], S_X est l'ensemble des éléments de S contenant X :

$$S_X = \{Y \in S / X \subset Y\}$$

S_X a un minimum sur S pour la relation d'ordre $\leq$ (qui est l'inclusion), ce minimum n'est pas forcément unique car $\leq$ est un ordre partiel.

L'univers de description S muni de la structure d'ordre partiel $\leq$ peut être représenté comme suit :



L'élément ? correspond à l'intervalle [0,20] et traduit l'indétermination.

Nous considérons que $\oplus$ et $\otimes$ sont l'addition et la multiplication définies dans l'ensemble des parties de IR :

- pour $\otimes$, nous plongeons S dans l'ensemble des parties de IR, $\boldsymbol{P}$(IR) :

$\otimes$: [0,1] x $\boldsymbol{P}$(IR) ----> $\boldsymbol{P}$(IR)

(p,A) ----> $p \otimes A = \{p*a \in IR / a \in A\}$

pour tout $k \in \{0,...,c-1\}$, on a : $sn_k/n \otimes I_k = \{sn_k/n * x \ / \ x \in I_k\} = \mathbf{J_k}$

- pour $\oplus$,, nous travaillons toujours dans $\boldsymbol{P}$(IR) :

$\oplus$: $\boldsymbol{P}$(IR) x $\boldsymbol{P}$(IR) ----> $\boldsymbol{P}$(IR)

(A,B) ----> $A \oplus B = \{a+b \in IR / a \in A \text{ et } b \in B\}$

pour tout $k,l \in \{0,...,c-1\}$, on a : $J_k \oplus J_l = \{x+y \in IR / x \in J_k \text{ et } y \in J_l\}$

Notons M le résultat final, $M = \sum_{k=0}^{c-1} J_k$

M est certainement un intervalle de [0,20], en effet :

$\forall k \in \{0,...,c-1\}$, on a : $0 \le sn_k / n \le 1$, $\sum_{k=0}^{c-1} sn_k / n = 1$ et $I_k \subset [0,20]$

On a donc bien : $M = \sum_{k=0}^{c-1} (sn_k / n) * I_k \subset [0,20]$

L'approximation n'est faite qu'une fois que nous disposons de M.

Pour que le résultat $\sim \bar{x}$ soit un élément de S et puisse ainsi être décrit qualitativement, nous le définissons par :

$\sim \bar{x} = \mathbf{Min\ S_M}$ (S_M a un sens car $M \subset [0,20]$)

$\sim \bar{x}$ est donc le plus petit élément de S contenant M.

Ce calcul approché assure l'appartenance de la moyenne exacte $\bar{x}$ à l'intervalle résultat obtenu $\sim \bar{x}$. Cet intervalle est assez significatif et "informatif", il est en général réunion de deux à trois intervalles I_k adjacents [Borgi 94]. Cependant la pertinence du résultat n'est pas garantie, elle dépend en effet du découpage du domaine [0,20] : c'est la rançon de la discrétisation !

6.3. Appréciation et comparaison qualitatives de la moyenne

L'***appréciation qualitative*** de la moyenne $\bar{x}$ est celle(s) associée(s) à l'élément $\sim \bar{x}$ de S, rappelons que $\sim \bar{x}$ est la réunion d'un ou de plusieurs intervalles adjacents de $\boldsymbol{I}$. Cela permet d'avoir une idée plus ou moins précise de la moyenne, selon le niveau de précision de l'intervalle $\sim \bar{x}$ obtenu. Le niveau le plus précis de l'arbre de représentation de l'univers de description correspond aux éléments de $\boldsymbol{I}$ et le moins précis à ?, c'est à dire à [0,20], [Piera et al. 92].

La ***comparaison*** des deux moyennes approchées $\sim \bar{x}$ et $\sim \bar{x}'$ (moyenne approchée de la série 2) peut se faire en utilisant l'égalité qualitative décrits dans [Piera et al. 92] :

$\sim \bar{x} \approx \sim \bar{x}'$ si et seulement si $\exists A \in S / A \le \sim \bar{x}$ et $A \le \sim \bar{x}'$.

Cette méthode de calcul approché de la moyenne fournit un résultat similaire à celui émis par les novices en statistiques. Son point faible est qu'elle ne simule pas la méthode employée par les novices, une perspective de recherche serait d'étudier le comportement cognitif des sujets lors du calcul de la moyenne et de le modéliser.

Remarque 1 :

Les novices en statistiques ne se limitent pas toujours à une approximation sur les valeurs observées de la distribution statistique, mais peuvent également prendre des valeurs approchées pour les effectifs. Cette double approximation conduit à un résultat moins précis que dans le cas d'une approximation sur les valeurs observées uniquement. Chaque approximation conduisant évidemment à une perte supplémentaire d'informations.

Pour modéliser l'approximation sur les effectifs, nous utilisons la discrétisation de [0,n] (n est l'effectif global de la série observée) en un nombre fini d'intervalles d'effectifs. Chaque effectif cumulé sn_k,(k=0,...,c-1), de la série approximative $\boldsymbol{A} = \{(I_k, sn_k), k=0,...,c-1\}$ est approché par l'intervalle d'effectif auquel il appartient. On obtient ainsi une nouvelle série *doublement approximative* à partir de laquelle se fait le calcul approché de la moyenne. Ce calcul est similaire à celui utilisé pour le cas d'une approximation sur les valeurs observées uniquement [Borgi 94].

Remarque 2 : *Le traitement du mode par les novices en statistiques*

L'extraction du mode est souvent traitée par les novices de la même manière que les experts, leurs conclusions sont alors identiques. Cependant les novices ont parfois une approche moins précise que celle des experts. Ils déterminent un intervalle modal au lieu d'une valeur modale. Cela se traduit par des verbalisations du type : "*La plupart des notes sont comprises entre a et b*". L'interprétation qualitative reste toujours possible, il suffit de déterminer le plus petit élément de S contenant l'intervalle [a,b], soit M* cet élément :

$$M^* = \text{Min } S_{[a,b]}.$$

L'appréciation qualitative du mode est alors celle(s) associée(s) à l'intervalle M*.

On peut également prendre en compte les effectifs : il suffit de traduire "*la plupart*" sur l'échelle des effectifs (échelle issue du découpage de [0,n] en un nombre fini d'intervalles d'effectifs).

7. CONCLUSION

Pour réaliser le travail présenté dans ce rapport, nous nous sommes basés sur des travaux de psychologie et plus précisément de psychologie cognitive décrits dans [Corroyer 92]. Pour les exploiter, nous avons fait appel aux notions de gradualité et de logique multi-valente. Des aspects de linguistique ont été abordé à travers les adjectifs graduables et les adverbes scalaires qui permettent de verbaliser la gradualité. Toutes ces notions ont contribué à la formalisation des verbalisations des sujets et à la représentation de connaissances qualitatives.

Nous avons tenté, dans cet esprit interdisciplinaire, de mieux refléter le raisonnement humain lors de la description qualitative de données statistiques.

Le travail réalisé présente également un intérêt de part ses applications possibles. Il pourrait par exemple être exploité dans le cadre d'un projet d'E.I.A.O. (Environnements Interactifs d'Apprentissage avec Ordinateur) pour l'enseignement des statistiques et contribuer à construire un modèle de l'élève [Bournaud et al. 94]. Il pourrait aussi aider à la simulation du raisonnement des experts ou des novices en statistiques lors de la résolution de problèmes et contribuer à une meilleure compréhension du raisonnement humain.

De nombreuses perspectives de recherche sont envisageables. L'une d'elles consiste à étudier le découpage en intervalles du domaine de la variable statistique, il s'agit de dégager le découpage le plus proche du raisonnement des experts et/ou des novices et ce en étudiant de près leurs comportements cognitifs : observations, réactions et analyses. Une deuxième voie de recherche serait de simuler la méthode cognitive du calcul approché de la moyenne par les novices en statistiques, ainsi la modélisation de leur raisonnement ne sera pas limité au résultat obtenu. On pourrait également définir des opérations d'addition et de multiplication purement qualitatives sur l'ensemble $\boldsymbol{I}$ des intervalles issus de la discrétisation du domaine de la variable statistique, on obtiendrait ainsi un modèle entièrement symbolique, purement qualitatif. D'autres améliorations du modèle proposé restent possibles, comme par exemple étendre la modélisation aux paramètres de dispersion et aux séries multivariées, ou envisager un ordre partiel et non plus total sur $\boldsymbol{I}$. Un ordre partiel traduirait mieux l'usage des adverbes scalaires lors de la description qualitative de données, en effet ces adverbes ne sont pas, en général, totalement ordonnés.

BIBLIOGRAPHIE

[Akdag et al. 92] : H. Akdag, M. De Glas et D. Pacholczyk. "A Qualitative Theory of Uncertainty". Fundamenta Informaticae 17, pp 333-362, 1992.

[Borgi 94] A. Borgi. "Le raisonnement qualitatif en statistiques". Rapport du D.E.A. I.A.R.F.A. Université Pierre et Marie Curie, Paris, 1994.

[Bournaud et al. 94] : I. Bournaud, J. Mathieu, D. Corroyer, H. Rouanet. "Using Artificial Intelligence in The Teaching of Statistics". 1994 International Symposium on Mathematics/Science Education and Technology, San Diego, CA USA, 1994.

[Burattini et al. 92] : E. Burattini, M. De Glas and M. De Gregorio. "A Symbolic Model for the Representation of Uncertain Knowledge. Theory and Applications". IEEE, International Conference on Systems, Man and Cybernetics, Chicago,Illinois, 1992.

[Corroyer 92] : Denis Corroyer. “Le traitement statistique des données en psychologie et son enseignement : de l'ère des tables à l'ère informatiques”. Thèse de l'Université René Descartes, Sciences Humaines, 1992.

[Dormoy 89] : J.L. Dormoy. "Nouvelles méthodes de calcul qualitatif", Revue d'intelligence artificielle, Vol. 3, n. 4, pp 39-53. 1989.

[Dormoy et al. 89] : J. L. Dormoy et L. Travé-Massuyès. Editorial. Revue d'intelligence artificielle, Vol. 3, n. 4, pp 7-10. 1989.

[Droesbeke 92] : J. Droesbeke. "Eléments de statistiques", éditions de l'Université de Bruxelles, éditions Ellipses, collection *Statistiques et Mathématiques Appliquées*, 1992.

[Fouché et al. 89] : P. Fouché, A. Charles, J-P. Barthès, C. Melin. "Qualitative physics : a survey". Revue d'intelligence artificielle, Vol. 3, n. 4, pp 10-38. 1989.

[Haton et al. 91] : J-P. Haton, N. Bouzid, F. Charpillet, M-C. Haton, B. Lâasri, H. Lâasri, P. Marquis, T. Mondot, A. Napoli. "Le Raisonnement en Intelligence Artificielle", InterEditions, Paris, 1991.

[Hayes 79] : P. J. Hayes. "The Naive Physics Manifesto". Expert Systems in the Micro Electronic Age, ed. D. Michie, Edinburgh University Press, Edinburgh 1979.

[Kuipers 86] : B. Kuipers. "Qualitative Simulation". Artificial Intelligence, Vol 29, pp 289-338, 1986.

[Missier et al. 89] : A. Missier, N. Piera and L. Travé-Massuyès. "Order of magnitude qualitative algebras". Revue d'intelligence artificielle, Vol. 3, n. 4, pp 95-109. 1989.

[Piera et al. 92] : N. Piera and N. Agell. "Binary Relations For Qualitative Reasoning", IEEE, 1992.

[Raiman 91] : O. Raiman. "Order of Magnitude Reasoning", Artificial Intelligence, Vol. 51, pp 11-38, 1991.

[Williams et al. 91] : B. C. Williams and J. de Kleer. "Qualitative reasoning about physical systems : a return to roots", Artificial Intelligence, Vol 51, pp 1-9, 1991.