Rapports scientifiques et techniques
de l'IFREMER

Nº 12 1988

BASES STATISTIQUES de la STRATÉGIE de SURVEILLANCE du MILIEU MARIN

Philippe GROS
Institut français de recherche pour l'exploitation de la mer

IFREMER

Rapports scientifiques et techniques
de l'IFREMER

Nº 12 1988

BASES STATISTIQUES de la STRATÉGIE de SURVEILLANCE du MILIEU MARIN

DÉTECTION DE L'IMPACT INDUIT par l'AMÉNAGEMENT DU LITTORAL

Philippe GROS
Institut français de recherche pour l'exploitation de la mer

IFREMER

Le rapport

BASES STATISTIQUES de la SURVEILLANCE du MILIEU MARIN

Détection de l'impact induit par l'aménagement du littoral

a été rédigé par
Philippe GROS

INSTITUT FRANÇAIS de RECHERCHE pour L'EXPLOITATION de la MER (IFREMER)
DIRECTION DE L'ENVIRONNEMENT ET DES RECHERCHES OCÉANIQUES
Département Environnement littoral - Centre de BREST

Service de la Documentation
et des Publications (S.D.P.)
IFREMER - Centre de Brest
BP 70 - 29263 PLOUZANÉ
Tél. 98 22 40 13 - Télex 940 627F

ISSN - 0761-3970

TABLE DES MATIERES

INTRODUCTION 7

I. LES OUTILS STATISTIQUES CLASSIQUES DE LA DECISION 9
1.1. RAPPEL PRELIMENAIRE DES CONCEPTS FONDAMENTAUX 9
1.2. DETECTION DE L'IMPACT EN TEMPS DIFFERE : APPROCHE PARAMETRIQUE 13
1.2.1. Modification de la statistique usuelle de comparaison de deux moyennes quand la taille de l'un des échantillons est fixée 13
1.2.2. Hétérogénéité des variances induite par la loi de Taylor : remèdes 21
1.3. RECAPITULATION ET DISCUSSION..... 31

II. LES LIMITES DES PROCEDURES PARAMETRIQUES - PALLIATIFS 39
2.1. LA NOTION DE ROBUSTESSE 39
2.2. DETECTION DE L'IMPACT EN TEMPS DIFFERE : APPROCHE NON PARAMETRIQUE . 41
2.2.1. Hypothèses statistiques 42
2.2.2. Puissance du test de Mann-Whitney 43
2.2.3. Limite des approximations : cas où n_1 est fixé 46

III. L'EVOLUTION DE L'OUTIL STATISTIQUE ECLAIREE PAR UNE ETUDE DE CAS : LA DETECTION DE L'IMPACT A L'AIDE D'OBSERVATIONS SYNCHRONES 47
3.1. DEFINITION DU PROBLEME 47
3.2. LA SOLUTION CLASSIQUE 47
3.3. LE TRAITEMENT PARAMETRIQUE DES GENERALISATIONS DE L'ALTERNATIVE 51
3.4. L'APPROCHE NON PARAMETRIQUE 54
3.4.1. Hypothèses statistiques 55
3.4.2. Emploi du test de Wilcoxon 55
3.4.3. Calcul approché de la puissance 57
3.5. LES TECHNIQUES DE REECHANTILLONNAGE ET LES STATISTIQUES ROBUSTES ... 60
3.5.1. L'estimation robuste d'une valeur centrale 63
3.5.2. Le bootstrap 69

IV. MODALITES D'APPLICATION : LES CHOIX ESSENTIELS 73
4.1. LE CHOIX DE LA VARIABLE INDICATIVE 74
4.2. LES ECHELLES D'OBSERVATIONS : CHOIX DES FENETRES SPATIO-TEMPORELLES. 80
4.3. CHOIX ECONOMIQUES ET CRITERE D'OPTIMALITE POUR LA DEFINITION D'UNE STRATEGIE 94

CONCLUSION 101
BIBLIOGRAPHIE 105
ANNEXE I - DISTRIBUTIONS USUELLES ET CONVERGENCES : RAPPELS 117
ANNEXE II - ECHANTILLONNAGE : PRESENTATION RESUMEE DES PRINCIPALES STRATEGIES 135

RESUME

Le rapport présente les éléments fondamentaux nécessaires à la définition et à la mise en oeuvre d'un protocole de surveillance du milieu marin au voisinage d'un aménagement de la façade littorale. L'objectif est de montrer comment doivent être appliqués dans ce contexte les tests d'hypothèses ; à cet égard, le premier chapitre rappelle que l'étape prioritaire est celle de la définition rigoureuse de l'alternative (opposée à l'hypothèse nulle d'absence d'impact), *i.e.* de l'ampleur et de la nature de l'éventuelle modification de l'écosystème devant être testée. Ce préalable autorise en particulier à déterminer l'effort d'échantillonnage qu'il convient d'engager pour statuer sur l'effet de l'aménagement avec des probabilités d'erreur fixées *a priori* . La première stratégie examinée de ce point de vue est celle qui vise à détecter une différence entre les deux situations consécutives : "avant" *vs.* "après" aménagement. La puissance de la comparaison paramétrique usuelle est d'abord calculée sous les hypothèses de normalité et d'homoscédasticité, et des solutions approchées sont ensuite proposées pour le cas où existe une dépendance stochastique entre moyenne et variance. Les limites du domaine d'application de la démarche paramétrique étant soulignées, le second chapitre est consacré au traitement des données acquises selon la même stratégie à l'aide d'une méthode non paramétrique.

Le troisième chapitre expose une stratégie différente, élaborée pour déceler l'influence spatiale de l'impact : les observations sont saisies simultanément d'une part dans le champ proche de l'aménagement ("zone impactée"), et d'autre part dans un secteur voisin mais indemne de toute perturbation imputable à ce même aménagement ("zone témoin"). Pour ce protocole est considérée une gamme de techniques qui conduisent à des inférences statistiques tributaires d'un corps d'hypothèses plus ou moins restrictif : ainsi sont présentées la solution classique, la procédure non paramétrique, et enfin la construction par rééchantillonnage (bootstrap) d'un intervalle de confiance attaché à l'estimation robuste d'une valeur centrale.

Au quatrième chapitre sont discutés les choix concrets essentiels : celui de la variable indicative (*i.e.* , la caractéristique de l'écosystème la plus adéquate à la formulation d'un diagnostic), celui des échelles d'observation spatio-temporelles, et celui du critère d'optimalité permettant de quantifier le meilleur rapport coût/précision d'une stratégie. Un protocole est en outre proposé, qui permet de traiter conjointement les deux aspects de l'impact (temporel et spatial) envisagés séparément aux chapitres précédents.

ABSTRACT

STATISTICAL BASIS FOR MARINE ENVIRONMENTAL MONITORING
Detection of ecological impact of a coastal engineering project.

This report is devoted to the problem of testing statistical hypotheses within the general framework of impact assessment. For any given engineering project in the coastal area, the aim is to supply some rules for the definition and implementation of an ecological survey program, which must be designed to have a specified probability of detecting a predicted change of specified nature and magnitude. In the first chapter, it is thus remembered that the main step is to identify the distinguishing ecological feature between "reference" and "affected" states, *i.e.* the difference between the null and alternative hypotheses respectively. Power calculations are then performed for the usual parametric test of comparison between two periods : pre-operational and operational ; this approach corresponds to the so-called "before *vs.* after strategy". Approximate solutions are also provided when the data do not match the classical assumption of homoscedasticity, and exhibit a stochastic variance-mean relationship.

In the second chapter, the limits of reliability of parametric procedures are outlined. The nonparametric treatment of data collected according to the same strategy is set out.

The "impacted area *vs.* control area strategy" constitutes the topic of the third chapter. The purpose is to assess the spatial component of impact ; accordingly, paired observations are sampled at control and affected stations. Several methods of statistical inference are presented ; they can be distinguished by the broadness of the required basic assumptions : the classical solution, the nonparametric procedure, and the use of a resampling technique (the bootstrap) for constructing a confidence interval of a robust location estimate.

Three fundamental options are discussed in the fourth chapter : (i)the choice of the diagnostic variable ; (ii)the choice of the spatio-temporal scales of observation ; and (iii)the choice of the objective function for optimizing the allocation of sampling (cost-power tradeoff). Guidelines are given for the definition of a monitoring strategy taking simultaneously location (impact *vs.* control) and condition (before *vs.* after) into account.

INTRODUCTION

La mise en évidence des effets du développement économique, activités industrielles et agricoles, urbanisation..., sur l'écosystème marin est l'un des problèmes posés aux gestionnaires et aménageurs de l'environnement côtier. D'un point de vue plus concret, la question traitée ici est celle des conséquences d'un aménagement implanté sur la façade littorale.

En la matière, l'expérience du CNEXO, puis de l'IFREMER, a été pour une large part acquise lors de l'évaluation de l'incidence écologique des effluents thermiques et chlorés rejetés par les centrales électro nucléaires installées en bordure de mer. Les enseignements tirés de ces programmes d'étude sont appelés à être transposés dans un contexte beaucoup plus général ; cela constitue l'objet du présent rapport, dans lequel il sera néanmoins souvent fait référence, à titre d'illustration, au problème de la détection de l'impact engendré par le fonctionnement des centrales nucléaires.

Au plan méthodologique, plusieurs démarches sont envisageables pour attester l'impact d'un aménagement : par exemple, une modélisation explicative des principales variables d'état qui décrivent l'écosystème, afin de simuler les conséquences les plus vraisemblables de la modification par l'aménagement d'une ou plusieurs des variables forçantes qui contrôlent l'évolution de ce même écosystème. Idéalement, les moyens d'obtenir des prévisions crédibles devraient être partie intégrante du projet d'aménagement. Cela ne constitue cependant que l'une des approches possibles ; au même titre que les autres, elle possède ses propres limitations (e.g., bornes de l'emprise spatio-temporelle du modèle, pertinence des hypothèses sur lesquelles il repose, sensibilité éventuelle à des paramètres estimés de façon insuffisamment précise). Les restrictions attachées à la portée des conclusions d'une méthode particulière conduisent à prôner l'utilisation conjointe d'approches complémentaires susceptibles de se valider mutuellement.

A cet égard, l'objectif des développements qui vont être présentés est très précisément défini : éclairer la contribution aux études de "surveillance du milieu" de la théorie statistique de la décision, et montrer de quelle(s) manière(s) un impact peut être décelé à l'aide de tests.

Par ailleurs, ce rapport s'adresse prioritairement à la communauté scientifique en charge du contrôle de la "qualité du milieu marin". Le public visé est donc principalement composé de biologistes, de chimistes, d'écotoxicologues..., possédant une connaissance statistique minimale ; par précaution, les bases intellectuelles de la décision statistique sont néanmoins rappelées au début du premier chapitre. De plus, les définitions et propriétés des lois de probabilité utilisées dans le cours du texte sont énoncées à l'annexe I, accompagnées du résumé de théorèmes limites qui fondent les approximations appliquées au cas des grands échantillons. Une brève présentation des principales stratégies d'échantillonnage fait l'objet d'un second appendice : le protocole de sondage contribue en effet à déterminer la structure statistique des données.

Les trois premiers chapitres présentent les grands types de procédures (paramétriques, non paramétriques, simulations), une attention toute particulière étant accordée à l'évaluation de la puissance des tests en vue d'une planification rationnelle de la saisie des informations. Enfin, conformément à la vocation appliquée du rapport, les aspects pratiques de l'élaboration d'une stratégie de surveillance sont discutés au quatrième chapitre.

CHAPITRE I

LES OUTILS STATISTIQUES CLASSIQUES DE LA DECISION

1.1. RAPPEL PRELIMINAIRE DES CONCEPTS FONDAMENTAUX

Un test statistique est une règle de décision visant à statuer sur une réalité inconnue à partir de données expérimentales : par exemple, une information quantitative sur une caractéristique du milieu telle que la concentration en chlorophylle, la biomasse du mésozooplancton... Concrètement, cela revient dans le cas présent à choisir entre deux hypothèses :

- l'hypothèse nulle H_o, selon laquelle l'aménagement n'a aucun effet sur la caractéristique étudiée,
- et l'alternative H_1 qui stipule l'existence d'un impact.

La décision en faveur de l'une ou l'autre de ces deux conceptions de la réalité repose sur la comparaison de la valeur prise par une statistique avec un seuil fixé *a priori*, et qui ne peut être dépassé qu'avec une probabilité faible si H_o est vraie. Le point fondamental est le suivant : la statistique du test, étant par construction une fonction des données expérimentales, est calculée sur l'échantillon ; pour évaluer la probabilité qu'elle dépasse un seuil donné, il est nécessaire de munir l'échantillon d'une structure statistique. Les tests présentés dans ce premier chapitre présupposent une structure gaussienne des données, hypothèse qui n'est guère contraignante dans la mesure où ils concernent des valeurs centrales.

Dans ce contexte, l'acte d'échantillonnage engendre une variable aléatoire dont les réalisations sont plus ou moins probables sous H_o ; l'observation du résultat obtenu permet en retour d'inférer que H_o est plus ou moins vraisemblable. Corrélativement, la composante stochastique de la procédure attache à tout choix fondé sur un test statistique deux probabilités d'opter pour une hypothèse fausse :

(1) La probabilité α de commettre l'erreur de première espèce, *i.e.* de rejeter à tort H_o :

$$\alpha = \text{Proba}\{H_o \text{ repoussée} \mid H_o \text{ vraie}\}$$

Cette probabilité est fixée *a priori* à une valeur habituellement faible, *e.g.* $\alpha = .05$. Sachant qu'il est quasi-certain qu'un aménagement produit un effet sur le milieu marin, l'objectif visé (qui est la détection statistique de l'impact) incite néanmoins à se préoccuper essentiellement du second risque d'erreur, défini ci-après :

(2) La probabilité β de commettre l'erreur de deuxième espèce, *i.e.* ne pas rejeter H_o quand elle est fausse.

$$\beta = \text{Proba}\{H_o \text{ non repoussée} \mid H_1 \text{ vraie}\}$$

Il est plus souvent fait référence au complément à 1 de cette probabilité, noté π, qui désigne la "puissance" du test :

$$\pi = 1 - \beta = \text{Proba}\{H_o \text{ repoussée} \mid H_1 \text{ vraie}\}$$

Concrètement, la puissance π représente la probabilité de décider en faveur de l'alternative H_1 stipulant l'existence d'un impact, et ceci lorsqu'il y a réellement impact.

Au plan expérimental, le contrôle du risque de seconde espèce est délicat ; en effet, à la différence de α, la probabilité β est fonction de plusieurs facteurs, à savoir :

(1) De la loi des variables échantillonnées.

(2) De "l'emplacement" de la région critique dans l'univers des résultats possibles sous H_o, la région critique étant l'ensemble des résultats qui entraîneront la décision de repousser H_o. Selon le formalisme développé par NEYMAN et PEARSON, le principe directeur de la construction de cette région est le suivant : ayant fixé *a priori* la probabilité α (*i.e.*, la taille de la région critique), minimiser la probabilité β. A cet égard, une région critique dont la puissance (relativement aux mêmes hypothèses H_o et H_1) n'est jamais inférieure à celle de toute autre région de même taille est appelée région critique optimale. Il est à noter qu'une telle région n'existe pas nécessairement, son existence éventuelle étant conditionnée par la nature des hypothèses en présence (*cf.* KENDALL & STUART, 1979, chap. 22-23).

(3) Du degré d'erreur dont est entachée l'hypothèse H_o quand elle est fausse. En pratique, cela correspond à l'amplitude de l'effet de l'aménagement. Par exemple, si le problème posé est celui de caractériser l'impact d'une centrale électronucléaire sur l'écosystème côtier : soit X_1 la variable aléatoire qui à un certain compartiment de l'écosystème associe la mesure de sa biomasse, ou bien encore de sa production, avant mise en service de la centrale ($X_1 = x_{11},\ldots,X_1 = x_1n_1$, où n_1 est le nombre d'observations), et soit μ_1 la vraie moyenne de X_1. Soient X_2, μ_2 et n_2 définis de la même manière, mais pour la phase durant laquelle la centrale est en fonctionnement. Dans ces conditions, la vraie différence $\Delta = \mu_1 - \mu_2$ est une caractérisation de l'un des effets, sur le compartiment étudié, des contraintes exercées (toutes choses égales par ailleurs) par les rejets de la centrale.

Si le sens de l'écart attendu est connu à l'avance (*e.g.*, $\mu_2 < \mu_1$), les hypothèses H_o et H_1 peuvent être formulées comme suit :

$H_o : \Delta = 0$, contre $H_1 : \Delta > 0$.

Lorsque le test utilisé est non biaisé, la puissance $\pi(\Delta)$ est une fonction croissante de Δ : la puissance est alors d'autant meilleure que l'hypothèse H_o est "d'autant plus fausse".

(4) Une fois choisies la variable "sensible" à l'impact, ainsi que la procédure de test, les trois points inventoriés ci-dessus constituent des propriétés intrinsèques du problème posé. Expérimentalement, il n'est alors possible d'intervenir que sur deux des trois quantités suivantes : le nombre d'observations $n = n_1 + n_2$, et les probabilités α et $\beta(H_1)$.

(i)Si n est donné, la réduction de α entraîne généralement l'augmentation de $\beta(H_1)$. Autrement dit, moins il est souhaité risquer de conclure à tort à l'effet de l'aménagement (erreur de première espèce), plus il devient probable d'accepter à tort l'hypothèse d'absence d'impact (erreur de seconde espèce). Sachant donc que le prix d'une diminution de la taille de la région critique est une dégradation de la puissance, la question à résoudre est celle de l'établissement d'un compromis entre les deux types d'erreur.

(ii)Inversement, s'il est possible de moduler à volonté l'effort d'échantillonnage, alors n peut être ajusté de telle sorte que α et $\beta(H_1)$ descendent à des niveaux fixés à l'avance. Face à un problème donné, il faut donc définir une

combinaison optimale de α, $\beta(H_1)$ et n, opération qui nécessite de disposer d'une information sur le coût de chacun des deux types d'erreur, de même que sur celui de la saisie d'une observation.

Cette discussion se prolonge au-delà de considérations strictement pratiques : si un test est effectué au seuil α avec n très grand, sa puissance devient très proche de 1 face à toute alternative. Du fait de ce comportement asymptotique, la procédure a été l'objet de la critique suivante : nul ne croit à la rigoureuse exactitude d'une quelconque hypothèse, qui n'est qu'un modèle approché d'une réalité inconnue. Un échantillon de très grande taille entraînera donc presque certainement le rejet (au seuil α) de l'hypothèse testée. D'où l'argument : dans ces conditions, pourquoi prendre la peine de tester l'hypothèse à partir d'un échantillon de faible taille, sachant que le diagnostic établi est moins fiable que la décision de rejet qui serait obtenue avec un échantillon beaucoup plus grand ?

Ce paradoxe appelle deux éclaircissements : tout d'abord, n étant donné, la question n'est pas celle de l'exactitude de l'hypothèse testée, mais celle de son aptitude à constituer une hypothèse de travail réaliste : ses qualités ressortent principalement à sa forte plausibilité et à son caractère opérationnel . Le fait que H_o ne soit qu'approximativement valide est alors pris en compte par la définition de l'alternative qui lui est opposée, en ménageant une "distance" suffisante entre celle-ci et celle-là. Pour reprendre l'exemple évoqué précédemment, la différence Δ doit être d'une amplitude telle qu'elle corresponde à l'expression d'une sensible perturbation du compartiment étudié dans l'écosystème, étant entendu que toute différence inférieure à cette amplitude sera considérée comme "une absence d'impact", même si cela ne recouvre pas strictement la situation décrite par H_o (*i.e.*, $\Delta = 0$).

En second lieu, lorsque n est autorisé à croître indéfiniment, l'hypothèse nulle n'est rejetée avec une probabilité voisine de 1 que si le seuil α est maintenu constant. Or il n'existe aucune raison d'imposer cette contrainte, et c'est précisément l'habitude de fixer α à une valeur conventionnelle (.05, .01, voire .001) qui engendre le paradoxe. Ce dernier disparaît si le gain de sensibilité dû à l'augmentation de n est réinvesti dans une diminution conjointe de α et $\beta(H_1)$: si α varie en sens inverse de n, il n'est alors plus assuré qu'un faible écart à H_o (*e.g.*, un petit Δ) entraîne le rejet systématique de H_o ; cette décision dépend de la vitesse de décroissance de α. En conséquence, KENDALL & STUART (1979) recommandent de s'en tenir à un choix quel-

que peu arbitraire, mais présentant néanmoins le mérite de la rationalité : poser $\alpha = \beta(H_1)$ pour l'écart à H_o au-delà duquel cette hypothèse sera en pratique considérée comme fausse. C'est la position qui sera adoptée dans la présente étude.

1.2. DETECTION DE L'IMPACT EN TEMPS DIFFERE : APPROCHE PARAMETRIQUE

Cette approche peut être illustrée par un exemple précis : c'est l'une des démarches qui fut en premier envisagée par le CNEXO en 1974, lorsque Electricité de France lui confia l'étude des conséquences écologiques engendrées par l'implantation de centrales électronucléaires sur le littoral de la Manche. L'idée consiste schématiquement à caractériser "l'état des lieux" à l'aide de descripteurs biologiques avant la mise en chantier de la centrale, pour se doter ainsi d'une référence par rapport à laquelle seraient définies les modifications éventuellement décelées dans le champ proche des rejets thermiques. A l'évidence, la validité de la comparaison repose sur la stabilité moyenne de la référence, aussi bien avant qu'après le début de l'impact. Cela nécessite donc :

(i) soit de s'appuyer sur une hypothèse de "bruit blanc pluriannuel" des variations de X (hors du champ proche), variations qui seront donc supposées dépourvues de tendance ;

(ii)soit de faire l'économie de cette hypothèse (ou tout au moins de se donner le moyen de la vérifier), et décider de contrôler simultanément une zone témoin localisée de telle manière qu'elle demeure hors de portée de l'impact : de la sorte peuvent être filtrées d'éventuelles variations "naturelles", *i.e.*, dont la causalité est indépendante du fonctionnement de la centrale (ou plus généralement des effets de l'aménagement).

Quelle que soit l'option retenue, les calculs qui vont suivre ne sont d'application pertinente que dans un contexte de stabilité du milieu aux limites du champ présumé affecté par l'impact de l'aménagement.

1.2.1. Modification de la statistique usuelle de comparaison de deux moyennes quand la taille de l'un des échantillons est fixée

Ne prenant pas en considération à ce niveau la question des échelles de temps et d'espace, et conservant les notations introduites plus haut, soit

X la variable aléatoire qui associe à l'un des compartiments de l'écosystème la mesure de l'une de ses caractéristiques, et soient $\mu = E(X)$, $\sigma^2 = V(X)$. L'indice 1 désigne la situation avant, et l'indice 2 après mise en place de l'aménagement. D'où la formulation des hypothèses statistiques :

$$H_o : \mu_1 = \mu_2 \quad , \text{ contre } H_1 : \mu_1 - \mu_2 = \Delta, \text{ avec par exemple } \Delta > 0,$$

i.e., l'impact entraîne une réduction de la moyenne de la caractéristique étudiée dans le champ proche du rejet.

Le développement ci-après suppose l'homogénéité des variances ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), qui peut être obtenue par transformation des données ; cette question sera examinée ultérieurement (§1.2.2.), de même que le problème de Behrens-Fisher (§3.3.). La dispersion étant soit stable, soit stabilisée (*i.e.*, indépendante de la moyenne), la variance commune σ^2 est habituellement estimée sans biais par s^2 :

$$s^2 = ((n_1 - 1)s_1^2 + (n_2 - 1)s_2^2) / (n_1 + n_2 - 2)$$

où s_1^2 et s_2^2 sont les variances empiriques calculées à partir de n_1 et n_2 réalisations indépendantes de X_1 et X_2 respectivement. Dans ces conditions, la loi de la quantité $\nu s^2/\sigma^2$ est d'autant mieux décrite par un χ^2 à $\nu = n_1 + n_2 - 2$ d.d.l. que les lois de X_1 et X_2 sont voisines de la normalité.

A ce niveau, il faut souligner l'asymétrie du problème traité, en observant que dans le cas d'une étude d'impact en temps différé, les nombres d'observations n_1 et n_2 ne jouent pas des rôles équivalents. En effets, n_1 est une quantité sur laquelle l'expérimentateur ne peut plus agir après la mise en service de la centrale : passé ce délai, n_1 est définitivement figé. Il est en revanche toujours possible de moduler n_2 après cette étape, et la question de la détection statistique peut alors être posée sous deux formes distinctes et complémentaires :

(1) Disposant de n_1 réalisations de X_1, et prévoyant d'effectuer n_2 observations durant la période dite de surveillance du milieu, quel est alors le plus petit écart Δ décelable à des niveaux de risque α et β fixés ? Le problème revient à exprimer Δ comme une fonction des facteurs contrôlés :

$$\Delta = f(n_2, \alpha, \beta \mid n_1)$$

(2) L'"état de référence" étant décrit à l'aide d'un nombre invariable d'observations n_1, quelle valeur attribuer à n_2 pour qu'un écart donné Δ soit détecté au seuil α avec une puissance $1 - \beta$ donnée ? Il s'agit ici d'expliciter la fonction g :

$$n_2 = g\ (\Delta,\ \alpha,\ \beta \mid n_1)$$

Les formulations précédentes font apparaître, sous l'hypothèse de bruit blanc pluriannuel, que le problème est conditionné par l'effort d'échantillonnage alloué à la définition de l'"état de référence". La procédure classique de comparaison de moyennes doit être aménagée en conséquence (HEILBRUN & Mc GEE, 1985) ; ainsi, la variance σ^2 est-elle seulement estimée par s_1^2, le nombre de degrés de liberté étant lui-même réduit à $\nu = n_1 - 1$.

La statistique du test permettant de décider en faveur de H_o ou bien de H_1 devient donc :

$$y = (\bar{x}_1 - \bar{x}_2)/(s_1\sqrt{1/n_1+1/n_2}) \qquad \{1\}$$

où $\bar{x}_1$ et $\bar{x}_2$ sont les moyennes empiriques qui estiment les paramètres inconnus μ_1 et μ_2. Lorsque la valeur de y dépasse un seuil C_α, l'hypothèse nulle ($\mu_1 = \mu_2$) est repoussée en faveur de l'alternative ($\mu_1 > \mu_2$). Pour déterminer la valeur de C_α, il est nécessaire de connaître la loi de y sous H_o. Soit $X^\dagger$ la variable aléatoire[1] :

$$X^\dagger = ((\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)) \ /\ (\sigma\sqrt{1/n_1+1/n_2})$$

---

(1) Les variables $X^\dagger$ et X^2 sont introduites ici sans justification. Cette opération n'est toutefois nullement arbitraire, car ces deux variables apparaissent dans l'expression du rapport des vraisemblances, qui permet de déterminer la région critique. La construction du test repose en effet sur la méthode de Neyman et Pearson.

Le théorème de la limite centrale garantit qu'il est légitime d'admettre, en première approximation :

$$X^{\dagger} \sim N(0,1)$$

Et soit X^2 : $X^2 = (n_1 - 1)s_1^2/\sigma^2$,

dont la loi est celle d'un χ^2 à n_1-1 d.d.l. dans le cadre gaussien. Par ailleurs, une caractéristique de ce cadre est l'indépendance entre $\bar{x}$ et s^2, qui entraîne l'indépendance entre $X^{\dagger}$ et X^2. En conséquence, le rapport $X^{\dagger}\sqrt{\nu}/\sqrt{X^2}$ suit une loi de Student à n_1-1 d.d.l. ; ce résultat est exact pour une distribution parente normale, et ne constitue plus qu'une approximation lorsque cette distribution est seulement "voisine" de la normalité. Il s'énonce formellement :

$$X^{\dagger}\sqrt{\nu}/\sqrt{X^2} = ((\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2))/(s_1\sqrt{1/n_1+1/n_2}) \sim t_{\nu}$$

Sous H_o (*i.e.*, $\mu_1-\mu_2=0$), la loi de y est donc un t Student à n_1-1 d.d.l. La valeur de C_{α} vaut donc :

$$C_{\alpha} = t_{\nu,\alpha}$$

où $t_{\nu,\alpha}$ est définie comme suit : soit F la fonction de répartition, de la loi de Student à ν d.d.l., alors $F(t_{\nu,\alpha}) = 1-\alpha$.

La quantité α désigne la probabilité de commettre l'erreur de première espèce, *i.e.* : $\alpha = \text{Proba}\{y>t_{\nu,\alpha} \mid \Delta=0\}$

et avec la règle de décision précédemment indiquée, le risque de seconde espèce s'exprime

$$\beta(H_1) = \text{Proba}\{y<t_{\nu,\alpha} \mid \Delta>0\}$$

Toutefois, la loi de y sous H_1 (*i.e.*, $\Delta>0$) n'est plus celle d'un t centré ; en effet :

$$(\bar{x}_1-\bar{x}_2)/(\sigma\sqrt{1/n_1+1/n_2}) \underset{\sim}{H_1} N(\Delta/\sigma,1)$$

Cela entraîne que la statistique du test suit sous H_1 une loi de Student non centrée à $n_1 - 1$ d.d.l., et de paramètre de décentrement Δ/σ. Le calcul de $\beta(H_1)$ nécessite donc le recours à la tabulation de la loi t non centrée. Cette approche ne sera pas retenue ici, car l'objectif du présent développement est d'aboutir à l'écriture formelle des fonctions f et g définies plus haut. Sachant que :

$$X^{\dagger}\sqrt{\nu}/\sqrt{X^2} = ((\bar{x}_1-\bar{x}_2) - \Delta)/(s_1\sqrt{1/n_1+1/n_2})$$

suit sous H_1 un t centré à $\nu = n_1 - 1$ d.d.l., et faisant apparaître ce rapport dans l'expression $\beta(H_1)$, il vient :

$$\beta(H_1) = \text{Proba}\{\frac{(\bar{x}_1-\bar{x}_2)-\Delta}{s_1\sqrt{1/n_1+1/n_2}} < t_{\nu,\alpha} - \frac{\Delta}{s_1\sqrt{1/n_1+1/n_2}} \mid \Delta>0\}$$

Ainsi qu'il vient d'être rappelé, le premier membre de l'inégalité suit une loi de Student centrée. Par conséquent, le second membre vérifie :

$$t_{\nu,\alpha} - (\mu_1-\mu_2)/(s_1\sqrt{1/n_1+1/n_2}) = t_{\nu,1-\beta} = -t_{\nu,\beta}$$

Dans la suite, la différence Δ sera exprimée relativement à la moyenne μ_1 de la caractéristique étudiée avant fonctionnement de la centrale. Soit donc :

$$\delta = \Delta/\mu_1$$

avec cette notation :

$$\mu_2 = \mu_1(1-\delta)$$

La situation décrite par H_o correspond à $\delta = 0$. Quant à la classe des alternatives unilatérales H_1, elle correspond aux valeurs de δ comprises dans l'intervalle $]0,1]$. Ainsi, si μ représente la biomasse moyenne de l'un des compartiments de l'écosystème, l'alternative $\delta = 1$ exprime la disparition de ce compartiment.

Dans ce système de notations, l'égalité précédemment établie devient :

$$t_{\nu,\alpha} + t_{\nu,\beta} = \delta.\mu_1/(s_1\sqrt{1/n_1+1/n_2}) \qquad \{2\}$$

Remplaçant μ_1 par son estimation non biaisée $\bar{x}_1$, et faisant apparaître le coefficient de variation empirique $\hat{CV}_1$:

$$\hat{CV}_1 = s_1/\bar{x}_1$$

il vient alors :

$$t_{\nu,\alpha}+t_{\nu,\beta} = \delta/(\hat{CV}_1\sqrt{1/n_1+1/n_2})$$

Cette relation permet de répondre aux deux questions adressées en préambule, à savoir expliciter les fonctions f et g.

(1) L'information acquise avant fonctionnement de la centrale étant résumée par $\hat{CV}_1$, la plus petite réduction relative de μ_1 qu'il est possible de déceler (avec des probabilités d'erreur α et β) à l'aide de n_2 observations réalisées pendant la période de surveillance vaut :

$$\delta = (t_{\nu,\alpha}+t_{\nu,\beta})\hat{CV}_1\sqrt{1/n_1+1/n_2} \qquad \{3\}$$

Il est à noter que le seuil minimal de détection de l'impact tend vers une limite non nulle lorsque l'effort d'échantillonnage alloué à la période de surveillance augmente indéfiniment :

$$n_2 \to +\infty \Rightarrow \delta \to \delta_L = (t_{\nu,\alpha}+t_{\nu,\beta})\hat{CV}_1/\sqrt{n_1} \qquad \{4\}$$

Autrement dit, le pouvoir de résolution demeure déterminé par la précision avec laquelle μ_1 est estimée avant que la centrale n'entre en fonctionnement, et ce quel que soit l'effort consenti par la suite.

(2) Pour qu'une réduction relative δ fixée *a priori* soit détectée au seuil α avec une puissance $\pi = 1 - \beta$ donnée, il faudra effectuer pendant la période de surveillance un nombre d'observations égal à :

$$n_2 = n_1\hat{CV}_1^2(t_{\nu,\alpha}+t_{\nu,\beta})^2/(n_1\delta^2-(t_{\nu,\alpha}+t_{\nu,\beta})^2\hat{CV}_1^2) \qquad \{5\}$$

Conformément à ce qui a été souligné, n_2 n'est positif que si $\delta>\delta_L$, et devient infini pour $\delta=\delta_L$.

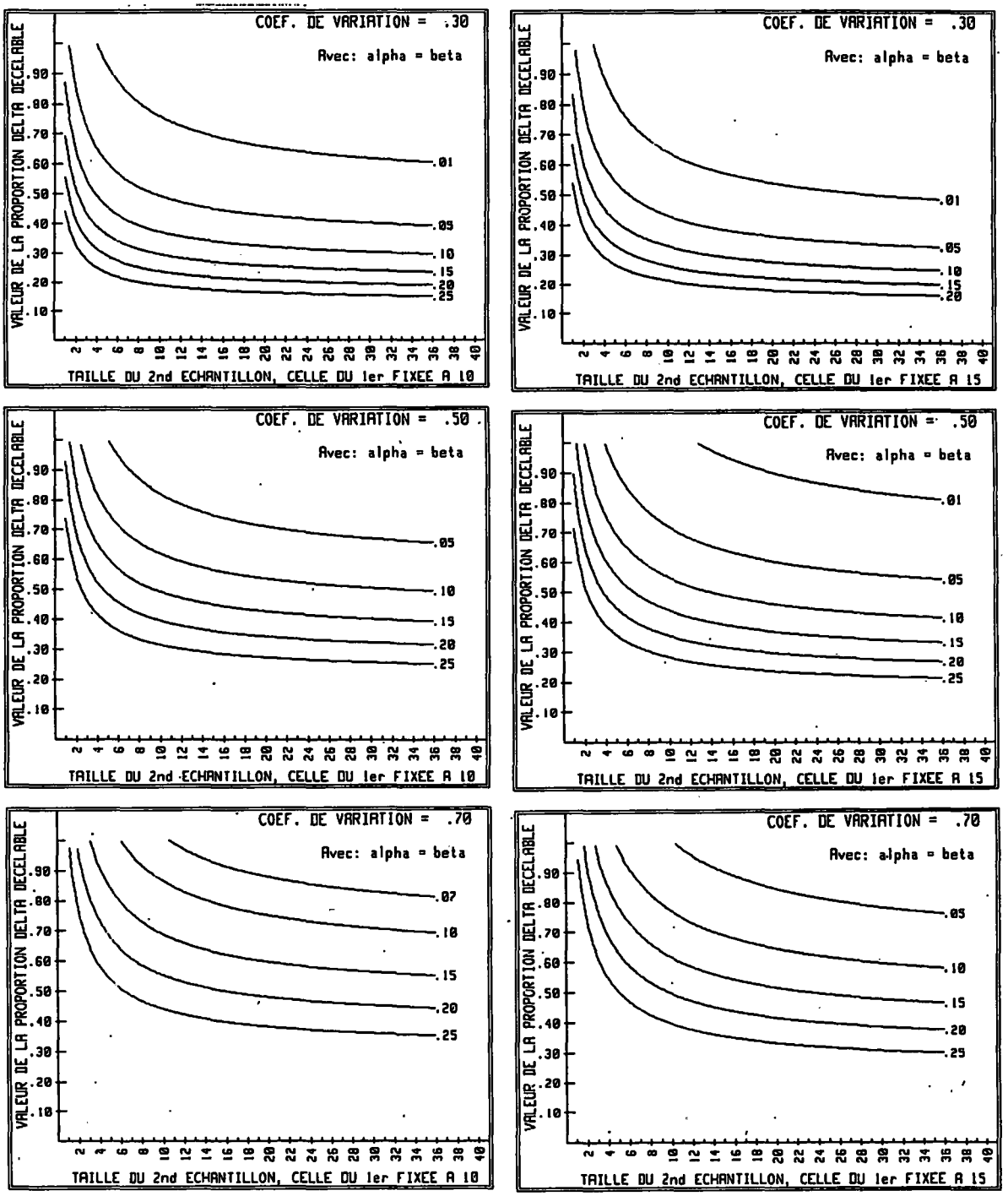


Figure 1 - Exemple d'application de l'équation {3}.

Courbes d'isorisques ($\alpha=\beta=.01,\ldots,.25$) associées à la détection de l'alternative δ (ordonnées) fonction de n_2 (abscisses), pour n_1 fixé à 10 (à gauche) ou bien à 15 (à droite), et pour trois valeurs de $\hat{CV}_1$ (.3, .5, .7).

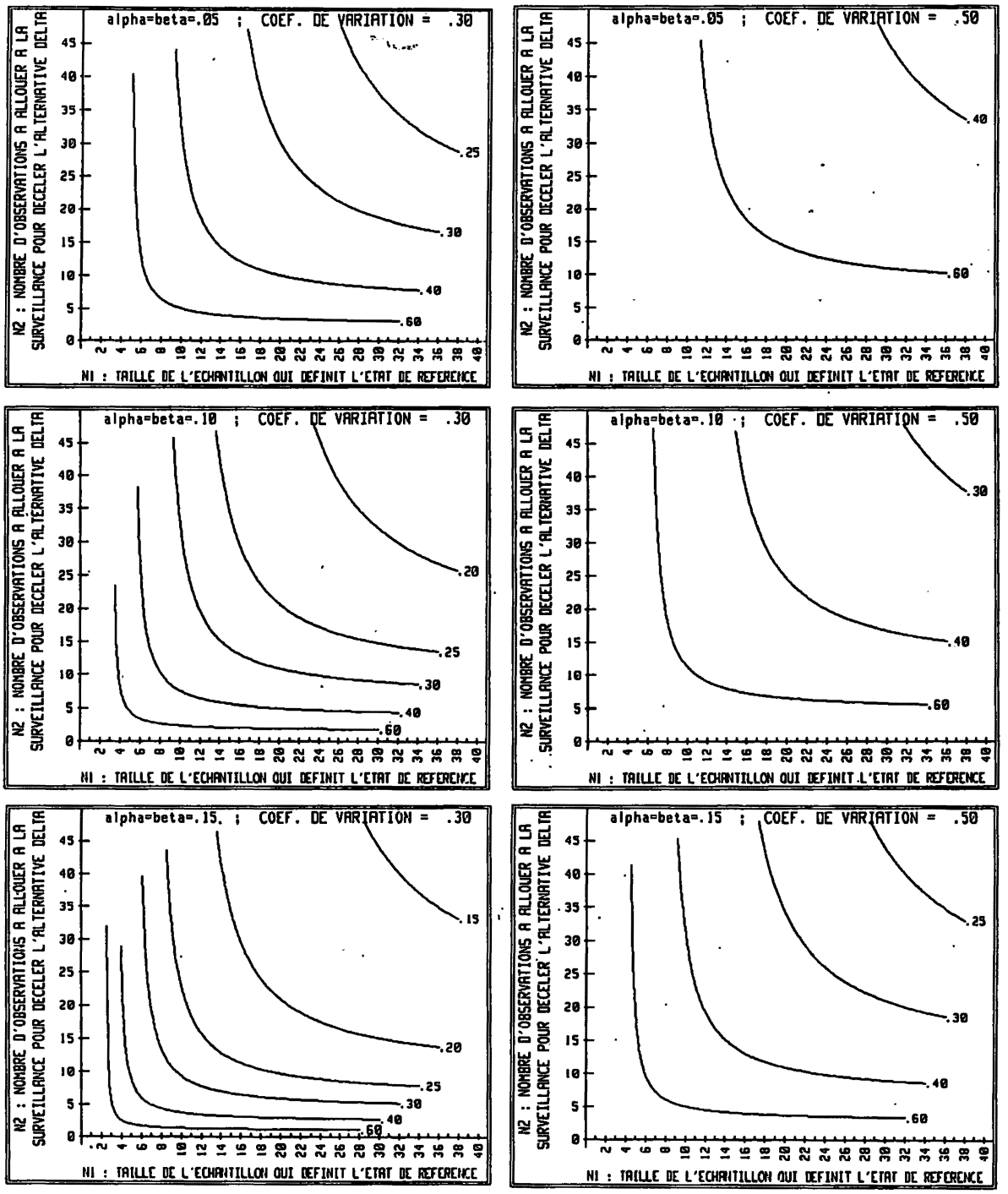


Figure 2 - Exemple d'application de l'équation {5}.

Effort d'échantillonnage à consacrer à la surveillance (ordonnées) en fonction de celui réalisé pour définir l'état de référence (abscisses), pour détecter l'alternative δ (isoplèthes .15, .20, ..., .60) avec des risques fixés a priori (α=β=.05, .10, .15). Deux valeurs de $\hat{CV}_1$ sont considérées : .30 (à gauche), et .50 (à droite).

1.2.2. Hétérogénéité des variances induite par la loi de Taylor : remèdes

Outre le fait que les lois de X_1 et X_2 doivent être proches de la normalité, le développement s'est jusqu'à présent appuyé sur l'hypothèse de la stabilité de la variance ($\sigma_1^2=\sigma_2^2=\sigma^2$). Si cette condition n'est plus respectée, c'est-à-dire si :

$$\sigma_1^2/\sigma_2^2=\theta, \quad \text{avec } \theta \neq 1$$

alors : $V((\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)) = \sigma_1^2/n_1+\sigma_2^2/n_2$

Et la variable $X^\dagger$ qui suit une loi N(0,1) s'exprime :

$$X^\dagger = ((\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2))/\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$$

$$= ((\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2))/\sigma_1\sqrt{1/n_1+1/(\theta.n_2)}$$

rapport au dénominateur duquel apparaît le paramètre θ.

Il est classiquement reconnu que pour de nombreux descripteurs écologiques, la dispersion croît en même temps que la tendance centrale (*cf.* par exemple FRONTIER, 1973). Cette dépendance stochastique entre la variance et la moyenne est généralement bien décrite par la loi de Taylor :

$\sigma^2=a\mu^b$, où a et b sont des constantes réelles, les valeurs de b les plus communément rencontrées se situant entre 1.5 et 2.5 (FRONTIER, *op. cit.*; DOWNING et al., 1987). Admettant cette loi, θ s'exprime :

$$\theta = \mu_1^b/(\mu_1-\Delta)^b$$

Soit encore :

$$\delta=\Delta/\mu_1 \quad , \quad 0\leq\delta\leq 1 \Rightarrow \theta = (1-\delta)^{-b}$$

Il sera de plus admis, comme précédemment, que la loi de la variable X^2

$$X^2 = (n_1-1)s_1^2/\sigma_1^2$$

peut être approchée par un χ^2 à n_1-1 d.d.l. Dans ces conditions, la quantité z dont la loi est un t de Student à $\nu = n_1-1$ d.d.l. s'écrit :

$$z = ((\bar{x}_1-\bar{x}_2)-\delta\mu_1)/(s_1\sqrt{1/n_1+(1-\delta)^b/n_2})$$

La vraie valeur de b n'est en général pas connue, aussi est-il nécessaire de l'estimer à partir des couples de moyennes et variances empiriques $(\bar{x}, s^2)$: cela attache à l'estimation $\hat{b}$ de b une incertitude, quantifiée par la variance d'échantillonnage de l'estimateur $\hat{b}$; en pratique, c'est la valeur estimée de ce dernier qui est introduite dans l'expression z. La statistique z est donc en toute rigueur une fonction des variables aléatoires $\bar{x}_1$, $\bar{x}_2$, s_1 et $\hat{b}$. Par souci de simplicité, il sera toutefois considéré par la suite que la loi de z ne dépend que de celles de $\bar{x}_1$, $\bar{x}_2$ et s_1, autrement dit que l'exposant b est connu sans erreur.

Sous H_o (*i.e.*, $\Delta = 0$), z est identique à la statistique y définie en {1} :

$$z = y = (\bar{x}_1-\bar{x}_2)/(s_1\sqrt{1/n_1+1/n_2})$$

et de même : $\alpha = \text{Proba}\{ y > t_{\nu,\alpha} \mid \delta=0\}$

En revanche, sous H_1 ($0<\delta\leq 1$), l'évaluation de la probabilité β fait apparaître la quantité z, et donc le rapport θ des variances ;

$$\beta(H_1) = \text{Proba}\{ y \leq t_{\nu,\alpha} \mid 0<\delta\leq 1\}$$

$$= \text{Proba}\{ z \leq t_{\nu,\alpha}+z-y \mid 0<\delta\leq 1\}$$

La variable de z obéissant à une loi de Student centrée, l'égalité ci-dessus implique :

$$t_{\nu,1-\beta} = t_{\nu,\alpha}+z-y$$

avec, en posant $\hat{\Delta} = \bar{x}_1-\bar{x}_2$:

$$z-y = ((\hat{\Delta}-\Delta)\sqrt{n_1+n_2}/\sqrt{n_1/\theta+n_2} - \hat{\Delta}) \;/\; (s_1\sqrt{1/n_1+1/n_2}) \qquad \{6\}$$

1.2.2.1. Solution analytique approchée

La différence z - y telle qu'exprimée ci-dessus n'est pas directement utilisable pour répondre soit à la question du δ minimum décelable, soit à celle de l'effort d'échantillonnage à engager pour la période de surveillance du milieu en vue d'y détecter un δ donné avec une puissance donnée. Le terme entre crochets sera donc simplifié à cette fin. Pour des valeurs de δ petites devant 1 :

$$\sqrt{n_1/\theta + n_2} = \sqrt{n_1 + n_2 + n_1(1/\theta - 1)}$$

$\delta << 1 \Rightarrow \theta$ voisin de 1, donc $n_1(1/\theta-1) << n_1+n_2$

et :

$$\sqrt{n_1/\theta + n_2} \simeq \sqrt{n_1+n_2} + n_1(1/\theta-1)/(2\sqrt{n_1+n_2})$$

Avec cette approximation, le numérateur N de la fraction :

$$z-y = N/(s_1\sqrt{1/n_1+1/n_2})$$

donnée précédemment par l'équation {6} devient :

$$N = \hat{\Delta}\left((1+(n_1(1/\theta-1))/(2(n_1+n_2)))^{-1} - 1\right) -2\Delta(n_1+n_2)/(n_1(1/\theta+1)+2n_2)$$

Pour les faibles valeurs de δ, le terme en facteur de $\hat{\Delta}$ peut être négligé ; de plus : $1/\theta = (1-\delta)^b \simeq 1-b\delta$

D'où la valeur approchée de N :

$$N \simeq -2\Delta(n_1+n_2) \;/\; (2(n_1+n_2) - n_1 b\delta)$$

Introduite dans l'équation {6}, cette simplification conduit à la relation :

$$t_{\nu,\alpha}+t_{\nu,\beta} \simeq 2\delta\mu_1(n_1+n_2) \;/\; (s_1\sqrt{1/n_1+1/n_2}(2(n_1+n_2)-n_1 b\delta))$$

relation qui permet de répondre à la première question du plus petit δ pouvant être détecté avec des probabilités d'erreur α et β fixées. En remplaçant μ_1 par son estimation $\bar{x}_1$, et en faisant apparaître le coefficient de variation $\hat{CV}_1$,

il vient :

$$\delta \simeq \frac{2.\hat{CV}_1\sqrt{1/n_1+1/n_2}(n_1+n_2)(t_{\nu,\alpha}+t_{\nu,\beta})}{2(n_1+n_2) + n_1.b.\hat{CV}_1\sqrt{1/n_1+1/n_2}(t_{\nu,\alpha}+t_{\nu,\beta})} \qquad \{7\}$$

L'examen de cette formule soulève plusieurs remarques :

- Lorsque b = 0, *i.e.* lorsque $\sigma_1^2 = \sigma_2^2 = \sigma^2$, la valeur de δ est alors identique à celle donnée par l'équation {3}.

- De même, lorsque l'effort d'échantillonnage alloué à la période de surveillance du milieu augmente indéfiniment, la valeur de δ tend vers une limite δ_L conditionnée par l'information acquise durant la période qui a précédé la mise en service de l'aménagement :

$$n_2 \to \infty \Rightarrow \delta \to \delta_L = \hat{CV}_1(t_{\nu,\alpha}+t_{\nu,\beta})/\sqrt{n_1} \qquad \{8\}$$

- La valeur de δ_L montre que l'influence de l'exposant b de la loi de Taylor s'estompe quand n_2 croît. Pour les faibles valeurs de n_2, l'effet de b dépend de son signe : si b > 0, c'est-à-dire si la dispersion σ_2 autour de μ_2 ($\mu_2 < \mu_1$) est plus faible que dans le cas où la variance serait constante, le δ minimum décelable est lui aussi plus faible. Toujours face à l'alternative unilatérale $\mu_1 - \mu_2 = \Delta > 0$, la résolution de la procédure est évidemment détériorée si au contraire b est négatif, et donc que $\sigma_2^2 > \sigma_1^2$.

Enfin, les simplifications qui ont permis d'expliciter δ en fonction de n_2, α et β supposent δ petit. Il ne pourra donc pas être envisagé une aussi large gamme d'alternatives que dans le cas où la variance est stable.

Il reste à répondre à la seconde question du nombre n_2 d'observations à réaliser après mise en service de la centrale, *i.e.* exprimer n_2 en fonction de δ, α et β fixés. Partant de :

$$(t_{\nu,\alpha}+t_{\nu,\beta})^2 \simeq \delta^2/(\hat{CV}_1^2(1/n_1+1/n_2)(1-n_1b\delta/(2(n_1+n_2)))^2)$$

et négligeant le terme en δ^2 dans le développement du carré qui apparaît au dénominateur :

$$(t_{\nu,\alpha}+t_{\nu,\beta})^2 \simeq n_1n_2\delta^2/(\hat{CV}_1^2(n_1(1-b\delta)+n_2))$$

D'où la valeur approchée de n_2 :

$$n_2 \simeq n_1(1-b\delta)\hat{CV}_1^2(t_{\nu,\alpha}+t_{\nu,\beta})^2/(n_1\delta^2-\hat{CV}_1^2(t_{\nu,\alpha}+t_{\nu,\beta})^2) \qquad \{9\}$$

avec $n_2>0$ si $\delta_L<\delta<\min(1,1/b)$ pour $b>0$

La structure de la formule {9} appelle les mêmes remarques que celles énoncées à propos de l'expression de δ (formule {7}).

1.2.2.2. Transformation des données

Les méthodes proposées ci-dessus en vue de pallier l'hétérogénéité des variances reposent sur une approche analytique, qui consiste à introduire le rapport θ dans la relation entre $\alpha,\beta(H_1),\delta$ et le nombre d'observations, cette relation étant établie pour les valeurs brutes des réalisations de X_1 et X_2. Toutefois, cette démarche n'est pas la seule possible : en effet, les variances peuvent être stabilisées par une transformation des données. Du point de vue le plus général, si la relation entre moyenne et variance de la variable aléatoire X est une fonction de la forme :

$$V(X) = \psi(E(X))$$

le corollaire d'un théorème dû à RAO énonce que la variable Y définie par :

$$Y = C\int_K^X dt/\sqrt{\psi(t)}$$

où $0<X<\infty$, et où K est une constante arbitraire, possède une variance asymptotiquement stabilisée à C^2. De plus, une telle transformation appliquée à X a non seulement pour effet de rendre V(Y) indépendante de E(Y), mais aussi le plus souvent de conférer à Y une distribution voisine de la normalité, par un phénomène de réduction d'asymétrie.

Il a été indiqué que pour de nombreux descripteurs écologiques, la relation entre σ^2 et μ peut être décrite par la loi de Taylor, modèle dans lequel ψ est une fonction puissance possédant un exposant b proche de 2. Dans les cas où b = 2 exactement, la transformation adéquate de la variable (strictement positive) X est la transformation logarithmique : Y = ln(X). Si la

relation n'est pas exactement quadratique, en toute rigueur doit être utilisée la transformation $Z = X^{1-b/2}$. Pour les valeurs de b peu différentes de 2, cette transformation a pour effet "d'écraser" les fortes valeurs de la variable X, *i.e.* d'atténuer la dissymétrie des distributions communément rencontrées. De ce point de vue, les conséquences des transformations logarithme et puissance sont assez semblables. Cela explique pourquoi, en l'absence d'une estimation précise de b, le recours systématique à la transformation $Y = \ln(X)$ constitue un choix raisonnable, bien qu'arbitraire. La prise en compte de cette option dans la procédure de décision reposant sur la statistique {1} sera donc examinée en premier.

Soient donc X_1 et X_2 définies comme précédemment, mais sous des hypothèses plus faibles en ce sens qu'il ne leur est pas imposé de suivre les lois voisines de la normalité. De même :

$$E(X_1) = \mu_1, \quad V(X_1) = \sigma_1^2 = a\mu_1^b$$

$$E(X_2) = \mu_2(1-\delta), \quad 0<\delta\leq 1, \quad V(X_2) = \sigma_2^2 = a(\mu_1(1-\delta))^b$$

Et soient $Y_1 = \ln(X_1)$, $Y_2 = \ln(X_2)$

L'exposant b étant voisin de 2, il sera considéré que les variables Y_1 et Y_2 ont même variance, et une distribution approximativement normale. Dans ces conditions, la décision en faveur de H_o ou bien de H_1 repose sur la valeur prise par la statistique {1}, qui vaut maintenant :

$$(\bar{y}_1-\bar{y}_2)/\sqrt{(1/n_1+1/n_2)\hat{V}(Y_1)}$$

où $\bar{y}_1$, $\bar{y}_2$ et $\hat{V}(Y_1)$ désignent les estimations non biaisées de $E(Y_1)$, $E(Y_2)$ et $V(Y_1)$ respectivement, calculées sur les données transformées $y_i = \ln(x_i)$. De même, l'équation {2} devient :

$$t_{\nu,\alpha}+t_{\nu,\beta}=(E(Y_1)-E(Y_2))/\sqrt{(1/n_1+1/n_2)\hat{V}(Y_1)}$$

Exprimer $E(Y_1)$ et $E(Y_2)$ en fonction des deux premiers moments centrés de X_1 et X_2 respectivement permet de faire apparaître la proportion δ, proportion dont est réduite sous H_1 la moyenne μ_1 des données non transformées. Pour cela, $\ln(X_1)$ et $\ln(X_2)$ sont développés jusqu'à l'ordre 2, au

voisinage de μ_1 et μ_2 respectivement. En passant ensuite aux espérances, il vient :

$$E(Y_1) \simeq \ln(\mu_1) - \sigma_1^2/(2\mu_1^2)$$

$$E(Y_2) \simeq \ln(\mu_1) + \ln(1-\delta) - \sigma_2^2/(2(\mu_1(1-\delta))^2)$$

Sachant que : $\sigma^2 = a\mu^b$

$$E(Y_1) - E(Y_2) \simeq -\ln(1-\delta) + \frac{a}{2}\mu_1^{b-2}((1-\delta)^{b-2} - 1) \qquad \{10\}$$

Premier cas : b=2

Il vient immédiatement : $E(Y_1)-E(Y_2) \simeq \ln(1/(1-\delta))$

D'où : $\delta \simeq 1 - \exp\{-(t_{\nu,\alpha}+t_{\nu,\beta})\sqrt{(1/n_1+1/n_2)\hat{V}(Y_1)}\}$ $\{11\}$

Et comme pour les autres configurations envisagées jusqu'à présent sont également obtenus les résultats suivants :

$$\delta_L \simeq 1 - \exp\{-(t_{\nu,\alpha}+t_{\nu,\beta})\sqrt{\hat{V}(Y_1)/n_1}\} \qquad \{12\}$$

$$n_2 \simeq n_1\hat{V}(Y_1)(t_{\nu,\alpha}+t_{\nu,\beta})^2/\ (n_1(\ln(1-\delta))^2-\hat{V}(Y_1)(t_{\nu,\alpha}+t_{\nu,\beta})^2) \qquad \{13\}$$

avec $n_2>0$ si $\delta_L<\delta<1$

L'étroite parenté entre ces formules et celles étudiées au §1.2.1. apparaît encore plus clairement en exprimant $V(Y_1)$ en fonction de la moyenne et de la variance de la variable non transformée X_1. En négligeant le biais et les termes d'ordre supérieur à 1 dans le développement de $\ln(X_1)$, il vient :

$$V(\ln(X_1)) \simeq E((\ln(X_1) - \ln(\mu_1))^2) \simeq V(X_1)/\mu_1^2$$

c'est-à-dire : $\hat{V}(Y_1) \simeq \hat{CV}^2(X_1)$, noté $\hat{CV}_1^2$ dans {3}, {4} et {5}.

En particulier, cela entraîne que pour les valeurs de δ voisines de zéro, *i.e.* celles pour lesquelles l'effet de la transformation est peu sensible, la formule {13} est alors équivalente à la formule {5} ; en effet, dans ce cas : $(\ln(1 - \delta))^2 \simeq \delta^2$.

Second cas : $b \neq 2$

Les résultats {11}, {12} et {13} ne constituent plus qu'une approximation dont la qualité se détériore lorsque b s'éloigne de 2 ; cela pour deux raisons :

- le terme $(1 - \delta)^{b-2} - 1$ devient non nul dans {10}, de l'ordre de $(2-b)\delta$,
- et en outre la transformation $Y = \ln(X)$ n'est plus dans ce cas la "meilleure" transformation stabilisante.

Or, ainsi qu'il a été rappelé, c'est la transformation logarithmique qui est employée "par précaution" quand les paramètres a et b de la loi de Taylor ne sont pas estimés. Dans ce cas, sachant que la valeur de b n'est pratiquement jamais extérieure à l'intervalle $[1,3]$, les résultats précédents peuvent être appliqués pour les petites valeurs de δ, d'autant qu'alors la différence entre μ_1 et μ_2 est faible, et donc que l'effet de la variation de dispersion due à la relation moyenne - variance n'est que très peu sensible.

Néanmoins, il n'est pas sans intérêt de considérer la situation où est connue une estimation de b différant significativement de 2. Soit donc Z la variable aléatoire :

$$Z = X^{1-b/2} \quad , \quad E(X) = a\, V^b(X) \quad , \quad b \neq 2$$

En admettant que la stabilisation de V(X) s'accompagne de surcroît d'un "effet normalisant", l'équation {2} devient :

$$t_{\nu,\alpha} + t_{\nu,\beta} = (E(Z_1) - E(Z_2)) / \sqrt{(1/n_1 + 1/n_2)\hat{V}(Z_1)} \qquad \{14\}$$

Afin d'alléger l'écriture, soit k = 1-b/2 ; avec cette notation, l'espérance de la variable transformée Z peut s'exprimer en fonction de l'espérance μ de la variable originale X :

$$E(Z) = E(X^k) \simeq \mu^k + ak(k-1)/(2\mu^k)$$

D'où :

$$E(Z_1)-E(Z_2) \simeq \mu_1^k(1-(1-\delta)^k) + ak(k-1)(1-(1-\delta)^{-k})/(2\mu_1^k)$$

En ne conservant que le premier terme du second membre, puis en l'introduisant dans {14} :

$$(1-\delta)^k \simeq 1-(t_{\nu,\alpha}+t_{\nu,\beta})\sqrt{(1/n_1+1/n_2)\hat{V}(Z_1)/\mu_1^{2k}}$$

Sachant que la variance de Z est stabilisée à la valeur ak^2, et que le carré du coefficient de variation $CV^2(X_1)$ de la variable non transformée vaut ici a/μ_1^{2k}, l'équation précédente peut s'écrire :

$$(1-\delta)^k \simeq 1 - (t_{\nu,\alpha}+t_{\nu,\beta})\hat{CV}_1\sqrt{k^2(1/n_1+1/n_2)}$$

Soit encore, en revenant aux notations initiales :

$$\delta \simeq 1 - (1-\hat{CV}_1(t_{\nu,\alpha}+t_{\nu,\beta})(1-b/2)\sqrt{(1/n_1+1/n_2)})^{1/(1-b/2)} \qquad \{15\}$$

L'écart relatif δ admet évidemment une limite δ_L non nulle quand $n_2 \to \infty$, et pour $\delta > \delta_L$:

$$n_2 \simeq \frac{n_1(\hat{CV}_1(1-b/2)(t_{\nu,\alpha}+t_{\nu,\beta}))^2}{n_1(1-(1-\delta)^{1-b/2})^2 - (\hat{CV}_1(1-b/2)(t_{\nu,\alpha}+t_{\nu,\beta}))^2} \qquad \{16\}$$

Il est bon de vérifier que l'expression ci-dessus devient équivalente à la formule {13} lorsque b se rapproche de la valeur 2. En posant :

$$A = \hat{CV}_1^2\,(t_{\nu,\alpha}+t_{\nu,\beta})^2$$

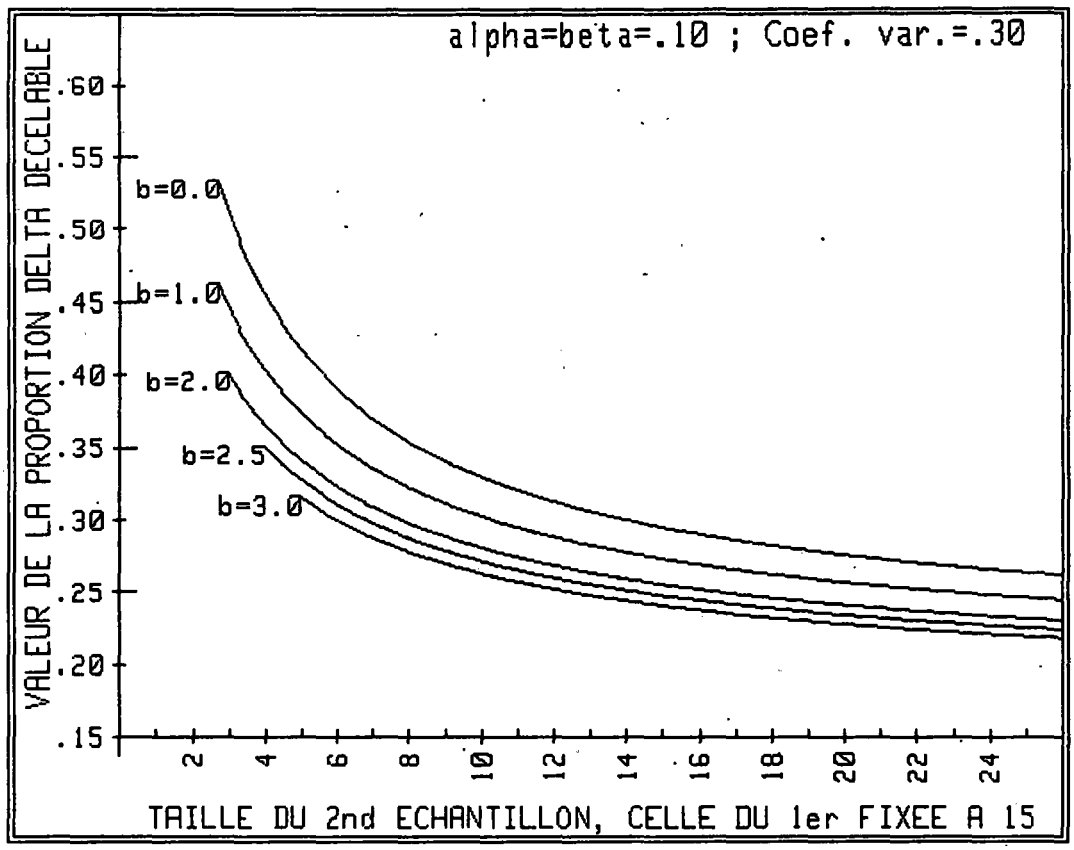


Figure 3 - Influence de l'exposant b de la loi de Taylor sur la valeur de la proportion δ décelable avec des probabilités d'erreur α et β fixées à .10, pour $\widehat{CV}_1 = .30$ et $n_1 = 15$.

Les graphes sont ceux de la fonction $\delta = f(n_2)$ définie par l'équation {15} ; b=0 correspondant au cas décrit par l'équation {3}, et b=2 à celui décrit par l'équation {11}.

et en reprenant la notation $k = 1-b/2$, n_2 s'écrit alors :

$$n_2 \simeq n_1.A \;/\; (n_1((1-(1-\delta)^k)/k)^2-A)$$

Sachant que : $(1-\delta)^k = \exp(k.\ln(1-\delta))$

et que : $b \to 2 \Rightarrow k \to 0$

$\Rightarrow k.\ln(1-\delta) \to 0$ pour tout δ fixé, $0<\delta<1$

Dans ces conditions : $\exp\{k.\ln(1-\delta)\} \simeq 1 + k.\ln(1-\delta)$

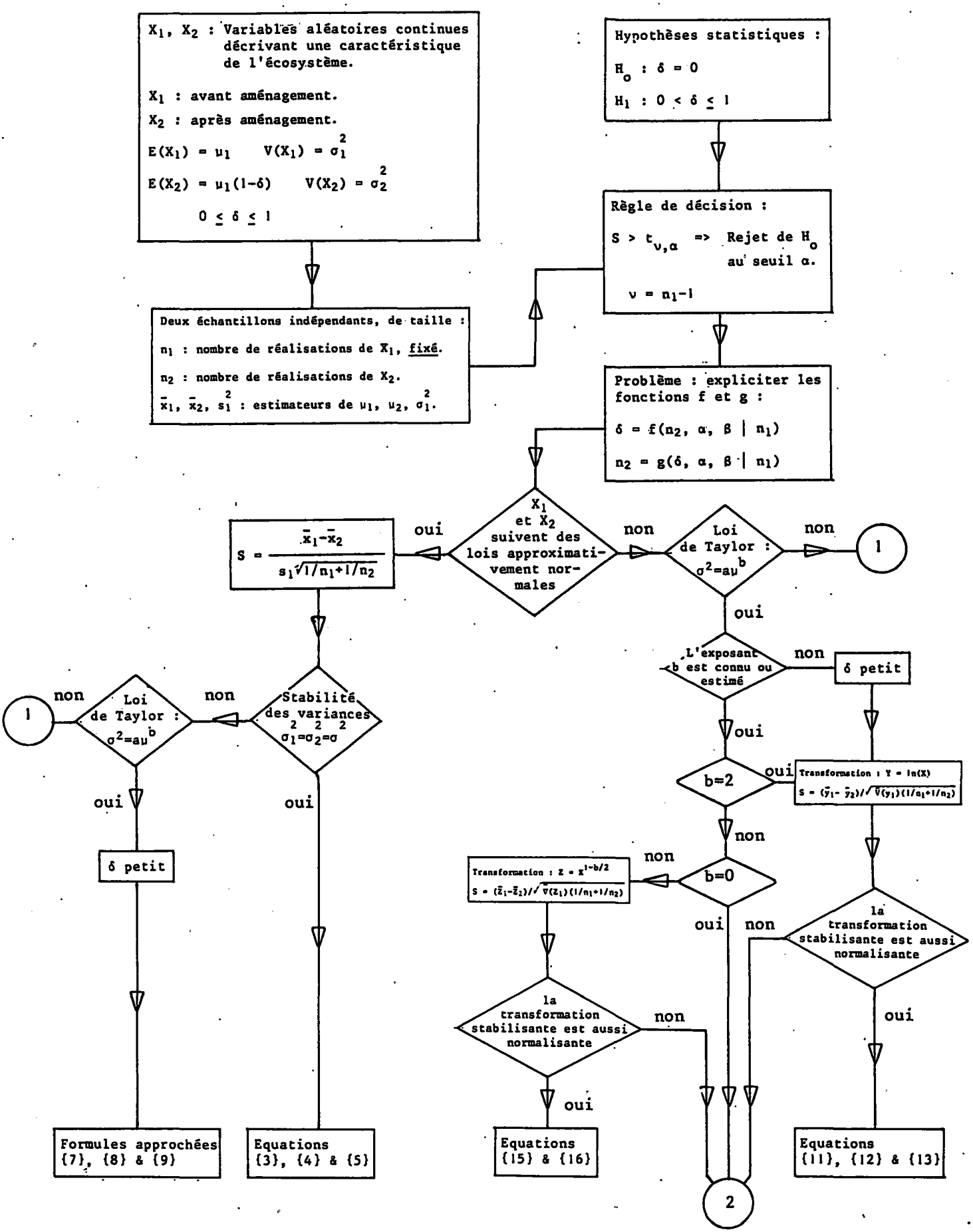
et : $$\lim_{k\to 0} \frac{1-(1-\delta)^k}{k} = -\ln(1-\delta)$$

résultat qui établit l'équivalence des formules {13} et {16} au voisinage de b=2. De même est-il possible de vérifier l'identité des formules {9} et {16} pour les valeurs de δ voisines de zéro. Dans ces conditions en effet, le terme $(1-(1-\delta)^k)^2$ est équivalent à $k^2\delta^2$.

1.3. RECAPITULATION ET DISCUSSION

L'organigramme présenté ci-après rassemble les divers critères conduisant au choix de l'une des expressions établies aux paragraphes 1.2.1. et 1.2.2. Cet arbre de décision, outre qu'il résume les résultats obtenus, met aussi en lumière les limites au-delà desquelles leur validité n'est pas établie. De la sorte apparaissent deux principaux verrous ; un troisième, évoqué à part, est en fait implicitement contenu dans celui noté (2) au bas de l'organigramme.

(i) Le premier tient à la formulation de la relation entre moyenne et variance (issue notée (1)) ; il est évidemment nécessaire de se référer à un modèle, et seule la loi de Taylor a été envisagée ici. Hors de ce cadre d'application très générale, les calculs doivent être adaptés au modèle *ad hoc* qui serait retenu pour ψ (il est rappelé que la fonction ψ est définie par $\sigma^2 = \psi(\mu)$). Il convient en effet de souligner que les solutions proposées dans le cadre de la loi de Taylor ne sont qu'une facette des investigations possibles :


X_1, X_2 : Variables aléatoires continues décrivant une caractéristique de l'écosystème.
X_1 : avant aménagement.
X_2 : après aménagement.
$E(X_1) = \mu_1$ $V(X_1) = \sigma_1^2$
$E(X_2) = \mu_1(1-\delta)$ $V(X_2) = \sigma_2^2$
$0 \leq \delta \leq 1$
Hypothèses statistiques :
$H_o : \delta = 0$
$H_1 : 0 < \delta \leq 1$
Règle de décision :
$S > t_{\nu,\alpha}$ => Rejet de H_o au seuil α.
$\nu = n_1 - 1$
Deux échantillons indépendants, de taille :
n_1 : nombre de réalisations de X_1, fixé.
n_2 : nombre de réalisations de X_2.
$\bar{x}_1$, $\bar{x}_2$, s_1^2 : estimateurs de μ_1, μ_2, σ_1^2.
Problème : expliciter les fonctions f et g :
$\delta = f(n_2, \alpha, \beta \mid n_1)$
$n_2 = g(\delta, \alpha, \beta \mid n_1)$
$S = \frac{\bar{x}_1 - \bar{x}_2}{s_1\sqrt{1/n_1 + 1/n_2}}$
oui
X_1 et X_2 suivent des lois approximativement normales
non
Loi de Taylor : $\sigma^2 = a\mu^b$
non
1
oui
L'exposant b est connu ou estimé
non
δ petit
oui
b=2
oui
Transformation : $Y = \ln(X)$
$S = (\bar{y}_1 - \bar{y}_2)/\sqrt{\hat{V}(y_1)(1/n_1 + 1/n_2)}$
non
b=0
non
Transformation : $Z = x^{1-b/2}$
$S = (\bar{z}_1 - \bar{z}_2)/\sqrt{\hat{V}(z_1)(1/n_1 + 1/n_2)}$
la transformation stabilisante est aussi normalisante
non
oui
Equations {15} & {16}
la transformation stabilisante est aussi normalisante
oui
non
Equations {11}, {12} & {13}
2
Stabilité des variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$
non
Loi de Taylor : $\sigma^2 = a\mu^b$
non
1
oui
δ petit
oui
Formules approchées {7}, {8} & {9}
oui
Equations {3}, {4} & {5}

le développement qui a été présenté repose sur une hypothèse de travail, à savoir le caractère entièrement stable et déterministe des paramètres a et b. Le modèle choisi exprime la relation stochastique entre la variance d'échantillonnage et la moyenne locale ; et dans la comparaison entre deux situations (avant-après), seule la variation de la moyenne (qui influence la variance) a été considérée, exclusivement d'une variation éventuelle de la nature de la relation. Cette option simplificatrice mérite d'être brièvement discutée.

Un impact possible est l'accroissement de l'hétérogénéité du milieu, qui devient une mosaïque de petits domaines sur l'ensemble desquels l'impact doit être apprécié en moyenne. *A contrario*, l'effet peut aussi être une réduction de l'hétérogénéité initiale. Dans un cas comme dans l'autre, l'augmentation ou bien la décroissance de la variance ne sont pas seulement dues à une variation de la moyenne. L'évocation de ces éventualités révèle les limites d'une démarche fondée sur la comparaison de valeurs centrales, et rappelle que des options distinctes (*e.g.*, une étude structurale) peuvent s'avérer tout autant fructueuses.

(ii) La seconde limitation tient aux hypothèses formulées relativement à la loi de la variable X, ou des variables transformées Y et Z : elle a été supposée normale. En effet, le point essentiel de la définition de la loi de Student (sur laquelle reposent les calculs des probabilités α et β) est l'indépendance entre le numérateur et le dénominateur, qui n'est exactement acquise que dans un contexte gaussien. Des résultats asymptotiques garantissent néanmoins qu'hors de ce contexte la distribution de la statistique $(\bar{x}-\mu)\sqrt{n}/s$ tend vers la normalité, et en principe d'autant plus rapidement que la loi de X est symétrique. Il demeure que le nombre d'observations ne légitime pas toujours le recours au comportement asymptotique de la statistique du test, et que le problème peut concrètement se poser au niveau de l'issue notée (2) sur l'organigramme : cette issue conduit à des variables aléatoires non gaussiennes, mais dont la dispersion a été stabilisée par une transformation adéquate.

Cette limitation met en exergue le fait que la conséquence prioritairement attendue d'une transformation des données est en général l'effet normalisant. Cette hiérarchie s'est trouvée ici quelque peu occultée, dans la me-

sure où la normalité était supposée *a priori* acquise, et où la difficulté envisageable était l'hétéroscédasticité. Il ne faut néanmoins pas perdre de vue qu'en dehors du cadre gaussien imposé dans ce chapitre, l'obtention de variables aléatoires normales constitue la motivation première du recours à une transformation des données. La démarche diffère alors sensiblement de celle envisagée au §1.2.2.2. (voir à ce sujet EFRON, 1982) ; elle nécessite au surplus le respect de certaines conditions : il serait aventureux de considérer que tout problème peut être résolu par le simple jeu d'une transformation.

(iii) Les lois de type delta demeureront réfractaires à toute tentative de normalisation de la distribution par une transformation des données. Il est rappelé que ces lois peuvent être considérées comme résultant d'un mélange à poids positifs τ et $1-\tau$.

$$\begin{aligned} &\text{Proba}(X<t)=0 && \text{si } t<0 \\ &\text{Proba}(X=0)=1-\tau && \text{avec } 0\leq\tau\leq 1 \\ &\text{Proba}(X<t)=\tau.F(t) && \text{si } t>0 \end{aligned}$$

où F est une fonction de répartition quelconque sur $]0, +\infty]$. En d'autres termes, les observations issues d'une loi de type delta contiennent, outre des valeurs strictement positives, une certaine proportion de valeurs nulles.

L'élargissement des propriétés envisageables pour X suggère de faire appel à des techniques s'appuyant sur un corps d'hypothèses moins restrictif. L'application d'une procédure dite non paramétrique sera par conséquent examinée ; elle ne sera toutefois pas plus efficace qu'une procédure paramétrique face à des distributions de type delta.

Auparavant, et concernant la mise en pratique des résultats jusqu'ici établis, une remarque importante doit être introduite. De l'examen des formules mentionnées sur l'organigramme, il ressort immédiatement que la seule information d'origine "expérimentale" qui est utilisée est fournie par les n_1 réali-

sations de X_1. Cela définit précisément le champ d'application de ces procédures de test dans la chronologie des études d'impact : elles sont le moyen de prendre une décision à l'échéance de la période au cours de laquelle est défini "l'état de référence", et avant que ne commencent les études dites "de surveillance". En ce sens, la question prioritaire est l'explicitation de la fonction g (question (2), §1.2.1.) :

$$n_2 = g\ (\Delta,\alpha,\beta|n_1)$$

En d'autres termes, le traitement de l'information acquise préalablement à la mise en service de la centrale sert à évaluer le nombre d'observations à réaliser pendant le fonctionnement de celle-ci, de sorte que soit garanti, avec des risques connus, un pouvoir de résolution donné de la procédure visant à déceler un éventuel impact.

De ce point de vue, la question complémentaire (question(1), déterminer f telle que $\Delta = f\ (n_2,\alpha,\beta|n_1)$), ne présente qu'un moindre intérêt pratique. En revanche, cette interrogation acquiert toute sa pertinence si elle est formulée non plus au début de la période de surveillance, mais à la fin de celle-ci : elle conduit alors à établir un bilan à l'issue des études, et à traiter le cas non envisagé par les questions (1) et (2) adressées au début du §1.2.1., *i.e.* résoudre le problème suivant :

(3) Disposant de n_1 réalisations de X_1 (faisant partie intégrante de la définition de l'état de référence), ainsi que n_2 réalisations de X_2 (acquises durant la période de surveillance), quel est alors le plus petit écart Δ pouvant être détecté avec des probabilités d'erreur α et β ? Le problème est d'exprimer analytiquement la fonction h :

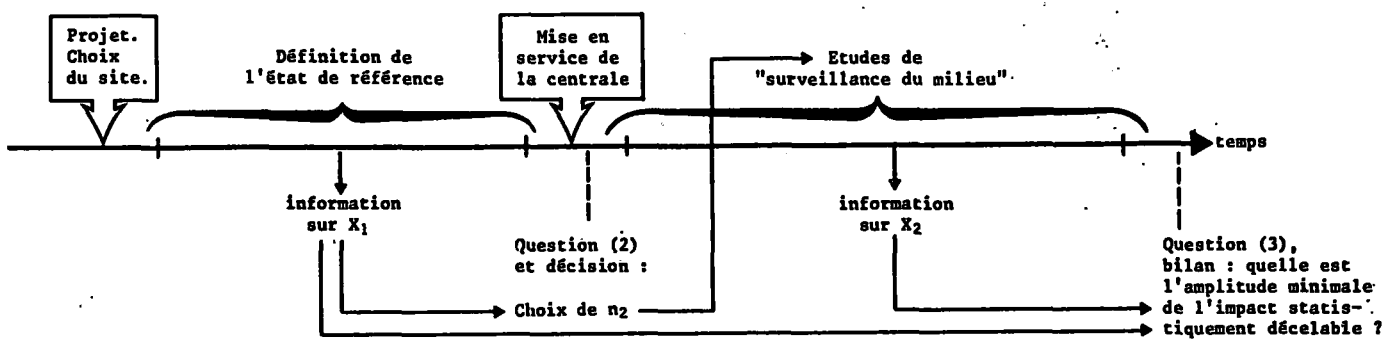
$$\Delta = h\ (\alpha,\ \beta,\ n_1,\ n_2)$$

Les deux échantillons indépendants (de taille n_1 et n_2) jouent maintenant des rôles équivalents[1]. En particulier, la statistique du test intègre

---

(1) S'il est souhaité être parfaitement rigoureux, l'indépendance entre les deux échantillons ne peut pas être totalement admise, dans la mesure où la taille n_2 du second est décidée à partir d'un coefficient de variation estimé d'après le premier. Ce ne serait pas le cas si n_1 et n_2 étaient fixés *a priori* au début de l'étude (*i.e.*, avant la définition de l'état de référence).

l'information obtenue non seulement sur la variable X_1, mais aussi sur la variable X_2, comme le montre le schéma suivant :



Acquisition et utilisation de l'information en vue de la détection statistique des effets d'un aménagement. Exemple : la surveillance du milieu au voisinage d'une centrale nucléaire.

A la différence de la réponse aux questions (1) et (2), la réponse à la question (3) procède de l'utilisation de la statistique classiquement employée pour la comparaison des moyennes de deux échantillons indépendants (et non plus de la statistique modifiée {1}). En particulier, l'estimateur de la variance commune σ^2 n'est plus s_1^2, mais une somme pondérée de s_1^2 et de s_2^2, le nombre de degrés de libertés étant ajusté en conséquence (n_1+n_2-2 au lieu de n_1-1). Les formules {3}, {7}, {11} et {15} peuvent alors aisément être aménagées pour intégrer ces modifications, et permettre ainsi de connaître le plus petit écart Δ statistiquement décelable au terme de la période d'étude de surveillance.

D'un point de vue plus général, il convient de s'attarder ici sur une question de fond. L'inférence statistique est du domaine de l'induction ; et au même titre que toute construction intellectuelle, elle est source de débats. Le lecteur doit donc être averti des controverses relatives aux procédures exposées dans le présent rapport.

Pour s'en tenir simplement aux tests statistiques, il faut d'abord rappeler que la construction d'un test d'hypothèses consiste fondamentalement à isoler de l'univers des résultats possibles un sous-ensemble appelé région critique ; l'appartenance du résultat expérimental à cette région critique entraîne la décision de repousser H_o.

Jusqu'à présent, et il en sera de même par la suite, n'ont été considérés que des tests bâtis selon le principe dit "de Neyman-Pearson". La logique de la définition de la région critique repose dans ce cas sur la maximisation de la puissance $1-\beta$, sous la contrainte que la probabilité que le résultat observé "tombe par hasard", si H_o est vraie, dans la région critique, soit précisément égale à un risque α fixé. Schématiquement, le respect de cette contrainte conduit à la définition d'un seuil, auquel est comparé un rapport de vraisemblances (la vraisemblance sous H_1 de l'échantillon, rapportée à sa vraisemblance sous H_o).

Une autre démarche doit être mentionnée, qui fonde la décision statistique d'une manière différente : la démarche bayésienne. De même que le critère de Neyman-Pearson, le critère de Bayes conduit à asseoir la décision sur une comparaison entre un rapport de vraisemblances et un seuil. La divergence entre les deux approches tient à l'introduction d'information *a priori* dans un test bayésien, sous forme de probabilités *a priori* des hypothèses en présence, ainsi que des coûts attachés à chacune des diverses décisions possibles. La région critique est alors déterminée de façon à minimiser l'espérance du coût entraîné par la décision qui sera prise. Au plan pratique, l'intérêt d'une telle procédure réside dans le fait que ce n'est pas le calcul du rapport des vraisemblances qui est concerné par l'information *a priori*, mais le calcul du seuil auquel ce rapport est comparé. Cela offre la possibilité d'actualiser les conclusions, dans la mesure où elles sont parfois redevables, dans une première étape, d'une information *a priori* acquise "au pifomètre".

CHAPITRE II

LES LIMITES DES PROCEDURES PARAMETRIQUES. PALLIATIFS

2.1. LA NOTION DE ROBUSTESSE

Les propriétés d'optimalité des méthodes paramétriques sont tributaires du respect d'un corps d'hypothèses bien défini ; soit donc M le modèle postulé :

$$M \begin{cases} X \text{ de loi } F ; \\ \text{échantillon : réplications indépendantes du modèle } (X,F). \end{cases}$$

la question qui se pose à ce niveau est de savoir avec quelles garanties peuvent être employées les procédures lorsque les conditions d'application ne sont plus exactement M, mais seulement "voisines" de M. Sachant qu'il y a plusieurs façons de s'éloigner de M : la condition d'indépendance des réplications peut n'être que partiellement respectée, les individus peuvent ne pas être seulement régis par F, mais par un mélange de différents processus ; deux dans le cas le plus simple, quand il y a "contamination" de F par un processus aléatoire Q :

$$X \sim (1-\varepsilon)F+\varepsilon Q \qquad 0<\varepsilon<1$$

Cet exemple permet d'illustrer les conséquences de l'écart à M : la statistique classique estime la dispersion de X par $s_n^2 = \Sigma(x_i-\bar{x})^2/(n-1)$. Un autre indicateur de dispersion possible est :

$$d_n = \Sigma|x_i-\bar{x}|/n$$

Il doit être souligné, d'un point de vue plus global, que la comparaison entre s_n et d_n ne constitue qu'une émanation d'un vaste et toujours actuel débat, articulé autour de la désignation du critère de base : la distance. Dans l'exemple présenté, la question sous-jacente est celle du choix entre l'écart quadratique et l'écart absolu ; autrement dit, entre la "norme L_2" et la "norme L_1". La plupart des tests reposent sur la norme L_2, sans que cette situation soit pour autant exclusive de la recherche des possibilités offertes par la norme L_1.

Dans le cadre du modèle $F = N(\mu,\sigma^2)$, s_n et d_n convergent en probabilité vers σ et $\sigma\sqrt{2/\pi}$ respectivement. S'il y a contamination par $Q = N(\mu,(3\sigma)^2)$, *i.e.* :

$$\text{Proba}(X<t) = (1-\varepsilon)\,\Phi((t-\mu)/\sigma) + \varepsilon\Phi((t-\mu)/(3\sigma))$$

alors d_n est meilleure que s_n pour $.002<\varepsilon<.5$, au sens du critère suivant : l'efficacité asymptotique relative ARE,

$$\text{ARE} = \lim_{n \to \infty} \{ (V(s_n)/E^2(s_n)) \,/\, (V(d_n)/E^2(d_n)) \}$$

devient supérieure à 1 dès que 2 observations sur 1000 sont "aberrantes", c'est-à-dire qu'elles proviennent de $N(\mu,(3\sigma)^2)$ et non du modèle $N(\mu,\sigma^2)$ (exemple emprunté à TUKEY, *in* LECOUTRE *et al.*, 1986). Ces résultats amènent à considérer que l'estimateur de dispersion d_n est plus "robuste" que s_n, en ce qu'il présente une plus forte stabilité dans un voisinage du corps d'hypothèses M. Plus généralement, une procédure (un estimateur, un test...) est dite robuste si ses propriétés ne perdent que peu de leur optimalité hors du contexte strictement défini par M. Par exemple : une procédure qui conserve une grande efficacité relative par rapport à une procédure optimale sous M, un estimateur de faible variance asymptotique sur un voisinage de M, une statistique de loi stable autour de M. Quelques autres propositions de définition de la robustesse (terme introduit en 1953 par BOX) sont présentées par HUBER (1972), qui résume ainsi la conception qu'il nourrit de cette notion : "... I am inclined to view robustness as a kind of insurance problem : I am willing to pay a premium (a loss of efficiency of, say, 5 to 10 % at the ideal model) to safeguard against ill effects caused by small deviations from it ; although I am happy if the procedure performs well also under large deviations, I do not really care - inferences based upon a grossly wrong statistical model may have little physical significance". (*op.cit.*, p. 1047).

Dans cet esprit, il est reconnu depuis les années soixante que certains des tests statistiques les plus classiques acquièrent un comportement instable en réponse à de faibles changements de la distribution parente postulée. Concrètement, une violation du modèle M se traduit par une altération de la loi de la statistique du test, qui change les probabilités d'erreur de pre-

mière et seconde espèces. A cet égard, le test t est réputé relativement robuste ; cette appréciation générale doit néanmoins être nuancée : le test t présente une robustesse modérée face à un écart à la normalité (à la différence, par exemple, d'un test de χ^2), en ce sens que le seuil effectif est peu différent du seuil nominal α, mais cela n'est plus vrai pour la puissance. Au surplus, le test t pour deux échantillons indépendants est particulièrement sensible à l'hétérogénéité des variances, et d'autant plus que les tailles d'échantillon n_1 et n_2 diffèrent entre elles (alors que pour $n_1 = n_2$, l'hétéroscédastiticité ne détériore pratiquement pas les qualités du test). Cela explique l'attention qui a été précédemment portée à la stabilisation des variances. Le lecteur trouvera dans BRADLEY (1968, chap. 2) une étude des variations du rapport seuil nominal/seuil effectif pour différents tests paramétriques usuels appliqués après dégradation des hypothèses de base. Il peut aussi être souligné que JOHNSON (1978) a proposé un aménagement de la statistique t en vue de la rendre moins sensible à une asymétrie de la loi de la population parente.

La relative fragilité des propriétés des tests paramétriques justifie l'intérêt suscité par les tests dits "non paramétriques" (encore appelés, bien que les deux dénominations ne soient pas exactement équivalentes, "distribution-free procedures"), souvent établis sur la base de statistiques de rangs, et qui pour nombre d'entre eux possèdent une bonne robustesse. Mais comme le souligne HUBER (1972), sans doute ne s'agit-il là que d'un heureux concours de circonstances ; la plus large définition admise pour la loi des variables échantillonnées (de nombreux tests non paramétriques la supposent simplement continue) stabilise le seuil, mais pas nécessairement la puissance. En outre, les développements qui vont suivre montrent que la décision de rejeter ou non H_o peut être prise sous des hypothèses assez lâches, mais qu'en revanche le calcul de β, qui nécessite la connaissance du comportement de la statistique du test sous H_1, implique le recours à un modèle paramétrique précis.

2.2. DETECTION DE L'IMPACT EN TEMPS DIFFERE : APPROCHE NON PARAMETRIQUE

Le problème est maintenant celui de la détection d'un éventuel décalage entre les positions de deux distributions quelconques, mais néanmoins équi-dispersées. Il est classiquement reconnu que le test de Wilcoxon offre, dans l'éventail des techniques non paramétriques, l'un des outils les plus efficaces pour traiter cette question.

2.2.1. Hypothèses statistiques

De façon plus formelle, et en conservant les notations utilisées jusqu'à présent, soient F_1 et F_2 les fonctions de répartition des deux variables aléatoires indépendantes X_1 et X_2 respectivement :

$$X_1 \sim F_1 \qquad i.e., \quad \text{Proba}(X_1 \leq x) = F_1(x)$$
$$X_2 \sim F_2 \qquad \text{Proba}(X_2 \leq x) = F_2(x)$$

Les propriétés requises sont désormais les suivantes :

(1) F_1 et F_2 sont continues,

(2) X_1 et X_2 possèdent d'emblée, ou bien après transformation, la même variance stable ; c'est-à-dire que cette variance est insensible aux changements de position affectant la tendance centrale de X_1 ou de X_2.

Conformément à cette nouvelle formulation, les hypothèses statistiques s'écrivent :

$$H_o : F_1 = F_2$$

i.e., la loi de X_1 n'est pas modifiée par l'aménagement.

A cette hypothèse nulle est opposée, comme précédemment, l'alternative unilatérale selon laquelle les n_1 réalisations de X_1 prennent en probabilité des valeurs supérieures à celles des n_2 réalisations de X_2 ; autrement dit, H_1 stipule que X_2 est stochastiquement inférieure à X_1. La diminution étant imputable, pour le compartiment considéré de l'écosystème, aux effets du fonctionnement de la centrale (par exemple). Cela s'exprime :

$$F_1(x) \leq F_2(x) \qquad \forall x, \text{ inégalité stricte pour un x au moins}$$

Plus précisément, le modèle qui sera ici retenu énonce que la densité de X_2 est translatée vers la gauche d'une quantité constante Δ par rapport à la densité de X_1 ("shift model", ou "slippage alternative", des auteurs anglo-saxons) :

$$H_1 : F_1(x) = F_2(x-\Delta) \qquad \forall\, x, \quad \Delta>0$$

D'où la présentation abrégée des hypothèses statistiques :

$H_o : \Delta=0$ contre $H_1 : \Delta>0$

2.2.2. Puissance du test de Mann-Whitney

La décision procède de la valeur prise par la statistique de Mann-Whitney :

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(x_{1i}, x_{2j})$$

où x_{1i} désigne la i-ième des n_1 réalisations indépendantes de X_1, et x_{2j} la j-ième des n_2 réalisations indépendantes de X_2, et où la fonction ϕ est définie par :

$$\phi(x_{1i}, x_{2j}) = \begin{cases} 1 \text{ si } x_{1i} > x_{2j} \\ 0 \text{ si } x_{1i} < x_{2j} \end{cases}$$

Cela signifie que U est égal au nombre de fois où une réalisation x_2 de X_2 précède une réalisation x_1 de X_1 dans l'interclassement des $n = n_1+n_2$ réalisations de ces deux variables indépendantes. Il est intuitivement aisé de percevoir que plus le nombre U des transpositions est élevé (une transposition étant l'apparition d'un x_2 plus petit qu'un x_1), et moins le modèle défini par H_o est compatible avec les données. De fait, H_o est repoussé dès que U dépasse une valeur critique C_α :

$U \geq C_\alpha$ => rejet de H_o au seuil α ;

La valeur C_α est fournie par les tables de la loi de permutation de U sous H_o, loi qui ne dépend que des tailles d'échantillon n_1 et n_2. Lorsque le nombre total n d'observations croît, il est habituellement recouru à "l'approximation normale", fondée sur le résultat asymptotique suivant :

$(U - E(U))/\sqrt{V(U)} \sim N(0,1)$, asymptotiquement

Ce résultat, établi par MANN et WHITNEY, vaut aussi bien pour $F_1=F_2$ que pour $F_1 \neq F_2$. Ces deux auteurs ont en outre donné les expressions des deux premiers moments de la statistique U :

$$E(U) = n_1 n_2 E(F_2(X_1))$$

$$V(U) = n_1 n_2 \{ E(F_2(X_1))E(F_1(X_2)) + (n_1-1)V(F_1(X_2)) + (n_2-1)V(F_2(X_1)) \}$$

Il peut être facilement vérifié que sous H_o :

$$E_o(U) = n_1 n_2/2 \quad , \; V_o(U) = n_1 n_2 (n_1+n_2+1)/12$$

D'où la formulation de la règle de décision avec l'approximation normale :

$$(U-n_1 n_2/2)/\sqrt{n_1 n_2 (n+1)/12} \geq \Phi^{-1}(1-\alpha) \Rightarrow \text{rejet de } H_o \text{ au seuil } \alpha \quad \{17\}$$

où Φ^{-1} désigne la fonction réciproque de la fonction de répartition Φ de l'aléa $N(0,1)$, et où $n = n_1+n_2$. La valeur critique C_α peut donc être approchée par :

$$C_\alpha \simeq n_1 n_2/2 + \sqrt{n_1 n_2 (n+1)/12}.\Phi^{-1}(1-\alpha)$$

Face à l'alternative composite unilatérale H_1 (*i.e.*, $\Delta>0$), la puissance $\pi(\Delta)$ du test s'exprime :

$$\pi(\Delta) = \text{Proba}(U \geq C_\alpha \mid \Delta>0)$$

Et avec l'approximation normale :

$$\pi(\Delta) \simeq 1 - \Phi\{(C_\alpha - 1/2 - E_\Delta(U))/\sqrt{V_\Delta(U)}\}$$

où $E_\Delta(U)$ et $V_\Delta(U)$ sont l'espérance et la variance de U sous H_1. En remplaçant C_α par sa valeur approchée, en négligeant la correction de continuité, et en utilisant la symétrie de la fonction Φ :

$$\pi(\Delta) \simeq \Phi\{ [E_\Delta(U) - n_1 n_2/2 + \sqrt{n_1 n_2 (n+1)/12}.\Phi^{-1}(\alpha)]/\sqrt{V_\Delta(U)} \} \quad \{18\}$$

L'approximation de la puissance $\pi(\Delta)$ nécessite donc l'évaluation de la variance $V_\Delta(U)$. Ce calcul constitue une difficulté pratique, qui peut cependant être tournée en utilisant une simplification suggérée par LEHMANN (1975) : elle consiste à négliger la différence entre $V_\Delta(U)$ et $V_o(U)$, quantités voisines lorsque Δ est petit.

Par ailleurs :

$$E_\Delta(U) = n_1 n_2 E(F_2(X_1)) = n_1 n_2 E(F_1(X_1+\Delta))$$

avec : $F_2(X_1) = \text{Proba}(X_2 < X_1) = \text{Proba}(X_2-(X_1-\Delta) < \Delta)$

où, bien entendu : $X_1-\Delta$, $X_2 \overset{iid}{\sim} F_2$

Soit $F^\dagger$ la fonction de répartition de la différence des deux variables aléatoires $X_1-\Delta$ et X_2 ; d'après ce qui précède :

$$\text{Proba}(X_2 < X_1) = F^\dagger(\Delta)$$

En développant à l'ordre 1 au voisinage de $\Delta=0$:

$$F^\dagger(\Delta) \simeq F^\dagger(0) + \Delta.f^\dagger(0)$$

La fonction de densité $f^\dagger$ étant impaire, il vient :

$$E_\Delta(U) \simeq n_1 n_2(1/2+\Delta.f^\dagger(0))$$

En introduisant ce résultat dans l'équation {18}, en négligeant la différence entre $V_\Delta(U)$ et $V_o(U)$, est alors obtenue une relation simple entre n_1, n_2, Δ et les probabilités d'erreur α et β :

$$\Phi^{-1}(1-\beta) \simeq \sqrt{12n_1n_2/(n_1+n_2+1)}.\Delta.f^\dagger(0) + \Phi^{-1}(\alpha)$$

Cette équation permet de répondre à la question (3) du §1.3. :

$$\Delta \simeq -\sqrt{(n+1)/(12n_1n_2)}(\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))/f^\dagger(0) \qquad \{19\}$$

2.2.3. Limite des approximations : cas où n_1 est fixé

Les approximations précédemment employées (*i.e.*, approximation normale, $V_o(U) \simeq V_\Delta(U)$ pour Δ petit) ont permis d'obtenir une équation pour la résolution du problème (3), symétrique en n_1 et n_2. Il reste à examiner si la même démarche pourrait conduire à un résultat analogue à celui établi dans le contexte paramétrique, et fournir une réponse aux questions (1) et (2). C'est-à-dire traiter le cas où n_1 observations réalisées avant mise en service de la centrale ne se voient pas accorder dans les calculs de même rôle que celui attribué aux n_2 observations à venir.

A cette fin sera utilisé un résultat dû à VAN DER WAERDEN (1967) : pour n_1 fixé et n_2 autorisé à croître indéfiniment, la grandeur U/n_2 tend à se comporter comme une somme de n_1 variables indépendantes distribuées chacune comme $F_2(X_1)$, où X_1 suit la fonction de répartition F_1.

Donc : $E(U/n_2) = n_1.E(F_2(X_1))$

$V(U/n_2) = n_1.V(F_2(X_1))$

Cela entraîne que sous l'hypothèse nulle ($\Delta=0$), la loi de la quantité :

$$(U-n_1n_2/2)/\sqrt{n_1n_2^2/12}$$

tend, en vertu du théorème de la limite centrale, vers la loi de l'aléa $N(0,1)$. La convergence est suffisamment rapide pour autoriser à invoquer cette propriété dès que $n_1>3$ et n_1+n_2 de l'ordre de 20 (VAN DER WAERDEN, *op. cit.*). Sous ces conditions, il est possible de modifier en conséquence la règle de décision {17} ainsi que la formule approchée {18} de la puissance $\pi(\Delta)$; cependant, l'utilisation en ce sens des résultats asymptotiques de VAN DER WAERDEN conduit à une approximation de la puissance qui n'inclut pas la taille d'échantillon n_2. Il n'existe donc pas dans ce cas de possibilité de réponse simple à la question (2) du §1.2.1., de même qu'à la question (1).

CHAPITRE III

L'EVOLUTION DE L'OUTIL STATISTIQUE ECLAIREE PAR UNE ETUDE DE CAS : LA DETECTION DE L'IMPACT A L'AIDE D'OBSERVATIONS SYNCHRONES.

3.1. DEFINITION DU PROBLEME

La stratégie exposée aux paragraphes 1.2 et 2.2 est de nature fondamentalement diachronique. Il est possible de concevoir le problème de la détection de l'impact d'un autre point de vue, en accordant une moindre attention à la dimension temporelle pour privilégier en contrepartie les disparités spatiales. Concrètement, il s'agit alors de mettre en évidence d'éventuelles différences entre caractéristiques de l'écosystème dans le champ proche de l'aménagement d'une part, et d'autre part dans un secteur voisin mais pouvant néanmoins être considéré comme indemne de toute perturbation attribuable à cet aménagement.

Cet objectif impose la conception d'une stratégie sensiblement différente de celle qui fut jusqu'ici présentée. A la notion d'état de référence est substituée celle de zone "témoin", *i.e.* hors d'atteinte d'un éventuel impact, et cela sans restrictions d'ordre chronologique.

Au plan statistique, la principale incidence de cette approche concerne la nature des échantillons d'observations effectuées sur les descripteurs du milieu. Alors qu'il fut auparavant traité d'échantillons indépendants (la taille du premier étant fixée), les données consisteront désormais en une série d'observations appariées : à différents instants, des mesures sont simultanément effectuées dans la zone "témoin" et dans le secteur soumis à l'impact. Et l'accent n'est plus mis sur l'écart entre tendances centrales estimées de deux échantillons indépendants, mais sur la position par rapport à zéro de la tendance centrale de la différence entre observations appariées. Ces aspects vont être maintenant précisés.

3.2. LA SOLUTION CLASSIQUE

Les notations déjà employées seront conservées, leur signification ne recouvrant cependant pas exactement les mêmes entités du fait du changement de stratégie. Ainsi, à la variable aléatoire X est associé un descripteur écologique choisi de telle sorte qu'il puisse être considéré comme sensible aux

conséquences de la mise en place de l'aménagement. Et de même qu'aux paragraphes 1.2. & 2.2. seront distingués, sans que cela aille au-delà d'une simple identité formelle :

X_1, qui fait correspondre au compartiment étudié de l'écosystème les valeurs observées dans la zone "témoin" ;

X_2 : idem, dans le champ proche des rejets.

La base des données rassemble cette fois n couples d'observations (x_{11}, x_{21}), ..., (x_{1n}, x_{2n}), où (x_{1i}, x_{2i}) est la i-ième réalisation du couple aléatoire (X_1, X_2). Pour tenter des inférences à partir de ces observations, il est nécessaire de se munir d'un corps d'hypothèses relatives à la loi de (X_1, X_2). L'exemple très classiquement présenté est celui où ce couple de variables suit une loi normale à deux dimensions :

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left\{\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right\}$$

Le problème sera examiné plus loin dans un contexte beaucoup moins contraignant. Dans l'immédiat, il sera traité dans le cadre gaussien ; soit donc Δ la différence entre les moyennes des 2 variables :

$$\mu_1 - \mu_2 = \Delta$$

Les hypothèses statistiques sont alors définies comme suit :

H_o : pas d'impact apparent, les valeurs prises par X_1 et X_2 sont en moyenne les mêmes ;

H_1 : il existe une différence entre la zone "témoin" et la zone "impactée", qui se manifeste (par exemple) par la tendance que montre la variable X_2 à prendre, en moyenne, des valeurs inférieures à celles observées dans le même temps pour X_1.

Autrement dit, $H_o : \Delta = 0$, contre $H_1 : \Delta > 0$

Pour tester ces hypothèses est créée une nouvelle variable aléatoire :

$$D = X_1 - X_2$$

Eu égard aux hypothèses relatives à la loi du couple (X_1, X_2), la variable D suit évidemment une loi de Laplace-Gauss :

$$D \sim N(\Delta, \sigma_1^2+\sigma_2^2-\rho\sigma_1\sigma_2)$$

et les données consistent quant à elles en n réalisations indépendantes d_i de D :

$$d_i = x_{1i}-x_{2i} \quad , \quad i = 1, \ldots, n$$

La vraie moyenne Δ de ces n réalisations est estimée sans biais par $\bar{d}$:

$$\bar{d} = (1/n)\sum_i d_i \quad , \text{ avec } E(\bar{d}) = \Delta \text{ et } V(\bar{d}) = V(D)/n$$

Par conséquent : $(\bar{d}-\Delta)\sqrt{n}/\sqrt{V(D)} \sim N(0,1)$

Pour tester l'hypothèse $\Delta=0$, il est nécessaire d'estimer V(D) ; cette variance est estimée sans biais par s_d^2 :

$$s_d^2 = \sum_i (d_i-\bar{d})/(n-1)$$

et dans le cas où $\rho=0$, avec $\sigma_1^2+\sigma_2^2$ indépendante de Δ, la quantité $\bar{d}\sqrt{n}/s_d$ suit une loi de Student à $\nu=n-1$ d.d.l. ; de sorte que la règle de décision s'énonce :

$$\bar{d}\sqrt{n}/s_d > t_{\nu,\alpha} \Rightarrow \text{rejet de } H_o \text{ au seuil de } \alpha.$$

A cette règle de décision sont associées les probabilités d'erreur :

$$\alpha = \text{Proba}(\bar{d}\sqrt{n}/s_d > t_{\nu,\alpha} \mid \Delta=0)$$

$$\beta(\Delta) = \text{Proba}(\bar{d}\sqrt{n}/s_d < t_{\nu,\alpha} \mid \Delta>0)$$

en remarquant que la probabilité de commettre l'erreur de deuxième espèce peut encore s'écrire :

$$\beta(\Delta) = \text{Proba}((\bar{d}-\Delta)\sqrt{n}/s_d < t_{\nu,\alpha}-\Delta\sqrt{n}/s_d \mid \Delta>0) \Rightarrow t_{\nu,\alpha}-\Delta\sqrt{n}/s_d = t_{\nu,1-\beta} = -t_{\nu,\beta} \quad \{20\}$$

De sorte que la plus petite différence Δ entre les moyennes μ_1 et μ_2 pouvant être décelée avec des probabilités α et β fixées, et avec une taille d'échantillon égale à n, est estimée par :

$$\hat{\Delta} = (t_{\nu,\alpha}+t_{\nu,\beta})s_d/\sqrt{n}$$

Il faut noter que l'écart Δ qu'il est possible de mettre en évidence (avec des risques d'erreur donnés) peut toujours être réduit en augmentant l'effort d'échantillonnage : la précision de l'estimation de $\bar{d}$ croît avec n, en même temps que $\hat{V}(\bar{d})$ diminue. Cela s'oppose à ce qui avait été souligné au §1.2., où l'écart relatif δ tendait vers une limite δ_L non nulle (conditionnée par l'information contenue dans le premier échantillon de taille fixée) lorsque le nombre d'observations était autorisé à croître indéfiniment au cours de la période de surveillance.

D'un autre point de vue, la relation {20} permet aussi de connaître le nombre n de couples (x_{1i}, x_{2i}) nécessaire à la détection d'un écart Δ donné, avec des probabilités α et β fixées *a priori* :

$$n = (s_d(t_{\nu,\alpha}+t_{\nu,\beta})/\Delta)^2 \qquad \{21\}$$

Les résultats qui viennent d'être rappelés procèdent d'un protocole qui, comparé à celui de la stratégie "temps différé", présente les avantages suivants :

- Il n'est pas requis que X_1 et X_2 aient même variance.
- Les deux échantillons ne sont pas nécessairement indépendants (ce sont les réalisations d_i qui doivent l'être).
- S'il existe une source de variation exogène, *i.e.* dont les effets se superposent à ceux dont la détection constitue l'objet de l'étude, alors l'appariement des observations "filtre" cette information parasite. Cette propriété est spécialement intéressante pour l'étude de l'impact des centrales électronucléaires, étant donnée l'existence dans le milieu marin d'une forte variabilité naturelle, par exemple saisonnière ou interannuelle.

Il demeure cependant nécessaire de respecter les contraintes suivantes :

- Le nombre n de réalisations doit être également partagé entre X_1 et X_2.

- Il faut surtout s'assurer que la variance V(D) est indépendante des valeurs prises par les x_{1i} et les x_{2j}. Cela implique que les variances de X_1 et X_2 soient stables (indépendantes des fluctuations de μ_1 et μ_2), et peut conduire le cas échéant à recourir à une transformation des données, comme cela avait été exposé au §1.2.2.2.

3.3. LE TRAITEMENT PARAMETRIQUE DES GENERALISATIONS DE L'ALTERNATIVE

Jusqu'à présent, l'attention n'a été consacrée qu'à l'emploi des tests d'hypothèses les plus usuels, et dans ce cadre a été imposée une restriction supplémentaire, qui tient à la formulation des hypothèses H_o et H_1 elles-mêmes. Concernant ce point, l'alternative opposée à H_o est celle d'un décalage constant, *i.e.* une simple translation de la distribution des valeurs de la variable indicative consécutive à l'aménagement. Il s'agit là de l'alternative la plus élémentaire, et des hypothèses plus larges pourraient être envisagées : considérer par exemple que l'écart Δ est une fonction des valeurs de X. En reprenant les notations employées jusqu'ici ($X_1 \sim F_1$, $X_2 \sim F_2$), F_1 est dans ce cas la fonction de répartition de la variable aléatoire $X_2+\Delta(X_2)$; ainsi, s'il y a "effet proportionnel" :

$$X_1 = X_2+\Delta(X_2) = (1+a)X_2$$

la constante a étant nulle sous H_o, non nulle sous H_1. En fait, l'alternative la plus générale est celle où F_1 et F_2 diffèrent non seulement de position, mais aussi de dispersion et éventuellement de forme. Le modèle statistique sous-jacent se complique d'autant plus que les hypothèses en présence acquièrent de généralité, et il n'est pas assuré que la solution exacte du problème soit accessible. Cela peut être illustré par un exemple classique : lorsque F_1 et F_2 sont normales, de variances inconnues et distinctes ($\sigma_1^2 \neq \sigma_2^2$), la détection de H_1 (μ_1 diffère de μ_2) constitue ce qu'il est convenu d'appeler le problème de BEHRENS-FISHER. Dans ce contexte gaussien, un ensemble de statistiques exhaustives pour $\{\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$ est $\{\bar{x}_1, \bar{x}_2, S_1^2, S_2^2\}$, où $S^2 = \Sigma(x_i-\bar{x})^2$, et une région critique invariante pourrait être définie par :

$$(\bar{x}_1-\bar{x}_2)/\sqrt{S_1^2+S_2^2} \geq q(S_1^2/S_2^2)$$

où q désigne une fonction convenablement choisie (LEHMANN, 1959) ; si en outre le test est non biaisé, la probabilité de réalisation de l'inégalité doit être égale à la taille α de la région critique quand $\mu_1=\mu_2$, et cela quelle que soit la valeur de σ_1/σ_2. La question n'est toujours pas résolue de savoir s'il existe une fonction q possédant ces propriétés. Une solution approchée a néanmoins été avancée par WELCH (1947), qui au plan des applications offre une solution satisfaisante : elle consiste à définir une statistique h, fonction des estimations s_1^2, s_2^2 et d'une probabilité P donnée, telle que :

$$\text{Proba}\{((\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2))<h(s_1^2,\ s_2^2,P)\} = P$$

pour toute valeur de σ_1/σ_2. WELCH (*op.cit.*, équation 21) exprime h sous forme d'un développement permettant de calculer une approximation de

$$(h(s_1^2,\ s_2^2,P))/\sqrt{s_1^2/n_1+s_2^2/n_2}$$

pour toute valeur de P, et donc de tabuler la fonction de répartition de v,

$$v = ((\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2))/\sqrt{s_1^2/n_1+s_2^2/n_2}$$

L'explicitation de h est un raffinement d'une solution antérieurement présentée par le même auteur, consistant à ajuster le nombre ν de d.d.l. pour que les deux premiers moments de $s^2(1/n_1+1/n_2)/(\sigma_1^2/n_1+\sigma_2^2/n_2)$ où $s^2 = (n_1s_1^2+n_2s_2^2)/(n_1+n_2-2)$, soient égaux à ceux d'un χ^2. De la sorte, la quantité v suit approximativement une loi de Student à :

$$\nu = (\sigma_1^2/n_1+\sigma_2^2/n_2)^2/(\sigma_1^4/(n_1^2(n_1-1))+\sigma_1^4/(n_2-1))),$$

degrés de liberté (WELCH, 1947, équaticns (26) et (27) ; typiquement, ce nombre de d.d.l. n'est pas entier).

En fait, plusieurs auteurs ont suggéré l'utilisation routinière de la procédure de WELCH, *i.e.* l'emploi systématique du test pour la comparaison des moyennes de deux lois normales, qu'il y ait ou non hétéroscédasticité. En effet, BEST et RAYNER (1987, tabl. 4 p. 209) ont montré que le test de WELCH est d'une puissance équivalente au test t usuel, dès lors que $\nu \geq 5$, où ν est le nombre de d.d.l. indiqué ci-dessus. Incidemment, les difficultés théoriques

sur lesquelles vient d'être mis l'accent justifient l'intérêt accordé à la stabilisation des variances(§1.2.2.2.).

Hors du cadre gaussien qui vient d'être évoqué existent de nombreuses autres formulations de l'alternative, plusieurs d'entre elles ayant été recensées par SAVAGE (1956). Parmi celles-ci, l'alternative de LEHMANN se prête à des calculs approchés de la puissance relativement simples pour un certain nombre de tests non paramétriques usuels (*cf.* LEHMANN, 1953) ; l'alternative H_1 considérée énonce que X_1 est stochastiquement supérieure à X_2, au sens où $F_1 = F_2^k$, avec $k > 1$. En supposant k entier, l'interprétation est immédiate : F_2^k est la fonction de répartition du maximum de k variables aléatoires, $Y_1, \ldots, Y_k \overset{iid}{\sim} F_2$. Pour autant, il n'est nullement évident que cette hypothèse statistique constitue un modèle satisfaisant de la réalité biologique étudiée. C'est en cas d'inadéquation du modèle simple de translation $F_1(x) = F_2(x-\Delta)$, $\Delta>0$, que des alternatives plus élaborées pourront être envisagées.

Par ailleurs, outre la définition des hypothèses statistiques, l'éventail des méthodes examinées n'est lui-même pas exhaustif. De ce point de vue, il peut être remarqué que l'ensemble des tests mentionnés requiert des échantillons de taille prédéterminée (le nombre d'observations est une constante fixée *a priori*). C'est oublier les méthodes dites "séquentielles", qui utilisent des échantillons de taille aléatoire : typiquement, ces méthodes se composent d'une règle de saisie des observations et d'une règle d'arrêt, la seconde permettant de décider à chaque tirage s'il y a lieu ou non de continuer à échantillonner. Leur intérêt majeur tient à l'économie qu'elles permettent d'espérer, car la décision peut souvent être prise avec moins d'observations que n'en demanderait une procédure non séquentielle.

Le principe des méthodes séquentielles est fort bien illustré par la procédure de WALD (1947), qui utilise pour critère le rapport des vraisemblances ("sequential probability ratio test"). Formellement :

$X, X', X'', \ldots \overset{iid}{\sim} F$, de densité f ;

Problème : tester $H_o : f = f_o$, contre $H_1 : f = f_1$;

Au jème tirage est calculé : $L_j = \prod_{i=1}^{j} f_1(x_i)/f_o(x_i)$

Décision : - accepter H_o si $L_j \leq B$
- repousser H_o si $L_j \geq A$
- continuer à échantillonner, *i.e.* saisir une j+1ème observation, si $B<L_j<A$. Calculer L_{j+1} et recommencer les comparaisons.

A et B sont deux constantes, telles que $0<B<A$, et fixées *a priori*. L'un des résultats obtenus par WALD (*op. cit.*) est d'en avoir donné une approximation simple en termes de probabilités d'erreur α et β du test :

$$A = (1-\beta)/\alpha \qquad B = \beta/(1-\alpha)$$

à l'évidence, $B<1<A$ dès lors que $\alpha+\beta<1$. Il est remarquable que cette approximation ne fasse pas intervenir la loi de la variable étudiée X ; la procédure n'est pas pour autant non paramétrique, car la connaissance de la loi de X est utilisée directement pour la définition du critère L_j.

L'attention portée à l'analyse séquentielle n'ira pas au delà de ces considérations élémentaires. Il convient en particulier de souligner que la technique précitée s'appuie sur une région de décision fixée, et que n'ont pas été évoquées les procédures qui en autorisent une actualisation. Quelques références à des développements modernes de ces méthodes sont données par GOVINDARAJULU (1985).

L'application aux études de surveillance des tests séquentiels est demeurée jusqu'à présent assez limitée. Certains exemples peuvent être néanmoins cités, ainsi ALLEN *et al.* (1972), FOWLER (1983), RESH & PRICE (1984), JACKSON & RESH (1988).

3.4. L'APPROCHE NON PARAMETRIQUE

De même qu'au paragraphe 3.2., les données consistent en un échantillon aléatoire simple de n couples d'observations $(x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})$, qui sont n réalisations indépendantes du couple aléatoire :

$$(X_1, X_2) \sim G \qquad , \textit{i.e.} \quad \text{Proba}(X_1 \leq x_1, X_2 \leq x_2) = G(x_1, x_2)$$

Et soit la variable aléatoire D : $D = X_1 - X_2$

dont la fonction de répartition H est supposée continue :

$$D \sim H \qquad , \textit{i.e.} \quad \text{Proba}(D \leq d) = H(d)$$

Dans le cas présent, la question pratique à résoudre est celle de savoir si D est distribuée de façon "équilibrée" autour de zéro (absence d'impact), ou bien si au contraire la loi de D est centrée sur une valeur positive (cas où la variable X_2 tend à prendre, en probabilité, des valeurs plus faibles que celles de la variable X_1).

3.4.1. Hypothèses statistiques

Selon l'hypothèse nulle, la loi de D est symétrique autour de zéro ; d'où la présentation formelle :

$$H_o : G(x_1,x_2) = G(x_2,x_1) \iff \text{Proba}(D \leq d) = \text{Proba}(-D \leq d)$$
$$\iff H(d) = 1-H(-d)$$

La fonction de répartition symétrique de D sous H_o sera notée S.

A l'hypothèse nulle est opposée l'alternative suivant laquelle X_1 est stochastiquement supérieure à X_2 ; le modèle simple choisi pour décrire cette situation est celui d'un décalage $\Delta>0$, constant, retranché à la variable X_1. En conséquence, sous l'alternative unilatérale H_1, la fonction de répartition de la variable aléatoire $(X_1-\Delta) - X_2$ est précisément S. D'où l'écriture de H_1 :

$$H_1 : \text{Proba}(D \leq d) = S(d-\Delta) \quad , \Delta>0$$

Dans ce contexte, les hypothèses statistiques peuvent être abrégées comme suit :

$$H_o : \Delta=0 \text{ contre } H_1 : \Delta>0$$

3.4.2. Emploi du test de Wilcoxon

La décision en faveur de l'une ou l'autre des hypothèses précédemment énoncées repose sur la statistique de Wilcoxon pour observations appariées ("Wilcoxon two-sample Statistic"), définie comme une somme de $n(n+1)/2$ termes :

$$W = \sum_{i \leq j} \sum \phi(d_i,d_j)$$

où d_i et d_j sont deux des n réalisations indépendantes de D, et où la fonction ϕ est définie par :

$$\phi(d_i,d_j) = \begin{cases} 1 \text{ si } d_i+d_j>0 \\ 0 \text{ sinon} \end{cases} \qquad i = 1, \ldots, j \; ; \; j = 1, \ldots, n$$

W est donc le nombre des moyennes positives $(d_i+d_j)/2$ qu'il est possible de former, avec $i \leq j$. Cette définition est équivalente à celle, plus habituelle, qui énonce que la statistique W est calculée :

- en classant par ordre croissant les n valeurs absolues $|d_i|$ des différences observées d_i,
- puis en donnant à W la valeur prise par la somme des rangs affectés dans ce classement aux différences d_i positives.

Quelle que soit la définition à laquelle il est fait référence, il demeure facile de percevoir que les fortes valeurs de W sont aussi celles qui sont le moins conformes à ce que prévoit le modèle défini par l'hypothèse nulle. Au demeurant, H_o est rejetée quand W est supérieur à une valeur critique C_α :

$$W \geq C_\alpha \Rightarrow \text{rejet de } H_o \text{ au seuil } \alpha$$

La valeur de C_α, qui ne dépend que du nombre n des différences observées d_i (et bien sûr aussi du seuil α), est fournie par les tables de la loi de permutation de W sous H_o. Toutefois, pour une taille d'échantillon suffisante, une approximation peut être employée, fondée sur le résultat asymptotique suivant :

$$(W-E(W))/\sqrt{V(W)} \sim N(0,1) \text{ asymptotiquement}$$

De la sorte, il est possible d'appliquer la règle de décision approchée :

$$(W-E_o(W))\sqrt{V_o(W)} \geq \Phi^{-1}(1-\alpha) \Rightarrow \text{rejet de } H_o \text{ au seuil } \alpha$$

où E_o et V_o désignent respectivement l'espérance et la variance sous H_o. L'application de l'approximation normale nécessite donc la connaissance des deux premiers moments de W ; ceux-ci valent respectivement (résultat classique) :

$$E(W) = n(n-1)p'/2+np$$

$$V(W) = n((n-1)((n-2)(p''-p')^2+(2(p-p')^2+3p'(1-p'))/2)+p(1-p))$$

avec : $p = \text{Proba}(D_i>0)$, $p' = \text{Proba}(D_i+D_j>0)$

$p'' = \text{Proba}(D_i+D_j>0 \text{ ET } D_i+D_k>0)$, où $D_i, D_j, D_k \overset{iid}{\sim} H$

Sous H_o : $p = p' = 1/2$, $p'' = 1/3$, et donc :

$$E_o(W) = n(n+1)/4 \quad , \quad V_o(W) = n(n+1)(2n+1)/24$$

D'où l'approximation retenue pour la valeur critique C_α :

$$C_\alpha \simeq n(n+1)/4 - \sqrt{n(n+1)(2n+1)/24}.\Phi^{-1}(\alpha)$$

3.4.3. Calcul approché de la puissance

Face à l'alternative de décalage $H_1(\Delta>0)$, la puissance du test vaut :

$$\pi(\Delta) = \text{Proba}(W \geq C_\alpha \mid \Delta > 0)$$

$$= \text{Proba}\{(W-E_\Delta(W))/\sqrt{V_\Delta(W)} \geq (C_\alpha-E_\Delta(W))/\sqrt{V_\Delta(W)}\}$$

où E_Δ et V_Δ sont l'espérance et la variance sous l'alternative. Utilisant le fait que le membre gauche de l'inégalité suit une loi proche de la normalité, il vient,en négligeant la correction de continuité :

$$\pi(\Delta) \simeq 1 - \Phi((C_\alpha-E_\Delta(W))/\sqrt{V_\Delta(W)})$$

Soit encore, en remplaçant C_α par sa valeur approchée, et compte tenu de la symétrie de la fonction de répartition Φ de l'aléa N(0,1) :

$$\pi(\Delta) \simeq \Phi(A/\sqrt{V_\Delta(W)}), \text{ avec :}$$

$$A = (n/2)(n-1)(p'-1/2)+n(p-1/2)+\sqrt{n(n+1)(2n+1)/24}.\Phi^{-1}(\alpha) \qquad \{22\}$$

Le calcul de $\pi(\Delta)$ sera maintenant poursuivi en utilisant une heuristique proposée par LEHMANN (1975), et qui conduit à des résultats valides pour de petits décalages Δ ; en reprenant les probabilités p et p' apparaissant dans l'expression {22} :

$$p = \text{Proba}(D>0) = \text{Proba}(D-\Delta>-\Delta) = S(\Delta)$$

car sous l'alternative, la fonction de répartition de $(X_1-\Delta) - X_2$ est S, ainsi qu'il l'a été souligné au §3.4.1. ; il est rappelé que S est la fonction de répartition symétrique autour de zéro de D sous H_o.

De même :

$$p' = \text{Proba}((D_i-\Delta)+(D_j-\Delta)>-2\Delta) = S^{\dagger}(2\Delta)$$

où $S^{\dagger}$ est la fonction de répartition de la somme de deux variables aléatoires $\overset{iid}{\sim}$ S ; si les densités correspondantes sont respectivement notées $s^{\dagger}$ et s, alors au voisinage de $\Delta=0$:

$$p \simeq S(0)+\Delta.s(0) \quad , \text{ et } p' \simeq S^{\dagger}(0)+2\Delta.s^{\dagger}(0)$$

et compte tenu de la symétrie de S et de $S^{\dagger}$, il vient :

$$p - 1/2 \simeq \Delta.s(0) \quad , \text{ et } p' - 1/2 \simeq 2\Delta.s^{\dagger}(0)$$

La variance $V_{\Delta}(W)$ dépend également de Δ ; toutefois, selon LEHMANN (*op. cit.*), sa valeur reste proche de celle de $V_o(W)$ pour les faibles valeurs de Δ ; d'où l'approximation de la puissance :

$$\pi(\Delta) \simeq \Phi\{\Delta(n(n-1)s^{\dagger}(0)+ns(0))/\sqrt{n(n+1)(2n+1)/24} + \Phi^{-1}(\alpha)\} \qquad \{23\}$$

L'expression {23} fait apparaître la relation entre la taille n de l'échantillon, l'amplitude Δ du décalage, et les probabilités α et β associées respectivement aux risques de première et de seconde espèce. Cette relation permet de répondre à la question du plus petit écart pouvant être détecté avec des risques d'erreur α et β donnés, la taille de l'échantillon étant égale à n :

$$\Delta \simeq -(\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))\sqrt{n(n+1)(2n+1)/24}/(n(n-1)s^{\dagger}(0)+n.s(0)) \qquad \{24\}$$

Il est à remarquer que lorsque n augmente, le radical de numérateur tend à se comporter comme $\sqrt{n^3/12}$, tandis que le dénominateur devient équivalent à $n^2.s^\dagger(0)$. D'où l'approximation, valide pour n grand et Δ petit :

$$\Delta \simeq -(\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))/(\sqrt{12n}.s^\dagger(0)) \qquad \{25\}$$

Quant à la détermination du nombre n d'observations à réaliser pour déceler un écart donné Δ avec des risques d'erreur α et β fixés *a priori*, elle passe par la résolution de l'équation en n donnée par la formule {23}. Cette équation devra en général être résolue numériquement, mais il existe cependant une solution analytique simple lorsque n est suffisamment grand (et Δ petit) pour que soit admise l'approximation précédente :

$$n \simeq (\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))^2/(12(\Delta.s^\dagger(0))^2)$$

Enfin le cas particulier où la variable aléatoire D suit une loi normale peut être succinctement examiné ; dans ces conditions :

$$S \equiv N(0, \sigma^2) \quad \Rightarrow \quad \begin{array}{l} s(0) = 1/(\sigma\sqrt{\pi}) \\ S^\dagger \equiv N(0, 2\sigma^2) \text{ et } s^\dagger(0) = 1/(2\sigma\sqrt{\pi}) \end{array}$$

Dans ce contexte gaussien, la formule {24} devient, en exprimant le décalage Δ relativement à l'écart-type de la différence entre X_1 et X_2 :

$$\Delta/\sigma \simeq -(\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))\sqrt{\pi(n+2)/(3(n(n+1)))}$$

Il existe de même, pour les grands échantillons, une expression analogue à {25} :

$$\Delta/\sigma \simeq -(\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))\sqrt{\pi/3n}$$

Pour obtenir la taille d'échantillon nécessaire à la détection d'un écart Δ avec des probabilités α et β données, il suffit de résoudre une équation du second degré en n ; ou plus simplement de faire appel à l'approximation qui vient d'être employée ci-dessus, légitime pour n grand et Δ petit :

$$n \simeq (\pi/3)((\Phi^{-1}(\alpha)+\Phi^{-1}(\beta))(\sigma/\Delta))^2$$

La comparaison de cette dernière équation à la formule {21} rappelle, sans en constituer aucunement la démonstration, que l'efficacité de Pitman (ou efficacité asymtotique relative) du test de Wilcoxon par rapport au test de Student vaut $3/\pi$ lorsque la loi de D est normale ; cette notion peut être concrétisée en énonçant que, sous cette condition, le test de Wilcoxon appliqué avec n = 22 observations est à peu près aussi puissant que le test de Student utilisé avec n = 21 observations (car $3/\pi \simeq 21/22$).

Lorsque la loi de D n'est plus normale, l'efficacité de Pitman du test de Wilcoxon par rapport au test t peut devenir supérieure à 1, et elle n'est en tout cas jamais inférieure à .864 pour toute loi de variance finie. Cela signifie qu'il n'y a pas nécessairement perte de puissance, et qu'au pire cette perte restera faible.

3.5. LES TECHNIQUES DE REECHANTILLONNAGE ET LES STATISTIQUES ROBUSTES

Ainsi qu'il fut rappelé (§2.1.), la relative fragilité des tests paramétriques a suscité le développement des tests non paramétriques ; l'intérêt accordé à ces procédures est attesté par la richesse de la littérature anglo-saxonne, qui leur a consacré d'excellents manuels : NOETHER (1967), BRADLEY (1968), HAJEK (1969), HOLLANDER & WOLFE (1973), LEHMANN (1975), CONOVER (1980), pour ne citer que les plus connus.

Afin d'éviter toute confusion, il apparaît opportun de faire ici référence au jugement de HAMPEL (1973). Cet auteur estime qu'aussi surprenant que cela puisse sembler au premier abord, les techniques non paramétriques n'ont fondamentalement rien à voir avec les techniques robustes développées dans le contexte de modèles paramétriques, et qu'en particulier elles n'apparaissent guère appropriées pour s'insérer dans des protocoles complexes. De sorte que HAMPEL (*op. cit.*) souligne le besoin de mise au point de nouvelles méthodes, robustes et spécifiques, et qui pourraient marquer l'essor des statistiques "de la troisième génération", faisant suite à la statistique paramétrique classique et à la statistique non paramétrique. La distinction entre ces différentes notions a été clairement précisée par HUBER (1981) :
(i) une procédure est dite non paramétrique quand elle est d'un usage adapté à un large ensemble non paramétré de distributions parentes ;

(ii) un test est dit "distribution-free" si la probabilité α demeure la même quelle que soit la loi (continue) sous-jacente. Cette qualité n'implique rien quant au comportement de la puissance, même si une certaine stabilité de cette dernière peut être empiriquement constatée dans une majorité de cas ;
(iii) les méthodes robustes sont conçues pour être utilisées dans un contexte paramétrique ; leur originalité tient à ce que le modèle postulé n'est plus considéré comme tout à fait exact, et que la prise en compte explicite de l'écart au modèle fait partie intégrante de la méthode.

En ce sens ont été proposés des tests robustes (HUBER, 1981) ; outre qu'ils semblent actuellement plus constituer un objet d'étude pour les mathématiciens qu'une technique d'usage courant, il est difficile d'en déterminer le seuil et la puissance. Ils ne seront pas examinés ici, l'attention étant en revanche consacrée à l'emploi conjoint d'estimateurs de position robustes et de techniques de rééchantillonnage. Cette idée peut être didactiquement exposée en empruntant à EFRON l'introduction de l'un de ses articles de présentation du "bootstrap" (EFRON, 1979b) ; l'objet principal de la statistique étant la comparaison d'ensemble de nombres (entre eux, avec des modèles théoriques, ...), l'archétype de la question posée peut se résumer ainsi : "les valeurs des nombres de l'ensemble A sont-elles supérieures à celles des nombres de l'ensemble B ?".

Avant 1950, la réponse standard était : calculer la statistique t, et comparer la valeur obtenue à la distribution tabulée sous le modèle gaussien. A partir des années cinquante fut utilisée une approche (dont la paternité peut être attribuée à FISHER, qui en exposé le principe en 1935) conduisant à une solution ne dépendant pas du modèle gaussien : regrouper les n_A et n_B éléments de A et B dans un unique ensemble C, puis considérer les $(n_A+n_B)!/(n_A!n_B!)$ partitions possibles (équiprobables sous H_o) de C en deux ensembles d'effectifs n_A et n_B, et calculer pour chacune d'elles la différence entre, par exemple, leurs moyennes : $\bar{a} - \bar{b}$ (l'une de ces différences coïncide avec la différence expérimentale observée) ; il est alors décidé que "A est plus grand que B" si la différence observée appartient au quantile supérieur de taille α de la loi de permutation de $\bar{a} - \bar{b}$. Pour se libérer de la condition de normalité, la technique non paramétrique paye le prix d'une augmenta-

tion du temps de calcul : la distribution de $\bar{a} - \bar{b}$, qui est en quelque sorte une idiosyncrasie de l'échantillon, doit être tabulée pour chaque nouveau lot de données[1]. Le gain ne se résume cependant pas à l'abandon de la condition de normalité, car il existe aussi une plus grande latitude pour le choix de la statistique du test : la méthode s'applique par exemple tout autant à la différence des médianes qu'à celle des moyennes.

L'"impensable" mentionné dans le titre de l'article d'EFRON (1979b, *cf.* bibliographie), c'est d'effectuer quelques centaines de milliers d'opérations arithmétiques pour analyser des échantillons dont la taille est de l'ordre de la dizaine de données chiffrées. L'algorithme de base sera exposé en partant du cas simple défini au §3.1. : l'objectif consiste à tenter des inférences sur la tendance centrale (décalée ou non par rapport à zéro) des variables aléatoires :

$$D_1, \ldots, D_n \overset{iid}{\sim} H$$

où $D = X_1 - X_2$. Soit donc θ un paramètre de position de la vraie fonction de répartition H. Au §3.2. a été choisi $\theta \equiv E(D)$, estimé sans biais par $\hat{\theta} = \bar{d}$, moyenne des n différences $x_{i1} - x_{i2}$. L'un des intérêts de l'algorithme du bootstrap, qui va être présenté plus loin, réside dans ce qu'il s'applique identiquement à des estimateurs plus robustes que la moyenne (*vide infra*), tout en permettant d'en évaluer aisément des caractéristiques de la distribution d'échantillonnage (*e.g.*, biais, variance) auxquelles il n'est pas toujours facile d'accéder par la voie analytique. Préalablement seront donc examinées les possibilités d'estimation robuste de la valeur centrale θ de la loi H (symétrique lorsque les densités de X_1 et X_2, quelle qu'en soit la forme, ne diffèrent que par une translation).

---

(1) Dans le cas de nombreux tests non paramétriques, cette difficulté est tournée en tabulant une fois pour toutes la distribution sous H_o d'une statistique calculée sur les rangs des valeurs observées (et non sur les valeurs elles-mêmes) : de la sorte, la loi de permutation ne dépend que du nombre d'observations.

3.5.1. L'estimation robuste d'une valeur centrale

Dans ce qui va suivre, tout paramètre θ apte à repérer la tendance centrale de la fonction de répartition H peut être considéré comme une fonction paramétrique associée à cette même loi par une fonctionnelle $T(.) : \theta = T(H)$.

Ainsi la moyenne arithmétique est-elle définie par la formule implicite :

$$\int (x-T(H))H(dx) = 0$$

Quant à l'estimateur naturel $\hat{\theta}$ de θ, il est tout simplement défini par :

$$\hat{\theta} = T(\hat{H})$$

où $\hat{H}$ désigne la fonction de répartition empirique de l'échantillon $\{D_1 = d_1, \ldots, D_n = d_n\}$, *i.e.* :

$$\hat{H} = (1/n) \sum_{i=1}^{n} \delta_{Di} \quad , \delta \text{ masse de Dirac}$$

Pour reprendre l'exemple de la moyenne arithmétique :

$$\bar{d} = T(\hat{H}) \quad \text{solution de} \quad \sum_{i=1}^{n} (d_i - T(\hat{H})) = 0$$

Ces notions de fonctionnelle statistique et d'estimateur naturel, conceptuellement très accessibles, sont à la base des techniques qui vont être exposées. En particulier, à partir de la fonctionnelle statistique est élaboré un outil mathématique privilégié pour qualifier la robustesse : la fonction d'influence (HAMPEL, 1974). Les développements qui président à sa définition débordent le cadre de ce rapport ; il suffira d'indiquer ici quelle en est l'utilité :

(i) quantifier l'impact d'une observation individuelle sur la valeur prise par l'estimation (ou la statistique du test). La valeur absolue du maximum de la fonction d'influence constitue la sensibilité globale ("gross error sensi-

tivity" de HAMPEL). Un maximum non borné révèle qu'une observation aberrante est potentiellement génératrice de dégâts ; de ce point de vue, la sensibilité globale infinie de la moyenne arithmétique la désigne comme un estimateur non robuste ;

(ii) fournir une heuristique pour aborder les propriétés asymptotiques d'un estimateur ; la fonction d'influence permet à ce titre de statuer sur la robustesse au sens de la variance asymptotique minimale.

Pour éclairer le lecteur d'une référence qui lui est sans doute plus familière peut être évoqué le lien entre jackknife et fonction d'influence : les différences algébriques entre les "pseudo-valeurs" de TUKEY et l'estimateur jackknife réalisent une fonction d'influence empirique (*cf.* HUBER, 1972, 1981).

Ces définitions de base étant rappelées, les grands traits des solutions proposées pour l'estimation robuste d'un paramètre de position peuvent être présentés. Bien qu'attestée par d'anciens écrits, l'idée de l'estimation robuste d'une valeur centrale a été systématisée il y a plus d'une vingtaine d'années seulement par HUBER, qui a étudié une classe d'estimateurs généralisant le maximum de vraisemblance, et nommés pour cette raison les M-estimateurs. Ils sont à l'origine conçus pour conserver de "bonnes qualités" en présence d'une loi dont les extrêmités sont plus épaisses que celles prévues par le modèle gaussien (loi normale "perturbée", *cf.* le modèle de contamination présenté au §2.1.). Un M-estimateur $\hat{\theta}$ est solution du problème :

$$\min_{\theta} \sum_{i=1}^{n} \rho(x_i ; \theta)$$

où ρ est une fonction arbitraire non constante. Cette définition englobe celle des moindres carrés, ainsi que celle du maximum de vraisemblance (en posant $\rho(x ; \theta) = -\ln(f(x ; \theta))$, où f désigne la densité de la variable). Si ρ est dérivable par rapport à θ, alors $\hat{\theta}$ est solution des équations du premier ordre :

$$\sum_{i=1}^{n} \psi(x_i,\theta) = 0 \quad , \text{ avec } \psi(x,\theta) = \partial\rho(x,\theta)/\partial\theta \qquad \{26\}$$

Les estimateurs des moindres carrés (*i.e.* $\rho(u) = u^2$, $\psi(u) = 2u$) acquièrent un comportement instable quand les observations sont engendrées par des distributions à queues épaisses. D'où la suggestion de HUBER : pour estimer la valeur centrale θ d'une loi supposée symétrique, choisir une fonction ψ impaire et strictement croissante sur $[-k, +k]$, et constante en dehors de cet intervalle (k est un nombre positif fixé). Ainsi l'effet des valeurs extrêmes est-il atténué en posant $\psi(u) = \max(-k,\min(u,k))$, *i.e.* :

$$\psi(u) = \begin{cases} u \text{ si } |u| \leq k \\ k.\text{sgn}(u) \text{ si } |u| > k \end{cases}$$

une option plus radicale étant : $\psi^{*}(u) = 0$ si $|u| > k$

Il peut être remarqué que la médiane est le M-estimateur défini par $\rho(u) = u$ (*i.e.*, $\psi(u) = \text{sgn}(u)$). Il ne saurait être question d'aller au delà de ces exemples simples, et d'inventorier, même succinctement, l'ensemble des M-estimateurs existant à l'heure actuelle. Plusieurs d'entre eux sont étudiés dans l'ouvrage de ANDREWS *et al.* (1972), leurs propriétés générales étant exposées dans celui de HUBER (1981). Par ailleurs, toute nouvelle formulation de la fonction ψ engendre l'étude du comportement de l'estimateur correspondant sur de petits échantillons, de ses propriétés asymptotiques, ..., et doit être assortie d'une méthode de calcul ; à cet égard, l'optimisation des paramètres impliqués dans l'expression de ψ (par exemple k, *vide supra*) dépend de façon complexe des observations, et la résolution du système {26} nécessite en général le recours à un procédé itératif.

Deux autres classes d'estimateurs robustes doivent être mentionnées : (i) celle des L-estimateurs, construits par combinaison linéaire de statistiques d'ordre, et dont le représentant le plus connu est la moyenne α-tronquée ($0<\alpha<1/2$). En reprenant le cas concret traité dans ce chapitre, elle est définie comme suit : dans l'échantillon $\{D_1 = d_1, \ldots, D_n = d_n\}$ de taille n, les valeurs observées d_i sont ordonnées par valeurs croissantes $d_{(i)}$:

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(n)}$$

En posant $g = n.\alpha$, et en supposant pour simplifier que g est entier, la moyenne α-tronquée est obtenue en supprimant les g premières et dernières observations de l'échantillon ordonné $\{d_{(1)}, \ldots, d_{(n)}\}$:

$$\bar{d}_\alpha = (1/(n-2g)) \sum_{i=g+1}^{n-g} d_{(i)}$$

A la même classe appartient la moyenne de Winsor d'ordre $\alpha(0<\alpha<1/2)$, qui est la moyenne arithmétique simple obtenue en remplaçant les g premières observations par $d_{(g+1)}$ et les g dernières par $d_{(n-g)}$:

$$\bar{w}_\alpha = \left(\sum_{i=g+1}^{n-g} d_{(i)} + g(d_{(g+1)}+d_{(n-g)}) \right)/n$$

L'intention sous-jacente à la définition de $\bar{w}_\alpha$ correspond au souci de ne pas élaguer aussi brutalement qu'avec $\bar{d}_\alpha$ l'information attachée aux valeurs extrêmes ; l'examen des fonctions d'influence montre combien l'intuition est trompeuse (HUBER, 1981) : la moyenne tronquée ne repousse pas toute l'information contenue dans les 2g valeurs éliminées, elle possède en fait le comportement que l'on attendrait de la moyenne de Winsor.

Il peut être enfin noté que pour une loi H symétrique, la fonctionnelle qui définit la moyenne α-tronquée est identique à celle du M-estimateur associé à la fonction ψ^{**} pour $k = H^{-1}(1-\alpha)$;

(ii) la troisième classe d'estimateurs robustes est celle des R-estimateurs. Un R-estimateur est une statistique linéaire de rang de la forme :

$$\sum_{i=1}^{n} C_i.a(r_i)$$

où r_i est le rang de l'observation d_i dans l'échantillon $\{d_1, \ldots, d_n\}$, et où $a(r_i)$ est un réel appelé "score". A cette catégorie appartient l'estimateur de Hodges-Lehmann. Appliqué au problème de l'estimation de la valeur centrale de la fonction de répartition H à partir de l'échantillon $\{D_1 = d_1, \ldots, D_n = d_n\}$, cet estimateur est défini par :

$$\hat{\theta} = \text{med}\{(d_i+d_j)/2\} \qquad i=1, \ldots, n \;\; ; \;\; j=1, \ldots, n$$

et n'est autre que la médiane des n^2 moyennes $(d_i+d_j)/2$.

Pour les trois classes d'estimateurs considérées précédemment ont été choisis des exemples exprimés sous la forme déduite de la fonctionnelle T de la fonction de répartition empirique $\hat{H}$: $\hat{\theta} = T(\hat{H})$. Plus généralement, les fonctionnelles qui définissent les représentants de chacune des trois classes sont :

. pour les M-estimateurs : $\int\psi(x-T(H))H(dx) = 0$

. pour les L-estimateurs : $T(H) = \int J(t)H^{-1}(t)dt$

. et pour les R-estimateurs : $\int J((H(x)+1-H(2T(H)-x))/2)H(dx) = 0$

A partir de ces équations sont calculées les fonctions d'influence, qui fournissent des critères quantitatifs permettant de comparer entre eux et à la moyenne arithmétique les quelques estimateurs robustes déjà cités ; ci-après sont brièvement résumés quelques résultats de HAMPEL (1974) :

	Sensibilité globale	Sensibilité locale	V_n	V_{cn}
moyenne arithmétique	∞	1	1	∞
médiane	1.25	∞	$\pi/2$	1.74
moyenne .05-tronquée	1.83	1.11	1.03	1.30
moyenne .10-tronquée	1.60	1.25	1.06	1.26
moyenne de Winsor d'ordre .05	2.13	∞	1.01	1.46
estimateur de Hodges-Lehmann	1.77	1.41	1.05	1.29

Quatre mesures de la robustesse ont été rassemblées dans ce tableau. La sensibilité globale a déjà été évoquée plus haut ; la sensibilité locale ("local-shift-sensitivity" de HAMPEL) est un révélateur de l'effet plus ou moins prononcé du déplacement local d'une observation, créé par exemple par un arrondissage, une troncature, ou encore par une erreur de mesure. Les deux dernières colonnes indiquent :

- la variance asymptotique V_n sous l'hypothèse de normalité N(0,1) ;
- la variance asymptotique V_{cn} sous la loi contaminée $.95N(0,1)+.05N(0,\sigma)$, $\sigma \to +\infty$.

A côté de ces indices quantitatifs existent des moyens d'apprécier qualitativement la robustesse (*e.g.*, la continuité de la fonctionnelle T au voisinage de la loi H, ou encore le "point de rupture", qui correspond très schématiquement à la proportion d'observations aberrantes que peut supporter l'estimateur avant que ses qualités ne se détériorent de manière catastrophique). Il ne sera pas insisté sur ces questions théoriques, car il semble plus judicieux de fournir au praticien des éléments de réponse à la traditionnelle interrogation : quel estimateur choisir ? Selon ANDREWS *et al.* (1972), il n'existe aucune panacée applicable à la diversité des situations rencontrées. De la soixantaine d'estimateurs éprouvés par ces auteurs, la moyenne arithmétique se révèle être le plus mauvais hors du cadre gaussien ; en présence de données suspectes ("outliers", *cf.* BECKMAN & COOK, 1983), elle devra être utilisée conjointement avec un procédé d'élimination des valeurs aberrantes. Cela introduit néanmoins la difficulté de savoir ce que représente exactement la variance des observations qui auront été conservées. Une méthode robuste en revanche, qui pondère explicitement les observations, fournit des estimations de variance asymptotiquement correctes.

Sans qu'il soit pour autant possible de la qualifier de "passe-partout", il existe cependant une classe d'estimateurs pourvus d'intéressantes propriétés de robustesse : faible réaction à de légères perturbations (robustesse qualitative), comportement satisfaisant en cas de contamination importante (point de rupture élevé), l'influence relative maximale de la contamination étant elle même bornée (faible sensibilité globale). Cette classe est celle des M-estimateurs "à paliers",qui conservent de surcroît une bonne efficacité sous le modèle gaussien. Avec les mêmes notations qu'en {26} :

$$\begin{aligned}
\psi(x) &= x && \text{si } |x| \leq a,\\
&= a.\mathrm{sgn}(x) && \text{si } a \leq |x| \leq b,\\
&= a(x-c.\mathrm{sgn}(x))/(b-c) && \text{si } b \leq |x| \leq c,\\
&= 0 && \text{si } c \leq |x|
\end{aligned}$$

Par exemple : a = 2e, b = 4e, c = 8e, où e est la médiane des valeurs absolues des écarts à la médiane empirique med(x) de l'échantillon :

$$e = \mathrm{med}|x_i - \mathrm{med}(x)| \qquad , i=1, \ldots, n$$

L'estimateur $\hat{\theta}$ de la valeur centrale est solution de :

$$\Sigma\psi((x_i-\theta)/e) = 0$$

Le résultat est obtenu en quelques itérations en prenant med(x) pour point de départ. Ce détail technique n'est pas dépourvu d'intérêt si la distribution d'échantillonnage de l'estimateur $\hat{\theta}$ est étudiée à l'aide d'une méthode de ré-échantillonnage telle que le bootstrap.

3.5.2. Le bootstrap

Soit donc $\hat{\theta}$, estimateur naturel de θ, défini par la fonctionnelle statistique T :

$$\hat{\theta} = T(\hat{H})$$

où $\hat{H}$ désigne la fonction de répartition empirique de l'échantillon $\{D_1 = d_1, \ldots, D_n = d_n\}$, avec, dans le problème traité ici, $d_i = x_{i1}-x_{i2}$:

$\hat{H}$: masse $1/n$ en d_i , $i = 1, \ldots, n$

Cette fonction $\hat{H}$ réalise une estimation non paramétrique, au sens du maximum de vraisemblance, de la vraie fonction de répartition H. Et pour obtenir de l'information sur la distribution d'échantillonnage de $\hat{\theta}$, la technique du bootstrap consiste "tout simplement" à créer des échantillons à partir de $\hat{H}$ plutôt que d'en saisir dans H. En d'autres termes, les "échantillons bootstrap", extraits de la fonction de répartition empirique, sont donc formés à partir de l'unique échantillon *s.s.* au lieu d'être tirés de la population. D'où le nom de bootstrap, qui signifie tirant de botte, et qui évoque le miracle consistant à se soulever soi-même en tirant sur ses propres lacets de chaussures ("pulling yourself up by your own bootstraps") ... En résumé, les "réplicats bootstrap" $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_b, \ldots, \hat{\theta}^*_B$ sont obtenus par la méthode de Monte Carlo décrite ci-après :

Etape 1 - Construire $\hat{H}$: masse $1/n$ à chaque d_i, $i = 1, \ldots, n$

Etape 2 - Tirer de $\hat{H}$ un échantillon bootstrap de taille n.

$$D_i^{\ast} = d_i^{\ast} \quad , \text{ avec } D_1^{\ast}, \ldots, D_n^{\ast} \overset{iid}{\sim} \hat{H}$$

Les valeurs $d_i^{\ast}$ sont simplement obtenues par échantillonnage aléatoire simple, et tirages avec remise, des valeurs de l'ensemble $\{d_1, \ldots, d_n\}$; elles permettent de calculer :

$$\hat{\theta}^{\ast} = T(d_1^{\ast}, \ldots, d_n^{\ast})$$

Etape 3 - Effectuer B répétitions indépendantes de l'étape 2, jusqu'à l'obtention de B (en pratique, quelques centaines) réplicats bootstrap $\hat{\theta}_1^{\ast}, \ldots, \hat{\theta}_B^{\ast}$.

Partant de ces réplicats, plusieurs caractéristiques de l'estimateur $\hat{\theta} = T(\hat{H})$ peuvent être aisément calculées ; ainsi, le biais (EFRON, 1981b) :

$$\text{biais de } \hat{\theta} \equiv E_{\hat{H}}(T(\hat{H})) - T(H)$$
$$\text{est-il estimé par : biais estimé} = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b^{\ast} - \hat{\theta}$$

De même la variance d'échantillonnage de $\hat{\theta}$ est-elle estimée par la variance des $\hat{\theta}_b^{\ast}$ (*cf.* EFRON, 1979a,b). Le problème posé est de savoir si la loi de D peut être considérée comme centrée ou non sur zéro (en un sens à définir : moyenne, médiane, ces exemples simples ne devant cependant pas occulter le domaine d'élection du bootstrap, qui est celui de problèmes hautement plus ardus - *cf.* ROCKE & DOWNS, 1981), et il est tentant de le résoudre en attribuant un intervalle de confiance à l'estimateur de position $\hat{\theta}$. En effet, les réplicats $\hat{\theta}_1^{\ast}, \ldots, \hat{\theta}_B^{\ast}$ permettent d'estimer la distribution d'échantillonnage de l'estimateur $\hat{\theta}$ (EFRON, 1981a) ; soit donc $\hat{C}$ la cumulée empirique obtenue pour la fonctionnelle statistique $\hat{\theta} = T(\hat{H})$:

$$\hat{C}(t) = \text{Proba}_{\ast}\{\hat{\theta}^{\ast} < t\} = (\#\{\hat{\theta}_b^{\ast} < t\})/B$$

où le symbole # signifie "nombre de fois où", la notation $\text{Proba}_{\ast}$ rappelant que c'est le statisticien (et non la nature) qui crée ici l'aléatoire en utilisant l'algorithme du bootstrap. L'intervalle de confiance non paramétrique à $100(1-2\alpha)\%$ est construit à l'aide des percentiles 100α et $100(1-\alpha)$ de $\hat{C}$, en ne choisissant pas un risque de première espèce trop faible (voir la remarque de NASH dans la discussion de l'article de EFRON, 1981a) :

en notant $\hat{\theta}(\alpha) = \hat{C}^{-1}(\alpha)$, $\hat{\theta}(1-\alpha) = \hat{C}^{-1}(1-\alpha)$

il vient alors : $\text{Proba}\{\theta \in [\hat{\theta}(\alpha), \hat{\theta}(1-\alpha)]\} \simeq 1-2\alpha$

intervalle à partir duquel peuvent être faites des inférences sur la tendance centrale de la variable D, en particulier tester $H_o : \theta = 0$. Les questions théoriques soulevées par cette approche sont discutées par EFRON (1982, chap.10). L'ouvrage qui vient d'être cité éclaire aussi, entre autres, les relations entre le bootstrap et la technique plus ancienne du jackknife (*cf.* MILLER, 1974), celle-ci n'étant qu'un cas particulier de celle-là (voir à ce sujet EFRON & GONG, 1983). Au surplus, ces travaux révèlent que les moyens modernes de calcul confèrent à la statistique un caractère expérimental, compte tenu des larges possibilités offertes par le recours aux techniques de simulation.

De ce qui précède, il ressort que les "études d'impact" sont encore loin de constituer une activité figée. Et l'interrogation d'apparence anodine, *e.g.* "quelle(s) caractéristique(s) de l'écosystème constitue(nt) un bon révélateur de perturbations statistiquement décelables ?" recouvre des problèmes difficiles aussi bien dans le domaine statistique que dans le domaine écologique. L'attention n'a été jusqu'à présent consacrée qu'au premier de ces domaines. Encore doit-il être remarqué qu'il ne saurait être question d'offrir un inventaire complet de la gamme des techniques statistiques utilisables pour étudier l'impact d'un aménagement : ainsi, nulle mention n'est faite du traitement des séries chronologiques, des analyses multivariables, ... La priorité va être maintenant accordée aux thèmes écologiques, afin d'éclairer dans quel contexte (et sous quelles contraintes) sont appliqués les tests d'hypothèses qui viennent d'être considérés.

CHAPITRE IV

MODALITES D'APPLICATION : LES CHOIX ESSENTIELS

La mise en oeuvre d'une stratégie d'étude d'impact doit s'attacher à éviter un double écueil : celui de n'être apte qu'à déceler une catastrophe, et celui d'échantillonner beaucoup plus qu'il n'est besoin pour tester une hypothèse raisonnable. Cette remarque liminaire amène à préciser la notion de "signifiance d'un impact" ; jusqu'à présent n'a été exclusivement spécifié que ce qu'il faut entendre par impact "statistiquement significatif". Pour décider que l'impact possède ou non cette qualité, il ressort des développements techniques présentés antérieurement que doivent être préalablement formulées des hypothèses statistiques : la priorité implicite réside donc dans la définition de l'impact "biologiquement (ou écologiquement) significatif". En reprenant la notation précédemment employée, la question est de savoir quelle est la plus petite valeur de Δ qui peut être tenue pour l'expression d'un remaniement de l'écosystème. Dans les définitions de base rappelées au §1.1., il a été insisté sur le fait que l'amplitude de l'écart Δ entre H_o et H_1 doit correspondre à une réelle perturbation du compartiment étudié dans l'écosystème, et qu'en dessous de cette valeur il serait conclu à une absence d'impact. La nuance qui sépare "statistiquement significatif" de "écologiquement non négligeable" peut alors être résumée comme suit, en notant d la valeur courante qui définit la classe des alternatives unilatérales considérées jusqu'ici :

$0<d<\Delta$: statistiquement significatif et écologiquement négligeable ;
$\Delta \leq d$: statistiquement significatif et écologiquement notable.

Le critère qui gouverne la règle de décision statistique dépend donc étroitement de la formulation écologique du problème ; cela appelle l'examen de cette dernière. Quelques critères susceptibles de guider le choix de la variable "sensible à l'impact" seront donc recensés.

A cet égard, il est opportun de signaler que PATTEN (1984) a récemment explicité la trame conceptuelle du problème de la protection de l'environnement, en usant d'une présentation formelle fondée sur la théorie des systèmes. Concernant plus particulièrement les critères de choix, cet auteur analyse, dans le contexte de la problématique écotoxicologique, les qualités que doivent posséder les "variables indicatives" ("diagnostic variables") pour que leur maintien entre certaines normes soit une condition nécessaire et suffisante de l'intégrité des autres parties de l'écosystème. Bien que la "nuisance"

considérée par PATTEN soit une émission de polluants chimiques, son raisonnement est généralisable à une grande variété de cas.

De la synthèse de PATTEN (*op. cit.*), il ressort à l'évidence que la définition de critères de surveillance qui soient à même d'offrir des garanties précises exige une connaissance elle-même assez exacte du fonctionnement du système soumis à perturbation. Cette connaissance fondamentale est rarement disponible lorsque doivent être arrêtées les méthodes de la surveillance (*cf.* ROSENBERG & RESH *et al.*, 1981). Ainsi, pour ne considérer que l'exemple du contrôle du milieu à proximité des centrales électronucléaires, il apparaît que l'information acquise sur les sites côtiers de la Manche est surtout de nature "statique" et descriptive, la compréhension de la dynamique des processus n'ayant pas été l'objectif primordial des études "d'état de référence" (*cf.* PARENT, 1981). Par conséquent, les critères de choix qui seront donnés plus loin n'ont pas été établis conformément à l'approche cognitive prônée par PATTEN (*op. cit.*), ils tentent plutôt d'établir un compromis entre une exigence scientifique minimale et des impératifs financiers, lesquels se verront accorder une sensible influence : la démarche est donc empreinte d'un certain pragmatisme.

4.1. LE CHOIX DE LA VARIABLE INDICATIVE

La variable indicative, notée X aux chapitres précédents, est celle dont une variation brutale et durable sera comprise comme l'indice d'une modification de l'écosystème, modification engendrée par l'aménagement. L'état de l'art ne permet de donner qu'une définition assez floue des éléments du diagnostic : les lois du comportement dynamique des écosystèmes sont encore insuffisamment élucidées pour pouvoir reconnaître à l'avance les subtiles variations qui seraient à même d'entraîner un profond changement d'état. De fait, les programmes de surveillance ne sont pas conçus pour attaquer le problème en amont, mais plutôt pour analyser la réponse de certaines composantes du système à une perturbation qui l'affecte dans son ensemble. Cela distingue ces programmes (dits de "surveillance") de ceux ayant pour objectif la prévision des modifications entraînées par une contrainte exercée sur le milieu, et qui globalement doivent s'appuyer sur la connaissance de la fonction de transfert du système. La préoccupation sous-jacente des études de surveillance n'est pas d'expliquer comment l'écosystème filtre le signal "mise en place de l'aménagement", mais d'apprécier si quelques uns de ses compartiments (voire un seul),

réagissent différemment lorsque ce stimulus s'ajoute à ceux émis par l'environnement en conditions non perturbées. Un linguiste dirait, par homologie, que c'est au signifiant et non au signifié que s'adressent les études de surveillance du milieu.

Sera donc considéré comme impact toute modification (de nature écologique) persistant au cours du temps, et dont l'amplitude est présumée traduire un changement d'état du système ; cela sous-entend que l'amplitude de la modification est telle que celle ci peut être distinguée de la variabilité naturelle. En ce sens, sa détection est tributaire du pouvoir de résolution des techniques statistiques, *i.e.* de leur puissance. En outre, et dans le même ordre d'idée, il faut souligner que les méthodes présentées jusqu'ici supposent que le changement Δ ne s'"installe" pas progressivement pendant la période de surveillance, mais existe comme tel dès son début : à l'échelle d'observation retenue, l'effet est supposé à la fois soudain puis chronique[1]. Cela n'est pas sans incidence sur le choix de la variable indicative ; en effet, la dynamique du compartiment étudié devra être telle que son délai de réaction soit faible en regard du pas de temps de l'échantillonnage. Cette exigence conduira, en première analyse, à s'intéresser aux compartiments à taux de renouvellement élevé ; compte tenu du fait que le rythme de l'échantillonnage des études de surveillance est couramment de l'ordre du mois, les microorganismes planctoniques (et spécialement le phytoplancton) constituent souvent une cible privilégiée. Néanmoins, la relative brièveté des générations successives n'est pas l'apanage exclusif de ces peuplements : ce critère est par exemple tout autant satisfait par la méiofaune, le microphytobenthos, et plus encore par le compartiment bactérien. Pour ceux-ci, le coût de l'échantillonnage et du dépouillement intervient en revanche de façon rédhibitoire.

Ce dernier aspect, qui sera précisé plus loin (§4.3.), rappelle que les choix sont par ailleurs subordonnés à des impératifs contradictoires : il est attendu de la variable indicative qu'elle délivre une certaine quantité

---

(1) Le protocole récemment proposé par WALTERS *et al.* (1988) permet de s'appuyer sur l'hypothèse plus subtile d'une modification graduelle, tout en estimant l'interaction temps x intensité de l'impact sur une période de régime transitoire.

d'informations sur la structure, voire le comportement, de l'écosystème ; en même temps, elle doit pouvoir être échantillonnée facilement, fréquemment, et à moindre coût. La première préoccupation tendrait à privilégier l'estimation d'un indice plutôt sensible aux variations de composition des peuplements (*e.g.*, richesse spécifique, diversité), ou mieux encore directement lié aux transferts d'énergie dans l'écosystème (mesures de flux, de taux de production). Quant aux contraintes techniques et financières, elles engagent peu ou prou à imposer le choix d'une variable qui peut être saisie *in situ* de manière automatisable (avec réplicats), sans nécessiter de dépouillement ultérieur des échantillons au laboratoire. Eu égard à la minceur des ambitions scientifiques des programmes de surveillance (qui ne peuvent que susciter des thèmes de recherche en vue d'une amélioration des procédures de contrôle, mais qui ne conservent pour domaine que le contrôle lui-même, par essence plus ou moins routinier), ce sont en l'état actuel les considérations d'ordre économique qui interviennent de façon prépondérante. En conséquence, le choix de X sera effectué conformément aux deux priorités suivantes :

(1) faible coût de saisie des échantillons ;
(2) dépouillement des échantillons "consommant" peu (ou pas) de temps de personnel spécialisé ;

tout en tenant compte des critères suivants :

(3) possible adéquation du pas de temps de l'échantillonnage au "turnover" du compartiment étudié ;
(4) et compatibilité de l'extension des corrélations spatiales (entre réalisations de X) avec le pas d'espace de l'échantillonnage.

Il sera revenu au § 4.2. sur les critères (3) et (4) ; leur intérêt est de garantir l'indépendance entre observations de deux campagnes d'échantillonnage consécutives, comme entre stations d'une même campagne.

L'inventaire ci-dessus offre une grille d'appréciation ne pouvant être que générale, chaque site concerné par un aménagement étant abordé dans ce cadre comme un cas particulier. Ainsi, pour reprendre l'exemple des études réalisées en France pour le compte d'EDF, il peut être observé que les cen-

trales électronucléaires littorales ont été majoritairement installées sur les côtes de la Manche. S'il fallait envisager une extension à la Méditerranée, alors la définition d'un programme de surveillance n'intègrerait que *pro parte* l'expérience acquise : les caractéristiques du milieu, tant biologiques qu'abiotiques, devraient induire des options originales.

Toute étude de cas devra cependant s'attacher à répondre aux impératifs (1) - (4) ; en ce sens, le choix de la variable X peut être guidé par les appréciations qualitatives rassemblées dans le tableau suivant :

	Variable indicative et compartiment caractérisé	Conformité aux critères : (1)	(2)	(3)	(4)
Domaine benthique	Macrofaune vagile	+/-	+/-	-	-
	Macrofaune sessile				
	Biomasse du peuplement total	+	+/-	-	+
	Production	+	-	-	+
	Richesse et diversité spécifique	+	-	-	+
	Méiobenthos	+/-	-	+	+
	Microphytobenthos	+/-	-	+	+
Domaine pélagique	Zooplancton (méso-, micro-)				
	Biomasse du peuplement total	+	+/-	+	+/-
	Production	+	-		
	Richesse et diversité spécifique	+	-		
	Phytoplancton				
	Biomasse chlorophyllienne	+	+	+	+/-
	Production	+	+	+	+/-
	Richesse et diversité spécifique	+	-		

Les réponses aux critères (3) et (4) sont étroitement dépendantes des caractéristiques de la stratégie d'échantillonnage ; ici est considérée la pratique la plus fréquente, *i.e.* un pas de temps de l'ordre du mois et un pas d'espace de l'ordre du kilomètre.

Comme toute classification, celle proposée ci-dessus procède d'une part d'arbitraire, et l'interprétation du tableau appelle plusieurs remarques :

- La conformité aux critères y apparaît comme relevant d'un jugement dépourvu de nuance (+ dans l'affirmative, - dans le cas contraire, +/- lorsqu'une appréciation par trop tranchée ne peut être raisonnablement défendue). Cette formulation abrupte ne cause guère d'embarras pour les critères (1) et (2), qui ne prêtent généralement pas à ambiguïté ; mais les positions affichées relativement aux critères (3) et (4), conditionnées par les enseignements tirés de l'étude des sites largement ouverts de la Manche, pourraient être sujettes à révision dans un contexte différent.

- Deux compartiments ne sont pas mentionnés : les poissons (bien que les poissons démersaux soient implicitement inclus dans la macrofaune benthique vagile), et les bactéries. Concernant les premiers, compte tenu de leur mobilité à partir d'un certain stade de développement, les études d'impact (de celui des centrales nucléaires en particulier) se sont généralement limitées à l'estimation des quantités de juvéniles piégés par les installations de pompage (de nombreuses autres voies de recherche peuvent toutefois être explorées ; voir à ce sujet SINDERMANN, 1984). Une conclusion imagée, généralement admise, énonce que l'influence d'une centrale sur un stock est en gros équivalente à celle d'un chalutier. Là encore, il est difficile d'affiner les conclusions en l'absence d'estimations précises susceptibles d'être introduites dans un modèle prévisionnel ; les coefficients de mortalité naturelle par classe d'âge sont de ce point de vue spécialement critiques (*cf.* RAGO, 1984). Quoiqu'il en soit, la présence de nourrisseries côtières à proximité d'un site devrait suffire à justifier une étude particulière. Quant aux bactéries marines, les problèmes techniques que posent à l'heure actuelle tant la saisie que le dépouillement des échantillons, l'interprétation des résultats ainsi que les coûts associés, conduisent à considérer qu'à moyen terme il resterait aventureux d'en faire la clef de voûte d'une étude de surveillance (pour des raisons analogues, le nanoplancton et le picoplancton ne sont pas considérés ici). Ajoutons que le macrophytobenthos ne figure pas non plus parmi les sous-systèmes examinés ; il montre souvent une colonisation irrégulière des substrats côtiers (du fait de la diversité des caractéristiques des affleurements rocheux, comme des conditions d'exposition à différents facteurs abiotiques). Cependant, ainsi qu'il l'a été souligné, chaque site doit motiver une étude de cas, qui ne doit pas systématiquement écarter ce compartiment.

- Enfin, la liste des définitions envisagées pour la variable indicative ne prétend nullement à l'exhaustivité. Elle n'offre qu'un reflet des choix les plus usuels, mais ne doit pas être comprise comme exclusive d'alternatives novatrices (*e.g.* indices biochimiques révélateurs de dérèglements physiologiques). Il ne faut pas non plus oublier que les arguments proposés sont orientés par le souci d'éclairer les choix préalables à la mise en oeuvre de tests statistiques ; ce n'est évidemment pas la seule manière d'entreprendre l'étude d'un impact, point sur lequel il avait été insisté en introduction.

Avec ces remarques présentes à l'esprit, plusieurs enseignements peuvent être tirés du tableau récapitulatif :

. Les indices qui nécessitent comptage et identification d'unités taxinomiques (*e.g.*, diversité) devront être écartés, eu égard au critère prioritaire (2). Cette conclusion s'applique spécialement au phytoplancton, dont les peuplements présentent une grande variété d'espèces, ainsi qu'à un moindre degré au mésozooplancton. De surcroît, l'analyse de la structure floristique (ou faunistique) des prélèvements fait intervenir plusieurs niveaux de sous-échantillonnage, qui entraînent une augmentation de la variance des estimations. De ce dernier point de vue, le problème est moins aigu pour la macrofaune benthique, qui par ailleurs peut être caractérisée par quelques espèces dominantes d'identification souvent assez aisée.

. La macrofaune benthique sessile offre une possibilité attractive, son handicap étant sa forte inertie, comparée à celle des organismes planctoniques (critère (3)). De sorte que ce compartiment constitue plutôt le matériel de choix pour l'étude des modifications progressives et des effets cumulés engendrés par l'aménagement, problématique différant assez sensiblement de celle traitée ici (voir par exemple l'étude de BAMBER & SPENCER, 1984).

. L'intérêt du domaine pélagique tient à la simplicité technique de l'échantillonnage, et au taux de renouvellement rapide des populations zoo- et surtout phytoplanctoniques. Etant admis qu'il ne peut pas être raisonnablement attendu de pousser l'analyse des échantillons jusqu'à l'identification des espèces (*vide supra*), les compartiments "mésozooplancton" et "microphytoplancton" satisfont assez bien les critères de choix. Lorsque les prélèvements ne sont pas trop contaminés par du matériel détritique, qui a pour effet de

polluer l'estimation de la biomasse mésozooplanctonique, cette dernière peut être retenue pour variable indicative. Le phytoplancton présente toutefois des avantages supplémentaires : brièveté de la durée des générations (*e.g.*, environ 5 jours pour un taux de croissance spécifique de 0.2 jour^{-1}), et caractérisation de sa présence par la chlorophylle. En outre, la quantité de chlorophylle (la "biomasse chlorophyllienne") peut être estimée *in situ* de manière automatique par fluorescence, et une bonne approximation des potentialités de production primaire peut être obtenue conjointement par fluorescence induite au DCMU (méthode appliquée entre autres par DEMERS *et al.*, 1985). Ces atouts s'accompagnent d'une contrepartie : la diffusion des cellules phytoplanctoniques dans le milieu contribue à engendrer des corrélations spatiales. L'ensemble de ces considérations amène à traiter la question des échelles d'observations.

4.2. LES ECHELLES D'OBSERVATIONS : CHOIX DES FENETRES SPATIO-TEMPORELLES

Les phénomènes qu'étudie l'écologiste sont spatialement structurés ; ce ne sont pas les structures elles-mêmes qui sont ici l'objet de l'étude, toutefois l'existence depuis longtemps reconnue d'une organisation dans l'espace des organismes vivants implique de considérer de quelle manière ils se déploient dans la zone d'étude avant de décider de l'emplacement des stations de prélèvement. En effet, pour employer le langage des géostaticiens, la variable indicative est "régionalisée".

Une abondante littérature est consacrée à la répartition spatiale des êtres vivants, et porte aussi bien sur les méthodes de détection, de description, et de quantification du phénomène, que sur son déterminisme. Ainsi est-il actuellement admis que l'agencement spatio-temporel des organismes planctoniques est contrôlé par les processus physiques. Concernant le phytoplancton, le rôle prépondérant de l'énergie turbulente et de la stabilité de la colonne d'eau ont été mis en évidence ; par conséquent, en milieu côtier, la morphologie des essaims planctoniques est gouvernée (entre autres) par le vent (*e.g.* THERRIAULT & PLATT, 1981), par la propagation de la marée (*e.g.* DEMERS & LEGENDRE, 1982 ; LEVASSEUR *et al.*, 1983). A plus basse fréquence temporelle, la dynamique de l'écosystème est influencée par la circulation résiduelle (NIHOUL & HECQ, 1984).

Ces quelques références ne constituent qu'une liste excessivement limitative des travaux publiés, et leur thématique commune (contrôle par les processus physiques) ne se rapporte qu'à l'une des nombreuses facettes du problème. Cette présentation restrictive est délibérée, elle est ici justifiée par la référence qu'offrent les conditions rencontrées au voisinage des centrales électronucléaires : la nécessité d'une dilution rapide du panache thermique a conduit à sélectionner les sites hydrologiquement favorables à l'installation d'une centrale d'après l'intensité des processus de mélange et l'amplitude de la dérive (*i.e.*, le déplacement résiduel de la masse d'eau en une marée). Ainsi, face aux quatre centrales construites sur le littoral de la Manche règnent d'assez forts courants alternatifs de marée, de l'ordre du mètre par seconde. Leur orientation est grossièrement parallèle à la côte au voisinage de celle-ci, et entraîne une "excursion" des masses d'eau de plusieurs kilomètres (*cf.* BOULOT & HAUGUEL, 1981).

Au plan des caractéristiques de l'écosystème (spécialement de sa composante pélagique), cette situation a vraisemblablement pour effet de propager les similitudes sur de plus grandes distances parallèlement à la côte que perpendiculairement à celle-ci. Pour statuer sur cette présomption, comme pour évaluer l'emprise des corrélations spatiales, une étude structurale des réalisations x_i de la variable indicative peut être entreprise. Cela vaut aussi bien pour l'exemple qui vient d'être évoqué que dans un contexte plus général. En considérant X comme une variable régionalisée (*cf.* DELHOMME, 1978), c'est en fait la réalisation x(u) de la fonction aléatoire X(u) qui est étudiée, où u désigne un point de l'espace, ici à 2 dimensions. Une fonction aléatoire (F.A. en abrégé) peut donc être comprise comme une variable aléatoire ayant une infinité de composantes, chacune correspondant à un point de l'espace.

Une méthode de détection des structures consiste à construire le variogramme expérimental ; l'un des attraits de cet outil mathématique est qu'il existe sous des hypothèses faibles, relativement à celles qui sont nécessaires à la définition d'une covariance (*vide infra*), à savoir que les accroissements de la F.A. X(u) doivent être stationnaires d'ordre 2. C'est-à-dire :

(1) les accroissements sont d'espérance nulle :

$$E\{X(u+h)-X(u)\} = 0 \quad \forall u$$

(2) et la variance des accroissements de la F.A. est indépendante de la position du point d'appui u, elle ne dépend que de l'éloignement h par rapport à ce point :

$$V\{X(u+h)-X(u)\} = 2\gamma(h) \quad \forall u$$

Ces deux conditions forment "l'hypothèse intrinsèque" (qui font de X(u) une F.A. intrinsèque), et la fonction $\gamma(h)$ est habituellement appelée variogramme (par abus de langage, car il faudrait en toute rigueur parler de demi variogramme). S'il arrive que l'espérance des accroissements soit localement non nulle, une hypothèse encore plus faible peut être formulée, portant sur la stationnarité des accroissements d'ordre 1, 2, ... ; elle ne sera pas évoquée ici.

Les conditions d'existence d'une covariance C(h) sont plus contraignantes : la stationnarité est cette fois requise pour les deux premiers moments de la F.A. elle-même (au lieu de ses accroissements). Formellement :

$$E\{X(u)\} = \mu \quad \text{, constante } \forall u \text{ ;}$$

$$E\{(X(u+h)-\mu)(X(u)-\mu)\} = C(h) \quad \text{, ne dépendant que de h.}$$

Il s'agit là d'une hypothèse beaucoup plus forte, qui implique une répétitivité spatiale de la F.A. (plus exactement, l'invariance par translation de ses deux premiers moments). Quand cette hypothèse est vérifiée, et que donc la covariance existe, alors cette dernière est liée à la variance C(o) des données et au variogramme par la relation simple :

$$\gamma(h) = C(o) - C(h)$$

L'estimation du variogramme se fait pas à pas, pour différentes valeurs de l'écartement h, les points représentant l'évolution de $\gamma(h)$ en fonction de h étant chacun obtenus à partir de N(h) couples d'observations $x(u_i)$ distantes de h :

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} (x(u_i+h) - x(u_i))^2$$

L'examen du comportement du variogramme expérimental $\hat{\gamma}(h)$ est riche d'information : il réalise une description graphique synthétique de la structure du phénomène étudié (quelques exemples commentés sont présentés par CLARK, 1979). En général, $\hat{\gamma}(h)$ croît avec h lorsque les observations diffèrent statistiquement d'autant plus qu'augmente la distance qui les sépare, la vitesse de croissance permettant d'apprécier la plus ou moins forte continuité de la variable (*cf.* DELHOMME, *op. cit.*). Quand cette dernière est fortement irrégulière, il existe une discontinuité du variogramme en h = 0 (effet de pépite), engendrée soit par des microstructures dont l'échelle est inférieure au pas de l'échantillonnage, soit par des erreurs de mesures.

Le variogramme expérimental doit aussi être analysé pour les grandes valeurs de h : lorsque h devient supérieur à la distance d'influence d'un point expérimental (la "portée" du variogramme), les valeurs de $\hat{\gamma}(h)$ se stabilisent autour d'un palier : pour les valeurs de la distance qui dépassent la portée (si elle existe, *i.e.* si le variogramme est borné), les corrélations s'annulent. Appliquée au problème du positionnement des stations de mesure, l'utilisation d'un variogramme borné est particulièrement intéressante : la valeur de la portée indique la séparation minimale à ménager entre la zone "impactée" et la zone témoin (*cf.* §3.1.) pour éliminer les corrélations spatiales.

Il peut être noté ici que MACKAS (1984) a proposé, pour étudier l'organisation spatiale des communautés pélagiques, de construire une "fonction de structure", laquelle présente quelque parenté conceptuelle avec le variogramme (elle en diffère toutefois radicalement si l'on considère les potentialités d'inférence et d'interpolation attachées à ce dernier). Ainsi qu'il vient d'être dit, le variogramme représente l'évolution de la variance des accroissements de la F.A. en fonction d'une distance physique ; et en pratique, le variogramme expérimental montre l'évolution de l'écart quadratique moyen entre réalisations d'un processus spatial (*i.e.*, de mesures d'une quantité scalaire) en fonction de leur éloignement géographique. La fonction de structure de MACKAS (*op. cit.*) en diffère en ce qu'elle représente la variation d'une dissimilitude de composition floristique (ou faunistique) entre unités d'échantillonnage, en fonction de la distance géographique qui les sépare ; cette dissimilitude entre deux prélèvements est elle-même une distance, mais définie entre les deux points qui les représentent dans un espace possédant autant de dimensions qu'il y a d'espèces reconnues dans les différents

échantillons. Autrement dit, la fonction de structure illustre l'atténuation de la ressemblance multivariée entre échantillons de plus en plus éloignés sur le terrain.

Dans le cadre d'une étude de surveillance, l'apport de la fonction de structure pourrait être pertinent si le profil du spectre des différentes espèces de la communauté pélagique était la cible de l'étude. Mais compte tenu des critères avancés au §4.1., la variable indicative sera plutôt une caractéristique du peuplement total (*e.g.* la biomasse chlorophyllienne, par nature "régionalisée"), auquel cas le variogramme, s'il existe, constituera l'outil le plus commode de l'étude structurale.

La stratégie de surveillance est aussi définie par son échelle d'observation temporelle. A cet égard, pour s'en tenir à l'exemple des sites côtiers de la Manche, la succession au cours de l'année des populations planctoniques a été particulièrement bien décrite. Des observations effectuées pour l'étude du milieu au voisinage des centrales nucléaires, il ressort une certaine régularité du "pattern" saisonnier des relais entre espèces holo- et mérozooplanctoniques (*e.g.* GROS, 1981), qui s'oppose à la non-reproductibilité des observations effectuées sur les cortèges floristiques phytoplanctoniques d'une année à l'autre, même si quelques tendances communes peuvent être dégagées (RYCKAERT *et al.*, 1983 ; GROS & RYCKAERT, 1983). Dans l'un comme dans l'autre cas, pour un pas de temps *ca.* mensuel, la source majeure de variabilité temporelle est la variation saisonnière, ainsi qu'il est de règle en domaine tempéré. Et le contrôle du milieu face au site de Gravelines a montré que depuis que cette centrale est en sercice, elle n'a pas eu pour effet de modifier les grands traits de cette forte "saisonnalité" de l'écosystème pélagique. Il s'est en fait révélé que l'impact n'est actuellement perceptible que dans le champ proche des rejets (*cf.* GENTIEN *et al.*, 1986), et qu'il faudra accorder au moins autant d'attention à sa détection spatiale qu'aux modifications temporelles. Au surplus, la chronologie de l'évolution annuelle des communautés pélagiques n'apparaissant pas perturbée, il peut être considéré qu'il n'est pas indispensable de surveiller le milieu douze mois par an : l'effort d'échantillonnage pourrait être ainsi recentré sur une ou deux périodes (il sera parlé plus loin de blocs temporels) de courte durée, par exemple une période "hors chloration" (lorsque la température de l'eau est inférieure à 10°C), et une période durant laquelle des ions hypochloreux sont injectés à l'entrée

des canalisation de refroidissement (quand la température du milieu dépasse 10°C). Une stratégie "économique" pourrait donc être élaborée sur la base de deux blocs spatiaux croisés avec, par an, deux blocs temporels.

D'un point de vue plus général, la littérature consacrée aux protocoles dont l'objectif est la mise en évidence statistique de l'impact fait toutefois état de stratégies plus élaborées, qui vont être maintenant discutées ; l'examen de ces travaux offre plus matière à réflexion qu'exemple à reproduire, les stratégies les plus complètes étant aussi assorties d'un certain nombre d'inconvénients.

Dans la gamme des méthodes examinées aux chapitres 1, 2 et 3, deux comparaisons sont privilégiées : "état de référence" *vs.* période de "surveillance" (approche diachronique, chapitres 1 et 2), et zone témoin *vs.* zone soumise à l'impact durant la période de surveillance (détection spatiale de l'impact, chapitre 3). Un choix approprié des fenêtres d'observation, voire des contraintes pratiques, peuvent légitimer l'étude préférentielle de l'un des aspects au détriment de l'autre ; c'est-à-dire adopter une stratégie focalisée soit sur la structure spatiale, soit sur la variabilité temporelle de la variable indicative. Il peut toutefois être considéré que l'appréhension de l'impact ne doit pas éluder au premier abord l'une de ses composantes. En conséquence, des protocoles ont été proposés afin d'aborder conjointement les deux questions traitées séparément jusqu'ici. Ces protocoles, du fait de leur caractère plus "intégré", requièrent un certain nombre de conditions d'application inventoriées et hiérarchisées dans un arbre de décision dû à GREEN (1979).

Globalement, la démarche consiste à évaluer l'influence de plusieurs facteurs (ou "critères") identifiés *a priori* sur les valeurs observées de la caractéristique X sensible à l'impact. Cela implique une nouvelle fois le recours à des tests d'hypothèses. Plus précisément, en supposant que les réalisations de X sont engendrées par un modèle simple dans lequel les effets des facteurs sont additifs, le problème revient à tester si les termes que ces critères ajoutent à l'espérance générale μ de X sont statistiquement significatifs.

Une large panoplie, voire une pléthore de techniques, regroupées sous le vocable collectif d'analyse de la variance, permet de résoudre le type

de problème qui vient d'être énoncé. Préalablement à la mise en application de l'une d'entre elles, des choix devront donc être opérés, concernant notamment :

- le nombre de facteurs distincts à prendre en compte ;

- la nécessité de leur adjoindre des termes d'interaction (et dans l'affirmative, jusqu'à quel ordre ?), phénomène qui complique la relation d'addivité simple des facteurs, pouvant interdire d'identifier leurs effets séparés ;

- plus fondamentalement, la nature des effets des critères doit être rigoureusement définie : si ces effets sont considérés comme certains, la décomposition de X suivant un modèle additif repose sur l'estimation des paramètres du modèle linéaire classique (modèle I). Si au contraire les effets sont aléatoires, alors toute réalisation x de X s'exprime comme la somme d'une moyenne générale et d'une combinaison linéaire de variables aléatoires, plus un terme d'erreur (modèle II) ; la résolution du problème consiste dans ce cas à estimer les coefficients des éléments de la matrice de covariance des observations. Il faut noter que la dichotomie modèle I *vs.* modèle II ne recouvre pas la totalité des méthodes existantes, et que des modèles mixtes peuvent également être envisagés.

Les quelques choix qui viennent d'être inventoriés sont évidemment conditionnés par la nature des données, ils interagissent avec le protocole de leur saisie, et demeurent assujettis à la vérification des hypothèses sur lesquelles seront fondées les inférences. Ce dernier point va être maintenant précisé à l'aide d'un exemple ; la présentation diffèrera de celle adoptée au chapitre 3 : dans ce qui va suivre ne sera mentionnée que l'approche paramétrique usuelle du problème.

Le modèle d'étude d'impact discuté par MILLARD et LETTENMAIER (1986) peut ainsi être présenté à titre d'illustration ; selon ces auteurs, les écarts des réalisations de X à sa moyenne générale μ sont redevables des effets suivants :

$$X_{jklmn} = \mu + A_j + B_k + C_l + AB_{jk} + AC_{jl} + BC_{kl} + ABC_{jkl} + \eta_{jklm} + \varepsilon_{jklmn} \qquad \{27\}$$

où :

A_j correspond à la composante spatiale de la caractéristique et désigne l'effet de la station j(j = 1, ..., J).

B_k exprime l'impact de l'aménagement ; l'indice k ne prend donc que deux valeurs : k = 1 (avant) et k = 2 (après mise en place).

C_l rend compte de la variation temporelle intra-annuelle. Cet effet "saisonnier" est estimé en allouant les campagnes d'échantillonnage à L blocs temporels (qui peuvent être définis de manière à réaliser une partition de l'année).

A ces trois effets (considérés comme fixés par MILLARD & LETTENMAIER, *op. cit.*, qui attribuent donc ce modèle à la catégorie "modèle I") sont ajoutés des termes d'interaction (AB, BC, ABC), ainsi que deux composantes aléatoires :

η_{jklm} : l'unité d'échantillonnage est ici la station (fixée) ; dans le bloc temporel l, et en situation k (*i.e.* "avant", ou bien "après"), chaque station j fait l'objet de M campagnes d'échantillonnage. La variabilité des observations entre ces différentes campagnes est traduite par la dispersion des réalisations de la variable aléatoire η_{jklm}.

ε_{jklmn} : une station n'est généralement pas étudiée exhaustivement ; cette unité d'échantillonnage est elle-même sondée par sous-échantillonnage : au cours d'une campagne donnée sont effectués N réplicats en une même station, dont la variabilité est exprimée par le "terme d'erreur" ε_{jklmn}.

Il faut souligner que l'erreur dont il est ici question n'est pas équivalente à celle commise dans les expériences contrôlées en laboratoire : dans ce dernier cas, les réplicats reflètent l'erreur de mesure. En revanche, dans l'étude des variations d'une caractéristique biologique en milieu marin, à l'erreur de mesure se superpose un bruit causé par la variabilité spatio-temporelle intervenant à petite échelle ; cela peut engendrer des difficultés, s'il arrive qu'en une station donnée la variance entre réplicats soit plus forte que la variance entre campagnes d'un même bloc temporel.

Pour tenter des inférences à partir du modèle {27}, il est nécessaire de respecter le corps d'hypothèses suivantes : (1) les termes "déterministes" (facteurs principaux et interactions) jouent de manière additive. Et concernant les composantes aléatoires, par ordre d'importance décroissante : (2) les variables η_{jklm} sont indépendantes, de même que les variables ε_{jklmn} ; (3) elles sont l'une et l'autre de variance stable (homoscédasticité) ; (4) elles sont gaussiennes. Formellement, (2), (3) et (4) s'énoncent :

$$\eta_{jklm} \overset{iid}{\sim} N(0,\sigma_\eta^2) \quad \text{et} \quad \varepsilon_{jklmn} \overset{iid}{\sim} N(0,\sigma_\varepsilon^2)$$

Si les données de base ne respectent pas (3) et (4), elles peuvent faire l'objet d'une transformation. Ce prétraitement ne permet cependant pas de se ramener sous la condition (2) lorsque celle-ci n'est pas vérifiée ; il existe en effet deux sources d'interdépendance possibles : une corrélation temporelle entre les erreurs, qui provient de ce qu'une observation saisie à un instant donné dépend, avec une rémanence plus ou moins accentuée selon la dynamique propre à la variable étudiée, des états antérieurs de l'écosystème. D'où la nécessité de l'adéquation du pas de temps de l'échantillonnage à la "mémoire" de la variable X. La seconde source est l'existence des corrélations spatiales, qui entraînent qu'en moyenne les résultats observés en deux stations sont, en deçà de la portée du phénomène, d'autant plus semblables entre eux que ces stations sont proches. La manière de se prémunir contre ces deux sources de corrélation a déjà été discutée.

Quant à l'influence de ces liaisons sur les conclusions tirées de l'analyse, elle a été étudiée par MILLARD *et al.* (1985), à partir de données réelles ainsi qu'à l'aide de simulations. Des résultats de ces auteurs, il ressort principalement que :

- Les corrélations temporelles et/ou spatiales ont pour effet d'accroître l'erreur de première espèce, *i.e.* la probabilité de décider à tort en faveur de l'existence d'un impact ; les tests sont effectués à un seuil supérieur au seuil nominal.

- Aux rythmes d'échantillonnage les plus couramment employés (*e.g.* mensuel ou bimensuel), l'incidence des corrélations spatiales l'emporte généralement sur celle des corrélations temporelles, dès lors que la variable étudiée caractérise un compartiment de l'écosystème possédant à la fois un "turnover" élevé et de grandes capacités de diffusion dans la zone d'étude.

Ces remarques ne concernent que la composante aléatoire du modèle {27}. La condition (1), relative à ses termes "déterministes", doit aussi être examinée avec attention : le modèle additif ne constitue qu'une simplification, car il existe en règle générale une confusion partielle entre les diverses sources de variation (*cf.* LAUREC *et al.*, 1981 : confusion des variations spatiales et temporelles, contamination des différentes fréquences temporelles). Ces phénomènes, qui sont le fait de limitations inhérentes à l'échantillonnage, incitent ces auteurs à parler de "sources de variations apparentes". En ce sens, l'hypothèse de "bruit blanc pluriannuel" introduite au §1.2. est implicitement contenue dans la formulation du modèle {27} : le terme C_1 n'exprime que la variabilité temporelle intra-annuelle (et non interannuelle), et cela suppose que dans chaque bloc temporel la variable X soit dépourvue de tendance.

Outre les difficultés engendrées par la nécessité d'une bonne conformité aux hypothèses requises par la technique, il demeure un certain nombre de complications propres à l'application du modèle complet (*i.e.* incluant les effets des critères plus l'ensemble des interactions possibles). Tout d'abord, la présence des termes d'interaction d'ordre 1 et 2 rend délicate la formulation et la compréhension des hypothèses testées. Si l'interaction ne peut être négligée, il n'est pas possible d'apprécier les effets séparés des facteurs ; or le problème consiste précisément à évaluer l'effet du critère B (impact de l'aménagement), c'est-à-dire tester H_o : $\Delta=0$, contre H_1 : $\Delta\neq0$, où $\Delta = B_1-B_2$. Cette décision simple ne peut s'appuyer sur la règle :

$$MS(B) \,/\, MS(\eta) > F_{\alpha,\nu1,\nu2} \Rightarrow \text{rejet de } H_o \text{ au seuil } \alpha$$

qu'à la condition que les interactions faisant intervenir B soit insignifiantes (ici, $\nu_1 = 1$, $\nu_2 = 2JL(M-1)$, et MS(.) désigne le carré moyen).

En pratique, lorsque le modèle complet est retenu d'emblée, cela implique les tests "en chaîne" de la séquence d'hypothèses : $ABC_{jkl} = 0, \forall j,k,l$; $AB_{jk} = 0\ \forall j,k$; $BC_{kl} = 0\ \forall k,l$; et enfin $\Delta = 0$. L'expérience montre qu'il n'est pas exclu d'aboutir à des résultats scabreux (*e.g.* AB non significatif, BC et ABC significatifs, ...), et de surcroît la faible robustesse de la procédure doit inciter à suspecter que la détection d'une interaction (spécialement d'ordre élevé) peut n'être causée que par une observation légèrement aberrante. Abstraction faite de ces inconvénients, s'il existe réellement une interaction saison x impact (qui entraîne le rejet de l'hypothèse $BC_{kl} = 0\ \forall k,l$), il devient

alors sans objet de tester l'hypothèse $\Delta = 0$, et les données devront être examinées séparément dans chaque bloc temporel ; cela conduit à envisager à nouveau la problématique évoquée au chapitre 3.

Les précédentes remarques incitent à privilégier le modèle n'incluant que le minimum d'interactions pertinentes. Selon ce point de vue, pour la commodité de l'interprétation autant que pour des raisons techniques, le terme ABC d'ordre 2 ne sera pas conservé. De même, il peut être admis (ou mieux, vérifié au cours d'une étude pilote) que la source de variation temporelle affecte pareillement les J stations, et donc que l'interaction A x C peut être négligée. En contrepartie, il serait imprudent de décider *a priori* que l'effet de l'aménagement, s'il existe, intervient de manière équivalente en toutes les stations et dans tous les blocs temporels : l'interaction B x C ne doit donc pas être écartée systématiquement, de même que l'interaction A (stations) x B ("avant"/"après"). Ce second terme correspond à l'influence spatiale de l'impact, effet différentiel admis dès le départ de l'étude, et dont l'emprise est évidemment limitée. Il n'est donc guère intéressant de tester l'interaction A x B (qui est une donnée du problème), et il est préférable d'aménager le protocole en y incluant des stations de contrôle situées dans la zone "témoin" (*cf.* §3.1.). Dans cette logique, la zone témoin est définie comme celle pour laquelle il peut être *a priori* considéré que $B_1 - B_2 = 0$. Les conséquences de l'aménagement dans le champ proche sont alors quantifiées par l'influence des critères sur les différences observées entre stations appariées : l'ensemble des J stations A_j est donc partagé en deux groupes égaux, de façon à former J/2 couples de stations : $j \in C$ (stations situées en zone témoin) et $J \in I$ (stations situées en zone "impactée", *i.e.* dans le champ proche des rejets). La variable analysée est dans ce cas :

$$D_{jklmn} = X_{\substack{jklmn \\ j\in C}} - X_{\substack{jklmn \\ j\in I}} \qquad j = 1, \ldots, J/2$$

dont les variations sont expliquées par le modèle :

$$D_{jklmn} = \mu + A_j + B_k + C_l + BC_{kl} + \eta_{jklm} + \varepsilon_{jklmn}$$

les différents termes du modèle admettant la même définition que pour l'équation {27}, bien qu'ils ne recouvrent plus tout à fait la même réalité. En particulier, si l'ensemble du secteur étudié est spatialement homogène (pas de différence systématique mise en évidence entre champ proche et zone témoin pendant la période d'état de référence), la moyenne μ peut être raisonnablement tenue pour nulle.

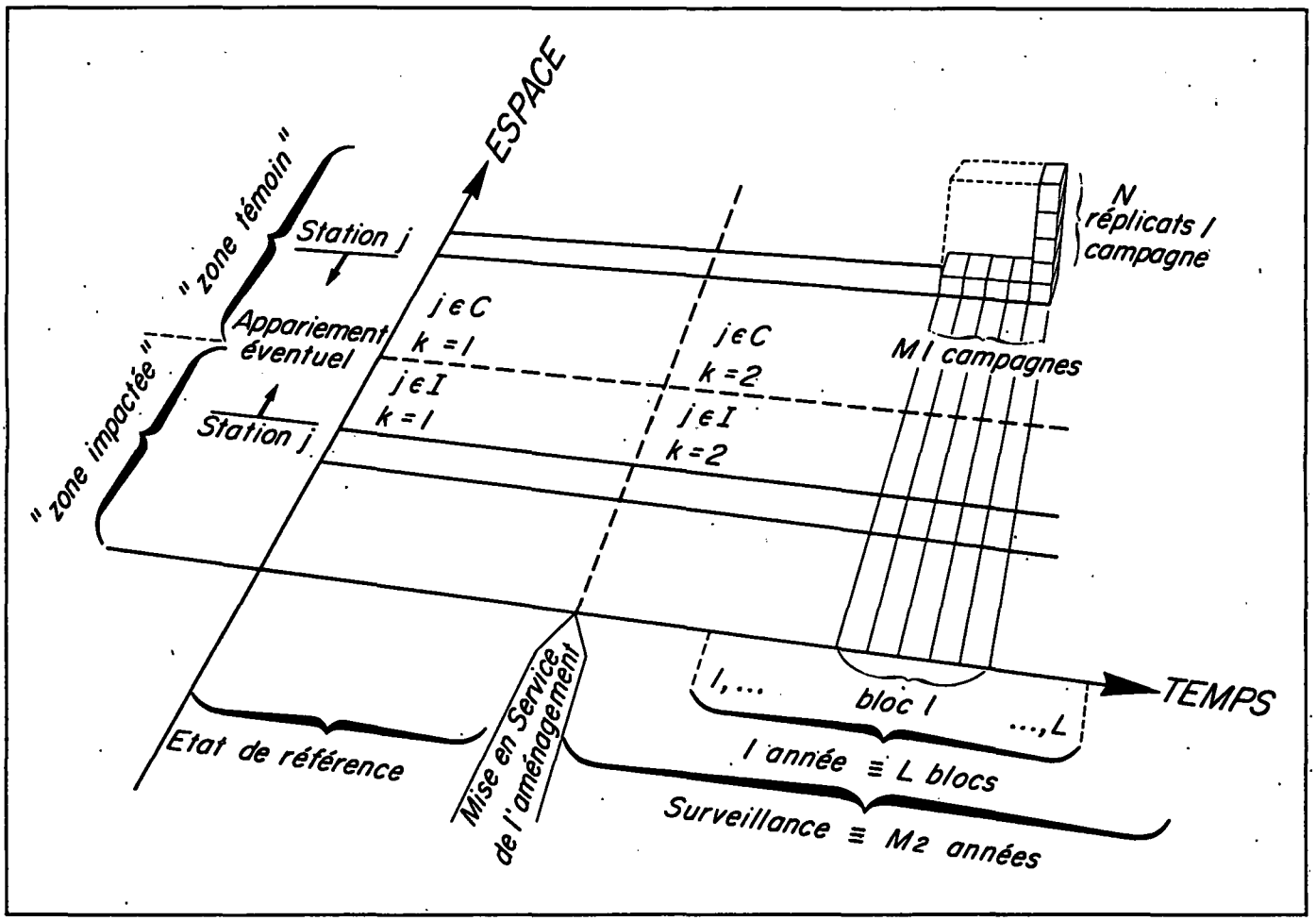


Figure 4 - Structure générale des protocoles discutés au paragraphe 4.2.

Cette option ne constitue pas le seul choix possible : la littérature fait état de diverses variantes (*e.g.* BERNSTEIN & ZALINSKI, 1983 ; EVANS & SELL, 1983), qui toutes procèdent fondamentalement du schéma conceptuel présenté à la figure 4 . L'ensemble des unités d'échantillonnage est scindé en quatre strates spatio-temporelles, les limites entre strates correspondant d'une part à la distinction stations "impactées" ($j \in I$) - stations de contrôle ($j \in C$), et d'autre part à la différence d'état "avant-après" ($k = 1$ ou 2 respectivement). Les nuances entre stratégies tiennent essentiellement à la nature des critères (modèle I, II ou bien mixte), au fait que les stations sont appariées ou non, et aussi à la manière de définir l'effet de l'aménagement (soit en tant que facteur, soit en tant que niveau du facteur temps).

La multiplicité des stratégies ne doit pas occulter leur objectif commun : tester l'effet éventuel d'un aménagement sur le milieu. Cela nécessite, comme pour les développements présentés au premier chapitre, le calcul de la puissance $\pi(\Delta)$ de la procédure, où $\Delta = B_1 - B_2$, et cela en vue d'une planification rationnelle de l'échantillonnage. Etant donnée la diversité des approches concevables, seules ne pourront être ici formulées que quelques règles très générales.

En premier lieu, et ce point a déjà été signalé, le calcul de $\pi(\Delta)$ n'est fondé qu'en l'absence de toute interaction significative incluant le facteur B. Ensuite, et sous cette condition, la règle de décision repose sur le calcul d'une statistique qui est un rapport de carrés moyens (celui du facteur B rapporté au carré moyen "intra-groupe"), pondérés par leurs d.d.l. respectifs ν_1 (égal à 1 pour $k = 1,2$) et ν_2 (qui dépend du nombre des stations, des blocs temporels, et des campagnes d'échantillonnage par bloc). Dans le contexte inférentiel usuel, cette statistique suit une loi de Fisher-Snedecor à ν_1 et ν_2 d.d.l., centrée sous H_o, et généralement non centrée sous H_1, de sorte que la probabilité d'erreur de seconde espèce est évaluée à l'aide d'une équation de la forme :

$$\beta(\Delta) = 1 - \text{Proba}\{NCF(\xi) > F_{\alpha,\nu_1,\nu_2} \mid H_1\}$$

où NCF est une variable aléatoire obéissant à une loi de Fisher non centrée, et dont le paramètre de décentrement ξ est une fonction de l'écart Δ exprimé relativement à l'écart-type "résiduel" (voir par exemple l'équation 6 de MILLARD & LETTENMAIER, *op. cit.*). Il s'ensuit que l'évaluation précise de $\beta(\Delta)$

est assez lourde, en ce qu'elle doit inclure un algorithme de calcul des valeurs de la fonction de répartition de Fisher non centrée. Pour pallier cette pesanteur, il peut être plus simplement fait appel à l'approximation proposée par LAUBSCHER (1960), à savoir appliquer une transformation normalisante à la variable NCF(ξ). Des trois transformations examinées par cet auteur (Argch, racine carrée, racine cubique), la seconde est celle qui apparaît fournir les résultats les plus satisfaisants dans la gamme des valeurs de ν_1, ν_2 et ξ considérées (*op. cit.*, tab. 2 p. 1110). Considérant que NCF(ξ) est définie comme le rapport de deux khi-deux indépendants :

$$NCF(\xi) = (NC\chi^2(\xi)/\nu_1)/(\chi^2/\nu_2)$$

celui du numérateur étant non centré à ν_1 d.d.l., avec ξ pour paramètre de décentrement, le choix de la transformation racine carrée est justifié comme suit : sachant que, approximativement

$$\sqrt{2NC\chi^2(\xi)} \sim N(\sqrt{2(\nu_1+\xi)-(\nu_1+2\xi)/(\nu_1+\xi)} , (\nu_1+2\xi)/(\nu_1+\xi))$$

$$\sqrt{2\chi^2} \sim N(\sqrt{2\nu_2-1} , 1)$$

et dans la mesure où ces deux approximations sont valides, alors la variable

$$R = \sqrt{\nu_1 NCF(\xi)/\nu_2}$$

se comporte comme le rapport de deux aléas normaux indépendants. LAUBSCHER applique à R le théorème qui énonce que si Y_1 et Y_2 sont indépendantes et de lois respectives $N(\mu_1,\sigma_1^2)$ et $N(\mu_2,\sigma_2^2)$, alors la quantité :

$$(\mu_1 Z-\mu_2)\sqrt{\sigma_1^2 Z^2+\sigma_2^2} \quad \text{où} \quad Z = Y_2/Y_1$$

suit une loi voisine de celle de l'aléa N(0,1). En posant :

$$\tau(R) = (R\sqrt{2\nu_2-1}-\sqrt{2(\nu_1+\xi)-(\nu_1+2\xi)/(\nu_1+\xi)})/\sqrt{R^2+(\nu_1+2\xi)/(\nu_1+\xi)}$$

il s'ensuit qu'approximativement : $\tau(R) \sim N(0,1)$. La dérivée de la fonction $\tau(R)$ est strictement positive, et la transformation τ conserve donc les relations d'ordre. Cela entraîne :

$$\beta(\Delta) \simeq \Phi\{\tau(\sqrt{\nu_1 F_{\alpha,\nu_1,\nu_2}/\nu_2})\}$$

où Φ désigne, comme d'usage, la fonction de répartition de l'aléa N(0,1). Cette simplification permet de calculer aisément (*e.g.* MORAN, 1980) la probabilité d'erreur β pour toute combinaison de ξ, ν_1, ν_2 et du seuil α. L'intérêt de pouvoir disposer de ce moyen d'évaluation tient au fait que β (ou Δ lui-même) entre dans la définition d'un critère (*vide infra*) dont l'optimisation nécessite un processus itératif. Cet aspect "économique" du problème va être maintenant considéré.

4.3. CHOIX ECONOMIQUES ET CRITERE D'OPTIMALITE POUR LA DEFINITION D'UNE STRATEGIE

Telle que conduite jusqu'à présent, la discussion n'a principalement abordé la définition de la stratégie d'étude d'impact que sous l'angle de la puissance des procédures de tests associées. Ce point de vue privilégie surtout la prise en compte du niveau maximal admis pour β ; ainsi furent présentés aux chapitres 1, 2 et 3 plusieurs résultats permettant de connaître le nombre d'observations nécessaires pour atteindre un seuil de risque donné.

En pratique, ce n'est pas sous une forme aussi succincte qu'est posée la question, dans la mesure où toute étude d'impact ne se voit impartir qu'un financement limité. La recherche de la "stratégie optimale" intègre donc *ipso facto* les coûts de ses différentes opérations, d'autant plus que les considérations d'ordre économique l'emportent généralement sur celles relatives à la précision des résultats. Ces préoccupations, déjà évoquées au §4.3., conduisent à déterminer la fonction de coût du programme de surveillance. En adoptant un formalisme suffisamment général, celle-ci peut s'écrire :

$$C(N_1, N_2, N_3) = C_o + f_1(N_1)(C_1 + f_2(N_2)(C_2 + f_3(N_3)C_3)) \qquad \{28\}$$

avec : C_o : frais fixes globaux ; outre l'achat (et/ou l'amortissement) du matériel, cette rubrique inclut entre autres le coût de la phase de définition du programme, de sa conduite, ainsi que du traitement et de l'interprétation des données acquises. Il peut être admis en première analyse que C_o est indépendant de N_1, N_2, N_3 définis plus loin.

C_1 : coût d'une campagne d'échantillonnage, comprenant en particulier l'accès à la zone d'étude et le prix de la location du navire.

C_2 : coût de prospection d'une station ; ce volet recouvre l'accès à la station, ainsi que le temps passé à la collecte des échantillons.

C_3 : coût d'un réplicat ; c'est essentiellement le prix à payer pour le dépouillement d'un échantillon au laboratoire.

N_1, N_2 et N_3 désignent respectivement le nombre de campagnes d'échantillonnage, de stations par campagne, et de réplicats par station. Les fonctions f_1, f_2 et f_3 sont monotones croissantes et peuvent admettre une définition simple, *e.g.* :

$$f_i(N_i) = N_i \qquad i = 1, 2, 3$$

Il n'y a toutefois guère intérêt à se restreindre d'emblée à ce cas, qui ne permet pas (pour s'en tenir à un exemple élémentaire) de prendre en compte les possibilités de tarifs forfaitaires ou dégressifs ; c'est donc la formulation générale qui sera retenue.

La probabilité d'erreur β dépend elle aussi des N_i (pour Δ fixé). Cela apparaît clairement au paragraphe précédent, où ces quantités définissent le nombre de d.d.l. ν_2 ; il en est de même des cas traités au premier chapitre, pour lesquels se pose le problème de la meilleure répartition de l'effort total d'échantillonnage entre les campagnes, les stations, ... La définition d'une stratégie d'étude d'impact relève donc typiquement d'un problème d'optimisation, ainsi, Δ étant donné :

$$\text{minimiser } \beta(N_1, N_2, N_3)$$
$$\text{Sous les contraintes : } C(N_1, N_2, N_3) = C_{max}$$
$$\text{et } N_{Li} \leq N_i \leq N_{Ui}$$

Il s'agit là d'une approche "à coût fixé", où Cmax représente le montant du budget consacré au programme. L'optimisation porte sur les paramètres N_i, chacun étant astreint à demeurer dans un intervalle de variation admissible (dont les bornes inférieure et supérieure sont définies par les indices L et U respectivement), défini *a priori*.

En théorie pourrait être conçue de même une approche "à puissance fixée", consistant à minimiser $C(N_1, N_2, N_3)$ sous la contrainte d'égalité $\beta(N_1, N_2, N_3) = \beta$ donné, et avec le même type de contraintes d'inégalité.

Bien qu'en toute logique aussi fondée que la précédente, cette seconde démarche pêche par son manque de réalisme : pour un risque β trop faible, le minimum calculé du coût peut excéder largement le financement disponible.

La première approche apparaît donc la plus judicieuse ; de fait, elle est préconisée par certains auteurs (*e.g.* MILLARD & LETTENMAIER, 1986). Elle peut néanmoins faire l'objet de la réserve suivante : β est une fonction des seuls N_i non seulement pour Δ fixé, mais aussi pour α fixé (*cf.* §1.1.). Et il n'est nullement garanti qu'après optimisation la valeur du β minimal sous contraintes soit voisine de la probabilité α prescrite, à défaut de lui être égale. Cela va à l'encontre de la résolution adoptée au §1.1., et interdit en outre de décider *a priori* de la hiérarchie à accorder aux probabilités d'erreur de première et seconde espèce (ici a été simplement choisi $\beta = \alpha$). Il paraît donc préférable de recommander une approche "à risques et coûts fixés", où le critère est cette fois la "distance" Δ devant être ménagée entre l'alternative et l'hypothèse nulle. D'où le problème :

$$\text{minimiser } \Delta(N_1, N_2, N_3) \quad , \quad \alpha \text{ et } \beta \text{ fixés, avec } \alpha = \beta$$

$$\text{sous les contraintes : } C(N_1, N_2, N_3) = Cmax$$

$$\text{et } N_{Li} \leq N_i \leq N_{Ui}$$

En d'autres termes, l'allocation de l'effort est optimisée de sorte que l'écart pouvant être décelé à des niveaux de risque α et β donnés, et pour un financement Cmax, soit le plus petit possible. Le critère est donc l'alternative H_1 elle-même.

La mise en oeuvre de cette démarche présente un double aspect de simplicité et de difficulté. La simplicité tient à la faible dimension de l'espace paramétrique : égale ici à 3, elle pourrait cependant être légèrement augmentée, par exemple en ajoutant des niveaux de sous-échantillonnage. Il ne faut pas perdre de vue qu'en toute éventualité, le nombre des paramètres

à optimiser restera grotesquement faible en regard de la taille de certains problèmes couramment résolus à l'heure actuelle et qui, suivant la plus ou moins forte non-linéarité de la fonction-objectif, font intervenir de l'ordre de 10^2 à 10^4 paramètres (SHANNO, 1983).

Les difficultés, qui sont d'ordre théorique, tiennent quant à elles d'une part à la nature des paramètres, d'autre part à la formulation du problème. Concernant le premier point : les N_i ne pouvant prendre que des valeurs entières, le problème devrait donc être traité à l'aide des techniques de la programmation entière. La complication peut néanmoins être tournée en considérant N_1, N_2, N_3 comme des nombres réels, puis en arrondissant à l'entier le plus proche les valeurs atteintes à l'optimum. Cet expédient doit cependant être utilisé avec une certaine prudence (*cf.* FLETCHER, 1981, chap. 13).

En second lieu, le problème est contraint. Une première idée consiste à éliminer les contraintes d'inégalité en transformant les paramètres (BOX, 1966) ; soit T une telle transformation :

$$x \in [x_L, x_U] \xrightarrow{T} \tilde{x} = T(x) \in \mathbb{R}$$

par exemple : $\tilde{x} = \mathrm{tg}(\pi((x - x_L)/(x_U - x_L) - 1/2))$

La contrainte d'égalité est ensuite introduite dans le critère en écrivant que le Lagrangien L

$$L(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3 ; \lambda) = \Delta(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3) - \lambda(C(\tilde{N}_1, \tilde{N}_2, \tilde{N}_3) - Cmax)$$

doit vérifier la condition nécessaire d'extrêmum local, *i.e.* être stationnaire :

$$\partial L/\partial \tilde{N}_i = 0 \quad , \quad i = 1, 2, 3 \quad , \text{ et } \partial L/\partial \lambda = 0$$

où λ est un multiplicateur de Lagrange. La résolution de ce système d'équations fournit les valeurs de $\tilde{N}_i$; le retour aux valeurs entières non transformées est obtenu par :

$$N_i = \mathrm{int}(T^{-1}(\tilde{N}_i)) \quad , \; i = 1, 2, 3$$

T^{-1} est la fonction réciproque de T, et l'opération int(.) désigne le passage à l'entier le plus proche ; il doit être vérifié que cette seconde opération n'éloigne pas sensiblement le point (N_1, N_2, N_3) de la région admissible de l'espace paramétrique.

Bien que séduisante, l'application de transformations, si ingénieuses soient-elles, n'est généralement pas prônée : l'introduction de fonctions de pénalité se révèle en pratique plus efficiente (*e.g.* BARD, 1974 ; FLETCHER, 1981). Pour éclairer le principe de la pénalisation (ou plus exactement, dans l'exemple qui va suivre, des "fonctions de barrière"), la solution proposée par FIACCO & Mc CORMICK (1966) peut être brièvement présentée : soit une fonction f d'un vecteur θ de n paramètres, devant être minimisée sous m contraintes d'inégalité $g_i(\theta) \geq 0$, (i = 1, ..., m), et p contraintes d'égalité $h_j(\theta) = 0$, (j = 1, ..., p). L'idée est de minimiser la combinaison P :

$$P(\theta, r_k) = f(\theta) + r_k \Sigma_i 1/g_i(\theta) + \Sigma_j h_j^2(\theta)/\sqrt{r_k} \quad , r_k > 0$$

suivant un algorithme du type :

(1) partant d'un point initial admissible, déterminer une séquence r_1, ..., r_k, ..., avec $r_k \to 0$ pour $k \to \infty$;

(2) pour chaque r_k, trouver le minimum local de $P(\theta, r_k)$;

(3) stopper la procédure quand le critère de convergence atteint le seuil requis, sinon retourner en (2).

L'intérêt majeur de ce procédé réside en ce qu'à l'étape (2) il suffit d'effectuer une minimisation sans contrainte. Plus généralement, l'utilisation des fonctions de pénalité (dont l'éventail des formulations possibles est beaucoup plus large que ne le laisserait penser l'exemple cité) revient à transformer un problème contraint en une succession de problèmes non contraints ; d'où une simplification technique, dont la contrepartie ne doit cependant pas être cachée : augmentation du temps de calcul, et surtout risque de passage progressif à un critère mal conditionné (*cf.* FLETCHER, 1981, chap. 12). Malgré cela, ces méthodes restent les plus attractives pour les non spécialistes, du fait de la large diffusion de bons algorithmes d'optimisation sans contrainte, d'emploi relativement aisé.

Ces derniers peuvent être schématiquement classés en deux groupes : celui des méthodes directes, qui n'utilisent que les valeurs prises par la fonction-objectif, et celui des méthodes analytiques, qui nécessitent en outre le calcul (ou l'approximation par différences finies) de son gradient, parfois aussi de son hessien (voir par exemple FLETCHER, 1980). Il est classi-

quement reconnu que les secondes sont plus performantes que les premières ; l'avantage n'est toutefois guère patent quand la dimension de l'espace paramétrique est faible (< 10), ce qui en l'occurrence correspond au problème discuté ici. Pour cette raison, et aussi par souci de commodité car le calcul des composantes du gradient de Δ n'est pas trivial, l'emploi d'une méthode directe peut être conseillée ; par exemple, l'algorithme des "directions conjuguées" de POWELL (1964), ou encore l'algorithme du "simplex" de NELDER & MEAD (1965).

Pour clore ce chapitre, il faut revenir sur un point qui y a été soulevé à plusieurs reprises : les trois grandes classes de choix fondamentaux examinés font appel à une connaissance préalable du terrain d'étude : caractéristiques biologiques (faunistiques, floristiques, ...), hydrologiques, sédimentologiques, conditions d'accès au site, ... Divers niveaux de variabilité doivent en particulier être quantifiés pour permettre ensuite de décider de l'effort à consentir en vue d'atteindre une précision souhaitée ; cela rejoint la recommandation formulée par GREEN (1979) dans le cinquième de ses "dix principes". La mise en place d'une stratégie de surveillance est tributaire de l'information acquise au cours d'une "étude pilote", dont le coût sera comptabilisé dans le terme C_o de l'équation {28}.

CONCLUSION

L'écologie appliquée est une science qui ne peut pas être considérée comme ayant aujourd'hui atteint sa maturité. En particulier parce que l'écologie dite fondamentale est un domaine qui lui-même appelle de nouveaux développements, et dans lequel restent à accomplir de sensibles progrès méthodologiques et conceptuels. Cet état de l'"épistémè" contribue à relativiser un certain nombre des arguments développés dans cet article, spécialement en ce qui concerne les choix relatifs à la variable indicative, et donc la qualité du canal informatif devant permettre de diagnostiquer un éventuel dysfonctionnement (ou tout au moins un changement d'état) de l'écosystème. Ne procédant que d'une connaissance partielle des lois qui régissent l'évolution de celui-ci, la tentative de mise en évidence de modifications induites par une perturbation ne peut qu'être teintée d'empirisme. Cela impose de bien cerner la portée et les limites de la "surveillance" du milieu, par exemple lorsque le révélateur étudié est une biomasse ; ce cas a été judicieusement commenté par BERNSTEIN & ZALINSKI (1983) : "We would like to introduce a note of caution. The impact criterion is based solely on changes in abundance, but measures of abundance provide only a limited view of the behavior of ecological systems. This focus ignores community structure, productivity and energy flow, and interspecific interactions such as competition and predation. This approach to defining and measuring impact is therefore useful and appropriate only in ecological systems where levels of abundance furnish meaningful information about the state on the system ... test procedures are powerful tools for detecting and documenting man-induced changes, but only when used with an awareness of their limitation".

S'il fallait établir un ordre de priorité dans les choix discutés au chapitre 4, celui de la variable indicative serait assurément placé au sommet de la hiérarchie. De fait, la nature du signal diagnostique retenu gouverne la sélection entre options débattues aux paragraphes 4.2. et 4.3. Au plan technique de surcroît, les caractéristiques de la distribution spatio-temporelle de la variable indicative orienteront le choix des méthodes statistiques de saisie et de traitement des données.

Plus profondément, la variable indicative devra être définie avec le souci d'enrichir la "fonction sémiotique" des acteurs en charge de la surveillance du milieu. Ainsi qu'il fut rappelé en préambule du chapitre 4, cela impose

notamment que la conception de la stratégie s'appuie sur une compréhension minimale des mécanismes cybernétiques de l'écosystème. Au plan opérationnel, il est donc nécessaire d'envisager corriger continûment, si besoin, la conduite de la surveillance en y intégrant les acquis d'une recherche à caractère plus fondamental . En ce sens doivent être régulièrement soumis à évaluation les programmes de surveillance tendant à devenir figés et pérennes, et dont la justification procède du souhait de disposer de chroniques suffisamment longues et qualitativement homogènes ; eu égard aux coûts de telles actions, elles ne peuvent être engagées qu'après une analyse critique du rapport qualité/prix des résultats espérés. Cela rejoint les recommandations formulées au §4.3.

Conformément à ce qui précède, l'élaboration d'une stratégie de surveillance doit reposer sur un modèle de l'objet de l'étude. Ainsi peut-il être fait appel au raisonnement analogique. Cependant, le recours à une représentation mathématique simplifiée du fonctionnement de l'écosystème sera plus fructueux : dans une première étape, il sera vérifié que le modèle reproduit correctement les observations acquises durant l'enquête-cadre. Si tel est le cas, l'idée que les phénomènes majeurs ont bien été cernés s'en trouve confortée. Dans un second temps pourra alors être évaluée la sensibilité de chaque variable d'état à l'effet supposé de l'aménagement : de la sorte seront obtenus par simulation des critères quantitatifs pour le choix de la variable indicative (incidemment, les échantillons collectés au cours de la surveillance serviront en retour à la validation du modèle). Autrement dit, et en reprenant la terminologie utilisée dans ce rapport, il est attendu de cette approche un choix raisonné de X d'une part, et d'autre part une idée *a priori* judicieuse de la valeur de l'écart Δ devant être testée. Par ailleurs, le champ d'application de la démarche ne doit pas être compris comme limité à l'exemple choisi : elle vaudrait tout autant pour un modèle des régulations physiologiques d'une catégorie taxinomique (*e.g.*, Mollusques bivalves, Crustacés, ...) à laquelle il serait souhaité accorder une attention particulière. Un dernier aspect du problème doit enfin être mentionné : pour l'essentiel, l'argumentation a porté sur la surveillance des variations du niveau moyen de descripteurs du milieu qui sont des entités biologiques (*cf.* §4.1.) ; pour autant, le contrôle ne se limite nullement à ces deux facettes (niveaux moyens, descripteurs biologiques). De nombreux exemples peuvent être cités pour lesquels la cible est de nature chimique (*e.g.*, éléments nutritifs, polluants d'origine industrielle, ...), et où la question posée est non pas celle du niveau moyen, mais celle de

"seuils de tolérance" (à ne pas franchir, voire à définir). L'une des motivations de ce type d'étude réside dans l'éventuelle forte non linéarité de certaines "réponses écologiques" engendrées par des modifications quantitativement mineures des facteurs abiotiques. Selon cette optique, et en complément de ce qui précède, pourraient être conçues des stratégies de surveillance ayant pour objectif de traquer les "valeurs extrêmes" (de la réponse). Cette philosophie diffère très sensiblement de celle qui a été exposée dans ce rapport : un retour aux paragraphes 2.1. et 3.5. montre que les valeurs extrêmes y sont plutôt considérées comme une gêne, leur effet étant de déstabiliser l'estimation des valeurs centrales.

La richesse de la problématique esquissée au travers de ces quelques réflexions désigne à l'évidence la nécessaire optimisation des activités de surveillance comme un catalyseur privilégié d'investigations méthodologiques.

Remerciements

Il nous est hautement agréable d'adresser ici nos plus vifs remerciements aux personnes auxquelles nous sommes redevables d'un précieux concours intellectuel. Ainsi avons nous eu le plaisir de débattre nombre de thèmes abordés dans ce rapport avec Alain LAUREC (IFREMER/DRV). La version finale du texte a en outre été enrichie des pertinentes suggestions formulées par Pierre CHARDY (IFREMER/DERO) et Jacques LABEYRIE (IFREMER/DIT).

Il convient par ailleurs de souligner que l'ingrate tâche dactylographique a été diligemment assumée par Madame Yvette CASSOU.

BIBLIOGRAPHIE

ALLEN, J., D. GONZALEZ, & D.V. GOKHALE, 1972
Sequential sampling plans for the bollworm, *Heliotis Zea*.
Environ. Entomol. 1(6) : 771-780.

ANDREWS, D.L., P.J. BICKEL, F.R. HAMPEL, P.J. HUBER, W.H. ROGERS & J.W. TUKEY, 1972
Robust estimates of location. Survey and advances.
Princeton University Press, Princeton, N.J., 373 p.

BAMBER, R.N., & J.F. SPENCER, 1984
The benthos of a coastal power station thermal discharge canal.
J. mar. Biol. Ass. U.K. 64 : 603-623

BARD, Y., 1974
Nonlinear parameter estimation.
Academic Press, New-York, San-Francisco, London, 341 p.

BERNSTEIN, B.B., & J. ZALINSKI, 1983
An optimum sampling design and power tests for environmental biologists.
J. Environ. Management 16 : 35-43

BECKMAN, R.J., & R.D. COOK, 1983
Outlier ...s.
Technometrics 25(2) : 119-149

BEST, D.J., & J.C.W. RAYNER, 1987
Welch's approximate solution for the Behrens-Fisher problem.
Technometrics 29(2) : 205-210

BOULOT, F. & A. HAUGUEL, 1981
Modélisation de la dilution des rejets thermiques en mer.
E.D.F., Bull. D.E.R. (A) 2 : 9-30

BOX, M.J., 1966
A comparison of several current optimization methods, and the use of transformations in constrained problems.
Comput. J. 9 : 67-77

BRADLEY, J.V., 1968
Distribution - free statistical tests
Prentice-Hall, Inc., Englewood Cliffs, N.J., 388 p.

CLARK, I., 1979
Practical Geostatistics.
Applied Science Publishers Ltd., Ripple Rd., Barking, Essex, England, 129 p.

CONOVER, W.J., 1980
Practical nonparametric Statistics
John Wiley & Sons, New York, Chichester, Brisbane, Toronto, 2nd. ed : 510 p.

DELHOMME, J.-P., 1978
Applications de la théorie des variables régionalisées dans les sciences de l'eau.
Bull. B.R.G.M. (2)III, 4 : 341-375

DEMERS, S., & L. LEGENDRE, 1982
Water column stability and photosynthetic capacity of estuarine phytoplankton : long term relationships.
Mar. Ecol. Prog. Ser. 7 : 337-340

DEMERS, S., J.-C. THERRIAULT, L. LEGENDRE & J. NEVEUX, 1985
An *in vivo* fluorescence method for the continuous *in situ* estimation of phytoplankton photosynthetic characteristics.
Mar. Ecol. Prog. Ser. 27 : 21-27

DOWNING, J.A., M. PERUSSE, & Y. FRENETTE, 1987
Effect of interreplicate variance on zooplankton sampling design and data analysis.
Limnol. Oceanogr. 32(3) : 673-680

EFRON, B., 1979a
The 1977 Rietz lecture. Bootstrap methods : another look at the jackknife.
Ann. Statist. 7(1) : 1-26

EFRON, B., 1979b
Computers and the theory of statistics : thinking the unthinkable.
SIAM Review 21(4) : 460-480

EFRON, B., 1981a
Nonparametric standard errors and confidence intervals.
Can. J. Statist. 9(2) : 139-172

EFRON, B., 1981b
Nonparametric estimates of standard error : the jackknife, the bootstrap and other methods.
Biometrika 68(3) : 589-599

EFRON, B., 1982a
The jackknife, the bootstrap, and other resampling plans. CMBS-NSF, Regional conference series in applied mathematics no. 38,
SIAM ed., Philadelphia, Pennsylvania, 92 p.

EFRON, B., 1982b
The 1981 Wald memorial lectures. Transformation theory : how normal is a family of distributions ?
Ann. Statist. 10(2) : 323-339

EFRON, B., & G. GONG, 1983
A leisurely look at the bootstrap, the jackknife, and cross-validation.
Amer. Stat. 37(1) : 36-48

EVANS, M.S., & D.W. SELL, 1983
Zooplankton sampling strategies for environmental studies.
Hydrobiologia 99 : 215-223

FIACCO, A.V., & G.P. Mc CORMICK, 1966
Extensions of SUMT for nonlinear programming : equality constraints and extrapolation.
Management Science 12(11) : 816-828

FLETCHER, R., 1980
Practical Methods of Optimization. Volume 1 : Unconstrained Optimization
John Wiley & Sons, Chichester, New York, Brisbane, Toronto, 120 p.

FLETCHER, R., 1981
Practical Methods of Optimization. Volume 2 : Constrained Optimization
John Wiley & Sons, Chichester, New York, Brisbane, Toronto, 224 p.

FOWLER, G.W., 1983
Accuracy of sequential sampling plans based on Wald's sequential probability ratio test.
Can. J. For. Res. 13 : 1197-1203

FRONTIER, S., 1973
Etude statistique de la dispersion du zooplancton.
J. exp. mar. Biol. Ecol. 12 : 229-262

GENTIEN, P., Ph. GROS & G. LE FEVRE-LEHOERFF, 1986
Evaluation des conséquences du fonctionnement de la centrale de Gravelines sur le milieu marin
Rapp. IFREMER/DERO/EL.07/86, 92 p + Annexes.

GOVINDARAJULU, Z., 1985
Recent developments in sequential analysis : testing hypotheses
Math. Scientist 10(1) : 51-64

GREEN, R.H., 1979
Sampling design and statistical methods for environmental biologists
John Wiley & Sons, New York, Chichester, Brisbane, Toronto, 257 p.

GROS, Ph., 1981
Gravelines : première étude de surveillance du site. Description statistique des données et interprétation écologique
Rapp. E.D.F. - CNEXO/COB/ELGMM, 121 p.

GROS, Ph. & M. RYCKAERT, 1983
Etude de la production primaire phytoplanctonique dans les eaux littorales de la côte normande (Manche orientale).
Oceanol. Acta 6(4) : 435-450

HAJEK, J., 1969
A course in nonparametric Statistics
Holden-Day, San Francisco, Cambridge, London, Amsterdam, 184 p.

HAMPEL, F.R., 1973
Robust estimation : a condensed partial survey.
Z. Wahrscheinlichkeitstheorie verw. Geb. 27 : 87-104

HAMPEL, F.R., 1974
The influence curve and its role in robust estimation.
J. Amer. Statist. Assoc. 69 (346) : 383-393

HEILBRUN, L.K., & D.L. Mc GEE, 1985
Sample size determination for the comparison of normal means when one sample size is fixed.
Comput. Stat. & Data Analysis 3 : 99-102

HOLLANDER, M., & D.A. WOLFE, 1973
Nonparametric Statistical methods
John Wiley & Sons, New York, London, Sydney, Toronto, 491 p.

HUBER, P.J., 1972
The 1972 Wald lecture. Robust statistics : a review.
Ann. Math. Statist. 43(4) : 1041-1067

HUBER, P.J., 1981
Robust statistics
John Wiley & Sons, New York, Chichester, Brisbane, Toronto, 308 p.

JACKSON, J.K., & V.H. RESH, 1988
Sequential decision plans in monitoring benthic macroinvertebrates : cost savings, classification accuracy, and development of plans.
Can. J. Fish. Aquat. Sci. 45(2) : 280-286.

JOHNSON, N.J., 1978
Modified t tests and confidence intervals for asymmetrical populations.
J. Amer. Statist. Assoc. 73(363) : 536-544

KENDALL, M., & A. STUART, 1979
The Advanced Theory of Statistics. Vol. II : Inference and Relationship
Charles Griffin & Co, Ltd., London & High Wycombe, 4th edition, 748 p.

LAUBSCHER, N.F., 1960
Normalizing the noncentral t and F distributions.
Ann. Math. Stat. 31 : 1105-1112

LAUREC, A., P. CHARDY, G. LEFEVRE & F. TOULARASTEL, 1981
Définition d'un état de référence écologique : problèmes d'inférences statistiques. *In* : 2es journées de la thermo-écologie. *Influence des rejets thermiques sur le milieu vivant en mer et en estuaire*, ISTPM, Nantes, 14-15 novembre 1979.
E.D.F., Direction de l'Equipement éd., pp. 157-194

LECOUTRE, J.P., P. TASSI, & A. TROGNON, 1986
La robustesse statistique. I : généralités, outils, définitions.
Doc. Trav. INSEE 8602, 31 p.

LEHMANN, E.L., 1953
The power of rank tests
Ann. Math. Statist. 24 : 23-42

LEHMANN, E.L., 1959
Testing Statistical Hypotheses
John Wiley & Sons, Inc., New York, London, Sydney, 369 p.

LEHMANN, E.L., 1975
Nonparametrics. Statistical Methods based on Ranks
Mc Graw-Hill International Book Company, Holden-Day, Inc., San Francisco, 445 p.

LEVASSEUR, M., J.-C. THERRIAULT, & L. LEGENDRE, 1983
Tidal currents, winds and the morphology of phytoplankton spatial structures.
J. Mar. Res. 41 : 655-672

MACKAS, D.L., 1984
Spatial autocorrelation of plankton community composition in a continental shelf ecosystem.
Limnol. Oceanogr. 29(3) : 451-471

MILLER, R.C., 1974
The jackknife : a review.
Biometrika 61(1) : 1-15

MILLARD, S.P., & D.P. LETTENMAIER, 1986
Optimal design of biological sampling programs using the analysis of variance.
Estuar., Coast. and Shelf Sci. 22 : 637-656

MILLARD, S.P., J.R. YEARSLEY, & D.P. LETTENMAIER, 1985
Space-time correlation and its effects on methods for detecting aquatic ecological changes.
Can. J. Fish. Aquat. Sci. 42(8) : 1391-1400

MORAN, P.A.P., 1980
Calculation of the normal distribution function.
Biometrika 67(3) : 675-676

NELDER, J.A., & R. MEAD, 1965
A simplex method for function minimization.
Comput. J. 7 : 308-313

NIHOUL, J.C.J., & J.H. HECQ, 1984
Influence of the residual circulation on the physico-chemical characteristics of water masses and the dynamics of ecosystems in the belgian coastal zone.
Cont. Shelf Res. 3(2) : 167-174

NOETHER, G.E., 1967
Elements of nonparametric statistics
John Wiley & Sons, Inc., New York, London, Sydney, 104 p.

PARENT, J.-F., 1981
Problèmes posés par l'évaluation prévisionnelle de l'impact d'une centrale électrique lors de l'établissement du dossier d'impact.
In : 2es journées de la thermo-écologie : *Influence des rejets thermiques sur le milieu vivant en mer et en estuaire*, ISTPM, Nantes, 14-15 novembre 1979
E.D.F., Direction de l'Equipement éd., : 761-776

PATTEN, B.C., 1984
System theory formulation of site-specific water quality standards and protocols.
Ecol. Modelling 23 : 313-340

POWELL, M.J.D., 1964
An efficient method for finding the minimum of a function of several variables without calculating derivatives.
Comput. J. 7 : 155-162

RAGO, P.J., 1984
Production forgone : an alternative method for assessing the consequences of fish entrainment and impingement losses at power plants and other water intakes.
Ecol. Modelling 24 : 79-111

RESH, V.H., & D.G. PRICE, 1984
Sequential sampling : a cost-effective approach for monitoring benthic macro-invertebrates in environmental impact assessments.
Environmental Management 8(1) : 75-80

ROCKE, D.M., & G.W. DOWNS, 1981
Estimating the variances of robust estimators of location : influence curve, jackknife and bootstrap.
Commun. Statist. simula. computa. B10(3) : 221-248

ROSENBERG, D.M., & V.H. RESH *et al.*, 1981
Recent trends in environmental impact assessment.
Can. J. Fish. Aquat. Sci. 38(5) : 591-624

RYCKAERT, M., Ph. GROS, & E. ERARD-LE DENN, 1983
Succession saisonnière des populations phytoplanctoniques des eaux côtières de la Manche.
Oceanol. Acta. Proceedings 17th European Marine Biology Symposium, Brest, France, 27 Sept.-1 Oct. 1982 : 171-175

SAVAGE, R.I., 1956
Contributions to the theory of rank order statistics - the two-sample case.
Ann. Math. Statist. 27 : 590-615

SHANNO, D.F., 1983
Large scale unconstrained optimization.
Comput. Chem. Engng. 7(5) : 569-574

SINDERMANN, C.J., 1984
Fish and environmental impacts.
Arch. Fisch Wiss. 35(1) : 125-160

THERRIAULT, J.-C., & T. PLATT, 1981
Environmental control of phytoplankton patchiness.
Can. J. Fish. Aquat. Sci. 38(6) : 638-641

VAN DER WAERDEN, B.L., 1967
Statistique Mathématique
Dunod éd., Paris, 371 p.

WALTERS, C.J., J.S. COLLIE, & T. WEBB, 1988
Experimental designs for estimating transient responses to management disturbances.
Can. J. Fish. Aquat. Sci. 45(1) : 530-538

WELCH, B.A., 1947
The generalization of "Student's" problem when several different population variances are involved.
Biometrika 34 : 28-35

A N N E X E I

DISTRIBUTIONS USUELLES ET CONVERGENCES : RAPPELS.

DISTRIBUTIONS USUELLES ET CONVERGENCES : RAPPELS

Le propos de la présente annexe a été délibérément limité à quelques rappels : d'une part la définition des quelques lois de probabilités fréquemment mentionnées dans le rapport, et d'autre part l'énoncé des théorèmes limites qui fondent les approximations appliquées au cas des grands échantillons. Les concepts de base de la théorie des probabilités sont supposés acquis, de même que les connaissances relatives aux variables aléatoires.

1. QUELQUES DISTRIBUTIONS ABSOLUMENT CONTINUES

1.1. GENERALITES

L'ensemble des valeurs possibles d'une variable aléatoire est soit fini, soit dénombrable, soit continu. Conformément au titre du paragraphe, nulle mention ne sera faite ici des lois discrètes classiques (loi binômiale, loi de Poisson, loi multinôminale...).

Soit donc X une variable aléatoire dont la fonction de répartition est notée F, *i.e.*,

$$\text{Proba}(X \leq x) = F(x)$$

La fonction de répartition F de l'aléa X est dite absolument continue s'il existe une fonction f telle que :

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Cette fonction f est appelée fonction de densité de la variable aléatoire X. Il est rappelé que :

$$\begin{cases} F(-\infty) = 0 \\ F(+\infty) = 1 \\ F(x_1) \leq F(x_2) \quad \text{si} \quad x_1 \leq x_2 \end{cases} \qquad \begin{cases} f(x) \geq 0 \qquad -\infty \leq x \leq +\infty \\ \text{et } \int_{-\infty}^{+\infty} f(t)dt = F(+\infty) - F(-\infty) = 1 \end{cases}$$

Ce formalisme permet d'introduire de façon concise deux notions fondamentales : celles de distribution jointe, et celle d'indépendance.

(i) Soient deux variables aléatoires continues X et Y. Soit F la fonction de répartition jointe du couple aléatoire (X,Y), et soit f la densité jointe correspondante. Alors :

$$\text{Proba}(X \leq x,\ Y \leq y) = F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(s,t)\, ds\, dt$$

La relation entre F et f s'exprime encore : $f(x,y) = \partial^2 F(x,y)/(\partial x \partial y)$.

Comme précédemment :

$$\begin{cases} F(-\infty,y) = F(x,-\infty) = F(-\infty,-\infty) = 0 \\ F(+\infty,+\infty) = 1 \end{cases} \qquad \begin{cases} f(x,y) \geq 0 \\ \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} f(x,y)dy = 1 \end{cases}$$

(ii) En considérant toujours le couple aléatoire (X,Y) :

$F(x,+\infty) = F_1(x)$ est la fonction de répartition marginale de X,

$F(+\infty,y) = F_2(y)$ est la fonction de répartition marginale de Y.

L'interprétation en termes de probabilités est immédiate ; par exemple :

$$F_1(x) = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f(s,t)\, ds\, dt = \text{Proba}(X \leq x)$$

Par ailleurs, l'indépendance statistique est définie comme suit : si et seulement si

$$F(x,y) = F(x,\infty)F(\infty,y) = F_1(x)F_2(y)$$

pour toute valeur de x et y, alors les deux variables aléatoires X et Y sont indépendantes.

Remarque : tout ce qui vient d'être énoncé pour un couple aléatoire se généralise aisément à un n-uplet de variables aléatoires.

1.2. LA LOI NORMALE (Gauss, 1809 ; Laplace, 1812)

Il s'agit d'une distribution absolument continue, encore appelée loi de Laplace-Gauss, et définie par la fonction de densité dépendant de deux paramètres μ et σ (avec $\sigma>0$) :

$$f(x) = 1/(\sigma\sqrt{2\pi}) \exp\{ ((x-\mu)/\sigma)^2/2 \} \qquad -\infty < x < \infty, \quad -\infty < \mu < \infty$$

Propriétés :

(i) $X \sim N(\mu,\sigma^2)$; alors : $E(X) = \mu$, $V(X) = \sigma^2$

(ii) Stabilité de la loi normale : soient n variables aléatoires X_i indépendantes et normalement distribuées, *i.e.*,

$$X_i \sim N(\mu_i, \sigma_i^2) \qquad i = 1,\ldots,n$$

La somme de ces n aléas indépendants suit elle-même une loi normale :

$$\sum_i X_i \sim N(\sum_i \mu_i, \sum_i \sigma_i^2)$$

Tabulation de la fonction de répartition :

L'usuelle "table de la loi normale" est en réalité celle de la variable gaussienne N(0,1), dont la fonction de répartition est classiquement notée Φ. Cette table est bien entendu utilisable dans le cas général, *i.e.* lorsque la moyenne est non nulle et la variance différente de l'unité :

$$X \sim N(\mu,\sigma^2) \quad \text{Proba}(X\leq x) = \Phi((x-\mu)/\sigma) = (1/\sqrt{2\pi}) \int_{-\infty}^{(x-\mu)/\sigma} \exp(-t^2/2)dt$$

La fonction Φ est souvent appelée fonction de répartition de la loi normale centrée et réduite.

Loi multinormale :

Soit X le (n,1) vecteur dont les composantes sont les n variables aléatoires X_i ; le (n,1) vecteur des espérences $\mu_i = E(X_i)$ est noté μ, et la

(n,n) matrice symétrique non singulière des covariances est notée W :

$$W = E\{ (X-\mu)(X-\mu)' \}$$

où le symbole ' désigne la transposition. Alors le vecteur aléatoire X suit dans $\mathbb{R}^n$ une loi normale à n dimensions si, par définition, sa fonction de densité s'écrit :

$$f(X) = ((2\pi)^n \det W)^{-1/2} \exp\{-(X-\mu)'W^{-1}(X-\mu)/2\}$$

Si les composantes X_i sont toutes mutuellement indépendantes, cela entraîne que la matrice W est diagonale. Une caractéristique du cadre gaussien est que la réciproque est vraie : si les covariances (ou les corrélations) entre variables aléatoires normales sont nulles, alors ces variables sont indépendantes.

1.3. DISTRIBUTIONS DERIVEES DE LA LOI NORMALE

1.3.1. Loi du χ^2 à n degrés de liberté

Soient $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(0,1)$

Par définition, la variable aléatoire : $\chi^2_n = \sum_{i=1}^{n} X_i^2$

suit une loi appelée distribution du CHI-2 à n degrés de liberté. Sa densité est de la forme :

$$f(x) = C\, x^{n/2-1} \exp(-x/2) \qquad 0 < x < \infty$$

où C est une constante telle que l'aire sous la courbe entre 0 et $+\infty$ soit égale à 1.

Propriétés :

(i) $E(\chi^2_n) = n \qquad V(\chi^2_n) = 2n$

(ii) Additivité : soient k aléas indépendants $\chi^2_{n_1},\ldots,\chi^2_{n_k}$ possédant respectivement $n_1,\ldots,n_k$ d.d.l. ; alors :

$$\sum_{i=1}^{k} \chi^2_{n_i} \sim \chi^2_{n_1+\ldots+n_k}$$

Remarques :

(1) $X_1,\ldots,X_n \overset{iid}{\sim} N(\mu,\sigma^2) \Rightarrow (1/\sigma^2) \sum_i (X_i-\bar{X})^2 \sim \chi^2_{n-1}$

où $\bar{X}$ désigne la moyenne arithmétique.

(ii) Soit X un vecteur aléatoire de $\mathbb{R}^n$, multinormal, caractérisé par $E(X) = \mu$ et par la (n,n) matrice de covariances W, supposée non singulière. Alors la forme quadratique :

$$\chi^2_n = (X-\mu)'W^{-1}(X-\mu)$$

possède une distribution du CHI-2 à n d.d.l.

CHI-2 non centré :

Soient $X_1,\ldots,X_n$ indépendantes, avec $X_i \sim N(\mu_i,1)$. La somme des carrés de ces n aléas suit une loi de χ^2 à n d.d.l., non centrée, et de paramètre de décentrement δ :

$$\delta = \left(\sum_i \mu_i^2\right)^{1/2}$$

La variable est notée $\chi^2_{n,\delta}$, avec $E(\chi^2_{n,\delta}) = n+\delta^2$, et $V(\chi^2_{n,\delta}) = 2n+4\delta^2$.

Il est souvent utile d'approcher la loi non centrée $\chi^2_{n,\delta}$ par la loi de CHI-2 centrée. Une démarche possible consiste à retenir pour approximation $c\chi^2_m$, les constantes c et m étant déterminées de telle sorte que soient égaux les deux premiers moments centrés de $\chi^2_{n,\delta}$ et de $c\chi^2_m$, *i.e.* :

$$cm = n + \delta^2 \quad , \text{ et } c^2m = n + 2\delta^2$$

En général, le nombre m de d.d.l. n'est pas entier.

1.3.2. Loi de "Student" à n degrés de liberté

Soient X et Y deux variables aléatoires indépendantes, et telles que :

$$X \sim N(0,1) \qquad Y \sim \chi^2_n$$

Par définition, la distribution du rapport t_n :

$$t_n = X/\sqrt{Y/n} \qquad X,Y \text{ indépendantes}$$

est une loi de Student à n d.d.l. (Student est le pseudonyme du statisticien britannique W.S. Gosset, 1876-1937). Sa densité est de la forme :

$$f(x) = C(1+x^2/n)^{-(n+1)/2} \qquad -\infty < x < +\infty$$

où C est une constante telle que l'aire sous la courbe soit égale à 1. Il s'en déduit :

$$E(t_n) = 0 \quad , \quad n > 1 \qquad \text{et} \quad V(t_n) = n/(n-2) \quad , \quad n > 2$$

Cas particulier : la distribution de t_1 (*i.e.*, de la variable de Student à 1 d.d.l.) est une loi de Cauchy, de densité $1/(\pi(1+x^2))$. En tendant vers l'axe des x, les extrémités de cette fonction de densité se comportent comme $1/x^2$. L'épaisseur des queues de distribution fait de cette loi un outil pour l'étude théorique de l'influence des événements extrêmes sur les procédures statistiques.

Loi de Student non centrée : soient deux aléas indépendants X et Y,

$$X \sim N(\delta,1) \quad , \quad Y \sim \chi^2_n$$

Le rapport $X/\sqrt{Y/n}$ suit une loi de t non centrée, de paramètre de décentrement δ. Etant donné que les tables de la loi de Student non centrée sont assez peu diffusées, il peut être commode d'avoir recours à l'approxi-

mation suivante de la fonction de répartition :

$$\text{Proba}(t_{n,\delta} \leq x) \simeq \Phi \{ (x-\delta)/\sqrt{1+x^2/(2n)} \}$$

où Φ désigne la fonction de répartition de l'aléa N(0,1).

1.3.3. Loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté

Soient X_1 et X_2 deux aléas indépendants, $X_1 \sim \chi^2_{n_1}$, $X_2 \sim \chi^2_{n_2}$.

La distribution du rapport :

$$F_{n_1,n_2} = (X_1/n_1)/(X_2/n_2)$$

suit une loi de Fisher-Snedecor à n_1 et n_2 d.d.l.. La fonction de densité est de la forme :

$$f(x) = C\, x^{n_1/2-1}\, (n_2+n_1x)^{-(n_1+n_2)/2} \quad , x > 0$$

C étant une constante telle que l'aire sous la courbe entre 0 et $+\infty$ soit égale à 1.

Propriétés :

(i) $E(F_{n_1,n_2}) = n_2/(n_2-2) \qquad n_2 > 2$

$$V(F_{n_1,n_2}) = (n_2^2(2n_2+2n_1-4))/(n_1(n_2-2)^2(n_2-4)) \qquad n_2 > 4$$

(ii) Soit $F_{1-\alpha;n_1,n_2}$ le réel défini par la relation :

$$\text{Proba}\{ (X_1/n_1)/(X_2/n_2) \geq F_{1-\alpha;n_1,n_2} \} = \alpha$$

alors : $F_{1-\alpha;n_1,n_2} = 1/F_{\alpha;n_2,n_1}$

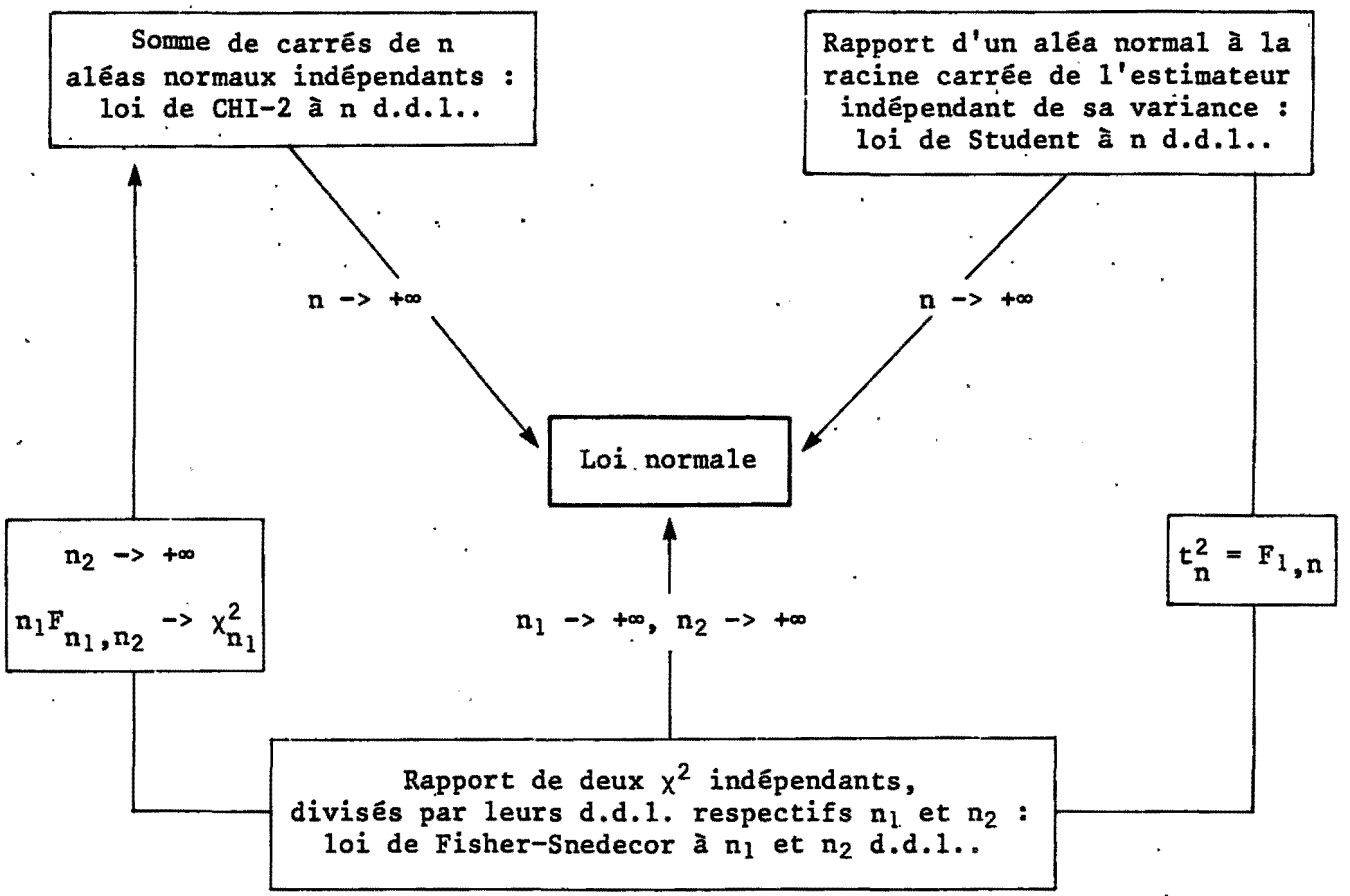
(iii) Avec les mêmes notations :

$$F_{\alpha;1,n} = t^2_{\alpha/2;n} \quad , \text{ et } nF_{\alpha;n,\infty} = \chi^2_{\alpha;n}$$

Remarque : la variable F est dite non centrée lorsque la variable X_1 suit elle-même une loi de χ^2 non centrée.

1.3.4. Tendance vers la loi normale

Avant de rappeler l'énoncé de théorèmes limites, qui établissent la convergence vers la loi de Laplace-Gauss dans un contexte plus large, la tendance vers la normalité des trois distributions (CHI-2, Student, Fisher) est résumée par le schéma qui suit. C'est là une première illustration du rôle central de la loi normale en statistique, spécialement en tant que loi limite.



2. THEOREMES LIMITES

La présentation de la loi des grands nombres et du théorème de la limite centrale établira le lien entre le modèle probabiliste et son utilisation dans le raisonnement statistique : des résultats classiques seront rappelés, qui montrent comment les fréquences relatives tendent vers les probabilités, les moyennes vers les espérances, et les sommes de variables aléatoires vers des variables normales.

2.1. INEGALITE DE BIENAYME (1853) - TCHEBYCHEFF (1867)

Soit une variable aléatoire X dont la fonction de répartition est inconnue. Seules sont connues son espérance E(X) et sa variance $V(X) = \sigma^2$. Soit α un réel positif arbitraire ; alors :

$$\text{Proba}(|x - E(X)| \geq \alpha\sigma) \leq 1/\alpha^2 \qquad \alpha > 0$$

L'interprétation concrète est évidente : pour toute variable aléatoire, la probabilité que ses réalisations s'éloignent de l'espérance d'une distance supérieure à 3 écart-types (par exemple), est inférieure à 1/9.

Ce résultat est très général, puisqu'il vaut aussi bien pour les variables aléatoires discrètes qu'absolument continues, et qu'il ne fait aucune hypothèse sur la nature ou la forme de la loi de X. Dans de nombreux cas cependant, la probabilité est très inférieure à $1/\alpha^2$: ainsi n'excède-t-elle pas .0027 pour $\alpha = 3$ si X est un aléa normal, valeur bien plus faible que 1/9. C'est d'ailleurs ce comportement du modèle gaussien qui a conduit à considérer qu'il doit être utilisé avec prudence pour décrire les phénomènes expérimentaux : si la partie centrale décrit en général correctement la distribution empirique des observations, en revanche ce modèle n'accorde bien souvent qu'une irréaliste, parce que trop insignifiante, probabilité d'occurrence aux valeurs "extrêmes".

Remarques :

(i) L'inégalité de Bienaymé-Tchebycheff est usuellement énoncée en langage probabiliste. Elle peut aussi s'exprimer à l'aide de la fonction de répartition F de la variable X :

$$F(x) - F(-x) \geq 1-\sigma^2/x^2$$

(ii) Sans entrer dans le détail, il peut être mentionné que l'inégalité peut être considérablement "resserrée" au prix de quelques restrictions assez faibles sur la loi de X ; par exemple, pour une distribution continue ne présentant qu'un unique mode m :

$$\text{Proba}(|X - m| > \alpha\tau) < 4/(9\alpha^2)$$

où τ désigne la racine carrée du moment d'ordre 2 centré sur le mode m, *i.e.*

$$\tau^2 = \sigma^2 + (m - E(X))^2$$

2.2. CONVERGENCE EN PROBABILITE

2.2.1. Cas des épreuves de Bernoulli

Une épreuve de Bernoulli n'a que deux résultats possibles : soit le "succès" avec la probabilité p, soit "l'échec" avec la probabilité 1 - p. Il lui est associée une variable aléatoire I, dite variable de Bernoulli, qui prend la valeur I = 1 en cas de succès, ou bien I = 0 dans le cas contraire. Soit K le nombre de succès dans une suite de n répétitions indépendantes de l'épreuve :

$$K = \sum_{i=1}^{n} I_i$$

A l'aide de l'inégalité de Bienaymé-Tchebycheff, il est montré que la différence entre la fréquence relative K/n d'un événement de probabilité p, et cette même probabilité p, peut être rendue très faible avec une probabilité très forte, pourvu que n soit assez grand. Formellement :

$$\forall \varepsilon > 0 \quad , \quad \lim_{n\to\infty} \text{Proba} (|K/n - p| < \varepsilon) = 1$$

Résultat qui s'exprime : la fréquence relative K/n converge en probabilité vers p lorsque n->∞ ; ou encore : K/n est une estimation convergente de p. Ce résultat fonde l'estimation statistique des probabilités ; c'est un cas particulier de la "loi des grands nombres", qui établit sa généralisation

à des variables aléatoires identiques et indépendantes quelconques.

2.2.2. Loi faible des grands nombres (Khintchine, 1929)

Soient $X_1, X_2, \ldots, X_n, \ldots$ des variables aléatoires indépendantes, et possédant la même distribution dont la moyenne $E(X_i) = \mu$ est finie, et soit :

$$\overline{X}_n = (1/n) \sum_{i=1}^{n} X_i$$

Quelle que soit la valeur de n, $\overline{X}_n$ a aussi pour espérance μ. La loi faible des grands nombres énonce que la dispersion de $\overline{X}_n$ autour de μ décroît au fur et à mesure que n augmente ; formellement :

$$\forall\, \varepsilon > 0, \quad \lim_{n \to \infty} \text{Proba} \left(|\overline{X}_n - \mu| < \varepsilon \right) = 1$$

Cette loi atteste que la variable "moyenne empirique" converge en probabilité vers la variable dégénérée (*i.e.*, constante) μ.

Si la distribution commune des X_i possède une variance $V(X_i) = \sigma^2$, la démonstration du résultat est aisément obtenue à l'aide de l'inégalité de Bienaymé-Tchebycheff. Il est aussi prouvé que le résultat demeure valide sans que soit imposée l'existence de σ^2.

Remarque : il existe une "loi forte des grands nombres", relative au comportement de la suite des $\overline{X}_n$ lorsqu'elle tend vers sa limite. Cette loi forte n'est que très rarement utilisée en statistique mathématique.

2.3. CONVERGENCE EN LOI

2.3.1. Définition préliminaire

Une variable aléatoire X, dont la fonction de répartition dépend d'un paramètre n, est dite asymptotiquement normale s'il existe deux nombres a et b, dépendant aussi de n, et tels que la fonction de répartition de (X-a)/b tende vers Φ, fonction de répartition de l'aléa N(0,1), lorsque $n \to \infty$.

Dans de nombreux cas, a et b ne sont autres que la moyenne et l'écart-type de X, respectivement ; cette configuration usuelle ne constitue cependant pas une nécessité.

2.3.2. Cas des variables binômiales

L'aléa K_n précédemment défini comme la somme de n variables de Bernoulli indépendantes et de même probabilité de succès p :

$$K_n = \sum_{i=1}^{n} I_i$$

suit, par définition, une loi binômiale de paramètres n et p, avec $E(K_n) = np$ et $V(K_n) = np(1-p)$. Soit S_n :

$$S_n = (K_n - np)/\sqrt{np(1-p)}$$

Le théorème de De Moivre-Laplace établit que la variable aléatoire S_n est asymptotiquement normale. Cela s'énonce encore : la suite S_n converge en loi vers la variable N(0,1). Formellement, si F_s désigne la fonction de répartition de S_n :

$$\lim_{n\to\infty} \{F_s(x)\} = \Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} \exp(-t^2/2)dt$$

Plus généralement, pour une suite $X_1, X_2, \ldots, X_n, \ldots$ d'aléas définis sur le même ensemble d'événements élémentaires, la convergence en loi de X_n vers X exprime que la fonction de répartition de X_n tend vers la fonction de répartition F(x) de X, en tout point x où F(x) est continue.

L'intérêt propre du théorème de De Moivre-Laplace est de fournir une approximation simple pour le calcul des probabilités binômiales (approximation d'autant meilleure que n est grand et que p est proche de 1/2). Par ailleurs, dans le présent contexte, il apparaît comme un cas particulier du théorème de la limite centrale, applicable à une beaucoup plus grande variété de situations.

2.3.3. Théorème de la limite centrale (Lindeberg-Lévy, 1922)

Le théorème de la limite centrale énonce que, sous certaines conditions, la distribution de la somme de n variables aléatoires indépendantes :

$$X_1 + X_2 + \ldots + X_n$$

est asymptotiquement normale. Les conditions imposées visent à garantir (i) qu'un unique terme X_i de la somme "ne l'emporte pas" sur les autres, et (ii) que les fonctions de répartition des X_i tendent "suffisamment vite" vers 0 et 1 au voisinage de $\pm\infty$.

Les démonstrations ont été faites sous des hypothèses plus ou moins faibles. Un seul résultat sera présenté ici : le cas où les X_i ont tous la même fonction de répartition, avec une moyenne et une variance finies.

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$$

μ et σ^2 existent, $\mu = E(X_i)$, $\sigma^2 = V(X_i)$, $i = 1,\ldots,n$

Soit : $Y_n = \left(\sum_{i=1}^{n} X_i - n\mu\right)/(\sigma\sqrt{n})$

et soit G_n la fonction de répartition de Y_n. Alors :

$$\lim_{n\to\infty} \{G_n(x)\} = \Phi(x) = (1/\sqrt{2\pi}) \int_{-\infty}^{x} \exp(-t^2/2)dt$$

Autrement dit, la somme (et donc aussi, bien entendu, la moyenne) de variables aléatoires iid converge en loi vers la variable normale. L'importance considérable de ce théorème limite tient à ce que la tendance vers la distribution normale apparaît quelle que soit, dans une large mesure, la loi commune des X_i. D'un point de vue concret, ce phénomène a souvent été schématiquement traduit de la façon suivante : dès qu'une variable aléatoire résulte de l'addition de multiples causes elles-mêmes aléatoires, indépendantes, et du même ordre de grandeur, il faut alors s'attendre à ce que sa distribution soit proche de la normalité.

Exemple : cas de la loi de χ^2.

Soient $X_1,\ldots,X_n \overset{iid}{\sim} N(0,1)$; par définition : $\chi^2_n = \sum_{i=1}^{n} X_i^2$

Sachant que $E(X_i^2) = 1$, et que $V(X_i^2) = 2$, alors, d'après le théorème de la limite centrale :

$$\chi^2_n \sim N(n,\sqrt{2n}) \quad \text{, asymptotiquement.}$$

En fait, dans ce cas particulier, l'approximation est encore meilleure pour la racine carrée :

$$\sqrt{2\chi^2_n} \sim N(\sqrt{2n-1},1) \quad \text{, asymptotiquement.}$$

Remarques : (i) Le théorème demeure valide si les termes de la somme sont des variables aléatoires dont les distributions sont différentes. Au surplus, des résultats du même type ont été obtenus pour des sommes d'aléas non indépendants.

(ii) Il n'a été ici question que d'aléas univariés ; la démonstration du théorème a aussi été étendue au cas multivarié, contexte dans lequel il garantit la convergence vers la multinormalité.

2.3.4. Théorème de Glivenko (1933)

Soient $X_1,X_2,\ldots,X_n \overset{iid}{\sim} F$

et soit l'échantillon { $X_1=x_1$,..., $X_n=x_n$ } de n observations x_i. Ces n observations sont rangées en ordre croissant, de la plus petite notée $x_{(1)}$, à la plus grande notée $x_{(n)}$:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

Les quantités $x_{(i)}$ sont des variables aléatoires, appelées statistiques d'ordre de l'échantillon. La fonction de répartition empirique liée à l'échantillon, notée $\hat{F}_n(x)$, est définie par :

$$\hat{F}_n(x) = \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ r/n & \text{si} \quad x_{(r)} \leq x < x_{(r+1)} \\ 1 & \text{si} \quad x_{(n)} \leq x \end{cases}$$

L'échantillon étant un n-uplet de variables aléatoires, la fonction $\hat{F}_n(x)$ est elle-même une fonction aléatoire. Sa réalisation, qui correspond à la réalisation $x_1,\ldots,x_n$ de l'échantillon, est aussi appelée fonction de répartition empirique, mais le contexte permet toujours de distinguer la fonction aléatoire d'une de ses réalisations.

La convergence des fréquences relatives vers les probabilités (loi faible des grands nombres) assure qu'en tout point x fixé, $\hat{F}_n(x)$ converge en probabilité vers F(x). Il existe cependant un théorème plus fort (théorème de Glivenko) qui garantit la convergence globale en tout x de $\hat{F}_n(x)$ vers F(x) :

$$\text{Proba}\{ \lim_{n\to\infty} (\sup_{-\infty<x<+\infty} | \hat{F}_n(x) - F(x)|) = 0 \} = 1$$

Ce résultat établit que pour n suffisamment grand, la fonction de répartition empirique restitue "une bonne image" de la vraie fonction de répartition F. Cela fonde les méthodes de rééchantillonnage (*i.e.*, bootstrap, jackknife), dans la mesure où ces dernières procèdent par saisie"d'échantillons" dans une estimation $\hat{F}_n$ de F, plutôt que dans F elle-même ; autrement dit, les "ré-échantillons" proviennent de l'échantillon et non de la population, étant créés par une technique de repondérations de la fonction de répartition empirique.

A N N E X E II

ECHANTILLONNAGE : PRESENTATION RESUMEE DES PRINCIPALES STRATEGIE.

ECHANTILLONNAGE : PRESENTATION RESUMEE DES PRINCIPALES STRATEGIES

INTRODUCTION : POPULATION-ECHANTILLON-INFERENCES

Les études expérimentales concernent souvent des populations trop vastes pour qu'elles soient appréhendées exhaustivement ; il est alors fait appel aux techniques de sondage, qui extraient de la population un groupe de ses éléments constitutifs : l'échantillon. Faire une inférence consiste ensuite à extrapoler à la population les propriétés observées sur l'échantillon. Une telle pratique est toujours plus ou moins entachée d'erreur. Lorsque l'échantillonnage est effectué "au jugé", i.e. lorsque les éléments entrant dans l'échantillon sont choisis d'après des critères subjectifs, il n'est pas possible d'évaluer l'amplitude de l'erreur commise. Pour pallier cette lacune, et afin de quantifier la précision des inférences, il est nécessaire d'élaborer un protocole attribuant à tout élément de la population une probabilité connue d'apparaître dans l'échantillon. Une telle stratégie d'**échantillonnage aléatoire**, qui associe à tout échantillon sa probabilité d'occurrence, permet d'estimer la fiabilité de l'extrapolation à la population parente : on parle alors d'**inférence statistique**.

Le préalable à toute inférence statistique inclut donc :

- la définition de la population étudiée, ou **population-cible** : c'est une collection d'objets identifiables, que l'on appellera individus, en nombre fini ou infini. Ses limites (dans le temps et/ou l'espace) doivent être clairement posées. Dans la pratique, on ne s'intéresse pas à l'ensemble des propriétés de chaque individu, mais seulement à certaines d'entre elles, nous dirons certaines caractéristiques ; l'objet de l'étude est donc une **population de caractéristiques**. Si à toute caractéristique de chaque individu est associée une valeur numérique, la population de caractéristiques peut être décrite par un ensemble de nombres ;

- la définition de la procédure de sélection **aléatoire** des individus qui vont constituer l'échantillon. Les procédures usuelles sont brièvement présentées ci-après.

1. EXEMPLE FONDAMENTAL : L'ECHANTILLONNAGE ALEATOIRE SIMPLE (EAS)

1.1. DEFINITION

Le protocole de formation de l'échantillon obéit aux trois règles suivantes :

a) Tous les individus sont **équiprobables**, i.e. ils ont chacun la même probabilité d'apparaître dans l'échantillon. En particulier, la sélection d'un individu n'est pas influencée par la valeur qu'il présente pour la caractéristique étudiée. Cette première règle est la condition des inférences simples sans biais ; elle est habituellement mise en pratique à l'aide d'une table de nombres aléatoires de distribution uniforme.

b) Les tirages sont mutuellement **indépendants** : la présence d'un individu dans l'échantillon ne modifie en rien les probabilités d'apparition des autres individus de la population dans ce même échantillon. En toute rigueur, si l'effectif de la population est fini, cette condition n'est respectée qu'en opérant des tirages **avec remise.**

La propriété d'indépendance entre les tirages est capitale : c'est elle qui autorise l'application des résultats établis pour les variables aléatoires indépendantes, tels que l'additivité des variances, ou encore ceux relatifs à l'espérance des produits de variables aléatoires.

Si la population est finie et qu'elle est échantillonnée **sans remise**, alors le tirage d'un individu augmente la probabilité de sélection de ceux qui ne sont pas encore entrés dans l'échantillon ; cela est à l'origine des termes de **correction pour population finie.**

c) La taille de l'échantillon est prédéterminée : c'est donc une quantité non aléatoire, i.e. certaine, et qui est traitée comme telle dans les calculs qui s'en trouvent de beaucoup simplifiés.

1.2. NOTION DE DISTRIBUTION D'ECHANTILLONNAGE : EXEMPLE DE LA MOYENNE

Soit une population finie de caractéristiques X_i, et d'effectif N ($i=1,\ldots,N$). La structure de cet ensemble de nombres peut être en partie résumée par des **paramètres** descripteurs de la population, i.e. des quantités inconnues que l'on souhaite estimer. Par exemple, la vraie moyenne : $\mu = (1/N) \sum_{i}^{N} X_i$, qui est un indice de la tendance centrale des X_i.

De même définit-on des paramètres qui expriment la dispersion de la caractéristique :

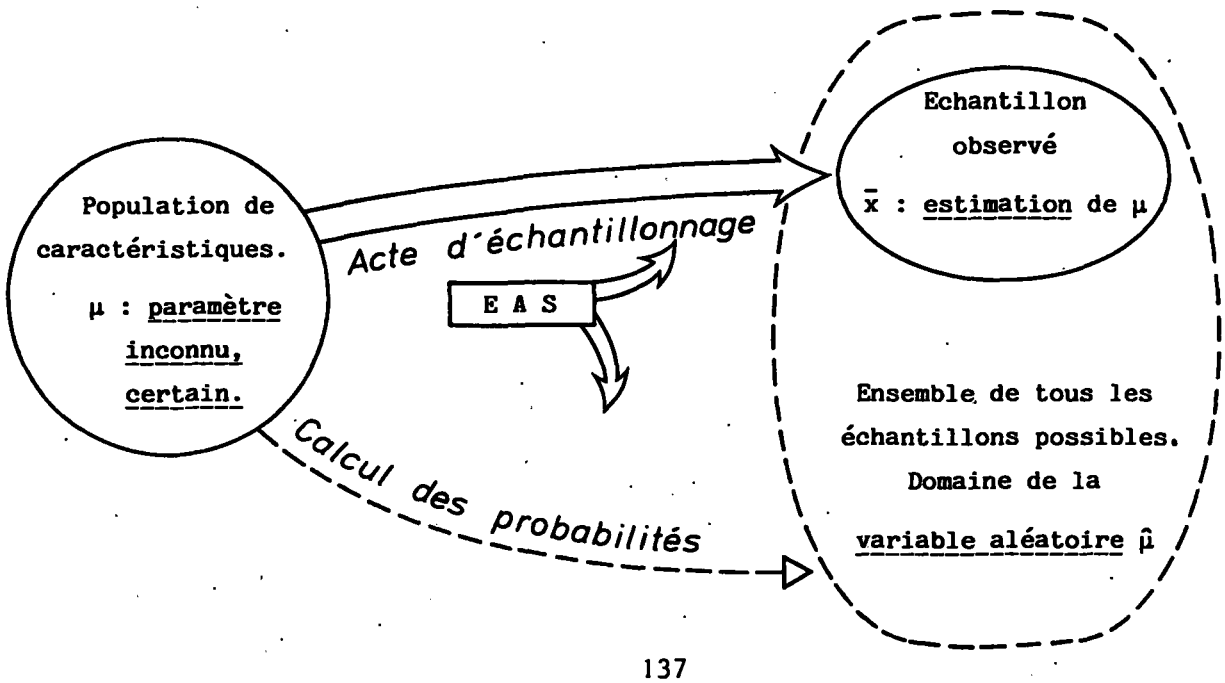
la vraie variance : $\sigma^2 = (1/N) \sum^{N} (X_i - \mu)^2$, ou bien encore la "variance corrigée" : $S^2 = (N/(N-1))\sigma^2$

Afin d'estimer μ, un échantillon aléatoire simple de n individus est formé, qui contient donc n observations indépendantes x_i (i=1,...,n) de la caractéristique X. Ces observations permettent de calculer les **estimations** suivantes :

la moyenne empirique : $\bar{x} = (1/n) \sum^{n} x_i$
ainsi que la variance empirique : $s^2 = (1/(n-1)) \sum^{n} (x_i - \bar{x})^2$

Le point fondamental de toute méthode d'échantillonnage aléatoire, qui autorise la quantification de la précision associée aux estimations, est le suivant : le protocole, en même temps qu'il conduit à l'obtention d'un échantillon, engendre une collection d'entités abstraites, formée de l'ensemble de tous les échantillons que l'on aurait pu obtenir. La définition de ce second ensemble repose sur les probabilités de sélection attachées aux individus de la population.

Par conséquent, lorsque la moyenne inconnue μ de la population est estimée par EAS, l'acte d'échantillonnage crée une variable aléatoire $\hat{\mu}$, que l'on appelle **l'estimateur** de la moyenne, et dont l'ensemble des valeurs possibles est constitué de celles que prendrait $\hat{\mu}$ dans chacun des échan-



tillons que l'on aurait pu former. Dans l'échantillon effectivement obtenu est observée l'estimation $\bar{x}$, qui est une réalisation particulière de la variable aléatoire $\hat{\mu}$ (schéma page précédente).

La variable aléatoire $\hat{\mu}$ est caractérisée par sa distribution, qui est appelée **distribution d'échantillonnage de l'estimateur** du paramètre inconnu μ. Dans le cas de la moyenne empirique, le théorème de la limite centrale garantit que la distribution d'échantillonnage de $\hat{\mu}$ se rapproche de la normalité quand la taille n de l'échantillon augmente. En effet, $\hat{\mu}$ est une somme de n variables aléatoires indépendantes (point b de la définition de l'EAS) et de même loi. Quelle que soit, dans une large mesure, cette loi commune de la répartition des valeurs de la caractéristique dans la population, la distribution d'échantillonnage de $\hat{\mu}$ tend vers une loi normale d'espérance $E(\hat{\mu}) = \mu$, et de variance $V(\hat{\mu})$ dont la valeur sera donnée plus loin. C'est cette variance qui quantifie la précision de l'estimation $\bar{x}$: plus $V(\hat{\mu})$ est faible, et plus on a de chances que l'estimation $\bar{x}$ soit proche de la vraie valeur μ.

L'exemple de la moyenne est particulièrement simple, en ce sens que la distribution d'échantillonnage de $\hat{\mu}$ est approximativement connue sans qu'il soit nécessaire de formuler des hypothèses sur la loi de la caractéristique X dans la population. Mais en général, pour déterminer la distribution d'échantillonnage d'un estimateur, il faut associer au protocole de sélection aléatoire des individus la définition d'un corps d'hypothèses probabilistes relatives à la caractéristique X étudiée. Ainsi, si l'on suppose que X est une variable aléatoire réelle continue normalement distribuée dans la population-cible, alors :

- la loi de $\hat{\mu}$ est elle-même normale, quelle que soit la taille de l'échantillon,
- et la quantité $\sum^{n} (x_i - \bar{x})^2 / \sigma^2$ suit une loi de χ^2 à n-1 d.d.l., propriété qui permet d'attribuer un intervalle de confiance à la variance empirique s^2.

1.3. PRINCIPAUX RESULTATS

La notation qui distingue la variable aléatoire ($\hat{\mu}$, ou bien $\hat{\sigma}^2$) de sa réalisation ($\bar{x}$, ou bien s^2) sera maintenant abandonnée. Ainsi qu'il est d'usage, $\hat{\mu}$ comme $\hat{\sigma}^2$ désigneront aussi bien l'estimateur que l'estimation, le contexte ne prêtant généralement pas à confusion.

Population finie de caractéristiques

X_i, d'effectif N, $i = 1, \ldots, N$.

Paramètres descripteurs

$$\mu = \sum_{1}^{N} X_i / N$$

$$\sigma^2 = \sum_{1}^{N} (X_i - \mu)^2 / N$$

$$S^2 = \frac{N}{N-1} \sigma^2$$

E A S avec TIRAGES SANS REMISE

Echantillon de n observations x_i

$$\hat{\mu} = \sum_{1}^{n} x_i / n$$

$$E(\hat{\mu}) = \mu$$

$$V(\hat{\mu}) = \frac{\sigma^2}{n} \left(\frac{N - n}{N - 1}\right) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right)$$

$$\hat{\sigma}^2 = \sum_{1}^{n} (x_i - \hat{\mu})^2 / (n - 1)$$

$$E(\hat{\sigma}^2) = S^2$$

$$\hat{V}(\hat{\mu}) = \frac{\hat{\sigma}^2}{n}\left(1 - \frac{n}{N}\right)$$

E A S et TIRAGES AVEC REMISE

Echantillon de n observations x_i

$$\hat{\mu} = \sum_{1}^{n} x_i / n$$

$$E(\hat{\mu}) = \mu$$

$$V(\hat{\mu}) = \sigma^2/n$$

$$\hat{\sigma}^2 = \sum_{1}^{n} (x_i - \hat{\mu})^2/(n - 1)$$

$$E(\hat{\sigma}^2) = \sigma^2$$

$$\hat{V}(\hat{\mu}) = \hat{\sigma}^2/n$$

Remarques :

(i) Pour des tirages avec remise, la variance de l'estimateur $\hat{\mu}$ est indépendante de la taille N de la population. Ce n'est pas le cas des tirages sans remise, où l'effectif N apparaît dans le terme de correction pour population finie 1-n/N.

(ii) Les tirages avec ou sans remise sont équivalents lorsque la population est infinie.

(iii) Le schéma qui définit les relations entre population, échantillon et estimateur peut être mis à profit pour présenter un point de vue distinct de celui adopté dans la présente annexe, comme d'ailleurs dans l'ensemble du rapport : il s'agit du point de vue bayésien ; selon cette seconde con-

ception, la moyenne inconnue μ n'est plus considérée comme une quantité fixée, mais comme une variable aléatoire caractérisée par une loi, dite "distribution a priori de μ", les paramètres de cette loi étant des constantes connues. La valeur observée $\bar{x}$ est donc elle-même issue d'une loi qui dépend de la valeur prise par l'aléa μ ; le caractère conditionnel du processus probabiliste qui régit $\bar{x}$ peut être rappelé par la notation $\bar{x}|\mu$.

Le théorème de Bayes permet alors de combiner la distribution a priori de μ et la distribution conditionnelle de $\bar{x}|\mu$ pour en déduire la distribution conditionnelle de $\mu|\bar{x}$: cette dernière est appelée "distribution a posteriori de μ pour la valeur observée $\bar{x}$". L'estimation bayésienne de μ est simplement la moyenne de la distribution a posteriori de $\mu|\bar{x}$.

Le point de vue bayésien ne prend donc en compte que la valeur expérimentale $\bar{x}$ effectivement observée, sans faire appel à l'ensemble des autres valeurs de $\bar{x}$ théoriquement possibles ; et c'est le paramètre μ, considéré comme variable, qui est concerné par le processus de moyennage. EFRON (1978) analyse les difficultés engendrées par la coexistence de solutions statistiques, fondamentalement différentes, au problème du choix de la "meilleure moyenne" devant fonder des inférences. Quoiqu'il en soit, ces controverses demeurent connexes au problème de l'échantillonnage, et ne seront désormais plus évoquées.

(iv) Le principal avantage de l'EAS tient à ce qu'il ne requiert aucune connaissance a priori de la population, sinon bien évidemment ses limites. Il autorise de plus des inférences simples, et fournit des estimateurs généralement non biaisés. En contrepartie, l'EAS est souvent difficile à mettre en oeuvre, car le protocole de tirage nécessite la liste complète des individus de la population. Au surplus, l'atout de n'utiliser aucune information a priori s'accompagne le plus souvent d'une perte d'efficacité. Afin de pallier ce dernier inconvénient ont été proposées des stratégies qui prennent explicitement en compte les "structures fortes" reconnues au sein de la population. Les stratégies les plus classiques sont succinctement présentées ci-après ; un développement relativement plus substantiel est toutefois consacré à une notion qui, pas plus que la définition de l'EAS, ne peut être ignorée : la stratification.

2. LES PRINCIPALES STRATEGIES D'ECHANTILLONNAGE

2.1. LA NOTION DE STRATIFICATION

Les valeurs de la caractéristique se répartissent en général de manière hétérogène dans la population-cible ; par conséquent, σ^2 prend une valeur élevée. Cela affecte la précision des estimations si l'objectif est par exemple d'estimer par EAS la moyenne μ : nous venons en effet de voir que $V(\hat{\mu})$ est proportionnelle à σ^2.

Le but de la stratification, comme celui de la plupart des autres stratégies, est de réduire la variance des estimateurs. Cela est possible quand les individus de la population peuvent être rassemblés en quelques blocs, formés de telle manière qu'ils regroupent les individus les plus semblables entre eux, et que les individus appartenant à deux blocs distincts soient aussi différents que possible. Ces blocs, appelés **strates**, réalisent une **partition** de la population en sous-ensembles relativement homogènes du point de vue de la caractéristique étudiée, l'hétérogénéité initiale étant transférée principalement entre les strates.

Les deux contraintes essentielles liées à l'échantillonnage stratifié sont les suivantes :

a) **Toutes** les strates sont échantillonnées.

b) L'échantillonnage dans une strate est **indépendant** de ceux effectués dans les autres strates.

Sous ces deux conditions, tout protocole d'échantillonnage aléatoire peut être appliqué dans les strates, les protocoles n'étant pas nécessairement identiques d'une strate à l'autre. Toutefois, seul le cas où chaque strate fait l'objet d'un EAS sera envisagé ici.

2.1.1. Echantillonnage aléatoire stratifié

La population d'effectif N est partitionnée entre k strates ($h = 1,\ldots,k$) d'effectif N_h :

$$\sum_{h=1}^{k} N_h = N$$

Un échantillon de n observations x_i est formé, en extrayant de chaque strate un EAS de n_h individus tirés **sans remise** :

$$\sum_{h}^{k} n_h = n$$

Estimation de la moyenne ; dans la strate h : $\hat{\mu}_h = \sum_{i=1}^{n_h} x_i/n_h$

dans la population : $\hat{\mu} = (1/N) \sum_{h=1}^{k} N_h\hat{\mu}_h$

Variance de l'estimateur $\hat{\mu}$; cas général :

$$V(\hat{\mu}) = (1/N^2) \sum^{k} N_h^2(\sigma^2/n_h)(1 - n_h/N_h)$$

où σ_h^2 est la vraie variance de la caractéristique dans la strate h. On l'estime par $\hat{\sigma}_h^2$:

$$\hat{\sigma}_h^2 = \sum^{n_h}(x_i-\hat{\mu}_h)^2/(n_h-1)$$

et $\hat{V}(\hat{\mu})$ est obtenue en remplaçant σ_h^2 par son estimation $\hat{\sigma}_h^2$ dans l'expression de $V(\hat{\mu})$.

Répartition de l'effort d'échantillonnage dans les différentes strates : deux cas particuliers sont à considérer.

- allocation proportionnelle : chaque strate est échantillonnée proportionnellement à son effectif N , <u>i.e.</u> : $n_h/N_h = n/N$ = une constante notée f, et appelée **taux d'échantillonnage**. Avec cette notation :

$$V(\hat{\mu}) = (1-f)/(fN^2) \sum^{k} N_h\sigma_h^2$$

Cette variance de l'estimateur $\hat{\mu}$ est, sauf rares exceptions, inférieure à celle qui serait obtenue avec un EAS sans stratification, et d'autant plus inférieure que les moyennes μ_h des strates diffèrent entre elles.

- allocation optimale : le taux d'échantillonnage n'est plus constant, mais dépend de l'effectif N_h de la strate et de sa variabilité, quantifiée par σ_h.

$$n_h/N_h = n\ \sigma_h\ /\ (\sum_h N_h\ \sigma_h) \qquad h = 1,\ldots,k$$

avec ces taux, la variance $V(\hat{\mu})$ est plus faible que celle obtenue avec l'allocation proportionnelle, et d'autant plus faible que les variances intra-strate σ_h^2 diffèrent entre elles.

2.1.2. Mise en oeuvre

Au plan pratique, la réalisation d'un EAS stratifié inclut les étapes suivantes :

1 - Le choix du critère de stratification, ou **stratificateur** : ce peut être la caractéristique X elle-même, si l'on possède une idée approximative de sa distribution, par exemple à la suite d'une **étude pilote** : i.e., une enquête préliminaire qui fournit une information générale sur la structure de la population. Le stratificateur peut aussi être une **variable auxiliaire**, i.e. une variable fortement corrélée à X, et dont la saisie est beaucoup moins coûteuse.

2 - Le choix du nombre de strates : l'expérience montre que le gain de précision s'atténue rapidement lorsque ce nombre croît, et qu'en pratique il n'est généralement pas utile de définir plus de 6 strates.

3 - Le choix des limites des strates : établies de telle sorte que soit minimisée la variance intra-strate, et donc maximisée la variance inter-strates. La fonction de répartition empirique du stratificateur peut être utilisée à cette fin.

4 - Le dénombrement des individus de chaque strate : préalable à leur sélection par EAS ; c'est habituellement l'étape la plus fastidieuse.

5 - Le choix du taux d'échantillonnage : quelle que soit l'allocation retenue, il est nécessaire d'extraire au minimum deux individus de chaque strate, afin de pouvoir estimer les variances intra-strate σ_h^2.

Enfin, le calcul de $\hat{\mu}$ et de $V(\hat{\mu})$ fait intervenir les **poids des strates** $W_h = N_h/N$. Lorsque ces poids sont inconnus, il est nécessaire de les estimer préalablement. On applique alors une procédure dite de **double échantillonnage** : la population fait d'abord l'objet d'un EAS (estimation des W_h), puis dans un second temps, de l'échantillonnage stratifié proprement dit (estimation de μ et de $V(\hat{\mu})$).

2.2. ECHANTILLONNAGE PAR GRAPPES ET STRATEGIES DERIVEES

La stratification consiste à reconnaître quelques grands ensembles homogènes dans la population, puis à extraire un échantillon de chacun d'eux. La définition des **grappes** procède d'une démarche opposée : la population-cible est partitionnée en sous-ensembles (aussi nombreux que l'impose l'étude), définis de telle sorte que chacun soit le plus hétérogène possible. Autrement dit, le partage de la variance totale de la caractéristique est tel que sa variance intra-grappe soit maximale, et sa variance inter-grappes minimale. Une fois ces grappes définies, un échantillon aléatoire (simple ou non, avec ou sans remise) de grappes est formé : l'unité d'échantillonnage est donc ici la grappe elle-même, et chaque grappe sélectionnée est étudiée exhaustivement (i.e., tous les individus de la grappe entrent dans l'échantillon).

Cette opposition théorique entre grappes et strates vaut pour les grappes rationnellement décidées. Formellement, pour une population partitionnée en sous-ensembles :

$$(N-1)\,\sigma^2 = \sum_h (N_h-1)\,\sigma_h^2 + \sum_h N_h(\mu_h-\mu)^2$$

Ecart quadratique total de la caractéristique autour de μ	Terme maximisé par la formation des grappes	Terme maximisé par la formation des strates

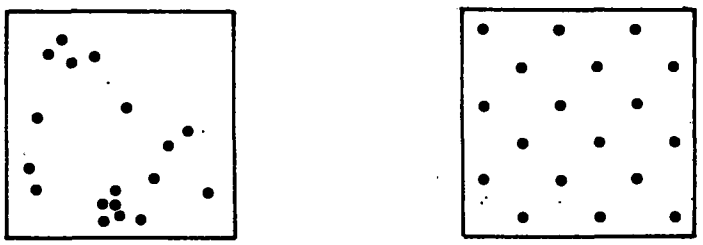
Mais en pratique, les grappes ne sont pas toujours décidées selon ce principe, il arrive qu'elles soient obligées : si la population est fractionnée en sous-ensembles trop nombreux pour que chacun puisse être échantillonné, il n'est pas possible de les considérer comme des strates, force est alors de les assimiler à des grappes, quelles que soient les composantes inter- et intra- de la variance totale.

Enfin, il est souvent difficile, voire impossible, d'étudier exhaustivement les grappes sélectionnées lorsque leurs effectifs sont trop élevés. On adopte alors une procédure d'échantillonnage des individus dans ces grappes, dite de **sous-échantillonnage**, ou **échantillonnage à deux niveaux**. Plus généralement peuvent être envisagées des procédures d'**échantil-**

lonnage à plusieurs niveaux, dès l'instant qu'à chaque niveau (sauf au dernier) les unités d'échantillonnage sont trop vastes pour être étudiées en totalité.

2.3. ECHANTILLONNAGE SYSTEMATIQUE. VARIABLES REGIONALISEES

De même que l'EAS, l'échantillonnage systématique peut être appliqué à une population dont on ne connaît pas a priori la structure. Mais la sélection des individus procède de manière radicalement différente dans les deux stratégies. Cela est illustré par le schéma ci-après, où sont identifiées les unités échantillonnées à l'intérieur des limites spatiales d'une population : par exemple, les parcs d'une concession ostréicole, les arbres d'une forêt, etc.



A la différence des unités sélectionnées par EAS, celles choisies de façon systématique ne sont pas indépendantes. En effet, cette dernière stratégie maximise les distances entre les unités d'échantillonnage : le choix de l'une d'elles détermine le choix des autres. En procédant ainsi, on évite les redondances et les lacunes de l'EAS (le schéma de gauche montre qu'à côté d'"essaims" d'unités voisines existent des zones non prospectées). En contrepartie, il est difficile d'obtenir avec un tel protocole une estimation réaliste de la précision des résultats. Pour éclairer ce point, considérons les unités sélectionnées du schéma de droite ; on peut considérer qu'elles forment une grappe spatialement discontinue, et qui, étant exhaustivement étudiée, donne une estimation de la variance intra-grappe. Mais c'est aussi la seule grappe tirée, ce qui ne permet pas d'estimer la variance inter-grappes. Il n'est donc pas possible d'estimer la variance des estimateurs.

Le recours est apporté par la théorie des variables régionalisées, qui sont des fonctions de chaque point de l'espace. Elles présentent à la fois un aspect aléatoire et un aspect structuré, et la mise en évidence du

second est le but de l'étude de ces variables. Cette approche est spécialement intéressante quand il existe une continuité spatiale (et/ou temporelle) dans la population, et qu'en moyenne deux individus se ressemblent d'autant plus qu'ils sont proches. Les méthodes issues de la théorie des variables régionalisées (estimation du variogramme s'il existe, techniques de krigeage, fonctions aléatoires intrinsèques) débordent le cadre de ce résumé.

2.4. ECHANTILLONNAGE AVEC VARIABLE AUXILIAIRE

La stratification est une stratégie qui s'adapte à la structure de la population-cible ; dans le cas des grappes, il s'agit même souvent d'une structure subie. L'échantillonnage avec variable auxiliaire relève d'un principe différent : au lieu de l'adéquation à la structure de la population, cette stratégie utilise la connaissance apportée par une variable auxiliaire Y, peu coûteuse si possible, et corrélée à la caractéristique étudiée X. Les aménagements apportés en vue d'améliorer l'estimation de la moyenne de X concernent non pas le protocole de saisie des individus (i.e., le choix des probabilités de sélection), mais directement l'estimateur lui-même : ainsi sont définies la famille des **estimateurs rapport**, et celle des **estimateurs par régression.**

Dans l'expression de ces estimateurs apparaît la vraie moyenne μ_y de la variable auxiliaire Y. Cela suppose connu μ_y. Si tel n'est pas le cas, ce paramètre doit être préalablement estimé, d'où une procédure de **double échantillonnage.**

2.5. L'OPTIMISATION DU COMPROMIS COUT-PRECISION

Jusqu'ici, la finalité de l'échantillonnage n'a été présentée que d'un seul point de vue : celui de la réduction de la variance des estimateurs. En pratique, le problème ne se pose pas seulement de cette manière, car à la réalisation de l'étude n'est alloué qu'un budget limité ; cette enveloppe financière devra en particulier couvrir l'accès aux unités d'échantillonnage (e.g. des grappes géographiquement éloignées), la saisie des individus sélectionnés, et pour chacun d'eux la mesure de la (ou des) caractéristique(s) étudiée(s). Cela impose d'intégrer dans la définition

d'une stratégie la prise en compte d'un double souci : celui de la précision du résultat, tout autant que celui de son coût. Formellement, si la variance d'estimation est notée V, et si C désigne la fonction de coût, la stratégie optimale est celle qui minimise le produit C.V.

Du point de vue le plus général, les variances d'estimation sont souvent de la forme :

$$V = v_o + \sum_{r=1}^{k} v_r/w_r$$

On peut se ramener à cette écriture pour une stratégie faisant intervenir k strates, ou bien k grappes, ou encore k niveaux. Dans l'expression ci-dessus, les écarts quadratiques v_r ne dépendent que des paramètres de la population, et les pondérations w_r, qui sont indépendantes des v, correspondent de manière plus ou moins directe à l'effort d'échantillonnage, i.e. au nombre d'observations.

De même la fonction de coût s'exprime-t-elle couramment :

$$C = c_o + \sum_r w_r c_r$$

où c_o représente les frais fixes globaux, et c_r le coût unitaire associé à la catégorie r (e.g. la strate, ou bien la grappe, ou bien le niveau r). Le produit C.V vaut donc :

$$(V-v_o)(C-c_o) = (\Sigma_r v_r/w_r)(\Sigma_r w_r.c_r)$$

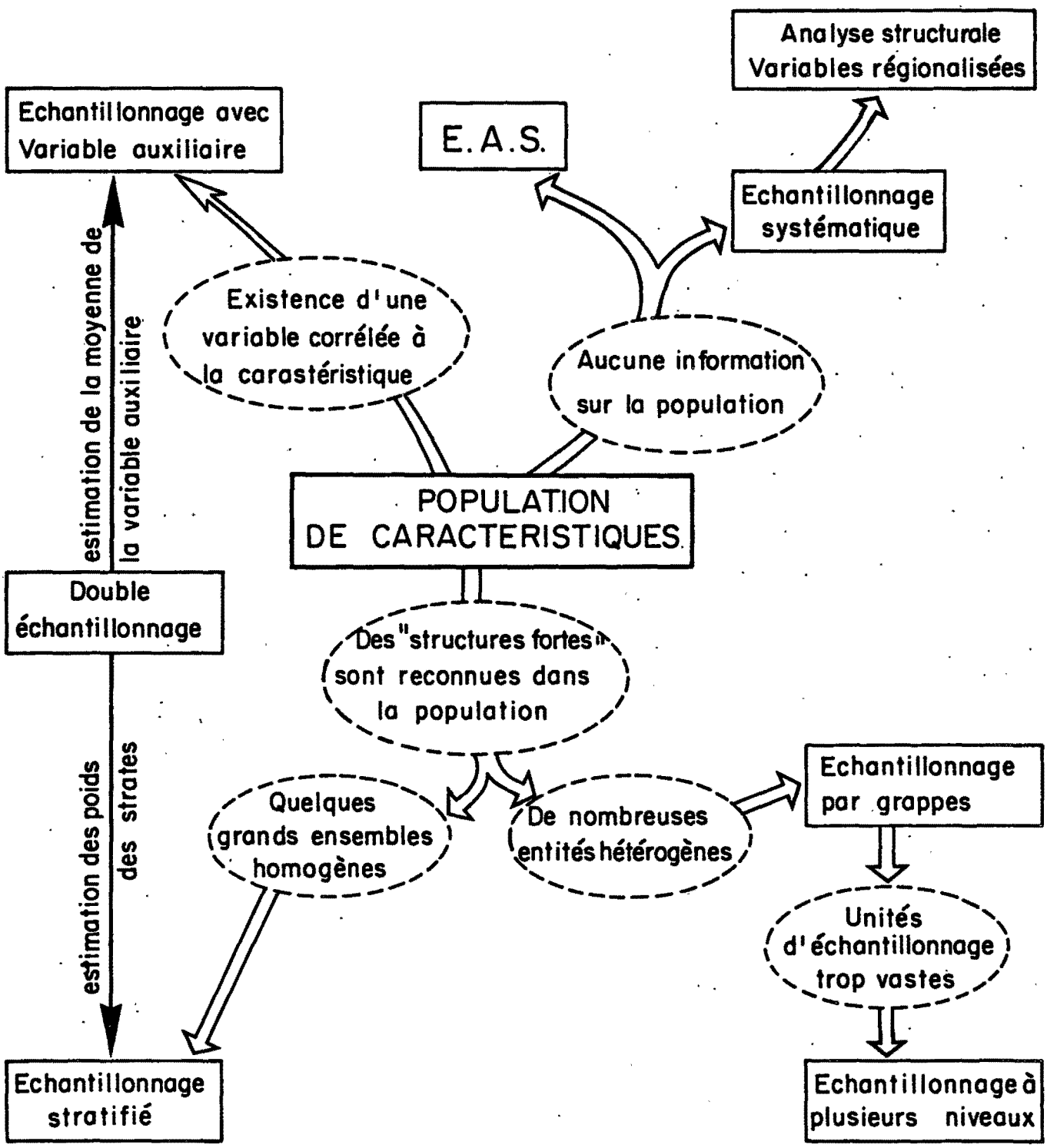
et la condition nécessaire et suffisante de minimum s'énonce :

$$w_r^2 \propto v_r/c_r \qquad \text{pour tout } r = 1,\ldots,k$$

Cette relation permet de répondre aux deux questions suivantes, dont l'intérêt pratique est évident :

- celle du budget à prévoir pour obtenir une précision fixée à l'avance,
- ou bien celle de la précision à attendre d'un coût donné.

3. RECAPITULATION



Il convient de souligner, pour conclure, que les quelques stratégies présentées l'ont été séparément par souci de didactisme. Dans la pratique, il est cependant rare que l'on soit amené à n'utiliser qu'un seul des protocoles précédemment décrits : ce sont plutôt des **combinaisons de stratégies** qui sont mises en oeuvre. Par exemple, on reconnaît d'abord quelques strates, certaines pouvant être elles-mêmes partitionnées en grappes, ces dernières faisant éventuellement l'objet d'un échantillonnage à plusieurs niveaux. Il ne faut toutefois pas perdre de vue que plus le protocole devient complexe, plus le calcul des estimateurs se complique, et plus difficiles sont les inférences (KISH & FRANKEL, 1974).

Par ailleurs, la présente annexe ne saurait offrir les connaissances nécessaires à la mise en application des techniques qui y sont succinctement présentées : elle a pour ambition de constituer une introduction à la lecture d'ouvrages spécialisés, par exemple COCHRAN (1977), SCHERRER (1982), KENDALL et al. (1983, chapitres 39 & 40), ou encore GILBERT (1987, chapitres 3 à 9).

REFERENCES ADDITIONNELLES

COCHRAN, W.G. (1977)
Sampling techniques.
John Wiley & Sons, 3rd ed., 428 p.

EFRON, B. (1978)
Controversies in the foundations of statistics.
Amer. Math. Monthly 85 : 231-246.

GILBERT, R.O. (1987)
Statistical methods for environmental pollution monitoring.
Van Nostrand Reinhold Company Inc., New York, 320 p.

KENDALL, M., A. STUART, & J.K. ORD (1983)
The advanced theory of statistics, Vol. 3, Design and analysis, and time-series.
Charles Griffin & Co. Ltd., London & High Wycombe, 780 p.

KISH, L., & M.R. FRANKEL (1974)
Inference from complex samples (with discussion).
J.R. Statist. Soc. B 36 (1) : 1-37.

SCHERRER, B. (1982)
Techniques de sondage en écologie ; pp. 63-162 in : Stratégies d'échantillonnage en écologie, S. FRONTIER éd., Masson, Paris & Les Presses de l'Université Laval-Québec, 494 p.

LEXIQUE FRANCAIS-ANGLAIS

EAS (échantillonnage aléatoire simple) avec (sans) remise	Simple random sampling with (without) replacement
Echantillonnage aléatoire stratifié	Stratified random sampling
Allocation proportionnelle	Uniform sampling fraction
Allocation optimale	Minimum variance allocation, optimum allocation
Echantillonnage par grappes	Cluster sampling
Sous-échantillonnage, ou échantillonnage à deux niveaux	Subsampling, two-stage sampling
Echantillonnage à plusieurs niveaux, ou échantillonnage par degrés	Multi-stage sampling
Echantillonnage systématique	Systematic sampling
Double échantillonnage	Two-phase sampling, double sampling
Estimateur rapport	Ratio estimator
Estimateur par régression	Regression estimator

Achevé d'imprimer
au Centre IFREMER-Brest

4ème trimestre 1988

Le rapport présente les éléments fondamentaux nécessaires à la définition et à la mise en œuvre d'un protocole de surveillance du milieu marin au voisinage d'un aménagement de la façade littorale. L'application de la théorie statistique de la décision est d'abord examinée, spécialement du point de vue de l'évaluation de la puissance des tests d'hypothèses (paramétriques, non paramétriques) ; des développements plus récents sont également abordés : notion de robustesse, techniques de rééchantillonnage. Les aspects opérationnels sont ensuite discutés, et des préceptes sont formulés pour le choix (i) de la variable indicative, (ii) des échelles spatio-temporelles d'observation, et (iii) du critère d'optimalité permettant de quantifier le rapport coût/précision d'une stratégie donnée.

mots clés : stratégie de surveillance, tests d'hypothèses, statistique paramétrique - non paramétrique - robuste, techniques de rééchantillonnage.

This report is devoted to the problem of testing statistical hypotheses within the general framework of impact assessment studies : the aim is to supply some rules concerning the definition and implementation of an ecological survey program. Accordingly, power calculations are performed for the usual parametric and nonparametric tests ; more recent developments are also examined : robustness, resampling techniques. Operational basic options are then discussed, and guidelines are provided for the choice (i) of the diagnostic variable, (ii) of the spatio-temporal scales of observation, and (iii) of the objective function for optimizing the allocation of sampling, i.e. the cost-efficiency tradeoff.

key words — monitoring strategy, statistical decision theory, parametric & nonparametric tests, robust statistics, resampling techniques.

Service de la Documentation
et des Publications (S.D.P.)
IFREMER - Centre de Brest
BP 70 - 29263 PLOUZANÉ
Tél. 98 22 40 13 - Télex 940 627F

ISSN - 0761-3970