UNIVERSITÉ LOUIS PASTEUR STRASBOURG

Thèse présentée pour obtenir le grade de
Docteur de l'Université Louis Pasteur
Strasbourg I

Discipline : Sciences Médicales

Recherche Clinique, Innovation Technologique, Santé Publique

par Nicolas MEYER

Méthodes statistiques d'analyse des données d'allélotypage en présence d'homozygotes

Soutenue publiquement le : 22 juin 2007

Membres du jury

Directeur de thèse : M. Pierre MEYER, Professeur, ULP Strasbourg
Rapporteur Interne : M. Daniel GRUCKER, Professeur, ULP Strasbourg
Rapporteur Externe : Mme Catherine QUANTIN, Professeur, UB Dijon
Rapporteur Externe : M. François KOHLER, Professeur, UHP Nancy
Examinateur : M. Pierre OUDET, Professeur, ULP Strasbourg

Remerciements

A Monsieur le Pr. Meyer, pour m'avoir aidé à creuser mon sillon...

A Madame le Pr. Quantin, pour votre bienveillance à mon égard,

A Monsieur le Pr. Kohler, pour votre jugement, riche d'enseignement,

A Monsieur le Pr. Oudet, pour la confiance que vous m'avez toujours témoigné,

A Monsieur le Pr. Grucker, pour l'honneur que vous me faite de juger mon travail,

... et à Marie Pierre Gaub, pour ton aide très précieuse. Sans toi, je n'aurais pas pu mener à bien ce travail !

— — —

A mon épouse et mes filles, pour leur infinie patience,

A ma famille.

— — —

« En réalité, nous ne savons rien, car la vérité est au fond de l'abime. »

Démocrite

Résumé Les données d'allélotypage contiennent des mesures réalisées par *Polymerase Chain Reaction* sur une série de microsatellites de l'ADN afin de déterminer l'existence d'un déséquilibre allélique pour ces microsatellites. D'un point de vue statistique, ces données sont caractérisées par un nombre important de données manquantes (en cas d'homozygotie du microsatellite), par des matrices carrées ou comportant plus de variables que de sujets, des variables biniomiales, des effectifs parfois faibles et éventuellement de la colinéarité. Les méthodes statistiques fréquentistes ont un nombre important de limites qui font choisir un cadre bayésien pour analyser ces données. En analyse univariée, l'intérêt du facteur de Bayes est exploré et différentes variantes selon l'absence ou la présence de données manquantes sont comparées. Différents types d'imputations multiples sont ensuite étudiés. Des modèles de type méta-analyses sont également évalués. En analyse multivariée, un modèle de type *Partial Least Square* est développé. Le modèle est appliqué sous une forme de modèle linéaire généralisé (régression logistique) et combiné avec l'algorithme *Non Iterative Partial Least Squares*, ce qui permet de gérer simultanément toutes les limites propres aux données d'allélotypage. Les propriétés de ce modèle sont explorées. Il est ensuite appliqué à des données d'allélotypage portant sur 33 microsatellites de 104 patients porteurs d'un cancer du colon pour prédire le stade Astler-Coller de la tumeur. Un modèle avec toutes les interactions possibles entre couples de microsatellites est également réalisé.

Title Considering homozygotes in Statistical analysis of allelotyping data.

Summary Allelotyping data contain measures done using Polymerase Chain Reaction on a batch of DNA microsatellites in order to ascertain the presence or not of an allelic imbalance for this microsatellites. From a statistical point of view, those data are characterised by a high number of missing data (in case of homozygous microsatellite), square or flat matrices, binomial data, sample sizes which may be small with respect to the number of variables and possibly some colinearity. Frequentist statistical methods have a number of shortcomings who led us to choose a bayesian framework to analyse these data. For univariate analyses, the Bayes factor is explored and several variants according to the presence or absence of missing data are compared. Different multiple imputations types are then studied. Meta-analysis models are also assessed. For multivariate analyses, a Partial Least Square model is developed. The model is applied under a generalised linear model (logistic regression) and combined with a Non Iterative Partial Least Squares algorithm which

makes it possible to manage simultaneously all the limits of allelotyping data. Properties of this model are explored. It is then applied on allelotyping data on 33 microsatellites of 104 patients who have colon cancer to predict the tumor Astler-Coller stage. A model with all possible microsatellites pairs interactions is also run.

Mots-clés : Polymerase Chain Reaction, Partial Least Squares, bayes, allélotypage, microsatellites,

Key-words : Polymerase Chain Reaction, Partial Least Squares, bayes, allelotyping, microsatellites,

Adresse : Laboratoire de Biostatistique
Faculté de Médecine
4, rue Kirschleger
67089 STRASBOURG

Table des matières

1 Introduction

Les données d'allélotypage sont des données issues de la biochimie moléculaire. Elles sont constituées des états dits normaux ou altérés d'une série de microsatellites qui sont des zones particulières de l'ADN. Ces données d'allélotypage sont couramment utilisées en cancérologie pour décrire les éventuelles lésions chromosomiques que peuvent présenter les cellules cancéreuses.

Les données d'allélotypage, comme toutes données biologiques, sont appelées à être traitées statistiquement afin d'étudier les propriétés des microsatellites ainsi que leurs relations avec d'autres facteurs relatifs soit au patient soit à la tumeur en elle-même. Comme nous le verrons par la suite, les données d'allélotypage présentent un certain nombre de caractéristiques rendant leur exploitation statistique difficile. L'objectif global de ce travail sera centré sur la proposition de méthodes permettant de traiter correctement d'un point de vue statistique les données d'allélotypage. Des méthodes nouvelles seront proposées et d'autres seront adaptées à la problèmatique présentée. L'utilisation de ces méthodes sera illustrée sur des jeux de données rélles. Cependant, nous soulignons d'emblée que l'objectif est ici non pas d'analyser spécifiquement ces données mais bien de proposer des méthodes *permettant* de réaliser ces analyses. L'inteprétation biologique et médicale qui pourra être faite des données utilisées ici ne sera que brièvement abordée, l'accent étant porté sur l'aspect statistique du problème. Les méthodes proposées seront donc utilisées mais l'analyse ne sera pas poussée outre mesure au-delà de la vérification de l'applicabilité des méthodes. En effet, notamment en raison du nombre très important de questions posées en pratique par les données d'allélotypage, il n'était pas envisageable de traiter à la fois la recherche de nouvelles méthodes et l'analyse des conclusions qui pourront être apportés par les-dites méthodes sur les données utilisées.

Nous allons maintenant aborder plus précisément les aspects techniques des données d'allélotypage et les problèmes liées à leur exploitation statistique.

1.1 Définition des microsatellites

Les microsatellites (MS) sont des éléments de l'ADN répété du génome. Les séquences d'ADN répétées sont fréquentes dans le génome humain puisqu'elles constituent environ 10% du génome. Un microsatellite est un polymorphisme de séquence simple, comportant habituellement des copies en tandem d'unité de répétition de un, deux, trois ou quatre nu-

cléotides. Ils sont également appelés simple répétition en tandem ou *Single Tandem Repeat*, STR [50, 204]. Un microsatellite est donc une séquence génétique composée de la répétition d'un motif élémentaire. Ce motif est constitué d'un petit nombre de nucléotides. Si la définition classique admet un nombre de répétitions allant de 1 à 4, pour d'autres, le nombre de répétitions peut aller jusqu'à 13 [57]. Par ailleurs la taille de la séquence répétée (du motif) varie d'un microsatellite à l'autre. Les différents allèles des microsatellites se distinguent donc non pas par une modification du motif de base, comme pour les gènes classiques, mais par une variation du nombre n de répétition de ce motif, ce nombre variant de 20 à 125 [50] et la plupart des microsatellites a une longueur inférieure à 150 paires de bases. Ainsi on distingue deux allèles d'un microsatellite donné par le nombre de répétitions du motif dans la séquence des deux allèles, un allèle ayant n répétitions, l'autre ayant m répétitions, n et m pouvant donc être différents. Le nombre de répétitions semble toutefois être limité par un seuil maximum et l'on ne peut avoir d'allèle de taille indéfiniment grande [235]. De façon synthétique, le microsatellite peut donc se noter de la façon suivante : $(w, x, y, z)_n$. où n représentent le nombre de répétitions et w,x,y,z représente l'une des 4 bases de l'ADN. Cette définition est sujet à débat.

Les microsatellites sont ubiquitaires dans le génome humain. On les trouve en effet sur tous les chromosomes sans qu'une localisation préférentielle ait pu être mise en évidence. Les microsatellites les plus fréquents sont les dinucléotides, au nombre d'environ 140 000. A noter que les répétitions d'un seul nucléotide (*monorepeat*) sont également très fréquentes puisque l'on en trouve environ 120 000 dans le génome. Au total le nombre de microsatellites est d'environ $6,5.10^5$. Les monorepeat peuvent être localisés dans les séquences codantes des gènes.

L'origine de ces microsatellites est probablement expliquée par des mutations neutres survenant dans des séquences non-informatives, lesquelles mutations sont ensuite répétées d'une génération à l'autre. Le nombre variable de répétitions d'un allèle à l'autre s'explique par la survenue d'erreurs de copie lors de la synthèse d'ADN, l'ADN-polymérase étant connue pour présenter des « bégaiements » lorsqu'elle parcourt des séquences répétitives. Ces erreurs de copie consistent donc à copier deux fois le même motif (de même indice i, $i \in \{1, \ldots, n\}$) aboutissant à une nouvelle séquence de longueur $n+1$ copies [50]. Une autre erreur de copie consiste pour l'ADN polymerase à « sauter » un motif ce qui provoque un raccourcissement de la séquence. Une conséquence de ces erreurs de copie est que ces séquences microsatellites sont très polymorphes d'un individu à l'autre car le nombre d'allèles

d'un microsatellite est en général assez élevé. Ceci est lié au fait que le niveau de polymorphisme dépend globalement de la longueur du microsatellite puisque plus le microsatellite est long plus le risque d'erreur de copie est grand. Donc, en général, plus le microsatellite est long plus il est polymorphe. Par ailleurs, les microsatellites sont faciles à typer en utilisant la Polymerase Chain Reaction (PCR).

Ces séquences n'ont pas de fonctions connues dans le sens où ce sont des régions non codantes et que ces microsatellites n'interviennent pas dans la transcription des séquences codantes. Leur rôle n'est pas encore élucidé. Ces microsatellites sont néanmoins des marqueurs de l'évolution du génome car, n'ayant pas d'activité fonctionnelle propre, leur mutation est neutre vis à vis de la survie(4) à l'exception de ceux localisés dans les séquences codantes des gènes.

Les microsatellites sont utilisés pour baliser l'ADN, car ils présentent des caractéristiques particulières leur permettant d'être facilement positionnés sur une carte génomique. Leur répartition est homogène sur l'ensemble du génome et il y a donc pratiquement toujours un microsatellites à proximité d'un gène d'intérêt. Le microsatellite en question peut alors jouer un rôle de balise pour le gène donné ce qui permet donc de suivre le transfert de ce gène d'une génération à l'autre ou encore d'une génération cellulaire à une autre. Enfin, les microsatellites sont d'un point de vue technique plus facile à analyser (repérer) que les gènes qu'ils sont censés marquer, d'où leur utilisation.

En dehors des microsatellites, il existe d'autres marqueurs tels que les fragments de restriction polymorphes (RFLP) et les polymorphismes portant sur un seul nucléotide (SNP). Les microsatellites, encore appelés dans ce cadre STR (Simple tandem Repeat), appartiennent quant à eux à la catégorie des séquences simples présentant un polymorphisme de longueur (SSLP) contenant également les *minisatellites*.

Intérêt théorique et pratique des microsatellites Les microsatellites sont utilisés actuellement dans trois situations. Dans le premier cas, les microsatellites sont utilisés dans la réalisation de l'arbre génétique d'une famille porteuse d'une pathologie génétique pour montrer la transmission du gène suspecté ([106]). Une seconde application, indépendante de tout gène, est celle que l'on en fait en Médecine légale dans le cadre de l'identification de sujets, encore appelée identification génétique ([106]). Le haut polymorphisme des microsatellites permet avec un nombre limité de microsatellites d'individualiser un sujet (exceptés bien sûr les jumeaux monozygotes) et donc de le distinguer d'un autre individu. Dans le

troisième cas, on utilise le microsatellite pour définir l'état du génome d'une cellule et donc pour décrire l'état de cet ADN et les modifications importantes qu'il aura pu subir. Ceci est notamment utilisé dans le cadre de la cancérologie.

1.2 Utilisation des microsatellites en cancérologie

De par leur caractère polymorphe, les microsatellites permettent donc de typer un segment chromosomique paternel ou maternel (typage allélique). On peut supposer que si l'on arrive à montrer le gain ou la perte d'une copie d'un ou de plusieurs microsatellites, on montrera par la même occasion le gain ou la perte du segment chromosomique correspondant, sur lequel ce microsatellite est situé. On sait par ailleurs, que les cellules cancéreuses sont caractérisées par des remaniements chromosomiques importants tels que, justement, des gains (par duplication, endo-reduplication ou erreur de ségregégation) ou des pertes (par erreur de ségrégation ou perte pure) de segments chromosomiques. L'utilisation concomitante de plusieurs microsatellites répartis sur l'ensemble du génome ou au moins sur les zones *a priori* intéressantes permet donc de cartographier simultanément les altérations présentes sur plusieurs segments chromosomiques dans une cellule cancéreuse. C'est cette application qui est exploitée dans les données que nous avons ici à traiter.

Les microsatellites ont déjà été utilisés pour étudier de nombreux cancers : cancer du colon, du sigmoïde, du sein, de la vessie, du poumon, tumeurs cérébrales, tumeurs de la sphère ORL et mélanomes entre autres [125, 135, 197, 232, 249, 271, 315, 335, 337, 348, 352, 353]. Les microsatellites ont fait la preuve de leur intérêt dans l'étude des cancers digestifs [192]. Certaines critiques ont cependant été faites à l'encontre des études de perte d'hétérozygotie [325]

La Polymerase Chain Reaction La technique utilisée pour typer les microsatellites est la Polymerase Chain Reaction (PCR) [148, 204, 229].

La PCR est une technique de biologie moléculaire basée sur une répétition de cycles de températures et permettant de faciliter l'analyse d'une courte séquence d'ADN. La méthode vise globalement à amplifier cette séquence présente dans un échantillon à l'aide d'amorce spécifique de l'ADN bordant cette séquence. A partir d'une très faible quantité d'ADN on peut alors obtenir une quantité exploitable de matériel génétique. Les différentes étapes de la PCR sont décrites ci-dessous.

Dans un milieu de réaction tamponné, on introduit l'ADN à analyser, les polymérases et les nucléotides (dNTPs). La réaction se déroule de la façon suivante :

1. conditions initiales : on dispose d'une quantité d'ADN natif, en configuration double brin ;
2. dénaturation initiale : étape de chauffage (à 95°C) ce qui permet de dénaturer l'ADN en séparant les deux brins ;
3. phase d'hybridation : après une baisse modérée de la température de la solution, les amorces introduites préalablement dans le milieu de réaction peuvent s'hybrider avec les simples brins d'ADN matrice ;
4. phase d'élongation à une température plus élevée (à 72°C ce qui augmente la spécificité de la réaction) : les polymérases (Taq polymérase résistante aux hautes températures) réalisent la synthèse de nouveaux brins d'ADN en utilisant les dNTPs libres présents ;
5. remontée de la température (à 95°C) pour dénaturer l'ADN double brin et débuter un nouveau cycle.

A la fin d'un cycle la quantité d'ADN est donc (approximativement) multipliée par 2 dans des conditions optimales de PCR. Une fois terminé, le cycle recommence ce qui permet de doubler une nouvelle fois le matériel de départ. On réalise en général une trentaine de cycles. Un grand nombre de variantes sont possibles mais nous ne les aborderons pas ici.

Le matériel génétique ainsi amplifié doit ensuite être séparé par éléctrophorèse pour pouvoir déterminer la longueur de chaque allèle du microsatellite ciblé.

Typage d'un microsatellite Après PCR, les fragments amplifiés sont analysés par électrophorèse ce qui permettra après quantification de définir si un segment chromosomique a été altéré. Le principe est le suivant : on utilise simultanément de l'ADN réputé sain (issu par exemple de lymphocytes du sujet, hors hémopathies malignes) et de l'ADN du tissu tumoral à l'étude (néoplasme, métastase, adénopathie cancéreuse etc). Lors de la migration des amplifiats d'ADN d'un sujet, des pics apparaissent mettant en évidence les deux allèles de chaque microsatellite : à un allèle correspond un pic. Les hauteurs des pics sont proportionnelles à la quantité d'amplifiat et donc au nombre de copies initialement présentes dans la réaction. On connaît la longueur de l'allèle en utilisant un étalon de taille connue. Pour un microsatellite hétérozygote, on obtient donc deux pics correspondant à la longueur des

deux allèles. Si le microsatellite est homozygote, on obtient un seul pic qui correspond à la superposition des deux pics de chacun des deux allèles [156, 167].

Afin de détecter une perte (ou un gain) de segment chromosomique, après avoir fait migrer simultanément l'ADN sain et l'ADN tumoral, on compare les hauteurs des pics obtenus pour chacun des deux ADN. On suppose évidemment que l'ADN de référence ne présente ni perte ni gain d'un segment chromosomique quelconque. On compare donc les quatre pics obtenus pour les deux allèles du microsatellite, deux pour l'ADN sain et deux pour l'ADN à l'étude. En général, le second pic, correspondant à l'allèle le plus long, est légèrement plus bas que le premier en raison d'un nombre moins important de copies pour les allèles les plus longs. La hauteur des deux pics pouvant être mesurée (par une intensité de fluorescence), on peut calculer le rapport R_s entre le second et le premier pic obtenus pour l'ADN sain. On obtient en général une valeur comprise entre 0,8 et 1. On calcule de la même façon le rapport R_t pour l'ADN du tissu tumoral. S'il n'y a pas de perte du segment chromosomique sur lequel se trouve le microsatellite, on observe le même rapport que pour le tissu sain $R_t = R_s$, aux variations aléatoires expérimentales près. Si en revanche le second rapport est nettement inférieur au rapport obtenu pour le tissu sain, on admet qu'il y a une perte de segments chromosomiques dans toutes ou une partie des cellules tumorales prélevées pour l'analyse. En toute rigueur, la méthode ne permet pas de distinguer une perte de segment chromosomique d'un gain de segment chromosomique (par exemple sur le premier allèle) car la méthode est relative dans le sens où la comparaison entre un tissu tumoral et un tissu sain ne se fait pas par rapport à une valeur absolue de référence mais se fait l'un par rapport à l'autre. Pour être rigoureux, il faut même admettre que si l'un des pics est abaissé dans le tissu tumoral, c'est peut-être en raison de la sur-expression de l'autre allèle. On ne peut donc savoir que de manière relative, si l'un des allèles est modifié dans le tissu tumoral par rapport au tissu sain. Par la suite, nous analyserons donc une diminution de pic d'un des allèles comme la présence d'une altération (gain ou perte) de cet allèle. On parle alors de façon plus générale de déséquilibre allélique. Le seuil à partir duquel on parle de déséquilibre allélique (DA) est défini en général expérimentalement pour un microsatellite donné. Pour cela, on calcule sur du tissu normal pour une série de n sujets le rapport suivant :

$$\frac{|R_s - R_t|}{R_s}$$

On obtient donc un n-échantillon d'une variable quantitative. Pour cet échantillon, on détermine ensuite un seuil de part et d'autre duquel est dichotomisée la variable. Ce seuil est défini par la valeur de la moyenne plus ou moins 3 écart-types et on parle donc d'*allelic imbalance* (AI) lorsque pour un tissu tumoral le rapport dépasse la valeur $s = |\bar{m} \pm 3 \cdot \sigma|$ défini sur un échantillon sain de référence. Sinon, on admet que le microsatellite est normal (N). Ainsi, bien que la mesure d'origine soit quantitative, on ne retient qu'une information qualitative : présence ou absence d'un déséquilibre allélique (AI ou normal), défini à partir du dépassement d'un seuil lors du calcul de la différence entre les pics sur les deux échantillons de tissu sain et tumoral. On a admis jusqu'ici que le sujet était hétérozygote pour le microsatellite en question. En cas de déséquilibre allélique, on parle ainsi de perte d'hétérozygotie, ou *Loss Of Heterozygozity (LOH)*. Comme cela a déjà été signalé plus haut, il s'agit bien d'une perte d'hétérozygotie et non pas d'une perte allélique car la technique de lecture de la PCR ne permet pas de savoir si en fait il y a eu perte d'un des deux allèles ou s'il y a eu surexpression de l'autre allèle.

En un second temps, l'interprétation des données est rendue plus complexe par la présence d'homozygotes. En effet si le microsatellite est homozygote, les deux allèles ont la même longueur et lors de la migration éléctrophorétique, ils donneront des pics qui seront superposés, tant pour le tissu sain que pour le tissu tumoral. Qu'une perte allélique survienne ou non dans le tissu tumoral analysé, la PCR fournira tout de même un pic, lié à l'allèle restant. La disparition éventuelle du deuxième allèle ne pourra pas être et ne sera pas observée. On ne peut alors pas calculer le rapport R_t, ni R_s. C'est pourquoi les sites homozygotes sont dit non-informatifs. *A contrario*, seuls les microsatellites hétérozygotes sont informatifs et donc utilisables pour recherche un déséquilibre allélique.

Dans le cas homozygote, on est donc amené à attribuer au rapport R_t une valeur *manquante*. Cette difficulté est à la base de la problématique qui nous occupe dans ce travail. En effet, la proportion de microsatellite homozygote est d'environ 1/3, quel que soit le microsatellite, ce qui génère des données manquantes en très grand nombre. Il ne s'agit pas ici d'un problème technique qui pourrait être amélioré par une modification adéquate des conditions de réalisation de la mesure. Le résultat *manquant* pour les homozygotes est inhérent à la méthode utilisée (la PCR) et ne peut donc pas être contourné (sauf évidemment à changer de technique).

1.3 Aspects généraux des données à étudier

Les jeux de données d'allélotypage se présentent classiquement comme la juxtaposition du typage allélique (ou allélotypage) d'un certain nombre de microsatellites pour un nombre variable de sujets. On ne retient ici que la situation où les mesures obtenues pour chaque microsatellite sont catégorisées dans des variables à deux classes : présence / absence de déséquilibre allélique ou *allelic imbalance* (AI). Pour des raisons liées aux habitudes de langage nous utiliserons par la suite l'abréviation AI pour désigner un déséquilibre allélique. Le tableau est donc constitué d'une série de « 0 » et de « 1 » auxquels se rajoute un certain nombre de données manquantes. Très souvent, à ces variables est ajoutée une variable de type résultat pour laquelle on cherche à prédire les variations en fonction de l'état des microsatellites. Il s'agit soit d'un type cancéreux, comme un stade, ou bien la survenue d'un évènement particulier dans le cours de la maladie (rechute, métastase etc.) ou encore la survie du sujet. Les objectifs sont le plus souvent doubles : tout d'abord décrire les microsatellites dans leur globalité et ensuite prédire au mieux la variable résultat à partir de tout ou d'un sous-ensemble des microsatellites [159].

Décrire les microsatellites L'intérêt du typage simultané de plusieurs microsatellites est de pouvoir décrire l'état du jeu de chromosomes d'une cellule. Cela permet donc d'envisager une description des mécanismes aboutissant à la cancérogénèse puis à l'évolution métastatique. Un modèle classique de Vogelstein pour le cancer du colon par exemple, fait l'hypothèse d'une accumulation séquentielle d'anomalie [105]. Les données d'allélotypage permettent de vérifier ou d'infirmer cette hypothèse. Le typage allélique a pour objectif de permettre de comprendre l'enchainement des événements délétères pour la cellule et de décrire la ou les voies de la cancérogénèse. Cela suppose une description simultanée de tous les microsatellites ce qui implique des analyses statistiques descriptives multivariées.

Association et prédiction d'un résultat Le second objectif de ce type d'étude est bien sûr de comprendre comment telle association de lésion pour tel groupe de microsatellites peut expliquer la survenue de tel évènement dans le cours de la pathologie. Il paraît naturel de vouloir utiliser les caractéristiques génétiques d'un sujet pour évaluer un risque de rechute ou une durée de survie. Cette estimation ne peut logiquement se faire qu'en considérant la structure globale des données, au moins dans une première étape, puisque c'est très certainement l'ensemble des lésions chromosomiques *et* l'interaction entre ces lésions qui

permettent d'expliquer une caractéristique clinique de la maladie.

De cette structure globale, découlent les caractéristiques mathématiques des données ainsi qu'un certain nombre de difficultés à gérer simultanément.

Difficultés techniques et statistiques soulevées par l'étude des microsatellites Les données étant catégorielles, les variables correspondant à chaque microsatellite sont modélisées par des variables binomiales $B_j(n, p)$, où n est l'effectif de l'échantillon et p la probabilité qu'un sujet soit en AI pour le microsatellite j. A cette situation simple se rajoutent les difficultés suivantes.

Le problème des manquants Il a été dit plus haut que la technique biochimique utilisée pour typer les microsatellites génère automatiquement et inévitablement une donnée manquante lorsque le microsatellite est homozygote. Ce problème est important en raison de la fréquence des homozygotes dans la population générale quel que soit le microsatellite considéré. En effet, le fait que les microsatellites choisis soient hautement polymorphes ne permet pas d'empêcher l'observation de microsatellite homozygote. Dans l'ensemble, la proportion d'homozygotes pour un microsatellite est de l'ordre de 30%, allant d'environ 5% jusqu'à 50%. Plus de détails sur les valeurs de ces proportions seront donnés dans la section des résultats. Un corollaire immédiat de cette situation est qu'il n'y a dans la plupart des jeux de données aucun sujet complet, ce qui a un impact profond sur les modèles utilisables. La totalité des méthodes statistiques est impactée par ce problème, qu'il s'agisse de méthodes univariées ou multivariées. Au moins dans une première approche, il faut utiliser des méthodes tenant compte des données manquantes pour pouvoir estimer l'impact de ces données manquantes sur les résultats et soit les supprimer si l'impact est faible ou lorsque les données sont manquantes aléatoirement, soit les inclure spécifiquement dans le modèle. Nous détaillerons ces éléments par la suite.

Le problème des dimensions de la matrice Tout d'abord, l'allélotypage (ou typage allélique) est habituellement réalisé sur un nombre important de microsatellites, important en tout cas relativement au nombre de sujets. Classiquement c'est entre 30 et 50 microsatellites qui sont typés, quelque soit le cancer étudié. Ce nombre est considéré comme élevé car le nombre de sujets est souvent limité, de l'ordre de 50 à 100, cet effectif pouvant être parfois de l'ordre d'une dizaine de sujets. Dans certains sous-groupes à analyser,

les effectifs peuvent être inférieurs à la demi-douzaine. On obtient alors des matrices où le nombre de variables est (très) supérieur au nombre de sujets. Les méthodes statistiques classiques ne permettent pas de travailler sur des matrices de ce type en raison d'un rang insuffisant de la matrice. Des méthodes de sélection de variables ou de projection sur des sous-espaces doivent alors être envisagées. La situation est donc proche de celle rencontrée dans les biopuces ou *microarrays* [345].

Le problème de la colinéarité Dans un certain nombre de cas, les biochimistes sont amenés à encadrer un gène ou une zone d'intérêt par deux ou plus de deux microsatellites. Ces microsatellites sont donc proches les uns des autres et il est possible d'observer une colinéarité entre deux ou plus de deux de ces microsatellites. La colinéarité est également une limitation classique des méthodes statistiques standards, les analyses multivariées sélectionnant des variables dites indépendantes, aux dépens d'une certaine cohérence dans les modèles [321], des variables pertinentes et explicatives pouvant être éliminées d'un modèle.

Le problème des dimensions de la matrice et le problème de la colinéarité limitent donc tous les deux l'usage de méthodes multivariées classiques tels que la régression logistique ou le modèle de Cox. Il faudra donc tenter soit de changer de modèle soit de les redéfinir de manière à ce qu'ils soient utilisables malgré ces deux limites.

Le problème des interactions Si la description des taux individuels d'AI pour les microsatellites est évidemment intéressante, il faut également envisager que ce soit surtout les interactions entre AI qui permettent d'expliquer les voies de la cancérogénèse. Exceptés certains oncogènes, la plupart des anomalies du génome ne sont pas suffisantes pour expliquer à elles seules la cancérisation d'une cellule. C'est donc dans l'association ou l'interaction entre microsatellites qu'il faut tenter de décripter les voies de la cancérogénèse. Ceci suppose donc des modèles capables d'introduire des interactions en nombre suffisant compte tenu des trois précédents (données manquantes nombreuses, dimensions de la matrice, colinéarité éventuelle).

Le problème des tests multiples Les données d'allélotypage présentent des caractéristiques proches de celles des biopuces ce qui fait qu'elles sont également confrontées au problème des tests multiples, bien que ce soit à une échelle plus petite que pour les biopuces. Le modèle à proposer devra tenir compte dans une certaine mesure de ce point technique.

Une solution simple et naturelle consiste à quitter le cadre fréquentiste pour utiliser le cadre inférentiel bayésien. Nous apporterons quelques éléments pour appuyer ce choix [322].

Limites des « modèles » actuels Dans la plupart des applications utilisant les microsatellites, les auteurs ignorent totalement le problème des homozygotes et travaillent sur chaque microsatellite de manière isolée, en univarié sur les seuls sujets dont le microsatellite étudié est informatif . Les travaux déjà cités ([58, 125, 135, 184, 194, 197, 213, 249, 271, 319, 335, 348, 353]), mais d'autres encore, utilisent ce raccourci ou ignorent dans des modèles multivariés de type analyse en cluster la présence des manquants [337]. Les homozygotes étant ignorés dans les analyses, le problème même de leur existence d'un point de vue statistique est ignoré : il n'existe, à notre connaissance, aucune publication consacrée spécifiquement à ce problème.

On trouve cependant dans la littérature quelques méthodes statistiques consacrées à des analyses portant sur les microsatellites. Newton [243] par exemple propose un modèle pour l'analyse des données de pertes alléliques mais l'analyse se porte au niveau d'un seul microsatellite : les auteurs proposent un modèle permettant de déterminer si une perte allélique observée est aléatoire, tenant compte d'un taux de perte de base et d'un modèle de sélection des allèles lié au fait que les anomalies s'accumulent et qu'elles ne peuvent être létales pour la cellule sous peine de ne pouvoir être observées. D'autres paramètres interviennent également. Cette méthode s'adresse donc plutôt à un microsatellite en particulier et ne traite pas des modifications de l'ensemble des microsatellites dans la cellule. Par ailleurs les auteurs ne traitent pas clairement le problème des homozygotes. Cependant, ce modèle peut-être utile dans une première approche et devrait idéalement être inclus dans le modèle recherché ici. Un autre modèle, proche, vise à localiser un gène suppresseur de tumeur [244]. Desai propose également un modèle basé sur des modèles de mélanges avec utilisation d'un facteur de Bayes pour déterminer le caractère aléatoire ou non d'une perte allélique [94]. Là aussi ce modèle est assez éloigné de nos préoccupations mais il mériterait éventuellement d'être inclus dans le modèle plus large qui sera proposé. D'autres modèles mériteraient également plus d'attention [336].

Slebos propose également une méthode statistique pour évaluer le déséquilibre allélique dans de petites quantités d'ADN [307]. La méthode consiste en fait à définir un seuil au delà duquel le ratio R_t/R_s indique un déséquilibre allélique. Là non plus, le problème des homozygotes n'est pas spécifiquement pris en compte dans l'incertitude supplémentaire

qu'ils apportent. Par ailleurs, la méthode ne traite pas de la description ni de l'inférence à partir de données d'allélotypage.

Un des objectifs des analyses d'allélotypage est de pouvoir prédire la présence d'un cancer à partir de cellules prélevées dans le tissu suspect dans un but de dépistage ou de surveillance thérapeutique [232]. Dans cette situation, on utilise l'ensemble des microsatellites pour définir un score dont une valeur seuil est supposée indiquer un risque accru ou diminué de lésion cancéreuse. C'est par exemple ce qui a été proposé par Panhard [260]. Le score proposé est construit sur une loi du χ^2 à partir d'une somme de lois gaussiennes tirées de chaque microsatellite individuel. Il est suggéré que ce score serait utile dans le dépistage du cancer de la vessie. Cependant, là aussi, le problème des homozygotes n'est pas pris en compte, le score étant construit uniquement sur les microsatellites informatifs pour un sujet donné. La variabilité et l'incertitude supplémentaire liées aux homozygotes ne sont pas incluses dans le score ce qui pourrait notablement modifier ses performances. Enfin, ce score est réalisé avec un objectif clairement différent de celui du présent travail.

1.4 Résumé sur le problème posé et les objectifs du modèle

L'analyse de données d'allélotypage, tant dans un objectif descriptif que dans un objectif inférentiel, suppose la réalisation de modèles capables de prendre en compte *simultanément* les quatre difficultés évoquées plus haut. Si des méthodes existent pour gérer l'un ou l'autre des problèmes, il n'existe actuellement pas de méthodes permettant de les gérer tous les quatre, notamment lorsque les données sont toutes qualitatives. L'objectif de ce travail est donc de fournir une solution générale au problème de l'analyse des données d'allélotypage, qui soit utilisable quel que soit le jeu de données et permettant donc de fournir des estimations fiables quel que soit le nombre et la simultanéité des difficultés énoncées. Le modèle devra donc fournir des résultats à la fois descriptifs et inférentiels, en prenant en considération d'une manière ou d'une autre les données manquantes, quelles que soient les dimensions de la matrice analysée et même en cas de colinéarité.

Pour toute méthode statistique, se pose le problème du choix du cadre conceptuel dans lequel devront se dérouler les analyses. La statistique est divisée en deux grandes branches depuis les origines : la statistique fréquentiste, issue des travaux de Neyman et Pearson d'une part et de Fisher d'autre part malgré leurs oppositions, et la statistique bayésienne qui tire son nom de Thomas Bayes, à l'origine du théorème éponyme. La théorie fréquentiste, si elle est très largement utilisée, présente un certain nombre de limitations que nous

nous efforcerons de décrire. La statistique bayésienne peut quant à elle s'adapter à plus grand nombre de situations avec plus de souplesse, notamment dans le cas des données manquantes. Par ailleurs, l'inférence statistique est bayésienne par essence, ce qui justifie l'utilisation de cette théorie pour analyser des données et tirer des conclusions sur des paramètres. L'analyse bayésienne ne s'est développée que récemment grâce aux progrès de l'informatique en raison du caractère intensif des calculs requis. Cependant, dans un certain nombre de situations, l'analyse fréquentiste reste utile justement en raison de la plus grande facilité à réaliser les calculs.

Après plusieurs rappels sur les notions théoriques de bases et sur l'ensemble des méthodes envisageables, nous présenterons les données, puis les résultats des méthodes classiques et des méthodes plus récentes avant de montrer leurs limites respectives. Nous proposerons enfin une méthode nouvelle permettant de répondre à l'ensemble des questions posées. Une ouverture sur un modèle plus large, englobant la totalité des contraintes biologiques connues sera faite.

1.5 Organisation du document

Ce document est organisé de la façon suivante. Après la présentation des données et de la problématique qui s'y rattache, les solutions explorées et proposées sont présentées en trois parties.

La première partie est une discussion formelle sur les fondements des deux principaux paradigmes actuellement utilisés en statistique. Le choix de l'un d'entre eux pour la suite du travail trouvera sa justification dans cette première partie.

La deuxième partie présente l'ensemble des méthodes pouvant être utilisées dans le cas des données d'allélotypage pour des analyses de type univarié uniquement. Certains points seront détaillés d'autres seulement brièvement abordés. Les résultats issus de l'utilisation des méthodes principales seront donnés dans cette même partie.

La troisième partie est consacré à l'analyse multivariée des données d'allélotypage avec une proposition d'amélioration de l'une des méthodes existantes. Les principaux résultats seront donnés.

Ces trois sections seront suivies d'une discussion et d'une conclusion.

2 Choix du cadre statistique

La plupart des analyses statistiques réalisées de nos jours se font en utilisant des tests statistiques.

La notion de test statistique est très large et ne présuppose pas une théorie en particulier. C'est une règle de décision spécifiant, à partir d'un échantillon donné, si une hypothèse peut être acceptée ou doit être rejetée à la vue de cet échantillon. Elle peut se définir comme une procédure permettant de vérifier à l'aide d'un échantillon si l'on peut ou non accepter une certaine hypothèse faite pour une population [171, 334]. Les approches fréquentistes et bayésiennes utilisent toutes les deux la notion de test statistique. En revanche la façon de réaliser ces tests diffèrent grandement d'une théorie à l'autre et même à l'intérieur de la statistique fréquentiste différentes écoles existent. Nous décrirons brièvement ici les deux principales théories fréquentistes du test statistique proposée pour l'une par Neyman & Pearson et pour l'autre par Fisher dans la première moitié du $20^{\text{ème}}$ siècle.

2.1 Le test d'hypothèse : concepts de base

Le test statistique, dans sa conception la plus large, vise à vérifier à l'aide d'un échantillon expérimental aléatoire l'hypothèse qu'un paramètre théorique de la population d'intérêt se trouve ou non dans une plage de valeurs attendues. La plage de valeurs peut éventuellement être réduite à une valeur unique. Il confronte donc un certain nombre d'hypothèses entre lesquelles il faudra choisir. Par la suite, nous considérerons qu'il n'y a que deux hypothèses à comparer. Avec cette simplification, un test statistique consiste en une subdivision de l'ensemble des échantillons de taille n en deux sous-ensembles mutuellement exclusifs associée à la règle définissant l'acceptation ou le rejet d'une hypothèse si l'échantillon observé se trouve dans l'un ou l'autre des sous-ensembles [334]. Par ailleurs, nous nous placerons essentiellement dans le cadre de la recherche médicale et biologique, ce qui a son importance sur la formulation des hypothèses.

2.1.1 La position de Neyman-Pearson

La théorie des tests d'hypothèse de Neyman-Pearson consiste en la formulation de deux hypothèses H_A et H_B. La première hypothèse est souvent appelée hypothèse nulle tandis que l'autre est généralement appelé hypothèse alternative [246, 247, 248, 312, 313]. Le cadre proposé permet de décider, en situation d'incertitude, quelle hypothèse doit être retenue.

En choisissant à l'issue du test une des deux hypothèses, on peut se tromper de deux façons : soit rejeter à tort l'hypothèse nulle soit accepter à tort l'hypothèse alternative. Pour déterminer la bonne hypothèse avec le plus petit risque d'erreur possible, Neyman & Pearson proposent la procédure suivante qui implique une planification expérimentale soigneuse :

- on définit les deux hypothèses nulle et alternative à comparer ;
- on divise l'espace d'échantillonnage en deux régions dénommées région critique (W) et région d'acceptation ($\bar{W}$) ;
- on utilise une statistique fournissant une valeur t pour l'échantillon ;
- on choisit un risque α de rejet à tort de l'hypothèse nulle tel que

$$\Pr\{t \in W | H_0\} = \alpha$$

- la valeur de α définit donc la zone de rejet W de l'hypothèse nulle ainsi que la région d'acceptation de l'hypothèse alternative par complément par rapport à l'espace des échantillons ;
- on choisit un risque β de non acceptation de l'hypothèse alternative (ce qui définit les effectifs) ;
- parmi toutes les statistiques possibles répondant à la définition de α, on choisit la plus puissante ;
- sur l'échantillon, on calcule la statistique et si la valeur observée tombe dans la zone de rejet on rejette l'hypothèse nulle, sinon, on l'accepte.

Le risque α est fixé à l'avance et il est donc indépendant des données [152]. Les valeurs de α et de β sont choisies en fonction du coût des erreurs. Le choix de β est un peu plus complexe car pour un α donné, β est une fonction de α et la taille d'échantillon à utiliser détermine la valeur de β.

Il ne faut pas confondre ici la statistique (de test), fonction des données et le test statistique qui est également une variable aléatoire mais qui prend deux valeurs seulement, selon que l'on conserve ou rejette l'hypothèse nulle.

Le terme de test d'hypothèse est généralement utilisé pour décrire cette procédure. Le terme plus courant de test d'hypothèse nulle (TNH) sera cependant retenu par la suite pour décrire cette procédure, même si l'adjectif « nulle » suppose que le paramètre définissant l'hypothèse nulle prend justement une valeur nulle ce qui n'est pourtant pas forcément le cas dans le cadre d'un test statistique quel qu'il soit. En pratique, c'est pourtant l'hypothèse

testée le plus fréquemment.

Cette procédure est typiquement fréquentiste. Supposons que l'on dispose d'un grand nombre d'échantillons, disons M, de taille n. Supposons que pour chaque échantillon de M on rejette l'hypothèse nulle si l'échantillon est dans la zone W et que l'on accepte cette hypothèse si l'échantillon n'est pas dans cette zone. On prend donc M décisions dont un certain nombre seront fausses. Si l'hypothèse nulle est vraie et que M est grand, la proportion de cas où l'on rejette cette hypothèse nulle à tort est de $\alpha\%$. Si l'hypothèse alternative est vraie, la proportion de conclusions fausses la concernant sera de $\beta\%$. Ces propriétées sont des propriétés au long cours supposant un échantillonnage répétitif [334].

2.1.2 La position Fisherienne

Pour Fisher, il n'y a qu'une seule hypothèse à poser, l'hypothèse nulle. On calcule alors sous cette hypothèse nulle la probabilité d'observer une différence aussi importante que celle que l'on a observée [109, 110, 111, 313]. Classiquement désignée par la lettre p, cette probabilité s'écrit : $p = \Pr(T \geq t_{obs}|HN)$, où $T = f(Y)$ est une statistique de test, fonction des données Y et t_{obs} est la valeur observée du test, obtenue *a posteriori* à partir des données observées. Selon Fisher, la valeur de p indique alors le niveau de confiance attribuable à l'hypothèse nulle. On peut aussi présenter p en disant que c'est une mesure de la différence entre les données et l'hypothèse nulle : la différence s'accroît à mesure que p diminue [312]. Soulignons à nouveau que la position de Fisher n'implique pas la formulation d'une hypothèse alternative. Fisher considère le p comme un indice informel de l'écart entre l'hypothèse nulle et les données. La valeur de p ne présente pas de seuil particulier même si lorsque la valeur de p est inférieure à 5%, Fisher la considère comme étant une preuve forte contre l'hypothèse nulle [110]. Fisher voyait le procédé inductif comme quelque chose de fluide, non fixé et non formel [123, 127, 129]. Plutôt que de parler de test d'hypothèse, on parle ici d'un test de significativité (TS).

Un exemple d'application du test statistique consiste à déterminer si deux populations diffèrent ou non l'une de l'autre pour un paramètre θ, chaque population ayant une valeur donnée de θ, par exemple θ_A et θ_B. La question est donc de savoir si la valeur de θ est ou n'est pas la même dans les deux populations desquelles sont extraits des échantillons utilisés pour effectuer cette vérification. Pour réaliser ce test, la conception neymanienne pose une hypothèse dite nulle (HN) qui spécifie que, *a priori*, les deux séries de valeurs ne diffèrent pas par le paramètre θ d'intérêt : $\theta_A = \theta_B$. Le cadre conceptuel de Fisher ne

pose pas d'autre hypothèse. En revanche, dans la conception de Neyman & Pearson, une hypothèse alternative (HA), complément logique de l'HN, spécifie qu'il existe une différence entre ces deux séries de valeurs et que $\theta_A \neq \theta_B$. La comparaison portant sur θ_A et θ_B est faite *via* l'estimation d'une statistique combinant ces deux paramètres observés sur l'échantillon correspondant. Il s'agit donc de réaliser une induction sur les paramètres, c'est-à-dire d'obtenir, à partir d'une observation résultant d'un phénomène, une information sur ce phénomène. La formulation donnée ici est la formulation la plus couramment utilisée, même si elle n'est pas toujours la plus pertinente.

2.1.3 Oppositions entre les deux approches

Une première opposition entre Neyman & Pearson et Fisher concerne l'aspect décisionnel du THN. Le paradigme de Neyman & Pearson est une approche décisionnelle (*decision based approach*) [248, 246, 247, 312]. Dans le concept de Neyman & Pearson, il y a deux risques d'erreur : rejeter HN quand elle est vraie et accepter HN quand elle est fausse. Fisher quant à lui ne s'intéresse qu'à la première de ces deux erreurs. Le test statistique proposé par Neyman & Pearson vise à se tromper le moins souvent possible lors de l'acceptation ou le rejet de l'hypothèse nulle. Par conséquent, pour Neyman & Pearson, un test statistique ne permet pas de savoir si une hypothèse est juste ou fausse mais seulement de *décider* quelle hypothèse est juste ou fausse en se trompant le moins souvent possible. Une autre divergence entre Neyman & Pearson et Fisher porte sur la notion de niveau de signification. Pour Neyman & Pearson, le niveau de signification est une propriété du test alors que pour Fisher, p ne peut dériver que des données [152] par opposition avec la valeur de α, et le niveau de signification est une propriété des données, c'est-à-dire du lien entre les données observées et une théorie [123]. Pour Fisher, si $p < \alpha$, cela ne mène pas à rejeter l'HN mais à mener une autre expérience [312]. Cependant une telle comparaison n'a pas lieu d'être puisqu'elle compare les deux concepts.

Le raisonnement de Neyman & Pearson s'articule autour des valeurs de α et β préalablement fixées pour minimiser le risque de se tromper lors du choix de l'une des deux alternatives. Mais cette mécanique, dépourvue de quantification de la véracité de l'hypothèse nulle et de l'hypothèse alternative (probabilité *a priori* des hypothèses), ne donne pas d'information sur la véracité de ces hypothèses. Elle présuppose que l'une des deux alternatives est obligatoirement juste tandis que l'autre est obligatoirement fausse et le test statistique de Neyman & Pearson permet alors « seulement » de trouver celle qui est vraie

(et donc celle qui est fausse) en se trompant le moins souvent possible. La possibilité de connaître et éventuellement de modifier la probabilité d'une hypothèse est pourtant fondamentale, surtout lorsque l'on dispose de peu de connaissance ou de plusieurs théories en compétition.

La conception de Neyman & Pearson s'applique « au long cours », sur un grand nombre d'épreuves similaires, en raison des propriétés fréquentistes du test. Par ailleurs, α représente en fait la probabilité d'un ensemble de valeurs possibles pour p, obtenues sur un grand nombre d'épreuves, et donc avant que les épreuves ne soient réalisées : α est une zone de rejet, définie *a priori* et c'est pour cela qu'il ne donne pas la même indication que p qui dépend lui d'une seule épreuve qui le détermine entièrement. La valeur de p n'est connue qu'*a posteriori*. Une comparaison entre p et α n'a donc pas de sens. L'habitude très répandue consistant à mélanger les deux indices (en faisant les calculs fixant une limite de significativité à $\alpha\%$ mais en donnant la valeur exacte de p) est donc fausse et source de nombreuses confusions [109, 129].

2.2 Les critiques de la théorie du test d'hypothèse

Le test statistique est devenu depuis longtemps un élément incontournable de la recherche médicale. S'il est très répandu, le test statistique fréquentiste d'hypothèse (selon Neyman & Pearson ou selon Fisher) a néanmoins souvent subi de multiples critiques vis-à-vis de son principe de fonctionnement théorique [18, 1, 75, 162, 177, 180, 205, 349] et plusieurs centaines d'articles critiquant l'utilisation des tests statistiques d'hypothèses ont été publiés dont les plus importants sont [18, 66, 75, 101, 123, 129, 332, 152, 162, 193, 205, 227, 312, 313, 2]. Ces critiques peuvent néanmoins être regroupées en quelques points principaux.

Le THN et le TS ne répondent pas à la question d'intérêt L'argument principal de la critique est que le THN et le TS ne répondent pas à la question qui est scientifiquement la plus pertinente : quelle est la probabilité que le traitement à l'épreuve soit efficace ? Lorsque l'hypothèse nulle est de type $\delta = 0$, le test statistique classique permet de savoir s'il y a une différence non nulle entre deux traitements ce qui répond donc très partiellement à la question, mais qui contrairement à une idée encore répandue dans la littérature médicale (clinique ou épidémiologique), ne permet pas de quantifier la probabilité que le traitement soit efficace. Globalement, le THN et le p ne nous disent pas ce que l'on veut savoir. En effet,

le p quantifie la probabilité d'observer sous l'hypothèse nulle des données D aussi extrêmes que celles obtenues d, soit $\Pr(D \geq d|HN)$ alors que ce qui nous intéresse est la probabilité des hypothèses en fonction des données observées (et uniquement celles-ci), soit $\Pr(HN|d)$ [75, 136, 162, 180]. Le THN nous dit quelle hypothèse conserver mais il ne nous dit pas quelle est la probabilité que cette hypothèse soit vraie. L'obtention de ce résultat suppose l'utilisation des méthodes bayésiennes. Pour que le THN permette d'évaluer la probabilité qu'une hypothèse soit vraie, il faudrait introduire des notions de probabilités *a priori* de cette hypothèse.

La valeur de p résulte de la combinaison de la taille de l'effet et de la taille de l'échantillon La valeur de p est proportionnelle à la taille de l'effet et à la puissance de l'essai (c'est-à-dire aux effectifs). Un effet de taille importante sur un petit effectif et un effet d'importance limitée sur un grand échantillon peuvent donner des p de même ordre de grandeur. Par conséquent, la seule valeur de p ne permet pas de distinguer l'une de l'autre ces deux situations et il ne renseigne donc pas sur l'effet clinique [47, 75, 162, 176, 205, 311]. Pour certains, l'analyse doit être décomposée en deux parties : d'abord réaliser le test et en cas de rejet de l'hypothèse nulle (notamment d'absence de différence), procéder à l'estimation de cette différence [171]. On peut donc toujours faire en sorte d'accepter l'hypothèse alternative en augmentant les effectifs, ce qui rend l'hypothèse nulle inutile. Boren [47] souligne un nombre important de confusions résultant de ce double aspect de p. La même critique peut être porté à l'encontre du THN.

L'hypothèse nulle n'a pas de sens la plupart du temps Dans les situations courantes de la recherche médicale, il n'est pas raisonnable de s'attendre à ce que deux groupes de sujets aient exactement la même valeur pour un paramètre donné. Une hypothèse nulle au sens stricte du terme n'a donc en général pas de sens (ce qui n'est pas forcément vrai ailleurs, par exemple dans le domaine du contrôle industriel de procédés pour vérifier qu'une pièce vient bien d'un lot de pièces similaires). Par exemple, il n'y a aucune raison que la proportion de fumeurs soit exactement la même chez les hommes et chez les femmes, même si ces proportions ne sont pas forcément très éloignées. Il existe, en pratique, toujours une petite différence entre deux populations à comparer et la plupart du temps l'HN peut-être reconnue comme fausse avant même que l'expérience ne soit lancée et que le test ne soit effectué. A l'extrême, il est inutile de recueillir des données pour tester l'HN puisque celle

ci est fausse d'emblée [253]. L'HN est donc le plus souvent sans intérêt d'autant plus que les HN sont très souvent testées de manière très précise ($\theta = 0$) ce qui en général n'a pas lieu d'être [47, 75, 180, 205, 327].

L'hypothèse nulle est fausse *a priori* Dans le même ordre d'idée, la plupart des hypothèses nulles sont fausses *a priori* [162, 205, 245]. Si un nouveau produit est testé dans un essai thérapeutique, la plupart du temps ce traitement a déjà été étudié de nombreuses fois et s'il arrive au stade de l'essai de phase II et *a fortiori* de phase III, c'est que l'on sait déjà qu'il a de forte chance d'avoir une efficacité supérieure au placebo. Par ailleurs, même si un nouveau traitement est inefficace *i.e.* pas plus efficace que le traitement de référence, il est quasi impossible que la différence d'effet observée soit exactement nulle. Il est donc peu pertinent de tester son effet en partant du principe qu'il n'a aucun effet, hypothèse pourtant la plus fréquemment testée.

Les tests utilisent des données non-observées Le THN et le TS sont basés sur des données observées mais aussi sur des données non observées, celles donnant une différence au moins aussi importante que celle effectivement observée. Soit z une statistique de test. Ainsi, la probabilité donnée par le TS est la probabilité que la valeur de z au seuil α dépassent la valeur observée z_{obs} de z. C'est donc : $\Pr(|z_\alpha| > z_{obs})$. Le p est donc aussi une probabilité cumulée sous l'HN [205]. En conséquence de quoi, la détermination de la zone de rejet W inclut des valeurs non-observées. En faisant intervenir des résultats qui n'ont pas été observés, la valeur de p exagère l'ampleur du rejet de l'hypothèse nulle [35, 75, 127]. Il en va de même pour α.

De ce fait, et en raison de l'incorporation de données non observées, le p viole le principe de vraisemblance [81, 255].

Le seuil α est arbitraire Au risque d'énoncer une évidence, rappelons que dans la conception de Neyman & Pearson la valeur seuil de décision, α est arbitraire et ne repose sur aucune base théorique [314]. Ceci pose inévitablement le problème de conclusions radicalement contraires selon que $p = 0{,}051$ ou $0{,}049$. Ce qui choque dans ce constat n'est pas tant le fait de fixer un seuil arbitraire, car l'arbitraire est inhérent à la recherche qui contient toujours une certaine part de subjectivité. Ce qui choque est que le seuil soit toujours le même : 5%. Ce seuil est le même, non seulement dans tous les domaines de la médecine,

mais aussi dans tous les domaines scientifiques ou presque. Les valeurs des risques α et β représente des coûts liés à des erreurs dans la décision. Il n'est pas concevable que ce coût soit systématiquement le même dans tous les domaines scientifiques et toutes les situations, ce qui montre bien une erreur quasi permanente dans l'utilisation de ces tests.

> Le plus étonnant est que les scientifiques, qui ne sont d'accord sur rien soient d'accord sur la chose la plus arbitraire qui soit [312].

Sterne rappelle que la division entre significatif et non significatif n'était pas dans les intentions des fondateurs des tests [312]. Cette division résulte d'une confusion très courante entre p et α.

Les tests multiples et la correction de α Afin d'assurer le maintien du risque α sur un nombre de tests supérieur ou égal à deux, le test selon Neyman & Pearson impose des procédures de correction du risque α. La difficulté de réaliser ces corrections est grande et actuellement aucune solution claire ne semble se dégager malgré la grande diversité des approches. Un exemple de controverse peut être trouvé dans [258] et [84]. Proschan propose une intéressante synthèse mettant en exergue les difficultés liées aux tests multiples [272]. D'autres difficultés sont présentées par Thomas ou Altman [16, 322].

2.3 Les erreurs d'interprétation

Au delà des limites théoriques de ces deux procédures statistiques, l'utilisation du test selon Neyman & Pearson ou selon Fisher n'est pas optimale car des erreurs d'interprétation sont très fréquentes. Ces erreurs d'interprétation s'expliquent entre autres par l'existence de deux approches fondamentalement différentes malgré leurs similitudes apparences, combinées à un certain nombre de limites inhérentes à chacune de ces deux approches.

Des erreurs d'interprétation se trouvent même dans des publications scientifiques même de haut niveau, comme par exemple [162, 168, 205, 313] :

- confondre le p avec la probabilité de l'hypothèse nulle ;
- considérer p comme la probabilité de l'HN ou encore comme la probabilité que l'HN soit correcte ;
- admettre que p est censé mesurer la force de la preuve à l'encontre de l'HN ;
- considérer l'absence de significativité comme la preuve de la véracité de l'HN [180] ;
- dire de p qu'il représente la probabilité que les résultats soient dus au hasard ;

Une autre liste d'erreurs d'interprétation est donnée dans un article de Gigerenzer [123].

Que le p soit fréquemment mal interprété n'est évidemment pas imputable à la théorie qui lui a donné naissance, mais cela en limite tout de même la pertinence et suggère de le remplacer par des indices de compréhension plus immédiate pour des personnes qui utilisent les statistiques sans les connaître.

Une autre erreur d'interprétation fréquente consiste à confondre la valeur de p et la probabilité de retrouver le même résultat lors d'une répétition ultérieure de l'expérience [127, 180, 297]. Ces deux valeurs sont en réalité souvent assez éloignées l'une de l'autre. On peut montrer qu'environ 50% des $p < 0,05$ sont issues d'HN qui en réalité sont vraies [127, 146, 206, 270, 313, 316, 329].

2.4 Les arguments en faveur du test d'hypothèse nulle

Malgré ces limitations, un certain nombre d'auteurs défendent néanmoins le THN : Frick argumente sa position en disant que la p-valeur peut donner des arguments d'ordre à défaut de donner des arguments quantitatifs pour ou contre une hypothèse [118]. Un autre de ces arguments consiste à dire qu'une expérience n'a pas toujours pour but de quantifier un résultat. Il reconnaît cependant certains défauts du THN.

Wein pense qu'il peut être fait un usage raisonné des p-valeurs et du THN, celui-ci permettant de prendre des décisions en raison du seuil de signification, ce que ne permettrait pas d'après lui les autres méthodes, mettant en relief le côté décisionnel du test selon Neyman & Pearson [338].

Stephens propose trois arguments en faveur de l'HN [311] : (1) il suggère de la voir comme le complément logique de l'HA alors qu'habituellement, on part de l'HN pour poser l'HA comme complément logique. (2) L'HN doit servir à définir un cadre conceptuel pour la taille d'effet et lui servir de borne de repère, de valeur de référence. (3) Enfin, Stephens rappelle que l'HN est plutôt une hypothèse à « nullifier » suivant l'expression de Fisher, dans le sens où la valeur du paramètre dans l'HN n'est pas forcément nulle. Par ailleurs, il estime que le THN, s'il peut avoir des aspects utiles dans certains cas, est d'un intérêt moindre qu'une approche basée sur la théorie de l'information et qu'un bon choix pourrait être un mélange du THN et de la théorie de l'information.

Les sciences humaines ne sont pas en reste et certains auteurs y défendent également le THN [136]. Enfin, certains auteurs contribuent indirectement au débat en étudiant les propriétés de la p-valeurs sous l'hypothèse alternative [154].

2.5 Les solutions possibles

De très nombreux auteurs, surtout des statisticiens ont dénoncé la vacuité du p et du THN, diverses solutions alternatives à leur utilisation ont été proposées, notamment dans la recherche biomédicale [18, 19, 39, 47, 75, 140, 176, 177, 190, 205, 208, 268, 296, 311, 312, 327, 326, 349].

Plutôt qu'une alternative aux THN, certains suggèrent d'utiliser les meilleures composantes de chacun des deux paradigmes fréquentiste et bayésien, en utilisant les quelques ponts théoriques permettant de les relier [36, 37, 38].

L'utilisation des intervalles de confiances (IC) est intéressante dans le principe car l'IC d'un paramètre quantifie l'incertitude liée à l'estimation du paramètre. Ces intervalles de confiances sont des alternatives utiles au p ce qui est reconnu même par les fréquentistes [210]. Mais en pratique, l'IC partage un certain nombre des problèmes du p [107, 129]. Ainsi, l'IC est une autre expression du THN et d'un point de vue formel, il est l'équivalent d'un test statistique. De plus, comme le THN, l'IC est souvent mal interprété : on lui donne une lecture typique de la théorie bayésienne même lorsqu'il est issu d'une estimation fréquentiste [75]. Enfin, il ne tient pas compte des probabilités *a priori* des données.

Les limites intrinsèques des théories de Neyman & Pearson ainsi que les très nombreuses erreurs d'interprétation qui découlent de leur mélange erroné plaident en faveur de la grande alternative au test fréquentiste qu'est le paradigme bayésien [195]. Ce concept qui permet de faire disparaitre un grand nombre des difficultés du THN, ne rencontre pourtant encore qu'un faible écho, malgré une popularité récemment croissante.

2.6 La théorie bayésienne

2.6.1 Cadre général de la théorie bayésienne

Dans l'approche fréquentiste, le paramètre θ pour lequel on souhaite réaliser une inférence est considéré comme une constante inconnue. Ce paramètre peut prendre n'importe quelle valeur dans l'espace du paramètre et les données seules peuvent donner une indication sur la valeur vraie de θ sans jamais néanmoins donner cette vraie valeur exactement. Les méthodes fréquentistes, basées sur les propriétés des échantillons aléatoires, permettent d'obtenir donc une estimation ponctuelle ou par intervalle de la vraie valeur de θ. Le paramètre est fixe, les données sont aléatoires. Dans les méthodes bayésiennes [81], on considère que le paramètre θ est issu d'une densité de probabilité $f_\Theta(\theta)$ dont la forme est précisée *a*

priori de même que les paramètres de cette densité de probabilité. En revanche, les données sont considérées comme fixes. Les données, combinées à cette information *a priori* vont servir à préciser la densité de probabilité de θ. Cette combinaison d'information se résume dans une densité de probabilité dite *a posteriori*. Cette densité de probabilité est donc conditionnelle aux données D. Cette formulation permet par ailleurs d'obtenir une distribution prédictive pour de nouvelles valeurs de Y, conditionnellement à la distribution *a posteriori* de θ. Ces résultats sont des conséquences du théorème de Bayes.

2.6.2 Le théorème de Bayes

La théorie bayésienne [121, 188, 255, 264] est née du théorème de Bayes, théorème éponyme du révérend père Thomas Bayes (1702-1761) qui publia de façon posthume sa formule dans le « *Essay Towards Solving a Problem in the Doctrine of Chances* », en 1763 dans les *Philosophical Transactions of the Royal Society* de Londres.

Le théorème de Bayes, encore appelé théorème des probabilités inverses par Laplace qui l'a redécouvert indépendamment de Bayes, est basé sur les probabilités conditionnelles et permet d'établir la probabilité d'une hypothèse parmi d'autres à partir de données observées. Soient k hypothèses A_j susceptibles d'expliquer une série de données D. La probabilité *a posteriori* de l'hypothèse A_j est la suivante :

$$Pr(A_j|D) = \frac{Pr(D|A_j) \cdot \Pr(A_j)}{\sum_{j=1}^{k} Pr(D|A_j) \cdot \Pr(A_j)}$$

Cette formulation est intéressante pour des hypothèses disjointes et en compétition, mais le plus souvent les « hypothèses » sont des variables aléatoires et il faut reformuler la formule précédente pour obtenir, pour une variable aléatoire Θ et des données X :

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{p(x)}$$

or

$$p(x) = \int p(x|\theta) \cdot p(\theta) d\theta$$

donc

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x|\theta) \cdot p(\theta) d\theta}$$

qui est l'expression complète formelle du théorème de Bayes (TB) pour une variable aléatoire continue Θ. On peut donc simplifier l'expression précédente et noter le théorème de la façon suivante :

$$p(\theta|x) \propto p(x|\theta) \cdot p(\theta)$$

Dans les expressions ci-dessus, $p(\theta)$ est la probabilité *a priori* de θ, $p(x|\theta)$ est la fonction de vraisemblance de θ et $p(\theta|x)$ est la probabilité *a posteriori* de θ. On voit donc que la densité de probabilité de θ a été mise à jour par les données *via* la fonction de vraisemblance. Dans le contexte bayésien, les données servent donc à moduler l'information dont on dispose sur la probabilité d'un paramètre d'intérêt. On peut juger de l'apport d'information apporté par les données en utilisant le facteur de Bayes qui est le rapport des ratios des probabilités *a priori* de deux hypothèses sur le rapport des probabilités *a posteriori*. Si l'on souhaite comparer deux hypothèses HN et HA dites respectivement hypothèses nulles et alternatives telles que :

$$HN : \theta \in \Theta_0$$

et

$$HA : \theta \in \Theta_1$$

On note π_0 et π_1 les probabilités *a priori* de ces deux hypothèses, exclusives l'une l'autre de sorte que $\pi_0 = 1 - \pi_1$. On calcule ensuite *via* le TB la valeur des probabilités *a posteriori* de chaque hypothèse, notées p_0 et p_1. Le facteur de Bayes FB vaut par définition :

$$FB = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0\pi_1}{p_1\pi_0}$$

Cette interprétation est simple lorsque les hypothèses comparées sont simples, c'est-à-dire lorsque :

$$\Theta_0 = \{\theta_0\} \text{ et } \Theta_1 = \{\theta_1\}$$

En se rappelant que $p_0 \propto \pi_0 p(x|\theta_0)$ et $p_1 \propto \pi_1 p(x|\theta_1)$, on peut remarquer que :

$$\frac{p_0}{p_1} = \frac{\pi_0}{\pi_1}\frac{p(x|\theta_0)}{p(x|\theta_1)}$$

et donc :

$$FB = \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

Donc le FB est le rapport de vraisemblance de HN contre HA. Dans le cas des hypothèses composites, l'interprétation est plus délicate car il faut incorporer des lois *a priori* dans le calcul du FB lequel ne peut plus être vu comme un simple rapport de vraisemblance. Cet aspect limite l'usage du FB à des cas simples et même à des situations ou seules deux hypothèses alternatives non composées sont en compétition [169]. On lui préfère donc fréquemment d'autres méthodes ou indices tels que le Bayesian Information Criterion (BIC) ou Deviance Information Criterion (DIC) lesquels peuvent être calculés numériquement par des logiciels. Citons encore le cas particulier de la comparaison d'une hypothèse simple et d'une hypothèse composite qui essaye de mimer la situation des tests fréquentistes. On compare une hypothèse ponctuelle (qui est à rejeter dans le cas fréquentiste) avec une hypothèse composite qui spécifie uniquement que le paramètre estimé n'est pas dans l'ensemble de référence de l'hypothèse ponctuelle. On a donc :

$$HN : \Theta_0 = \{\theta_0\}$$

et

$$HA : \theta_1 \in \Theta_1$$

La situation est décrite entre autre par Berger [35] et Spiegelhalter [309].

La théorie bayésienne suppose que l'on connaisse donc la densité de probabilité *a priori* du paramètre θ. En l'absence d'information, on choisit une loi uniforme ou quasi uniforme. En présence d'information, des procédures d'élicitation, décrites par exemple dans [166, 256] et de façon générale dans [254], permettent de déterminer les paramètres des lois *a priori*.

2.7 Intérêt de la statistique bayésienne dans le domaine biomédical

Par rapport à la statistique fréquentiste, la statistique bayésienne présente un certain nombre d'avantages qui découlent de ses propriétés.

2.7.1 L'absence d'hypothèse nulle

La conception bayésienne est purement probabiliste ce qui implique que la connaissance d'un paramètre θ est représentée par une densité de probabilité $p(\theta)$. Pour une distribution

classique telle que la loi de Gauss, il n'est pas pertinent de calculer la probabilité que la valeur de θ prenne une valeur donnée θ_0 car cette probabilité est nulle : $\Pr(\theta = \theta_0) = 0$ [123, 188, 254]. La situation peut être différente pour des distributions particulières basées sur des lois de Dirac ou des mélanges de loi incluant une loi de Dirac mais cet exemple est artificiel dans la mesure où il ne se rencontre jamais dans le domaine médical. Le fait de poser une hypothèse nulle de la même manière que pour un fréquentiste n'est donc pas pertinent pour un bayésien. La formulation des hypothèses se fait d'une manière équivalente à l'approche unilatérale des tests fréquentistes. Le bayésien pose donc des hypothèses du type suivant : $\Pr(\theta > \theta_0)$. Dans cette situation, la probabilité *a posteriori*, pour une loi *a priori* non informative, correspond à la p-valeur unilatérale du test.

Cette façon de faire est la plus simple et la plus classique. Cependant, un certain nombre d'auteurs proposent, comme cela a déjà été dit plus haut, de s'approcher de la formulation fréquentiste des tests en utilisant pour l'hypothèse nulle un mélange entre une loi *a priori* de type habituelle (loi de Gauss, loi Beta etc., selon la situation) et une densité de probabilité ponctuelle formant une masse *a priori* pour $\theta = \theta_0$ (hypothèse nulle « ponctuelle »). Pour tester la différence entre deux proportions p_1 et p_2, cela revient à placer une loi uniforme sur l'espace où $\theta_1 = \theta_2$. Cela a été proposé et étudié par Berger [35, 36, 37, 38, 39] lequel finit par conclure que les deux approches ne peuvent être réconciliées : la p-valeur est une mauvaise mesure si l'on cherche à s'en servir pour tester une hypothèse, qu'elle soit ponctuelle ou non. Une solution alternative proche mais complètement bayésienne consiste à remplacer la distribution ponctuelle par une distribution très pointue ce qui permet de se situer en pratique autour de la valeur de référence que l'on souhaite tester sans tomber dans l'écueil d'une impossibilité mathématique. Précisons que certains auteurs ([40]) utilisent à des fins didactiques cette méthode mais sans clairement l'identifier ce qui peut poser problème d'un point de vue pédagogique. Enfin, Spiegelhalter l'illustre dans le cas d'un essai thérapeutique [309].

2.7.2 L'absence de seuil α

Le seuil α intervient dans le test fréquentiste en raison de son rôle dans le processus décisionnel de Neyman & Pearson. Ce seuil est défini de manière arbitraire et une fois fixé, il permet d'assurer que pour une puissance donnée, la procédure de test utilisée aura un risque de rejet à tort de l'hypothèse nulle inférieur ou égal à α. L'arbitraire de ce seuil est donc un contre-argument à la critique de fréquentiste vis-à-vis de l'aspect subjectif de la

distribution *a priori* de la statistique bayésienne. Il n'est pas moins subjectif de choisir un seuil α que de choisir une distribution *a priori*. Ce seuil n'existe pas dans le cadre bayésien puisqu'il n'y a pas à faire le choix entre deux hypothèses concurrentes. Les deux hypothèses en compétition subsistent à la fin de la confrontation.

2.7.3 L'absence de p-valeurs

Un des arguments majeurs utilisé à l'encontre de la statistique fréquentiste est l'utilisation de p-valeurs en lieu et place d'une inférence sur les paramètres d'intérêt. Comme cela a été dit plus haut (voir notamment Cohen [75]), la p-valeur représente $\Pr(X|\theta)$ et non pas $\Pr(\theta|X)$ qui est en fait l'information importante. Le paradigme bayésien fournit cette valeur $\Pr(\theta|X)$ et ne fait quasiment aucun usage des p-valeurs.

2.7.4 Les tests multiples

Dans la conception bayésienne, toute inconnue est représentée par une densité de probabilité. On peut donc placer une densité de probabilité sur tous les paramètres d'un problème statistique. On peut donc placer une densité de probabilité sur des comparaisons entre moyennes. Les différentes moyennes correspondantes aux différents groupes comparés dans une expérimentation sont considérées alors comme étant issues d'une distribution d'ordre supérieur. Cette distribution est dotée d'une densité de probabilité à partir de laquelle on peut calculer les paramètres classiques tels que la probabilité qu'une moyenne m_i d'un groupe i soit supérieure à une moyenne m_j d'un groupe j, $i \neq j$. Les modèles sont soit des modèles à effets fixes soit des modèles à effets aléatoires selon la situation.

2.7.5 La quantification directe de l'effet du traitement

Il a été dit que dans la conception bayésienne, tout paramètre est doté d'une densité de probabilité. Ainsi un risque relatif ou un odds-ratio est doté d'une densité de probabilité *a priori* qui sera modifié par les données observées. On obtient alors, après mise à jour de l'*a priori* par les données, la distribution *a posteriori* qui permet de calculer directement la probabilité que le paramètre d'intérêt soit supérieur à un seuil donné, par exemple qu'un risque relatif soit supérieur à 1 ou encore la probabilité qu'un odds-ratio soit compris entre deux limites l_1 et l_2 données. La méthode bayésienne modélise directement la densité de probabilité de la taille d'effet comme le RR [138]. L'effet du traitement est donc directement quantifié et ne passe pas par l'intermédiaire d'un test statistique dont la valeur n'est en

général pas indicative de la taille de l'effet. L'inférence sur la taille d'effet est donc beaucoup plus directe. Par ailleurs, cela suppose que pour pouvoir résumer de manière commode cette distribution on utilise des intervalles de confiance (intervalle de crédibilité dans le vocabulaire bayésien usuel).

A contrario, le test t de Student ne renseigne pas sur la différence des moyennes observées bien que cette différence soit incluse dans son calcul. Il faut calculer séparément la valeur de la différence ainsi que son intervalle de confiance. De même, le test d'un risque relatif repose sur le test du χ^2 de la table associée ou par des modèles tel que le modèle de Cox et non pas sur une analyse directe de la valeur du risque. En fréquentiste, le raisonnement tenu est souvent le suivant : si le test est significatif, alors on s'intéresse à la valeur du risque et à son intervalle de confiance. Même si ce raisonnement est fautif dans la mesure où l'on peut s'intéresser à l'intervalle de confiance d'un paramètre quelle que soit la significativité du test associé, c'est néanmoins une attitude très fréquemment rencontrée dans la littérature bio-médicale.

2.7.6 La confrontation des hypothèses n'en élimine aucune

Le fait que la méthode bayésienne exprime toutes les inconnues *via* une densité de probabilité fait que les différentes hypothèses en compétition ne sont jamais totalement éliminées. Dans la mécanique fréquentiste, à partir du moment où l'on rejette une hypothèse nulle par un test significatif, on admet définitivement que cette hypothèse nulle est fausse. Ceci amène évidemment des difficultés d'interprétation quand deux expérimentations identiques aboutissent l'une au rejet de l'hypothèse nulle et l'autre à l'acceptation (temporaire au moins) de celle-ci. La comparaison de deux hypothèses par un facteur de Bayes n'élimine pas directement l'hypothèse qui n'est pas soutenue par les données : elle est dotée d'une probabilité faible indiquant qu'elle n'est sans doute pas celle qui explique au mieux les données. Par ailleurs, la combinaison des données des deux expériences est tout à fait naturelle dans le cadre bayésien, ce qui aboutit à une inférence unique.

2.7.7 L'utilisation de connaissances antérieures

La théorie bayésienne reposant essentiellement sur la notion de loi *a priori* mise à jour par des données observées, et tenant compte du fait qu'une probabilité peut traduire un sentiment subjectif, il est facile d'introduire dans cette loi *a priori* l'avis d'experts ou des données recueillies antérieurement pour préciser la densité de probabilité *a posteriori* des

paramètres. Ceci permet de façon manifeste une accumulation des connaissances. La difficulté de définition d'une loi *a priori* à partir de l'avis d'experts est largement décrite dans l'ouvrage de O'Hagan [166, 254] . Cette définition de la loi *a priori* s'appelle l'élicitation. En linguistique, l'élicitation est l'incitation d'un locuteur à un autre à statuer sur différentes hypothèses, c'est-à-dire à introduire chez lui le recours à sa compétence/performance. On recourt à des stratégies pour connaître la réaction des locuteurs. (définition de Wikipedia.org, site accédé le 16/11/2006)

La notion bayésienne d'*a priori* est la cible la plus fréquente des attaques fréquentistes contre la théorie bayésienne ce qui est étonnant puisque le fréquentiste fait également usage d'information *a priori*. De plus, face à un problème scientifique, un chercheur est rarement dans l'inconnu complet [312]. Préalablement à toute étude, un auteur établit toujours une bibliographie pour évaluer ce que d'autres auteurs ont fait et dit sur le même sujet. Des données antérieures, même partielles, sont donc presque toujours disponibles et une méthode bayésienne permet de les combiner aux résultats de l'expérimentation réalisée. La meilleure preuve de la présence de cet *a priori* est bien l'exemple cité précédemment du calcul d'effectifs nécessaires dans le concept de Neyman & Pearson.

L'utilisation de ces connaissances *a priori* permet donc d'accumuler des données, ce qui a une signification scientifique importante. La théorie bayésienne est par ailleurs très utile dans le cadre des essais controlés, sa souplesse lui permettant de s'adapter plus facilement que la théorie fréquentiste à un nombre important de situations. Sur ces sujets on pourra se reporter aux références [40, 263] ou, pour des comptes rendus plus techniques sur le sujet, aux références [82, 308, 331].

La valeur de l'*a priori* étant dépendante de l'utilisateur, les résultats finaux d'une analyse bayésienne peuvent différer entre deux personnes travaillant sur les mêmes données. Contrairement à l'idée très répandue que la science est objective, la théorie bayésienne traduit la subjectivité inhérente à la recherche. Cette apparente faiblesse de la théorie bayésienne est en fait une force dans le sens où elle permet de vérifier la robustesse des résultats en jouant sur la possibilité de faire varier les données *a priori*. Pour faciliter l'incorporation par chacun de ses propres connaissances sur un domaine donné, Van Houwelingen [332] a proposé d'utiliser le réseau Internet de sorte que le lecteur puisse interagir avec un article en modifiant lui-même l'*a priori* dans une analyse bayésienne qui deviendrait ainsi plus « personnelle ». Cette proposition a été mise en place par [189]

2.7.8 Le respect du principe de vraisemblance

Nous avons déjà abordé ce problème plus haut. Nous n'y reviendrons donc pas ici.

2.7.9 Conclusion intermédiaire sur les méthodes bayésiennes

Les articles en faveur de l'utilisation des méthodes bayésiennes, comme par exemple [100, 127, 129, 130, 128, 268, 281] pour n'en citer que quelques-uns d'une très longue liste, semblent rester sans écho. Si les épidémiologistes sont de plus en plus enclins à utiliser la statistique bayésienne, à en juger par une revue telle que *Epidemiology* qui a totalement supprimé les p-valeurs de ses publications, les revues médicales des spécialités cliniques ne semblent pas avoir intégré ce changement de théorie. Plus précisement, l'utilisation de la statistique bayésienne est de plus en plus courante mais en général pour des applications particulières, pour lesquelles une approche fréquentiste serait particulièrement délicate à réaliser. Dans cet ordre d'idée, on trouvera par exemple les travaux de Clough ou de Dexter [74, 95] ou encore Eilstein [102]. Un bon exemple d'application de la théorie bayésienne peut être trouvé dans [277]. Il faut cependant rester conscient des limites et difficultés de la méthode bayésienne qui n'est pas non plus une panacée [130, 308].

La théorie bayésienne est parfois jugée complexe. Sa compréhension ne pose pourtant pas de problèmes particuliers en comparaison de la statistique fréquentiste [41, 188]. Il est d'ailleurs étonnant de constater que le théorème de Bayes soit si souvent utilisé pour valider une campagne de dépistage ou pour appliquer un test diagnostique en clinique alors qu'il n'est pratiquement jamais utilisé dans le cadre de l'inférence statistique [100]. Le nombre limité de logiciels disponibles utilisant les analyses bayésiennes est sans doute une des explications de la faible fréquence de ces méthodes dans la recherche bio-médicale [310].

Par la suite, nous utiliserons donc de manière prédominante les méthodes bayésiennes.

3 L'analyse d'une table de contingence

Par table de contingence, nous sous-entendons ici qu'il s'agit d'une table de contingence de taille 2×2, contenant donc deux lignes et deux colonnes, la table résultant du croisement de deux variables aléatoires ayant chacune deux modalités et suivant donc des lois binomiales

.

3.1 Forme générale de la table de contingence

On note Ber_X et Ber_Y deux variables aléatoires de Bernoulli, telles que :

$$B_X = \begin{cases} B_X = 1 & \text{avec une probabilité } p_X \\ B_X = 0 & \text{avec une probabilité } 1 - p_X \end{cases}$$

et

$$B_Y = \begin{cases} B_Y = 1 & \text{avec une probabilité } p_Y \\ B_Y = 0 & \text{avec une probabilité } 1 - p_Y \end{cases}$$

Dans un échantillon de taille n, le nombre de sujets prenant la première modalité de la loi de Bernoulli Ber_X (resp. Ber_Y) suit une loi Binomiale $B_X(n, p_X)$ (resp. $B_Y(n, p_Y)$).

Le croisement des deux variables aléatoires est représenté dans une table de contingence qui contient des effectifs totaux, marginaux et par cases prenant respectivement les valeurs $n_{..}$, $n_{i.}$ ou $n_{.j}$ et n_{ij}, pour $i, j \in \{1; 2\}$, i et j étant les indices des lignes et des colonnes.

La table se présente donc de la façon suivante :

Sujets	$Y = 1$	$Y = 0$	Total
$X = 1$	n_{11}	n_{12}	$n_{1.}$
$X = 0$	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

TAB. 1 – Forme générale d'une table de contingence

A cette présentation suivant les effectifs, on fait correspondre une présentation en terme de proportion, plus appropriée lorsque l'on étudie des paramètres issus de lois binomiales, les proportions de chacune des cases étant justement le paramètre p d'une loi binomiale, ou d'une loi multinomiale, selon le type d'échantillonnage considéré. En divisant les effectifs observés de chaque case et chaque marge de la table de contingence par l'effectif total de la table, on obtient les paramètres suivants : $p_{..}$, $p_{i.}$ ou $p_{.j}$ et p_{ij}, pour $i, j \in \{1; 2\}$, i et j étant les indices des lignes et des colonnes.

La table se présente alors de la façon suivante :

Sujets	$Y = 1$	$Y = 0$	
$X = 1$	p_{11}	p_{12}	$p_{1.}$
$X = 0$	p_{21}	p_{22}	$p_{2.}$
	$p_{.1}$	$p_{.2}$	1

TAB. 2 – Présentation d'une table de contingence sous forme de proportion

En statistique, que le paradigme soit bayésien ou fréquentiste, les tables de contingences de taille 2×2 peuvent être analysées de différentes façons selon le plan expérimental. Les marges d'une table 2×2 peuvent être soit fixes soit aléatoires. On appelle marge fixée une marge dont la répartition est connue ou déterminé avant la réalisation de l'expérience. On appelle marge aléatoire une marge dont la répartition n'est pas connue avant l'expérience mais qui résulte de l'observation des résultats de l'expérience. Prenant en compte les deux marges, on a donc soit (1) deux marges fixées, soit (2) deux marges aléatoires, soit (3) une marge fixée et une marge aléatoire. Ces trois types de combinaison de marges correspondent à trois types de plan expérimental. Dans le cas (1), on définit à l'avance pour chacune des deux variables qualitatives d'intérêt la répartition des deux modalités. Cela revient à fixer par exemple la proportion d'hommes et la proportion de fumeurs dans un échantillon. Ce plan est très rarement utilisé en raison de la rareté des situations où il peut être utilisé. Dans le cas (3), on fixe à l'avance la répartition de l'une des variables (par exemple la répartition des sexes ou bien, dans un essai thérapeutique la proportion de sujets se voyant attribuer le traitement A et le traitement B). L'autre variable est aléatoire et correspond au résultat auquel on s'intéresse tout particulièrement et qui est censé être lié au facteur fixé. C'est une situation très courante en expérimentation bio-médicale puisqu'elle inclut les cas des essais thérapeutiques, des études de type exposé-non-exposé et des études cas-témoins. Le cas (2) est relativement fréquent également puisqu'il correspond aux études d'observation où un facteur de risque est croisé avec une pathologie donnée sans que l'on connaisse à l'avance dans l'échantillon la répartition ni du facteur de risque ni de la pathologie.

L'analyse fréquentiste distingue théoriquement le cas (2) du cas (3). Dans le cas (2), le test de Barnard devrait être utilisé. Un long débat entre Barnard et Fisher [212] a abouti à l'acceptation par Barnard de l'idée (de Fisher) que les marges devaient quand même être considérées comme fixées. Cette conclusion fait que, d'un point de vue fréquentiste, les trois types de tables selon l'échantillonnage des marges sont analysées de la même manière, soit par l'utilisation du test du χ^2 soit par le test exact de Fisher. Le test de Barnard, qui n'est proposé que dans le logiciel `StatXact`, n'est en pratique plus utilisé. La raison en est essentiellement la reconnaissance par Barnard lui même que ce test n'était pas bon [212].

En pratique, les controverses sur ce sujet se poursuivent et toutes les tables sont analysées soit par un test exact de Fisher soit par un test du χ^2 [7].

3.2 Les paramètres d'intérêts dans une table de contingence

Différentes mesures peuvent être définies pour analyser une table de contingence. Il s'agit principalement du risque attribuable, de la différence de risque (DR), du risque relatif (RR) et de l'odds-ratio (OR). Seuls les trois derniers indices seront rappelés, dans les paragraphes suivants. De nombreux auteurs ont traités le problème de l'estimation d'une proportion et le problème de la comparaison de deux proportions. Nous n'utilisons ici que les principes les plus courants mais les variantes possibles sont données dans la bibliographie, tant pour l'analyse fréquentiste [8, 10, 67, 81, 99, 207, 211, 215, 236, 237, 241, 240, 275, 276, 298, 304] que bayésienne [5, 6, 7, 12, 15, 56, 81, 134, 138, 150, 265].

3.2.1 La différence de risque

En reprenant la notation introduite dans le tableau 1, la différence de risque Δ se calcule de la façon suivante :

$$\Delta = \frac{n_{11}}{n_{1.}} - \frac{n_{21}}{n_{2.}}$$

3.2.2 Le risque relatif

Le risque relatif RR est le rapport entre le risque dans la population « exposée » et le risque dans la population de référence (« non exposée »). On a donc :

$$RR = \frac{n_{11}/n_{1.}}{n_{21}/n_{2.}}$$

3.2.3 L'odds-ratio

C'est la mesure qui nous intéressera principalement dans ce travail. L'OR peut être utilisé comme une mesure d'association (même si des articles récents critiquent cette utilisation [179, 242]). Le fait qu'il soit lié directement au modèle de régression logistique lui donne une généralité intéressante et utile. En épidémiologie, pour un sujet exposé, la cote est le rapport entre le risque de développer la maladie et le risque complémentaire de ne pas développer la maladie soit $a/(1-a) = a/b$. Le plus intéressant est de comparer cet odd à l'odd d'une population témoin (non exposée par exemple) en calculant le rapport de la cote chez les exposés et de la cote chez les non-exposés. On obtient donc le rapport des cotes : l'odds-ratio. De manière générale, l'OR se calcule de la façon suivante :

$$OR = \frac{x/(1-x)}{y/(1-y)}$$

Dans cette formule, et dans un contexte épidémiologique, x représente le risque d'être malade lorsque l'on est exposé et y le risque d'être malade lorsque l'on n'est pas exposé. Sur le même tableau que précédemment, on peut donc réécrire cette formule de la façon suivante :

$$OR = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

Cet OR doit évidemment être assorti de son intervalle de confiance, soit fréquentiste, soit bayésien. L'intervalle de confiance fréquentiste est donné par deux formules, selon que l'on utilise une approximation normale ou une formule exacte.

3.3 Analyse fréquentiste

Approximation gaussienne On trouve dans Fleiss [115] que l'erreur standard de l'OR peut-être estimé par :

$$s_{OR} = OR \cdot \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

On peut donc en déduire un intervalle de confiance pour l'OR par :

$$IC_{OR} = OR \pm z_{\alpha/2} \cdot s_{OR}$$

Méthode exacte Avec les notations précédentes, le calcul de l'intervalle de confiance exact de l'OR est basé sur une formule donnée par Cox [80].

On retient ici comme convention d'écriture que, pour des marges fixées, la valeur observée de n_{11} peut varier, conditionnellement aux marges, entre n_{min} et n_{max}. Par ailleurs, on note N, la variable aléatoire associée au nombre de malades exposés. On cherche à estimer β tel que $OR = e^{\beta}$. Alors :

$$Pr_N(n_{11};\beta) = \frac{\binom{n_{.2}}{n_{1.}-n_{11}}\binom{n_{.1}}{n_{11}}e^{\beta n_{11}}}{\sum_u \binom{n_{.2}}{n_{1.}-u}\binom{n_{.1}}{u}e^{\beta u}} \quad (1)$$

Dans cette équation, u varie de $n_{11,min}$ à $n_{11,max}$. Les queues de distribution supérieure $Pr_N^{sup}(n_{11};\beta)$ et inférieure $Pr_N^{inf}(n_{11};\beta)$ se calculent en sommant respectivement $Pr_N(n_{11};\beta)$ de n_{11} à $n_{11,max}$ pour la queue supérieure et de $n_{11,min}$ à n_{11} pour la queue inférieure. Pour

trouver les valeurs des deux bornes p_{inf} et p_{sup} de l'intervalle de confiance de p, on parcourt l'espace de définition de p pour trouver les valeurs telles que :

$$Pr_N^{sup}(n_{11};\beta|p_{inf}) = \frac{\alpha}{2} \tag{2}$$

$$Pr_N^{inf}(n_{11};\beta|p_{sup}) = \frac{\alpha}{2} \tag{3}$$

En pratique, un algorithme convergent basé sur la méthode des simplex donne rapidement la solution attendue.

Rappel et définition du modèle logistique Le modèle logistique, basé sur la fonction logistique, s'écrit de la façon suivante :

$$\begin{aligned} P(Y=1) &= P(M^+) \\ P(M^+|X) &= f(X) = \tfrac{1}{1+e^{-(\alpha+\beta x)}} \end{aligned}$$

Une variable d'exposition codée comme une variable qualitative à deux classes, E+ et E-, peut être recodée de la façon suivante : Si E+ $\Rightarrow X = 1$ et si E$-$ $\Rightarrow X = 0$. Le modèle s'écrit alors, pour ces deux modalités de l'exposition :

$$P(M^+|X=1) = P_1 = \frac{1}{1+e^{-(\alpha+\beta)}} \tag{4}$$

$$P(M^-|X=1) = 1 - P_1 = \frac{e^{-(\alpha+\beta)}}{1+e^{-(\alpha+\beta)}} \tag{5}$$

et :

$$P(M^+|X=0) = P_0 = \frac{1}{1+e^{-(\alpha)}} \tag{6}$$

$$P(M^-|X=0) = 1 - P_0 = \frac{e^{-(\alpha)}}{1+e^{-(\alpha)}} \tag{7}$$

L'OR peut être réécrit de la façon suivante.

On rappelle que :

$$OR = \frac{P_1/(1-P_1)}{P_0/(1-P_0)}$$

Soit, en introduisant les probabilités conditionnelles :

$$OR = \frac{P(M^+|X=1)/P(M^-|X=1)}{P(M^+|X=0)/P(M^-|X=0)}$$

En remplaçant dans cette expression les probabilités conditionnelles par les formules présentées plus haut (4 à 7), on obtient :

$$OR = \frac{1/(e^{-(\alpha+\beta)})}{1/(e^{-\alpha})} = \frac{e^{(\alpha+\beta)}}{e^{\alpha}} = e^{\beta}$$

$$ln(OR) \quad = \quad \beta$$

On voit donc que l'OR est l'exponentielle du paramètre β d'une régression logistique.

3.4 Analyse bayésienne

L'OR se définit de la même manière dans le cadre bayésien. L'intervalle de confiance ou intervalle de crédibilité est un peu différent. Il s'obtient soit en utilisant un intervalle de plus haute probabilité (HDR : *Highest Density Region*) soit avec un intervalle empirique basé sur les percentiles $\alpha/2$ et $1 - \alpha/2$. L'expression analytique de l'OR, incluant donc les distributions *a priori*, est relativement complexe (voir [138]). Le calcul de l'intervalle de crédibilité par la méthode de l'HDR s'obtient soit analytiquement soit par intégration numérique. En pratique, on se reposera le plus souvent sur des estimations empiriques obtenues par simulations avec les méthodes de type Markov Chain Monte Carlo (MCMC) comme dans `WinBUGS` par exemple.

3.5 Cas des données binomiales

Soit une série de n observations suivant chacune un schéma de Bernoulli de paramètre θ. Si x est le nombre de succès observés dans cet n-échantillon et θ la probabilité de succès, le modèle naturel pour décrire le phénomène est le modèle binomial. c'est-à-dire que, conditionnellement à θ, la probabilité d'observer x succès parmi n essais s'écrit :

$$\Pr(X = x|\theta) = C_n^x \cdot \theta^x (1-\theta)^{n-x}$$

3.5.1 Rappel sur la loi Beta

L'utilisation du théorème de Bayes suppose l'utilisation et donc la spécification d'une loi *a priori*. Concernant la distribution *a priori* de θ, un bon choix est une loi de la famille Beta. Ces lois, à deux paramètres α et β, sont définies dans l'intervalle $[0, 1]$

Une variable aléatoire X présente une distribution β si sa densité de probabilité s'exprime de la manière suivante :

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}(1-x)^{\beta-1}}{Be(\alpha,\beta)} & \text{si} \quad 0 < x < 1 \\ \\ 0 & \text{ailleurs} \end{cases}$$

où $Be(\alpha, \beta)$ est la fonction Beta qui s'écrit :

$$Be(\alpha,\beta) = \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}du \quad \text{avec} \;\; m > 0, \; n > 0$$

On peut montrer que $Be(\alpha, \beta)$ vaut :

$$Be(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

La fonction Γ vaut :

$$\Gamma(n) = \int_0^\infty t^{n-1}e^{-t}dt \quad n > 0$$

La moyenne et la variance de la loi Beta sont respectivement :

$$\mu = \frac{\alpha}{\alpha+\beta} \tag{8}$$

$$\sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{9}$$

Par ailleurs, si $\alpha > 0$ et si $\beta > 0$, alors il y a un mode unique, dont la valeur est :

$$x_{\text{mode}} = \frac{\alpha-1}{\alpha+\beta-2}$$

Selon les valeurs de α et β, les lois Beta peuvent avoir des formes très différentes. Cette souplesse rend les lois Beta très adaptées à la formalisation de la connaissance préliminaire sur une variable bornée, telle qu'une proportion. En vertu de ce choix, l'expression de la loi *a priori* de θ est :

$$\Pr(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

D'après la formule de Bayes, la loi *a posteriori* de θ, $[\theta|y]$, est proportionnelle au produit entre loi *a priori* et vraisemblance :

$$\Pr(\theta|x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}$$

qui est alors encore une loi Beta de paramètres $(x+\alpha, n-x+\beta)$. Les lois *a priori* et *a posteriori* appartiennent ainsi à la même famille Beta. Elles sont donc conjuguées, c'est-à-dire que la loi *a posteriori* est de la même famille que la loi *a priori*. Ces propriétés sont évidemment tout à fait intéressantes puisque, notamment dans le cas d'un *a priori* non informatif, l'estimation la plus probable de la valeur du paramètre est alors la valeur empirique de l'estimateur, ce qui correspond à l'estimation du maximum de vraisemblance.

3.6 Rappel sur la loi de Dirichlet

La loi de *Dirichlet* de taille $k \in \mathbb{N}$ et de paramètre $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k) \in]0, +\infty[^k$ k est la loi du vecteur :

$$\left(\frac{Z_1}{Z_1 + \cdots + Z_k}, \frac{Z_2}{Z_1 + \cdots + Z_k}, \ldots, \frac{Z_k}{Z_1 + \cdots + Z_k} \right)$$

$Z_1, Z_2, \ldots, Z_k$ sont des variables aléatoires i.i.d. distribuées selon des lois exponentielles de paramètres $\alpha_1, \alpha_2, \ldots, \alpha_k$ respectivement. La densité de probabilité de $(x_1, x_2, \ldots, x_k)$ est donnée par :

$$\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdot \Gamma(\alpha_2) \cdot \cdots \cdot \Gamma(\alpha_k)} \left(\prod_{i=1}^{k-1} x_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} x_i \right)^{\alpha_k - 1}$$

ou

$$\alpha_0 = \Sigma_{i=1}^{k} \alpha_i$$

Les espérances et variances d'une loi de Dirichlet de paramètre α_i sont :

$$\mu(X_i) = \frac{\alpha_i}{\alpha_0}$$

et

$$\sigma^2(X_i) = \frac{\alpha_i \cdot (\alpha_0 - \alpha_i)}{\alpha_0^2 \cdot (\alpha_0 + 1)}$$

Les lois de Dirichlet sont normalement utilisées en statistique bayésienne comme lois *a priori* conjuguées du paramètre p_i d'une loi multinomiale. Avec ce choix, la loi *a posteriori* issue de la formule de Bayes, proportionnelle au produit entre une densité de probabilité de Dirichlet et une vraisemblance multinomiale est encore une Dirichlet de paramètre $\alpha_i + x_i$, avec $i \in \{1, ..., k\}$.

3.6.1 Choix de la loi *a priori* et de ces paramètres

Si les lois Beta sont fréquemment choisies pour modéliser le paramètre θ d'une loi binomiale, il reste à préciser les valeurs *a priori* des paramètres de cette loi. Dans le cadre de l'estimation d'une proportion, plusieurs choix sont possibles. Les variantes les plus couramment proposées sont décrites ci-dessous.

Loi uniforme Le choix le plus spontané est d'utiliser une loi uniforme ne spécifiant rien ou plus exactement spécifiant une absence d'information *a priori* telle que toutes les valeurs que peut prendre p soient également probables. On désigne par le terme de « non informatif » cette loi uniforme. C'est la loi qui avait été proposé par Bayes lui-même dans sa publication [33, 188]. De façon générale, et sauf indications contraires, nous utiliserons dans ce travail des *a priori* non informatifs. Dans le cas de binomiales, les paramètres des lois Beta seront donc $\alpha = \beta = 1$. Cette loi n'est pas sans poser certaines difficultés notamment du fait qu'une fonction de la densité uniforme n'est pas forcément uniforme. Par ailleurs, le choix d'autres valeurs serait tout à fait pertinent dans le cadre de l'analyse des données d'un point de vue strictement biologique. Si l'on dispose d'une information préalable, telle que les résultats d'un essai thérapeutique ou d'une série d'observations antérieures, il serait naturel d'utiliser une loi Beta dont les paramètres α et β reflètent ces connaissances. Logiquement, la valeur de α correspond au nombre de succès observés et β correspond au nombre d'échecs observés précédemment [188, 308]. Mais le présent travail s'attache essentiellement à comparer différents modèles. L'introduction de différents paramètres pour les lois *a priori* obligerait alors à prendre en compte les différentes valeurs possibles des lois Beta, ce qui alourdirait l'analyse sans en modifier les conclusions quant aux modèles. Par ailleurs, c'est la loi qui correspond à des résultats fréquentistes, d'où son intérêt lorsqu'il s'agit de comparer les deux méthodes. En conséquences de quoi, et sauf mention contraire, les *a priori* seront ceux indiqués plus haut.

Cependant d'autres lois *a priori* sont fréquemment présentées dans la littérature, bien qu'elles ne soient quasiment jamais utilisées. Il s'agit essentiellement de l'*a priori* de Haldane et de l'*a priori* de Jeffrey.

Loi *a priori* de Haldane L'*a priori* de Haldane correspond à une loi Beta(0,0) dont la densité est :

$$\Pr(\theta) \propto \theta^{-1}(1-\theta)^{-1}$$

Cette loi est une densité impropre mais elle est équivalente à une loi uniforme dans l'échelle des logarithmes pour l'odds de θ. Un argument pour l'utilisation de cette loi est le fait que la moyenne de la distribution $Beta(\alpha + x, \beta + n - x)$ vaut :

$$(\alpha + x)/(\alpha + \beta + n)$$

laquelle coïncide avec l'estimation du maximum de vraisemblance $E(\theta) = x/n$ uniquement quand $\alpha = \beta = 0$. Lors de l'utilisation d'un autre *a priori*, c'est le mode de la distribution *a posteriori* qui correspond à l'estimation du maximum de vraisemblance de la proportion.

Loi *a priori* de Jeffrey La loi *a priori* de Jeffrey est la suivante :

$$\Pr(\theta) =\propto \sqrt{I(\theta|x)}$$

où I représente la matrice d'information de Fisher. La motivation essentielle de l'utilisation de cette loi est basée sur sa propriété d'invariance. En effet, quelle que soit l'échelle de mesure utilisée pour le paramètre, on obtient la même loi *a priori* ce qui assure un grand confort d'utilisation.

3.7 Le choix du modèle d'échantillonnage

Nous avons dit que le traitement bayésien des tables 2×2 tient théoriquement compte de l'échantillonnage des marges [5, 12, 56, 150]. Après avoir choisi les lois *a priori* pour les paramètres d'intérêt, il reste à choisir le modèle utilisé pour mener l'analyse. Les situations (1) et (3) présentées plus haut sont modélisées comme des comparaisons de lois binomiales utilisant comme *a priori* des lois Beta. Dans le cas (2), on utilise un modèle multinomial avec comme loi *a priori* des densités de Dirichlet $D(\alpha_1, \ldots, \alpha_k)$ qui généralise les lois Beta. En pratique, une grande confusion peut s'observer dans la littérature et même dans les ouvrages de références, le choix d'un modèle multinomial ou binomial n'étant pas toujours clairement justifié.

Quel que soit le modèle sous-jacent (deux binomiales ou une multinomiale) il est possible de calculer la valeur de l'OR pour la table 2×2 [5, 6]. Les programmes Winbugs correspondants sont très simples. On utilisera par exemple l'un des deux programmes donnés en annexes 2 (prog. 1 et 2.) En pratique, les résultats des deux modèles (binomial ou

multinomial) sont souvent très proches voire identiques et nous n'utiliserons que l'approche basée sur la comparaison de deux binomiales sous l'hypothèse d'une marge fixée.

3.8 Le calcul du facteur de Bayes dans un tableau 2×2

La statistique fréquentiste formule pour tout test un couple d'hypothèses : une hypothèse nulle, spécifiant la nullité d'un paramètre, et une hypothèse alternative. Cette formulation n'est pas naturelle dans le raisonnement bayésien mais il est néanmoins possible de la transposer en terme bayésien. Elle sera donc étudiée afin de faciliter les comparaisons entre les deux théories statistiques. Le paragraphe suivant présentera une formulation plus typiquement bayésienne. Pour une table 2×2 croisant deux binomiales, dont les nombres de succès sont respectivement x_1 et x_2, les deux hypothèses ou modèles à comparer sont l'hypothèse nulle d'égalité des taux de succès et l'hypothèse alternative de différence des taux de succès. Les deux hypothèses correspondent à deux modèles pour un même jeu de données. Une manière simple de comparer ces deux modèles afin d'en choisir le meilleur est d'utiliser le Facteur de Bayes (FB) [169]. Le FB quantifie le changement de l'information *a priori* apporté par les données. Plus le changement est important, plus les données soutiennent l'une des deux hypothèses. Nous rappelons ici rapidement que, sous sa forme générale, le FB est le rapport des vraisemblances des données sous les deux hypothèses :

$$FB = \frac{p(x|\theta_0)}{p(x|\theta_1)}$$

Facteur de Bayes sous un modèle binomial pour une hypothèse ponctuelle Dans le cas de la comparaison de deux proportions, sous un modèle binomial, le Facteur de Bayes (BF) se calcule de la façon suivante [131, 134, 138, 139, 188] :

$$FB = \frac{p_1(x_1, x_2)}{p_2(x_1, x_2)}$$

avec

$$p_1(x_1, x_2) = C_{n_1}^{x_1} C_{n_2}^{x_2} \frac{Be(\alpha_1 + \alpha_2 - 1, \beta_1 + \beta_2 - 1)}{Be(\alpha_1^0 + \alpha_2^0 - 1, \beta_1^0 + \beta_2^0 - 1)}$$

et

$$p_2(x_1, x_2) = C_{n_1}^{x_1} C_{n_2}^{x_2} \frac{Be(\alpha_1, \beta_1) Be(\alpha_2, \beta_2)}{Be(\alpha_1^0, \beta_1^0) Be(\alpha_2^0, \beta_2^0)}$$

Le coefficient du numérateur et celui du dénominateur s'annulent ce qui laisse un produit de loi Beta facile à calculer. Le programme R correspondant est donné en annexe (programme N°3).

Albert propose une autre formule pour le calcul du FB dans le cadre d'une table 2×2. Dans le cadre de la comparaison de deux lois binomiales avec détermination de la valeur de l'OR, le FB est défini de manière à tester une hypothèse nulle ponctuelle de type $\theta = \theta_0$. Pour Hashemi, l'hypothèse nulle est de façon naturelle l'hypothèse que $\theta_1 = \theta_2$. Ceci revient à spécifier une loi unique pour les deux binomiales qui suivent donc chacune une loi binomiale de même paramètre θ et $x\ B(n_i, \theta), \quad i = 1, 2$. Il spécifie donc une loi *a priori* unique pour les deux paramètres :

$$p = Be(\alpha_1 + \alpha_2 - 1, \beta_1 + \beta_2 - 1)$$

Albert propose lui d'utiliser des lois *a priori* indépendantes pour les probabilités de chaque binomiale. Sa formulation est donc un peu plus complexe mais un peu plus générale. On peut se ramener à la même valeur du FB que pour Hashemi en utilisant un choix adéquat des lois Beta *a priori*.

Le programme indiqué précédemment (programme N°3) peut être légèrement amélioré. On peut intégrer le calcul de la probabilité *a posteriori* de l'hypothèse nulle à partir de la valeur de la probabilité *a priori* de HN et du FB. En effet [188], on sait que la probabilité *a posteriori* $\Pr(HN|D)$ vaut :

$$\Pr(HN|D) = \frac{1}{1 + \frac{1}{FB} \cdot \frac{\Pr(HA)}{\Pr(HN)}}$$

Par ailleurs, Altham [15, 188] donne une formule permettant dans le cas de deux lois binomiales de calculer la probabilité que la proportion p_1 de la première loi soit inférieure à la probabilité p_2 de la seconde loi.

Pour deux lois binomiales dont les paramètres π et ρ suivent respectivement une loi $Be(\alpha, \beta)$ et $Be(\gamma, \delta)$, cette probabilité se calcule de la manière suivante :

$$\Pr(\pi < \rho) = \sum_{\kappa=max(\gamma-\beta,0)}^{\gamma-1} \frac{C_{\gamma+\delta-1}^{\kappa} C_{\alpha+\beta-1}^{\alpha+\gamma-1-\kappa}}{C_{\alpha+\beta+\gamma+\delta-2}^{\alpha+\gamma-1}}$$

Il est clair que la probabilité que $\Pr(\pi < \rho) = \Pr(OR < 1)$, ce qui permet d'utiliser cette formule pour tester un OR. Par ailleurs, Altham [15] montre qu'il existe un lien élégant entre le test exact de Fisher et la théorie bayésienne puisque la valeur de $\Pr(\pi < \rho)$ n'est

autre que la valeur du test exact de Fisher dans sa version unilatérale en supposant que l'on utilise une loi *a priori* spécifiant une relation négative entre les lignes et les colonnes. Cette loi est une loi de Dirichlet $D(0, 1, 1, 0)$ équivalente à un tableau à 4 cases dans lequel les effectifs de la première diagonale sont nuls et les effectifs de la seconde diagonale égaux à 1. Ceci correspond à un *a priori* favorisant l'hypothèse nulle. Dit autrement, le test exact de Fisher correspond à un test bayésien avec un *a priori* conservateur ce qui explique la nature conservative (puissance faible) du test de Fisher [7], en dehors de l'aspect discret de la densité de probabilité de la loi multinomiale utilisée pour le test exact de Fisher.

En utilisant ces deux éléments, on aboutit au programme N°4 (en annexe) permettant de calculer la valeur du FB, les probabilités *a posteriori* de HN, HA ainsi que la probabilité que p_1 soit inférieure à p_2. Il faut noter que ce dernier calcul se fait dans le cadre du test d'une hypothèse nulle ponctuelle, hypothèse quelque peu artificielle dans le cadre bayésien. Nous verrons par la comparaison avec la « vraie » probabilité que p_1 soit inférieure à p_2 que cette formulation est peu réaliste.

Facteur de Bayes sous un modèle binomial pour une hypothèse composite Le paragraphe précédent était consacré au calcul du FB pour une formulation de type fréquentiste du test d'hypothèse. Dans la conception bayésienne, cette formulation est peu pertinente car elle revient à calculer la probabilité qu'un paramètre à valeur dans $\mathbb{R}$ soit exactement égal à une valeur donnée. Or, $\Pr(\theta = \theta_r) = 0$. Une hypothèse nulle ponctuelle a une probabilité nulle d'être vraie. Il est plus naturel pour un bayésien de s'intéresser à des hypothèses dites composites dans lesquelles le paramètre d'intérêt appartient à l'une ou l'autre région de l'espace possible du paramètre. Donc, ici on a : $\theta \in \Theta$ où Θ n'est pas un singleton. Dans ce cas, la distribution a priori de θ intervient directement dans le calcul de FB. Ainsi, on a :

$$FB = \frac{\int_{\theta \in \Theta_0} p(x|\theta)\rho_0(\theta)d\theta}{\int_{\theta \in \Theta_1} p(x|\theta)\rho_1(\theta)d\theta}$$

avec $\rho_i = p(\theta)/\pi_i$, et $\theta \in \Theta_i, i \in \{0; 1\}$. Le FB est alors le ratio des vraisemblances de Θ_0 et Θ_1 pondérées par ρ_0 et ρ_1 [188]. Pour être plus précis, la situation présentée au paragraphe précédent est en fait un test entre une hypothèse nulle qui est ponctuelle et une hypothèse alternative qui est composite. (voir programme N°5)

4 Analyses statistiques pour données incomplètes

Les méthodes statistiques relatives aux données manquantes sont extrêmement nombreuses et une description exhaustive, même très synthétique de la littérature traitant de ce sujet est hors de portée. Nous ne présentons ici que les définitions générales et les méthodes ayant un certain intérêt pour notre propos. Plus de détails pourront être trouvés pour des situations particulières dans les références suivantes : [14, 4, 24, 89, 91, 98, 155, 173, 182, 183, 191, 209, 224, 279, 280], notamment dans le cas de données répétées, catégorielles ou non [68, 112, 113, 196, 198, 221, 222, 223], pour des modèles statistiques en particuliers [278] ou dans le cadre des essais thérapeutiques [25, 26, 22, 158, 231] ou encore de modèle originaux [59], parfois dans le cadre bayésien [165]. Les sources principales se trouvent dans les publications de Little et Rubin [199, 200, 201, 202, 203, 214, 283, 284, 283, 284]

4.1 Définitions générales

Les données manquantes peuvent être classées dans différentes catégories selon la façon dont elles se répartissent parmi l'ensemble des données recueillies. Il faut distinguer la répartition « géographique » de la répartition « probabiliste » des données.

Soit une matrice de données $\mathbf{X}$, de taille $s \times p$, où s est le nombre de sujets et p le nombre de variables relevées sur ses n sujets. La matrice $\mathbf{X}$ contient donc $n = s \times p$ données au sens large du terme. On définit alors :

1 les données ;
2 les données observées (DO) ;
3 les données complètes (DC) ;
4 les données manquantes (DM) ;
5 les données imputées (DI) ;
6 les données complétées (DCé) ;

Les données On appelle *donnée* (au sens large du terme) une variable aléatoire à valeur dans $\mathbb{R}$ ou $\mathbb{N}$ (éventuellement après recodage lorsqu'il s'agit d'une variable qualitative) pour laquelle l'obtention d'une mesure (le recueil des données) est prévue dans le plan d'expérience, que ce recueil ait, *in fine* eu lieu ou pas.

Les données observées On appelle *donnée observée* une donnée pour laquelle on souhaite disposer d'une mesure. Le nombre de données observées est égal au nombre de données complètes plus le nombre de données manquantes $n = n_p + n_m$.

Les données complètes On appelle *donnée complète* une mesure pour laquelle on dispose d'une valeur. Le nombre de données complètes est n_p. Il faut noter que le nombre de sujets complets diffèrent du nombre de données complètes.

Les données manquantes On appelle *donnée manquante* une donnée pour laquelle on ne dispose pas de la valeur de la mesure. On note x_* cette valeur. Le nombre de DM de la matrice $\mathbf{X}$ est n_m. Les différentes situations menant à une DM seront détaillées dans un paragraphe ultérieur.

Les données imputées On appelle *donnée imputée* une valeur n_{i*} attribuée à une valeur manquante. La façon de réaliser cette imputation sera vue dans un paragraphe ultérieur. Le nombre de données imputées est de $n_{i^*} \times i$ où i est le nombre d'imputations réalisées.

Les données complétées On appelle *données complétées* l'ensemble des valeurs tel qu'on les observe après imputation des valeurs (quel que soit le mode d'imputation). On verra que les données complètes peuvent éventuellement exister en plusieurs exemplaires pour un même jeu de données. Le nombre de données complétées est $n = n_p + n_m$.

On peut donc à partir de ces éléments réaliser un chaînage dans l'ordre d'apparition des différents types de données. On part de mesures devant être réalisées afin d'aboutir à un tableau complet contenant les données uniquement complètes. On obtient le plus souvent une matrice de données contenant des données complètes *et* des données manquantes, lesquelles peuvent remplacées par des données imputées pour aboutir, après addition au tableau des données complètes, à des données complétées.

4.2 Mécanisme des manquants

On distingue pour les DM le mécanisme « matériel » ayant généré la DM et le mécanisme « statistique ».

4.2.1 Mécanisme matériel menant aux données manquantes

D'un point de vue matériel, le DM survient de l'absence de recueil, soit par impossibilité de mesurer la valeur (valeur hors étendue de l'appareil ou appareil en panne ou méthode prise en défaut), valeur mesurée mais perdue avant la saisie, valeur non applicable (pour les données qualitatives en général), soit encore par simple défaut de mesure, ce qui est très fréquent lors des enquêtes rétrospectives. Dans le cas des données qui nous intéressent dans le cadre de ce travail, les données manquantes sont générées par l'impossibilité de mesurer la valeur : la méthode de mesure (ici la PCR) est prise en défaut pour obtenir la mesure dans le cas des sites homozygotes pour le microsatellite considéré.

4.2.2 Mécanisme statistique menant aux données manquantes

Les différents mécanismes statistiques ont abouti à une classification proposée en 1976 par Rubin [283] et présentée par Little et Rubin [284]. Une présentation plus pragmatique adapté à l'épidémiologie est donnée par Chavance [64]. Nous rappelons ici cette classification.

Soient deux variables aléatoires X et Y dont on recueille n réalisations. Admettons que X soit complètement observée et que Y comporte *in fine* un certain nombre de valeurs manquantes. Par ailleurs, dans cet exemple, les deux variables aléatoires X et Y peuvent être chacune soit qualitative soit quantitative sans perte de généralité. On peut schématiser cette situation à l'aide du tableau 3 ci-dessous :

sujet	X	Y
1	x_1	y_1
...	...	...
i	x_i	y_i
...	...	...
n-m+1	x_{n-m+1}	$*$
n	x_n	$*$

Tab. 3 – Tableau pour la classification de LR

Cette structure de données manquantes, dans laquelle les valeurs de Y sont manquantes de façon ordonnée par rapport aux valeurs de X pour un certain agencement des sujets est dite monotone. Rubin propose à partir de cette situation, trois types de mécanisme pour les DM :

(1) si la probabilité d'avoir une valeur non manquante est indépendante de X et de Y ;

(2) si la probabilité d'avoir une valeur non manquante dépend de X mais pas de Y ;

(3) si la probabilité d'avoir une valeur non manquante dépend de X et de Y ;

Dans le cas (1) on dit que les valeurs manquantes sont Manquantes Aléatoirement (*Missing at random* : MAR) et que les données observées sont Observées Aléatoirement (*Observed at random* : OAR). De façon plus synthétique, on dit que les données sont manquantes complètement aléatoirement (*missing completely at random* : MCAR). Dans ce cas, les valeurs Y observées forment un sous-échantillon aléatoire des valeurs de Y (comme dans le cas du tableau 3).

Dans le cas (2) , on dit que les données sont manquantes aléatoirement (*missing at random* : MAR). Dans cette situation, les valeurs observées de Y ne sont pas forcément un sous-échantillon aléatoire des valeurs échantillonnées de Y mais elles sont un sous-échantillon aléatoire de Y dans des sous-classes définies par les valeurs de X.

Enfin, dans le cas (3) les valeurs ne sont ni manquantes aléatoirement (MAR) ni obervées aléatoirement (OAR). Elles sont dites *missing not at random* : MNAR . Dans les cas (2) et (3) le mécanisme des manquants peut être ignoré pour les méthodes d'inférences basées sur la vraisemblance. Dans le cas (3), il peut être ignoré à la fois pour les approches basées sur la vraisemblance et pour les approches basées sur l'échantillonnage. Dans le cas (1) le mécanisme ne peut être ignoré.

Les données manquantes de types MNAR sont celles posant les problèmes les plus sérieux, ce qui explique le nombre relativement faible de publications traitant de ce problème [69, 79, 201, 262]. De plus, la plupart des méthodes pour données MNAR ont été développées pour des cas particuliers.

4.3 Nécessité d'analyse pour données manquantes

La nécessité de développer et d'utiliser des méthodes statistiques spécifiques aux données manquantes est issu d'un besoin tant théorique que pratique. Sur un plan théorique, les données manquantes posent des problèmes particuliers pour l'inférence, essentiellement une baisse de puissance et un risque de biais. La nécessité pratique de prendre en considération les données manquantes découle de ces aspects mais aussi du besoin de prendre une décision qui soit la moins entachée d'erreur possible.

4.3.1 Nécessité théorique

Soit une matrice de j variables dont chacune contient la même proportion p_m de valeurs manquantes

$$\Pr(X_i = *) = \Pr(X_j = *), \forall\ i \neq j$$

En admettant que les manquants soient indépendants d'une variable à l'autre,

$$\Pr(X_i = * \text{ et} X_i = *) = \Pr(X_i = *) \times \Pr(X_j = *), \forall\ i \neq j$$

on calcule la probabilité $p_{\bar{m}}$ d'avoir un sujet complet (n'ayant aucune valeurs manquantes sur les j variables) de la manière suivante :

$$p_{\bar{m}} = (1 - p_m)^j$$

Le tableau ci-dessous montre cette proportion en fonction du nombre de variables j incluses dans l'analyse et de la proportion de manquants pour chaque variable j.

valeur de p_m	nombre de variables incluses									
	1	2	3	4	5	6	7	8	9	10
0,01	0,990	0,980	0,970	0,961	0,951	0,941	0,932	0,923	0,914	0,904
0,02	0,980	0,960	0,941	0,922	0,904	0,886	0,868	0,851	0,834	0,817
0,03	0,970	0,941	0,913	0,885	0,859	0,833	0,808	0,784	0,760	0,737
0,04	0,960	0,922	0,885	0,849	0,815	0,783	0,751	0,721	0,693	0,665
0,05	0,950	0,903	0,857	0,815	0,774	0,735	0,698	0,663	0,630	0,599
0,10	0,900	0,810	0,729	0,656	0,591	0,531	0,478	0,430	0,387	0,349
0,15	0,850	0,723	0,614	0,522	0,444	0,377	0,320	0,272	0,232	0,197
0,20	0,800	0,640	0,512	0,410	0,328	0,262	0,210	0,168	0,134	0,107
0,30	0,700	0,490	0,343	0,240	0,168	0,118	0,082	0,058	0,040	0,028
0,40	0,600	0,360	0,216	0,130	0,078	0,048	0,028	0,017	0,010	0,006
0,50	0,500	0,250	0,125	0,063	0,031	0,016	0,008	0,004	0,002	0,001

TAB. 4 – Proportion de sujets complets dans une analyse statistiques selon le nombre de variables incluses et la proportion de manquants par variables.

La perte d'effectif devient vite très importante, de l'ordre de 20 à 30% même pour des probabilités de manquer faibles et pour un nombre de variables limitées. Par exemple pour une probabilité de manquer $p_m = 0,05$ et un nombre de variables égal à 4 ou 5, la proportion de sujets complets n'est que de 81,5 et 77,4% respectivement. Bien sûr, la situation précise dépendra de la validité des hypothèses utilisées mais ce simple calcul donne au moins une première estimation de l'envergure du problème.

L'impact sur la puissance dans le cas d'une proportion est simple à calculer : pour une proportion p à estimer, pour un nombre n_p de données présentes et un nombre n_m de données manquantes, la variance de la proportion lorsque l'on utilise une approximation gaussienne est multipliée par un facteur $(n_p+n_m)/n_p$ en supposant que la proportion parmi les manquants est la même que parmi les non manquants. Dans le cas contraire, le rapport peut être plus grand encore.

4.3.2 Nécessité pratique

De façon générale, les données manquantes ne sont pas prises en compte dans la littérature bio-médicale [55, 158, 341, 346, 354]. Ce constat général se vérifie particulièrement pour les données d'allélotypage. Les microsatellites ne sont pas analysés correctement dans la littérature biologique. Plus exactement, le problème des homozygotes n'est jamais spécifiquement traité ni même évoqué. Avant de savoir si l'on peut ignorer le problème des homozygotes il faut s'assurer de la validité de cette hypothèse ce qui, à notre connaissance, n'a jamais été fait. On trouve un grand nombre d'articles dans lesquelles le problème des homozygotes n'est pas pris en considération : dans la plupart des cas, les sujets homozygotes sont retirés de l'analyse et donc l'incertitude supplémentaire n'est pas incluse dans les résultats. Les problèmes de traitement des données d'allélotypage et les erreurs qui en découlent peuvent être classées en trois catégories :

1 Dans les cas les plus simples tels que l'estimation d'un taux d'AI, les estimations sont faites sans prise en compte des homozygotes, alors qu'il existe des solutions analytiques satisfaisantes qui devraient donc être utilisées.

2 Dans certains cas, il n'y a pas actuellement de solutions satisfaisantes, comme par exemple lorsque l'on cherche à réaliser des modèles multivariés prédictifs (régression logistique ou modèle de Cox). Plus précisément, il existe de nombreuses solutions, dont certaines, telle que l'imputation multiple, sont efficaces mais qui ne sont pas utilisées. Le problème se complexifie lorsque l'analyse implique des matrices ou le nombre de variables dépasse le nombre de sujets, auquel cas les auteurs tendent généralement à simplifier le problème en réduisant l'espace des variables le plus souvent de manière arbitraire.

3 Dans certains cas, les auteurs utilisent de « fausses bonnes » solutions, telles que les analyses en cluster où le calcul de la matrice des distances par couples de variables peut se faire sur les données présentes dans chaque couple : on aboutit parfois à des matrices mal conformées, même si en général cela ne pose pas de problème pour les analyses en cluster. En revanche, si l'on adjoint une analyse en composante principale à une analyse en cluster [103] l'ACP peut échouer en raison de cette matrice non conforme. Par ailleurs, dans la plupart des logiciels gérant les données manquantes de cette façon (comme Cluster et Treeview), le traitement se fait de façon transparente pour l'utilisateur qui n'est alors pas conscient des problèmes et des hypothèses posées lors de l'analyse.

4.4 Une classification des méthodes d'analyses en présence de données manquantes

Face au problème des DM, un nombre important de méthodes statistiques a été mis au point. On peut classer ces méthodes en quatre grandes catégories selon leur façon de traiter les DM.

4.4.1 Les méthodes sur données observées

La technique du cas complet Cette technique est une des plus simples qui puisse être envisagée. Elle consiste à ne travailler que sur les unités statistiques (les lignes de la matrice) qui sont complètes.

Les défauts de cette méthode Cette méthode est à éviter pour deux raisons. D'abord, elle créé plus de données manquantes qu'il n'en existe dès lors que l'on utilise plus d'une variable dans l'analyse.

En effet, les n sujets pour lesquels on observe par exemple m_1 valeurs manquantes sur une variable X_1 ont donc $(n-m)_1$ valeurs observées pour cette variable. Pour une variable X_2 il y a m_2 valeurs manquantes et $(n-m)_2$ valeurs présentes. A moins que les manquants ne soient en nombre et en positions identiques pour les deux variables, le fait de ne retenir que les valeurs non-manquantes induit une perte de valeurs parmi les valeurs présentes puisque l'on ne conserve que $r = min((n-m)_1, (n-m)_2)$ valeurs. Le nombre de valeurs non manquantes supprimées de l'analyse vaut donc : $max((n-m)_1, (n-m)_2) - r$. Cette valeur est non nulle dès que $m_1 \neq m_2$. Ceci accroît donc la proportion de valeurs

manquantes. L'effet peut être majeur en fonction des valeurs de m_1 et m_2 ainsi que de leurs répartitions respectives dans la matrice de données.

Voyons un exemple extrême. Un nombre faible de données manquantes peut avoir un effet drastique en réduisant considérablement le nombre de données présentes. Pour un nombre p de variables, il suffit de p valeurs manquantes réparties sur l'ensemble de la matrice de façon à ce que chaque donnée manquante apparaisse chacune dans une des p colonnes pour supprimer p sujets de l'analyse. On perd donc $(p/n)\%$ des sujets alors que le nombre de valeurs manquantes n'est que de $p/(p*n)$. On peut quantifier l'impact en faisant le rapport des deux valeurs. Pour 100 sujets et 10 variables, s'il y a 10 valeurs manquantes sur 10 sujets différents, le nombre de sujets utilisables est de 90, soit 900 valeurs au lieu de 1000. Chaque valeur manquante induit donc 9 autres valeurs manquantes, soit une perte totale de 10% de l'information utilisable alors que l'information manquante ne représente que 1% des valeurs.

Enfin, la technique du cas complet introduit presque certainement un biais, les données présentes n'étant généralement pas un échantillon représentatif des données complètes.

La technique du cas disponible Cette technique consiste à tirer partie au maximum des données présentes. C'est le principe utilisé par défaut dans les logiciels de statistiques car il est facile à mettre en oeuvre en construisant une sous-matrice en retirant les lignes où au moins une valeur est manquante. Dans l'analyse statistique d'une expérience, on procède généralement à plusieurs analyses différentes sur des sous-matrices de taille $p' < p$. Soit i analyses A. On utilise pour une analyse A_i un nombre p_i de variables. La technique du cas disponible consiste alors à utiliser pour chaque analyse A_i, les sujets complets sur les p_i variables utilisées. On se retrouve alors en fait dans le cas de la technique précédente mais sur une sous-matrice p'. L'avantage de cette technique est qu'elle perd moins de sujets que la technique du cas complet. En effet, le nombre de valeurs utilisées sur l'ensemble des A_i analyses est plus important et le nombre de données présentes supprimées par la méthode est plus faible. Par contre, l'inconvénient majeur de la méthode est de créer des sous-matrices différentes pour chaque analyse, ce qui fait que les sous-bases de données ainsi obtenues ne sont pas comparables. De plus, lors de procédures ascendantes (comme lors de la modélisation pas-à-pas ascendant faite par l'utilisateur), la base change plusieurs fois de taille et de contenu, ce qui pose des problèmes sérieux d'interprétation des résultats.

4.4.2 Les techniques de pondération

Les différentes techniques de sondage utilisent toutes un plan de sondage faisant appel à des poids de sondage pour chaque unité échantillonnée. L'estimation d'un pourcentage, en l'absence de non réponse, utilise le poids de chaque individu échantillonné :

$$\sum(\pi^{-1}) \cdot y_i \ / \sum \pi^{-1} \qquad (1)$$

où la somme est calculée sur l'ensemble des unités statistiques échantillonnées et π est la probabilité d'inclusion de l'unité i dans l'échantillon et $\pi - 1$ est le poids pour l'unité i. Les procédures de pondération modifient les poids des unités i de manière à ajuster les estimations des paramètres en prenant en considération les non-réponses. L'estimateur (1) est remplacé par l'estimateur suivant (2) :

$$\sum(\pi_i \hat{p}_i)^{-1} y_i / \sum(\pi_i \hat{p}_i)^{-1} \qquad (2)$$

où la sommation est faite sur l'ensemble des sujets ayant répondu[1] et π est une estimation de la probabilité de réponse pour l'unité i, cette estimation étant en général la proportion de sujets répondant dans une sous classe de l'échantillon.

4.4.3 Les techniques de modélisation

Ces procédures consistent à définir un modèle pour les données partiellement manquantes et à baser les inférences sur la vraisemblance de ce modèle, les paramètres étant estimés à partir par exemple de la technique du maximum de vraisemblance. Ces approches ont plusieurs avantages, le premier étant la flexibilité. Par ailleurs, on évite l'utilisation de méthodes *ad hoc*, dans le sens où les hypothèses sous-jacentes au modèle peuvent être évaluées. Par ailleurs, sur de grands échantillons, l'utilisation de la dérivé seconde de la log-vraisemblance permet d'avoir des estimations qui prennent en considération la variabilité supplémentaire liée aux manquants. Dans l'une des parties suivantes, nous serons amenés à utiliser des estimations des données manquantes basées sur la méthode du maximum de vraisemblance pour déterminer la valeur attendue des effectifs dans chaque modalité de la variable étudiée. Dans le cas de données incomplètes, l'utilisation du maximum de vraisemblance consiste à maximiser la vraisemblance de l'estimateur sur l'ensemble des données

[1] On sous-entend ici que les sujets ont répondu à un questionnaire, cette méthode étant utilisée essentiellement dans le cadre des enquête par sondage.

complètes *et* incomplètes. L'obtention de ces estimations est décrite ci-dessous pour le cas d'une proportion et pour le cas d'une table 2×2.

Dans le cas de l'estimation d'une proportion, l'estimation du maximum de vraisemblance correspond à l'estimation faite sur les données complètes. L'estimation d'une proportion revient à estimer les effectifs et les proportions des cellules d'une table 1×2 de terme général n_i, ou p_i lorsque l'on s'interesse directement à la proportion. La valeur estimée en l'absence de données manquantes est p_i et la valeur estimée sur l'ensemble des données complètes et incomplètes est $\hat{p_i}$. Dans la situation présente, $p_i = \hat{p_i}$. L'estimation est donc directe.

Dans le cas d'une table 2×2, l'estimation procède de la façon suivante en utilisant un algorithme EM (Expectation-Maximisation) [92]. Soit deux variables catégorielles X et Y avec des modalités allant de 1 à I et de 1 à J respectivement. Les données sont formées de n observations dont n_p sont complètes et n_m sont manquantes. On admet que seule la variable Y présente des valeurs manquantes ce qui aboutit à une marge supplémentaire, extérieure au tableau contenant les données connues uniquement sur X. Les données sont présentées dans la table 5.

	Y_1	Y_2	$Total$	Marge sup.
X_1	n_{11}	n_{12}	$n_{1.}$	r_1
X_2	n_{21}	n_{22}	$n_{2.}$	r_2
Total	$n_{.1}$	$n_{.2}$	n_p	n_m

TAB. 5 – Données complètes et marge supplémentaire pour la méthode EM.

Little et Rubin montrent que par factorisation de la vraisemblance sur les données complètes et incomplètes, l'estimation d'une proportion $\hat{p}_{ij}$ du tableau ci-dessus, se fait par l'estimateur suivant :

$$\hat{p}_{ij} = \frac{n_{ij} + (n_{ij}/n_{i.})r_i}{n}$$

L'estimateur EM du maximum de vraisemblance distribue donc une proportion $(n_{ij}/n_{i.})$ des données non-classées r_i sur la cellule i, j. Lorsqu'une seule variable présente des données manquantes, comme c'est le cas pour la situation type retenue ici sur les données d'allélotypage, l'estimation est immédiate, c'est-à-dire que la convergence est atteinte à la première étape. Lorsque les deux variables présentent des données manquantes, l'obtention

du résultat peut nécessiter plusieurs étapes itératives avant convergence. Par ailleurs, cet estimateur justifie l'affirmation donnée au paragraphe suivant concernant l'estimation d'une proportion.

4.4.4 Les techniques d'imputation

Dans les techniques d'imputation, on remplace les valeurs manquantes par une valeur générée suivant un processus caractéristique de la méthode. La richesse des variantes possibles laisse entrevoir la difficulté de choisir une méthode systématiquement supérieure aux autres. Nous présentons ci-après les méthodes « historiques », les méthodes plus récentes étant détaillées par la suite pour les plus importantes ou seulement indiquées dans la bibliographie [34]

L'imputation d'une valeur unique Parmi les règles d'imputation unique, la plus simple, dans le cas de variables binomiales, consiste à attribuer à toutes les valeurs manquantes la même valeur de variable binomiale. La méthode est biaisée par construction. Il existe différentes variantes. Pour une variable binaire qualitative pouvant prendre soit la valeur a soit la valeur b, on donne à toute valeur manquante soit la valeur a, soit la valeur b. Toutes les solutions intermédiaires faisant varier la proportion de valeurs a attribuées sont possibles, la difficulté étant de choisir correctement cette proportion par rapport à la situation.

La technique du biais maximum Il s'agit d'un cas particulier de la méthode précédente. Dans les essais thérapeutiques, il est d'usage de réaliser une imputation un peu particulière qui consiste à considérer les données manquantes comme défavorisant le plus possible le nouveau traitement à l'étude. Dans ce cas, une conclusion significative de l'essai malgré ce handicap permet de conclure à l'efficacité du traitement. Cette méthode fonctionne uniquement pour le cas où le critère de jugement est manquant et non pas lorsque l'une des covariables est manquante.

Le *Last Observation Carried Forward* Cette méthode est fréquemment employée dans l'industrie pharmaceutique [29]. Elle consiste, dans une série de mesures répétées, à remplacer chaque valeur manquante par la valeur qui la précède immédiatement dans les mesures, attribuant donc à la valeur manquante t_i^* la valeur t_{i-1} qui est la dernière valeur présente [300]. Les limites de la méthode sont évidentes : faire l'hypothèse que les données

non-observées ne varient pas n'est pas biologiquement crédible. De plus l'impact sur la variance est du même type que pour toutes les méthodes d'imputation unique [60, 301].

Le hot-deck Le hot-deck consiste à remplacer la valeur manquante par une valeur présente issue des données d'un autre sujet apparié sur le sujet à compléter à partir des variables ayant des valeurs non manquantes pour le sujet à compléter.

Le cold-deck Le cold-deck consiste à remplacer la valeur manquante par une valeur présente issue de données obtenues sur une autre étude. L'absence totale de théorisation de cette méthode ainsi que son principe même rend son utilité très douteuse.

La substitution La méthode consiste à remplacer un sujet ayant des valeurs manquantes par un autre sujet similaire n'ayant pas de données manquantes. Le plus souvent, on réalise un tirage au sort d'une unité statistique supplémentaire et on recueille l'ensemble des valeurs. Cette technique est assez logique puisqu'elle utilise les propriétés de l'échantillonnage mais implique une logistique et des budgets d'étude relativement extensibles.

L'imputation de la moyenne non conditionnelle La méthode consiste à remplacer la valeur manquante par la moyenne des valeurs non manquantes de la variable. Elle comporte à la fois un biais et une augmentation artificielle de la puissance par une baisse de la variance de l'estimateur de tendance centrale.

L'imputation de la moyenne conditionnelle La méthode est un raffinement par rapport à la méthode précédente et elle consiste à imputer une donnée manquante en lui attribuant la moyenne des valeurs non manquantes de la variable par niveaux d'une variable de stratification. Le biais et la baisse de la variance sont diminués mais encore présents [51].

L'imputation par régression L'imputation par régression consiste à réaliser une régression de la variable à imputer sur un certain nombre de variables sans valeurs manquantes et à attribuer aux valeurs manquantes l'espérance conditionnelle du modèle. On peut éventuellement travailler sur une sous-matrice de données excluant certaines données manquantes. La méthode est alors utilisée de façon itérative et chaque valeur imputée est réutilisée pour réaliser l'imputation d'autres valeurs.

L'imputation par régression stochastique Le principe est identique au précédent mais l'on rajoute un aléa autour de chaque valeur pour éviter une baisse trop importante de la variance de l'estimateur. Cette méthode et la précédente ont le défaut de ne pas donner les mêmes résultats selon les variables utilisées dans le modèle initial et les suivants car l'ordre des variables pour lesquelles on va successivement réaliser l'imputation est arbitraire. Les résultats des régressions dépendent forcément de cet ordre ce qui donne des résultats variables selon les choix de l'utilisateur. Il est donc difficile d'estimer l'effet réel d'une telle méthode.

Les méthodes combinées Un grand nombre de méthodes ont été construites sur des combinaisons des méthodes précédentes [274]. Leur efficacité est difficile à apprécier.

En pratique, aucune de ces méthodes n'est utilisée en routine, exceptées la méthode du cas complet et la méthode du cas présent.

L'étude de sensibilité Le principe de l'étude de sensibilité vise à évaluer la sensibilité d'un résultat ou d'une conclusion en fonction des modifications apportées aux hypothèses de départ [288, 330]. Dans le cas des données manquantes, une étude de sensibilité vise à estimer l'impact des données manquantes sur le résultats en évaluant les variations apportées dans les résultats selon les hypothèses faites sur les manquants. L'étude de sensibilité n'est pas une modélisation des données manquantes mais elle constitue une première approche des données et des résultats permettant de se faire une idée de la nécessité ou non de réaliser une modélisation plus fine des données manquantes [24]. Dans le cas des données qualitatives, les études de sensibilité sont relativement simples à réaliser puisqu'il suffit de réaliser quelques imputations et d'analyser les données suite à ces imputations. Nous verrons plus loin quelques méthodes utilisant ce principe [63, 172, 174, 220, 273].

Les techniques bayésiennes Ces techniques s'appuient sur la conception bayésienne de la statistique qui place sur chaque valeur inconnue une densité de probabilité. Une valeur manquante étant une valeur inconnue, la méthode bayésienne considère la valeur manquante comme issue d'une distribution idoine. On pose alors un *a priori* sur les manquants et on intègre la vraisemblance du modèle sur l'ensemble de l'espace d'échantillonnage pour obtenir la loi *a posteriori* du paramètre, en incluant l'incertitude liée aux manquants. Nous discuterons plusieurs méthodes plus loin.

Les techniques d'imputation multiple L'imputation multiple a été inventée par Rubin ([283, 284, 285, 286, 287]) et développée par Schafer [291, 292, 293, 294, 295]. Différentes publications ont étudié et illustré cette méthode [14, 28, 104, 226, 282, 305]. Des alternatives ont également été proposées [178]. L'imputation multiple (IM) consiste à remplacer chaque valeur manquante par un vecteur de $M \geq 2$ valeurs imputées. La taille du vecteur imputé correspond donc au nombre d'imputations réalisées. On utilise ensuite les méthodes standards pour analyser les données complétées, combinant les données initialement observées et les données imputées. L'imputation multiple consistant à répéter l'imputation, on obtient donc M jeux de données que l'on combine lors des analyses. L'avantage de cette méthode est qu'elle permet de prendre en compte la variabilité supplémentaire liée à l'imputation, ce que ne permet pas l'imputation simple, qui considère comme connue la donnée imputée, ce qui n'est évidemment pas vrai. L'imputation multiple permet donc d'utiliser des méthodes standards (comme l'imputation simple) tout en tenant compte de l'incertitude sur les manquants liée à l'imputation. L'imputation multiple peut se faire soit avec un seul modèle soit avec plusieurs modèles pour les manquants. Dans ce second cas, l'imputation multiple permet de prendre en compte l'incertitude lié au modèle ce qui est un avantage remarquable sur toutes les autres méthodes d'analyse en présence de manquants. En effet, tout modèle n'est qu'une certaine représentation de la réalité[2] ([81]) et il est souvent difficile de valider ou de justifier un modèle dans une situation donnée, même lorsque les données sont complètes. En présence de données non observées, le modèle est constitué de deux parties : une portant sur les observées et une portant sur les manquants. Cette deuxième partie a le défaut majeur de ne pouvoir jamais être validée [283, 290]. Il est donc tout à fait opportun de pouvoir évaluer l'impact du modèle choisi sur les résultats des analyses, d'autant plus que les résultats peuvent différer d'un logiciel à l'autre [13].

L'imputation est réalisée par une méthode bayésienne, utilisant la prédiction prédictive *a posteriori* de Y_{miss}. Rubin[283] suggère le protocole suivant : Pour chaque modèle considéré, les M imputations sont M répétitions de la *posterior predictive distribution* de Y_{miss} c'est-à-dire la distribution *a posteriori* de Y_{miss}, chaque répétition correspondant à un tirage indépendant des paramètres du modèle et des valeurs manquantes. L'analyse d'un jeu de données ayant subi plusieurs imputations est assez directe : chaque jeu de données est analysé avec la méthode que l'on aurait utilisé en l'absence de données manquantes. Soit $\hat{\theta}_i, \hat{W}_i$, pour $l = 1, \cdots, M$, les estimations ponctuelles et leurs variances pour chacune des

[2]"All models are wrong, some are usefull", DR. Cox

M imputations réalisées. L'estimation combinée de ces $\hat{\theta}_i$ est :

$$\bar{\theta} = \sum_{i=1}^{M} \frac{\hat{\theta}_i}{M}$$

La variabilité associée à cette estimation a deux composantes : une variance intra-imputation $\bar{W}_M$et une variance inter-imputation B_M, avec :

$$\bar{W}_M = \sum_{i=1}^{M} \frac{\hat{W}_i}{M}$$

et

$$B_M = \frac{\sum(\hat{\theta}_i - \bar{\theta}_M)^2}{M-1}$$

La variabilité totale de $\bar{\theta}_M$ est alors :

$$T_W = \bar{W}_M + \frac{M+1}{M} B_M$$

La valeur de $\frac{M+1}{M}$ est un ajustement lié au nombre fini de tirage. Le paramètre θ a alors une distribution t de Student : $(\theta - \bar{\theta}_M)T_M^{-1/2} \sim t_\nu$, expression dans laquelle le nombre de degré de liberté ν est le suivant :

$$\nu = (M-1)\left[1 + \frac{1}{M+1}\frac{\bar{W}_M}{B_M}\right]$$

Ce nombre de degré de liberté est basé sur une approximation de Satterthwaite. Rubin fait remarquer que le rapport $\bar{W}_M/B_M$ est une estimation de la quantité $(1-\gamma)/\gamma$ dans laquelle γ est pour θ la fraction d'information manquante liée à la non réponse.

Cette méthode peut laisser croire qu'elle crée des données supplémentaires mais ce n'est pas le cas : elle estime les composantes manquantes de données incomplètes de sorte que ces données puissent être analysées en utilisant les méthodes standards pour données complètes [29].

Les méthodes d'imputation multiple basées sur la régression De nouvelles méthodes basées sur des variantes de l'imputation multiple ont été récemment présentées par Barnes [29]. Il s'agit des moindres carrés bayésiens, de la méthode de la correspondance prédictive moyenne, des résidus aléatoires locaux, du score de propension modifié et de la méthode du score de complétion. Ces méthodes cependant ne peuvent être utilisées que

dans le cadre de valeurs manquantes suivant un schéma dit monotone, c'est-à-dire que pour un sujet i, les valeurs sont manquantes à partir d'un temps t_j et pas pour les temps t_k tels que $k < j$. Cette structure n'étant pas celle des données d'allélotypage, nous ne décrirons pas plus ces méthodes.

4.5 L'imputation multiple en pratique : le `module CAT` de R

Le logiciel gratuit R dispose d'une vaste bibliothèque de modules (*package* dans le langage R) dont l'un, `CAT`, implémente la méthode d'imputation multiple pour des données catégorielles. Ce module est basé sur les techniques de simulation présentées dans le livre de Schafer [290]. Ce module s'applique sur des tables de contingences multivariées.

Les fonctions de `CAT` Le module `CAT` contient un certain nombre de fonctions qui sont détaillées ci-dessous, ainsi que leur rôle :

`prelim.cat` cette fonction permet de préparer le jeu de données contenant des données manquantes. Il formate les données pour qu'elles soient utilisables par les autres fonctions ;

`ecm.cat` cette fonction fournit une estimation du maximum de vraisemblance *i.e.* le mode *a posteriori* des probabilités de chaque cellule du tableau suivant un modèle log-linéaire hiérarchique. ;

`em.cat` cette fonction fournit une estimation du maximum de vraisemblance *i.e.* le mode *a posteriori* des probabilités de chaque cellule du tableau suivant un modèle log-linéaire saturé contrairement à `ecm.cat` ;

`dabipf` cette fonction réalise un *data-augmentation* [318] pour données catégorielles incomplètes. Il utilise les résultats de `ecm.cat` pour réaliser des tirages aléatoires dans la distribution *a posteriori* des observations des probabilités des cellules de la table sous un modèle log-linéaire hiérarchique ;

`da.cat` cette fonction réalise un *data-augmentation* pour données catégorielles incomplètes. Il utilise les résultats de `ecm.cat` pour réaliser des tirages aléatoires dans la distribution *a posteriori* des observations des probabilités des cellules de la table sous un modèle log-linéaire saturé, contrairement à `dabipf` ;

`imp.cat` cette fonction réalise une imputation unique de données manquantes dans un jeu de données pour une série de paramètres (probabilité de chaque cellule) fournie par

l'utilisateur. La valeur du vecteur de paramètre est issu d'une des fonctions `ecm.cat`, `em.cat`, `dabipf` ou `da.cat`. Elle fournit une matrice de données identique à la matrice de données de départ excepté que les données manquantes ont été remplacées par des imputations, ce qui permet de visualiser les valeurs imputées ;

`mi.inference` cette fonction permet de réaliser l'inférence globale sur l'ensemble des M imputations réalisées avec les fonctions précédentes. Son usage requiert donc la mise en place d'une boucle dans le programme, le nombre d'itérations de la boucle étant le nombre d'imputations désirées.

D'autres fonctions sont implémentées dans le module CAT mais elles sont d'intérêt secondaire.

L'enchaînement des instructions se fait alors suivant l'une de deux chaînes possibles selon que l'on utilise un modèle hiérarchique (sous-entendu non saturé) ou saturé.

Dans le premier cas, la chaîne des fonctions est la suivante :

`prelim.cat` → `ecm.cat` → `dabipf` → `imp.cat` → `mi.inference`

ou dans le second cas :

`prelim.cat` → `em.cat` → `da.cat` → `imp.cat` → `mi.inference`

4.6 Déterminer le mécanisme des manquants

Un autre objectif de l'analyse en présence de données manquantes est de chercher à déterminer le mécanisme des manquants et de savoir si ce mécanisme peut avoir influencé les résultats. Dans le problème présent, la question est de savoir si les homozygotes sont manquants de type MCAR, MAR ou MNAR. Le nombre de tests permettant de connaître le type de manquants est remarquablement faible dans la littérature [143, 225]. Il est cependant possible de proposer une modélisation des données telles que l'on puisse tenter d'estimer la probabilité de réponse de type AI et de type normal pour les sujets homozygotes. L'estimation de ces proportions pour chaque microsatellite permet de se faire une idée du mécanisme des manquants. Si les résultats sont similaires sur l'ensemble des microsatellites, cela renforce l'hypothèse retenue.

Un mécanisme de type MNAR doit être évoqué si le taux (vrai et non observé) d'homozygote est plus élevé parmi les AI car alors le fait d'être manquant dépendrait du fait d'être homozygote. On peut par ailleurs trouver un effet du microsatellite si le taux d'AI estimé parmi les homozygotes varie d'un microsatellite à l'autre.

Principe du modèle Soit un microsatellite i pour lequel on observe un taux d'homozygotie τ_i. Le taux d'AI global, estimé sur l'ensemble des sujets hétérozygotes et homozygotes pour le microsatellite est π. On note $\pi_0 = \Pr(HTZ|MS : N)$ la probabilité d'être hétérozygote sachant que le sujet n'est pas en AI et on note $\pi_1 = \Pr(HTZ|MS : AI)$ la probabilité d'être hétérozygote sachant que le sujet est en AI. On définit également les probabilités complémentaires $1-\pi_0 = \Pr(HMZ|MS : N)$ et $1-\pi_1 = \Pr(HMZ|MS : AI)$. Ces définitions sont résumées dans le tableau ci-dessous. Ce tableau se lit horizontalement.

sujet	HTZ	HMZ	total
N	π_0	$1-\pi_0$	$1-\pi$
AI	π_1	$1-\pi_1$	π
Total	$1-\tau$	τ	

Tab. 6 – Notation pour la détermination du caractère MCAR ou MNAR des homozygotes (HMZ) pour un microsatellite donné.

Alors, la probabilité d'être homozygote est égale à la probabilité d'être homozygote sachant que l'on est AI plus la probabilité d'être homozygote sachant que l'on est normal :

$$\tau = \pi \cdot (1-\pi_1) + (1-\pi) \cdot (1-\pi_0)$$

La probabilité d'être à la fois homozygote et en AI est calculée comme le produit de la probabilité d'être AI par la probabilité d'être homozygote sachant que l'on est AI :

$$s = \pi \cdot (1-\pi_1)/\tau$$

Cette probabilité est en fait le paramètre d'une loi binomiale définissant le nombre de sujets AI parmi les sujets homozygotes. On peut en déduire les effectifs de sujets étant à la fois normaux et homozygotes. Cette analyse est faite avec le programme N°6 donné en annexe, adapté de [76].

4.7 La méthode de Dellucchi

Le programme Delluchi [90] propose une alternative simple à l'analyse de données qualitatives incomplètes. Il réalise une énumération de toutes les configurations possibles de données manquantes puis après avoir combiné ces configurations avec les données complètes, il calcule pour chaque tableau complété la statistique d'intérêt. Sa démarche ne va

pas plus loin et cela revient donc à réaliser une étude de sensibilité dont le seul avantage est ici d'être exhaustive. Il ne s'agit pas vraiment d'une modélisation car elle ne tient pas compte de la probabilité d'apparition de chaque configuration possible de manquants et il ne fait aucune hypothèse sur cette répartition. Sa méthode est à rapprocher de celle de Hollis qui sera vue plus loin. Le programme de cette méthode est donnée en annexe (programme N°7).

Un exemple La figure 1 montre les résultats de cette méthode pour un microsatellites (D2S138) dans sa relation avec le stade d'Astler-Coller dans l'échantillon étudié. Parmi les 240 tables possibles, 27 sont significatives au seuil de 5%. La ligne rouge indique une p-valeur à 0,05. La valeur minimum de p est de 0,0033 et la valeur maximum de 1. La grande majorité des imputations possibles mène à un test non-significatif. L'absence de calcul de la probabilité de survenue de chaque imputation ne permet pas de savoir quelle est la probabilité que la relation entre le stade et le microsatellite soit significatif.

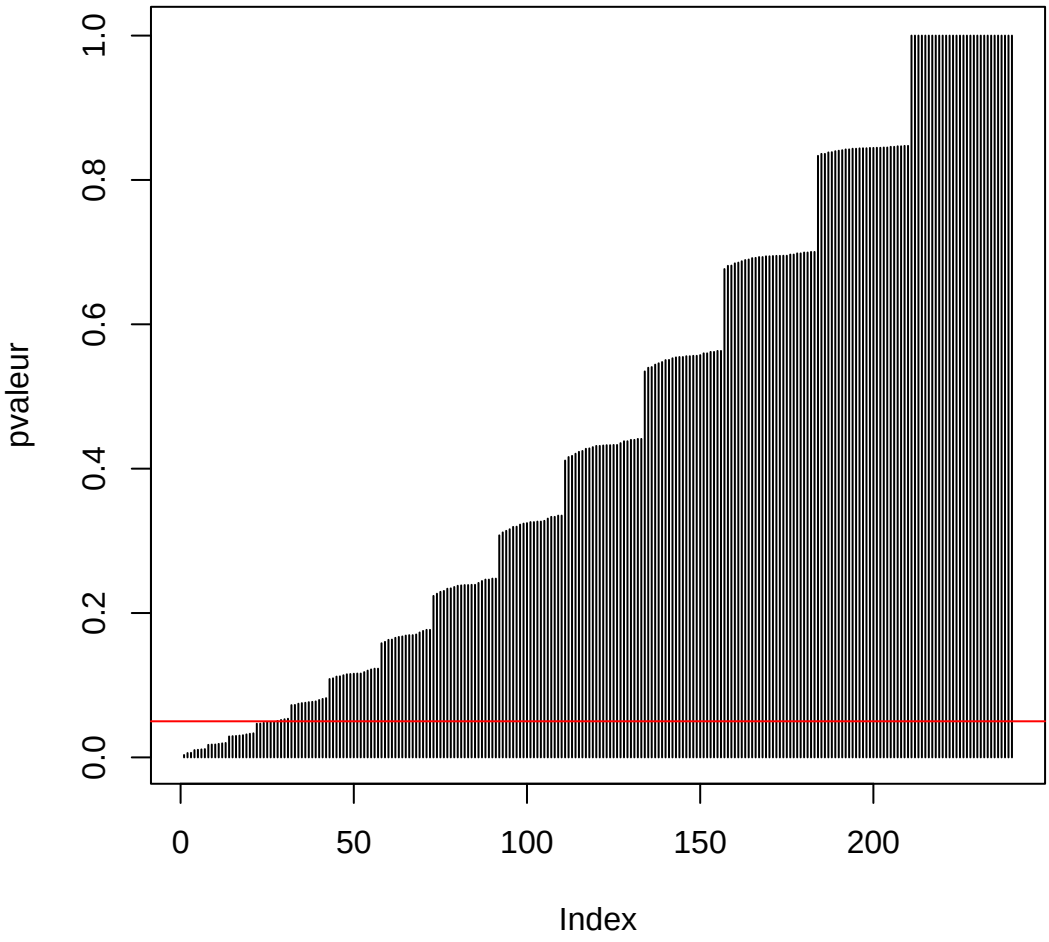

FIG. 1 – Analyse de la relation microsatellites D2S138-stade avec la méthode de Delucchi. La ligne horizontale rouge indique le seuil de significativité de 0,05.

4.8 La méthode de Shadish

Shadish propose d'étendre la méthode de Delucchi en faisant remarquer que la méthode de ce dernier ne permet pas de faire d'inférence et que de plus elle laisse l'utilisateur face à une série importante de résultats qu'il est difficile de visualiser et de calculer [299]. La remarque concernant la difficulté du calcul est un peu secondaire en raison de l'automatisation possible de calculs mais cela suppose tout de même une certaine maîtrise d'un langage de programmation, ce qui n'est pas forcément à la portée de tous les utilisateurs.

La méthode de Shadish consiste à calculer la liste des valeurs que peut prendre un paramètre d'intérêt tel que l'OR en énumérant toutes les tables complétées compatibles avec les données puis à calculer la probabilité que l'OR soit supérieur à 1, $\Pr(OR > 1)$. Bien entendu, cette probabilité n'est pas la simple proportion d'OR énuméré dont la valeur est supérieure à 1. Pour calculer cette probabilité il faut calculer la probabilité de chaque aspect de la table t ayant généré l'OR_t. Ce calcul repose sur des lois binomiales. Nous ne développerons pas plus cette méthode car elle est elle-même limitée par plusieurs éléments : Shadish ne propose pas d'intervalle de confiance de l'OR et sa méthode ne fonctionne qu'avec les hypothèses habituelles du test du χ^2 en terme d'effectifs. De plus l'aspect inférentiel suggéré par Shadish est ambigu : il sous-entend que la distribution des OR obtenus permet de conclure valablement sur l'OR mais il n'inclut pas vraiment l'incertitude liée aux manquants et s'il calcule la probabilité que $OR > 1$, il calcule en fait $\Pr(OR_i > 1)$, c'est-à-dire la probabilité que l'un de OR simulé soit supérieur à 1 alors que ce qui nous intéresse est la probabilité que l'OR de la table incomplète soit supérieur à 1, compte-tenu des données manquantes, ce qui n'est pas la même chose.

4.9 La méthode de Hollis

Une solution systématiquement préconisée pour l'analyse de données incomplètes est la réalisation d'étude de sensibilité qui consiste à faire varier les hypothèses sur les données manquantes afin de juger de l'impact de ces hypothèses sur les résultats de l'analyse. Dans le cadre des données qualitatives de type binaire et plus spécifiquement dans le cas d'un essai thérapeutique, Hollis propose une procédure de type étude de sensibilité avec une présentation graphique des résultats. Pour un facteur d'exposition et un résultat binaires, l'énumération exhaustive des résultats permet d'estimer l'influence des données manquantes sur les conclusions du tests. On peut considérer cette méthode comme un raffinement par rapport à la méthode du cas le plus défavorable, puisqu'il s'agit ici d'examiner tous les cas, du plus défavorable au moins défavorable. La méthode est très proche de celle de Dellucchi et elle en partage donc une des limites, à savoir l'absence de conclusion globale prenant en compte l'incertitude supplémentaire liée aux manquants. Le programme est mis à disposition sur internet par Hollis [147].

4.10 Une méthode « pré-bayésienne »

La méthode de Delucchi est relativement simple. Suite à une énumération exhaustive des configurations possibles des valeurs manquantes, on réalise autant de tests que de configurations possibles. On obtient donc autant de statistiques de tests et autant de p-valeurs. Cette méthode ne permet cependant pas de conclusion globale. L'incertitude liée aux données manquantes est en fait déplacée et reportée sur les p-valeurs qui présentent une certaine variabilité. Shadish calcule la probabilité $\Pr(OR > 1)$ en calculant la probabilité de chaque configuration de données manquantes sous une hypothèse donnée [299]. Il serait plus utile de pouvoir combiner ces comparaisons pour avoir une conclusion globale tenant compte de cette variabilité. Nous proposons ici une variante des méthodes de Delucchi et Shadish qui justement permet de réaliser cette inférence globale. Le principe de la méthode est basé sur une énumération exhaustive des configurations possibles avec calcul d'une zone de rejet de l'hypothèse nulle, pondérée par la probabilité d'apparition de chaque configuration sous certaines hypothèses [126].

La méthode peut s'appliquer à l'estimation de l'intervalle de confiance d'une proportion, à la réalisation de tests exacts de comparaisons de deux proportions ainsi qu'à l'estimation de l'intervalle de confiance d'un OR, les trois situations se comprenant évidemment en présence de données manquantes.

4.10.1 Formalisation du problème : imputation exhaustive

Plaçons-nous dans le cadre du calcul d'un pourcentage et de son intervalle de confiance exact en présence de données manquantes. Soit un vecteur V de n tirages d'une variable aléatoire de Bernoulli, dont certaines valeurs ne sont pas observées. Ce vecteur V est formé d'un vecteur V_m de longueur n_m contenant les valeurs manquantes et d'un vecteur observé V_o de longueur $n_o = n - n_m$. On a : $V = V_o + V_m$. Les valeurs du vecteur V_m peuvent se voir attribuer une valeur de manière systématisée suivant un modèle spécifié. Le nombre de façons d'attribuer des valeurs d'une variable aléatoire de Bernoulli à n_m sujets est de $A = \sum_{x=0}^{n_m} C_{n_m}^{x}$. La somme marginale du vecteur V peut alors prendre $n_m + 1$ valeurs distinctes suivant une loi binomiale de paramètre $(p_i; n_m)$. On cherche a estimer l'IC de p, proportion observée de valeur de V telles que $V = 1$, sur la combinaison de $n_m + 1$ vecteurs imputés $V_{imp,i}$ tels que :

$$V_{imp,i} = V_o + V_i \qquad \text{for } i \text{ in } 1, ..., m + 1$$

Dans le cas d'une table 2×2, les éléments sont les suivants : Soit un ensemble de n sujets et un couple de variables aléatoires X et Y de type binomial, où X est la variable explicative et Y la variable expliquée. Ces deux variables peuvent être croisées et le résultat placé dans un tableau T de taille 2×2 . Supposons par ailleurs que le vecteur X contenant les réalisations de la variable aléatoire X présente un certain nombre de données manquantes. Ce vecteur X est formé d'un vecteur X_m de longueur n_m contenant les valeurs manquantes et d'un vecteur observé X_o de longueur $n_o = n - n_m$. On a : $X = X_o + X_m$. On sépare alors l'ensemble des valeurs en deux groupes, les $n_o = n - n_m$ valeurs complètes (X et Y connues) et les n_m valeurs incomplètes, pour lesquelles on ne dispose que de la valeur de Y. On présente les $n_o = n - n_m$ valeurs complètes dans un tableau T_o (table 7).

Sujets	$Y = 1$	$Y = 0$	Total
$X = 1$	n_{11}	n_{12}	$n_{1.}$
$X = 0$	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

TAB. 7 – Tableau T_o

On présente ensuite les n_m valeurs manquantes dans un tableau supplémentaire Y_{sup} (table 8) formé d'une ligne et de deux colonnes contenant les n_{mqt} valeurs pour lesquelles Y est connu et X est inconnu. Les Y étant connus, on peut écrire $n_m = n_{..,m} = n_{.1,m} + n_{.2,m}$ avec $n_{.1,m}$ le nombre de sujets pour lesquels $(Y = 1; X = manquant)$ et $n_{.2,m}$ le nombre de sujets pour lesquels $(Y = 0; X = manquants)$. La notation $n_{..,m}$ est introduite ici pour indiquer que les n_m valeurs manquantes sont en fait à répartir sur l'ensemble des cases du tableau croisé de X et Y.

Sujets	$Y = 1$	$Y = 0$	Total
$X = *$	$n_{.1,m}$	$n_{.2,m}$	$n_{..,m}$

TAB. 8 – Tableau Y_{sup}

Cette marge des Y supplémentaires Y_{sup} peut être redistribuée sur la marge X_m des X manquants et complétée par imputation. Suivant le principe décrit dans la section sur les pourcentages, on peut, pour le vecteur X_m, attribuer des valeurs à chaque sujet aboutissant à $n_m + 1$ sommes marginales possibles pour les X_m. On admet pour les X_m une loi uniforme

ou binomiale ou autre. On croise ensuite les vecteurs X_m et Y_{sup} dans une table T_m (table 9). La table T_m est donc une table Y_{sup} « étalée » sur les quatre cases.

Sujets	$Y=1$	$Y=0$	Total
$X=1$	$n^*_{11,m}$	$n^*_{12,m}$	$n^*_{1.,m}$
$X=0$	$n^*_{2.,m}$	$n^*_{22,m}$	$n^*_{2.,m}$
$Total$	$n_{.1,m}$	$n_{.2,m}$	$n_{..,m}$

TAB. 9 – Tableau $T_m = Y_{sup}$ étendu

On note avec une astérisque les valeurs variant à chaque imputation. Les valeurs sans astérisque sont constantes tout au long du processus d'imputation exhaustive. Dans la table T_m, $n^*_{1.,m}$ varie de 0 à $n_m = n_{..,m}$ et $n^*_{2.,m}$ varie inversement de $n_m = n_{..,m}$ à 0. On note X^*_m le vecteur X_m pour une imputation particulière. Une fois les valeurs marginales de X^*_m fixées à $n^*_{1.,m}$ et $n^*_{2.,m}$, et s'agissant d'un tableau 2×2, il existe plusieurs tables compatibles avec ses marges X^*_m et Y_{sup}. Ces tables sont au nombre de $J = min(n^*_{1.,m}, n^*_{2.,m}, n_{.1,m}, n_{.2,m}) + 1$. Pour chaque imputation m de la marge des manquants, et donc chaque table générique T_m, on a donc $T_{j|m}$ tables, j allant de 1 à J. La probabilité de chaque table $T_{j|m}$ suit une loi hypergéométrique. On énumère donc systématiquement pour chacune des marges X^*_m toute les tables possibles $T_{j|m}$ à marge fixées. On obtient après imputation de toutes les tables $T_{j|m}$ pour toutes les marges X^*_m, un nombre de tables imputées égal à $F = \sum_{i=1}^{n_{..,m}+1}(min(n_{1.,m}, n_{2.,m}, n_{.1,m}, n_{.2,m})_i + 1)$. On les ajoute à la table des données complètes T_o pour obtenir une table finale T_f (table 10).

Sujets	$Y=1$	$Y=0$	Total
$X=1$	$n_{11} + n^*_{11,m}$	$n_{12} + n^*_{12,m}$	$n_{1.} + n^*_{1.,m}$
$X=0$	$n_{21} + n^*_{21,m}$	$n_{22} + n^*_{22,m}$	$n_{2.} + n^*_{2.,m}$
$Total$	$n_{.1} + n_{.1,m}$	$n_{.2} + n_{.2,m}$	$n_{..} + n_{..,m}$

TAB. 10 – Tableau T_f

Ces tables T_f sont également au nombre de F. Sur chaque table finale T_f on peut calculer un OR et son intervalle de confiance exact. On obtient une distribution des valeurs des OR qui peut être décrite comme n'importe quelle distribution. L'objectif de la méthode est de calculer un OR unique tenant compte de la variabilité supplémentaire liée aux données manquantes.

4.10.2 La méthode proposée.

Calcul dans le cas d'un pourcentage En l'absence de données manquantes, le calcul de l'intervalle de confiance exact d'un pourcentage repose sur l'utilisation de la loi binomiale. Soit t le nombre de succès dans n tirages indépendant d'une loi de Bernoulli. Soit p la vraie valeur du taux de succès. Alors les limites de l'intervalle de confiance au risque α % sont données par p_{inf} et p_{sup} telles que [73, 212] :

$$Pr(T \geq t|p_{inf}) = \frac{\alpha}{2} \tag{10}$$

et :

$$Pr(T \leq t|p_{sup}) = \frac{\alpha}{2} \tag{11}$$

En présence de données manquantes, ce pourcentage et son intervalle de confiance ne peuvent plus être calculés sans faire d'hypothèses sur la distribution des manquants. Dans la méthode de l'imputation multiple, on attribue plusieurs fois des valeurs aux données manquantes en se basant sur un modèle raisonnable pour ces données manquantes. On pousse le principe de l'imputation multiple au bout de l'idée pour aboutir à une technique d'imputation exhaustive des données manquantes. En suivant ce principe et après avoir spécifiée la loi de probabilité des données manquantes (uniforme, binomiale ou autre), on connaît la probabilité d'apparition de tous les vecteurs de données compatibles avec les données de départ. En utilisant la technique décrite dans le paragraphe précédent, un vecteur de taille $n = n_o + n_m$ subit $n_m + 1$ imputations, ce qui génère $n_m + 1$ vecteurs complétés, chacun de ces vecteurs ayant une probabilité $\Pr(m)$ d'être observé, cette probabilité $\Pr(m)$ étant la probabilité d'avoir une imputation donnée sous la loi définie pour les données manquantes. Les imputations sont générées en utilisant une loi binomiale $\mathcal{B}(n, p)$ dont le paramètre $n = m_m$. Le paramètre p est spécifié par l'utilisateur. On peut bien sûr utiliser une autre forme de distribution comme par exemple une loi uniforme lorsqu'aucune hypothèse ne peut être formulée sur les manquants ou toute autre distribution pertinente dans le contexte. L'imputation multiple permet alors d'obtenir une inférence globale sur le paramètre ou le vecteur de paramètre d'intérêt. S'agissant d'une imputation exhaustive, on connaît la loi de distribution des valeurs et on peut donc obtenir un paramètre et son intervalle de confiance en cherchant la valeur du paramètre qui vérifie les équations (3) et (4) sur l'ensemble des imputations. On cherche donc p_{inf} et p_{sup} telles que :

$$\sum_{m=1}^{n_m+1} Pr(T \geq t|p_{inf}, m) \cdot Pr(m) = \frac{\alpha}{2} \tag{12}$$

et :

$$\sum_{m=1}^{n_m+1} Pr(T \leq t|p_{sup}, m) \cdot Pr(m) = \frac{\alpha}{2} \tag{13}$$

où m est le numéro de l'imputation, pour m allant de 1 à $n_m + 1$, $n_m + 1$ étant le nombre total d'imputations possibles et Pr(m) est la probabilité de l'imputation m sous la loi de distribution préalablement choisie. Si on admet une loi binomiale pour Pr(m) on a alors :

$$\sum_{m=1}^{n_m+1} Pr(T \geq t|p_{inf}, m) \cdot C_{n_m}^{m-1} \cdot p^{m-1} \cdot (1-p)^{n_m-m+1} = \frac{\alpha}{2} \tag{14}$$

et :

$$\sum_{m=1}^{n_m+1} Pr(T \leq t|p_{sup}, m) \cdot C_{n_m}^{m-1} \cdot p^{m-1} \cdot (1-p)^{n_m-m+1} = \frac{\alpha}{2} \tag{15}$$

La recherche de la valeur se fait par une méthode convergente telle que la méthode des simplex. Cette méthode permet donc d'obtenir un intervalle de confiance exact de la proportion estimée, intervalle tenant compte de la variabilité supplémentaire liée aux données manquantes.

Application au cas de l'Odds-Ratio Le calcul de l'OR et l'obtention d'un IC exact par la formule de Cox ont été présentés dans un paragraphe précédent.

Soit une étude pour laquelle on souhaite analyser le rôle d'un facteur F sur l'apparition d'une maladie M. On peut résumer les données dans un tableau de la façon suivante :

Sujets	Malades	Sains	Total
Exposés	n_{11}	n_{12}	$n_{1.}$
Non Exposés	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

TAB. 11 – Tableau pour le calcul de l'OR par la formule de Cox

Comme dans le cas d'une proportion, on peut obtenir un intervalle de confiance exact pour un OR. Avec les notations du tableau ci-dessus et en l'absence de données manquantes, le calcul de cet intervalle est basé sur une formule donnée par Cox [80].

On retient ici comme convention d'écriture que, pour des marges fixées, la valeur observée de n_{11} peut varier, conditionnellement aux marges, entre n_{min} et n_{max}. Par ailleurs, on note N, la variable aléatoire associée au nombre de malades exposés. On cherche à estimer

β tel que $OR = e^{\beta}$. Alors :

$$Pr_N(n_{11};\beta) = \frac{\binom{n_{.2}}{n_{1.}-n_{11}}\binom{n_{.1}}{n_{11}}e^{\beta n_{11}}}{\sum_u \binom{n_{.2}}{n_{1.}-u}\binom{n_{.1}}{u}e^{\beta u}} \tag{16}$$

Dans cette équation, u varie de $n_{11,min}$ à $n_{11,max}$. Les queues de distribution supérieure $Pr_N^{sup}(n_{11};\beta)$ et inférieure $Pr_N^{inf}(n_{11};\beta)$ se calculent en sommant respectivement $Pr_N(n_{11};\beta)$ de n_{11} à $n_{11,max}$ pour la queue supérieure et de $n_{11,min}$ à n_{11} pour la queue inférieure. Pour trouver les valeurs des deux bornes p_{inf} et p_{sup} de l'intervalle de confiance de p, on parcourt l'espace de définition de p pour trouver les valeurs telles que :

$$Pr_N^{sup}(n_{11};\beta|p_{inf}) = \frac{\alpha}{2} \tag{17}$$

$$Pr_N^{inf}(n_{11};\beta|p_{sup}) = \frac{\alpha}{2} \tag{18}$$

On peut ici aussi calculer un intervalle de confiance prenant en compte la variabilité supplémentaire liée à des données manquantes. La méthode est la même que pour les proportions et on aboutit donc pour les formules (10) et (11) à une modification similaire à celle réalisée sur les formules (3) et (4). Après avoir spécifié une loi de distribution pour les données manquantes, on cherche p_{inf} et p_{sup} telles que :

$$\sum_{m=1}^{n_m+1} Pr(n_{11};\beta|p_{inf},m)\cdot \Pr(m)\cdot \Pr(T_{j|m}) = \frac{\alpha}{2} \tag{19}$$

$$\sum_{m=1}^{n_m+1} Pr(n_{11};\beta|p_{sup},m)\cdot \Pr(m)\cdot \Pr(T_{j|m}) = \frac{\alpha}{2} \tag{20}$$

4.10.3 Deux exemples

Intervalle de confiance d'un pourcentage Nous donnons deux exemples : l'un montre le détail des calculs et l'autre présente une version graphique des résultats pour un microsatellite donné. Soit un vecteur binaire $V = (V_o, V_m)$ avec $V_o = \{1,1,1,1,0,0,0,0,0,0,0\}$. On a alors $n_o = 11$, $n_m = 4$ et $p_o = 4/11$. Le tableau 10 présente les données, les imputations possibles pour le vecteur des données manquantes, les intervalles de confiance exacts de chaque proportion intermédiaire, la probabilité de chaque imputation suivant une loi $\mathcal{B}(11, p_o = 4/11)$ et l'intervalle de confiance exact global de la proportion $p = 36,36\%$ est [11,66 - 67,25].

V_o	possible V_m	p_{imp}	% CI	$\Pr(P = p_{imp})$
$1,1,1,1,0,0,0,0,0,0,0$	$0,0,0,0$	$4/15$	$[7,79 - 55,10]$	0,164
$1,1,1,1,0,0,0,0,0,0,0$	$1,0,0,0$	$5/15$	$[11,82 - 61,62]$	0,375
$1,1,1,1,0,0,0,0,0,0,0$	$1,1,0,0$	$6/15$	$[16,34 - 67,71]$	0,321
$1,1,1,1,0,0,0,0,0,0,0$	$1,1,1,0$	$7/15$	$[21,27 - 73,41]$	0,122
$1,1,1,1,0,0,0,0,0,0,0$	$1,1,1,1$	$8/15$	$[26,59 - 78,73]$	0,018
		global CI	$[11,66 - 67,25]$	

TAB. 12 – Exemple détaillé du calcul de l'IC avec données manquantes pour une proportion

Pour le second exemple d'application sur une proportion avec données incomplètes, on souhaite estimer la proportion d'AI dans le cas du microsatellite D2S138 dans le tissu cancéreux colique. Le nombre de patients[3] est de 37, le nombre d'homozygotes est de 9 et le nombre d'AI observé est de 17. La proportion de données manquantes est donc de 24,3%. La proportion observée d'AI est de $p_o = 17/(37-9) = 0,6071$, *i.e.* 60,71%. L'intervalle exact de ce pourcentage est : $[40,58 - 78,50]$. En appliquant la méthode proposée, avec comme distribution pour les manquants une loi uniforme, on obtient un IC exact de $[35,43 - 79,17]$, soit un intervalle plus large de 5,82%. En utilisant comme densité de probabilité pour les manquants une loi binomiale $\mathcal{B}(9, p_o = 17/28)$, on obtient un IC exact de $[41,56 - 77,89]$ soit un intervalle plus étroit de 1,59% que l'intervalle de p_o. Le graphique 2 montre l'énumération des intervalles de confiances pour chaque imputation ainsi que l'estimation globale de l'intervalle de confiance. Le graphique montre les intervalles de confiance intermédiaires calculés sur chacun des vecteurs imputés (lignes horizontales fines) ainsi que l'intervalle de confiance global (barre verticale épaisse).

[3]Les données utilisées ici sont issues d'une autre base de données obtenue sur 37 sujets porteurs d'un cancer colorectal pour lesquelles l'allélotypage a été réalisé de manière synchrone sur la tumeur primitive et sur une métastase [337].

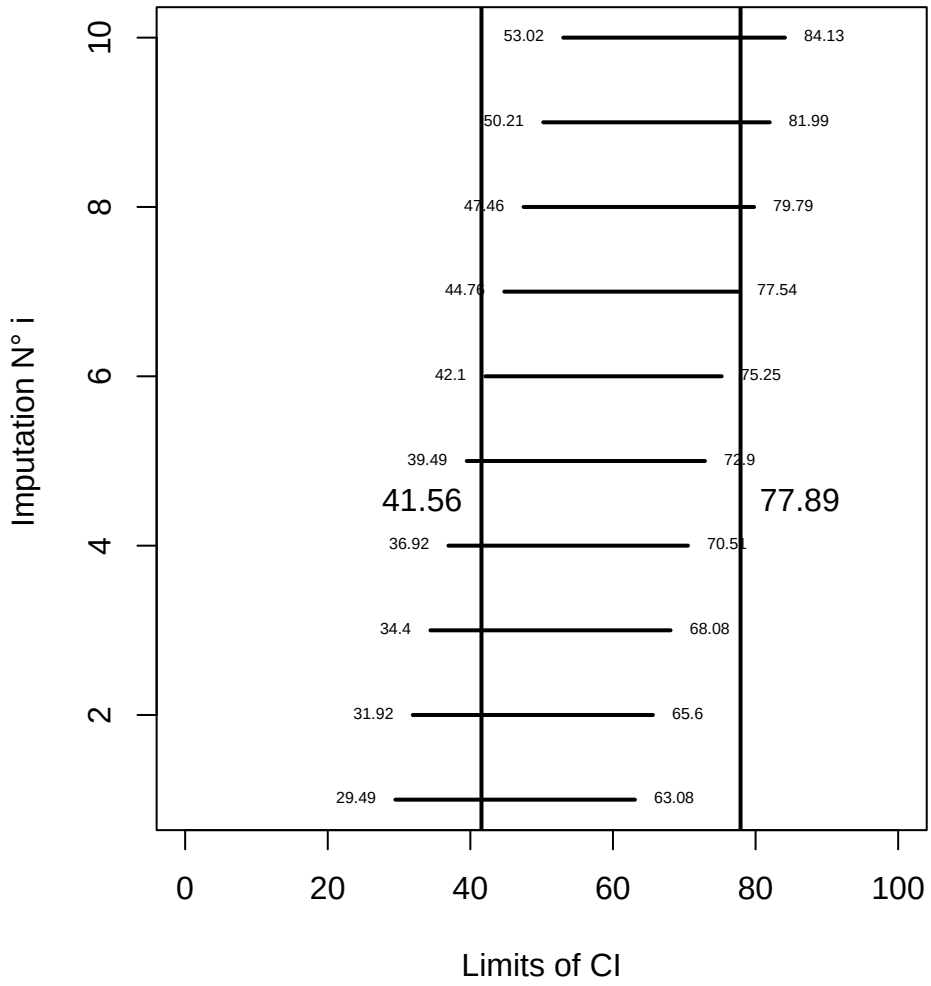

FIG. 2 – Estimation de l'intervalle de confiance du taux d'AI pour D2S138 après prise en compte des homozygotes.

Intervalle de confiance d'un Odds-Ratio en présence de données manquantes.

Une question d'intérêt pour les biologistes est de savoir si la présence d'AI sur l'un des microsatellite est associée au type diffus ou local de la tumeur [4]. L'association entre ces deux éléments est ici quantifiée par la valeur de l'OR. On obtient le tableau 2×2 suivant :

[4]On utilise également ici les données issues de la base des métastases synchrones

Sujets	Diffus	Local	Total
$X = AI$	4	13	17
$X = N$	6	5	11
Total	10	18	28

TAB. 13 – Tableau croisé pour D2S138 et le type de la tumeur.

La table supplémentaire des données manquantes contient les valeurs suivantes :

Sujets	Diffus	Local	Total
$X = *$	6	3	9

TAB. 14 – Tableau croisé pour D2S138 et le type de la tumeur : marge supplémentaire.

Cette marge est redistribuée sur la marge des X. On obtient ainsi les marges X supplémentaires possibles suivantes : $(X_1 = 0, X_0 = 9)$ à $(X_1 = 9, X_0 = 0)$ avec des valeurs intermédiaires de type : $(X_1 = 5, X_0 = 4)$. Ces marges supplémentaires doivent être additionnées aux marges X observées pour obtenir les marges X imputées montrées dans la table 15.

imputations i	1	2	3	4	5	6	7	8	9	10
X_i=1	17	18	19	20	21	22	23	24	25	26
X_i=0	20	19	18	17	16	15	14	13	12	11
Total	37	37	37	37	37	37	37	37	37	37

TAB. 15 – Liste de toutes les marges X imputées pour le microsatellite D2S138.

Dans le cas d'une table 2×2, il faut également énumérer toutes les tables possibles pour un couple donné de marges X et Y. Par exemple, la marge X imputée $(2, 7)$ est énumérée sous toutes les tables 2×2 possibles (table 16). Alors, ces tables imputées sont additionnées à la table des valeurs complètes donnant l'une des trois tables montrées dans le tableau 17.

Imputation i	1			2			3		
	$Y=1$	$Y=0$	$Total$	$Y=1$	$Y=0$	$Total$	$Y=1$	$Y=0$	$Total$
X_i=1	0	2	2	1	1	2	2	0	2
X_i=0	6	1	7	5	2	7	4	3	7
Total	6	3	9	6	3	9	6	3	9

TAB. 16 – Exemple : Liste de toutes les tables étendues pour la marge X imputée $(2, 9)$ pour le microsatellite D2S138.

Imputation i	1			2			3		
	$Y=1$	$Y=0$	$Total$	$Y=1$	$Y=0$	$Total$	$Y=1$	$Y=0$	$Total$
X_i=1	4	15	19	5	14	19	6	13	19
X_i=0	12	6	18	11	7	18	10	8	18
Total	16	21	37	16	21	37	16	21	37

TAB. 17 – Liste de toutes les tables imputées pour le microsatellite D2S138 (données observées ajoutées à la table 16).

Pour cet exemple, une présentation graphique peut être donnée, d'une manière similaire à celle proposée pour l'intervalle de confiance d'une proportion. Pour l'OR, on obtient l'intervalle de confiance [0,02-2,09]. Le graphique 3 montre les intervalles de confiance intermédiaires calculés sur chaque table obtenue à chaque imputation (lignes horizontales noires) ainsi que l'intervalle de confiance global (barre verticale rouge).

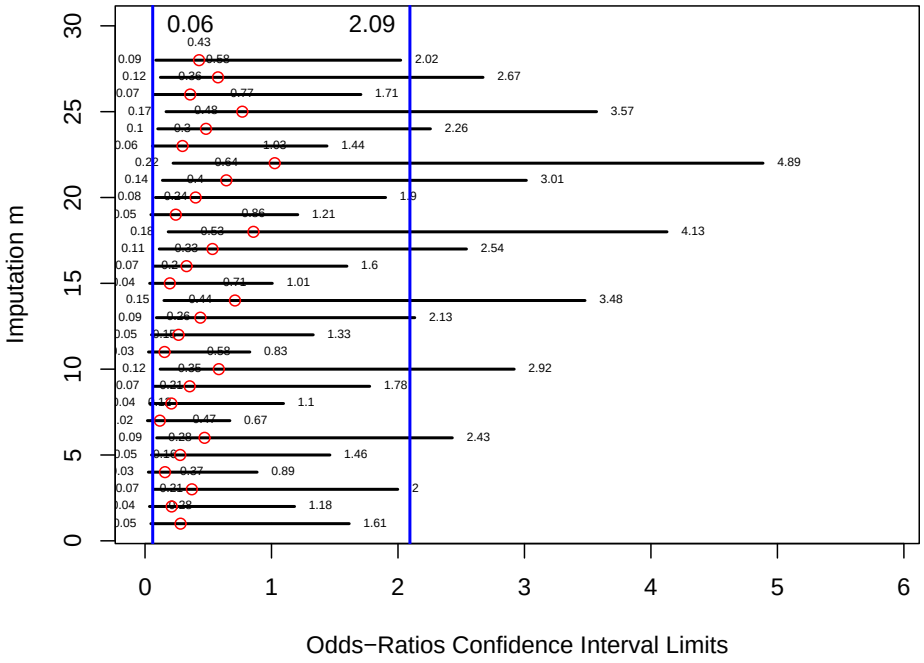

FIG. 3 – Estimation de l'intervalle de confiance exact de l'OR pour D2S138 prenant en compte les homozygotes.

Plusieurs des intervalles de confiance intermédiaires excluent la valeur « 1 » mais l'intervalle de confiance global inclut cette valeur.

Limites de la méthode La méthode proposée présente néanmoins un certain nombre de limites. En premier lieu, telle qu'elle est présentée, elle suppose que les données manquantes ont la même probabilité d'être en AI que les données observées. Même si cela ne semble pas infirmé par des analyses bayésiennes (voir plus loin), il s'agit d'une hypothèse forte qui peut cependant facilement être modulée. Il serait ensuite certainement plus pertinent de modifier la méthode pour définir une probabilité d'être en AI différente (éventuellement) selon les niveaux de la variable complète Y. Ensuite, en définissant cette probabilité (ou ces probabilités si l'on sépare l'estimation par niveau de Y), on fait l'hypothèse extrêmement forte que cette probabilité est connue, ce qui est douteux. Cela suppose en effet que l'estimation de cette proportion à partir des données dont on dispose est faite sans incertitude, ce qui ne peut pas être le cas. Cela introduit un *a priori unique* sur la valeur de cette proportion ce

qui permet de faire l'analyse mais qui revient à considérer que les données manquantes sont connues. Il faut donc introduire un aléa autour de cette valeur pour tenir compte de cette incertitude et utiliser une *distribution a priori* pour la valeur de la probabilité et non plus une valeur *unique*. Cela revient à réaliser soit une imputation multiple au sens de Little, Rubin et Schafer soit à utiliser une méthode bayésienne complète, avec une loi *a priori* sur le paramètre d'intérêt et sur les valeurs manquantes. Ces méthodes sont le sujet des paragraphes suivants. Cette méthode a été présenté à l'ISCB en 2002 [216].

5 Gestion des données manquantes dans les modèles bayésiens

L'imputation multiple suppose (voir paragraphe sur l'imputation multiple) que l'on puisse ajuster, de manière itérative, pour les données manquantes, un modèle obtenu à partir des données complètes de façon à imputer des valeurs crédibles aux valeurs manquantes tout en restant « réaliste », c'est-à-dire en conservant une variabilité dans les résultats. Ce principe se positionne très bien dans l'optique bayésienne. En fait l'imputation multiple est fondamentalement bayésienne (voir [284]). Il est donc naturel d'envisager ces modèles dans ce travail. La présentation des programmes montrera que la rédaction des modèles est plus naturelle en inférence bayésienne qu'en inférence fréquentiste.

L'inférence bayésienne suppose pour la plupart des modèles réalistes, donc d'une certaine complexité, le recours à la simulation par la méthode MCMC [43, 124]. Ces simulations peuvent être mises à profit dans le cadre de l'imputation multiple en étendant simplement le nombre d'imputations réalisées. En présence de données manquantes, l'analyse bayésienne traite les valeurs manquantes comme des paramètres inconnus à inclure en même temps que les autres paramètres du modèle. Ces données manquantes sont donc imputées à partir d'une distribution que l'on spécifie ou, plus rarement, dont on cherche les paramètres. Ainsi, si dans une imputation multiple habituelle le nombre d'imputations est de l'ordre de 3 à 10, dans une simulation basée sur un MCMC, il est aisé de réaliser plusieurs dizaines de milliers de tirages en quelques instants ce qui donne une approximation très fine de toutes les distributions et donc des distributions attribuées aux données manquantes. Il est donc naturel d'utiliser des méthodes MCMC pour réaliser de l'imputation multiple puisqu'elles sont le corollaire des méthodes bayésiennes auxquelles appartient l'imputation multiple. Les seuls éléments qui distinguent les versions MCMC de la version « habituelle » de l'imputation

multiple sont d'une part le nombre d'imputations et d'autre part l'outil informatique pour les réaliser mais fondamentalement, le concept est identique.

Dans un programme classique de WinBUGS, un certain nombre de sujets a des valeurs manquantes, représentées par des `NA` (« *Non-Available* ») dans la liste des valeurs. Les données manquantes doivent alors être soit modélisées, soit définies (remplacées) par une densité de probabilité *a priori*.

Plusieurs situations sont possibles.

Première situation On souhaite simplement estimer l'impact des valeurs manquantes sur les conclusions du modèle. Un exemple en sera donné pour chercher la présence d'un biais en vérifiant l'hypothèse MCAR que l'on peut faire pour les AI parmi les homozygotes.

Deuxième situation On souhaite estimer un paramètre à partir des données (par exemple une proportion p à partir d'une série de valeurs binaires. Dans ce cas, on utilise une densité de probabilité *a priori* pour p, laquelle va être mise à jour par les données. Dans ce cas, les manquants seront pris en compte par une loi *a priori* propre qui sera combinée à la densité *a priori* des données non-manquantes. L'opération est répétée autant de fois que nécessaire pour chaque variable ayant des données manquantes. On pourra par exemple calculer un risque relatif sous une hypothèse de données manquantes en comparant les p de deux variables, chaque p étant obtenu par une combinaison des lois *a priori* sur les manquants et sur la proportion obtenue à partir des non-manquants.

Troisième situation On souhaite établir une relation entre deux variables dont l'une présente des valeurs manquantes. Le cas classique en est la régression logistique dans laquelle la variable explicative présente des données manquantes. Dans ce cas, soit on impute des valeurs aux données manquantes, en spécifiant une distribution *a priori* pour les manquants, soit on impute les valeurs manquantes à partir des données présentes, par exemple à partir de la variable expliquée. Un exemple sera donné plus loin. Il s'agit ici plus de modéliser les données manquantes que de donner une loi *a priori*, même si la modélisation revient *in fine* à spécifier une loi pour les manquants. La modélisation peut être extrêmement complexe si besoin.

Pour un certain nombre d'auteur, la solution bayésienne est utilisée sans pour autant directement modéliser les données manquantes : les solutions se rapprochent des méthodes

fréquentistes qui sont alors transposées dans un cadre bayésien [233, 234]

5.1 L'imputation multiple bayésienne

Les analyses avec des modèles d'imputation multiple se font ici avec le logiciel WinBugs. Les différentes étapes peuvent se décrire de la façon suivante :

1. créer un modèle qui prédise chaque donnée manquante à partir de n'importe quelle information disponible qui soit utile pour cette prédiction. On utilise ici le modèle le plus adéquat en fonction des besoins : loi *a priori*, régression linéaire, logistique, non linéaire, etc. ;
2. on utilise le modèle précédemment créé pour fabriquer un jeu de données complètes. Cette étape comporte deux sous-étape :
   2.a on tire une série de valeurs pour le paramètre du modèle utilisé,
   2.b on utilise la valeur de ce paramètre pour prédire la valeur de la donnée manquante ;
3. à chaque fois que l'on a créé un jeu de données complétées, on réalise l'analyse voulue du jeu de données en conservant la moyenne et l'écart-type de chaque paramètre d'intérêt ;
4. on répète les étapes précédentes entre 2 et 4 fois en principe mais d'éventuels problèmes de convergence dans WinBugs liés à l'utilisation des MCMC amène à réaliser en fait plusieurs centaines d'itération, ce qui ne modifie pas le principe de l'imputation multiple et n'a pas d'implication sur les résultats autre qu'une durée de calcul plus longue et une meilleure précision des estimations ;
5. pour obtenir l'inférence finale, c'est-à-dire la valeur du paramètre assorti de son intervalle de confiance tenant compte des DM, on calcule (voir imputation multiple) la moyenne des moyennes du paramètre, ainsi que la variance intra et inter imputation des écart-types.

Paramétrage de Winbugs Pour chaque programme, on utilise 100 000 updates après un burn-in de 5000 itérations, sur une seule chaîne. La convergence est vérifiée visuellement sans utiliser les tests habituels de Keweke ou de Gelman-Rubin modifiés par Brooks et Gelman [49], la convergence étant en général réalisée pour des modèles simples du type utilisé dans la gamme d'*a priori* utilisés. Les lois *a priori* sont dans tous les cas des lois uniformes $B(1, 1)$. Ceci permet de conserver une comparabilité des résultats bayésiens entre

eux ainsi qu'avec les modèles fréquentistes, notamment avec la procédure pré-bayésienne proposée.

5.2 Les différents modèles d'imputation

Nous utiliserons les versions bayésiennes des différentes méthodes classiquement utilisées en présence de données manquantes. Les principes sont les mêmes lorsque l'on souhaite estimer une proportion (comme par exemple un taux d'AI pour un microsatellite) et lorsque l'on souhaite estimer un OR, comme l'OR quantifiant l'association entre un microsatellite normal ou AI et un résultat comme le décès du patient ou le stade du cancer.

Ces différentes méthodes sont les suivantes :

1. la méthode du cas complet : on supprime les sujets ayant des valeurs manquantes ;
2. les manquants sont remplacés par la modalité de référence ;
3. les manquants sont remplacés par l'autre modalité ;
4. la répartition des modalités est la même parmi les manquants que parmi les observés ;
5. la répartition des modalités suit la répartition obtenue par EM ;
6. les valeurs manquantes sont imputées suivant une loi de Bernoulli dont le paramètre est fixe ;
7. les valeurs manquantes sont imputées suivant une loi de Bernoulli dont le paramètre est variable et suit une loi uniforme, comme par exemple une $Be(1,1)$;
8. les valeurs manquantes sont imputées suivant une loi de Bernoulli $Ber(p)$ dont le paramètre est variable et suit une loi non uniforme, en général une loi $Be(\alpha,\beta)$ où α et $\beta \neq 1$.

Il faut distinguer les cinq premières situations des suivantes. Dans ces cas là, les données sont imputées de manière unique et l'analyse procède comme si l'on connaissait les valeurs manquantes. La sixième situation n'apporte pas de variabilité bien que les données soient générées par une densité de probabilité car définir précisément la valeur de p, fixe, revient à définir la proportion de sujets ayant l'une des deux modalités. Les deux dernières solutions réalisent une vraie imputation multiple et intègrent dans l'estimation des proportions une certaine incertitude liée aux manquants. Elles consistent en fait à donner des hyperparamètres à p avec éventuellement des hyperparamètres de second niveau aux hyperparamètres de p.

5.3 Estimation d'une proportion

Dans cette partie nous donnons les programmes permettant d'estimer une proportion suivant différentes méthodes (analyse naïve, imputation simple, imputation multiple). Dans les exemples, nous prenons le cas d'une série construite pour l'exemple et formée de 150 sujets pour lesquels un déséquilibre allélique a été cherché pour un microsatellite, et dont 50 sont homozygotes. La proportion de sujets en AI est fixée à 50%, parmi les observés et parmi les manquants. L'estimation globale de la proportion d'AI vaut donc également 50%. Nous avons choisi des proportions identiques parmi les présents et les manquants pour pouvoir nous focaliser sur l'estimation de la variance de la proportion et ne pas avoir à tenir compte du biais éventuel de l'estimation ponctuelle. On note les données de la façon suivante :

Le vecteur des données est $\mathbf{x} = \{\mathbf{x}_c; \mathbf{x}_m\}$ où $\mathbf{x}_c$ représente les données observées complètes et $\mathbf{x}_m$ représente les données observées manquantes ou incomplètes. $\mathbf{x}$ est donc le vecteur de longueur n contenant les valeurs du n-échantillon observé.

Première remarque Dans le cas de l'estimation d'une proportion, les programmes de Winbugs comportent une particularité à prendre en considération : il n'est pas possible d'imputer les valeurs des x_i, notés `x[i]`. En effet, l'estimation de la proportion se fait à partir des valeurs des `x[i]`. Cette estimation se fait en spécifiant que les données observées sont issues d'une loi de Bernoulli de paramètre p, lequel paramètre va justement être estimé à partir des données. Si l'on impute les valeurs des `x[i]` manquants, cela revient à dire que ces valeurs sont issues de deux distributions différentes à la fois, ce qui n'est pas possible. Il faut donc contourner le problème en travaillant directement sur la valeur de p_m, valeur du paramètre de la loi de Bernoulli ayant généré les données manquantes. Comme p_m est une variable aléatoire continue, sa loi *a priori* est une loi Beta. Cette loi n'est en fait pas mise à jour par les données puisqu'elle représente les données manquantes ou plutôt ce que l'on pense connaître des données manquantes. La loi de p_m est combinée par un simple calcul avec la valeur de p, estimée elle sur les données complètes. La description de la répartition des valeurs de p_t qui combine la valeur de p et de p_m donne l'estimation de la proportion sur l'ensemble de l'échantillon à partir des données complètes et des données manquantes. De façon générale, la proportion estimée p_t se calcule de la manière suivante :

$$p_t = (p \cdot n + p_m \cdot m)/(n + m)$$

où n et m sont les effectifs respectifs des données complètes et manquantes. Les modèles incluant cette méthode sont décrits à partir du paragraphe 5.3.7.

Seconde remarque Dans la mesure où il existe une relation directe entre les paramètres d'une loi Beta et les données, il est logique d'introduire les valeurs manquantes dans l'estimation de la proportion à partir des lois *a priori* en répartissant l'ensemble des données manquantes sur les deux valeurs α et β de la loi de p_t. Cette méthode permet d'estimer facilement la valeur du paramètre d'intérêt. Cependant, elle ne permet pas d'introduire d'incertitude dans les données car, justement en raison de l'équivalence entre les données et les paramètres des lois Beta, cette méthode revient à dire que l'on connaît exactement les données manquantes, ce qui bien sûr n'est pas le cas. L'introduction des manquants *via* les lois *a priori* revient donc à réaliser une imputation unique.

Dans les loi *a priori*, il faut alors distinguer deux parties très différentes. La première partie concerne les connaissances *a priori* que l'on peut avoir sur les résultats. On peut avoir un *a priori* fort ou faible selon la situation. La seconde partie concerne la (mé-)connaissance que l'on a des manquants et qui donc modélise ces manquants. L'*a priori* final incorpore donc ces deux aspects. La partie sur les manquants ne peut faire autrement que de spécifier que les manquants sont en nombre égal au nombre observé de manquants. Ensuite, on peut avoir un *a priori* fort sur la valeur de p que l'on cherche à estimer, *a priori* valable pour l'ensemble des données, présentes et manquantes. On peut appeler cela la loi *a priori* générale. Donc, l'*a priori* incorpore ces deux données. Plus exactement, la proportion à estimer suit une $Be(\alpha, \beta)$ pour laquelle α (et β symétriquement) incorpore le « vrai » *a priori* (*a priori* général) valable pour l'ensemble des données (complètes et manquantes), valable qu'il y ait ou pas des données manquantes plus la modélisation des manquants, qui n'intervient que s'il y a des données manquantes. Il est intéressant de remarquer que plus l'*a priori* général est fort, moins l'*a priori* des manquants a d'importance. L'*a priori* général sert donc tout le temps, de façon générale alors que la partie de l'*a priori* pour les manquants ne sert que lorsqu'il y a des manquants.

5.3.1 Méthode du cas complet

La méthode du cas complet repose sur l'exploitation exclusive des sujets complets. Dans le cas d'un microsatellite, les estimations du taux d'AI ne se font que sur les hétérozygotes.

Les données et le modèle s'écrivent de la façon suivante :

$$\mathbf{x} = \mathbf{x}_c \text{ et } p \sim Ber(\pi)$$

Les résultats Comme attendu,les paramètres observés correspondent aux valeurs théoriques. La moyenne et la médiane sont égales à 0,5 et l'écart-type (notée `sd` ci dessous) vaut pratiquement 0,05. L'intervalle de confiance correspondant evidemment à la théorie.

Le programme N°8 donne les résultats suivants :

```
Résultats :
     node     mean     sd      MC error   2.5%     median   97.5%
     p       0.5002   0.04932 1.534E-4    0.4037   0.5002   0.5963
```

5.3.2 Méthode d'imputation simple N°1

Dans cette première méthode, on estime le taux d'AI en considérant tous les hétérozygotes comme normaux. Les données et le modèle s'écrivent de la façon suivante :

$$\mathbf{x} = \{\mathbf{x}_c; \mathbf{x}_m\} \text{ où } \mathbf{x}_m = \mathbf{0} \text{ et } p \sim Ber(\pi)$$

Le programme N°9 donne les résultats suivants :

Les résultats

```
     node     mean     sd   MC error   2.5%     median   97.5%
    p    0.3354   0.03824 1.191E-4     0.2628   0.3345   0.412
```

On obtient évidemment une estimation très biaisée de la proportion (qui est théoriquement de 0,5).

5.3.3 Méthode d'imputation simple N°2

On estime le taux d'AI en considérant tous les hétérozygotes comme AI. Les données et le modèle s'écrivent de la façon suivante :

$$\mathbf{x} = \{\mathbf{x}_c; \mathbf{x}_m\} \text{ où } \mathbf{x}_m = \mathbf{1} \text{ et } p \sim Ber(\pi)$$

Le programme N°10 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
p	0.6646	0.03822	1.189E-4	0.5877	0.6654	0.7372

On obtient évidemment une estimation très biaisée vers le haut de la proportion. Le biais est le symétrique par rapport à 0,5 de la valeur obtenue dans la variante précédente.

5.3.4 méthode d'imputation simple N°3

Les valeurs manquantes sont ici imputées suivant une répartition en normal et AI identique à celle obtenue par le maximum de vraisemblance, soit 50%.

Le modèle Le programme N°11 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
p	0.5	0.04041	1.189E-4	0.4211	0.5001	0.5789

On obtient le résultat attendu qui est une estimation égale à 0,5 avec un écart-type plus petit que dans la première situation (ou l'on supprimait les données manquantes) puisque les effectifs sont ici augmentés. On trouve bien que

$$\sigma = \sqrt{0,5 \cdot 0,5/150} = 0,0408$$

5.3.5 Méthode d'imputation simple N°4

Ici, on utilise les lois *a priori* pour introduire les manquants. Les données manquantes sont spécifiées directement *via* leur répartiion dans les paramètres des lois *a priori* et non pas dans les données comme ce fut le cas précédemment.

Le programme N°12 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
p	0.5	0.04047	1.196E-4	0.4209	0.5001	0.5789

Les résultats sont identiques (à l'erreur de simulation près) aux résultats précédents comme attendu puisque les deux façons de spécifier les valeurs des manquants sont équivalentes de point de vue de l'information : les valeurs introduites dans les lois *a priori* sont strictement équivalentes aux données observées. Les paramètres de la loi *a posteriori* sont donc les mêmes dans les deux cas.

5.3.6 Méthode d'imputation simple N°5

On estime le taux d'AI ici en raffinant la version précédente et en considérant que pour les homozygotes le taux d'AI suit une loi de Bernoulli de paramètre p suivant lui-même une loi $Be(\alpha, \beta)$ dont les paramètres sont précisés suivant les principes bayésiens. La loi *a priori* globale est une loi $Be(\alpha_T, \beta_T)$ où

$$\alpha_T = \alpha_m + \alpha$$

et

$$\beta_T = \beta_m + \beta$$

dans lesquelles α_m et β_m sont les paramètres de la densité de probabilité *a priori* attribuée aux valeurs manquantes et α et β sont les paramètres de la loi *a priori* des données non-manquantes. L'inférence globale se fait sur les données combinées aux lois *a priori*. Il faut noter qu'il n'y a pas réellement de données pour les valeurs manquantes mais uniquement des données imputées qui en fait ne comptent que *via* la loi *a priori* qui les détermine. Les données imputées ne sont pas incluses deux fois dans les calculs. Le taux d'AI suit donc *in fine* une loi $Be(\alpha_f, \beta_f)$ dont les paramètres valent :

$$\alpha_f = \alpha + \alpha_m + n_1$$

$$\alpha_f = \beta + \beta_m + n_0$$

On utilise le programme N°12bis qui en pratique est le même que le programme N°12 puisque l'écriture finale de cette forme d'imputation aboutit à un résultat identique. La seule différence tiens dans la spécification explicite des paramètres liés aux manquants avant de les additionner aux paramètres généraux de la loi *a priori*. Cette méthode donne évidemment exactement les mêmes résultats que la méthode précédente. Il est donc indifférents d'imputer de manière unique les valeurs manquantes en les spécifiant dans les données ou en les spécifiant dans les *a priori* du paramètre.

Dans cette variante, le taux d'AI peut également être défini individuellement pour chaque sujet en précisant les valeurs du paramètre *a priori* à mettre. La loi *a priori* globale est une loi $Be(\alpha_T, \beta_T)$ où

$$\alpha_T = \sum_{i=1}^{m} \alpha_i + \alpha$$

et

$$\beta_T = \sum_{i=1}^{m} \beta_i + \beta$$

Le programme correspondant serait une simple extension du programme précédent (N°12bis). Il suffirait d'introduire les valeurs individuelles de α et de β dans une loi *a priori* pour chaque unité statistique avant d'en faire la somme et d'introduire cette somme dans les paramètres de la loi *a priori*.

5.3.7 Méthode d'imputation probabiliste N°1

On estime le taux d'AI en considérant que pour les homozygotes le taux d'AI suit une loi de Bernoulli de paramètre $p = 0,5$ défini. L'*a priori* pour p est neutre. La méthode consiste ici à faire les calculs à partir des valeurs des proportions et pas à partir des x_i.

Le programme N°13 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
p	0.5002	0.04932	1.534E-4	0.4037	0.5002	0.5963
pt	0.5001	0.03288	1.023E-4	0.4358	0.5002	0.5642

On constate que l'intervalle de confiance de p_t est plus étroit que celui de p alors qu'il devrait être plus large.

5.3.8 Méthode d'imputation probabiliste N°2

Ici on spécifie moins précisément la valeur de p_m qui reste néanmoins égale à 0,5 en moyenne. Il faut noter que la valeur de p_m n'est plus dans les données (« list »). L'utilisation de $dbeta(25, 25)$ [5] autorise la valeur de p_m à varier légèrement autour de 0,5 ce qui introduit donc une certaine incertitude comme on peut le constater en notant l'élargissement de l'intervalle de confiance de p_t. En fait, on retrouve ici pour p_t les résultats de l'estimation obtenue avec les programmes de méthode d'imputation simple N°3 et N°4.

Le programme N°14 donne les résultats suivants :

[5]=Be(25,25) en langage Winbugs.

node	mean	sd	MC error	2.5%	median	97.5%
p	0.5001	0.04916	1.552E-4	0.4034	0.5001	0.596
pm	0.4999	0.06998	2.165E-4	0.3636	0.5	0.6359
pt	0.5	0.04019	1.254E-4	0.4213	0.5	0.578

5.3.9 Méthode d'imputation probabiliste N°3

Dans cette variante, le programme est identique au précédent excepté en ce qui concerne la loi *a priori* de p_m. Une représentation plus réaliste serait de considérer une ignorance totale sur la valeur de p_m et il faut alors admettre que la probabilité que chacune des valeurs manquantes de x puisse prendre n'importe quelle valeur entre 0 et 1 avec une même probabilité. On aboutit alors aux résultats suivants en omettant les points similaires à ceux du programme précédent : Le programme N°15 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
p	0.5	0.04925	1.669E-4	0.4034	0.5002	0.5962
pm	0.4995	0.2888	8.951E-4	0.02478	0.4979	0.9741
pt	0.4998	0.1019	3.208E-4	0.322	0.4995	0.6777

On note que l'intervalle de confiance de p_t est maintenant beaucoup plus large. Il présente même une largeur maximum traduisant une incertitude maximum sur les données manquantes. Ce modèle permet donc d'établir un intervalle de confiance d'une longueur maximale pour une proportion estimée en tenant compte des données incomplètes.

D'autres variantes sont évidemment possibles pour préciser plus ou moins d'une part l'information disponible sur le paramètre parmi les non manquants et d'autre par l'information disponible sur les manquants. Il est possible à l'extrême de préciser la valeur de p_m pour chaque donnée manquante, ce qui aboutirait à une valeur de p_t prenant en compte au plus près l'information disponible. La valeur de p_t se calculerait alors de la manière suivante :

$$p_t = (p \cdot n + m \cdot \sum_{j=1}^{m} p_{m_i})/(n+m)$$

5.3.10 Prise en compte des manquants dans le calcul du facteur de Bayes

Le calcul de la valeur du FB suppose bien entendu l'utilisation de lois *a priori* pour les paramètres. La méthode bayésienne réalise une mise à jour de la distribution *a priori* par les données observées et dans ce cas là, les données introduites dans la loi *a priori* ont le même poids que si elles étaient introduites par imputation dans les données observées. Il pourrait alors sembler relativement facile de tenir compte des données manquantes lors de ce calcul du FB. Il suffirait en effet de spécifier comme cela a été indiqué plus haut pour différentes situations, les données manquantes dans la loi *a priori* du paramètre. Mais la valeur du FB dépend dans le cas d'hypothèses composites de la loi *a priori* des paramètres et pas des seules données observées ce qui rend son interprétation parfois complexe [169]. Le fait d'introduire les manquants *via* les lois *a priori* ne prend pas en compte l'incertitude liée aux manquants de la même manière que par l'introduction des données manquantes directement dans les données observées par une imputation simple. En effet, si l'on ajoute les manquants directement dans les données, le FB n'a pas la même valeur que si on tient compte des manquants par les lois *a priori*. Supposons que l'on cherche à calculer le FB pour une table 2×2 contenant les valeurs 15, 17, 21 et 22. On utilise des lois *a priori* non informatives pour l'exemple. Dans la fonction BF dont le programme est donné plus haut, on introduit les valeurs des 4 cases du tableau ainsi que les 2×2 paramètres des lois *a priori*. Cet exemple montre que les trois calculs suivants ne donnent pas les mêmes résultats :

```
BF(15,17,21,22,1,1,1,1) -->  3,48
BF(16,18,22,23,1,1,1,1) -->  3,57
BF(15,17,21,22,2,2,2,2) -->  2,98
```

La première ligne est l'exemple de base. Dans la seconde ligne on met les manquants directement dans les données (en rajoutant une unité dans chaque case) et dans la troisième ligne on met les manquants dans les lois *a priori* (en rajoutant une unité dans les paramètres des lois *a priori*). On observe que la valeur du FB diminue par rapport à la deuxième ligne ce qui montre une plus grande incertitude. L'imputation de valeurs manquantes dans les données observées a un impact plus important que l'introduction des données manquantes dans les lois *a priori*.

5.4 Estimation de l'Odds-Ratio

L'estimation d'un OR peut se faire de différentes manières : soit en utilisant des lois binomiales soit en utilisant le modèle logistique, tel que cela a été décrit plus haut. Les résultats sont en général très proches. L'utilisation de l'une ou l'autre méthode, notamment lorsque l'on utilise WinBUGS dépend de la façon dont sont introduites les données.

Les méthodes utilisables sont les suivantes :

1 : on supprime les manquants ;

2 : on met les manquants en « 0 » ;

3 : on met les manquants en « 1 » ;

4 : on met les manquants en « 0 » ou « 1 » selon les proportions du maximum de vraisemblance ;

5 : on met les manquants en « 0 » ou « 1 » selon autre ;

6 : les manquants suivent une $Ber(p = p_f)$ de paramètre fixé à une valeur de référence ;

7 : les manquants suivent une $Ber(p), p \sim Be(1, 1)$;

8 : les manquants suivent une $Ber(p), p \sim Be(\alpha_{\text{EM}}, \beta_{\text{EM}})$;

9 : les manquants suivent une $Ber(p), p \sim Be(\alpha_{\text{autre}}, \beta_{\text{autre}})$;

10 : la proportion d'AI parmi les manquants p_m est modélisée dans chaque stade en utilisant la même méthode que pour une proportion. La valeur de l'OR est estimée à partir des deux valeurs p_{m1} et p_{m2} en faisant varier les paramètres de la loi de probabilité déterminant les valeurs de p_{m1} et p_{m2} ;

11 : les manquants sont modélisés par imputation multiple en utilisant une régression logistique à partir des Y (le stade) sur les données complètes.

5.4.1 Les données complètes

Pour les exemples sur la régression logistique, nous avons retenu le jeu de données suivant :

	$Y = 1$	$Y = 0$	Total
$X = 1$	45	30	75
$X = 0$	30	45	75
Total	75	75	150

Tab. 18 – Données complètes de référence pour la régression logistique.

Les données manquantes sont construites artificiellement en retirant 10 sujets de chacune des cases de ce tableau, soit un total de 40 sujets. Le programme N°16 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	2.412	0.8395	0.004849	1.185	2.278	4.426
b[1]	-0.4127	0.2368	0.001418	-0.8832	-0.4103	0.04574
b[2]	0.8239	0.3361	0.001976	0.1695	0.8233	1.488

L'estimation de l'OR est bonne, proche de la valeur attendue de 2,25, sachant que cette estimation est faite ici à partir de la médiane et non pas de la moyenne, la distribution de l'OR étant asymétrique. De plus, la médiane est ici prise comme une approximation du mode qui correspond dans le cas bayésien à l'estimation du maximum de vraisemblance.

5.4.2 Méthode du cas complet

Dans cette situation, on retire des données les sujets ayant au moins une valeur manquante. On se retrouve donc dans la situation proposée par défaut dans les logiciels « fréquentistes » tels que SPSS, STATA, SAS, ou MINITAB. Le programme de calcul de l'OR et son IC est donc le programme de base.

Le programme Le programme est identique au précédent, excepté pour les données dont on retire pour les X et les Y les 40 dernières valeurs. Les résultats sont les suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	3.385	1.426	0.007975	1.438	3.116	6.894
b[1]	-0.5694	0.2822	0.001587	1.132	-0.5648	-0.02548
b[2]	1.139	0.4002	0.00225	0.3634	1.136	1.931

L'estimation de l'OR est biaisée à la hausse comme on pouvait s'y attendre et l'intervalle de confiance s'est notablement élargi.

5.4.3 Méthode d'imputation simple N°1

Dans cette situation, on attribue aux manquants une valeur « 0 », soit en terme de microsatellite, la valeur « normale », ou absence d'AI.

Le programme Le programme est le même que dans le cas précédent sauf pour les données. On ne met pas de données manquantes mais des valeurs choisies *a priori* pour les X, ici la valeur « 0 ».

node	mean	sd	MC error	2.5%	median	97.5%
OR	2.594	0.9503	0.0045	1.232	2.434	4.89
b[1]	-0.3216	0.2082	9.869E-4	-0.733	-0.3203	0.08386
b[2]	0.8911	0.3522	0.001677	0.2089	0.8897	1.587

L'estimation est légèrement biaisée vers le haut et l'intervalle de confiance est modérément élargit.

5.4.4 Méthode d'imputation simple N°2

Dans cette situation, on attribue aux manquants la valeur « 1 », c'est-à-dire qu'on les considère comme des AI. C'est la situation inverse de la précédente.

Le programme Le programme est le même que dans le cas précédent sauf pour les données. On ne met pas de données manquantes mais des valeurs choisies *a priori*, ici la valeur « 1 ».

node	mean	sd	MC error	2.5%	median	97.5%
OR	2.604	0.9535	0.006838	1.235	2.441	4.901
b[1]	-0.5732	0.2827	0.002064	-1.139	-0.5696	-0.02949
b[2]	0.8951	0.3514	0.002545	0.2114	0.8924	1.589

L'estimation est légèrement biaisée vers le haut et l'intervalle de confiance est modérément élargi et l'on retrouve des résultats proches de la version précédente où tous les manquants étaient recodés en « 0 ». Ceci s'explique aisément par les propriétés de l'OR.

5.4.5 Méthode d'imputation simple N°3

Une solution naturelle consiste à répartir les manquants en « 0 » ou « 1 » selon une répartition pertinente mais fixe, par exemple proche de celle obtenue par maximum de vraisemblance sur la table 2×2 croisant la variable X et la variable Y.

Le programme Le programme est le même que dans le cas précédent sauf pour les données. On ne met pas de données manquantes mais des valeurs choisies *a priori*, ici les valeurs « 0 » et « 1 » suivant une répartition pertinente. Nous ne donnons pas de résultats ici.

5.4.6 Méthode d'imputation simple N°4

Il est courant d'utiliser une catégorie *de novo* pour les données manquantes, ce qui transformerait la variable binomiale X en une variable multinomiale. Quoique courante, cette méthode présente systématiquement un biais lequel peut parfois être important. Nous ne donnons pas de programme ni de résultat pour ce type de modèle qui ne constitue pas une véritable modélisation des données manquantes.

5.4.7 Méthode d'imputation probabiliste N°1

Dans cette situation, on attribue aux manquants non plus une valeur fixe mais une probabilité d'être en « 1 » égale à celle observée sur les présents. On réalise donc ici une prise en compte de l'incertitude liée aux manquants. On met une loi de Bernoulli de paramètre p ($p = 0,5$ dans l'exemple ci-dessous). Il faut par ailleurs donner des valeurs initiales ou les faire générer par le programme. Le programme N°17 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	3.39	1.422	0.00987	1.444	3.117	6.879
b[1]	-0.5706	0.2639	0.001805	-1.1	-0.5672	-0.0635
b[2]	1.141	0.3982	0.002811	0.3677	1.137	1.928

5.4.8 Méthode d'imputation probabiliste N°2

Dans la méthode précédente, la valeur du paramètre de la loi de Bernoulli définissant la probabilité pour chaque valeur de prendre la valeur « 1 » est fixe, ce qui manque de réalisme :

admettre qu'il est connu revient à admettre que l'on connaît (au moins asymptotiquement) la valeur des manquants, ce qui n'est évidemment en général pas le cas. Il faut donc impérativement apporter un élément d'incertitude dans la valeur de p, lequel élément est imposé en donnant une loi *a priori* sur le paramètre p. Ceci est implémenté dans le programme WinBugs ci-dessous en plaçant une loi *a priori* sur la valeur de p pour caractériser l'incertitude le concernant. Différentes solutions sont bien sûr possibles. La plus simple, spécifiant une incertitude maximum consiste à donner à p comme loi *a priori* une loi $Be(1;1)$.

Le programme N°18 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	3.147	1.31	0.008385	1.368	2.894	6.359
b[1]	-0.5312	0.2702	0.00199	-1.084	-0.5243	-0.02393
b[2]	1.069	0.3928	0.00256	0.313	1.063	1.85
pm	0.4925	0.2509	0.003977	0.04308	0.4901	0.954

5.4.9 Méthode d'imputation probabiliste N°3

On fait ici varier la valeur de p suivant une loi Beta dans les proportions EM.

Le programme Il est identique au précédent. Seul change le paramétrage de la loi *a priori* de p, soit $x[i] \sim Ber(p_m), p_m \sim Be(\alpha_{EM}, \beta_{EM})$.

5.4.10 Méthode d'imputation probabiliste N°4

On fait ici varier la valeur de p suivant une loi Beta dans des proportions autres, en précisant par exemple que la valeur de p est proche de 0,5 mais avec une certaine variabilité.

Le programme Il est identique au programme de la première méthode probabiliste avec comme changement, la ligne suivante :

```
pm~dbeta(20,20)
```

5.4.11 Méthode d'imputation probabiliste N°5

On utilise la même construction que dans le programme de la méthode d'imputation probabiliste N°1 pour une proportion. On obtient alors une densité de probabilité pour

chaque proportion estimée et on en déduit lors des simulations la densité de probabilité de l'OR correspondant. La méthode n'utilise donc pas une régression logistique mais des lois binomiales. Le programme N°19 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	1.424	0.9543	0.002135	0.3594	1.173	3.93
ORn	1.415	0.7223	0.001579	0.4943	1.261	3.235
pm1	0.4983	0.2887	6.514E-4	0.02478	0.4971	0.9744
pm2	0.5	0.2886	6.252E-4	0.02536	0.4999	0.975
pt1	0.4722	0.1179	2.715E-4	0.2595	0.4717	0.6868
pt2	0.5107	0.0902	2.742E-4	0.3367	0.5111	0.683

La méthode a l'avantage de placer une contrainte minimum sur la valeur de la proportion d'AI parmi les manquants et l'on obtient des intervalles de confiance pour l'OR particulièrement larges ce qui permet de les prendre comme des valeurs extrêmes des bornes inférieures et supérieures de l'OR, tenant compte des données manquantes.

5.4.12 Méthode d'imputation probabiliste N°6

Dans les modèles précédents, la proportion de manquants prenant la valeur « 1 » est spécifiée soit sans incertitude soit avec un certain niveau d'incertitude en utilisant une loi Beta mais dans tous les cas, les manquants étaient modélisés sans faire intervenir les autres données disponibles. Une autre possibilité consiste à faire intervenir ces autres données afin de tenir compte de la structure globale des données dans l'imputation. Cette technique consiste à considérer que les manquants dépendent d'une certaine manière des autres variables et à faire usage de cette dépendance pour réaliser le modèle voulu tenant compte d'une certaine incertitude sur les données.

Dans le cas des données d'allélotypage, faire usage des autres valeurs consisterait à utiliser soit les autres microsatellites soit d'autres variables telles que le stade du cancer, le type histologique, l'âge du patient, etc, ou encore une combinaison de ces différentes variables. Il semble peu pertinent voire impossible d'utiliser les autres microsatellites. Bien que séduisante sur un plan théorique et biologique, la méthode ne peut pas fonctionner car elle devient rapidement circulaire puisqu'elle consiste à imputer des valeurs aux données manquantes d'une variable MS_1 à partir d'une variable MS_2 présentant elle-même des données manquantes qui pourraient être imputées à partir des données de la variable MS_1.

Même s'il n'est pas impossible que l'algorithme converge, une telle procédure nécessiterait validation avant utilisation. Il est nettement plus cohérent de réaliser l'imputation à partir de variables complètes. Dans l'exemple ci-dessous, nous utilisons le stade de Astler-Coller pour réaliser l'imputation. Plus précisément, deux techniques complémentaires sont utilisées. La première est une technique déjà décrite visant à imputer des valeurs à partir d'une loi Beta *a priori* pour p, la probabilité pour un homozygotes d'être AI. La seconde réalise une imputation à partir de la valeur du stade. Il est bien sûr possible de ne réaliser l'imputation qu'avec la seconde technique.

Le programme N°20 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR16	0.8786	0.3789	0.001951	0.3596	0.8076	1.797
OR18	2.527	1.66	0.007214	0.7209	2.111	6.798
alpha16	0.5391	0.2921	0.001573	-0.02273	0.5352	1.123
alpha18	1.313	0.3467	0.001677	0.6664	1.302	2.025
beta.s16	-0.2141	0.4113	0.002195	-1.023	-0.2137	0.5863
beta.s18	0.7593	0.5712	0.002673	-0.3273	0.7469	1.917

Ces procédés permettent donc d'introduire un certaine incertitude sur la valeur des manquants. A partir de ces méthodes élémentaires, le nombre de variantes possibles pour réaliser l'imputation de données manquantes est quasiment sans limite, même si l'impact sur l'inférence ne diffère pas forcément beaucoup d'une variante à l'autre.

Les deux principales variantes possibles sont les suivantes : soit on spécifie une loi *a priori* pour les manquants de chaque microsatellite soit on spécifie un *a priori* global pour l'ensemble des manquants avec éventuellement, dans cette seconde option, des hyperparamètres. Les résultats ne sont pas forcément très différents mais les hypothèses sous-jacentes sont très différentes. Dans le premier cas on admet que chaque microsatellite a une évolution qui lui est propre et que le type statistique de manquant qui lui correspond (MCAR, MAR, MNAR et valeurs des paramètres régissant la loi des manquants) n'est valable que pour lui. Dans le second cas, au contraire, on admet que tous les microsatellites sont équivalents en ce qui concerne les homozygotes et donc les manquants. La vérité est sans doute à mi-chemin entre les deux, certains microsatellites ayant certainement un comportement qui leur est propre en terme d'homozygotie et d'autres ayant des comportements similaires.

La diversité des approches possibles est assez caractéristique des problèmes de gestion des données manquantes : elle reflète la diversité des hypothèses possibles sur le mécanisme des manquants. Par ailleurs, cette diversité rajoute un niveau de complexité à la modélisation globale des données. La détermination du meilleur modèle dans une situation donnée est relativement complexe comme l'ont montrée plusieurs auteurs. La prise en compte des données manquantes rajoute un degré supplémentaire de complexité dans cette modélisation, d'autant plus que l'on ne dispose pas de critère permettant de juger de l'ajustement du modèle aux données puisqu'une partie des données est absente. Il faut donc rester très prudent dans l'analyse d'un modèle lorsqu'il y a des données manquantes.

5.5 Méta-analyses des relations microsatellite-Stade

Les données d'allélotypage sont constituées d'une série de mesures sur un ensemble de microsatellites. Jusqu'à présent, dans ce travail, les microsatellites ont été considérés individuellement, indépendamment les uns des autres. Une analyse globale fait cependant partie des objectifs de ce travail. Il faut donc envisager d'analyser simultanément l'ensemble des microsatellites dans un modèle global. Une solution consisterait à utiliser une méta-analyse. Il existe deux catégories de modèles de méta-analyse : d'une part les modèles à effet fixe et d'autre part sur les modèles à effet aléatoire (hiérarchique). L'intérêt des modèles à effet aléatoire est présenté dans cette section.

On appelle modèle hiérarchique un modèle statistique dans lequel les paramètres d'intérêts individuels sont issus d'une distribution de niveau supérieur ayant une distribution dont la forme est à déterminer. Dans cette famille de modèles, on trouve les modèles mixtes dont les modèles à effets aléatoires fréquemment utilisés pour les méta-analyses. Ces modèles permettent d'obtenir une inférence globale sur un ensemble d'expériences i considérées comme indépendantes mais similaires de sorte que l'on puisse chercher un effet commun sur l'ensemble de ces expériences.

Une méta-analyse basée sur un modèle aléatoire suppose que les différentes études (les différents microsatellites en fait) dont on souhaite faire la synthèse présentent un effet doté d'une partie fixe, commune à toutes les études, à laquelle s'ajoute une partie aléatoire, variable, propre à l'étude (propre au microsatellite concerné).

Appliqué au cas présent, ce modèle suppose donc que chaque microsatellite possède un effet θ_i, que les θ_i sont des variables aléatoires distribuées normalement autour d'un effet commun θ avec une variance τ^2.

$$\theta_i \sim \mathcal{N}(\theta, \tau^2)$$

L'estimation $\hat{\theta}_i$ obtenue sur un microsatellite MS_i est distribuée autour de ce θ_i, aléatoire et spécifique du microsatellite MS_i, avec une variance σ_i^2 qui représente les fluctuations d'échantillonnage :

$$\hat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$$

Le but de la méta-analyse est d'estimer le paramètre θ. La distribution marginale des θ_i est :

$$\hat{\theta}_i \sim \mathcal{N}(\theta, \tau^2 + \sigma^2)$$

avec τ^2 la variabilité inter-microsatellites et σ^2 la variabilité intra-microsatellite.

Dans le cas d'une régression linéaire généralisée de type régression logistique, la partie aléatoire porte soit sur la pente, soit sur l'ordonnée à l'origine, soit sur les deux. Nous utilisons ici un modèle rajoutant un paramètre b aléatoire tel que

$$logit(p) = \alpha + \beta x_i + b_i$$

avec

$$b_i \sim \mathcal{N}(0; 1)$$

Ce modèle fait donc l'hypothèse d'une pente et d'une ordonnée à l'origine commune avec un résidu propre à chaque série de mesures donc ici à chaque microsatellite.

Dans le cas d'une méta-analyse à partir de régression logistique, on admet que chaque pente est indépendante (il y a une pente par expérimentation donc par microsatellite) mais que la valeur β_i de la pente pour le microsatellite MS_i est issue d'une distribution $\mathcal{N}(\mu; \sigma)$ dont les paramètres sont déterminés lors de l'analyse. Les valeurs de ces paramètres donnent une mesure de l'effet global exprimé ici en $\log(OR)$.

Pour mettre en place ces différents modèles, on peut considérer dans un premier temps que les microsatellites sont des réalisations *indépendantes* d'un même processus cancérogène qui pourrait donc être analysé en isolant les différentes réalisations les unes des autres. Cette modélisation serait donc la même que celle utilisée pour faire une synthèse d'essais thérapeutiques par une méta-analyse. Ici, le microsatellite MS_i représenterait une étude e_i

et la méta-analyse définirait un modèle global dans lequel e_i est une réalisation indépendante d'une variable aléatoire E dont la loi est à déterminer. Ce modèle suppose donc une indépendance des résultats ce qui biologiquement est peu pertinent mais qui permet au moins une première approche globale sur la cancérogénèse telle qu'elle peut être comprise par l'utilisation des microsatellites.

Différentes approches de complexité croissante sont envisagées.

Modèle N°1 Le premier modèle possible consiste à réaliser une régression logistique aléatoire avec définition d'un effet commun et d'une valeur de b_i pour chaque microsatellite MS_i. Les microsatellites sont analysés en n'utilisant que les données complètes. On peut faire un modèle simple pour cette situation qui peut être écrite de différentes façons dans Winbugs.

Le programme N°21 donne les résultats suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	1.138	0.09293	5.002E-4	0.9659	1.134	1.33
alpha	-0.08657	0.05712	3.133E-4	-0.1981	-0.08684	0.02502
beta	0.126	0.08154	4.412E-4	-0.03469	0.1261	0.2854
b[1]	0.001825	0.06409	1.989E-4	-0.1305	0.001034	0.1371
b[66]	0.004811	0.06488	2.333E-4	-0.1261	0.003044	0.1455

L'intervalle de crédibilité de l'OR est donc [0,97 - 1,33] qui chevauche la valeur 1. La probabilité que l'OR soit supérieur à 1 est de l'ordre de 0,90. L'effet est globalement faible. Mais ce modèle suppose cependant qu'un déséquilibre allélique amenant un OR négatif compenserait un déséquilibre allélique associé positivement (par un OR>1) au résultat (le stade). Ce résultat n'est pas pertinent biologiquement. Il faut donc considérer chaque écart à 0 de l'OR qu'il soit positif ou négatif comme un effet « positif ». Une variante intéressante consiste donc à rechercher plus généralement un effet (positif *ou* négatif) en inversant chaque fois que nécessaire les valeurs de la variable X pour n'obtenir que des pentes positives (c.-à-d. des OR supérieurs à 1) qui s'interprètent alors dans l'effet global comme l'importance de l'effet d'un microsatellite sur le processus de cancérogénèse. Le modèle donne donc la probabilité qu'un microsatellite quelconque en AI soit associé de façon positive ou négative au résultat. Les données sont écrites de manière à obtenir un effet positif.

Le programme est identique au précédent, seules changent les données introduites. Les données et les résultats sont les suivants :

node	mean	sd	MC error	2.5%	median	97.5%
OR	1.439	0.1172	6.284E-4	1.223	1.435	1.682
beta	0.3609	0.08131	4.374E-4	0.2013	0.361	0.5201
b[1]	0.00208	0.05684	1.864E-4	-0.1135	0.00125	0.1214
...						
b[66]	5.308E-4	0.05669	1.735E-4	-0.1167	2.833E-4	0.1184

L'objectif du modèle est atteint puisqu'il met en évidence un OR à 1,44 [1,223 - 1,682], compatible avec les connaissances actuelles et donc l'intervalle de crédibilité exclut la valeur 1. La probabilité que la valeur de l'OR soit supérieure à 1 et donc que l'effet d'un microsatellite existe, dépasse les 0,99.

Modèle N°2 Une deuxième approche consiste à introduire les homozygotes dans les distributions *a priori* dans les programmes précédents mais c'est un procédé relativement complexe dans le cas d'une régression logistique lorsque l'on utilise Winbugs.

Modèle N°3 Une troisième approche est celle de la méta-analyse pour laquelle il existe deux variantes possibles : soit on n'inclut que les données complètes soit on inclut les données manquantes en donnant une distribution *a priori* pour chaque valeur manquante.

Limite de ces méthodes dans le cadre bayésien Pour les modèles N°2 et 3 et leurs variantes, une difficulté majeure intervient lorsque l'on souhaite réaliser ces modèles avec Winbugs.Lorsque l'on donne pour chaque valeur manquante une distribution *a priori*, le logiciel détermine les paramètres des différentes distributions en minimisant les erreurs sur l'ensemble des distributions, donc y compris celles utilisées comme loi *a priori* pour les manquants ce qui aboutit en général à des résultats biaisés. Cette minimisation n'a pas lieu d'être sur les lois *a priori* des manquants qui ne doivent pas être mise à jour par les données. C'est pourtant ce qui est réalisé par le logiciel lorsque l'écriture du modèle ne spécifie pas

explicitement le statut « manquant » des valeurs manquantes imputées par rapport aux valeurs non-manquantes. Dans le principe, il est techniquement possible d'introduire des sortes de « valves » dans le programme de manière à ce que l'information sur les données manquantes ne soit pas mise à jour par les données non-manquantes mais la mise en place de ce système dans le cas d'une méta-analyse est relativement complexe et n'a pas pu être réalisée. En effet, dans ce cas précis, il s'agit d'introduire les manquants *via* des lois *a priori* qui leur sont propres et qui ne sont donc pas les lois des paramètres d'intérêt (β et OR). La difficulté n'a pas pu être contournée pour l'instant. Soulignons encore pour plus de clarté que ce problème ne se pose que lorsque l'on souhaite donner une distribution *a priori* directement sur chaque valeur manquante et non pas lorsque l'on intègre les manquants dans les lois *a priori* ou que l'on modélise un paramètre dépendant d'une loi *a priori* des manquants comme cela a été réalisé plus haut pour l'estimation d'une proportion et d'un OR et de leur intervalle de confiance.

Les programmes correspondants aux modèles cités (N°2 et N°3) ont été réalisés mais donnent des résultats incohérents et très proches les uns des autres (résultats non montrés). Nous ne pouvons donc pas utiliser cette approche ici.

5.6 Les données

Les données analysées sont issues d'une étude prospective menée au CHU de Strasbourg depuis 1997 portant sur les cancers colo-rectaux. L'ensemble des patients opérés d'un cancer du colon ou du rectum dans les services de chirurgie de l'Hôpital de Hautepierre est inclu dans l'étude. Sur les biopsies de chaque tumeur, un allélotypage est réalisé. Les données retenues ici sont celles des patients porteurs d'un cancer du colon ayant des données validées pour l'allélotypage et dont le stade suivant la classification de Astler-Coller a été vérifié. Il y a 104 patients (une tumeur par patient) pour lesquels 33 microsatellites ont été mesurés. Ces microsatellites sont les suivants :

```
D2S138  D14S65  D9S179  D16S408 D15S127 D1S207  D22S928
D18S61  D18S53  D5S430  D5S346  D1S305  D10S192 TP53
D16S422 D17S790 D8S283  D10S191 D4S394  D3S1283 D9S171
D17S794 D1S225  D11S916 D13S173 D20S107 D4S414
D6S264  D3S1282 D2S159  D6S275  D1S197  D8S264
```

Le choix de ces microsatellites a été fait à partir de l'étude initiale de Delattre [87]. Celui ci a décrit la répartition et la fréquence d'altération de 400 microsatellites dans le cancer du colon. A partir de ses résultats, les 80 microsatellites les plus fréquemment altérés ont été retenu. Parmi ceux-ci, un sous-groupe de 20 microsatellites a été constitué de sorte que parmi ces 20 microsatellites, l'un au moins soit altéré lorsque la biopsie contient au moins 20% de cellule tumorale. Cette base de 20 microsatellites, a ensuite été augmentée jusqu'à 44 en incluant des microsatellites marqueur de gènes d'intérêt soit pour la réponse à un traitement (comme le gène de la topoisomérase) soit pour la compréhension de la cancérogénèse colique. Les 33 MS retenus ici sont ceux pour lesquels le recueil est réalisé depuis le début de l'étude, les 11 autres ayant été rajouté par la suite et n'étant donc pas disponible pour tous les sujets. Cette sélection est celle utilisée notamment dans l'article de Weber [337].

Le détail et la structure des microsatellites peuvent être obtenus sur la banque de données du GÉNÉTHON sur le site internet suivant : `http ://www.genlink.wustl.edu`.

Ces 33 microsatellites sont également ceux utilisés dans la base de données relatives aux tumeurs avec métastases synchrones utilisées antérieurement dans ce travail. Cette base de données concernent 37 patients (différents des 104 pré-cités) pour lesquel l'allélotypage a été réalisé simultanément sur la tumeur primitive et sur une métastase présente lors de l'acte chirurgical, d'où le terme de métastase synchrone.

Stades et Classification des tumeurs Différentes classifications sont utilisées de façon standard pour classifier le stade d'avancement d'une tumeur colique. Ces classifications ont un rôle pronostique important. Les principales classifications sont le stade TNM, celle de Dukes et celle de Astler-Coller.

Classification TNM Rappelons brièvement que la classification TNM est composée du T qui indique le niveau d'évolution de la tumeur, du N qui indique le niveau d'envahissement ganglionnaire et du M qui indique le niveau d'envahissement métastatique. Les différents niveaux d'envahissement sont regroupés en stades de la manière suivante :

- stade I : T1 N0M0, T2N0M0,
- stade Il : T3 N0M0, T4 N0M0,
- stade III : quel que soit le T, N1, N2 ou N3, M0,
- stade IV : quel que soit le T ou le N, M1.

La classification de Dukes Elle a été adaptée pour le cancer du côlon par Kirklin.

- Stade A : atteinte de la muqueuse ou de la sous- muqueuse ou de la musculeuse sans atteinte de la sous-séreuse,
- Stade B : atteinte transpariétale au-delà de la sous-séreuse,
- Stade C : envahissement ganglionnaire.

La classification de Astler-Coller ou Dukes modifié Elle a été adaptée par Turnbull pour ajouter le stade D (invasion de voisinage ou métastases).

- Stade A : atteinte muqueuse ou sous-muqueuse.
- Stade B1 : atteinte de la musculeuse sans atteinte de la sous-séreuse .
- Stade B2 : atteinte de la musculeuse avec atteinte de la sous-séreuse ou de la séreuse ou au-delà .
- Stade C1 : B1 avec envahissement ganglionnaire.
- Stade C2 : B2 avec envahissement ganglionnaire.
- Stade D : métastases.

Correspondances entre Astler-Coller et TNM Il est possible de faire correspondre un ou plusieurs niveau TNM à chaque niveau Astler-Coller. Les regroupements se font de la manière suivante :

- A : T1N0M0
- B1 : T2N0M0
- B2 : T3N0M0, T4N0M0,
- C1 : T1N1 ou N2 ou N3M0, T2N1 ou N2 ou N3M0,
- C2 : T3N1 ou N2 ou N3M0, T4N1 ou N2 ou N3M0,
- D : M1, quel que soit le T ou le N

Dans les applications et développements proposés, nous regrouperons les niveaux A, B1 et B2 d'une part (groupe I) et C et D d'autre part (groupe II) pour obtenir une variable binomiale.

Dans les données relatives aux 104 sujets, la répartition des stades Astler-Coller est la suivante :

Stade	Effectifs	Fréquence (en %)
A	6	5,8
B1	13	12,5
B2	36	34,6
C	27	26,0
D	22	21,2
Total	104	100

Tab. 19 – Fréquence des stades de Astler-Coller dans l'étude

6 Résultats

6.1 Taux d'AI sur les données complètes (hétérozygotes)

L'analyse débute par la description des proportions de microsatellites normaux, d'AI et d'homozygotes. La proportion d'AI est estimée ici sur les données complètes de trois manières différentes : en utilisant l'approximation normale (sur données complètes), en utilisant un intervalle de confiance exact basé sur la loi binomiale (sur données complètes) et en utilisant une estimation bayésienne. Les estimations fréquentistes basées sur la loi normale et sur la méthode exacte (tableau 20) donnent des résultats proches, un peu plus larges dans le second cas, ce qui était attendu. Les résultats du tableau 21 obtenus suivant une méthode bayésienne montrent des intervalles un peu plus étroits. Pour ces derniers, on note que le mode de la densité *a posteriori* correspond bien à l'estimation du maximum de vraisemblance du taux d'AI. Le mode est l'indice le plus pertinent pour caractériser ce taux dans le concept bayésien, la moyenne étant ici un mauvais indicateur du fait de l'asymétrie des distributions Beta lorsque les deux paramètres de la loi sont tels que $\alpha \neq \beta$.

Microsatellite	Effectifs				Intervalle de Confiance	
	n_N	n_{AI}	m	taux d'AI	asymptotique	exact
D2S138	43	32	29	0,427	[0,315 - 0,539]	[0,313 - 0,546]
D18S61	18	66	20	0,786	[0,698 - 0,874]	[0,683 - 0,868]
D16S422	42	28	34	0,400	[0,285 - 0,515]	[0,285 - 0,524]
D17S794	25	33	46	0,569	[0,442 - 0,696]	[0,432 - 0,698]
D6S264	37	27	40	0,422	[0,301 - 0,543]	[0,299 - 0,552]
D14S65	43	30	31	0,411	[0,298 - 0,524]	[0,297 - 0,532]
D18S53	27	47	30	0,635	[0,525 - 0,745]	[0,515 - 0,744]
D17S790	30	40	34	0,571	[0,455 - 0,687]	[0,447 - 0,689]
D1S225	51	36	17	0,414	[0,310 - 0,517]	[0,309 - 0,524]
D3S1282	32	31	41	0,492	[0,369 - 0,616]	[0,364 - 0,621]
D9S179	40	24	40	0,375	[0,256 - 0,494]	[0,257 - 0,505]
D5S430	28	22	54	0,440	[0,302 - 0,578]	[0,300 - 0,587]
D8S283	35	49	20	0,583	[0,478 - 0,689]	[0,471 - 0,690]
D11S916	46	32	26	0,410	[0,301 - 0,519]	[0,300 - 0,527]
D2S159	50	32	22	0,390	[0,285 - 0,496]	[0,284 - 0,504]
D16S408	44	25	35	0,362	[0,249 - 0,476]	[0,250 - 0,487]
D5S346	38	51	15	0,573	[0,470 - 0,676]	[0,464 - 0,677]
D10S191	55	29	20	0,345	[0,244 - 0,447]	[0,245 - 0,457]
D13S173	31	46	27	0,597	[0,488 - 0,707]	[0,479 - 0,708]
D6S275	49	23	32	0,319	[0,212 - 0,427]	[0,214 - 0,440]
D15S127	40	40	24	0,500	[0,390 - 0,610]	[0,386 - 0,614]
D1S305	43	36	25	0,455	[0,346 - 0,566]	[0,343 - 0,572]
D4S394	47	23	34	0,329	[0,219 - 0,439]	[0,221 - 0,451]
D20S107	26	44	34	0,629	[0,515 - 0,742]	[0,505 - 0,741]
D1S197	50	39	15	0,438	[0,335 - 0,541]	[0,333 - 0,547]
D1S207	59	32	13	0,352	[0,253 - 0,450]	[0,254 - 0,459]
D10S192	44	38	22	0,463	[0,355 - 0,571]	[0,353 - 0,577]
D3S1283	41	38	25	0,481	[0,371 - 0,591]	[0,367 - 0,596]
D4S414	49	27	28	0,356	[0,248 - 0,463]	[0,249 - 0,473]
D8S264	25	59	20	0,702	[0,605 - 0,800]	[0,593 - 0,797]
D22S928	36	43	25	0,544	[0,435 - 0,654]	[0,428 - 0,657]
TP53	22	63	19	0,742	[0,648 - 0,834]	[0,635 - 0,830]
D9S171	35	33	36	0,485	[0,367 - 0,604]	[0,362 - 0,610]

Tab. 20 – Taux d'AI pour chacun des 33 microsatellites estimés de manière fréquentiste, par une approximation gaussienne et par la méthode exacte basée sur la loi binomiale sur les données complètes uniquement.

Microsatellite	Effectifs					
	n_N	n_{AI}	m	moyenne	mode	IC
D2S138	43	32	29	0,429	0,427	[0,321 - 0,540]
D18S61	18	66	20	0,779	0,786	[0,686 - 0,860]
D16S422	42	28	34	0,403	0,400	[0,293 - 0,517]
D17S794	25	33	46	0,567	0,569	[0,441 - 0,688]
D6S264	37	27	40	0,424	0,422	[0,308 - 0,544]
D14S65	43	30	31	0,413	0,411	[0,305 - 0,526]
D18S53	27	47	30	0,632	0,635	[0,521 - 0,736]
D17S790	30	40	34	0,569	0,571	[0,454 - 0,681]
D1S225	51	36	17	0,416	0,414	[0,316 - 0,519]
D3S1282	32	31	41	0,492	0,492	[0,372 - 0,613]
D9S179	40	24	40	0,379	0,375	[0,267 - 0,498]
D5S430	28	22	54	0,442	0,440	[0,311 - 0,578]
D8S283	35	49	20	0,581	0,583	[0,476 - 0,683]
D11S916	46	32	26	0,413	0,410	[0,308 - 0,521]
D2S159	50	32	22	0,393	0,390	[0,292 - 0,499]
D16S408	44	25	35	0,366	0,362	[0,259 - 0,481]
D5S346	38	51	15	0,571	0,573	[0,469 - 0,671]
D10S191	55	29	20	0,349	0,345	[0,252 - 0,452]
D13S173	31	46	27	0,595	0,597	[0,485 - 0,700]
D6S275	49	23	32	0,324	0,319	[0,223 - 0,434]
D15S127	40	40	24	0,500	0,500	[0,393 - 0,607]
D1S305	43	36	25	0,457	0,456	[0,350 - 0,565]
D4S394	47	23	34	0,333	0,329	[0,230 - 0,445]
D20S107	26	44	34	0,625	0,629	[0,511 - 0,732]
D1S197	50	39	15	0,440	0,438	[0,340 - 0,542]
D1S207	59	32	13	0,355	0,352	[0,261 - 0,454]
D10S192	44	38	22	0,464	0,463	[0,359 - 0,571]
D3S1283	41	38	25	0,481	0,481	[0,374 - 0,590]
D4S414	49	27	28	0,359	0,355	[0,257 - 0,468]
D8S264	25	59	20	0,698	0,702	[0,597 - 0,790]
D22S928	36	43	25	0,543	0,544	[0,435 - 0,650]
TP53	22	63	19	0,736	0,741	[0,639 - 0,822]
D9S171	35	33	36	0,486	0,485	[0,370 - 0,602]

Tab. 21 – Estimation des taux d'AI pour chacun des 33 microsatellites sur les données complètes, en utilisant une méthode bayésienne avec *a priori* non informatif.

6.2 Description des données manquantes

Avant toute analyse portant sur des données incomplètes, les données manquantes doivent bénéficier d'une analyse propre visant à caractériser leur répartition par rapport à l'ensemble des données. Rappelons qu'il est difficile de tester le mécanisme des manquants ([290, 292]) mais on peut néanmoins essayer de formuler des hypothèses sur ce mécanisme. La répartition univariée des homozygotes pour chaque microsatellite est ici déterminée en utilisant les méthodes bayésiennes. Les taux d'homozygotes de chaque microsatellite, correspondant donc à la proportion d'information manquante pour ce microsatellite, sont donnés dans le tableau 22. Le taux le plus bas est de 0,144 (D5S346 et D1S197) et le taux le plus haut est de 0,519, pour le microsatellite D5S430. Le taux d'homozygote vaut en moyenne 0,272, avec une médiane à 0,260, un premier quartile à 0,192 et un troisième quartile à 0,327.

Microsatellite	Effectifs (total=104)			taux d'HMZ	Intervalle de confiance du taux
	n_N	n_{AI}	n_{HMZ}		
D2S138	43	32	29	0,279	[0,202 - 0,372]
D18S61	18	66	20	0,192	[0,128 - 0,279]
D16S422	42	28	34	0,327	[0,244 - 0,422]
D17S794	25	33	46	0,442	[0,350 - 0,538]
D6S264	37	27	40	0,385	[0,297 - 0,481]
D14S65	43	30	31	0,298	[0,219 - 0,392]
D18S53	27	47	30	0,288	[0,210 - 0,382]
D17S790	30	40	34	0,327	[0,244 - 0,422]
D1S225	51	36	17	0,163	[0,105 - 0,247]
D3S1282	32	31	41	0,394	[0,306 - 0,491]
D9S179	40	24	40	0,385	[0,297 - 0,481]
D5S430	28	22	54	0,519	[0,424 - 0,613]
D8S283	35	49	20	0,192	[0,128 - 0,279]
D11S916	46	32	26	0,250	[0,177 - 0,341]
D2S159	50	32	22	0,212	[0,144 - 0,300]
D16S408	44	25	35	0,337	[0,253 - 0,432]
D5S346	38	51	15	0,144	[0,090 - 0,225]
D10S191	55	29	20	0,192	[0,128 - 0,279]
D13S173	31	46	27	0,260	[0,185 - 0,352]
D6S275	49	23	32	0,308	[0,227 - 0,402]
D15S127	40	40	24	0,231	[0,160 - 0,321]
D1S305	43	36	25	0,240	[0,169 - 0,331]
D4S394	47	23	34	0,327	[0,244 - 0,422]
D20S107	26	44	34	0,327	[0,244 - 0,422]
D1S197	50	39	15	0,144	[0,090 - 0,225]
D1S207	59	32	13	0,125	[0,075 - 0,202]
D10S192	44	38	22	0,212	[0,144 - 0,300]
D3S1283	41	38	25	0,240	[0,169 - 0,331]
D4S414	49	27	28	0,269	[0,193 - 0,362]
D8S264	25	59	20	0,192	[0,128 - 0,279]
D22S928	36	43	25	0,240	[0,169 - 0,331]
TP53	22	63	19	0,183	[0,120 - 0,268]
D9S171	35	33	36	0,346	[0,262 - 0,442]

TAB. 22 – Taux d'homozygote pour chaque microsatellites : estimation brute par la méthode bayésienne avec loi *a priori* non informative.

6.3 Description multivariée des manquants

Une analyse descriptive multivariée des valeurs manquantes permet de rechercher une répartition particulière des valeurs manquantes dans le jeu de données. Nous utilisons ici une

analyse factorielle des correspondances multiple. Le résultat de cette analyse est présenté dans la figure 4. La répartition des manquants ne semble pas suivre un schéma particulier, suggérant l'indépendance du mécanisme des manquants par rapport aux données.

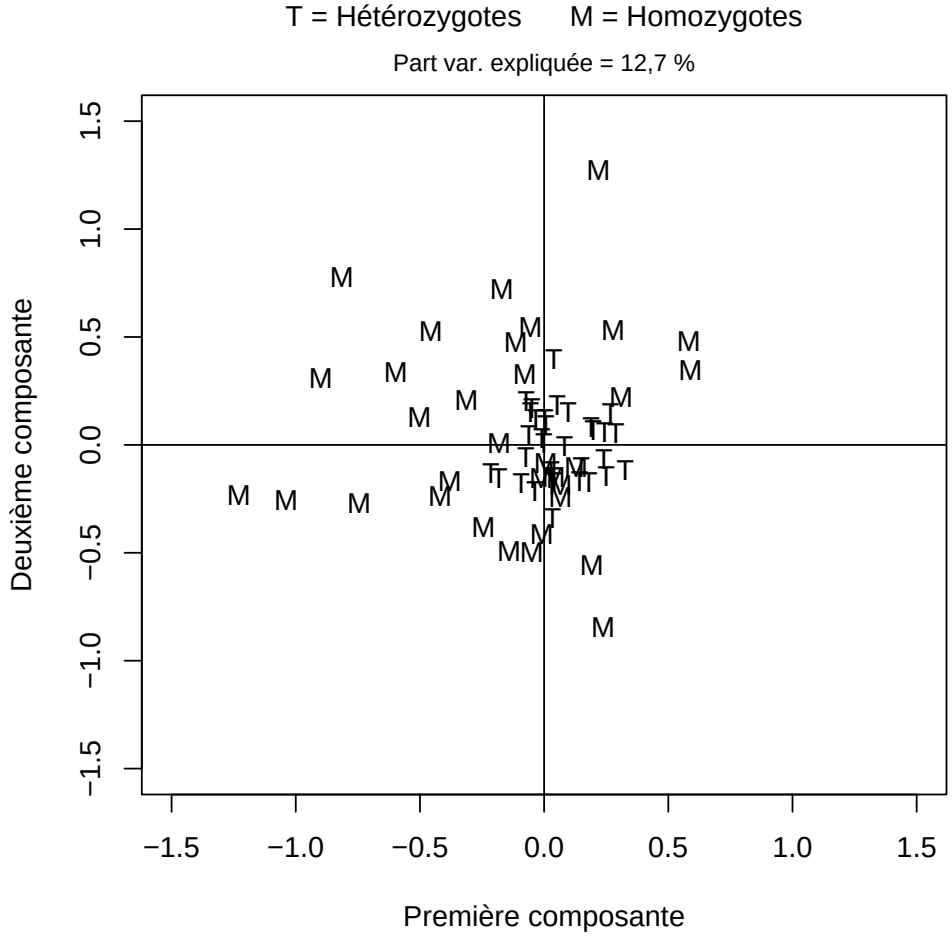

FIG. 4 – AFCM des données, représentées en fonction du caractère hétérozygote ou homozygotes (manquant) des microsatellites.

Une figure modifiée (5) fait apparaître la liste des microsatellites pour les homozygotes ce qui permet de mieux les visualiser. En raison de l'absence manifeste de structure visible dans les données, les hétérozygotes ne sont pas étiquetés afin de ne pas alourdir le graphique. La part de variance du premier plan est faible, de l'ordre de 12,7% ce qui concourt à montrer l'absence de structure dans la répartition multivariée des homozygotes par rapport aux hétérozygotes.

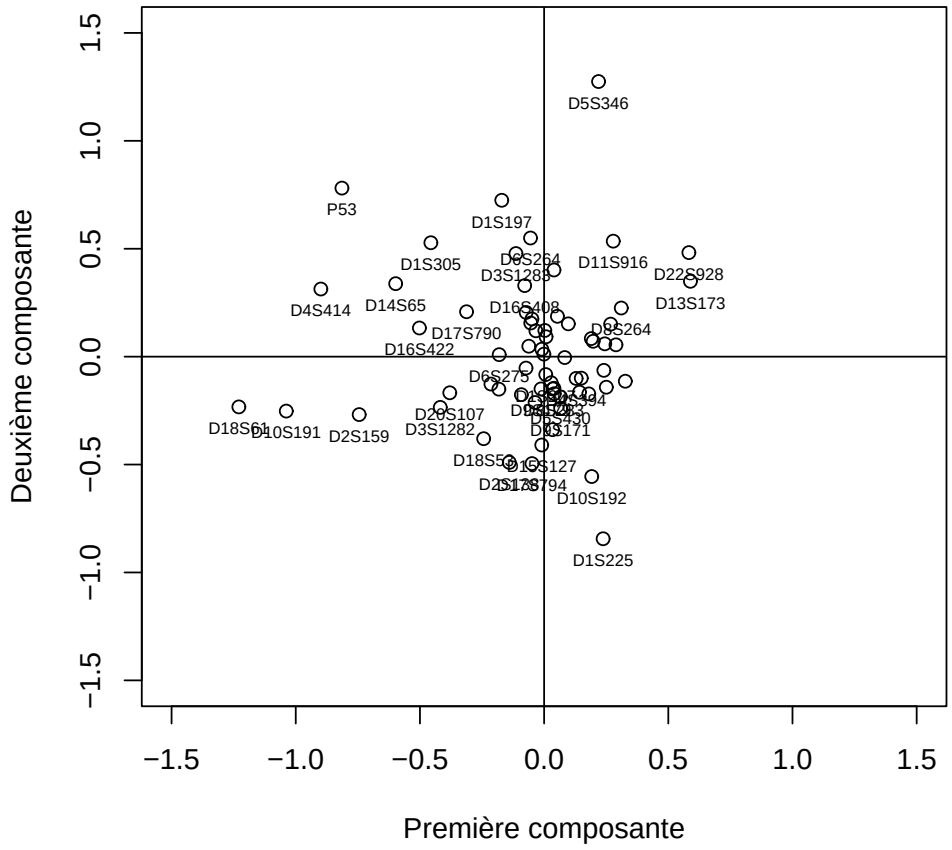

FIG. 5 – AFCM des données, représentées en fonction du caractère hétérozygote ou homozygotes (manquant) des microsatellites avec les microsatellites indiqués pour les homozygotes.

Cette absence de structure peut être confirmée par une analyse de type bootstrap d'une analyse en cluster. Le principe consiste à réaliser une analyse en cluster des données. On retient pour chaque microsatellite uniquement le statut hétérozygote et homozygote, objet de cette exploration. Après réalisation du cluster (avec comme mesure de distance un coefficient de corrélation et comme mesure d'agrégation la méthode de Ward), on réalise une série de tirages bootstrap [85] d'après la méthode de Shimodaira [302, 303]. Les analyses sont faites à l'aide du module `pvclust` dans `R`. La figure 6 confirme l'absence de structure dans les données. L'analyse en cluster avec validation bootstrap est présentée avec des indices numériques rouges et verts indiquant une probabilité de retrouver la même branche du dendrogramme lors d'une réplication bootstrap de l'analyse. Suivant les indications de

Shimodaira, seuls les indices rouges doivent être pris en considération. Une valeur supérieure ou égale à 95 confirme la stabilité de la branche, ce qui correspond grossièrement à une p-valeur de 0,05. Une seule de ces valeurs est égale à 95, aucune n'est supérieure à cette valeur seuil. Ce résultat est donc en faveur de l'absence de structure dans la répartition des homozygotes.

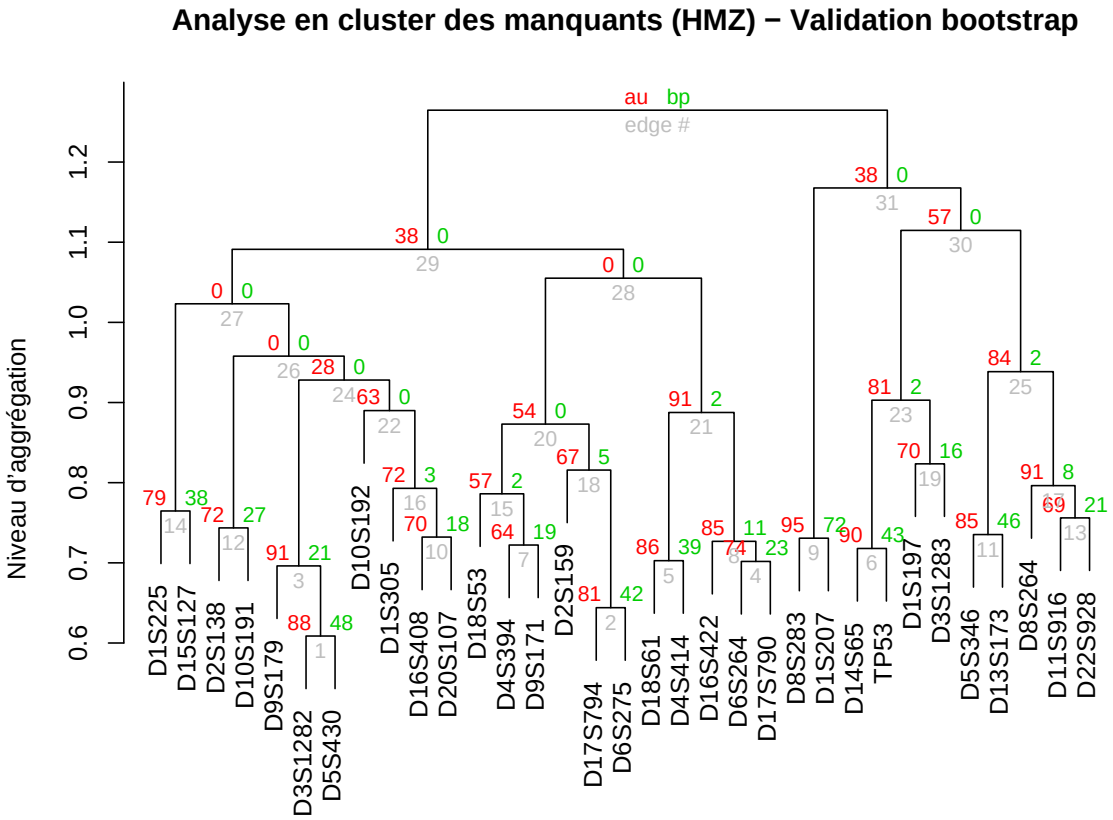

FIG. 6 – AFCM des données, représentées en fonction du caractère hétérozygote ou homozygotes (manquant) des microsatellites avec le numéro du microsatellite indiqué pour les homozygotes uniquement.

6.3.1 Détermination du type de manquants

La description et l'analyse des données manquantes supposent une tentative de détermination du type de manquants. Ce paragraphe contient les résultats des modèles proposés

dans le paragraphe 4.7.

Les résultats montrent que pour l'ensemble des microsatellites, la probabilité d'être hétérozygote parmi les normaux et d'être hétérozygote parmi les AI est proche voire équivalente. Seul le microsatellite D18S61 présente un écart notable entre ces deux probabilités, la première valant 0,680 et la seconde 0,885, soit une différence de près de 0,2. Pour d'autres microsatellites, les deux valeurs sont très proches, (D3S1283) ou quasi-identiques (D15S127). Le mécanisme ne serait donc pas de type MNAR. On pourrait globalement admettre que les données sont homozygotes suivant un modèle MCAR ce qui permet d'estimer le taux d'AI à partir des hétérozygotes seuls, ce qui suppose tout de même une perte d'effectif. Il vaut mieux donc utiliser l'estimation du taux d'AI obtenue sur l'ensemble des homozygotes et hétérozygotes soit par exemple par une des deux méthodes du paragraphe précédent soit à l'aide du programme du paragraphe 4.7. Une certaine prudence invite à admettre par ailleurs l'hypothèse MAR plutôt que l'hypothèse MCAR.

De façon complémentaire, on observe que la probabilité d'être hétérozygote parmi les normaux vaut 0,737 en moyenne et varie peu d'un microsatellite à l'autre. Notons que ce résultat est très cohérent avec le résultat obtenu au paragraphe 6.2 concernant le taux d'homozygote où le taux moyen d'homozygote était de 0,272.

Microsatellite		Moyenne	DS	2.5%	Médiane	97.5%
D2S138	$\pi_0[1]$	0,749	0,124	0,538	0,736	0,981
D18S61	$\pi_0[2]$	0,680	0,166	0,394	0,667	0,978
D16S422	$\pi_0[3]$	0,721	0,138	0,496	0,703	0,980
D17S794	$\pi_0[4]$	0,572	0,188	0,302	0,532	0,961
D6S264	$\pi_0[5]$	0,669	0,156	0,428	0,644	0,975
D14S65	$\pi_0[6]$	0,739	0,129	0,523	0,724	0,981
D18S53	$\pi_0[7]$	0,675	0,160	0,415	0,656	0,976
D17S790	$\pi_0[8]$	0,669	0,161	0,412	0,646	0,976
D1S225	$\pi_0[9]$	0,848	0,085	0,685	0,848	0,990
D3S1282	$\pi_0[10]$	0,639	0,166	0,386	0,610	0,971
D9S179	$\pi_0[11]$	0,677	0,149	0,445	0,652	0,974
D5S430	$\pi_0[12]$	0,556	0,189	0,294	0,510	0,959
D8S283	$\pi_0[13]$	0,781	0,117	0,566	0,776	0,986
D11S916	$\pi_0[14]$	0,780	0,115	0,578	0,771	0,986
D2S159	$\pi_0[15]$	0,814	0,099	0,633	0,809	0,988
D16S408	$\pi_0[16]$	0,720	0,134	0,501	0,702	0,979
D5S346	$\pi_0[17]$	0,830	0,097	0,641	0,831	0,989
D10S191	$\pi_0[18]$	0,837	0,088	0,672	0,833	0,989
D13S173	$\pi_0[19]$	0,714	0,144	0,470	0,699	0,979
D6S275	$\pi_0[20]$	0,750	0,122	0,547	0,736	0,981
D15S127	$\pi_0[21]$	0,774	0,120	0,561	0,766	0,985
D1S305	$\pi_0[22]$	0,777	0,117	0,570	0,768	0,986
D4S394	$\pi_0[23]$	0,738	0,129	0,525	0,723	0,980
D20S107	$\pi_0[24]$	0,646	0,170	0,377	0,621	0,974
D1S197	$\pi_0[25]$	0,861	0,080	0,703	0,862	0,992
D1S207	$\pi_0[26]$	0,890	0,065	0,759	0,892	0,994
D10S192	$\pi_0[27]$	0,798	0,108	0,601	0,793	0,986
D3S1283	$\pi_0[28]$	0,769	0,119	0,559	0,761	0,984
D4S414	$\pi_0[29]$	0,772	0,113	0,578	0,761	0,982
D8S264	$\pi_0[30]$	0,734	0,142	0,479	0,728	0,983
D22S928	$\pi_0[31]$	0,751	0,128	0,526	0,742	0,983
TP53	$\pi_0[32]$	0,720	0,148	0,456	0,711	0,982
D9S171	$\pi_0[33]$	0,682	0,151	0,441	0,660	0,975

Tab. 23 – Estimation des taux d'HTZ parmi les normaux pour chaque microsatellite obtenus par le modèle du paragraphe 4.7.

Avec le même modèle, la probabilité d'être hétérozygote parmi les AI vaut en moyenne 0,730, soit quasiment la même valeur que pour les normaux. Notons là aussi la cohérence avec le résultat obtenu par la méthode plus simple sur les taux d'hmozygotes.

Microsatellite		Moyenne	DS	2.5%	Médiane	97.5%
D2S138	$\pi_1[1]$	0,712	0,145	0,467	0,698	0,979
D18S61	$\pi_1[2]$	0,855	0,077	0,711	0,853	0,990
D16S422	$\pi_1[3]$	0,655	0,165	0,393	0,631	0,974
D17S794	$\pi_1[4]$	0,621	0,173	0,368	0,584	0,972
D6S264	$\pi_1[5]$	0,615	0,176	0,350	0,582	0,970
D14S65	$\pi_1[6]$	0,686	0,154	0,432	0,669	0,977
D18S53	$\pi_1[7]$	0,757	0,120	0,553	0,745	0,982
D17S790	$\pi_1[8]$	0,711	0,140	0,484	0,693	0,979
D1S225	$\pi_1[9]$	0,810	0,106	0,607	0,809	0,989
D3S1282	$\pi_1[10]$	0,634	0,167	0,378	0,605	0,971
D9S179	$\pi_1[11]$	0,599	0,184	0,323	0,565	0,969
D5S430	$\pi_1[12]$	0,513	0,201	0,245	0,458	0,954
D8S283	$\pi_1[13]$	0,826	0,094	0,648	0,824	0,989
D11S916	$\pi_1[14]$	0,728	0,140	0,487	0,716	0,981
D2S159	$\pi_1[15]$	0,756	0,130	0,522	0,750	0,984
D16S408	$\pi_1[16]$	0,632	0,173	0,361	0,606	0,972
D5S346	$\pi_1[17]$	0,863	0,078	0,710	0,864	0,992
D10S191	$\pi_1[18]$	0,757	0,131	0,518	0,752	0,985
D13S173	$\pi_1[19]$	0,775	0,116	0,573	0,766	0,985
D6S275	$\pi_1[20]$	0,639	0,174	0,360	0,614	0,973
D15S127	$\pi_1[21]$	0,773	0,120	0,561	0,766	0,985
D1S305	$\pi_1[22]$	0,752	0,129	0,525	0,742	0,984
D4S394	$\pi_1[23]$	0,619	0,176	0,344	0,591	0,968
D20S107	$\pi_1[24]$	0,725	0,132	0,508	0,707	0,978
D1S197	$\pi_1[25]$	0,834	0,095	0,649	0,835	0,990
D1S207	$\pi_1[26]$	0,825	0,100	0,627	0,828	0,990
D10S192	$\pi_1[27]$	0,781	0,117	0,569	0,775	0,986
D3S1283	$\pi_1[28]$	0,761	0,124	0,541	0,752	0,984
D4S414	$\pi_1[29]$	0,691	0,156	0,430	0,676	0,978
D8S264	$\pi_1[30]$	0,844	0,083	0,689	0,841	0,990
D22S928	$\pi_1[31]$	0,777	0,116	0,572	0,769	0,984
TP53	$\pi_1[32]$	0,858	0,077	0,711	0,856	0,991
D9S171	$\pi_1[33]$	0,687	0,153	0,438	0,667	0,978

TAB. 24 – Estimation des taux d'HTZ parmi les AI pour chaque microsatellite obtenus par le modèle du paragraphe 4.7.

Les probabilités s d'être AI sachant qu'on est homozygote sont très proches d'un microsatellite à l'autre. On peut donc admettre que la probabilité d'être AI quand on est homozygote ne dépend pas du microsatellite ce qui suggère au moins un mécanisme de type MCAR ou MAR mais pas MNAR puisque cela ne dépend pas du fait d'être manquant. On

note que s vaut en moyenne 0,496, soit 0,5, ce qui va également dans le sens d'un mécanisme qui ne soit pas de type MNAR.

Microsatellite		Moyenne	DS	2.5%	Médiane	97.5%
D2S138	$s[1]$	0,481	0,289	0,022	0,473	0,972
D18S61	$s[2]$	0,539	0,291	0,031	0,558	0,981
D16S422	$s[3]$	0,487	0,295	0,021	0,479	0,976
D17S794	$s[4]$	0,508	0,300	0,019	0,515	0,979
D6S264	$s[5]$	0,491	0,296	0,020	0,488	0,976
D14S65	$s[6]$	0,484	0,292	0,021	0,475	0,974
D18S53	$s[7]$	0,521	0,291	0,026	0,531	0,978
D17S790	$s[8]$	0,512	0,294	0,023	0,519	0,979
D1S225	$s[9]$	0,486	0,289	0,022	0,480	0,973
D3S1282	$s[10]$	0,498	0,294	0,022	0,497	0,977
D9S179	$s[11]$	0,476	0,295	0,018	0,465	0,973
D5S430	$s[12]$	0,494	0,302	0,019	0,491	0,978
D8S283	$s[13]$	0,506	0,288	0,026	0,509	0,976
D11S916	$s[14]$	0,487	0,293	0,022	0,482	0,975
D2S159	$s[15]$	0,482	0,291	0,022	0,471	0,973
D16S408	$s[16]$	0,477	0,292	0,019	0,466	0,973
D5S346	$s[17]$	0,505	0,289	0,026	0,508	0,976
D10S191	$s[18]$	0,475	0,291	0,021	0,462	0,971
D13S173	$s[19]$	0,508	0,291	0,023	0,513	0,977
D6S275	$s[20]$	0,466	0,292	0,019	0,451	0,971
D15S127	$s[21]$	0,499	0,293	0,023	0,499	0,976
D1S305	$s[22]$	0,494	0,293	0,022	0,491	0,977
D4S394	$s[23]$	0,480	0,293	0,021	0,470	0,973
D20S107	$s[24]$	0,522	0,292	0,027	0,534	0,979
D1S197	$s[25]$	0,492	0,29	0,023	0,489	0,974
D1S207	$s[26]$	0,489	0,288	0,023	0,483	0,974
D10S192	$s[27]$	0,492	0,291	0,022	0,489	0,974
D3S1283	$s[28]$	0,494	0,290	0,023	0,493	0,974
D4S414	$s[29]$	0,470	0,289	0,020	0,457	0,970
D8S264	$s[30]$	0,529	0,289	0,027	0,545	0,979
D22S928	$s[31]$	0,505	0,289	0,026	0,506	0,976
TP53	$s[32]$	0,527	0,290	0,028	0,544	0,980
D9S171	$s[33]$	0,488	0,294	0,020	0,482	0,975

TAB. 25 – Estimation de la probabilité s d'être AI lorsque l'on est homozygotes, $\Pr(AI|HMZ)$.

La différence δ_π entre les taux d'HTZ parmi les AI et les taux d'HTZ parmi les normaux (sur l'ensemble des valeurs hétérozygotes et homozygotes) est proche de 0 quel que soit

le microsatellite considéré, comme on pouvait s'y attendre à la vue des probabilités d'être hétérozygote pour les AI d'une part et les normaux d'autre part. On confirme donc que le mécanisme des manquants ne dépend pas du microsatellite et ne dépend pas de la vraie valeur du microsatellite. Donc, ce n'est probablement pas un mécanisme de type MNAR.

Microsatellite		Moyenne	DS	2.5%	Médiane	97.5%
D2S138	$\delta_\pi[1]$	-0,038	0,256	-0,480	-0,040	0,401
D18S61	$\delta_\pi[2]$	0,175	0,228	-0,227	0,185	0,564
D16S422	$\delta_\pi[3]$	-0,064	0,286	-0,547	-0,070	0,435
D17S794	$\delta_\pi[4]$	0,051	0,340	-0,548	0,057	0,624
D6S264	$\delta_\pi[5]$	-0,049	0,314	-0,581	-0,056	0,498
D14S65	$\delta_\pi[6]$	-0,055	0,269	-0,512	-0,060	0,412
D18S53	$\delta_\pi[7]$	0,080	0,266	-0,383	0,086	0,531
D17S790	$\delta_\pi[8]$	0,043	0,284	-0,450	0,046	0,528
D1S225	$\delta_\pi[9]$	-0,040	0,178	-0,355	-0,040	0,270
D3S1282	$\delta_\pi[10]$	-0,0057	0,317	-0,552	-0,005	0,541
D9S179	$\delta_\pi[11]$	-0,082	0,315	-0,610	-0,091	0,476
D5S430	$\delta_\pi[12]$	-0,039	0,366	-0,668	-0,045	0,612
D8S283	$\delta_\pi[13]$	0,043	0,198	-0,304	0,045	0,390
D11S916	$\delta_\pi[14]$	-0,053	0,240	-0,463	-0,056	0,363
D2S159	$\delta_\pi[15]$	-0,058	0,215	-0,432	-0,060	0,315
D16S408	$\delta_\pi[16]$	-0,085	0,292	-0,578	-0,094	0,428
D5S346	$\delta_\pi[17]$	0,032	0,162	-0,251	0,031	0,322
D10S191	$\delta_\pi[18]$	-0,078	0,205	-0,439	-0,080	0,275
D13S173	$\delta_\pi[19]$	0,061	0,246	-0,369	0,066	0,479
D6S275	$\delta_\pi[20]$	-0,114	0,279	-0,583	-0,124	0,380
D15S127	$\delta_\pi[21]$	-5,6E-4	0,225	-0,388	-9,322	0,389
D1S305	$\delta_\pi[22]$	-0,026	0,231	-0,425	-0,027	0,376
D4S394	$\delta_\pi[23]$	-0,112	0,288	-0,596	-0,122	0,399
D20S107	$\delta_\pi[24]$	0,078	0,286	-0,422	0,084	0,564
D1S197	$\delta_\pi[25]$	-0,027	0,161	-0,314	-0,027	0,256
D1S207	$\delta_\pi[26]$	-0,063	0,151	-0,340	-0,063	0,200
D10S192	$\delta_\pi[27]$	-0,020	0,211	-0,386	-0,022	0,348
D3S1283	$\delta_\pi[28]$	-0,012	0,231	-0,410	-0,015	0,388
D4S414	$\delta_\pi[29]$	-0,086	0,255	-0,520	-0,093	0,358
D8S264	$\delta_\pi[30]$	0,109	0,211	-0,258	0,112	0,479
D22S928	$\delta_\pi[31]$	0,027	0,230	-0,374	0,030	0,424
TP53	$\delta_\pi[32]$	0,139	0,211	-0,232	0,145	0,506
D9S171	$\delta_\pi[33]$	1,4E-5	0,288	-0,497	6,427	0,493

Tab. 26 – Différence de probabilité d'être hétérozygote entre les sujets normaux et les sujets AI.

6.4 Taux d'AI sur l'ensemble des données : hétérozygotes et homozygotes

Les résultats ci-dessous donnent également les estimations du taux d'AI mais cette fois-ci en prenant en compte la totalité des données, c'est-à-dire en modélisant les manquants. Sur l'ensemble des méthodes décrites aux chapitres précédents, seules les deux les plus pertinentes sont utilisées. Elles constituent dans l'ensemble des méthodes raisonnables dont les estimations sont de bonnes limites des taux d'AI, c'est-à-dire qu'elles correspondent respectivement aux estimations les plus optimistes et aux estimations les plus prudentes que l'on peut faire pour ce taux. Dans le premier cas (tableau N°27), les données manquantes sont introduites dans les lois *a priori* du paramètre p du taux d'AI à estimer. Le tableau suivant (tableau N°28) donne les estimations basées sur une estimation dans laquelle la valeur de p_m, taux d'AI parmi les manquants, est combinée avec le taux d'AI parmi les non-manquants pour obtenir une estimation sur l'ensemble des données. La valeur de p_m est issue d'une loi $Be(1;1)$. Il faut noter que les valeurs choisies pour les paramètres α et β de la loi Beta dans le premier cas sont les paramètres obtenus par une estimation du maximum de vraisemblance, qui ne donne pas forcément des valeurs entières. Ceci ne pose pas de problème puisque les paramètres de la loi Be peuvent prendre des valeurs continues, tant qu'elles sont positives : $\alpha, \beta \in \mathbb{R}^+$.

Les résultats du deuxième type de modélisation ont été obtenus avec le programme N°22. On observe que conformément à la théorie, les intervalles de confiance sont nettement plus larges et donc plus prudents avec la seconde méthode. Ces résultats montrent parfaitement l'ampleur que peut prendre l'incertitude liée aux manquants lors de l'estimation d'une proportion.

Microsatellites	Effectifs			Taux ponctuel		Intervalle de confiance
	n	n_{AI}	m	moyenne	mode	IC
D2S138	43	32	29	0,428	0,427	[0,336 - 0,523]
D18S61	18	66	20	0,780	0,786	[0,697 - 0,853]
D16S422	42	28	34	0,402	0,400	[0,311 - 0,496]
D17S794	25	33	46	0,568	0,569	[0,473 - 0,660]
D6S264	37	27	40	0,423	0,422	[0,331 - 0,518]
D14S65	43	30	31	0,413	0,411	[0,321 - 0,507]
D18S53	27	47	30	0,633	0,635	[0,539 - 0,721]
D17S790	30	40	34	0,570	0,571	[0,475 - 0,662]
D1S225	51	36	17	0,415	0,414	[0,324 - 0,510]
D3S1282	32	31	41	0,492	0,492	[0,398 - 0,587]
D9S179	40	24	40	0,377	0,375	[0,288 - 0,471]
D5S430	28	22	54	0,441	0,440	[0,348 - 0,536]
D8S283	35	49	20	0,582	0,583	[0,487 - 0,674]
D11S916	46	32	26	0,412	0,410	[0,320 - 0,507]
D2S159	50	32	22	0,392	0,390	[0,302 - 0,487]
D16S408	44	25	35	0,365	0,362	[0,276 - 0,458]
D5S346	38	51	15	0,572	0,573	[0,477 - 0,664]
D10S191	55	29	20	0,348	0,345	[0,261 - 0,441]
D13S173	31	46	27	0,596	0,597	[0,501 - 0,687]
D6S275	49	23	32	0,323	0,319	[0,238 - 0,414]
D15S127	40	40	24	0,500	0,500	[0,405 - 0,595]
D1S305	43	36	25	0,457	0,456	[0,363 - 0,551]
D4S394	47	23	34	0,332	0,329	[0,246 - 0,424]
D20S107	26	44	34	0,626	0,629	[0,532 - 0,715]
D1S197	50	39	15	0,439	0,438	[0,347 - 0,534]
D1S207	59	32	13	0,354	0,352	[0,267 - 0,447]
D10S192	44	38	22	0,464	0,463	[0,370 - 0,559]
D3S1283	41	38	25	0,481	0,481	[0,387 - 0,576]
D4S414	49	27	28	0,358	0,355	[0,270 - 0,451]
D8S264	25	59	20	0,699	0,702	[0,608 - 0,782]
D22S928	36	43	25	0,543	0,544	[0,449 - 0,637]
TP53	22	63	19	0,737	0,741	[0,649 - 0,816]
D9S171	35	33	36	0,486	0,485	[0,391 - 0,580]

Tab. 27 – Estimation du taux d'AI pour chaque microsatellite en introduisant les manquants dans les lois *a priori*. Les paramètres de la loi Beta sont basés sur les estimations du maximum de vraisemblance pour la proportion d'AI parmi les manquants.

Microsatellites	Effectifs			Taux ponctuel		Intervalle de confiance
	n	n_{AI}	m	moyenne	mode	IC
D2S138	43	32	29	0,448	0,448	[0,286 - 0,612]
D18S61	18	66	20	0,725	0,725	[0,601 - 0,846]
D16S422	42	28	34	0,434	0,434	[0,254 - 0,616]
D17S794	25	33	46	0,537	0,537	[0,310 - 0,765]
D6S264	37	27	40	0,453	0,453	[0,250 - 0,658]
D14S65	43	30	31	0,439	0,439	[0,269 - 0,610]
D18S53	27	47	30	0,593	0,593	[0,428 - 0,759]
D17S790	30	40	34	0,546	0,546	[0,365 - 0,728]
D1S225	51	36	17	0,429	0,429	[0,307 - 0,552]
D3S1282	32	31	41	0,495	0,495	[0,287 - 0,704]
D9S179	40	24	40	0,425	0,425	[0,223 - 0,630]
D5S430	28	22	54	0,472	0,472	[0,212 - 0,733]
D8S283	35	49	20	0,565	0,565	[0,433 - 0,697]
D11S916	46	32	26	0,434	0,434	[0,283 - 0,587]
D2S159	50	32	22	0,415	0,415	[0,278 - 0,554]
D16S408	44	25	35	0,411	0,411	[0,228 - 0,596]
D5S346	38	51	15	0,561	0,561	[0,443 - 0,677]
D10S191	55	29	20	0,377	0,377	[0,249 - 0,509]
D13S173	31	46	27	0,570	0,570	[0,414 - 0,725]
D6S275	49	23	32	0,378	0,378	[0,207 - 0,551]
D15S127	40	40	24	0,500	0,500	[0,354 - 0,646]
D1S305	43	36	25	0,467	0,467	[0,318 - 0,616]
D4S394	47	23	34	0,388	0,388	[0,209 - 0,568]
D20S107	26	44	34	0,584	0,584	[0,403 - 0,764]
D1S197	50	39	15	0,448	0,448	[0,331 - 0,565]
D1S207	59	32	13	0,373	0,372	[0,265 - 0,482]
D10S192	44	38	22	0,471	0,471	[0,332 - 0,611]
D3S1283	41	38	25	0,486	0,486	[0,336 - 0,635]
D4S414	49	27	28	0,397	0,397	[0,239 - 0,556]
D8S264	25	59	20	0,659	0,659	[0,530 - 0,786]
D22S928	36	43	25	0,532	0,532	[0,383 - 0,682]
TP53	22	63	19	0,692	0,692	[0,568 - 0,813]
D9S171	35	33	36	0,490	0,490	[0,301 - 0,680]

Tab. 28 – Estimation du taux d'AI pour chaque microsatellite. Les données manquantes sont incluses par l'intermédiaire d'une proportion p_m suivant une loi *a priori* $Be(1;1)$.

6.5 Relations entre microsatellites et stade : calcul des Odds-Ratio

Les microsatellites sont ici analysés sur les données complètes dans leurs relations avec le stade de la tumeur. La relation est quantifiée par l'OR. Celui-ci est calculé ainsi que son IC suivant les approximations fréquentistes classiques, suivant la version exacte [80] et suivant le concept bayésien. Les résultats pour chaque microsatellite sont donnés dans les tableaux suivants. Comme dans le cas de l'estimation d'une proportion, l'IC exact est plus large que l'IC asymptotique. Un seul microsatellite montre une relation avec le stade : il s'agit du microsatellite D5S346 (OR = 2,66, IC exact = [1,03 - 6,94]). TP53 tend à avoir une relation inverse avec le stade, la présence d'un AI sur TP53 étant plutôt associée avec un stade peu avancé.

Microsatellites	OR	$IC_{\mathcal{N}}$	IC exact	$p_{\mathcal{N}}$	p exact
D2S138	0,92	[0,37 - 2,31]	[0,33 - 2,55]	0,924	1,000
D18S61	2,26	[0,76 - 6,73]	[0,68 - 8,18]	0,138	0,186
D16S422	1,15	[0,44 - 3,00]	[0,40 - 3,36]	0,770	0,811
D17S794	1,20	[0,42 - 3,40]	[0,37 - 3,87]	0,735	0,795
D6S264	0,84	[0,31 - 2,29]	[0,28 - 2,56]	0,739	0,803
D14S65	1,37	[0,53 - 3,56]	[0,48 - 4,00]	0,521	0,631
D18S53	0,69	[0,27 - 1,78]	[0,24 - 1,98]	0,440	0,476
D17S790	0,57	[0,22 - 1,47]	[0,19 - 1,63]	0,241	0,334
D1S225	1,22	[0,52 - 2,87]	[0,47 - 3,12]	0,652	0,669
D3S1282	0,73	[0,27 - 1,98]	[0,24 - 2,21]	0,535	0,616
D9S179	1,55	[0,56 - 4,30]	[0,50 - 4,89]	0,401	0,447
D5S430	1,11	[0,36 - 3,42]	[0,31 - 3,94]	0,854	1,000
D8S283	1,44	[0,60 - 3,47]	[0,55 - 3,81]	0,415	0,506
D11S916	2,16	[0,85 - 5,34]	[0,77 - 5,91]	0,107	0,162
D2S159	1,27	[0,52 - 3,10]	[0,48 - 3,40]	0,595	0,654
D16S408	1,39	[0,52 - 3,74]	[0,47 - 4,21]	0,509	0,618
D5S346	2,66	[1,12 - 6,32]	[1,03 - 6,94]	0,025	0,033
D10S191	1,11	[0,45 - 2,73]	[0,41 - 3,01]	0,818	1,000
D13S173	1,80	[0,72 - 4,52]	[0,65 - 5,02]	0,209	0,250
D6S275	0,75	[0,28 - 2,02]	[0,25 - 2,26]	0,564	0,619
D15S127	0,74	[0,31 - 1,79]	[0,28 - 1,95]	0,501	0,654
D1S305	1,94	[0,79 - 4,77]	[0,72 - 5,26]	0,145	0,178
D4S394	0,62	[0,23 - 1,70]	[0,20 - 1,90]	0,352	0,447
D20S107	0,65	[0,25 - 1,73]	[0,22 - 1,93]	0,388	0,461
D1S197	0,97	[0,42 - 2,25]	[0,39 - 2,44]	0,946	1,000
D1S207	1,11	[0,47 - 2,61]	[0,43 - 2,86]	0,817	0,830
D10S192	1,81	[0,75 - 4,35]	[0,69 - 4,78]	0,184	0,268
D3S1283	0,78	[0,32 - 1,90]	[0,29 - 2,08]	0,587	0,655
D4S414	1,29	[0,50 - 3,33]	[0,45 - 3,74]	0,603	0,638
D8S264	2,52	[0,94 - 6,74]	[0,86 - 7,78]	0,062	0,094
D22S928	1,47	[0,60 - 3,58]	[0,55 - 3,94]	0,400	0,498
TP53	0,46	[0,17 - 1,24]	[0,15 - 1,38]	0,121	0,143
D9S171	1,27	[0,49 - 3,30]	[0,44 - 3,67]	0,622	0,637

TAB. 29 – Estimation des OR [IC] sur les données non manquantes par la méthode asymptotique et par la méthode exacte.

6.6 Calcul de l'Odds-Ratio par imputation multiple

6.6.1 Incorporation des manquants par la probabilité p_m

Dans ces résultats les données manquantes sont traitées par modélisation directe de la proportion de manquants en lui donnant une loi *a priori* non informative et en la combinant avec la proportion observée sur les données non-manquantes. Aucune relation ne semble apparaître, avec seul le microsatellite D5S346 ayant un IC à 95% pour l'OR excluant presque la valeur 1.

Le programme utilisé est le programme N°23.

Ce programme permet de définir une distribution pour la proportion d'AI parmi les manquants et de la combiner avec la proportion estimée chez les hétérozygotes. Cette procédure est utilisée d'une part pour les sujets en stade AB et d'autre part chez ceux en stade CD. Les taux d'AI ainsi construits sont ensuite combinés pour calculer la valeur de l'OR mesurant l'association entre le microsatellite étudié et le stade.

Lorsque l'on intègre les manquants directement dans les lois *a priori* des taux d'AI pour les deux stades, le tableau 31 montre que seuls deux microsatellites ont une relation avec le stade : il s'agit de D5S346 avec un OR à 2,72 [1,22 - 6,18] et D8S264 avec un OR à 2,58 [1,08 - 6,43].

Microsatellites	OR	IC(OR)
D2S138	0,962	[0,309 - 2,936]
D18S61	1,672	[0,581 - 5,469]
D16S422	1,075	[0,308 - 3,863]
D17S794	1,091	[0,226 - 5,461]
D6S264	0,925	[0,222 - 3,765]
D14S65	1,106	[0,333 - 4,112]
D18S53	0,742	[0,227 - 2,419]
D17S790	0,692	[0,191 - 2,363]
D1S225	1,202	[0,489 - 2,924]
D3S1282	0,834	[0,191 - 3,483]
D9S179	1,256	[0,307 - 5,527]
D5S430	1,080	[0,170 - 6,857]
D8S283	1,270	[0,496 - 3,413]
D11S916	1,845	[0,608 - 5,688]
D2S159	1,253	[0,459 - 3,368]
D16S408	1,204	[0,334 - 4,591]
D5S346	2,281	[0,967 - 5,593]
D10S191	1,064	[0,407 - 2,839]
D13S173	1,551	[0,524 - 4,740]
D6S275	0,773	[0,222 - 2,740]
D15S127	0,802	[0,286 - 2,196]
D1S305	1,638	[0,586 - 4,774]
D4S394	0,717	[0,192 - 2,602]
D20S107	0,744	[0,204 - 2,638]
D1S197	0,957	[0,403 - 2,290]
D1S207	1,110	[0,473 - 2,591]
D10S192	1,575	[0,600 - 4,270]
D3S1283	0,830	[0,289 - 2,355]
D4S414	1,063	[0,346 - 3,603]
D8S264	1,849	[0,693 - 5,593]
D22S928	1,315	[0,470 - 3,797]
TP53	0,587	[0,198 - 1,573]
D9S171	1,176	[0,360 - 3,938]

TAB. 30 – Valeurs de l'OR pour chaque microsatellite en donnant pour les manquants un *a priori* plat et en intégrant les manquants en modélisant la valeur de p_m, taux d'AI hypothétique chez les manquants.

Microsatellites	OR	IC(OR)
D2S138	0,927	[0,426 - 2,007]
D18S61	2,174	[0,862 - 5,771]
D16S422	1,157	[0,526 - 2,563]
D17S794	1,203	[0,550 - 2,648]
D6S264	0,842	[0,381 - 1,843]
D14S65	1,373	[0,624 - 3,056]
D18S53	0,685	[0,302 - 1,522]
D17S790	0,561	[0,250 - 1,223]
D1S225	1,221	[0,555 - 2,689]
D3S1282	0,726	[0,332 - 1,568]
D9S179	1,565	[0,700 - 3,522]
D5S430	1,113	[0,508 - 2,429]
D8S283	1,451	[0,659 - 3,225]
D11S916	2,155	[0,972 - 4,825]
D2S159	1,280	[0,581 - 2,851]
D16S408	1,403	[0,625 - 3,174]
D5S346	2,712	[1,220 - 6,179]
D10S191	1,112	[0,491 - 2,519]
D13S173	1,823	[0,819 - 4,091]
D6S275	0,744	[0,320 - 1,706]
D15S127	0,736	[0,337 - 1,587]
D1S305	1,965	[0,896 - 4,363]
D4S394	0,617	[0,266 - 1,406]
D20S107	0,648	[0,286 - 1,439]
D1S197	0,971	[0,443 - 2,128]
D1S207	1,108	[0,493 - 2,504]
D10S192	1,828	[0,834 - 4,038]
D3S1283	0,779	[0,357 - 1,683]
D4S414	1,291	[0,573 - 2,928]
D8S264	2,575	[1,077 - 6,432]
D22S928	1,479	[0,677 - 3,238]
TP53	0,449	[0,176 - 1,099]
D9S171	1,277	[0,569 - 2,874]

TAB. 31 – Valeurs de l'OR pour chaque microsatellite en intégrant les manquants dans les lois *a priori*.

6.6.2 Imputation multiple via une régression logistique

Ces résultats donnent la valeur de l'OR estimé en utilisant une imputation multiple avec WinBUGS dans laquelle on utilise le stade de Astler-Coller pour imputer les valeurs man-

quantes de chaque microsatellite. Le programme utilisé est celui du paragraphe « méthode d'imputation probabiliste N°5 » mais pour moitié seulement, la partie pour laquelle l'imputation est faite à partir d'une loi *a priori* n'a pas été incluse ici. Les données complètes et les données imputées (différentes à chaque imputation) sont ensuite introduites dans l'analyse de régression logistique pour estimer la valeur de l'OR. Il y a donc deux régressions logistiques successives. Les résultats sont donnés pour chaque microsatellite individuellement, les analyses étant univariées. Le tableau 32 montre que seul 4 microsatellites ont un OR ayant une probabilité d'être supérieure à 1 dépassant les 95%. Il s'agit des microsatellites D11S916, D5S346, D1S305 et D8S264.

Microsatellites	OR	IC(OR)
D2S138	0,883	[0,326 - 2,394]
D18S61	2,926	[0,960 - 9,499]
D16S422	1,317	[0,446 - 3,861]
D17S794	1,880	[0,414 - 6,815]
D6S264	0,659	[0,203 - 2,280]
D14S65	1,750	[0,604 - 5,082]
D18S53	0,537	[0,193 - 1,528]
D17S790	0,360	[0,128 - 1,014]
D1S225	1,276	[0,528 - 3,074]
D3S1282	0,447	[0,142 - 1,551]
D9S179	2,672	[0,814 - 8,320]
D5S430	2,687	[0,232 - 11,53]
D8S283	1,634	[0,666 - 4,042]
D11S916	3,070	[1,194 - 8,080]
D2S159	1,403	[0,553 - 3,561]
D16S408	1,894	[0,622 - 5,650]
D5S346	3,210	[1,343 - 7,919]
D10S191	1,147	[0,452 - 2,942]
D13S173	2,424	[0,920 - 6,397]
D6S275	0,598	[0,201 - 1,801]
D15S127	0,645	[0,254 - 1,636]
D1S305	2,561	[1,021 - 6,550]
D4S394	0,420	[0,140 - 1,283]
D20S107	0,452	[0,156 - 1,338]
D1S197	0,962	[0,407 - 2,279]
D1S207	1,129	[0,467 - 2,735]
D10S192	2,238	[0,909 - 5,573]
D3S1283	0,693	[0,275 - 1,775]
D4S414	1,492	[0,531 - 4,212]
D8S264	3,345	[1,247 - 9,515]
D22S928	1,742	[0,688 - 4,478]
TP53	0,363	[0,126 - 0,993]
D9S171	1,608	[0,533 - 4,750]

TAB. 32 – Valeurs de l'OR pour chaque microsatellite en utilisant une méthode d'imputation multiple à partir du stade Astler-Coller.

6.7 Résultats des analyses par Facteur de Bayes

Ces premiers résultats concernent les données sans les manquants : les paramètres de la loi de Dirichlet *a priori* sont mis à 1,1,1,1. Les facteurs de Bayes ont été calculés avec 3

valeurs *a priori* de Pr(HN) : 0,1, 0,5 et 0,9. Pour Pr(HN) $= 0,1$, les résultats montrent que seul D5S346 a une probabilité *a posteriori* d'être lié au stade inférieure à 0,05. Le microsatellites D8S264 a une probabilité *a posteriori* de 0,0658. Ces deux exemples montrent que pour que la valeur d'un test fréquentiste rejette l'hypothèse nulle, il faut déjà que la probabilité *a priori* de cette hypothèse soit faible. En effet, quand Pr(HN) =0,5, aucun microsatellite ne semble être associé au stade, comme on peut l'observer en notant que pour tous les microsatellites, Pr($HN|D$) $> 0,24$, et que pour la très grande majorité des microsatellites, Pr($HN|D$) $> 0,7$. Pour une probabilité *a priori* de HN égale à 0,9, les probabilités *a posteriori* de l'hypothèse nulle sont toutes supérieures à 0,7, voire 0,9 pour 31 d'entre eux. Par ailleurs, pour plusieurs microsatellites, on constate que Pr($HN|D$) $>$ Pr(HN), ce qui signifie que les données ont renforcé l'hypothèse nulle, ce qui se traduit donc par une augmentation de la probabilité de l'hypothèse nulle d'absence de relation entre le microsatellite et le stade. Le microsatellite D2S138 est dans cette situation.

Microsatellites $\Pr(HN) = 0,1$	FB	$\Pr(HN\|D)$	$\Pr(p_1 < p_2\|D)$	$\Pr(p_1 > p_2\|D)$
D2S138	3,477	0,278	0,407	0,313
D18S61	1,086	0,107	0,065	0,826
D16S422	3,245	0,265	0,284	0,450
D17S794	2,968	0,248	0,278	0,473
D6S264	3,103	0,256	0,466	0,277
D14S65	2,889	0,243	0,201	0,555
D18S53	2,555	0,221	0,606	0,172
D17S790	1,744	0,162	0,734	0,103
D1S225	3,410	0,274	0,237	0,487
D3S1282	2,737	0,233	0,559	0,207
D9S179	2,901	0,243	0,234	0,521
D5S430	2,898	0,243	0,322	0,433
D8S283	2,696	0,230	0,162	0,606
D11S916	1,010	0,100	0,049	0,849
D2S159	3,163	0,260	0,220	0,518
D16S408	2,673	0,228	0,199	0,571
D5S346	0,325	0,034	0,012	0,952
D10S191	3,491	0,279	0,295	0,424
D13S173	1,638	0,154	0,091	0,754
D6S275	2,785	0,236	0,546	0,216
D15S127	2,949	0,246	0,562	0,190
D1S305	1,282	0,124	0,065	0,809
D4S394	2,136	0,191	0,662	0,145
D20S107	2,323	0,205	0,638	0,156
D1S197	3,824	0,298	0,369	0,332
D1S207	3,638	0,287	0,290	0,421
D10S192	1,559	0,147	0,080	0,771
D3S1283	3,161	0,259	0,521	0,219
D4S414	3,042	0,252	0,229	0,517
D8S264	0,628	0,066	0,030	0,904
D22S928	2,576	0,222	0,158	0,619
TP53	1,037	0,103	0,839	0,056
D9S171	3,020	0,251	0,235	0,513

TAB. 33 – Valeur du Facteur de Bayes pour une probabilité *a priori* de HN de 0,1. HN est ici une hypothèse ponctuelle de type $\theta = \theta_r$.

Microsatellites $\Pr(HN) = 0,5$	FB	$\Pr(HN\|D)$	$\Pr(p_1 < p_2\|D)$	$\Pr(p_1 > p_2\|D)$
D2S138	3,477	0,776	0,126	0,097
D18S61	1,086	0,520	0,035	0,444
D16S422	3,245	0,764	0,091	0,144
D17S794	2,968	0,748	0,093	0,158
D6S264	3,103	0,756	0,152	0,090
D14S65	2,889	0,742	0,068	0,188
D18S53	2,555	0,718	0,218	0,062
D17S790	1,744	0,635	0,319	0,045
D1S225	3,410	0,773	0,074	0,152
D3S1282	2,737	0,732	0,195	0,072
D9S179	2,901	0,743	0,079	0,176
D5S430	2,898	0,743	0,109	0,147
D8S283	2,696	0,729	0,057	0,213
D11S916	1,010	0,502	0,027	0,470
D2S159	3,163	0,759	0,071	0,168
D16S408	2,673	0,727	0,070	0,201
D5S346	0,325	0,245	0,010	0,744
D10S191	3,491	0,777	0,091	0,131
D13S173	1,638	0,621	0,040	0,338
D6S275	2,785	0,735	0,189	0,074
D15S127	2,949	0,746	0,189	0,064
D1S305	1,282	0,561	0,032	0,405
D4S394	2,136	0,681	0,261	0,057
D20S107	2,323	0,699	0,241	0,059
D1S197	3,824	0,792	0,109	0,098
D1S207	3,638	0,784	0,088	0,127
D10S192	1,559	0,609	0,036	0,353
D3S1283	3,161	0,759	0,169	0,071
D4S414	3,042	0,752	0,076	0,171
D8S264	0,628	0,386	0,020	0,593
D22S928	2,576	0,720	0,056	0,222
TP53	1,037	0,509	0,459	0,031
D9S171	3,020	0,751	0,078	0,170

TAB. 34 – Valeur du Facteur de Bayes pour une probabilité *a priori* de HN de 0,5. HN est ici une hypothèse ponctuelle de type $\theta = \theta_r$.

Microsatellites $\Pr(HN) = 0,9$	FB	$\Pr(HN\|D)$	$\Pr(p_1 < p_2\|D)$	$\Pr(p_1 > p_2\|D)$
D2S138	3,477	0,969	0,017	0,013
D18S61	1,086	0,907	0,006	0,085
D16S422	3,245	0,966	0,012	0,020
D17S794	2,968	0,963	0,013	0,022
D6S264	3,103	0,965	0,021	0,012
D14S65	2,889	0,962	0,009	0,027
D18S53	2,555	0,958	0,032	0,009
D17S790	1,744	0,940	0,052	0,007
D1S225	3,410	0,968	0,010	0,021
D3S1282	2,737	0,960	0,028	0,010
D9S179	2,901	0,963	0,011	0,025
D5S430	2,898	0,963	0,015	0,021
D8S283	2,696	0,960	0,008	0,031
D11S916	1,010	0,900	0,005	0,093
D2S159	3,163	0,966	0,010	0,023
D16S408	2,673	0,960	0,010	0,029
D5S346	0,325	0,745	0,003	0,251
D10S191	3,491	0,969	0,012	0,018
D13S173	1,638	0,936	0,006	0,056
D6S275	2,785	0,961	0,027	0,010
D15S127	2,949	0,963	0,027	0,009
D1S305	1,282	0,920	0,005	0,073
D4S394	2,136	0,950	0,040	0,008
D20S107	2,323	0,954	0,036	0,008
D1S197	3,824	0,971	0,014	0,013
D1S207	3,638	0,970	0,012	0,017
D10S192	1,559	0,933	0,006	0,060
D3S1283	3,161	0,966	0,023	0,010
D4S414	3,042	0,964	0,010	0,024
D8S264	0,628	0,849	0,004	0,145
D22S928	2,576	0,958	0,008	0,032
TP53	1,037	0,903	0,090	0,006
D9S171	3,020	0,964	0,011	0,024

TAB. 35 – Valeur du Facteur de Bayes pour une probabilité *a priori* de HN de 0,9. HN est ici une hypothèse ponctuelle de type $\theta = \theta_r$.

Les résultats suivants concernent les données avec les manquants inclus dans les lois *a priori*, ce qui permet de juger de leur importance. Les conclusions obtenues sont similaires à celles obtenues sur les données complètes : les données tendent globalement à soutenir l'hypothèse nulle et seuls trois microsatellites ont des probabilités *a posteriori* inférieures à 0,05 : D11S916, D5S346, D8S264, rejetant donc l'hypothèse nulle. Il faut noter que cela ne s'observe que pour des probabilités *a priori* de 0,1, ce qui présuppose déjà la fausseté de l'hypothèse nulle. Pour les deux autres valeurs *a priori* de l'hypothèse nulle, les conclusions sont similaires à celles obtenues pour les données complètes, à savoir que dans l'ensemble les données soutiennent l'hypothèse nulle.

Microsatellites $\Pr(HN) = 0,1$	FB	$\Pr(HN\|D)$	$\Pr(p_1 < p_2\|D)$	$\Pr(p_1 > p_2\|D)$
D2S138	1,851	0,170	0,446	0,382
D18S61	0,753	0,077	0,024	0,897
D16S422	1,651	0,155	0,265	0,579
D17S794	1,428	0,136	0,261	0,601
D6S264	1,515	0,144	0,492	0,363
D14S65	1,199	0,117	0,136	0,746
D18S53	1,359	0,131	0,714	0,153
D17S790	0,872	0,088	0,814	0,097
D1S225	2,150	0,192	0,184	0,622
D3S1282	1,260	0,122	0,617	0,259
D9S179	1,513	0,143	0,178	0,677
D5S430	1,357	0,131	0,366	0,502
D8S283	1,508	0,143	0,091	0,765
D11S916	0,448	0,047	0,013	0,939
D2S159	1,829	0,168	0,171	0,659
D16S408	1,370	0,132	0,123	0,743
D5S346	0,200	0,021	0,002	0,975
D10S191	2,182	0,195	0,348	0,456
D13S173	0,876	0,088	0,063	0,847
D6S275	1,492	0,142	0,553	0,303
D15S127	1,607	0,151	0,601	0,246
D1S305	0,693	0,071	0,026	0,902
D4S394	1,183	0,116	0,781	0,102
D20S107	1,197	0,117	0,696	0,186
D1S197	2,345	0,206	0,441	0,351
D1S207	2,686	0,229	0,254	0,515
D10S192	0,879	0,089	0,035	0,875
D3S1283	1,647	0,154	0,629	0,215
D4S414	1,511	0,143	0,167	0,688
D8S264	0,376	0,040	0,012	0,947
D22S928	1,396	0,134	0,091	0,773
TP53	0,693	0,071	0,890	0,037
D9S171	2,493	0,216	0,291	0,491

TAB. 36 – Valeur du Facteur de Bayes pour une probabilité *a priori* de HN de 0,1. HN est ici une hypothèse ponctuelle de type $\theta = \theta_r$. Les manquants sont inclus dans les distributions *a priori* et on utilise une estimation du maximum de vraisemblance pour spécifier ces manquants.

Microsatellites $\Pr(HN) = 0,5$	FB	$\Pr(HN\|D)$	$\Pr(p_1 < p_2\|D)$	$\Pr(p_1 > p_2\|D)$
D2S138	1,851	0,649	0,188	0,161
D18S61	0,753	0,429	0,015	0,554
D16S422	1,651	0,622	0,118	0,258
D17S794	1,428	0,588	0,124	0,287
D6S264	1,515	0,602	0,228	0,168
D14S65	1,199	0,545	0,070	0,384
D18S53	1,359	0,576	0,348	0,075
D17S790	0,872	0,465	0,477	0,056
D1S225	2,150	0,682	0,072	0,244
D3S1282	1,260	0,557	0,311	0,130
D9S179	1,513	0,602	0,083	0,314
D5S430	1,357	0,575	0,179	0,245
D8S283	1,508	0,601	0,042	0,356
D11S916	0,448	0,309	0,009	0,680
D2S159	1,829	0,646	0,072	0,280
D16S408	1,370	0,578	0,060	0,361
D5S346	0,200	0,166	0,002	0,831
D10S191	2,182	0,685	0,136	0,178
D13S173	0,876	0,467	0,037	0,495
D6S275	1,492	0,598	0,259	0,142
D15S127	1,607	0,616	0,272	0,111
D1S305	0,693	0,409	0,016	0,573
D4S394	1,183	0,542	0,404	0,053
D20S107	1,197	0,544	0,359	0,096
D1S197	2,345	0,701	0,166	0,132
D1S207	2,686	0,728	0,089	0,181
D10S192	0,879	0,467	0,020	0,511
D3S1283	1,647	0,622	0,281	0,096
D4S414	1,511	0,601	0,077	0,320
D8S264	0,376	0,273	0,009	0,716
D22S928	1,396	0,582	0,044	0,372
TP53	0,693	0,409	0,566	0,023
D9S171	2,493	0,713	0,106	0,179

TAB. 37 – Valeur du Facteur de Bayes pour une probabilité *a priori* de HN de 0,5. HN est ici une hypothèse ponctuelle de type $\theta = \theta_r$. Les manquants sont inclus dans les distributions *a priori* et on utilise une estimation du maximum de vraisemblance pour spécifier ces manquants.

Microsatellites $\Pr(HN) = 0,9$	FB	$\Pr(HN\|D)$	$\Pr(p_1 < p_2\|D)$	$\Pr(p_1 > p_2\|D)$
D2S138	1,851	0,943	0,030	0,026
D18S61	0,753	0,871	0,003	0,125
D16S422	1,651	0,936	0,019	0,043
D17S794	1,428	0,927	0,021	0,050
D6S264	1,515	0,931	0,039	0,029
D14S65	1,199	0,915	0,013	0,071
D18S53	1,359	0,924	0,062	0,013
D17S790	0,872	0,887	0,100	0,012
D1S225	2,150	0,950	0,011	0,037
D3S1282	1,260	0,919	0,057	0,023
D9S179	1,513	0,931	0,014	0,054
D5S430	1,357	0,924	0,031	0,043
D8S283	1,508	0,931	0,007	0,061
D11S916	0,448	0,801	0,002	0,195
D2S159	1,829	0,942	0,011	0,045
D16S408	1,370	0,925	0,010	0,064
D5S346	0,200	0,642	0,000	0,356
D10S191	2,182	0,951	0,020	0,027
D13S173	0,876	0,887	0,007	0,104
D6S275	1,492	0,930	0,044	0,024
D15S127	1,607	0,935	0,045	0,018
D1S305	0,693	0,861	0,003	0,134
D4S394	1,183	0,914	0,075	0,009
D20S107	1,197	0,915	0,066	0,017
D1S197	2,345	0,954	0,025	0,020
D1S207	2,686	0,960	0,013	0,026
D10S192	0,879	0,887	0,004	0,107
D3S1283	1,647	0,936	0,047	0,016
D4S414	1,511	0,931	0,013	0,055
D8S264	0,376	0,772	0,003	0,224
D22S928	1,396	0,926	0,007	0,065
TP53	0,693	0,861	0,132	0,005
D9S171	2,493	0,957	0,015	0,026

TAB. 38 – Valeur du Facteur de Bayes pour une probabilité *a priori* de HN de 0,9. HN est ici une hypothèse ponctuelle de type $\theta = \theta_r$. Les manquants sont inclus dans les distributions *a priori* et on utilise une estimation du maximum de vraisemblance pour spécifier ces manquants.

Les résultats suivants (tableau 39) concernent les tests pour une hypothèse nulle composite définie à partir des lois *a priori* sous l'hypothèse nulle. Dans ces calculs, l'hypothèse nulle est composite, c'est-à-dire que l'on teste des hypothèses telles que $\theta \in \Theta = \{\theta_{inf}, \ldots, \theta_{sup}\}$ où Θ est en général une demi-droite dont l'une des bornes vaut 0 (ou 1 par exemple pour un OR).

Plusieurs microsatellites présentent des valeurs de FB modérément élevées ou élevées, comme par exemple le D18S61, le D1S380, le D11S916, le D5S346 ou encore le D8S264. La dernière colonne du tableau donne la probabilité que $p_1 < p_2$, ce qui revient à donner la probabilité que l'OR soit inférieur à 1, $\Pr(OR < 1)$. Pour les 4 microsatellites cités, cette probabilité est inférieure à 0,05, ce qui tendrait à soutenir l'hypothèse d'une relation entre le microsatellite et le stade. Si l'on utilise l'échelle des FB proposée par Kass [169], seul le microsatellite D5S346 présente un FB notable.

Dans le tableau suivant (tab. N° 40), on teste également une hypothèse composite mais en spécifiant les manquants dans les paramètres *a priori* de la fonction. Les valeurs des FB sont toutes diminuées par rapport aux valeurs notées dans le tableau 39. Les résultats en terme d'OR supérieur à 1 sont néanmoins encore soutenus pour les microsatellites D11S916, D8S264 et D5S346. Mais là aussi, si l'hypothèse *a posteriori* est en faveur d'une relation, c'est surtout parce que l'hypothèse *a priori* était déjà en faveur d'une relation.

Microsatellites	FB	$\Pr(p_1 < p_2 \mid HN)$	$\Pr(p_1 < p_2 \mid D)$
D2S138	0,769	0,5	0,565
D18S61	12,600	”	0,073
D16S422	1,581	”	0,387
D17S794	1,696	”	0,370
D6S264	0,594	”	0,627
D14S65	2,751	”	0,266
D18S53	0,285	”	0,778
D17S790	0,140	”	0,876
D1S225	2,057	”	0,327
D3S1282	0,371	”	0,729
D9S179	2,222	”	0,310
D5S430	1,343	”	0,426
D8S283	3,724	”	0,211
D11S916	17,110	”	0,055
D2S159	2,348	”	0,298
D16S408	2,868	”	0,258
D5S346	74,179	”	0,013
D10S191	1,435	”	0,410
D13S173	8,294	”	0,107
D6S275	0,396	”	0,716
D15S127	0,339	”	0,746
D1S305	12,380	”	0,074
D4S394	0,218	”	0,820
D20S107	0,244	”	0,803
D1S197	0,897	”	0,526
D1S207	1,447	”	0,408
D10S192	9,581	”	0,094
D3S1283	0,420	”	0,704
D4S414	2,250	”	0,307
D8S264	29,55	”	0,032
D22S928	3,919	”	0,203
TP53	0,067	”	0,936
D9S171	2,185	”	0,313

TAB. 39 – Valeurs du Facteur de Bayes pour tester une hypothèse composite pour les 33 microsatellites. La probabilité *a priori* de HN est définie implicitement par les paramètres des lois *a priori* : $\Pr(HN) = 0,5$.

Microsatellites	FB	$\Pr(p_1 < p_2 \mid HN)$	$\Pr(p_1 > p_2 \mid D)$
D2S138	0,530	0,382	0,538
D18S61	3,231	0,082	0,027
D16S422	1,383	0,387	0,313
D17S794	1,234	0,349	0,302
D6S264	0,604	0,449	0,574
D14S65	1,430	0,206	0,154
D18S53	0,494	0,696	0,822
D17S790	0,260	0,686	0,893
D1S225	0,795	0,190	0,228
D3S1282	0,456	0,520	0,704
D9S179	2,120	0,359	0,208
D5S430	1,176	0,462	0,422
D8S283	1,188	0,123	0,106
D11S916	11,40	0,140	0,014
D2S159	1,048	0,214	0,206
D16S408	1,314	0,179	0,142
D5S346	8,337	0,020	0,002
D10S191	1,543	0,541	0,433
D13S173	3,738	0,219	0,070
D6S275	0,244	0,308	0,645
D15S127	0,343	0,455	0,709
D1S305	3,636	0,095	0,028
D4S394	0,446	0,773	0,884
D20S107	0,343	0,562	0,788
D1S197	1,124	0,585	0,556
D1S207	0,662	0,246	0,330
D10S192	2,710	0,098	0,038
D3S1283	0,619	0,643	0,744
D4S414	1,132	0,216	0,195
D8S264	9,916	0,118	0,013
D22S928	1,428	0,144	0,105
TP53	0,153	0,784	0,959
D9S171	2,316	0,578	0,372

TAB. 40 – Valeurs du Facteur de Bayes pour tester une hypothèse composite pour les 33 microsatellites. La probabilité *a priori* de HN est définie implicitement par les paramètres des lois *a priori* qui incluent ici les données manquantes. Cette valeur est différente pour chaque microsatellite.

La limite principale de la solution bayésienne exposée ici tient en fait dans la nature du problème. Rappelons que l'objectif est ici de décrire et de modéliser les relations entre eux

des différents microsatellites. Le modèle bayésien ne gère pas spécifiquement la colinéarité des données et encore moins les problèmes liés aux dimensions de la matrice. L'estimation des paramètres, qu'elle soit fréquentiste ou bayésienne, ne peut se faire que sur des matrices de rang plein. Il faut donc adopter d'autres méthodes (fréquentiste ou bayésienne) pour franchir ces deux obstacles.

Les méthodes descriptives multivariées sont ici toute indiquées pour, au minimum, décrire les données puis pour tenter de réaliser de l'inférence. Différentes méthodes et variantes existent. Nous en citerons les principales avant de décrire l'une d'entre elle en particulier pour laquelle nous avons combiné deux variantes afin de répondre au problème.

7 Les Méthodes Multivariées

Les méthodes descriptives multivariées sont très nombreuses. Les approches sont très variées mais leur point commun est de réaliser une projection des données sur des sous-espaces, réduisant la dimensionalité des données [3, 52, 289, 333, 347]

7.1 Analyses en cluster

Une analyse en cluster permet de chercher la présence d'une structure dans les données. Le cluster présenté dans la figure suivante a été réalisé en recodant les sujets normaux en « 0 », les AI en « 1 » et les homozygotes sont notés comme manquants. La mesure de distance est le coefficient de corrélation de Pearson (valide sur des valeurs 0/1) et la méthode d'agrégation choisie est la méthode du lien moyen. Cette analyse, réalisée avec le logiciel `CLUSTER`, (Eisensoftware®), permet d'effectuer les calculs malgré les données manquantes en travaillant pour chaque couple de microsatellites sur les sujets complets. Cela fait donc l'hypothèse que les manquants sont de type MAR ou MCAR. Les résultats peuvent être instables en raison des effectifs différents pour chaque couple de microsatellites. La structure hiérarchique est visualisée avec le logiciel `JavaTreeView`. Cette analyse a permis de montrer l'existence de trois sous-groupes de patients se distinguant par une fréquence différente de taux d'AI. Ces résultats ont été détaillés dans l'article de Weber, Meyer *et al.* [337]. Ce type d'analyse fournit une réponse incomplète au problème de départ. En effet, les données manquantes ne sont pas véritablement prises en considération. Ceci n'est pas tout à fait un problème puisqu'une des analyses précédentes à montré que les manquants sont probablement de type MAR ou MCAR. Néanmoins cela ne peut être affirmé de façon définitive et

la possibilité d'inclure les manquants directement serait un plus pour l'analyse. Ensuite, les analyses descriptives multivariées sont des méthodes descriptives comme leur nom l'indique ce qui rappelle qu'il n'est pas possible de réaliser des modèles de régression pour les données. Or l'un des objectifs de ce travail est justement de pouvoir établir des modèles de régression permettant de prédire un résultat tel que le stade ou la survie. Pour cela, il faut donc utiliser des modèles ayant cette double aptitude de faire à la fois des analyses descriptives et des modèles de régression en conservant l'aspect multivarié de l'analyse. Une telle méthode est décrite dans le chapitre suivant dans lequel nous proposons une adaptation d'un modèle de base au cas particulier des données d'allélotypage.

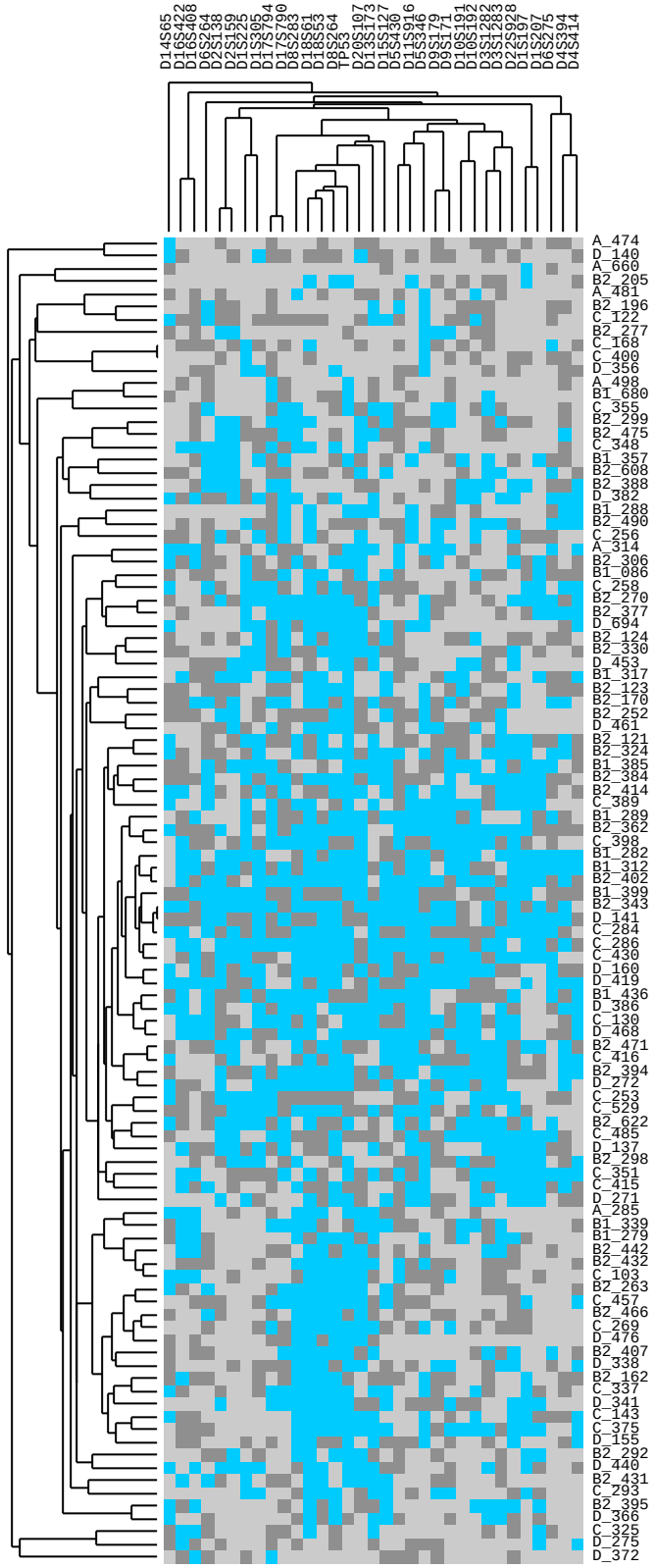

FIG. 7 – Analyse en cluster des données.

7.2 La régression PLS

7.2.1 La *méthode* PLS

Rappel sur la régression linéaire Une régression linéaire simple est un modèle adapté à l'analyse des relations entre une variable quantitative résultat Y et une série de variables explicatives quantitatives ou qualitatives X_i [289]. Le modèle s'écrit de la façon suivante :

$$y = \alpha + \sum_i \beta_i x_i + \varepsilon$$

pour $i \in \{1, \ldots, n\}$ et :

$$\varepsilon \rightsquigarrow N(0, 1)$$

et l'estimation de l'espérance de Y, conditionnellement aux valeurs de x_i s'écrit :

$$\mathbb{E}(y|x_i) = \hat{y} = \alpha + \sum_i \beta_i x_i$$

Rappelons que dans ce modèle, pour un échantillon de taille n avec p variables, pour pouvoir estimer les paramètres du modèle, il faut que : $n > p$ sinon il n'y a pas d'estimation possible. Si la condition est remplie, alors on estime les paramètres de la façon suivante, dans une régression linéaire simple :

Estimation b du paramètre β par :

$$b = \frac{Cov(x, y)}{Var(x)}$$

Estimation de r :

$$r = \frac{Cov(x, y)}{\sqrt{var(x)var(y)}}$$

De façon générale, pour une matrice de taille $n \times p$, on obtient les paramètres par :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

et la valeur attendue de Y, $\hat{\mathbf{y}}$ s'obtient, en l'absence de pondération particulière, par :

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

La source principale d'instabilité dans l'estimation du vecteur β est la multicolinéarité [71]. On a :

$$\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Si les prédicteurs sont très corrélés entre eux , $\mathbf{X'X}$ est mal conditionné, son déterminant est proche de 0 et son inverse aura des valeurs très grandes. De ce fait, l'estimation des paramètres β est imprécise. On peut avoir une mesure globale de la multicolinéarité basée sur le facteur d'inflation de la variance (FIV). Ce FIV vaut :

$$\frac{1}{1 - R_j^2}$$

où R_j^2 est le carré du coefficient de corrélation multiple de x_j et des $p-1$ autres variables. La moyenne des p facteurs d'inflation est une mesure globale de la multicolinéarité.

Lorsque la matrice des variables explicatives est « rectangulaire » ($n << p$), la matrice de covariance $\mathbf{X'X}$, de taille $p \times p$ est de rang $n - 1$ au maximum et est singulière. De ce fait on ne peut pas utiliser les moindres carrés ordinaires.

Pour contourner ces problèmes de la colinéarité et de dimensions de la matrice, il existe plusieurs solutions [71, 289]. Deux d'entre elles seront décrites brièvement dans les paragraphes suivants. Une troisième solution sera détaillé et modifiée dans un paragraphe subséquent. La sélection des variables dans le modèle final de régression PLS a été étudiée par Gauchi [119] mais nous ne discuterons pas ici de cet aspect du problème.

La régression sur composantes principales L'analyse en composante principale permet de décomposer une matrice de données X contenant p variables corrélées en une série de composantes principales qui sont des combinaisons linéaires des p variables d'origine et qui ne sont pas corrélées entre elles [289]. On peut alors introduire dans une régression linéaire multiple ces composantes principales en lieu et place des variables d'origine. Si l'interprétation en est plus délicate, le modèle permet néanmoins d'utiliser toute l'information utile de la matrice X puisque les composantes engendrent le même espace. En pratique, le modèle revient à réaliser p régressions simples :

$$\hat{y} = \sum_{j=i}^{p} \alpha_i \mathbf{c}_i$$

où

$$\alpha_j = \frac{r(y; c_j) s_y}{\sqrt{\lambda_j}}$$

En retenant les j premières composantes principales, on obtient une solution approchée. On peut ensuite réexprimer la régression sur les composantes en fonction des variables X_i d'origine. Notons enfin que les composantes de fortes variances ne sont pas forcément les

plus explicatives puisque la variance des composantes est déterminée uniquement à partir de la décomposition de X et non pas à partir de la corrélation entre la composante j et le vecteur y, contrairement à la régression PLS qui sera vue plus loin. En conséquence de quoi, la solution obtenue n'est pas forcément optimale en terme d'explication de y. Par ailleurs, par construction, cette solution s'adresse au traitement de données quantitatives et ne semble pas indiquée pour les données qui nous concernent ici. Notons que Tenenhaus utilise la méthode PLS dans un cas similaire à l'ACP sur des données qualitatives [321]. Son argument consiste à dire que la méthode ne fait pas d'hypothèse distributionnelle et en effet, il n'y a pas dans une ACP de résidus qui devraient suivre une loi de Gauss par exemple. Jollife [163] confirme que si l'ACP n'est utilisé qu'en tant que méthode descriptive et pas comme une méthode inférentielle, elle peut traiter des données qualitatives. Le fait de résumer des parts de la variance du jeu de données peut se faire quel que soit le type de données. Par ailleurs, on sait que l'analyse factorielle des correspondances et l'analyse factorielle des correspondances multiples sont une forme d'ACP pour variable qualitative. La méthode suppose tout de même un minimum d'aménagement. Pour revenir à l'utilisation de la PLS linéaire sur des variables qualitatives [89, 289], elle sera comparée avec le modèle que nous proposons plus loin.

La *Ridge* régression Cette méthode utilise l'estimateur $\mathbf{b}k$ suivant :

$$\mathbf{b}(k) = (\mathbf{X}\mathbf{X}' + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

où k est une constante positive à déterminer. Si $k = 0$, alors $\mathbf{b}(k)$ est l'estimateur des moindres carrés ordinaires. La valeur de k doit être déterminée en fonction d'un paramètre inconnu β. Pour contourner le problème, on choisit la valeur de k à partir d'un graphe étudiant la variation de k. Cette méthode est peu employée.

7.2.2 La régression linéaire PLS

La méthode PLS (Partial Least Squares) est une méthode permettant d'établir les relations entre une matrice Y et une matrice X observées sur le même ensemble d'individus et où la matrice Y est soit une matrice à expliquer comme le résultat des variables explicatives contenues dans X soit une matrice superposable à X. Dans ce deuxième cas, on se trouve dans la situation d'analyse de tableaux multiples telles que l'analyse canonique, les méthodes procrustéennes et la modélisation de relations structurelles sur variables latentes.

La méthode prend alors le nom d'*approche PLS*. En l'absence de matrice Y, la méthode PLS se réduit à une analyse en composante principale [97]. Le terme de Moindres Carrés Partiels (Partial Least Squares) traduit le fait que les paramètres du modèle sont obtenus par la réalisation de plusieurs régressions par moindres carrés successives [52, 53, 145, 321].

La méthode est née en 1966 de l'algorithme NIPALS développé par Wold pour réaliser une ACP en présence de données manquantes [342, 344] et de l'approche PLS proposée par Wold en 1975 [321]. Elle très couramment utilisée dans le domaine de la chémométrie [20, 120, 133, 259, 321, 342].

Lorsque l'on parle de PLS, on parle en fait d'une famille de méthodes dans laquelle on mélange et confond les principes, les modèles, les variantes de modèles et les méthodes d'estimation. On trouve globalement :

La méthode PLS : c'est un équivalent de l'ACP ;

La PLS1 ou régression PLS1 : il s'agit d'une régression linéaire d'une variable y sur une matrice X utilisant l'algorithme PLS ;

La PLS2 ou régression PLS2 : il s'agit d'une régression linéaire d'au moins deux variables y sur une matrice X utilisant l'algorithme PLS ;

L'Analyse discriminante PLS : c'est l'application de la méthode PLS à l'analyse discriminante ;

Le SIMPLS : c'est l'acronyme de « Straightforward Implementation of a statistically inspired Modification of the PLS ». Il s'agit d'un algorithme développé par De Jong [86] pour la PLS1. Ces résultats sont très proches de la régression PLS même si elle s'en distingue par des contraintes différentes dans la recherche des composantes ;

Le NIPALS : c'est l'algorithme le plus important, presque synonyme de PLS, car il permet d'estimer les composantes même lorsque les matrices X et Y présentent des données manquantes [238, 239, 321]. En l'absence de données manquantes, les résultats obtenus sont évidemment identiques à ceux obtenus avec les autres algorithmes. Il est donc plus simple de n'utiliser que celui-ci puisqu'il répond à toutes les situations rencontrées. Nous l'avons adapté au cas des données catégorielles à partir du travail de Bastien [31]. Il est très étonnant que cet algorithme utilisable pour la régression PLS et pour l'ACP ne soit pas plus connu [321]. C'est le seul que nous développerons par la suite. D'autres variantes existent [42, 52, 53, 137, 149, 157, 251, 252, 289, 340, 343, 351], combinant par exemple analyse en cluster et modèle PLS [30], PLS et ana-

lyse discriminante [27, 251, 252, 324] ou explorant ses performances sur des données qualitatives [89, 116, 122, 321], sur des données de survie [250, 262] ou dans le *path modelling* [320]. Elle a également été utilisée pour déterminer les facteurs d'expansion de séquences répétées de l'ADN [350].

Enfin, signalons que dans la grande famille des méthodes multivariées descriptives ou inférentielles, la PLS fait partie des méthodes non-supervisées lorsque l'on ne prend pas en compte la matrice Y (comme l'ACP) et fait partie des méthodes supervisées lorsque l'on prend la matrice Y en considération et que l'on réalise une régression.

Formalisation de la régression PLS Soit une matrice X de variables explicatives et y un vecteur (univarié) de résultats, tous deux observés sur n sujets. La matrice X est de taille $n \times p$ sans contrainte sur la valeur de n par rapport à celle de p. On suppose dans cette partie que les matrices X et y sont centrées et réduites. Si ce n'est pas le cas, il faut se ramener à cette situation. L'expression des résultats peut se faire *in fine* en fonction des variables d'origine afin de faciliter pour l'utilisateur l'interprétation des résultats.

La régression PLS est basée sur une décomposition en variable latente :

$$\mathbf{y} = \mathbf{T}'\mathbf{X} + \mathbf{E}$$

avec

$$\mathbf{T} = \mathbf{XW}$$

T est la matrice des composantes latentes, dont les colonnes sont indépendantes (orthogonales). **E** est une matrice de résidus. **W** est une matrice de poids. La matrice **T** contient les c composantes. On a :

$$T_1 = w_{11}X_1 + \cdots + w_{p1}X_p$$

$$\cdots = \ldots$$

$$T_c = w_{1c}X_1 + \cdots + w_{pc}X_p$$

- T est une matrice de t composantes orthogonales
  qui maximisent la covariance entre X et y ;
- le plus souvent comme en ACP : $t < p$;
- les composantes sont donc explicatives des X et prédisent y.

Présentation de l'algorithme NIPALS en l'absence de données manquantes pour la PLS1 Le principe général de cet algorithme consiste à réaliser une séquence itérative de régressions simples qui permettent la prise en compte des valeurs manquantes. Les résultats de ces régressions simples sont combinés entre eux pour définir les composantes principales successives expliquant une part décroissante de l'information contenue dans les données. Les régressions simples sont des régressions sans ordonnées à l'origine (ou sans intercept) ce qui permet de comprendre pourquoi les données manquantes n'interviennent pas dans l'obtention des paramètres du modèle.

En reprenant les notations du paragraphe précédent, on cherche à réaliser une régression d'une variable y sur des variables explicatives x_i qui peuvent être très corrélées entre elles, qui peuvent être en nombre plus grand que le nombre d'unités statistiques et en veillant à ce que les coefficients de la régression soient interprétables, c'est-à-dire que le coefficient d'une variable doit permettre à l'utilisateur de juger du rôle de la variable dans la constitution de y en ayant un sens concret. Ce point est important dans les conditions d'utilisation spécifiées car en cas de colinéarités, la régression linéaire multiple tend à produire pour deux variables colinéaires des coefficients de signes inverses ce qui n'est pas cohérent. La présentation reprend le plan de Tenenhaus [321]. Soit une matrice X de rang a.

Étape 1 : $X_0 = X$; $y_0 = y$

Étape 2 : pour $h = 1, 2, \ldots, a$;

Étape 2.1 : $w_h = X'_{h-1} y_{h-1} / y'_{h-1} y_{h-1}$

Étape 2.2 : Normer w_h à 1

Étape 2.3 : $t_h = X_{h-1} w_h / w'_h w_h$

Étape 2.4 : $p_h = X'_{h-1} t_h / t'_h t_h$

Étape 2.5 : $X'_h = X_{h-1} - t_h p'_h$

Étape 2.6 : $c_h = y'_{h-1} t_h / t'_h t_h$

Étape 2.7 : $u_h = y_{h-1} / c_h$

Étape 2.8 : $y_h = y_{h-1} - c_h t_h$

Les coordonnées des vecteurs w_h, t_h, p_h et c_h représentent des pentes de droites des moindres carrés passant par l'origine et peuvent donc être calculées lorsqu'il y a des données manquantes.

w_{hj} : c'est le coefficient de régression de y_{h-1} dans la régression de la j-ième colonne de X_{h-1} sur la variable y_{h-1} ;

t_{hi} : c'est le coefficient de régression de w_h dans la régression de la i-ième ligne de X_{h-1}

sur la variable définie par w_h ;

p_{hj} : c'est le coefficient de régression de t_h dans la régression de la j-ième colonne de X_{h-1} sur la variable t_h ;

c_h : c'est le coefficient de régression de t_h dans la régression de la y_{h-1} sur la variable t_h.

S'il n'y a pas de données manquantes, l'algorithme peut être simplifié dans certaines étapes mais ces simplifications ne sont pas nécessaires.

Une autre présentation de la méthode PLS permet de mettre en relief ses caractéristiques. On cherche une série de composantes orthogonales. La première s'écrit :

$$t_1 = w_{11}x_1 + \cdots + w_{1p}x_p \tag{21}$$

où le coefficient w_{1j} s'obtient de la façon suivante :

$$w_{1j} = \frac{cov(y, x_j)}{\sqrt{\sum_{j=1}^{p} cov^2(y, x_j)}} \tag{22}$$

Le coefficient w_{1j} est égal à la covariance entre le vecteur y et le vecteur x_j, standardisé par l'ensemble des coefficients w_{1j}. L'ensemble des coefficients w_{1j} est obtenu en réalisant de manière itérative toutes les régressions linéaires simples de y sur chacune des variables x_j, indépendamment les unes des autres. Une fois obtenue la première composante, on calcule la régression suivante :

$$y = ct_1 + y_1$$

Le vecteur y_1 est un vecteur de résidus. On peut donc réécrire cette formule de la façon suivante :

$$y = c_1w_{11}x_1 + \cdots + c_1w_{1p}x_p + y_1 \tag{23}$$

On calcule ensuite les résidus des régressions x_{1j} des variables x_j sur la composante t_1. La composante t_2 est une combinaison linéaire des résidus x_{1j} :

$$t_2 = w_{21}x_1 + \cdots + w_{2p}x_p \tag{24}$$

et :

$$w_{2j} = \frac{cov(y_1, x_{1j})}{\sqrt{\sum_{j=1}^{p} cov^2(y_1, x_{1j})}} \tag{25}$$

On calcule ensuite :

$$y = c_1 t_1 + c_2 t_2 + y_2 \tag{26}$$

On peut bien sûr réexprimer y en fonction des variables d'origine pour donner plus de sens au modèle. Cette présentation met bien en évidence que la régression PLS est une méthode combinant la recherche de composantes orthogonales de X (on explique au fur à mesure les résidus de la matrice X) avec l'explication de la variable résultat puisque les composantes sont construites à partir de la covariance entre les x_i et la variable résultats. La régression PLS est donc un modèle plus avancé que la régression sur composantes principales puisque les composantes sont ici obtenues sous contrainte de maximisation de la covariance avec y. Par ailleurs, toutes les régressions sont des régressions sans ordonnées à l'origine ce qui en cas de données manquantes facilite l'obtention des paramètres. En effet, s'il y a des données manquantes, les équations (22) et (25) sont modifiées de façon à n'utiliser que les données présentes. On utilise alors :

$$t_{1i} = \frac{\sum_{j:x_{ji}\ existe} w''_{1j} x_{ji}}{\sum_{\{j:x_{ji}\ existe\}} (w''_{1j})^2} \tag{27}$$

dans laquelle le coefficient w''_{1j} est défini par :

$$w'_{1j} = \frac{\sum_{j:x_{ji}\ et\ y_i\ existent} x_{ji} y_i}{\sum_{\{j:x_{ji}\ et\ y_i\ existent\}} y_i^2} \tag{28}$$

Le produit $x_{ji}y_i$ est égal à la covariance entre x_{ji} et y_i puisque toutes les variables sont centrées réduites. Comme les estimations se font de façon itératives, la seule perte d'information est la perte liée aux données manquantes pour la variable x_i et non pas pour les sujets ayant au moins une donnée manquante sur l'une des variables. La méthode ne crée donc pas plus de données manquantes qu'il n'y en a vraiment (contrairement à la méthode du cas complet ou la méthode du cas disponible). La perte de données est donc celle qui existait dans le jeu de données *avant* l'application du modèle contrairement à la plupart des autres méthodes. Il n'est pas nécessaire d'estimer ces données manquantes et de plus il est possible, une fois le modèle établi, de « retourner » l'équation pour estimer (de manière ponctuelle) les valeurs manquantes.

En résumé :

– la régression PLS est un algorithme d'estimation des paramètres du modèle ;
– au cours duquel on cherche t composantes orthogonales ;
– qui maximisent la covariance entre y et X (var. centrées - réduites) ;

– la première composante t_1 est définie par :

$$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} cov(y, x_j)^2}} \sum_{j=1}^{p} cov(y, x_j) x_j$$

En pratique, l'implémentation se fait de la façon suivante :

– la quantité $cov(y, x_j)$ est aussi le coefficient de régression de la régression simple de y sur x_j ;

– on peut donc réaliser un test pour évaluer le rôle de la variable x_j dans la constitution de t_h.

Puis :

– on calcule la régression simple $y = a_{ij} x_j + \varepsilon_1$;

– puis les p régressions : $x_j = p_{1j} t_1 + x_{1j}$;

– puis on calcule la seconde composante t_2.

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} cov(y_1, x_{1j})^2}} \sum_{j=1}^{p} cov(y_1, x_{1j}) x_{1j}$$

Le procédé est itéré sur autant de composantes qu'on le souhaite.

Pourquoi la PLS a-t-elle tant d'avantages ?

– elle ne fait pas d'hypothèse probabiliste ce qui n'est pas gênant tant que l'on n'estime pas d'intervalle de confiance pour les paramètres et que l'on se contente d'estimations ponctuelles ;

– en cas de données manquantes l'estimation se fait sur les seules données présentes, x_j par x_j (algorithme NIPALS) ;

– on peut travailler sur une matrice x_j de taille quelconque même si $n << p$ puisque les t_i sont des combinaisons linéaires des x_j ;

– on peut donc utiliser le modèle même lorsque les variables sont colinéaires ;

– les coefficients des variables colinéaires sont cohérents.

7.2.3 La régression linéaire généralisée PLS

Que faire si l'on a des données (uniquement) qualitatives ?

Les hypothèses de la régression linéaire n'interdisent pas que la variable y soit qualitative si les résidus suivent une loi gaussienne. On peut donc dans cette situation utiliser une régression PLS1 voire même une PLS2 en travaillant sur une matrice de variables

indicatrices de q colonnes si y est une variable multinomiale à q catégories. Il n'est pas totalement impossible d'utiliser également des variables X qualitatives, mais l'on risque beaucoup plus certainement de s'éloigner des conditions d'utilisation du modèle linéaire, les moindres carrés étant des mauvais estimateurs pour données qualitatives [81].

Pour contourner ces difficultés, il a été proposé d'étendre les principes et algorithmes de la méthode PLS à des modèles linéaires généralisés ce qui permet de construire des versions PLS de modèles classiques tels que la régression logistique ou le modèle de Cox. On peut également l'étendre à d'autres modèles ayant d'autres fonctions de liens [31]. Deux grandes approches ont été utilisées jusqu'à maintenant pour réaliser de la PLS-GLM. Nous commencerons par décrire la méthode de Nguyen et Rocke puis celle de Bastien.

Méthode de Nguyen Une approche en deux étapes a été proposée par Nguyen [250, 251] pour des variables réponses binomiales ou multinomiales. La première étape est l'étape de réduction des dimensions par recherche des composantes et la seconde étape est l'étape de classification utilisant les composantes PLS comme prédicteur. L'auteur compare trois méthodes de classification : la régression logistique, l'analyse linéaire discriminante et l'analyse quadratique discriminante. La seconde méthode semble donner les meilleurs résultats. La régression logistique qui pourrait sembler indiquée de prime abord, se révèle instable en pratique, en raison de l'absence d'estimation du maximum de vraisemblance lorsqu'il y a séparation ou quasi-séparation des données à partir des composantes. Ce phénomène est assez fréquent lorsque l'on utilise des composantes latentes par PLS. Une solution aurait été d'utiliser la méthode de Firth [108] présentée par [141, 142]. Nous y reviendrons plus loin.

Nguyen propose en fait deux méthodes différentes. La première consiste à calculer les composantes T_i en utilisant une série de régressions linéaires simples puis d'intégrer ces composantes dans une régression logistique en lieu et place des prédicteurs X d'origine. Le principe est utilisé avec plusieurs variantes pour faire des modèles de régression, de l'analyse linéaire discriminante et de l'analyse quadratique discriminante. La méthode est simple mais elle présente un inconvénient : on suppose que les données sont quantitatives. Dans une publication postérieure, il propose de remplacer les régressions linéaires par des régressions logistiques mais cela n'apparaît pas très clairement dans la publication [251]. Cette seconde méthode est très proche de celle décrite, beaucoup plus clairement, par Bastien.

Méthode de Bastien La méthode de Bastien [31] est plus cohérente que celle de Nguyen [250, 251] dans le sens où les composantes t_i sont obtenues dans le cadre du modèle linéaire généralisé. La méthode consiste à remplacer les régressions linéaires individuelles utilisées lors de la construction de chaque composante par des régressions logistiques. Le paramètre retenu alors n'est plus $cov(x_j, y)$ mais $\beta = ln(OR)$. Les composantes T obtenues sont ensuite intégrées en lieu et place des prédicteurs X dans un modèle de régression logistique. La régression logistique est donc utilisée aux deux étages de la construction du modèle. La méthode s'étend facilement au modèle de Cox en utilisant ce modèle aux deux étages du processus. On utilise donc le logarithme du risque relatif $\beta = ln(RR)$ comme mesure de la relation entre chaque x_j et Y. Dans cette publication, le modèle a été présenté sur un jeu de données complètes. Cependant la plupart du temps et surtout dans la situation qui nous intéresse, il faut pouvoir prendre en considération le problème des données manquantes. Nous proposons donc ici d'étendre le modèle de Bastien en combinant l'algorithme PLS-GLM avec la méthode NIPALS d'estimation des paramètres de manière à pouvoir l'appliquer à un jeu de données incomplètes.

Voyons d'abord la PLS-GLM selon Bastien.

Présentation déroulée de l'algorithme On cherche une série de composantes orthogonales. La première s'écrit :

$$t_1 = w_{11}x_1 + \cdots + w_{1p}x_p \tag{29}$$

où le coefficient w_{1j} s'obtient de la façon suivante :

$$w_{1j} = \frac{\beta_j}{\sqrt{\sum_{j=1}^{p} \beta_j}} \tag{30}$$

Le coefficient w_{1j} est égal au coefficient de la régression logistique simple entre le vecteur y et le vecteur x_j, standardisé par l'ensemble des coefficients w_{1j}. L'ensemble des coefficients w_{1j} est obtenu en réalisant de manière itérative toutes les régressions logistiques simples de y sur chacune des variables x_j, indépendamment les unes des autres. Une fois obtenue la première composante, on calcule la régression logistique suivante :

$$\Pr(Y = 1|x_j) = \frac{exp^{\alpha+\beta_h t_h}}{1 + exp^{\alpha+\beta_h t_h}}$$

On place dans un vecteur y_1 l'ensemble des résidus de ce modèle.

On calcule ensuite les résidus des régressions *linéaires* x_{1j} des variables x_j sur la composante t_1. La composante t_2 est une combinaison linéaire des résidus x_{1j} :

$$t_2 = w_{21}x_{11} + \cdots + w_{2p}x_{1p} \tag{31}$$

et :

$$w_{2j} = \frac{\beta_{x_{1j}}}{\sqrt{\sum_{j=1}^{p} \beta_{x_{1j}}}} \tag{32}$$

On calcule ensuite :

$$y = c_1 t_1 + c_2 t_2 + y_2 \tag{33}$$

On peut bien sûr réexprimer y en fonction des variables d'origine pour donner plus de sens au modèle. L'équation incluant uniquement la première composante, s'écrit alors :

$$\Pr(Y = 1|X) = \frac{exp^{\alpha + c_1 \times w_{11}x_1 + \cdots + c_1 \times w_{1p}x_p}}{1 + exp^{\alpha + c_1 \times w_{11}x_1 + \cdots + c_1 \times w_{1p}x_p}}$$

Lorsque l'on utilise les h premières composantes, on obtient l'équation suivante :

$$\Pr(Y = 1|X) = \frac{exp^{\alpha + c_1 \times w_{11}x_1 + \cdots + c_h \times w_{hp}x_{hp}}}{1 + exp^{\alpha + c_1 \times w_{11}x_1 + \cdots + c_h \times w_{hp}x_{hp}}}$$

Standardisation des variables Lorsque l'on utilise des variables continues, il faut préalablement les standardiser pour éviter que les variables ayant les plus grandes variances « n'écrasent » les autres variables. Le problème est similaire à celui de l'ACP où pour trouver les composantes on travaille soit sur la matrice de variance-covariance soit sur la matrice de corrélation selon que ces variables sont mesurées respectivement sur des échelles similaires ou non. Dans ce cas il faut réexprimer les composantes en fonction des variables d'origine pour garder au modèle sa cohérence et l'interprétation physique que l'on peut en faire.

Lorsque l'on travaille sur des variables qualitatives, il est inutile de standardiser les variables ce qui permet de conserver l'interprétation des variables directement en terme d'OR en prenant l'exponentiel des paramètres [122].

Extension au cas des données incomplètes S'il y a des données manquantes, le principe est le même que précédemment et les équations sont modifiées de manière à n'utiliser que les données présentes. On calcule alors :

$$t_{1i} = \frac{\sum_{j:x_{ji}\ existe} w''_{1j} x_{ji}}{\sum_{\{j:x_{ji}\ existe\}} (w''_{1j})^2} \tag{34}$$

dans laquelle le coefficient w''_{1j} est défini par :

$$w''_{hj} = \frac{\beta_{hj}}{\sqrt{\sum_{j=1}^{p} (\beta_{hj})^2}} \tag{35}$$

Problèmes particuliers liés aux données qualitatives Le fait d'utiliser des données qualitatives peut mener à un certain nombre de problèmes. Rappelons que la situation qui nous intéresse ici est une situation où le nombre de sujets est inférieur au nombre de variables à analyser. Ceci suppose en général que les effectifs soient petits, de l'ordre de 10 à 50. Dans ce cas, il est possible d'observer des problèmes de séparation ou de quasi-séparation des données. Le terme de séparation des données désigne la situation où l'une des catégories de la variable Y est parfaitement prédite par l'une des modalités du prédicteur (ou une combinaison de modalités de plusieurs prédicteurs). Dans ce cas, l'estimation du maximum de vraisemblance ne peut pas être obtenue. Lors d'une quasi-séparation, l'estimation est délicate et donne des paramètres ayant des valeurs non pertinentes avec des intervalles de confiance trop larges pour être réalistes.

Pour contourner ce problème, on peut utiliser les méthodes dites exactes basées sur l'énumération complète des combinaisons de variables. Ce sont les méthodes utilisées dans les logiciels `StatXact` et `LogXact`. Ces méthodes sont très lourdes d'un point de vue calculatoire et peut-être encore hors de portée dans le cadre d'un modèle PLS. Une solution plus simple est celle proposée par Heinze [141, 142] qui implémente la technique de Firth [108]. Il s'agit d'une pénalisation de la vraisemblance qui permet de toujours obtenir une estimation finie du paramètre de régression même en cas de séparation des données. En cas de quasi-séparation, les estimations sont moins biaisées qu'en l'absence de pénalisation de la vraisemblance. Cette méthode permet donc en théorie de réaliser une régression logistique PLS quels que soient les nombres de sujets et de variables même lorsque celles-ci sont qualitatives. Cependant, nous n'avons pas exploré ici les propriétés de cette méthode et nous étudierons la PLS-GLM sur des effectifs plus grands correspondant aux données dont nous disposons.

Le programme N°24 permet de réaliser une régression logistique PLS suivant la méthode de Bastien modifiée.

7.3 Propriétés de la PLS-GLM

L'utilisation de la PLS-GLM dans le cas de l'allélotypage est conditionnée par la validité de son utilisation de façon générale. D'un point de vue théorique, les propriétés de la régression PLS-GLM sont satisfaisantes notamment avec une absence de biais et des estimations présentes même en présence de colinéarité. Nous avons néanmoins souhaité vérifier les propriétés de la régression PLS-GLM avec une série de simulation afin de mieux en évaluer les points forts et les limites. Plus précisément, l'objectif de ces simulations est de vérifier la stabilité des estimations des différents paramètres (t_i et w_j notamment) par rapport à une situation de référence. Le résultat attendu est que les estimations ne sont pas biaisées et que par ailleurs elles sont proches de leurs vraies valeurs connues en l'absence de données manquantes.

Nous avons déjà dit plusieurs fois que l'algorithme NIPALS fonctionne en présence de données manquantes. L'influence de la proportion des données manquantes a été étudiée et l'on observe qu'une proportion de valeurs manquantes inférieure à 30% n'a que peu d'effet sur l'estimation des paramètres [32]. Des proportions de 50% ont été données comme n'ayant pas d'influence sur les résultats d'une régression linéaire PLS utilisant NIPALS [175]. Ces résultats font référence à la proportion de données manquantes et en aucun cas au type de manquants. Les deux travaux cités ne semblent pas avoir étudié l'impact d'un mécanisme de type MCAR ou MAR sur le résultat des analyses. On peut s'attendre *a priori* à un effet limité de ces mécanismes. Nous avons tenté de vérifier cette hypothèse en comparant les résultats d'un modèle PLS-GLM sur un mécanisme MCAR par rapport à une situation de référence sans valeurs manquantes et sur un mécanisme MAR, là-aussi par rapport à une situation de référence avec un jeu de données complètes.

Une simulation se déroule de la façon suivante, en trois étapes principales :

(i) - dans le jeu de données portant sur le cancer du colon, la corrélation entre chaque couple de microsatellites est estimée en utilisant toutes les paires complètes de données par sujet afin d'obtenir une matrice de corrélation réaliste pour les simulations. Les coefficients de corrélation observés vont de -0,2 à 0,9 ;

(ii) - une matrice de données binaires de taille $n \times p$ est construite pour simuler un jeu de données de microsatellites dans lequel le statut AI ou normal des microsatellites homozygotes serait connu. Nous prenons ici, pour simplifier, $n = 100$ $p = 30$;

(iii) - une matrice de données incomplètes est ensuite créée en se basant sur la distribution

marginale des homozygotes observée dans le jeu de données original des cancers du colon. Deux situations sont ensuite distinguées :

(iii-a) - dans le premier cas, les manquants sont générés de manière complètement aléatoire pour approximer un mécanisme de type MCAR,

(iii-b) - dans le second cas, la probabilité de manquer pour le microsatellite MS_i, $i \in \{1, \ldots, 33\}$, est différente selon le résultat (le stade Astler-Coller), avec une probabilité d'être manquant $\Pr(MS_i = *)$ plus grande pour un sujet ayant un stade de Astler-Coller C ou D et plus faible quand le sujet est en stade A ou B. Nous avons retenu :

$$\Pr(MS_i = *|Stade = CD) = \Pr(HMZ + 0,1)$$

$$\Pr(MS_i = *|Stade = AB) = \Pr(HMZ - 0,1)$$

Il y a donc une différence de probabilité de 0,20 entre les deux niveaux de stade.

Cinq cents simulations sont faites dans chaque situation. Pour chacune d'entre elles, on calcule les vecteurs t_1, t_2 des coordonnées des sujets sur les deux premières composantes et les deux vecteurs w_1, w_2 de poids de chaque microsatellite sur ces mêmes premières composantes. Les vecteurs sont estimés d'abord sur le jeu de données complètes puis sur le jeu de données avec données manquantes. Si les estimations ne sont pas biaisées, la différence moyenne Δ entre les deux estimations est nulle et il doit y avoir une forte corrélation et une forte reproductibilité entre les deux vecteurs. Dans une situation idéale où les manquants n'auraient aucune influence, les valeurs de ces trois indices seraient les suivantes :

$$\Delta = E(\mathbf{v}_{C_i} - \mathbf{v}_{M_i}) = 0$$

$$\rho = cor(\mathbf{v}_{C_i}, \mathbf{v}_{M_i}) = 1$$

$$\rho_{icc} = icc(\mathbf{v}_{C_i}, \mathbf{v}_{M_i}) = 1$$

Dans ces formules, C représente les données complètes et M représente les données incomplètes, qu'elles soient de type MCAR ou de type MAR ; avec $i \in \{1, 2\}$, les deux premières composantes et $\mathbf{v}$ étant soit $\mathbf{w}$ soit $\mathbf{t}$. Le coefficient de corrélation intraclasse (ICC) est calculé comme le rapport entre la variance inter-situation (données complètes et données incomplètes) et la variance totale.

Dans une seconde série de simulations, l'une des 30 covariables de la matrice $\mathbf{X}$ fabriquée comme ci-dessus est dupliquée et remplace 10 colonnes de cette même matrice. On obtient donc une matrice avec une colinéarité totale entre 10 des variables. Cette colinéarité est légèrement diminuée par l'ajout des données manquantes mais elle reste forte. Les structures MCAR et MAR sont ensuite générées comme précédemment. Cette modification dans les données permet d'évaluer l'impact de la colinéarité sur les résultats.

Les problèmes de sélection de variables, de validation des composantes et de façon plus général, les problèmes de modélisation ne sont pas abordés ici, le but étant surtout de montrer que la PLS-GLM permet d'apporter une réponse satisfaisante à la question posée compte tenu des contraintes imposées au modèle. La modélisation pose des problèmes particuliers mais qui ne sont pas liés spécifiquement aux données d'allélotypage.

Les calculs sont faits avec R, en utilisant le package `bindata` pour générer les matrices de données binaires et le package `psy` pour calculer les coefficients de corrélations intraclasses ρ_{icc}.

7.4 Résultats des simulations

Les simulations montrent que les estimations des vecteurs de paramètres $\mathbf{t}$ and $\mathbf{w}$ ont un biais très faible voire nul dans l'ensemble. Les corrélations entre les valeurs sur les données complètes et les valeurs sur les données incomplètes correspondantes sont bonnes pour les premières composantes. Les très faibles différences constatées entre les valeurs de ρ et de ρ_{icc} montrent que le biais est faible quelle que soit la valeur du coefficient. Les performances sont moins bonnes pour les deuxièmes composantes, sauf pour le biais qui reste très faible. On note en effet des valeurs plus faibles pour ρ et ρ_{icc}. En outre, les résultats sont similaires pour les deux situations MCAR et MAR (table N°41 et N°42).

	t_1	t_2	w_1	w_2
Δ	-0,053 [-1,027 ; 0,929]	-0,051 [-1,303 ; 1,390]	0,006 [-0,167 ; 0,205]	-0,005 [-0,356 ; 0,312]
ρ	0,714 [0,559 ; 0,827]	0,467 [0,017 ; 0,748]	0,834 [0,702 ; 0,925]	0,502 [0,132 ; 0,777]
ρ_{icc}	0,709 [0,554 ; 0,822]	0,440 [0,015 ; 0,717]	0,834 [0,702 ; 0,924]	0,501 [0,132 ; 0,776]

Tab. 41 – Performances de la PLS-GLM sur des structures de type MCAR. Les valeurs moyennes de Δ, ρ et ρ_{icc} sur 500 simulations sont données pour $\mathbf{t}$ et $\mathbf{w}$ sur les deux premières composantes avec les quintiles [0,025 - 0,975].

	t_1	t_2	w_1	w_2
Δ	-0,027 [-0,924 ; 0,921]	0,030 [-1,371 ; 1,485]	-0,0001 [-0,193 ; 0,190]	-0,010 [-0,360 ; 0,328]
ρ	0,715 [0,522 ; 0,825]	0,457 [-0,098 ; 0,763]	0,831 [0,654 ; 0,927]	0,467 [0,090 ; 0,784]
ρ_{icc}	0,710 [0,522 ; 0,820]	0,423 [-0,089 ; 0,734]	0,831 [0,654 ; 0,927]	0,466 [0,090 ; 0,782]

TAB. 42 – Performances de la PLS-GLM sur des structures de type MAR, Les valeurs moyennes de Δ, ρ et ρ_{icc} sur 500 simulations sont données pour **t** et **w** sur les deux premières composantes avec les quintiles [0,025 - 0,975].

Lorsque le degré de colinéarité est élevé les performances du modèle s'améliorent comme on peut s'y attendre, notamment pour les deuxièmes composantes (tableaux N°43 et 44). En effet, par définition, pour des variables colinéaires ou presque, les coefficients w doivent être très proches ce qui augmente la corrélation des valeurs de w entre les estimations sur les données complètes et les estimations sur les données incomplètes. Par exemple, dans la situation MCAR, les coefficients de corrélation de t_1 entre les données complètes et les données incomplètes sur la première composante sont plus élevés que pour la situation MAR avec une faible colinéarité (0,803 *vs.* 0,714). Sur la deuxième composante, les différences sont encore plus fortes : 0,644 *vs.* 0,467. Les conclusions sont similaires lorsque l'on considère les vecteurs t_i et w_i,

	t_1	t_2	w_1	w_2
Δ	0,007 [-0,760 ; 0,748]	0,017 [-1,104 ; 1,099]	-0,001 [-0,046 ; 0,039]	0,001 [-0,089 ; 0,101]
ρ	0,803 [0,525 ; 0,956]	0,644 [-0,504 ; 0,915]	0,811 [0,593 ; 0,923]	0,788 [-0,007 ; 0,978]
ρ_{icc}	0,775 [0,516 ; 0,927]	0,620 [-0,338 ; 0,882]	0,810 [0,592 ; 0,922]	0,786 [-0,007 ; 0,977]

TAB. 43 – Performances de la PLS-GLM sur des structures de type MCAR avec forte colinéarité. Les valeurs moyennes de Δ, ρ et ρ_{icc} sur 500 simulations sont données pour **t** et **w** sur les deux premières composantes avec les quintiles [0,025 - 0,975],

	t_1	t_2	w_1	w_2
Δ	0,035 [-0,757 ; 0,765]	0,031 [-0,954 ; 1,201]	0,002 [-0,044 ; 0,050]	-0,024 [-0,207 ; 0,103]
ρ	0,795 [0,510 ; 0,947]	0,568 [-0,574 ; 0,899]	0,804 [0,613 ; 0,923]	0,621 [-0,812 ; 0,983]
ρ_{icc}	0,765 [0,494 ; 0,919]	0,538 [-0,485 ; 0,865]	0,803 [0,611 ; 0,923]	0,620 [-0,811 ; 0,983]

TAB. 44 – Performances de la PLS-GLM sur des structures de type MAR avec forte colinéarité. Les valeurs moyennes de Δ, ρ et ρ_{icc} sur 500 simulations sont données pour **t** et **w** sur les deux premières composantes avec les quintiles [0,025 - 0,975].

Ces résultats montrent de bonnes performances pour la PLS-GLM lorsqu'il y a de 10 à 50% de valeurs manquantes à la fois pour un mécanisme de type MCAR et pour un mécanisme de type MAR. Par ailleurs, le biais est faible et il ne dépend pas de la proportion de valeurs manquantes du microsatellite considéré. Ceci est vrai surtout dans la situation MCAR. Dans la situation MAR, le biais est négligeable pour la première dimension. En revanche, la seconde dimension montre des particularités. Les valeurs de biais sont nettement réparties en deux groupes, avec certaines valeurs de biais très faibles et d'autres d'ampleur modérée, sans valeur intermédiaire. Les graphiques suivants montrent la valeur du biais pour les valeurs de w sur les deux premières composantes lorsque l'on analyse ces valeurs pour chaque microsatellite individuellement, d'une part pour le modèle MCAR et d'autre part pour le modèle MAR.

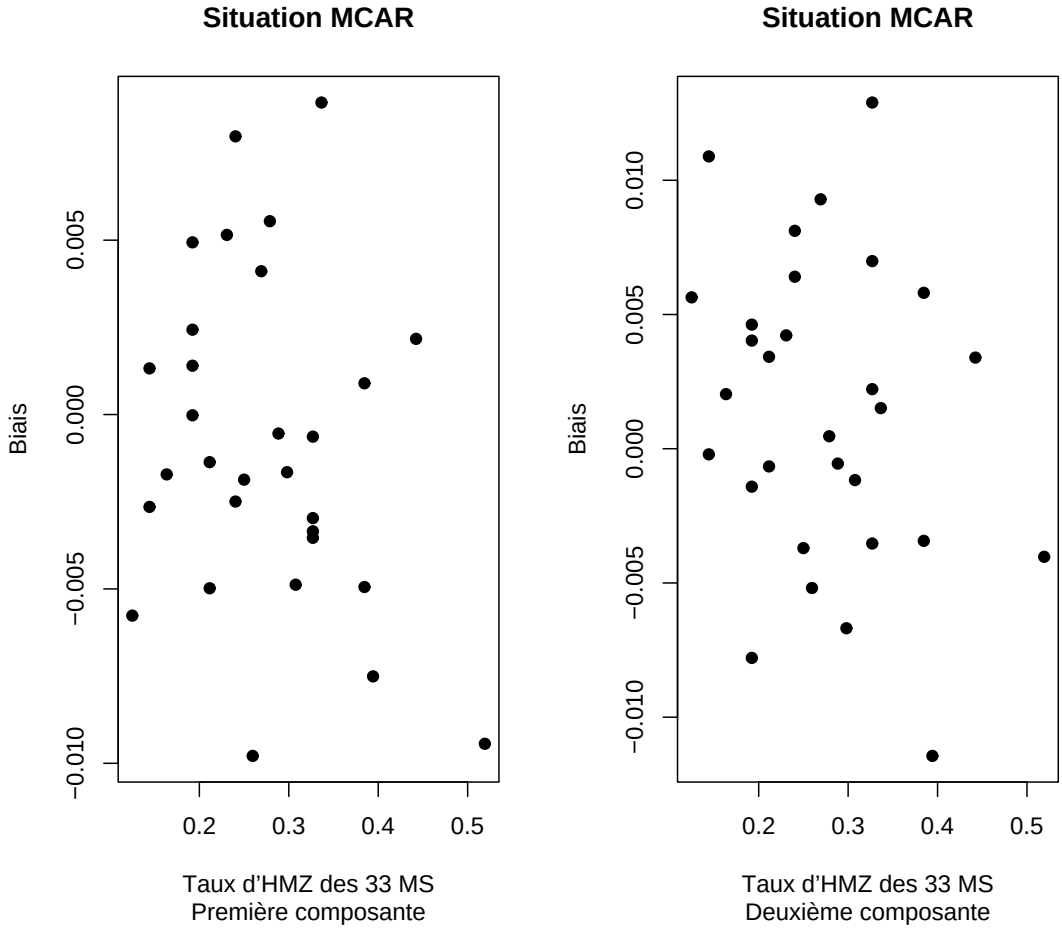

FIG. 8 – Valeurs du biais pour chaque microsatellite, situation MCAR, deux premières composantes.

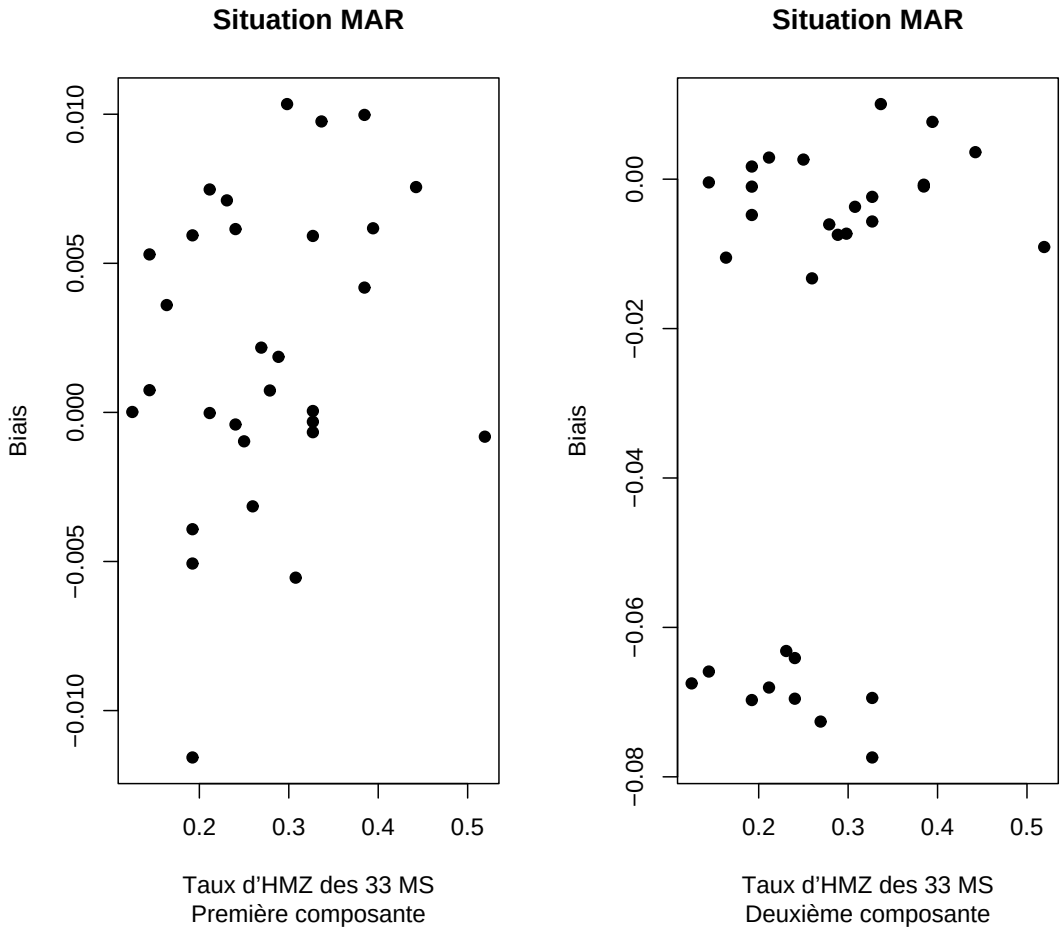

FIG. 9 – Valeurs du biais pour chaque microsatellite, situation MAR, deux premières composantes.

La régression PLS-GLM est un outil remarquable qui présente cependant un certain nombre de caractéristiques à connaître si l'on souhaite en faire un bon usage.

La PLS-GLM est idéale lorsque les variables incluses dans la matrice X sont homogènes, c'est-à-dire sont toutes de même type. La PLS-GLM est donc particulièrement indiquée pour les données d'allélotypage. L'interprétation en est relativement aisée.

En revanche, lorsque les variables incluses dans le modèle sont hétérogènes, l'interprétation des résultats, comme pour toute analyse descriptive peut être plus délicate. Les graphiques permettent de faciliter grandement cette interprétation.

Une solution possible à cette interprétation en cas d'hétérogénéité des variables consiste à combiner les variables homogènes et les autres variables dans un modèle classique de type : $Y = A + T'$. La matrice T' contient ici l'ensemble des composantes principales des

données homogènes et on peut alors ajuster un modèle de régression standard (non-PLS) sur ces composantes et les données hétérogènes.

A titre d'illustration, notons que les modèles PLS sont maintenant régulièrement utilisés dans le cadre de l'analyse des biopuces.

7.5 Application aux données d'allélotypage

La méthode développée ci-dessus permet donc d'analyser les données d'allélotypage puisqu'elle gère à la fois les problèmes de données manquantes, de dimension de matrice, de données qualitatives, en réalisant une régression dont les résultats peuvent s'exprimer de la même manière que pour des analyses descriptives multivariées. Nous appliquons ici cette méthode. Les résultats sont visualisés dans la figure 10.

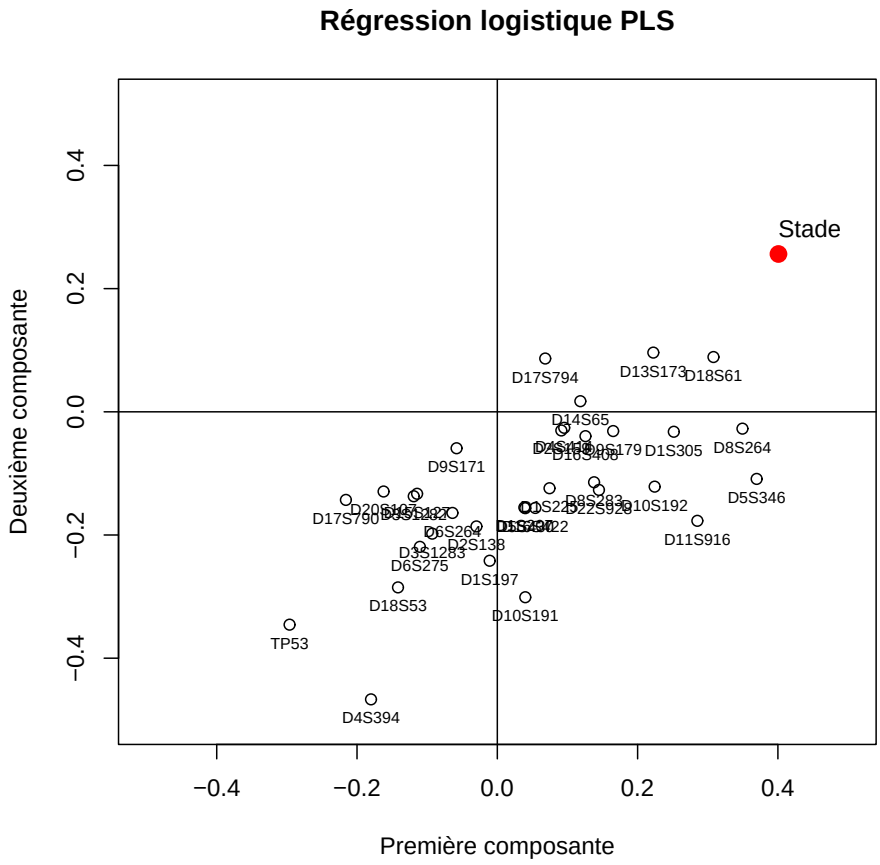

FIG. 10 – Régression logistique PLS par algorithme NIPALS. Deux premières composantes principales.

Le graphique permet de constater que sur le premier axe, le microsatellite D5S346 présente la corrélation la plus forte avec le Stade, ce qui est cohérent avec un certain nombre des résultats obtenus précédemment avec d'autres méthodes. D'autres microsatellites tels que D8S264, D18S61 et D13S173 semblent être associés au stade. Le TP53 quant à lui montre une relation inverse, comme cela avait été là aussi déjà constaté précédemment.

La méthode permet de traiter des matrices ou le nombre de variables est supérieure au nombre de sujets. Cette propriété peut être mise en évidence en utilisant la méthode pour inclure la totalité des interactions entre microsatellites. Pour 33 microsatellites, il y a 538 interactions de premier degré ce qui mène à une matrice de taille $104 \times (538 + 33)$. Les résultats sont donnés sur le graphique 11. Sur ce graphique, un point apparaît pour chaque microsatellite ainsi que pour chaque interaction possible. Seuls les effets principaux des microsatellites sont représentés, afin de ne pas alourdir le graphique. Cette présentation permet néanmoins de se rendre compte que la prise en compte de l'ensemble des interactions dans l'analyse modifie passablement les résultats, puisque par exemple le point TP53 qui précédemment était le plus éloigné du stade est désormais perdu au centre du nuage de points, dans une position montrant l'absence de relation entre ce microsatellite et le stade. Un autre exemple est celui du microsatellite D5S346 qui lui non plus ne partage plus de relation manifeste avec le stade, contrairement à ce qui a été observé sur l'analyse sans interaction. L'analyse et l'interprétation précise et complète de ces résultats appartiennent aux biochimistes. Nous préférons ici insister sur le fait que l'utilisation de la méthode PLS-GLM combinée avec l'algorithme NIPAL permet de mettre en évidence ce résultat, ce qu'aucune autre méthode ne permettait avant.

7.6 Codage des variables et interprétation du modèle

La formulation de la PLS logistique peut se faire de différentes façons selon le codage retenu pour les variables. Lorsque les variables sont quantitatives, le modèle fournit une valeur unique pour le coefficient de cette variable. La présentation graphique est alors similaire à celle d'une analyse en composante principale. Dans le cas de variables qualitatives, deux solutions sont possibles. Une première solution consiste à considérer une variable binaire comme une variable quantitative prenant des valeurs « 0 » et « 1 ». L'estimation du maximum de vraisemblance est alors la même que s'il s'agissait vraiment d'une variable quantitative. L'autre solution, plus en accord avec les habitudes des analyses descriptives multivariées consiste à utiliser autant de variables qualitatives que la variable a de moda-

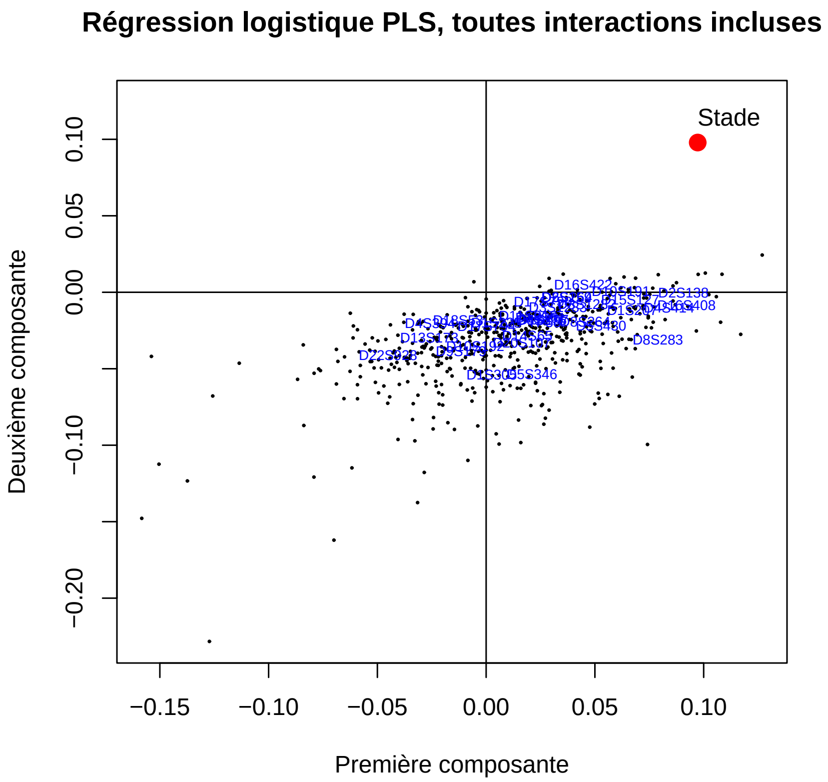

FIG. 11 – Régression logistique PLS par algorithme NIPALS avec inclusion de toutes les interactions possibles. Deux premières composantes principales.

lités, soit ici deux variables indicatrices. Cette modification aboutit à un doublement du nombre de variable ce qui ne pose aucun problème. Cette modification est non seulement pertinente mais obligatoire dans le cas de variables qualitatives multinomiales, mais dans le cas présent, elle est inutile puisque le modèle fournit deux estimations pour chaque microsatellite, ces deux estimations étant de même valeurs absolues mais de signe différent. Ceci découle évidemment des propriétés de la régression logistique. Sur un graphique, les points correspondant à chacune des deux modalités du microsatellite sont symétriques par rapport à l'origine et la totalité du graphique est donc symétrique. Il est donc inutile de placer toutes les modalités sur le graphique, ce qui permet de l'alléger sans en modifier l'interprétation. En revanche, ceci n'est valable que lorsque les coefficients sont exprimés sur l'échelle des logits qui est linéaire. Si l'on représente les données par les OR associant chaque variable au stade, les deux coefficients ne sont pas opposés mais inverses l'un de l'autre. Il faut alors présenter tous les points sur le graphique, points qui sont alors disposés en une double couronne plus ou moins complète de part et d'autre d'un cercle de rayon 1.

Cette méthode et ses principaux résultats ont été présentés en 2006 à l'ISCB [218] et à l'ADELF [217].

8 Discussion

8.1 Comparaison des méthodes

D'un point de vue statistique, les données d'allélotypage posent un certain nombre de problèmes en raison de l'existence pour certains sujets du caractère homozygote du ou des microsatellites étudiés et de l'incapacité de la PCR à classer ces microsatellites en normaux ou en déséquilibre allélique. Cet indéterminisme dans le résultat de la mesure aboutit à considérer chaque homozygote comme une donnée manquante. Le problème est rendu plus complexe par les objectifs des études d'allélotypage qui imposent de prendre en considération la structure des données de manière globale, c'est-à-dire en considérant non pas chaque microsatellite isolément mais tous les microsatellites simultanément. Cette analyse globale est incontournable si l'on veut comprendre comment les anomalies chromosomiques observées dans les pathologies cancéreuses se combinent pour aboutir à une tumeur maligne. Cet aspect dimensionnel des données est particulièrement prégnant si l'on souhaite tenir compte des interactions potentielles entre les microsatellites [153].

Pour apporter une solution à cette large problématique, nous avons présenté un certain

nombre de méthodes statistiques qui permettent de répondre à une plus ou moins grande partie du problème de façon plus ou moins complète à la question de départ.

Ce travail montre qu'il est tout à fait possible d'analyser des données malgré la présence de données manquantes. Ce problème étant très ancien, le nombre de méthodes utilisables est très grand et nous nous sommes plus particulièrement intéressés à certaines d'entre elles. À l'intérieur même de la méthode la plus justifiée (l'imputation multiple), le nombre de variantes possibles est aussi très grand. Néanmoins, ce qui peut de prime abord sembler être un avantage présente également un risque : si la souplesse d'utilisation de l'imputation multiple fait qu'il est très facile de choisir un modèle d'imputation, (*facile* sur le plan de la programmation), il reste néanmoins difficile de définir le bon modèle pour réaliser cette imputation multiple. Le nombre de variantes possibles est quasi infini et finalement, lorsque l'on est confronté à des données manquantes, l'imputation multiple ne fait que déplacer le problème. Si l'inférence sous un modèle est correctement réalisée par l'imputation multiple, la validité du modèle est invérifiable [283, 284, 290] et par conséquent la validité de l'inférence est elle-même invérifiable. L'intérêt et la validité de l'imputation multiple reposent donc sur la confiance que l'on peut apporter au modèle utilisé dans l'imputation multiple. Toute méthode de prise en compte des données manquantes suppose un certain nombre d'hypothèses qui pour la plupart ne sont pas vérifiables et l'imputation multiple n'échappe pas à cette difficulté. Il faut alors bien comprendre que même si l'imputation multiple sait prendre en compte la variabilité supplémentaire liée au manquant, elle ne fournit pas un résultat définitif, mais plutôt une aide utile à la décision lors de l'exploration d'un jeu de données.

Une bonne pratique statistique consiste alors à réaliser l'analyse souhaitée sur le jeu de données en l'état, avec les données manquantes, puis à réaliser non pas une mais *plusieurs* analyses avec imputation multiple afin d'évaluer la sensibilité des résultats aux hypothèses faites et donc la robustesse de ces résultats [174, 288].

Dans ce contexte d'évaluation de plusieurs modèles concurrents, l'inférence bayésienne semble beaucoup plus judicieuse que l'inférence fréquentiste en raison de l'absence de p-valeurs, lesquelles amènent souvent à des conclusions trompeuses, voire contradictoires lors de la comparaison de plusieurs modèles dans une analyse de sensibilité. En effet, le plus important dans ces analyses de sensibilité est d'estimer la variabilité des estimateurs, ce qui est beaucoup plus facile à réaliser en bayésien qu'en fréquentiste.

A l'intérieur des méthodes bayésiennes (rappelons que l'imputation multiple est fon-

damentalement une méthode bayésienne), un grand nombre de variantes est possible : la prise en compte des données manquantes peut se faire soit directement dans les données, ce qui est la plus mauvaise solution, quelle que soit la variante ou bien elle peut se faire en modifiant de façon adéquate les lois *a priori* du modèle. Là encore, différentes solutions sont possibles et il faut confronter la possibilité de spécifier une loi *a priori* directement pour les manquants et la possibilité d'adapter la loi *a priori* du paramètre d'intérêt de manière à refléter indirectement la répartition des données manquantes. A un niveau encore plus fin, pour chacune de ces variantes il est possible de spécifier de différentes manières ces lois *a priori*. La comparaison de toutes ces méthodes n'est pas forcément intéressante car elle suppose la fabrication d'un certain nombre de situations standards définissant différents mécanismes de manquants afin d'étalonner ces différentes méthodes. Les résultats aboutiraient certainement à montrer la supériorité de certaines méthodes sur d'autres dans certains cas, la supériorité n'étant sans doute pas constante sur l'ensemble des situations. Cela ne permettraient pas de conclure de manière générale et surtout ne permettraient pas de faire une utilisation systématique de telle ou telle méthode face à telle ou telle situation. En effet, dans la réalité, le vrai mécanisme des manquants est toujours inconnu et le choix d'une méthode dépend crucialement de ce vrai mécanisme. En pratique, il est certainement plus utile de réfléchir au cas par cas et de choisir une méthode pour une situation donnée.

Malgré la difficulté de définir la méthode qui puisse avoir un caractère générique, deux méthodes sont très certainement d'un plus grand intérêt que les autres. La première est celle qui consiste à combiner la proportion d'AI estimée sur les hétérozygotes avec la proportion générée par une loi Beta, le mélange des deux distributions de proportions formant la distribution finale (méthode « p_m »). Cette technique a l'avantage d'être particulièrement souple, car elle peut être déclinée en autant de variantes que nécessaire. Il est par exemple possible de définir une proportion d'AI pour tout sous-groupe d'homozygote voulu. Si par exemple on dispose d'information sur un sous-groupe G_1 de sujets homozygotes quant à la probabilité d'être AI et que l'on dispose d'autres informations pour un autre sous-groupe G_2, il est très simple de réaliser un mélange de trois distributions pour estimer la probabilité d'être en AI sur l'ensemble des trois groupes de sujets. Le modèle peut être varié à loisir en fonction des situations. La seconde méthode, plus élégante mais moins souple consiste à intégrer les manquants dans la distribution *a priori* de la proportion à estimer en faisant varier les paramètres α et β de cette loi.

L'imputation multiple, dans son principe repose sur l'ajustement d'un modèle sur l'en-

semble des données, complètes et incomplètes, de manière à réaliser une imputation des valeurs manquantes. Cela suppose la présence d'une ou de plusieurs variables complètes ou quasiment complètes de manière à pouvoir prédire raisonnablement les valeurs manquantes. Dans le cas des données d'allélotypage, cette méthode semble peu pertinente dans la mesure où les variables disponibles sont nettement moins nombreuses que les variables incomplètes, ce qui amène à une « surexploitation » des variables complètes, utilisées plusieurs fois. Même si le principe de l'imputation multiple ne l'interdit pas, cette méthode devrait être explorée avant de pouvoir être utilisée.

Quelle que soit la méthode utilisée, il faut se rappeler ici que le problème majeur des données manquantes n'est pas dans le fait que la donnée soit manquante mais plus exactement que l'on ne sache pas de quelle manière elle est manquante, ce qui est évidemment tautologique mais qui constitue le noeud du problème. Rappelons ici une remarque de P. Armitage [21] :

> However, such discussions seem relatively pedantic in view of the fact that missing data are more likely to be « non-ignorable » ; that is, the missingness is likely to be related to the hypothetical missing reading itself. An obvious example is the failure of a patient to appear at a clinic visit because of an exacerbation of illness which would have been likely to affect the clinical observations and test results. At first sight this seems an insoluble problem. How can we estimate the extent to which missingness depends on the intended observation, when the latter is unknown ? Diggle,Kenward and others [96, 172] have shown that some progress can be made by modelling the missingness mechanism and the distributional form of the response. However, as Kenward [172] emphasizes, the result may depend crucially on features of the model that may be impossible to verify from the data under analysis. The problem seems to call for sensitivity analysis,to assess the robustness of conclusions to a range of plausible assumptions.

Il faut enfin garder à l'esprit que l'utilisation de l'imputation multiple ne saurait se substituer à un jeu de données complet [89]. La nécessité de réaliser des études de sensibilité a été montrée dans la partie résultat où l'on voit bien que selon la distribution *a priori* utilisée pour les données, on obtient des résultats notablement différents à la sortie.

Dans le même esprit, Schafer souligne la chose suivante [290] :

> When data are missing for reasons beyond the investigator's control, one can never be certain whether MAR holds. The MAR hypothesis in such datasets cannot be formally tested unless the missing values, or at least a sample of them, are available from an external source. When such an external source is unavailable, deciding whether or not MAR is plausible will necessarily involve some guesswork, and will require careful consideration of conditions specific to the problem at hand.

Notons également que dans ce travail nous ne nous sommes intéressés qu'aux situations où les covariables sont manquantes. Ni les analyses ni la bibliographie ne traitent des méthodes adaptées au cas où le résultat Y est manquant (missing outcome). De même, nous n'avons pas détaillé les situations où l'on traite des données binaires répétées. Les méthodes relatives à ces deux situations sont extrêmement nombreuses et n'apportent que très peu d'éléments de réponse au problème traité [172, 185, 221, 222, 223, 339].

Un autre élément très important lors de l'analyse de données incomplètes est celui de l'exploration du mécanisme des manquants. Nous avons dit que les tests permettant de déterminer le type de mécanisme sont très rares. Une méthode bayésienne a été ici présentée qui permet d'approcher un résultat sur ce mécanisme. Il semblerait que les données d'allélotypage présentent des homozygotes dont la répartition soit indépendante du statut normal ou en déséquilibre allélique du microsatellite notamment parmi les homozygotes. C'est un résultat biologiquement cohérent, au moins avec le degré de finesse des mesures utilisées. Le mécanisme ne serait donc pas MNAR. La distinction restant à faire est donc celle entre un mécanisme MCAR et un mécanisme MAR. Si l'on considère un microsatellite de façon individuelle, sans chercher à établir de corrélation avec un résultat quelconque, le fait que le mécanisme ne soit pas de type MNAR implique que le mécanisme soit de type MCAR puisque pour une variable isolée, seuls ces deux mécanismes sont possibles. Dans le cas où l'on s'intéresse aux relations entre un microsatellite et une autre variable, par exemple le stade, la distinction doit être faite entre les possibilités. Le cas MCAR supposerait que le fait d'être en déséquilibre allélique serait indépendant du stade, ce qui est difficile à établir étant donnée la variété des taux d'AI parmi les microsatellites et parmi les stades. Une attitude plus prudente semble donc de devoir admettre que les données sont manquantes de types MAR, ce qui est une situation relativement confortable puisqu'elle permet d'appliquer la très grande majorité des méthodes pour données manquantes, dont celles présentées ici.

Il faut néanmoins garder à l'esprit les remarques de Armitage et Schafer rappelées plus haut.

Une autre approche proposée ici consiste à utiliser le Facteur de Bayes comme alternative au test d'hypothèse nulle [169]. Le FB est une mesure quantifiant le poids de la preuve apportée par les données en faveur d'un modèle ou d'une hypothèse. Différentes variantes ont été utilisées sur les données d'allélotypage apportant des résultats variables en fonction des hypothèses faites. En résumé, seul un petit nombre de microsatellites semble faire preuve d'un certain niveau d'association avec le stade. Par ailleurs, les conclusions sont assez variables selon les hypothèses faites, ce qui montre bien la fragilité des conclusions obtenues. Ces conclusions sont d'autant plus délicates à interpréter que la valeur du FB est connue pour dépendre parfois fortement des hypothèses faites dans la loi *a priori* des paramètres [169, 306]. Sur l'ensemble des microsatellites, il n'y a que le microsatellite D5S346 qui semble être associé au stade quelles que soient les hypothèses faites sur les manquants et quel que soit le modèle bayésien d'analyse retenu, c'est-à-dire, selon que l'on teste une hypothèse ponctuelle ou une hypothèse composite.

Les données d'allélotypage sont caractérisées par leur aspect multivarié : le nombre de microsatellites est important. La compréhension des voies d'altération du génome impliquées dans la cancérogénèse oblige à réaliser des modèles multivariées et prenant donc en compte la totalité des microsatellites. Plus précisément, ces modèles doivent être capables de fournir des résultats en incluant un nombre éventuellement grand de microsatellites même si le modèle final peut se révéler plus simple. Un premier modèle remplissant ces conditions peut être construit sur le principe des méta-analyses. Si on fait l'hypothèse que le rôle de chaque microsatellite dans la cancérogénèse est indépendant des autres microsatellites, une méta-analyse permet d'estimer l'effet propre de chaque microsatellite et de combiner l'ensemble des effets individuels pour obtenir une estimation globale. La méta-analyse repose soit sur des modèles à effets fixes soit sur des modèles à effets aléatoires. Nous avons retenu ici des modèles à effets aléatoires car les microsatellites contenus dans le jeu de données forment un sous-ensemble de tous les microsatellites pouvant théoriquement être inclus dans le modèle. Ils constituent donc un échantillon de l'ensemble des microsatellites. Cet échantillon peut être considéré en première approximation comme aléatoire et représentatif de l'ensemble des microsatellites. Ces hypothèses d'indépendance et de représentativité des microsatellites sont évidemment fausses mais permettent une première approche du phénomène à étudier malgré sa complexité.

Une méta-analyse réalisée directement sur les valeurs de chaque microsatellite montre un effet global pratiquement nul, ce qui était attendu car les relations positives et négatives entre les microsatellites et le stade sont de même ampleur et en nombre à peu près équivalent. L'effet global est donc faible ou nul. Mais la recherche des voies d'altération du génome implique plus la recherche de l'existence d'un effet et la taille de cet effet que le sens de cet effet, au moins dans un premier temps. Qu'un microsatellite soit lié positivement ou négativement au stade implique dans les deux cas qu'il soit lié au stade. Il faut donc envisager une méta-analyse montrant un effet et la taille de cet effet quel que soit le sens de cet effet. C'est pourquoi un second modèle de méta-analyse a été réalisé. Ce modèle inclut tous les microsatellites mais le codage des valeurs de chaque microsatellite est inversé lorsque la relation brute entre le microsatellite et le stade est négative. Ainsi, l'effet devient artificiellement positif et la méta-analyse permet d'estimer de façon globale l'ampleur de la relation entre chaque microsatellite et le stade. Les résultats obtenus montrent l'existence d'un effet faible mais quasi certain, quantifié par un OR global valant environ 1,5. Cette valeur confirme la difficulté à définir un modèle pour les voies d'altération du génome, les effets à prendre en considération étant certainement très nombreux et d'amplitude relativement faible. Les modèles doivent donc être particulièrement performants en raison d'un « bruit de fond » très important. Un des avantages des modèles de méta-analyse est la facilité avec laquelle ils peuvent être réalisés dans WinBUGS, même lorsqu'il y a des données manquantes, ce qui est ici le cas. Cette modélisation suppose tout de même que les données manquantes soient de type MAR ou MCAR ce qui est probablement le cas comme cela a été montré. On peut alors se passer de modéliser les données manquantes, avec toutes les réserves que cela appelle.

Les méthodes bayésiennes offrent donc un certain nombre de solutions pour traiter les données d'allélotypage, c'est-à-dire pour analyser les relations entre une série de microsatellites donnés et un résultat d'intérêt malgré le problème des homozygotes. Les analyses se font facilement en univarié ou avec des modèles de méta-analyses. Les analyses univariées sont pourtant d'un intérêt limité lorsque l'on considère le problème dans son ensemble. Ensuite, les méta-analyses, si elles prennent en considération l'ensemble des microsatellites, font des hypothèses très fortes sur l'indépendance et la représentativité des microsatellites. Pour aller plus loin et répondre plus complètement à l'objectif des analyses d'allélotypage, il faut donc souligner l'importance d'utiliser des méthodes multivariées, tenant compte de la vraie structure des données d'allélotypage avec leur haute dimensionalité, l'existence éven-

tuelle de colinéarités et une corrélation entre microsatellites qui ne peut être ignorée. Parmi les méthodes à envisager, de nombreuses solutions sont là aussi possibles.

Les analyses en cluster ont pour objectif de trouver des groupes dans les données et non pas d'attribuer des sujets ou des variables à un groupe préétabli [170]. L'utilisation de ces techniques dans le cadre présent des allélotypages vise donc à faire apparaître des regroupements de sujets et/ou de variables afin d'isoler des voies d'altération du génome ou de segment du génome menant à la tumorogénèse. On ne dispose pas ici de connaissances *a priori* fermement établies : on connaît partiellement le rôle de certains gènes tels que APC ou TP53 mais la plupart des microsatellites sont associés à des segments chromosomiques dont le rôle dans la cancérogénèse est seulement hypothétique. Le but de l'analyse est donc d'abord de confirmer ou d'infirmer l'implication d'un microsatellite dans cette cancérogénèse, et ensuite, dans la mesure où l'on a confirmé ce rôle, d'articuler de la façon la plus plausible ces différents éléments. Une des difficultés liée à l'utilisation des analyses en cluster est que les groupes définis par cette méthode ne correspondent pas forcément aux groupes attendus. La découverte d'une structure dans les microsatellites, aussi claire soit-elle n'implique aucunement que ces groupes soient stochastiquement liés à la cancérogénèse et encore moins qu'ils soient liés causalement. La structure obtenue n'est donc pas forcément corrélée au résultat auquel on s'intéresse, ce qui oblige l'investigateur à utiliser dans un second temps une méthode de type régression pour étudier justement l'existence d'un lien entre les données explicatives et le résultat. La plus grande prudence est donc de mise pour ces analyses en cluster. En revanche, les analyses en cluster permettent sans aucun doute de faciliter la compréhension d'un phénomène aussi complexe que les voies de l'altération génomique en proposant des schémas clairs, faciles à appréhender par l'esprit humain.

Les analyses factorielles apportent des solutions et des problèmes proches des analyses en cluster. Les Analyses Factorielles des Correspondances, simples ou multiples permettent aisément de décrire l'organisation des données et donc de faire des hypothèses sur les mécanismes biologiques sous-jacents au phénomène étudié mais leur validation (la structure existe-t-elle vraiment ?) est encore difficile. Des méthodes basées sur le bootstrap ont été récemment proposées mais elles restent encore confidentielles en raison d'une part de l'absence de logiciel disponible pour réaliser de telles études et d'autre part de la diversité des approches dans le cadre des analyses factorielles. Ensuite, pour ces analyses en bootstrap, un certain nombre de problèmes théoriques se posent encore [186]. Enfin, les analyses fac-

torielles ne permettent pas non plus d'analyser directement les relations avec une variable résultat. Certes, il est possible de réaliser des projections de variables supplémentaires (de la variable résultat par exemple) dans l'espace des variables explicatives, donnant des résultats souvent très suggestifs, mais il ne s'agit pas véritablement d'une modélisation des relations entre les deux types de variables. Cette approche, très utile, doit donc obligatoirement être complétée par des modèles de type régression pour répondre aux questions abordées ici. Enfin, ces méthodes (analyses en cluster ou analyses factorielles) ne prennent en considération aucunes connaissances antérieures et n'existent pas encore dans des versions bayésiennes.

Ces considérations ont donc amené à explorer l'intérêt de méthodes plus récentes, à savoir essentiellement la méthode PLS. Cette méthode inventée dans les années 50 dans le monde de la chémométrie a connu récemment un (petit) regain d'intérêt dans la littérature bio-médicale suite à l'apparition des biopuces [230, 250, 251, 252]. La méthode PLS permet en effet facilement de traiter des données de ce type, qui correspondent dans une bonne mesure aux données d'allélotypage. Néanmoins, pour pouvoir profiter de tous les avantages de la PLS, nous avons été amenés à en développer certains aspects notamment en combinant deux éléments de la méthode restés jusqu'ici distincts : la régression PLS linéaire généralisée - régression logistique plus précisément - et l'algorithme NIPALS. Cet algorithme permet de traiter une série de données malgré l'existence de données manquantes dans la matrice des variables explicatives. La combinaison de cet algorithme et de la régression logistique PLS (PLS-GLM) rend en effet possible l'analyse directe des données d'allélotypage. La méthode proposée a été explorée dans sa capacité à résister à différents mécanismes de manquants et pour des proportions de manquants variant dans des limites raisonnables (5 à 50 %). Les simulations ont permis de montrer que la méthode proposée permet en effet de traiter correctement, c'est-à-dire avec un biais faible ou nul, des données incomplètes, que le mécanisme des manquants soit de type MCAR ou de type MAR. La méthode peut donc être raisonnablement appliquée aux données d'allélotypage utilisées ici. Il a été montré que la régression logistique PLS peut traiter des matrices composées de données incomplètes quelles que soient les dimensions de la matrice (notamment quand $n < p$) et quel que soit le niveau de colinéarité. Un exemple extrême en est donné dans l'analyse PLS-GLM de la matrice des 33 microsatellites et de toutes les 528 interactions de premier ordre possibles entre les microsatellites. Lorsque les effectifs sont faibles, la méthode peut être prise en défaut du fait que les estimations utilisées sont basées sur la méthode du maximum de vraisemblance, méthode qui ne fournit pas de résultat en cas de séparation ou de quasi-

séparation des données. La solution n'a pas été explorée ici. On se reportera aux travaux d'Heinze [141, 142] sur la méthode de Firth de pénalisation de la vraisemblance. Un avantage de la méthode PLS est surtout de pouvoir envisager des modèles globaux de la cancérogénèse dans lesquels l'environnement et caractéristiques biologiques autres du patient pourraient être incluses, à l'image de certains modèles de la littérature [228]. Il faut également tenir compte du fait, très fréquemment oublié, que ces données sont observationnelles et que la quasi-totalité des modèles actuellement utilisée ne tient pas compte d'un certain nombre d'incertitude dans l'analyse des données que ce soit au niveau des données manquantes ou à d'autres niveaux [61, 132, 144].

Une limite importante de la méthode PLS, quelle que soit sa forme (PLS linéaire ou linéaire généralisé) tient justement à l'algorithme NIPALS utilisé. Cet algorithme a la propriété de pouvoir travailler sur des matrices incomplètes mais il est fondamental de comprendre que cet algorithme ne constitue en rien une modélisation des données manquantes. C'est en fait la seule méthode qui soit capable d'ignorer véritablement les données manquantes pour obtenir un résultat. Son fonctionnement revient en fait à utiliser un mélange de la méthode du cas complet et de la méthode du cas disponible puisqu'elle utilise tous les cas disponibles de chaque variable pour finalement travailler comme si les sujets étaient complets. Toutes les autres (bonnes) méthodes supposent une modélisation plus ou moins poussée des données manquantes, que ce soit par les méthodes basées sur le maximum de vraisemblance ou par les méthodes d'imputations simple et multiple. Ce fonctionnement très particulier rend la méthode très utile car utilisable virtuellement dans n'importe quelle situation de données incomplètes mais fait une hypothèse extrêmement forte sur le mécanisme des manquants. Dans le cas présent des données d'allélotypage, nous avons montré que le type de manquants et les propriétés de la méthode la rendaient utilisable mais ce ne sera pas forcément vrai pour toutes données d'allélotypage. Une meilleure solution serait certainement de combiner les méthodes PLS et une véritable modélisation des données manquantes. Parmi les solutions possibles, au regard de ce qui a été dit plus haut, il serait naturel de combiner les processus bayésiens d'inférence et les méthodes PLS. Le principe consisterait donc à attribuer une ou des distributions *a priori* aux données manquantes et de réaliser dans un cadre complètement bayésien l'estimation des paramètres. Si l'énoncé de la solution est simple, sa réalisation pratique l'est moins. Rappelons que l'estimation des paramètres se fait sur l'ensemble des lois de probabilité ce qui suppose que les lois des données manquantes soient également mises à jour, ce qui ne devrait pas être le cas dans

cette situations. Ceci impose la mise en place de « valves » dans les programmes, permettant à une partie de l'information d'être utilisée par le modèle mais sans qu'elle ne subissent de mise à jour. Cette programmation est relativement complexe à réaliser.

Si des modèles bayésiens existent théoriquement pour des analyses factorielles de type ACP [44, 45, 46], nous n'en avons trouvé aucune réalisation concrète, laquelle reste donc à faire. Le cas de la PLS fait partie des modèles à réaliser en écrivant le logiciel *ad hoc*. Ces modèles bayésiens de PLS-GLM sur données incomplètes formeraient en pratique le modèle ultime dans lequel l'inférence serait scientifiquement correcte et pour laquelle il serait possible d'évaluer l'influence des données manquantes, d'estimer la dimension effective des données et d'utiliser les modèles pour tirer des conclusions valides.

Conclusions

Les approches utilisées ici sont essentiellement de deux types différents : univariées avec un accent mis sur l'utilisation des méthodes bayésiennes et des méthodes multivariées avec le développement d'un modèle basé sur la régression PLS. Il faut ici souligner que les résultats obtenus avec ces méthodes sur les jeux de données utilisés ne doivent pas être considérés comme définitifs. Dans le cas des méthodes univariés, le calcul d'OR et l'utilisation de facteur de Bayes visaient surtout à montrer l'utilité et l'applicabilité des méthodes bayésiennes pour traiter des données incomplètes. Pour pouvoir comparer les résultats, des lois *a priori* non informatives ont été utilisées à chaque fois mais cet usage de la théorie bayésienne est très réducteur. Une analyse correcte de ces données supposerait l'incorporation de données antérieures chaque fois qu'elles sont disponibles. Par ailleurs, le calcul des facteurs de Bayes n'a pas pu incorporer l'incertitude sur les valeurs manquantes. Ces valeurs sont donc également à considérer avec une certaine prudence. Les résultats obtenus visaient plus à montrer la richesse des possibilités de traitements des données incomplètes dans le cas des données d'allélotypage qu'à analyser spécifiquement ces données, utilisées pour illustrer ces méthodes.

En ce qui concerne les méthodes multivariées utilisées et développées, la conclusion essentielle de ce travail est également liée plus à la possibilité d'utiliser la méthode décrite qu'à l'analyse spécifique des données d'allélotypage. En effet, ces données peuvent être analysées de manière très différentes selon la question que l'on se pose (recherche de sous groupes homogènes de microsatellites, aspect prédictif ou pronostique de ces données, etc).

Si la formulation précise de la réponse dépend donc de la question précise, le principe général de la PLS-GLM peut être adapté pour permettre d'apporter une réponse à chacune de ces questions. L'apport du modèle de PLS-GLM par l'algorithme NIPALS est essentiellement de pouvoir offrir une réponse dans le traitement des données d'allélotypage et c'est cet aspect là du problème qui a été ici présenté. Le résultat particulier obtenu lors de l'illustration de la méthode ne constitue pas une analyse complète des données et les relations entre tel microsatellite et le stade doivent pour l'instant être considérées avec une certaine réserve.

Comment se positionne la PLS-GLM par rapport aux méthodes existantes ? Les méthodes multivariées « traditionnelles » sont prises en défaut de différentes manières lorsqu'il s'agit de traiter des données d'allélotypage. Leurs limites tiennent essentiellement à la difficultés de gérer les données manquantes. Si les méthodes de classification permettent facilement de traiter des données incomplètes (car ces données ne perturbent pas les algorithmes), quelles que soient les dimensions des matrices de données, elles ne peuvent pas être utilisées pour réaliser des modèles de régression. Ainsi, dans l'article de Weber [337], des analyses en cluster on pu être effectuées sur les données d'allélotypage. En revanche, pour étendre les résultats de cette études, des modèles de type PLS-GLM devront être utilisés. Les méthodes de types analyse factorielle des correspondances (simple ou multiple) ne peuvent pas répondre complétement à la problématique en raison de la difficulté des ces méthodes à inclure des données incomplètes. Des solutions ont été proposées mais elles n'existent pas sous forme de logiciels ce qui imposent une programmation au cas par cas. De plus, les méthodes factorielles n'apportent qu'une réponse très indirecte à des problèmes de type régression. La encore, lorsque l'aspect prédictif du problème est le plus important, la méthode PLS-GLM montre sa supériorité puisqu'elle permet de réaliser des équivalents de modèle logistique et de modèle de Cox. La production par le modèle de paramètres de type odds-ratio ou risque relatif est un des avantages essentiels de la méthode. L'aptitude de la PLS-GLM à gérer simultanément les problèmes de dimensions de la matrice, les aspects descriptif et inférentiels d'un modèle de régression, les données manquantes et l'éventuelle collinéarité des données en fait une méthode irremplaçable qui n'a pas actuellement d'équivalent pour traiter les données d'allélotypage.

A Liste des abréviations et symboles

$\mathbf{X}$	matrice de données
n, p	dimensions de la matrice $\mathbf{X}$
OR	Odds-ratio
RR	Risque Relatif
MS	Microsatellite
N	Normal (pour un microsatellite)
AI	*Allelic Imbalance* : Déséquilibre allélique
HMZ	Homozygote
HTZ	hétérozygote
B	(loi) Binomiale
Be	(loi) Beta
Ber	(loi de) Bernoulli
D	(loi de) Dirichlet
FB	Facteur de Bayes
PCR	*Polymerase Chain Reaction*
DM	Données manquantes
IM	Imputation multiple
MCMC	Markov Chain Monte Carlo

Liste des tableaux

Table des figures

B Annexes

B.1 Critiques du test d'hypothèse nulle sur internet

Différents sites internet proposent une liste importante de citations et/ou d'article ayant trait au THN. Les principales adresses sont les suivantes :

`http://www.indiana.edu/~stigtsts/`

`http://www.npwrc.usgs.gov/perm/hypotest/hypotest.htm`

`http://www.npwrc.usgs.gov/resource/methods/statsig/litcite.htm`

`http://www.npwrc.usgs.gov/resource/1999/statsig/statsig.htm`

`http://www.cnr.colostate.edu/~anderson/null.html`

La dernière adresse en particulier donne une liste de références sur le test statistique et sa critique compilée par Nester.

B.2 Programmes WinBUGS pour les données manquantes

Programme N°1 Progamme Winbugs pour deux binomiales :

```
model{
      for (i in 1:n1){
            x[i] ~ dbern(theta1);
            }
      theta1 ~ dbeta(1,1);
      for (i in 1:n2){
            y[i]~ dbern(theta2);
            }
      theta2 ~ dbeta(1,1);
      propdiff <- theta1-theta2
      rr <- theta1/theta2
      or<- theta1*(1-theta2)/((1-theta1)*theta2)
}

# Données

list(x = c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1),n1 = 20,
     y = c(0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n2 = 25)
```

Dans le cas d'une loi multinomiale, le programme est le suivant :

Programme N°2

```
model{
      for (i in 1:n){
            x[i]   ~ dcat(p[1:4]);
            }
      p[1:4] ~ ddirch(alpha[]);  #  Densité \textit{a priori} pour theta
      or<-p[2]*p[3]/(p[1]*p[4])
}
# Données
list(n=45,alpha = c(1,1,1,1),x=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,
         3,3,3,3,3,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4,4))
```

Programme N°3

```
#########################################################
# dans cette fonction, les valeurs observées sont :     #
# a : nombre de succès de la première binomiale         #
# b : nombre d'echec de la     «         «              #
# c : nombre de succès de la seconde     «              #
# d : nombre d'échec de la     «         «              #
#                                                       #
# et les valeurs \textit{a priori} des lois beta sont   #
# les ao1,bo1,ao2,bo2                                   #
#########################################################


BF<-function(a,b,c,d,ao1,bo1,ao2,bo2){
n1<-a+b
n2<-c+d
a1<-ao1+a
a2<-ao2+c
b1<-bo1+b
b2<-bo2+d
p1<-beta(a1+a2-1,b1+b2-1)/beta(ao1+ao2-1,bo1+bo2-1)
p2<-beta(a1,b1)*beta(a2,b2)/(beta(ao1,bo1)*beta(ao2,bo2))
BF<-p1/p2
return(BF)
}
BF(1,40,9,33,1,1,1,1) #exemple
```

Dans ce programme, a et c représentent les nombres de « succès », b et d les nombres d'« échecs ». Les paramètres $ao1$, $bo1$, $ao2$, $bo2$ représentent les paramètres α et β respectivement de la première et de la seconde binomiale présentes dans le tableau.

Programme N°4

```
#########################################################################
# fonction FB                                                           #
# pour calculer la proba a posteriori de HN                             #
# à partir du FB et de la proba a priori de HN                          #
# version améliorée : on calcule Pr(p1<p2) par la formule d'Altham      #
#########################################################################

BFPe<-function(a,b,c,d,ao1,bo1,ao2,bo2,pi0){
pi1<-1-pi0
n1<-a+b
n2<-c+d
a1<-ao1+a
a2<-ao2+c
b1<-bo1+b
b2<-bo2+d
p1<-beta(a1+a2-1,b1+b2-1)/beta(ao1+ao2-1,bo1+bo2-1)
p2<-beta(a1,b1)*beta(a2,b2)/(beta(ao1,bo1)*beta(ao2,bo2))
BF<-p1/p2
p0<-1/(1+1/BF*pi1/pi0)
som<-0
for(k in (max(a2-a1,0)):(a2-1)){
som<-som+choose(a2+b2-1,k)*choose(a1+b1-1,a1+a2-1-k)/
        choose(a1+b1+a2+b2-2,a1+a2-1)
}
probsup<-som
probsup<-probsup*(1-p0)
probsup2<-1-p0-probsup
cat("\n","BF=",BF,"\n","P(H0|D)=",p0,"\n","Pr(p1<p2|D)",
                probsup,"\n","Pr(p1>p2|D)",probsup2,"\n")
}
BFPe(36,15,17,5,1,1,1,1,5/12)
```

Programme N°5

```
#####################################################################
# Calcul de la proba que OR<1                                        #
# à partir de la formule d'Altham                                   #
# pour une hypothèse composite, donc HN non ponctuelle              #
#####################################################################
PORS<-function(a,b,c,d,ao1,bo1,ao2,bo2){
n1<-a+b
n2<-c+d
a1<-ao1+a
a2<-ao2+c
b1<-bo1+b
b2<-bo2+d
som<-0
for(k in (max(ao2-ao1,0)):(ao2-1)){
som<-som+choose(ao2+bo2-1,k)*choose(ao1+bo1-1,ao1+ao2-1-k)/
        choose(ao1+bo1+ao2+bo2-2,ao1+ao2-1)
}
probinf1<-som
som<-0
for(k in (max(a2-a1,0)):(a2-1)){
som<-som+choose(a2+b2-1,k)*choose(a1+b1-1,a1+a2-1-k)/
        choose(a1+b1+a2+b2-2,a1+a2-1)
}
probinf2<-som
BF<-(probinf1/(1-probinf1))/(probinf2/(1-probinf2))
cat("&",BF,"&",probinf1,"&",probinf2,"&","\n")
}
```

Programme N°6

```
model{
# loi a priori des beta pour les proba de réponse et de résultat
  alpha0 <- 1; beta0<- 1;alpha1 <- 1; beta1<- 1;a <- 1;b <- 1;

for (i in 1:33){
     n[i]  <-  HTZ.N[i]+HTZ.AI[i]+HMZ[i];
# proba totale de non réponse
   p.HMZ[i] <- p[i]*(1-pi1[i])+(1-p[i])*(1-pi0[i]);# Pr(AI)*Pr(HMZ|AI)
                                                   #     + Pr(N)*Pr(HMZ|N)
   s[i] <- p[i]*(1-pi1[i])/p.HMZ[i];               # Pr(AI et HMZ|HMZ)
   HMZ.AI[i] ~ dbin(s[i],HMZ[i]);                  # nb d'HMZ qui sont AI
   HMZ.N[i] <- HMZ[i] - HMZ.AI[i];                 # nb d'HMZ qui sont N

# éléments pour densités \textit{a posteriori} des proba de réponse
   fc0.alpha[i] <- alpha0+HTZ.N[i];
   fc0.beta[i]  <- beta0+HMZ.N[i];
   fc1.alpha[i] <- alpha1+HTZ.AI[i]
   fc1.beta[i]  <- beta1+HMZ.AI[i]
   pi0[i] ~ dbeta(fc0.alpha[i],fc0.beta[i]);  # Pr(HTZ parmi les N)
   pi1[i] ~ dbeta(fc1.alpha[i],fc1.beta[i]);  # Pr(HTZ parmi les AI)
   dpi[i]<-pi1[i]-pi0[i];                     # différence entre pi0 et pi1

# éléments pour proba \textit{a posteriori} d'être AI
   fc.alpha[i] <- HMZ.AI[i]+HTZ.AI[i]+a;
   fc.beta[i] <- n[i]-HMZ.AI[i]-HTZ.AI[i]+b
   p[i] ~ dbeta(fc.alpha[i],fc.beta[i]);
   AItot[i] <- p[i]*n[i] }
               # Pr(AI) sur l'ensemble des HTZ et HMZ
               # Nb d'AI sur l'ensemble des (HTZ+HMZ)
}


#DATA
HTZ.N[] HTZ.AI[] HMZ[]
43 32 29
18 66 20
42 28 34
... ... ...
END
```

L'objectif est donc de déterminer la valeur de s.

Programme N°7

```
delucchi<-function(x,y){
tab<-table(x,y)
tabmqt<-matrix(table(is.na(x),y),ncol=2,byrow=F)[2,]
tabmqt1<-tabmqt[1]
tabmqt2<-tabmqt[2]
nbtab<-(tabmqt1+1)*(tabmqt2+1)
pvaleur<-c()
for(i in 0:tabmqt1){
    for(j in 0:tabmqt2){
        tadsim<-tab+rbind(tabmqt,c(0,0))+cbind(c(-i,i),c(-j,j))
        pvaleur<-c(pvaleur,fisher.test(tadsim)$p.value )
            }
        }
pvaleur<-sort(pvaleur)
plot(pvaleur,type="h")
nsig<-sum(pvaleur<=0.05)
names(nbtab)<-"Le nombre de tables possibles est de :"
names(nsig)<-"le nombre de tables significatives est de :"
minp<-min(pvaleur)
names(minp)<-"la valeur minimum de p est  :"
maxp<-max(pvaleur)
names(maxp)<-"la valeur maximum de p est  :"
pvaleur<-as.data.frame(pvaleur)
names(pvaleur)<-"Les valeurs de p sont :"
return(list(pvaleur,nbtab,nsig,minp,maxp))
#return(list(pvaleur,minp,maxp))
}
delucchi(D2S138,ASTLER)
```

Commentaires sur la fonction Cette fonction permet d'obtenir les calculs selon les propositions de Delucchi. Les lignes rendues muettes par des # sont là uniquement pour faire apparaître explicitement le nombre de tableaux possibles si on le souhaite. En pratique, le nombre de tableaux possibles est donné indirectement par la longueur du vecteur contenant les p-valeurs et apparaissant à l'écran.

Programme N°8

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);}
p ~ dbeta(1,1);
}
list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=100)
```

Programme N˚9

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);}
p ~ dbeta(1,1);
}

list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=150)
```

Programme N°10

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);}
p ~ dbeta(1,1);
}

list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=150)
```

Programme N°11

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);}
p ~ dbeta(1,1);
}

list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=150)
```

Programme N°12

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);}
p ~ dbeta(26,26);
}

list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=100)
```

Il faut noter que les paramètres de la loi *a priori* incluent à la fois l'*a priori* sur la proportion et l'opinion sur les manquants, d'où des valeurs de $\alpha = \beta = 26$ et non pas $\alpha = \beta = 25$.

Programme N°12bis

```
model{
alpham<-25;  # a priori des manquants
betam<-25;   # a priori des manquants

alphanm<-1; # a priori des non-manquants
betanm<-1;  # a priori des non-manquants

a<-alphanm+alpham;  # a priori global
b<-betanm+betam     # a priori global

for (i in 1:n)
{x[i] ~ dbern(p);}

p ~ dbeta(a,b);

}

list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=100)
```

Programme N°13

```
model {
for (i in 1:n)
{x[i] ~ dbern(p);
}
p ~ dbeta(1,1);

pt<-(pm*50+p*100)/150
}

list(pm=0.5,
x=c(0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
    1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
    1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),n=100)
```

Programme N°14

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);
}
p ~ dbeta(1,1);
pt<-(pm*50+p*100)/150
pm~dbeta(1,1);
}

list(pm=0.5,x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
n=100)
```

Programme N°15

```
model{
for (i in 1:n)
{x[i] ~ dbern(p);
}
p ~ dbeta(1,1);
pm~dbeta(25,25)
pt<-(pm*50+p*100)/150;
}
list(x=c(
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),
n=100)
```

Programme N°16

```
model{
    for (i in 1:N){
    Y[i] ~ dbern(p[i]); logit(p[i]) <- b[1]+ b[2]*X[i]
    }
    for (j in 1:2){b[j] ~ dnorm(0,0.0001)}
    OR<-exp(b[2])
    }
#Valeurs initiales
list(b= c(0, 0))
list(N=150,
X=c(
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),
Y=c(
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0))
```

Programme N°17

```
model{
    for (i in 1:N){
    Y[i] ~ dbern(p[i])
    logit(p[i]) <- b[1]+ b[2]*X[i]
    }
    for(k in 111:150){X[k]~dbern(0.5)}
    for (j in 1:2){b[j] ~ dnorm(0,0.0001)}
    OR<-exp(b[2])
    }

#Valeurs initiales
list(b= c(0, 0),
X=c(
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA, etc

# Données :
list(N=150,
X=c(
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1, etc
```

Programme N°18

```
model{
    for (i in 1:N){
    Y[i] ~ dbern(p[i])
    logit(p[i]) <- b[1]+ b[2]*X[i]
    }
    for(k in 111:150){X[k]~dbern(pm)}
    pm~dbeta(1,1)
    for (j in 1:2){b[j] ~ dnorm(0,0.0001)}
    OR<-exp(b[2])
    }
```

Programme N°19

```
model{
    for (i in 1:2){r[i] ~ dbin(p[i],n[i])
    p[i] ~ dbeta(1,1);
    }
    pm1~dbeta(1,1);
    pm2~dbeta(1,1);
    pt1<-(pm1*19+p[1]*33)/52;pt2<-(pm2*10+p[2]*35)/45;
    OR<-pt2*(1-pt1)/(pt1*(1-pt2))
    ORn<-p[2]*(1-p[1])/(p[1]*(1-p[2]))
    }
# Les données :
list(r=c(15,18),n=c(33,35))
```

Programme N°20

```
model{
  for (i in 1:N){
    stade[i] ~ dbern(p[i])
    logit(p[i]) <- beta[1]+D18S61imp[i]*beta[2]+D16S422imp[i]*beta[3]

# ci-dessous on impute les valeurs manquantes des deux microsatellites
# (1) pour le premier on donne une loi a priori indépendante de Y à D18S61
# (2) pour le second on utilise les valeurs du stade pour imputer la
# valeur du microsatellites (D16S422)

    D18S61imp[i] ~ dbern(p18[i])
    p18[i]~dbeta(50,50)
    D16S422imp[i] ~ dbern(p16[i])
    logit(p16[i]) <- alpha16 + beta.s16*stade[i]
        }

for (j in 1:3){
  beta[j] ~ dnorm(0,0.0001)
  }

    alpha16~dnorm(0,0.00001)
    beta.s16~dnorm(0,0.00001)
    alpha18~dnorm(0,0.00001)    # on peut ne pas l'utiliser puisque l'on
    beta.s18~dnorm(0,0.00001)   # impute à partir d'une loi a priori et pas de Y
    OR16<-exp(beta.s16)
    OR18<-exp(beta.s18)
}
# valeurs initiales puis générer les valeurs initiales pour les DM
list(beta = c(0, 0, 0),beta.s18=0,beta.s16=0,alpha16=0,alpha18=0)
# data
list(N=104)
D18S61imp[] D16S422imp[] stade[]
  1   1   0
  0   1   1
  1   NA   0
  0   NA   0
  0   1   0
  0   0   0
  1   0   0
... ...
END
```

Programme N°21

```
model{          for( i in 1 : N ) {
                r[i] ~ dbin(p[i],n[i])
                logit(p[i]) <- alpha + beta* x[i] + b[i]
                b[i] ~ dnorm(0.0,tau)
            }
            alpha ~ dnorm(0.0,1.0E-6)
            beta ~ dnorm(0.0,1.0E-6)
            tau ~ dgamma(0.001,0.001)
            sigma <- 1 / sqrt(tau)
            OR<-exp(beta)
        }


list(alpha = 0, beta = 0, tau = 1)

list(N=66,
r=c(15,21, 6,35,21,15,11,16,12,18,24,19,20,14,17,17,23,18,16,19,19,14,12,10,
14,24,16,17,22,16,21,14,14,31,27,15,13,26,11,27,17,20,18,21,10,26,19,14,20,
26,28,16,19,22,19,23,26,16,8,32,15,22,28,14,17,18),
n=c(32,43,18,66,42,28,25,33,27,37,43,30,47,27,40,30,51,36,31,32,40,24,28,22,
35,49,46,32,50,32,44,25,38,51,55,29,31,46,23,49,40,40,43,36,23,47,44,26,39,
50,59,32,44,38,38,41,49,27,25,59,36,43,63,22,35,33),
x=c(0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,
0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1))

# et variantes 21'
list(N=66,
r=c(21,15, 6,35,21,15,11,16,18,12,24,19,14,20,17,17,23,18,19,16,19,14,12,10,
14,24,16,17,22,16,21,14,14,31,27,15,13,26,27,11,20,17,18,21,26,10,14,19,26,
20,28,16,19,22,23,19,26,16, 8,32,15,22,14,28,17,18),
n=c(43,32,18,66,42,28,25,33,37,27,43,30,27,47,30,40,51,36,32,31,40,24,28,22,
35,49,46,32,50,32,44,25,38,51,55,29,31,46,49,23,40,40,43,36,47,23,26,44,50,
39,59,32,44,38,41,38,49,27,25,59,36,43,22,63,35,33),
x=c(0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,
0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0,1))
```

Programme N°22

```
model{
x[1] ~ dbin(p,x[2]);
p ~ dbeta(1,1);
pm~dbeta(1,1);
pt<-(pm*29+p*75)/104
}

list(x=c(32,75))
```

Les valeurs de p_t et de `x` sont à modifier pour chaque microsatellite.

Programme N°23

```
model{
for (i in 1:2){x[i] ~ dbin(p[i],n[i])

p[i] ~ dbeta(1,1);
}
pm1~dbeta(1,1);
pt1<-(pm1*50+p[1]*100)/150

pm2~dbeta(1,1);
pt2<-(pm1*50+p[2]*100)/150
OR<-pt1*(1-pt2)/(pt2*(1-pt1))
ORn<-p[1]*(1-p[2])/(p[2]*(1-p[1]))
}

list(x=c(60,40),n=c(100,100))
```

Programme N°24

```
PLSGLM<-function(y,X,nt){
#options(na.action=na.exclude)
normer<-function(coef){coef<-coef/sqrt(sum(coef*coef,na.rm=T))}
    X<-data.frame(X)
    Xres<-data.frame(X)
    nc <- ncol(X)
    nr <- nrow(X)
    dim1 <- dim(X)[1]
    dim2 <- dim(X)[2]
    th<-mat.or.vec(0,nr)
    cff <- mat.or.vec(0,nc)
    x <- list(T=matrix(1,nr,1),W=matrix(0,nc,nt),Wp=matrix(0,nc,nt),
              Tc=matrix(NA,nt,(nt+1)),Tp=matrix(NA,nt,(nt+1)))
    if(nc>dim2){stop("Nombre de composantes demandées trop grand !",
              «\n","Le nb de composantes demandées doit être
              inférieur ou égal au nombre de variable de X.")}
    row.names(x$T)<-row.names(X)
    row.names(x$W)<-names(X)

for(j in 1:nt){
    for(i in 1:nc){ modele <- glm(y~X[,i]+x$T-1,family=binomial)
                    cff[i] <- modele$coefficients[1]
                    x$Wp[i,j] <- summary(modele)$coefficients[1,4]
                    }
    cff<-normer(cff)
    if(j>1){for(i in 1:nc){Xres[,i]<-residuals(lm(X[,i]~x$T[,-1]-1,
                                          na.action=na.exclude))}
              }
    for(i in 1:nr){th[i]<-sum(Xres[i,]*cff,na.rm=T)}
    x$T<-cbind(x$T,th)
    x$W[,j]<-cff
    x$Tc[j,]<-replace(x$Tc[j,],1:(j+1),summary((glm(y~x$T-1,
              family=binomial)))$coefficients[,1])
    x$Tp[j,]<-replace(x$Tp[j,],1:(j+1),summary((glm(y~x$T-1,
              family=binomial)))$coefficients[,4])

    }
    return(x)
}
```

Références

[1] Collectif. Special issue on alternative methods of data interpretation. *Canadian Journal of Experimental Psychology* 2003 ;57(3) :139-264.

[2] Collectif. Research in the Schools. Special issue. Statistical Significance testing. *Research in the Schools.* 1998 ;5(2) :1-65.

[3] Abraham B, Merola G. Dimensionality reduction approach to multivariate prediction. Computational Statistics & Data Analysis 2005 ;48 :5-16.

[4] Afifi A, Elashof R. Missing observations in multivariate statistics I : Review of the literature. Journal of the American Statistical Association 1966 ;61 :595-604.

[5] Agresti A. On logit confidence intervals for the odds ratio with small samples. Biometrics 1999 ;55(2) :597-602. doi :10.1111/j.0006-341X.1999.00597.x.

[6] Agresti A, Min Y. Frequentist performance of bayesian confidence intervals for comparing proportions in 2×2 contingency tables. Biometrics 2005 ;61 :515-523. DOI : 10.1111/j.1541-0420.2005.031228.x.

[7] Agresti A, Hitchcock DB. Bayesian inference for categorical data analysis. Statistical Methods & Applications 2005 ;14 :297-330.

[8] Agresti A, Min Y. Unconditional small-sample confidence intervals for the odds ratio. Biostatistics 2002 ;3 :379-386.

[9] Agresti A. On logit confidence intervals for the odds-ratio with small samples. Biometrics 1999 ;55 :597-602.

[10] Agresti A. Exact inference for categorical data : recent advances and continuing controversies. Statist. Med. 2001 ;20 :2709-2722 (DOI : 10.1002/sim.738).

[11] Albert J. Teaching inference about proportions using bayes and discrete models. Journal of Statistics Education 1995 ;3(3). Accès en ligne le 20 02 06 : http ://www.amstat.org/publications/jse/v3n3/albert.html

[12] Albert JH. Bayesian testing and estimation of association in a two-way contingency table. JASA 1997 ;92(438) :685-693.

[13] Allison PD. Multiple imputation for missing data : a cautionary tale. Sociological methods and Research 2000 ;28 :301-309.

[14] Allison PD. Missing Data. Thousand Oaks, CA : Sage. 2001.

[15] Altham PME. Exact bayesian analysis of a 2×2 contingency table, and fisher's « exact » significance test. J of the Royal Statistical Society, Series B 1969, 31 :261-269.

[16] Altman DG. Statistics in medical journals : some recent trends. *Statistics in Medicine* 2000 ;19 :3275-3289.

[17] Matthews JNS, Altman DG. Statistics notes : interaction 2 : compare effect sizes not p. *BMJ* 1996 ;313 :808.

[18] Anderson DR, Burnham KP, Thompson WL. Null hypothesis testing : problems, prevalence, and an alternative. *Journal of Wildlife management* 2000 ;64(4) :912-923.

[19] Anderson DR, Link WA, Johnson DH, Burnham KP. Suggestions for presenting the results of data analyses. Journal of Wildlife Management 2001 ;65(3) :373-378.

[20] Antti H, Ebbels TMD, Keun HC, Bollard ME, Beckonert O, Lindon JC, Nicholson JK, Holmes E. Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects. Chemometrics and Intelligent Laboratory Systems 2004 ;73 :139-149.

[21] Armitage P. Theory and practice in medical statistics. Statistics in Medicine 2001 ;20 :2537-2548.

[22] Armitage P. Attitudes in clinical trials. Statistics in Medicine 1998 ;17,2675-2683.

[23] Baker S, Rosenberger W. DerSimonian R. Closed-form estimates for missing counts in two-way contingency tables. Statistics in Medicine 1992 ;11 :643-657.

[24] Baker SG, Ko CW, Graubard BI. A sensitivity analysis for nonrandomly missing categorical data arising from a national health disability survey. Biostatistics 2003 ;4 :41-56.

[25] Baker SG, Freedman LS. A simple method for analyzing data from a randomized trial with a missing binary outcome. BMC Medical Research Methodology 2003 ;3 :8. [online](http ://www.biomedcentral.com/1471-2288/3/8)

[26] Baker SG, Freedman LS. *Correction* : A simple method for analyzing data from a randomized trial with a missing binary outcome. BMC Medical Research Methodology 2004 ;4 :1.

[27] Barker M, Rayens W. Partial Least Squares for discrimination. Journal of Chemometrics 2003 ;17 :166-173.

[28] Barnard J, Meng XL. Applications of multiple imputation in medical studies : from AIDS to NHANES. Statistical Methods in Medical Research 1999 ;8 :7-36.

[29] Barnes SA, Lindborg SR, Seaman JW Jr. Multiple imputation techniques in small sample size. Statistics in Medicine 2006 ;25 :233-245.

[30] Barros AS, Rutledge DN. PLS Cluster : a novel technique for cluster analysis. Chemometrics and Intelligent Laboratory Systems 2004 ;70 :99-112.

[31] Philippe Bastien, Vincenzo Esposito Vinzi, Michel Tenenhaus. PLS generalised linear regression. Computational Statistics & Data Analysis 2005 ;48 :17-46.

[32] Bastien P. and Tenenhaus M. PLS regression and multiple imputations, Proceedings of the PLS'03 International Symposium, Vilares M. et al. editors. CISIA, Paris, 2003, pp. 497-498.

[33] Bayes T. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 1763 ;53 :370-418.

[34] Belin TR. Hu MY, Young AS, Grusky O. Using multiple imputation to incorporate cases with missing items in a mental health services study. Health Services & Outcomes Research Methodology 1 :1 (2000) : 7-22.

[35] Berger JO, Delampady M. Testing precise hypotheses. Statistical Science 1987 ;2 :317-352.

[36] Berger JO, Boukai B, Wang Y. Unified frequentist and bayesian testing of a precise hypothesis, with comments. Statistical Science 1997 ;12(3) :133-160.

[37] Bayarri MJ, Berger JO. The interplay of bayesian and frequentist analysis. Statistical Science 2004 ;19(1) :58-80. DOI 101214/088342304000000116.

[38] Berger JO. Could Fisher, Jeffreys and Neyman have agreed on testing ? *Statistical Science* 2003 ;18(1) :1-32.

[39] Berger VW. On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials* 2004 ;25 :613-619.

[40] Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. *Statistical Science* 2004,19(1) :175-187.

[41] Berry DA. Teaching elementary bayesian statistics with real applications in science. The American Statistician 1997 ;51(3) :241-246.

[42] Bertrand D, Qannari EM, Vigneau E. Latent root regression analysis : an alternative method to PLS. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :227-234.

[43] Best NG, Spiegelhalter DJ, Thomas A, Brayne CEG. Bayesian analysis of realistically complex models. Journal of the Royal Statistical Society Series A (Statistics in Society) 1996 ;159 :323-342.

[44] Bishop CM, Tipping ME. A hierarchical latent variable model for data visualization. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998 ;20 :281-293.

[45] Bishop CM. Latent variable models. Learning in Graphical Models, M. I. Jordan (Ed.), MIT Press 1999 ;371-403.

[46] Bishop CM. Variational principal components. Proceedings Ninth International Conference on Artificial Neural Networks ICANN99,1999 ;1 :509-514.

[47] Borenstein M. Hypothesis testing and effect size estimation in clinical trials. *Ann Allergy Asthma Immunol* 1997 ;78(1) :5-16.

[48] Brooks RJ, Cottenden AM, Fader MJ. Sample sizes for designed studies with correlated binary data. The Statistician 2003 ;52 :539-551.

[49] Brooks SP, Gelman A. Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 1998 ;7 :434-455.

[50] Brown TA. Génomes. Ed. Flammarion. Collection Médecine-Science. Paris 2004.

[51] Buck SF. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society Series B (Statistical Methodology) 1960 ;22 :302-306.

[52] Burnham AJ, Viveros R. Framework for latent variable multivariate regression. Journal of chemometrics 1996 ;10 :31-45.

[53] Burnham AJ, MacGregor JF, Viveros R. Latent variable multivariate regression modeling. Chemometrics and Intelligent Laboratory Systems 1999 ;48 :167-180.

[54] Bro R, Smilde AK, de Jong S. On the difference between low-rank and subspace approximation : improved model for multi-linear PLS regression. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :3-13.

[55] Burton A. Altman DG. Missing covariate data within cancer prognostic studies : a review of current reporting and proposed guidelines. British Journal of Cancer 2004 ;91 :4-8.

[56] Casella G, Moreno E. Objective bayesian analysis of contingency tables. Technical report, 2002 ; University of Florida, Department of Statistics.

[57] Chambers GK, MacAvoy ES. Microsatellites : consensus and controversy. Comparative Biochemistry and Physiology Part B 2000 :126 ;455-476.

[58] Chang HW, Lee SM, Goodman SN, Singer G, Cho SKR, Sokoll LJ, Montz FJ, Roden R, Zhang Z, Chan DW, Kurman RJ, Shih IM. Assessment of plasma dna levels, allelic imbalance, and CA 125 as diagnostic tests for cancer. Journal of the National Cancer Institute 2002 ;94 :1697-703.

[59] Carpenter J, Pocock S, Lamm CJ. Coping with missing data in clinical trials : a model based approach applied to asthma trials. Statistics in Medicine 2002 ;21 :1043-1066.

[60] Carpenter J, Kenward M, Evans S, White I. Letter to the editor : Last observation carry forward and last observation analysis by J. Shao and B. Zhong, Statistics in Medicine 2003, 22, 2429-2441. Statistics in Medicine 2004 ;23 :3241-3244.

[61] Carpenter JR, Kenward MK. Contribution to the discussion of Greenland. Multiple bias modelling for analysis of observational data. Journal of the Royal Statistical Society, Series A, 2005 ;168 :267-306.

[62] Carpenter J, Kenward M, Evans S, White I. Letter to the editor. Comment on : Last observation carry-forward and last observation analysis by J. Shao and B. Zhong, Statistics in Medicine 2004 ; 22 :2429-2441. Statistics in Medicine 2004 ;23 :3241-3244.

[63] Minini P, Chavance M. Sensitivity analysis of longitudinal binary data with non-monotone missing values. Biostatistics 2004 ;5 :531-544.

[64] Chavance M, Manfredi R. Modélisation d'observation incomplètes. Rev Epidémiol Santé Publique 2000 ;48 :389-400.

[65] Cheng C, Kimmel R, Neiman P, Zhaoa LP. Array rank order regression analysis for the detection of gene copy-number changes in human cancer. Genomics 2003 ;82 :122-129.

[66] Cherry S. Statistical tests in publications of The Wildlife Society. Wildlife Society Bulletin 1998 ;26(4) :947-953.

[67] Cheung YK. Exact two-sample inference with missing data. Biometrics 2005 ;61 :524-531. DOI : 10.1111/j.1541-0420.2005.00332.x.

[68] Choi L, Dominici F, Zeger S, Ouyang P. Estimating treatment efficacy over time : a logistic regression model for binary longitudinal outcomes. Statistics in Medicine 2005 ;24 :2789-2805 (DOI : 10.1002/sim.2147).

[69] Choi SC, Lu IL. Effect of non-random missing data mechanisms in clinical trials. Statistics in Medicine 1995 ;14 :2675-2684.

[70] Choisy M, Franck P, Cornuet JM. Estimating admixture proportions with microsatellites : comparison of methods based on simulated data. Molecular Ecology 2004 ;13 :955-968.

[71] Chong IG, Jun CH. Performance of some variable selection methods when multicollinearity is present. Chemometrics and Intelligent Laboratory Systems 2005 ;78 :103-112.

[72] Cifarelli DM, Regazzini E. De Finetti's contribution to probability and statistics. Statistical Science 1996 ;11(4) :253-282.

[73] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 1934 ; **26** :404-413.

[74] Clough HE, Clancy D, O'Neill PD, Robinson SE, French NP. Quantifying uncertainty associated with microbial count data : a bayesian approach. Biometrics 2005,61 ;610-616.

[75] Cohen J. The earth is round (p<0.05). *American Psychologist* 1994 ;49(12) :997-1003

[76] Congdon P. Bayesian Statistical Modelling. John Wiley & Sons. 2001.

[77] Congdon P. Bayesian Model for Categorical Data. John Wiley & Sons. 2001.

[78] Congdon P. Applied Bayesian Modelling. John Wiley & Sons. 2003.

[79] Copas JB, Li HG. Inference for non-random samples (with discussion). Journal of the Royal Statistical Society Series B (statistical methodology) 1997 ;59 :55-77.

[80] Cox DR. *Analysis of Binary Data*. Methuen : London, 1970.

[81] Cox DR, Hinkley DV. Theoretical Statistics. London : Chapman and Hall. 1974.

[82] Cronin KA, Freedman LS, Lieberman R, Weiss HL, Beenken SW, Kelloff GJ. Bayesian monitoring of phase II trials in cancer chemoprevention. *J Clin Epidemiol* 1999 ;52(1) :705-711.

[83] Raoul-Sam Daruwala RS, Rudra A, Ostrer H, Lucito R, Wigler M, Mishra B. A versatile statistical analysis algorithm to detect genome copy number variation. PNAS 2004 ;101 :16292-16297.

[84] Davis CE. Secondary endpoints can be validly analyzed, even if the primary endpoint does not provide clear statistical significance. *Controlled Clinical Trials* 1997 :18 :557-560.

[85] Davison AC. Hinkley DV. Bootstrap methods and their application. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press 1999.

[86] De Jong S. SIMPLS : An alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems 1993 ;18 :251-263.

[87] Delattre O, Olschwang S, Law DJ, Melot T, Remvikos Y, Salmon RJ, Sastre X, Validire P, Feinberg AP, Thomas G. Multiple genetic alterations in distal and proximal colorectal cancer. Lancet 1989 ;2 :353-5.

[88] De Leeuw E. Reducing missing data in surveys : an overview of methods. *Quality & Quantity* 2001 ;35 :147-160.

[89] De Leeuw J. Principal component analysis of binary data by iterated singular value decomposition. Computational Statistics & Data Analysis 2006 ;50 :21-39.

[90] Delucchi KL. Methods for the analysis of binary outcome results in the presence of missing data. *Journal of Consulting and Clinical Psychology* 1994 ; **62** :569-575.

[91] Demissie S, LaValley MP, Horton NJ, Glynn RJ, Cupples LA. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. Statistics in Medicine 2003 ;22 :545-557 (DOI : 10.1002/sim.1340).

[92] Dempster AP, Laird NM. Rubin DB. Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the Royal Statistical Society Series B (Statistical Methodology) 1977 ;39 :1-38.

[93] Denham MC, Whittaker JC. A Bayesian approach to disease gene location using allelic association. Biostatistics 2003 ;4 :399-409.

[94] Desai M, Emond MJ. A new mixture model approach to analyzing allelic-loss data using Bayes Factor. BMC Bioinformatics 2004 ;5 :182.

[95] Dexter F, Ledolter J. Bayesian prediction bounds and comparisons of operating room times even for procedures with few or no historic data. Anesthesiology 2005 ;103 :1259-67.

[96] Diggle PJ, Kenward MG. Informative dropout in longitudinal data analysis (with discussion). Applied Statistics 1994 ;43 :49-94.

[97] Dingstada GI, Westada F, Næsa T. Three case studies illustrating the properties of ordinary and partial least squares regression in different mixture models. Chemometrics and Intelligent Laboratory Systems 2004 ;71 :33-45.

[98] Dodge Y. Analysis of Experiments with Missing Data. New York : Wiley.1985.

[99] Duchesne P. Estimation of a proportion with survey data. Journal of Statistics Education 2003,11(3).[Online](www.amstat.org/publications/jse/v11n3/duchesne.pdf)

[100] Dunson DB. Commentary : Practical advantages of bayesian analysis of epidemiologic data. *American Journal of Epidemiology* 2001 ;153(12)1222-1226.

[101] Efron B. Why isn't everyone a bayesian. *The American Statistician* 1986 ;40 :1-11.

[102] Eilstein D, Uhry Z, Cherie-Challine L, Isnard H. Mortalité par cancer du poumon chez les femmes françaises. Analyse de tendance et projection à l'aide d'un modèle âge-cohorte bayésien, de 1975 à 2014. Revue d'Epidémiologie et de Santé Publique 2005 ;53(2) :167-181.

[103] Everitt. Cluster analysis. Arnold Ed. London 2001.

[104] Fay RE. When are inferences from multiple imputation valid ? Proceedings of the Survey Research Methods Section of the American Statistical Association, 1992 ;227-232.

[105] Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990 Jun 1 ;61 :759-67.

[106] Feingold J. Principes de génétique humaine. Hermann Éditions, Paris 1999.

[107] Feinstein AR. P-values and confidence intervals : two sides of the same unsatisfactory coin. *J Clin Epidemiol* 1998 ;51(4) :355-360.

[108] Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993 ;80 :27-38.

[109] Fisher RA. Statistical methods for research workers. Oliver & Boyd, Edinburgh, Scotland, 1951.

[110] Fisher RA. The design of experiments. New York, : Hafner Publishing Company, 1966.

[111] Fisher RA. Statistical methods and scientific induction. Journal of the Royal Statistical Society,B,1955 ;17 :69-78.

[112] Fitzmaurice G, Heath G, Clifford P. Logistic regression models for binary data panel data with attrition. Journal of the Royal Statistical Society Series A (Statistics in Society) 1996a ;159 :249-264.

[113] Fitzmaurice G, Laird N, Zahner G. Multivariate logistic models for incomplete binary response. Journal of the American Statistical Association 1996b ;91 :99-108.

[114] Fleiss JL. (letter to the editor). Confidence Intervals vs. Significance Tests : Quantitative Interpretation. *Am. J. Public Health* 1986 ;76 :587.

[115] Fleiss JL. Statistical methods for rates and proportions, John Wiley & Sohns.

[116] Fort G, Lambert-Lacroix S. Classification using partial least squares with penalized logistic regression. Bioinformatics 2005 ;21 :1104-1111.

[117] Freireich EJ, Gehan E, Frei, *et al.* The effect of 6-mercaptopurine on the duration of steroïd-induced remissions in acute leukemia : a model for evaluation of other potentially useful therapy. Blood 21 ;699-716.

[118] Frick RW. The appropriate use of the null hypothesis testing. *Psychological methods* 1996 ;1(4) :379-390.

[119] Gauchi JP, Chagnon P. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :171-193.

[120] Geladi P, Hadjiiski L, Hopke P. Multiple regression for environmental data : nonlinearities and prediction bias. Chemometrics and Intelligent Laboratory Systems 1999 ;47 :165-173.

[121] Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis, Chapman & Hall/CRC, Texts in statistical science, 1995.

[122] Gifi A. Non linear multivariate Analysis. Wiley & Sohns etc.

[123] Gigerenzer G. Mindless statistics. *The journal of socio-economics* 2004 ; 33 :587-606.

[124] Gilks WR, Richardson S. Spiegelhalter DJ. Markov Chain Monte-Carlo in practice. London : Chapman and Hall.1996

[125] Goldberg EK, Glendening JM, Karanjawala Z, Sridhar A, Walker GJ, Hayward NK, Rice AJ, Kurera D, Tebha Y, Fountain JW. Localization of multiple melanoma tumor-suppressor genes on chromosome 11 by use of homozygosity mapping-of-deletions analysis. Am. J. Hum. Genet. 2000 ;67 :417-431.

[126] Good P. *Permutation Tests, A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer, 2nd Ed., 2000.

[127] Goodman SN. A comment on replication, p-values and evidence. Statistics in Medicine 1992 ;11 :875-879

[128] Goodman SN. Of p-values and Bayes : a modest proposal. *Epidemiology* 2001 ; 12(3) :295-297.

[129] Goodman SN. Toward Evidence-based Medical Statistics. 1 : The P value fallacy. *Annals of internal medicine* 1999 ;130(12) :995-1004.

[130] Goodman SN. Toward Evidence-based Medical Statistics. 2 : The Bayes Factor. *Annals of internal medicine* 1999 ;130(12) :1005-1013.

[131] Goodman S, Sengul H. Letters to the Editor : Bayesian analysis of a single 2×2 table. Statistics in Medicine 1998 ;17 :2147-2148.

[132] Greenland, S. Multiple-bias modelling for analysis of observational data (with discussion). Journal of the Royal Statistical Society, Series A. 2005 ;168 :267-306.

[133] Guinot C, Latreille J, Tenenhaus M. PLS Path modelling and multiple table analysis. Application to the cosmetic habits of women in Ile-de-France. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :247-259.

[134] Gupta RC, Albanese RA, Penn JW, White TJ. Bayesian estimation of relative risk in biomedical research. Environmetrics 1997 ;8 :133-143.

[135] Hagg JD, Brasic GM, Shepel LA, Newton MA, Grubbs CJ, Lubet RA, Kelloff GJ, Gould MN. A comparative analysis of allelic imbalance events in chemically induced rat mammary, colon, and bladder cancer. Molecular Carcinogenesis 1999 ;24 :47-56.

[136] Hagen RL. In praise of the Null Hypothesis Statistical Test. *American Psychologist* 1997 ;52(1) :15-24.

[137] Hanafi M, Qannari EM. An alternative algorithm to the PLS B problem. Computational Statistics & Data Analysis 2005 ;48 :63-67.

[138] Hashemi L, Nandram B, Goldberg R. Bayesian analysis of a single 2×2 table. Statistics in Medicine 1997 ;16 :1311-1328.

[139] Hashemi L, Nandram B, Goldberg R. Author's reply : Bayesian analysis of a single 2×2 table. Statistics in Medicine 1998 ;17 :2148.

[140] Healy MJR. Is statistics a science ? J. R. Statist. Soc. 1978 ;141(3) :385-393.

[141] Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Statistics in Medicine 2002 ;21 :2409-2419.

[142] Heinze G, Ploner M. Fixing the nonconvergence bug in logistic regression with SPLUS and SAS. Computer Methods and Programs in Biomedicine 2003 ;71 : 181-187.

[143] Heitjan DF, Basu A. Distinguishing « missing at random » and « missing completely at random ». Journal of the American Statistical Association, 1996 ;50 :207-213.

[144] Heitjan DF, Rubin DB. Ignorability and coarse data. The Annals of Statistics 1991 ;19 :2244-2253.

[145] Helland IS. Some theoretical aspects of partial least squares regression. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :97-107.

[146] Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nature Review Genetics 2005 ;6 :95-108.

[147] Hollis S. A graphical sensitivity analysis for clinical trials with non-ignorable missing binary outcome. *Statistics in Medicine* 2002 ; **21** :3823-3834. DOI : 10.1002/sim.1276

[148] Hoque MO, Lee CCR, Cairns P, Schoenberg M, Sidransky D. Genome-wide genetic characterization of bladder cancer : a comparison of high-density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis. Cancer Research 2003 ;63,2216-2222.

[149] Höskuldsson A. Causal and path modelling. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :287-311.

[150] Howard JV. The 2×2 table : A discussion from a bayesian viewpoint. Statistical Science 1998 ;13(4) :351-367.

[151] Howard JV. The 2×2 table : a discussion from a bayesian viewpoint. *Statistical Science* 1998 ;13(4) :351-367.

[152] Hubbard R, Bayarri MJ. Confusion over Measures of evidence (p's) versus errors (α's) in classical statistical testing (with comments). *The American Statistician* 2003 ;57 :171-82.

[153] Hughes TR. Universal Epistastis analysis. Nature Genetics, 2005 ;37(5) :457.

[154] Hung HMJ, O'Neill RT, Bauer P, Köhne K. The behavior of the p-value when the alternative hypothesis is true. *Biometrics* 1997 ;53 :11-22.

[155] Hunt LA. Fitting a mixture model to three-mode three-way data with missing information. Journal of clasification 2001 ;18 :209-226.

[156] Hupé P, Stransky N, Thiery JP, Radvanyi F, Barillot E. Analysis of array CGH data : from signal ratio to gain and loss of DNA regions. Bioinformatics 2004 ;20 :3413-3422.

[157] Husson F, Pagès J. INDSCAL model : geometrical interpretation and methodology. Computational Statistics & Data Analysis 2006 ;50 :358-378.

[158] ICH E9 Expert Working Group. Statistical principles for clinical trials : ICH harmonised tripartite guideline. Statistics in Medicine 1999 ;18 :1905-1942.

[159] Janne PA, Li C, Zhao X, Girard L, Chen TH, Minna J, Christiani DC, Johnson BE, Meyerson M. High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. Oncogene 2004 ;23 :2716-2726.

[160] Jaynes ET. Probability Theory : The Logic of science. Cambridge University Press. 2003.

[161] Jeffreys HS. Theory of probability. Oxford : University Press. 1961.

[162] Johnson DH. The insignificance of statistical significance testing. *J. wildl. manage.* 1999 ;63(3) :763-772.

[163] Jollife IT. Principal Component Analysis. Springer Series in Statistics. Springer-Verlag. New-York 1986.

[164] Joseph L, Reinhold C. Statistical inference for continuous variables. AJR 2005 ;184 :1047-1056.

[165] Kadane JB. Subjective bayesian analysis for surveys with missing data. Journal of the Royal Statistical Society, Series D (The Statistician) 1993 ;42 :415-426.

[166] Kadane JB, Wolfson LJ. Experiences in elicitation. Statistician 1998 ;47 :3-19 .

[167] Kano M, Nishimura K, Ishikawa S, Tsutsumi S, Hirota K, Hirose M, Aburatani H. Expression imbalance map : a new visualization method for detection of mRNA expression imbalance regions. Physiol. Genomics 2003 ;13 :31-46.

[168] Kanter MH, Poole G, Garratty G. Misinterpretation and misapplication of p-values in antibody identification : the lack of value of a p-value. *Immunohematology* 1997 ;37(8) :816-822.

[169] Kass RE. Raftery AE. Bayes Factor. Journal of the American Statistical Association 1995 ;90 ;773-795.

[170] Kaufmann L, Rousseeuw PJ. Finding groups in data : an introduction to cluster analysis, John Wiley & Sons, Inc., NY, 1990.

[171] Kendall MG, Stuart A. The Advanced Theory of Statistics. Vol. 2 : Inference and Relationship. 2nd ed. London, 1967. Charles Griffin.

[172] Kenward MG. Selection models for repeated measurements with non-random dropout : an illustration of sensitivity. Statistics in Medicine 1998 ;17,2723-2732.

[173] Kenward MG, Molenberghs G. Likelihood based frequentist inference when data are missing at random. Statistical Science 1998 ;13 :236-247.

[174] Kenward MG, Molenberghs G, Goetghebeur E. Sensitivity analysis for incomplete categorical data. Statistical Modelling 2001 ;1 :31-48.

[175] Kettaneh N, Berglund A, Wold S. PCA and PLS with very large data sets. Computational Statistics & Data Analysis 2005 ;48 :69-85.

[176] Kieser M, Hauschke D. Assessment of clinical relevance by considering point estimates and associated confidence intervals. *Pharmaceutical Statistics* 2005 ;4 :101-107. DOI : 10.1002/pst.161.

[177] Killeen PR. An alternative to null-hypothesis significance tests. *Psychological Science* 2005 ;16(5) :345-353.

[178] King G, Honaker J, Joseph A, Scheve K. Analysing incomplete political science data : an alternative algorithm for multiple imputation. American Political Science Review, 2001 ;95 :49-69.

[179] Kraemer HC. Reconsidering the odds-ratio as a measure of 2×2 association in a population. Statistics in Medicine 2004 ;23 :257-270.

[180] Krueger J. Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist* 2001 ;56(1) :16-26. DOI : 10.1037//0003-066X.56.1.16.

[181] Kruskal WH. Tests of significance. *In* International Encyclopedia of Statistics. Kruskal WH and Tanur JM, eds. 1978. Free Press, New York : 944-958.

[182] Kuttatharmmakul S, Smeyers-Verbeke J, Massart DL, Coomans D, Noack S. The mean and standard deviation of data, some of which are below the detection limit : an introduction to maximum likelihood estimation. Trends in analytical chemistry 2000 ;19 :215-222.

[183] Lachin JM. Worst-rank score analysis with informatively missing observations in clinical trials. Controled Clinical Trials 1999 ;20 :408-422.

[184] Laia Y, Zhaob H. A statistical method to detect chromosomal regions with DNA copy number alterations using SNP-array-based CGH data. Computational Biology and Chemistry 2005 ;29 :47-54.

[185] Lavori PW, Dawson R, Shera D. A multiple imputation strategy for clinical trials with truncation of patient data. Statistics in Medicine 1995 ;14 :1913-1925.

[186] Lebart L, Morineau A, Piron M. Statistique exploratoire multidimensionnelle. Édition Dunod, Paris, 1997.

[187] Lee P. Bayesian Statistics : An Introduction. Troisième édition. Hodder & Stoughton 2004.

[188] Lee MLT, Schoenfeld D, Wang X, Penfornis A, Faustman D. Bayesian analysis of case control polygenic etiology studies with missing data. Biostatistics 2001 ;2(3) :309-322.

[189] Lehmann HP, Goodman SN. Bayesian communication : a clinically significant paradigm for electronic publication. J Am Med Inform Assoc. 2000 ;7 :254-266.

[190] Leung WC. Balancing statistical and clinical significance in evaluating treatment effects. Postgrad Med J 2001 ;77 :201-204.

[191] Liao TF. Estimating household structure in ancient China by using historical data : a latent class analysis of partially missing patterns. J. R. Statist. Soc. A 2004 ;167 :125-139.

[192] Lièvre A. Laurent-Puig P. Apport de la biologie moléculaire dans la recherche clinique en cancérologie : exemples des cancers digestifs. Rev Épidemiol Sante Publique 2005 ;53 :267-282.

[193] Lilford RJ, Braunholtz D. For debate : The statistical basis of piblic policy : a paradigm shift is overdue. BMJ 1996 ;313 :603-607.

[194] Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, Li C.dChipSNP : significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 2004 ;20 :1233-1240.

[195] Lindley DV. Why should clinicians care about Bayesian methods ? Discussion. Journal of Statistical Planning and Inference 2000 ;94 :59-60.

[196] Lindsey JK. Obtaining marginal estimates from conditional categorical repeated measurements models with missing data. Statistics in Medicine 2000 ;19 :801-809.

[197] Listgarten J, Damaraju S, Poulin B, Cook L, Dufour J, Driga A, Mackey J, Wishart D, Greiner R, Zanke B. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clinical Cancer Research 2004 ;10 :2725-2737.

[198] Listing J, Schlittgen R. Tests if dropouts are missed at random. Biometrical Journal 1998 ;40 :929-935.

[199] Little RJA. Pattern-mixture models for multivariate incomplete data. Journal of the American Statistical Association 1993a ;88 :125-134.

[200] Little RJA, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. Biometrics 1996 ;52 :98-111.

[201] Little RJA. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 1988 ;83 :1198-1202.

[202] Little RJA. Regression with missing X's : A Review. Journal of the American Statistical Association, 1992 ;87 :1227-1237.

[203] Little RJA. A class of pattern-mixture models for multivariate incomplete data. Biometrika 1994 ;81 :471-483.

[204] Liu BH. Statistical Genomics. Linkage, Mapping and QTL Analysis. CRC Press, 1998. p-35 et suivantes.

[205] Loftus GR. Psychology will be a much better science when we change the way we analyse data. *Current Directions in Psychological Science* 1996 ;5(6) :161-171.

[206] Lohmueller KE, Pearce CL, Piket M. *et.al.*. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nature genetics 2003 ;33 :177-182.

[207] Lydersen S, Laake P. Power comparison of two-sided exact tests for association in 2×2 contingency tables using standard, mid-p, and randomized test versions. Statistics in Medicine 2003 ;22 :3859-3871 (DOI : 10.1002/sim.1671)

[208] Lykken DT. Statistical significance in psychological research. *Psychological Bulletin* 1968 ;70 :(3) ;151-159.

[209] Marshall G, Warner B, MaWhinney S, Hammermeister K. Prospective prediction in the presence of missing data. Statistics in Medicine 2002 ;21 :561-570 (DOI : 10.1002/sim.966).

[210] Matthews RAJ. Why should clinicians care about Bayesian methods ? Journal of Statistical Planning and Inference 2000 ;94 :43-58.

[211] May WL, Johnson WD. Confidence intervals for differences in correlated binary proportions. Statistics in Medicine 1997 ;16 :2127-2136.

[212] Mehta C, Patel N. *StatXact 3, User manual*. Cytel Software Corporation, 1995. Cambridge MA, USA.

[213] Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Research 2000 ;10 :1126-1137

[214] Meng XL, Rubin DB. Performing likelihood ratio tests with multiply-imputed data sets. Biometrika 1992 ;79 :103-111.

[215] Meulepas E. A two-tailed p-value for fisher's exact test. Biometrical Journal 1998 ;40 :3-10.

[216] Meyer N, Meyer P, Oudet P. Analysis of incomplete categorical data by complete enumeration of missing patterns. 23^{rd} Annual Conference of the International Society for Clinical Biostatistics. 9-13 Septembre 2002, Dijon, France.

[217] Meyer N, Gaub MP, Oudet P, Meyer P. Intérêt de la régression logistique Partial Least Squares (PLS) pour des données d'allélotypage. Congrès Association Des Épidémiologistes de Langue Française (ADELF), 30-31 Août 2006, Dijon, France.

[218] Meyer N, Gaub MP, Oudet P, Meyer P. Partial Least Squares logistic regression for allelotyping data : interests, limits. 27^{rd} Annual Conference of the International Society for Clinical Biostatistics. 27-31 Août 2006, Genève, Suisse.

[219] Michiels B, Molenberghs G, Lipsitz S. Selection models and pattern-mixture models for incomplete categorical data with covariates. Biometrics 1999 ;55 :978-983.

[220] Molenberghs G, Kenward MG, Goetghebeur E. Sensitivity analysis for incomplete contingencey tables : the Slovenian plebiscite case. *Appl. Statist.* 2001 ; **50**,Part 1 :15–29.

[221] Molenberghs G, Williams PL, Lipsitz SR. Prediction of survival and opportunistic infections in HIV-infected patients : a comparison of imputation methods of incomplete CD4 counts. Statistics in Medicine 2002 ;21 :1387-1408 (DOI : 10.1002/sim.1118).

[222] Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ. Analyzing incomplete longitudinal clinical trial data. Biostatistics 2004 ;5 :445-464.

[223] Molenberghs G, Kenward MG, Goetghebeur E. Sensitivity analysis for incomplete contingency tables : The slovenian plebiscite case. Applied Statistics 2001 ;50(1) :15-29.

[224] Molenberghs G, Burzykowski T, Michiels B, Kenward MG. Analysis of incomplete public health data. Rev. Epidém. et Santé Publ. 1999a ;46 :499-514.

[225] Molenberghs G, Goetghebeur E, Lipsitz SR, Kenward MG. Non-random missingness in categorical data : strengths and limitations. American Statistician 1999b ;53 :110-118.

[226] Molenberghs G, Goetghebeur E, Lipsitz SR, Kenward MG, Lesaffre E, Michiels B. Missing data perspectives of the fluvoxamine data set : a review. Statistics in Medicine 1999c ;18 :2449-2464.

[227] Moran JL, Solomon PJ. A farewell to *p*-values ? *Critical Care and Resuscitation* 2004 ;6 :130-137.

[228] Morris JS, Wang N, Lupton JR, Chapkin RS, Turner ND, Hong M, Carroll RJ. A Bayesian analysis of colonic crypt structure and coordinated response to carcinogen exposure incorporating missing crypts. Biostatistics 2002 ;3 :529-546.

[229] Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H. Specific enzymatic amplification of DNA in vitro : the polymerase chain reaction. Cold Spring Harb Symp Quant Biol. 1986 :51 ;263-73.

[230] Musumarra G, Barresi V, Condorelli DF, Fortuna CG, Scire S. Potentialities of multivariate approaches in genome-based cancer research : identification of candidate genes for new diagnostics by PLS discriminant analysis. J. Chemometrics 2004 ;18 :125-132.

[231] Myers WR. Handling missing data in clinical trials : an overview. Drug Information Journal 2000 ;34 :525-533.

[232] Naidoo R, Chetty R. The applications of microsatellites in molecular pathology. Pathology Oncology Research 1998 ;4(4) :310-315.

[233] Nandram B. Choi JW. A Bayesian analysis of a proportion under non-ignorable nonresponse. Statistics in Medicine 2002 ;21 :1189-1212 (DOI : 10.1002/sim.1100).

[234] Nandram B, Liu N, Choi JW, Cox L. Bayesian non-response models for categorical data from small areas : an application to BMD and age. Statistics in Medicine 2005 ;24 :1047-1074. DOI : 10.1002/sim.1985.

[235] Nauta MJ, Weissing FJ. Constraints on allele size at microsatellite loci : implications for genetic differentiation. Genetics 1996 :143 ;1021-1032.

[236] Nelder JA. Multi-dimensional contingency table with one factor as a response. Statistician 1977 ;26,41-42.

[237] Nelder JA. The analysis of contingency tables with one factor as the response : round two. The Statistician 2000 ;49,pp. 383-388.

[238] Nelson PRC, Taylor PA, MacGregor JF. Missing data methods in PCA and PLS : Score calculations with incomplete observations. Chemometrics and Intelligent Laboratory Systems 1996 ;35 :45-65.

[239] Nelson PRC, MacGregor JF, Taylor PA. The impact of missing measurements on PCA and PLS prediction and monitoring applications. Chemometrics and Intelligent Laboratory Systems 2006 ;80 :1-12.

[240] Newcombe RG. Two-sided confidence intervals for the single proportion : comparison of seven methods. Statistics in Medicine 1998 ;17 :857-872

[241] Newcombe RG. Interval estimation for the difference between independent proportions : comparison of eleven methods. Statistics in Medicine 1998 ;17 :873-890

[242] Newcombe RG. A deficiency of the odds ratio as a measure of effect size. Statistics in Medicine *In press.*

[243] Newton MA, Gould MN, Reznikoff CA, Haag JD. On the statistical analysis of allelic-loss data. Statistics in Medicine 1998 ;17 :1425-1445.

[244] Newton MA, Lee Y. Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. Biometrics 2000 ;56 :1088-1097.

[245] Nester N. An applied statistician's creed. *Applied Statistics*, 1996 ;45(4) :401-410.

[246] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika 1928,20A :175-240.

[247] Neyman J, Pearson ES. On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. Biometrika 1928,20A :263-94.

[248] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A* 1933 ;231 :289-337.

[249] Ng IOL, Xiao L, Lam KY, Yuen PW, Ng M. Microsatellite alterations in squamous cell carcinoma of the head and neck : clustering of loss of heterozygosity in a distinct subset. Oral Oncology 2000 ;36 :484-490.

[250] Nguyen DV, Rocke DM. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics 2002 ;18 :1625-1632.

[251] Nguyen DV, Rocke DM. On partial least squares dimension reduction for microarray-based classification : a simulation study. Computational Statistics & Data Analysis 2004 ;46 :407-425.

[252] Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics 2002 ;18 :1216-1226.

[253] Nunnally J. The place of statistics in psychology. *Educational and Psychological Measurement* 1960 ;20(4) :641-650.

[254] O'Hagan A, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T. Uncertain judgements. Eliciting experts' probabilities. Wiley & Sons, Coll. Statistics in Practice. Chichester. 2006.

[255] O'Hagan A, Forster J. Kendall's advanced theory of statistics. Vol. 2B. 2nd ed. Arnold, London, 2004.

[256] O'Hagan A. Eliciting expert beliefs in substantial practical applications. The statistician 1988 ;47 :21-35.

[257] Ollivier L. Éléments de génétique quantitative. Masson Edition. Collection INRA Actualités scientifiques et agronomiques (volume 5), 1981.

[258] O'Neill RT. Secondary endpoints cannot be validly analyzed if the primary endpoint does not demonstrate clear statistical significance. *Controlled Clinical Trials* 1997 ;18 :550-556.

[259] Pages J, Tenenhaus M. Multiple factor analysis combined with PLS path modelling. Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgements. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :261-273.

[260] Panhard X, Dominique S, GAub MP, Ravery V, Grandchamp, Mentré F. Construction of a global score quantifying allelic imbalance among biallelic SIDP markers in bladder cancer. Statistics in Medicine 2003 ;22 :3771-3779.

[261] Park PJ, Tian L, Kohane IS. Linking gene expression data wih patient survival times using partial least squares. Bioinformatics 2002 ;18(S1) :S120-S127.

[262] Park T, Brown MB. Models for categorical data with nonignorable nonresponse. Journal of the American Statistical Association, 1994 ;89 :44-52.

[263] Parmar MK, Griffiths GO, Spiegelhalter DJ, Souhami RL, Altman DG, van der Scheuren E ; CHART steering committee. Monitoring of large randomized clinical trials : a new approach with bayesian methods. *Lancet* 2001 ;358 :375-81.

[264] Pasanisi A. Aide à la décision dans la gestion des parcs de compteurs d'eau potable. ENGREF, Ecole Nationale du Génie Rural, des Eaux et des Forêts. Thèse, 2004.

[265] Pham-Gia T, Turkkan N. Determination of exact sample sizes in the bayesian estimation of the difference of two porportions. The Statistician 2003 ;52(2) :131-150.

[266] Picard F, Robin S, Lavielle M, Vaisse C, Daudin JJ. A statistical approach for array CGH data analysis. BMC Bioinformatics 2005, 6 :27.

[267] Pinheiro JC, Bates D. Mixed-Effects Models in s and S-Plus. Series : Statistics and Computing . Springer 2002.

[268] Poole C. Low p-values or narrow confidence intervals : which are more durable ? *Epidemiology* 2001 ;12(3) :291-294.

[269] Popper K. Logik der Forschung 1934.

[270] Pritchard JK, Cox NJ. The allelic architecture of human disease genes : common disease-common variant... or not ? Human Molecular Genetics 2002,11(20) :2417-2423.

[271] Primdahl H, Wikman FP, Von der Maase H, Zhou X, Wolf H, Ørntoft TF. Allelic imbalances in human bladder cancer : genome-wide detection with high-density single-nucleotide polymorphism arrays. Journal of the National Cancer Institute 2002 ;94 :216-23.

[272] Proschan MA, Waclawiw MA. Practical guideliness for multiplicity adjustment in clinical trials. *Controlled Clinical Trials* 2000 :21 :527-539.

[273] Proschan MA, McMahon RP, Shih JH, Hunsberger SA, Geller NL, Knatterud G, Wittes J. Sensitivity analysis using an imputation method for missing binary outcome in clinical trials. *Journal of Statistical Planning and Inference* 2001 ; **96** :155–165.

[274] Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 2001 ;27 :85-95.

[275] Rahme E, Joseph L. Exact sample size determination for binomial experiments. J Stat Plan Infer 1998 ;66 :83-93.

[276] Reiczigel JO. Confidence intervals for the binomial parameter : some new considerations. Statistics in Medicine 2003 ;22 :611-621 (DOI : 10.1002/sim.1320).

[277] Roberts KA, Dixon-Woods M, Fitzpatrick R, Abrams KR, Jones DR. Factors affecting uptake of childhood immunisation : a Bayesian synthesis of qualitative and quantitative evidence. *Lancet* 2002 ;360 :1596-1599.

[278] Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 1995 ;89 :846-866.

[279] Robins JM. Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine 1997 ;16 :21-37.

[280] Robins JM, Gill RD. Non-response models for the analysis of non-monotone ignorable missing data. Statistics in Medicine 1997 ;16 :39-56.

[281] Rothman KJ. Writing for Epidemiology. *Epidemiology* 1998 ;9 :3.

[282] Royston P. Multiple imputation of missing values. The Stata Journal 2004 ;3 :227-241.

[283] Rubin DB. Inference and missing data. *Biometrika* 1980 ; **63** :581-592.

[284] Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. John Wiley & Sons : New York, 1987.

[285] Rubin D. Multiple imputation after 18 years. Journal of the American Statistical Association 1996 ;91 :473-490.

[286] Rubin DB. Inference and missing data. Biometrika, 1976 ;63 :581-592.

[287] Rubin DB. Multiple imputation for nonresponse in surveys. New York : Wiley. 1987.

[288] Saltelli A, Chan K, Scott EM. Sensitivity analysis. Chichester : Wiley. 2000.

[289] Saporta G. Probabilités, Analyse des données et statistique. Édition TECHNIP. 1990. Paris.

[290] Schafer JL. *Analysis of Incomplete Multivariate Data*. Chapman & Hall/CRC 2000.

[291] Schafer JL. Multiple imputation in multivariate problems when the imputation analysis models differ. Statistica Neerlandica 2003 ;57 :19-35.

[292] Schafer JL, Olsen MK. Multiple imputation for multivariate missing-data problems : A data analyst's perspective. Multivariate Behavioral Research 1998 ;33 :545-571.

[293] Schafer JL. Multiple imputation : a primer. Statistical methods in medical Research 1999 ;8 :3-15.

[294] Collins ML, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods 2001 ;6 :330-351.

[295] Schafer JL, Graham JW. Multiple imputation : our view of the state of the art. Psychological Methods 2002 ;7 :147-177.

[296] Sellke T, Bayarri MJ, Berger JO. Calibration of P-values for Testing Precise Null Hypotheses. *The American Statistician* 2001 ;(55) :62-71.

[297] A comment on replication, p-values and evidence, S.N.Goodman, Statistics in Medicine 1992 ;11 :875-879.

[298] Seneta E, Phipps MC. On the comparison of two observed frequencies. Biometrical Journal 2001 ;43 :23-43.

[299] Shadish WR, Hu X, Glaser RR, Kownacki R, Wong S. A method for exploring the effects of attrition in randomized experiments with dichotomous outcomes. *Psychological Methods* 1998 ;3 :3-22.

[300] Shao J, Zhong B. Last observation carry-forward and last observation analysis. Statistics in Medicine 2003 ;22 :2429-2441.

[301] Shao J, Zhong B. Author's reply. Statistics in Medicine 2004 ;23 :3241-3244.

[302] Shimodaira H. Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. Annals of Statistics 2004 ;32 :2616-2641.

[303] Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Systematic Biology 2002 ;51 :492-508.

[304] Shoukri MM, Chaudhary MA, Mohamed GH. Evaluating normal approximation confidence intervals for measures of 2 x 2 association with applications to twin data. Biometrical Journal 2003 ;45 :20-33.

[305] Sinharay S, Stern HS. Russell D. The use of multiple imputation for the analysis of missing data. Psychological Methods 2001 ;6 :317-329.

[306] Sinharay S, Stern HS. On the sensitivity of the Bayes Factor to the prior distributions. The American Statistician 2002 ;56 :196-201.

[307] Slebos RJC, Umbach DM, Sommer CA, Horner GA, Choi JY, Taylor JA. Analytical and statistical methods to evaluate microsatellite allelic imbalance in small amounts of DNA. Laboratory Investigation 2004 ;84 :649-657.

[308] Spiegelhalter DJ. Incorporating bayesian ideas into health-care evaluation. *Statistical Science* 2004 ;19(1) :154-174.

[309] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. Wiley & Sons, Ltd. Coll. Statistics in Practice. 2004.

[310] Spiegelhalter DJ, Thomas A, Best NG, Lunn D. Winbugs User Manual.Version 1.4. Cambridge, England : MRC Biostatistics Unit.

[311] Stephens PA, Buskirk SW, Hayward GD, Martinez Del Rio C. Information theory and hypothesis testing : a call for pluralism. *Journal of Applied Ecology* 2005 ;42 :4-12.

[312] Sterne JAC. Sifting the evidence-what's wrong with significance tests ? *BMJ* 2001 ;322 :226-31.

[313] Sterne JAC. Teaching hypothesis testing-time for a significant change ? Statistics in Medicine 2002 ;21 :985-994. DOI : 10.1002/sim.1129.

[314] Stoehr AM. Are significance thresholds appropriate for the study of animal behaviour ? *Animal Behaviour* 1999 ;55 :22-25.

[315] Stoler DL, Datta RV, Charles MA, Block AW, Brenner BM, Sieczka EM, Hicks WL, Loree TR, Anderson GR. Genomic instability measurement in the diagnosis of thyroid neoplasms. Head Neck 2002 ;24 :290-295.

[316] Tabor HK, Risch NJ, Myers R. Candidate-gene approaches for studying complex genetic traits : practical considerations. Nature Review Genetics 2002 ;3 :1-7.

[317] Tan W, Fang HB, Tian GL, Weib G. Testing multivariate normality in incomplete data of small sample size. Journal of Multivariate Analysis 2005 ;93 :164-179.

[318] Tanner M, Wong W. The calculation of posterior distributions by Data Augmentation (with discussion). Journal of the American Statistical Association 1987 ;82 :528-550.

[319] Teeguarden JG, Newton MA, Dragan YP, Pitot HC. Genome-wide loss of heterozygosity analysis of chemically induced rat hepatocellular carcinomas reveals elevated frequency of allelic imbalances on chromosomes 1, 6, 8, 11, 15, 17, and 20. Molecular Carcinogenesis 2000 ;28 :51-61.

[320] Tenenhaus M, Vinzia VE, Chatelin YM , Lauro C. PLS path modeling. Computational Statistics & Data Analysis 2005 ;48 :159-205.

[321] Tenenhaus M. La regression PLS. Théorie et Pratique. Édition Technip Paris 1998.

[322] Thomas DC, Clayton DG. Betting Odds and genetic association. Journal of the National Cancer Institute, Editorials 2004 ;96(6) :421-423.

[323] Tipping ME, Bishop CM. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 1999 ;61 :611-622.

[324] Tominaga Y. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. Chemometrics and Intelligent Laboratory Systems 1999 ;49 :105-115.

[325] Tomlinson IPM, Lambros MBK, Roylance RR. Loss of heterozygosity analysis : practically and conceptually flawed ? Genes, Chromosomes & Cancer 2002 ;34 :349-353.

[326] Tufte ER. The visual display of quantitative information. Graphics Press ; 2nd edition (May 2001). ISBN : 0961392142.

[327] Tukey JW. The philosophy of multiple comparisons. *Statistical Science* 1991 ;6 :100-116.

[328] Tunaru R. Models of association *versus* causal models for contingency tables. The Statistician 2001 ;50(3) :257-269.

[329] Van Driessche N, Demsar J, Booth EO. *et.al.*. Epistasis analysis with global transcriptional phenotypes. Nature Genetics 2005 ;35(5) :471-477.

[330] Vach W, Blettner M. Logistic regression with incompletely observed categorical covariates - investigating sensitivity against violation of the Missing at Random assumption. Statistics in Medicine 1999 ;14 :1315-1329.

[331] Vail A, Hornbuckle J, Spiegelhalter DJ, Thornton JG. Prospective application of bayesian monitoring and analysis in an "open" randomized clnical trial. Statistics in Medicine 2001 ;20 :3777-3787.

[332] Van Houwelingen HC. The future of biostatistics : expecting the unexpected. *Statistics in Medicine* 1997 ;16 :2773-2784.

[333] Vinzi VE. Explanatory methods for comparative analyses. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :275-286.

[334] Wald A. Sequential analysis. John Wiley & Sons, New York 1966.

[335] Wang ZC, Lin M, Wei LJ, Cheng Li C, Miron A, Lodeiro G, Harris L, Ramaswamy S, Tanenbaum DM, Meyerson M, Iglehart JD, Richardson A. Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers. Cancer Research 2004 ;64,64-71.

[336] Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R. A method for calling gains and losses in array CGH data. Biostatistics 2005 ;6,1,pp.45-58.

[337] Weber JC, Meyer N, Pencreach E, Schneider A, Guerin E, Neuville A, Stemmer C, Brigand C, Bachellier P, Rohr S, Kedinger M, Meyer C, Guenot D, Oudet P, Jaeck D, Gaub MP. Allelotyping analyses of synchronous primary and metastasis CIN colon cancers identified different subtypes. Int. J. Cancer 2007 ;120 :524-32.

[338] Weinberg CR. It's time to reabilitate the p-value. *Epidemiology* 2001 ; 12(3) :288-290.

[339] West CP, Dawson JD. Complete imputation of missing repeated categorical data : one-sample applications. *Statistics in Medicine* 2002 ;21 :203-217. DOI : 10.1002/sim.982

[340] Westerhuis JA, Kourti T, Macgregor JF. Analysis of multiblock and hierarchical PCA and PLS models. J. Chemometrics 11998 ;2 :301-321.

[341] White I, Carpenter J, Evans S. Schroter S. Eliciting and using expert opinions about non-response bias in randomised controlled trials. Technical report, Medical Statistics Unit, London School of Hygiene and Tropical Medicine. 2004.

[342] Wold S, Sjöström M, Eriksson L. PLS-regression : a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :109-130.

[343] Wold S, Trygg J, Berglund A, Antti H. Some recent developments in PLS modeling. Chemometrics and Intelligent Laboratory Systems 2001 ;58 :131-150.

[344] Wold H. Estimation of principal components and related models by iterative least squares. Multivariate Analysis, Krishnaiah PR (Ed.) Academic Press, New-York, 391-420.

[345] Wong KK, Tsang YTM, Shen J, Cheng RS, Chang YM, Man TK, Lau CC. Allelic imbalance analysis by high-density singlenucleotide polymorphic allele (SNP) array with whole genome amplified DNA. Nucleic Acids Research 2004 ;32,9,e69.

[346] Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled ? a review of published randomized controlled trials in major medical journals. Clinical Trials 2004 ;1 :368-376.

[347] Yang L. Distance-preserving projection of high-dimensional data for nonlinear dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence 2004 ;26 :1243-6..

[348] Yoshimoto K, Iwaki T, Inamura T, Fukui M, Tahira T, Hayashi K. Multiplexed analysis of post-PCR fluorescence-labeled microsatellite alleles and statistical evaluation of their imbalance in brain tumors. Jpn. J. Cancer Res. 2002 ;93,284-290.

[349] Zakzanis KK. Statistics to tell the truth, the whole truth, and nothing but the truth : formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of clinical neuropsychology* 2001 ;16(7) :653-667.

[350] Zander C, Thelaus J, Lindblad K, Karlsson M, Sjöberg K, Schalling M. Multivariate analysis of factors influencing repeat expansion detection. Genome Research 1998 ;8 :1085-1094.

[351] Zhai HL, Chen XG, Hu ZD. A new approach for the identification of important variables. Chemometrics and Intelligent Laboratory Systems 2006 ;80 :130-135.

[352] Zhou W, Galizia G, Goodman SN, Romans KE, Kinzler KW, Vogelstein B, Choti MA, Montgomery EA. Counting alleles reveals a connection between chromosome 18q loss and vascular invasion. Nature Biotechnology 2001 ;19 :78-81.

[353] Zhou X, Mok SC, Chen Z, Li Y, Wong DTW. Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. Hum Genet 2004 ;115 :327-330.

[354] Zhu JJ, Santarius T, Wu X, Tsong J, Guha A, Wu JK, Hudson TJ, McLBlack P. Screening for loss of heterozygosity and microsatellite instability in oligodendrogliomas. Genes, Chromosomes & Cancer 1998 ;21 :207-216.