UNIVERSITÉ DE ROUEN

U.F.R. DE PSYCHOLOGIE, SOCIOLOGIE ET DES SCIENCES DE L'ÉDUCATION

Méthodologie de l'analyse des données expérimentales : Étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive

THÈSE pour l'obtention du grade de DOCTEUR DE L'UNIVERSITÉ DE ROUEN
Discipline : Psychologie

présentée par Jacques POITEVINEAU

sous la direction de Bruno LECOUTRE, Directeur de Recherche au C.N.R.S.

Soutenue publiquement le 11 mars 1998 devant le jury composé de :

Présidente	:	Danièle DUBOIS,	Directeur de Recherche au C.N.R.S.
Rapporteurs	:	Jean-Michel HOC,	Directeur de Recherche au C.N.R.S.
		Jacques LAUTREY,	Professeur à l'université de Paris V
Examinateurs	:	Claude LEMOINE,	Professeur à l'université de Rouen
	:	Alain VOM HOFE,	Professeur à l'université de Rouen
Directeur	:	Bruno LECOUTRE,	Directeur de Recherche au C.N.R.S.

Remerciements

Le travail achevé, voici venu le moment des remerciements. Remerciements ou ... malédictions ? Car, il faut le savoir, cette thèse est le fruit d'un complot, odieux comme il se doit, perpétré par Bruno Lecoutre et ma femme et leurs dénégations ne changent rien à l'affaire. Leur succès n'a été dû qu'à la simultanéité de leur attaque (ils osent encore parler de coïncidence) combinée à la surprise : cela fait à peine plus de dix ans que Bruno Lecoutre, au cours de conversations de travail, glissait perfidement des "grff mmhh bbll THÈSE ???".

Enfin, après consultation des entrailles d'un 386 (sans coprocesseur, faut pas gâcher) sacrifié pour l'occasion, les augures m'ont conseillé d'opter pour les remerciements. Soit ! De toute façon, ma vengeance est d'ores et déjà accomplie puisque Bruno Lecoutre a dû lire les n + 1 versions de ce travail, n étant très grand et 1 également.

Je remercie donc très chaleureusement Bruno Lecoutre pour m'avoir proposé ce sujet, pour m'avoir guidé avec patience pendant ces trois ans, pour m'avoir consacré son temps sans compter, pour m'avoir fait profiter de sa très grande compétence qui n'a d'égale que sa modestie. "Nobody's perfect directeur of thèse" peut-être, mais presque. C'est un euphémisme de dire que cette thèse lui doit beaucoup; sauf en ce qui concerne les erreurs, abus, incorrections, omissions, fautes (d'orthographe, de style, de syntaxe, de grammaire) et autres que j'assume totalement (quoique ...; pour l'orthographe, Monsieur Word a sa part de responsabilité). J'associe à ces remerciements Marie-Paule Lecoutre pour avoir bien voulu m'associer à certains de ses travaux mais aussi pour avoir lu cette thèse sans y être obligée (!), pour les remarques et conseils pertinents qui ont découlé de cette lecture, et pour ses constants encouragements et sa gentillesse (c'est de famille).

En m'acceptant dans le Groupe Mathématique et Psychologie, Henry Rouanet m'a permis de côtoyer une pensée statistique de pointe; qu'il en soit remercié.

Je suis profondément reconnaissant à Danièle Dubois de m'avoir accueilli dans son laboratoire du C.N.R.S., de s'être intéressée à mes travaux et de m'avoir fourni des conditions de travail idéales. Je la remercie également pour l'ambiance conviviale et particulièrement propice à la réflexion qu'elle sait faire régner dans son équipe, ainsi que pour son accord spontané à présider ce jury de thèse.

Merci à Claude Lemoine et à Alain Vom Hofe d'avoir accepté d'examiner ce travail.

Jean-Michel Hoc et Jacques Lautrey ont bien voulu être mes rapporteurs; je leur sais gré de s'être chargés de cette lourde tâche. Qu'ils soient assurés également de ma respectueuse considération.

Je remercie bien sincèrement tous ceux qui m'ont plusieurs fois communiqué des références bibliographiques intéressantes, en particulier Guy Denhière et Gérard Derzko.

Mille mercis à tous les chercheurs qui ont joué le rôle de cobayes dans nos expériences.

Enfin, je remercie mon grand petit bonhomme et sa maman. Leur amour a été un soutien inépuisable et remettait promptement les choses en place dans les moments de lassitude.

TABLE DES MATIÈRES

TABLE DES MATIÈRES 5

INTRODUCTION 11

1ÈRE PARTIE APPROCHE NORMATIVE 15

CHAPITRE 1 DEUX THÉORIES FRÉQUENTISTES DES TESTS STATISTIQUES 19

1.1. LA THÉORIE DE FISHER 19

1.2. LA THÉORIE DE NEYMAN ET PEARSON 21

1.3. LE CADRE FRÉQUENTISTE 24

1.4. LA QUESTION DE LA PROBABILITÉ DES HYPOTHÈSES 25

1.5. UN AMALGAME DES DEUX THÉORIES 26

1.6. UN PEU DE TERMINOLOGIE 27

CHAPITRE 2 CRITIQUES ET ABUS DES TESTS 29

2.1. CRITIQUE DES TESTS 29
2.1.1. Quelle population ? 29
2.1.2. *Bis repetita* 29
2.1.3. De l'arbitraire… 31
2.1.4. Critique du "postulat α" 31
2.1.5. Et le principe de vraisemblance ? 32
2.1.6. Être ou ne pas être… observé 32
2.1.7. Chassez le naturel… 33
2.1.8. Décision ou jugement 33
2.1.9. Discontinu ou continu 34
2.1.10. Un problème dérivé : un biais de publication 34
2.1.11. Une hypothèse inutile 35
2.1.12. Corruption de méthode 35
2.1.13. Le statut de l'hypothèse de recherche 35
2.1.14. La question de l'intensité de l'effet 36
2.1.15. Le paradoxe fondamental 38

2.2. LES ABUS OU ERREURS D'INTERPRÉTATION 38
2.2.1. Le renversement des conditions : de *Pr*(*Données*|*Hypothèse*) à *Pr*(*Hypothèse*|*Données*) 38
2.2.2. 1 - *p* (ou 1 - α) considéré comme probabilité de reproduction du résultat 39
2.2.3. Significativité statistique et significativité substantielle 40
2.2.4. L'acceptation de l'hypothèse nulle 41
2.2.5. L'omission de la condition 41
2.2.6. Tel est pris… 42

2.3. DES RAISONS DES ABUS ET DE LEUR PERSISTANCE 42
2.3.1. La popularité des tests et l'existence des abus 43
2.3.2. La persistance de la popularité des tests et des abus 44

2ÈME PARTIE APPROCHE PRESCRIPTIVE 48

CHAPITRE 3 LES SOLUTIONS DE RECHANGE PRÉCONISÉES 52

3.1. LA MESURE DE LA GRANDEUR DE L'EFFET 52
3.1.1. Le plus simple 52
3.1.2. Les indicateurs de grandeur d'effet de Cohen (1962, 1969) 52
3.1.3. Les indicateurs en “part de variance expliquée” 53
3.1.4. L'étude de tableaux de contingence 53
3.1.5. Effets bruts ou effets relatifs ? 54
3.1.6. Une étape indispensable mais insuffisante 55

3.2. L'ÉTUDE DE LA PUISSANCE 55
3.2.1. La puissance peut être un guide utile pour la planification des expériences 55
3.2.2. Les problèmes posés par l’utilisation de la puissance pour interpréter les données 55

3.3. L'INTERVALLE DE CONFIANCE 57
3.3.1. Les mêmes critiques que les tests de signification 57
3.3.2. Des conclusions surprenantes 57
3.3.3. Des intervalles qui ne sont pas toujours disponibles 58
3.3.4. Est-ce le bon intervalle ? 58
3.3.5. Les abus d’interprétation : une situation paradoxale 59

3.4. LES MÉTHODES DE VRAISEMBLANCE 60

3.5. LES MÉTHODES BAYÉSIENNES 60
3.5.1. Un changement de l'objet de la probabilité 61
3.5.2. La question de l'objectivité 61
3.5.3. D'autres arguments à leur encontre 61
3.5.4. Un objet de rejet 62
3.5.5. Une disponibilité nouvelle 62
3.5.6. Les abus d'interprétation revisités à la lumière du cadre bayésien 62

3.6. LES AUTRES PROPOSITIONS 63
3.6.1. La répétition des expériences 63
3.6.2. La diminution des erreurs de mesure 64
3.6.3. La manipulation des variables 64
3.6.4. Les méta-analyses 64
3.6.5. L'analyse fiduciaire 64

CHAPITRE 4 QUELQUES OUVRAGES DE RÉFÉRENCE 66

4.1. J.-M. FAVERGE : MÉTHODES STATISTIQUES EN PSYCHOLOGIE APPLIQUÉE. (1950/1975) 66

4.2. S. SIEGEL : NONPARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES. (1956) 68

4.3. B. J. WINER : STATISTICAL PRINCIPLES IN EXPERIMENTAL DESIGN. (1962/1971) 70

4.4. W. L. HAYS : STATISTICS FOR THE SOCIAL SCIENCES. (1963/1973) 72

4.5. M. REUCHLIN : PRÉCIS DE STATISTIQUE. (1976) 74

4.6. R. E. KIRK : EXPERIMENTAL DESIGN: PROCEDURES FOR THE BEHAVIORAL SCIENCES. (1982) 76

3ÈME PARTIE APPROCHE DESCRIPTIVE 81

CHAPITRE 5 RÉANALYSES D'ARTICLES PUBLIÉS 85

5.1. RÉANALYSES ANTÉRIEURES 85
5.1.1. L'étude de Cohen (1962) 85
5.1.2. La réplique de Seldmeier et Gigerenzer (1989) 87
5.1.3. L'étude de Haase *et al.* (1982) 87
5.1.4. L'étude de Clark-Carter (1997) 88
5.1.5. L'étude de Freiman *et al.* (1978) 89
5.1.6. La réplique de Moher *et al.* (1994) 90

5.2. UNE RÉANALYSE FIDUCIO-BAYÉSIENNE 91
5.2.1. Méthode 91
5.2.2. Commentaires sur la présentation des tests 94
5.2.3. Résultats des réanalyses 95
5.2.3. Conclusion 106

CHAPITRE 6 EXPÉRIENCES AUPRÈS DES CHERCHEURS 108

6.1. QUESTIONNAIRES SUR L'INTERPRÉTATION DES TESTS 108

6.2. ÉTUDES EN SITUATION 109

6.3. EXPÉRIENCE 1 : "PSYCHOLOGUES ET STATISTICIENS" 111
6.3.1. Buts de l'expérience 111
6.3.2. Matériel 112
6.3.3. Consigne 112
6.3.4. Sujets 114
6.3.5. Passation 114
6.3.6. Résultats 114
6.3.7. Conclusion 118

6.4. L'ÉTUDE DE ROSENTHAL ET GAITO (1963) 119

6.5. EXPÉRIENCE 2 : PERCEPTION DES SEUILS OBSERVÉS 126
6.5.1. Méthode 127
6.5.2. Consigne "hypothèse alternative" : courbes moyennes 130
6.5.3. Consigne "hypothèse alternative" : identification de classes de sujets 131
6.5.4. Consigne "hypothèse alternative" : modèle "psychophysique" 136
6.5.5. Consigne "hypothèse nulle" 136
6.5.6. Discussion 138

CONCLUSION 143

BIBLIOGRAPHIE 149

ANNEXES 161

A. QUELQUES RAPPELS A-1

B. RÉANALYSE FIDUCIO-BAYÉSIENNE D'ARTICLES PUBLIÉS B-3

C. RÉSULTATS DE L'EXPÉRIENCE SUR LES SEUILS OBSERVÉS C-15

INTRODUCTION

"Ce n'est pas parce que l'erreur est répandue qu'elle devient vérité." Gandhi

INTRODUCTION

Si le chercheur scientifique accorde une part importante de son temps à la description des phénomènes, des données, il ne peut pour autant s'en tenir là et il lui faut généraliser ses résultats. À partir de faits particuliers il va donc procéder à une induction. Cette induction a pu s'opérer sur la base de l'intuition du chercheur, mais, dans un domaine où le souci d'objectivité est proclamé, on comprend que l'introduction par la statistique d'outils formels répondant à cette visée inductive n'ait pu que rencontrer un accueil favorable.

Ces outils inductifs sont connus comme les méthodes d'*inférence statistique*; en statistique, l'usage a établi que le terme inférence renvoie à l'inférence inductive. Ils se répartissent en deux classes : d'une part les méthodes de test, et d'autre part les méthodes d'estimation, ponctuelle et par intervalle.

Ce sont surtout les tests statistiques qui ont envahi un grand nombre de domaines scientifiques, sous la forme des tests de signification. Un "résultat significatif" à l'un des seuils fatidiques 0.05 ou 0.01 fait maintenant incontestablement partie des normes en vigueur dans la communauté scientifique, au point qu'il est devenu quasiment obligatoire dès lors qu'il s'agit de publier dans une revue expérimentale. Le test de signification sert pour le moins à renforcer la portée des arguments utilisés pour convaincre de l'intérêt des résultats présentés; plus encore, il apparaît souvent comme un label de "scientificité" d'une recherche (utilisé, par exemple, par la célèbre revue *Science*).

Un observateur extérieur pourrait donc croire que le test de signification est un outil parfaitement adapté à la méthodologie de la recherche expérimentale.

La thèse que nous défendons ici est qu'il n'en est rien et que l'apparente adaptation de la pratique du test de signification par les chercheurs est illusoire, ce qui aboutit à une situation de fait pour le moins étrange, voire paradoxale.

La recherche expérimentale peut en effet être rapprochée d'un jeu, d'un combat (Freeman, 1993, utilise l'adjectif "gladiatorial"), dans lequel seuls les résultats significatifs sont gagnants, alors que les résultats non significatifs sont en principe des constats d'ignorance, donc des échecs. Mais les règles de ce jeu sont inadaptées et par suite constamment transgressées, ce qui se traduit par d'innombrables abus d'interprétation et entraîne des distorsions considérables, notamment dans la conduite des expériences, dans la sélection des résultats publiés et dans la présentation de ceux-ci.

Une des conséquences est que le chercheur qui présente un résultat significatif, tel le vainqueur d'une épreuve sportive, fait souvent l'objet de suspicion et doit satisfaire à un contrôle avant que son résultat soit homologué (publié). C'est le rôle des éditeurs et rapporteurs des revues aux réserves desquels l'expérimentateur est souvent confronté. Malheureusement, la norme est si bien établie que ces réserves portent presque exclusivement sur la validité des tests (A-t-on utilisé le bon test ? Les conditions d'application sont-elles satisfaisantes ? *Etc.*) et non sur leur pertinence (Le test répond-il vraiment à la question posée ?).

Pour défendre notre thèse nous étudierons ici la pratique des tests statistiques chez les chercheurs en psychologie. Notre démarche visera à montrer l'inadaptation fondamentale de l'usage des tests de signification traditionnels; elle se structurera selon trois approches complémentaires.

L'approche normative

La discussion de l'inadaptation d'une pratique nécessite d'abord la connaissance de l'outil utilisé. La première partie sera donc consacrée à une étude critique de la norme statistique constituée par les théories sur lesquelles est fondé l'usage des tests de signification. Étant donné notre propos, ces théories seront jugées essentiellement sur leurs implications méthodologiques, et non selon un point de vue purement mathématique. Cette étude fournira la référence indispensable par rapport à laquelle il sera possible d'évaluer les distorsions constatées dans la pratique.

Se pose également la question de l'adaptation de l'outil à l'usage auquel on le destine. C'est d'abord l'examen des critiques méthodologiques dont fait l'objet l'usage des tests de signification qui permettra de répondre à cette question. En retour ces critiques éclaireront les propriétés des tests et l'apport réel de leur usage à la recherche expérimentale en psychologie. Toujours dans une perspective normative, nous présenterons également les erreurs et abus d'utilisation auxquels ils conduisent; nous verrons d'ailleurs que ces abus renvoient souvent aux critiques précédentes. Enfin nous nous demanderons comment on peut expliquer la persistance de l'usage des tests de signification, et leurs abus, en dépit de toutes les critiques formulées.

L'approche prescriptive

Si une pratique inadaptée peut venir de l'outil lui-même, elle peut aussi résulter de l'utilisation d'un mode d'emploi inadéquat. La deuxième partie abordera cette question. Nous nous situerons donc au niveau des prescriptions qui sont faites de l'usage des tests de signification; ces prescriptions consistent à la fois en une traduction des théories, en une adaptation à des problèmes particuliers pour un certain public (ici les psychologues), et en une illustration à partir d'applications dans le domaine concerné. L'approche prescriptive comportera donc l'étude de quelques manuels classiques de statistique appliquée s'adressant directement au psychologue et lui enseignant quoi appliquer et comment. La présentation des tests de signification devrait en principe y être une adaptation fidèle des théories; mais nous verrons, à l'examen de quelques uns de ces manuels, que cela n'est pas vraiment le cas et qu'il existe déjà à ce niveau des distorsions de l'outil statistique, qui sont encore aggravées dans la présentation des exemples d'applications.

Mais cet examen ne serait pas complet si nous n'abordions pas également la prescription de méthodes de rechange. La plupart des auteurs qui ont critiqué l'usage des tests de signification ont également prescrit des méthodes de rechange. La plupart d'entre elles renvoient à d'autres méthodes d'inférence, notamment l'intervalle de confiance et les procédures bayésiennes, dont l'étude normative dépasserait le cadre de ce travail. Nous nous contenterons de passer brièvement en revue les principales approches proposées et d'examiner leurs implications méthodologiques ainsi que les difficultés qu'elles soulèvent. Cet examen sera un élément important pour savoir si l'usage des tests peut véritablement être remis en cause par la prescription de procédures mieux appropriées et praticables, et pour expliquer quelle évolution on peut attendre des pratiques des chercheurs dans les années à venir.

L'approche descriptive

L'usage d'un outil est jugé sur les résultats qu'il donne. L'adaptation de la pratique des tests de signification peut être appréhendée, *in fine*, par l'examen des articles scientifiques que les chercheurs produisent. Nous compléterons la présentation de travaux antérieurs consistant en des réanalyses statistiques des résultats contenus dans des publications par celle d'une étude que nous avons effectuée, dans une perspective plus explicitement descriptive, en relation avec la recherche des abus d'interprétation des tests effectivement commis.

Enfin l'étude expérimentale de la démarche du chercheur dans des situations d'analyse statistique des données, à laquelle nous contribuons par deux expériences, permet d'une part de mieux comprendre les difficultés liées à l'usage des tests de signification et de cerner les conditions d'apparition de leurs abus d'utilisation, et d'autre part de préciser les attentes des utilisateurs eux-mêmes envers l'inférence statistique.

Plan de l'exposé

Chacune des parties comprend deux chapitres :

- 1ère partie : Approche normative
  - Chapitre 1 - Deux théories fréquentistes des tests statistiques
  - Chapitre 2 - Critiques et abus des tests
- 2ème partie : Approche prescriptive
  - Chapitre 3 - Les solutions de rechange préconisées
  - Chapitre 4 - Quelques ouvrages de référence
- 3ème partie : Approche descriptive
  - Chapitre 5 - Réanalyses d'articles publiés
  - Chapitre 6 - Expériences auprès des chercheurs

Néologismes

Nous emploierons souvent le terme "significativité" pour évoquer la propriété d'un résultat d'être "significatif" ou non et éviter ainsi le terme signification, qui même dans le contexte des tests, peut être ambigu, englobant parfois un début d'interprétation (par exemple, par signification d'un résultat on peut vouloir évoquer le fait que l'existence d'un effet est "démontrée", en cas de résultat significatif).

Nous utiliserons le terme de "mésusage" pour signifier un usage abusif (du test de signification).

Nous utiliserons également les néologismes courants suivants :

méthodologiste,
prescriptif,
réanalyse,
réplicabilité.

1ère PARTIE

APPROCHE NORMATIVE

Dans le chapitre 1 nous présenterons les deux principales théories fréquentistes sur lesquelles est fondé l'usage des tests de signification, celle de Fisher et celle de Neyman et Pearson. Il existe à propos de ces théories un nombre considérable d'idées fausses, dont peuvent être victimes mêmes les experts. C'est donc à partir de l'étude des textes fondateurs de Fisher et de Neyman et Pearson eux-mêmes que nous examinerons les caractéristiques de ces deux théories.

Dans le chapitre 2 nous passerons en revue et discuterons les critiques des tests. Ces critiques sont apparues très tôt, et de vives controverses ont constamment opposé Fisher à Neyman et Pearson. Dans le domaine de la psychologie, elles se sont surtout multipliées dans les années soixante.

L'exposé portera essentiellement sur les principes et nécessitera donc peu d'outils mathématiques. Nous fournissons cependant en annexe A quelques définitions rudimentaires sur les notions de distribution d'échantillonnage et de vraisemblance ainsi que sur le théorème de Bayes.

Afin de faciliter l'exposé, il nous paraît utile ici de préciser le statut de l'hypothèse statistique en jeu dans un test de signification et de rappeler brièvement les grandes interprétations de la probabilité qui sous-tendent les différentes théories de l'inférence statistique (on en trouvera un exposé plus détaillé dans Oakes, 1986).

L'hypothèse statistique et l'hypothèse de recherche

Ce qui est en jeu dans un test de signification, c'est une certaine *hypothèse statistique*, c'est-à-dire un énoncé concernant la loi de distribution d'une population statistique. Souvent il s'agit d'un énoncé relatif à la valeur d'un ou plusieurs paramètres (constantes inconnues) spécifiant cette distribution, par exemple que la moyenne d'une variable pour une population donnée est égale à telle valeur précise. Cette hypothèse statistique est dérivée de l'hypothèse de recherche à laquelle le chercheur s'intéresse en premier lieu, et qui concerne sa discipline, par des voies plus ou moins (im)pénétrables. Il est évident que cette dérivation est cruciale pour la validité d'une inférence portant sur l'hypothèse de recherche. Il est tout aussi évident que les théories des tests statistiques ne concernent que l'inférence portant sur l'hypothèse statistique et que la question de la pertinence de la dérivation qui y mène n'est pas du ressort de la statistique (du moins pas directement; il peut se faire que la théorie psychologique en jeu soit elle-même de nature statistique, mais ceci est un autre problème). Aussi, nous laisserons de côté cette question, qui n'est en rien spécifique de l'utilisation du test de signification et se pose quel que soit la méthode d'inférence utilisée.

Les grandes interprétations de la probabilité

Si la probabilité en tant qu'objet mathématique est clairement définie par un système d'axiomes, son interprétation est diverse et l'on distingue trois grandes conceptions.

La conception objectiviste ou fréquentiste
Dans cette conception qui remonte à Venn et Von Mises, la probabilité de réalisation d'un événement est définie comme la limite de sa fréquence d'apparition quand le nombre d'épreuves (répétitions) tend vers l'infini. De ce point de vue, la probabilité d'un événement singulier n'a pas de sens.

La conception logiciste
Dans ce cadre, la probabilité exprime le degré de vérité d'une proposition incertaine. La logique classique en est un cas particulier où toute proposition ne peut être que vrai ou fausse. C'est une conception normative : le calcul de la probabilité d'un énoncé ne dépend pas des opinions de celui qui effectue le calcul. Le premier exposé systématique de cette conception est dû à Keynes.

La conception subjectiviste
Elle est très proche de la conception précédente. Cette fois la probabilité est conçue comme une mesure du degré d'incertitude d'un individu *rationnel* à l'égard d'un énoncé. Par individu rationnel il faut entendre un

individu qui appliquerait rigoureusement les lois du calcul des probabilités. Cette conception n'est donc plus normative dans le sens où chaque individu est libre d'attribuer à un événement la probabilité qu'il souhaite. Simplement, deux individus qui attribuent une même probabilité à un certain événement et qui sont ensuite confrontés aux mêmes données le concernant doivent réviser leur probabilité initiale de la même manière. Clairement, un événement singulier peut recevoir une probabilité dans ce cadre. Cette conception n'exclut pas que le degré d'incertitude soit parfois déterminé à partir de fréquences, mais elle refuse la réduction à ce seul cas. Ramsey, de Finetti, Jeffreys, Savage ont marqué cette conception.

Pour qualifier une probabilité concernant le degré d'incertitude à l'égard d'un énoncé singulier, on rencontre parfois l'adjectif *épistémique*, selon la proposition faite par Shafer (1976; cité par Oakes, 1986, p. 97). Pour caractériser les probabilités associées aux résultats d'un processus répétitif (jets de dés, *etc.*), Shafer parle alors de probabilités *aléatoires*.

CHAPITRE 1

DEUX THÉORIES FRÉQUENTISTES DES TESTS STATISTIQUES

1.1. LA THÉORIE DE FISHER

Elle s'inscrit dans la lignée des travaux de Yule et de Karl Pearson. Fisher cristallise, en quelque sorte, les notions et les pratiques jusque là en vigueur, et les développe, tout en leur conférant le statut de méthode d'inférence incontournable. C'est en 1925 que paraît la première édition de *Statistical Methods for Research Workers*, puis en 1935 celle de *The Design of Experiments* qui connaissent un succès considérable (14 éditions pour le premier ouvrage, 8 pour le second).

- Pour Fisher, le test est un moyen d'apprendre à partir des données expérimentales (voir, par exemple, 1990b/1935, pp. 8, 25, 1990c/1956, p. 96). Dans cette perspective, il s'adresse principalement aux chercheurs scientifiques et leur propose une procédure qui vise explicitement à se prononcer sur une hypothèse statistique au vu de résultats expérimentaux et qui soit un gage d'objectivité :

> "Though recognizable as a psychological condition of reluctance, or resistance to the acceptance of a proposition, the feeling induced by a test of significance has an *objective basis* in that the probability statement on which it is based is a fact communicable to, and verifiable by, other rational minds." (Fisher, 1990c/1956, p. 46) (Italiques ajoutés.)

En fait il s'engage encore plus loin et vise le raisonnement inductif dans son ensemble quand il écrit que les progrès de la statistique permettent dorénavant de traiter l'induction aussi bien et aussi complètement que la déduction l'a été par les méthodes traditionnelles :

> "That such a process of induction existed and was possible to normal minds, has been understood for centuries; it is only with the recent development of statistical science that an analytic account can now be given, about as satisfying and complete, at least, as that given traditionally of the deductive process." (Fisher, 1955, p. 74)

- Les données sont regardées comme un échantillon supposé provenir aléatoirement d'une population purement hypothétique, infinie et inconnue :

> "Whereas, the only populations that can be referred to in a test of significance have no objective reality, being exclusively the product of the statistician's imagination through the hypotheses which he has decided to test, or usually indeed of some specific aspects of these hypotheses." (Fisher, 1990c/1956, p. 81)

On est donc loin de l'idée d'une population bien définie dans laquelle on pourrait à l'envi tirer au hasard des échantillons.

- Une seule hypothèse est mise à l'épreuve. Elle est appelée hypothèse nulle, dans le sens de *to be nullified*; c'est-à-dire à réfuter, et non nécessairement, comme on le trouve écrit dans certains manuels[1], d'une valeur de zéro pour le paramètre testé, même si c'est le cas le plus fréquent. En général l'hypothèse nulle sera la négation de l'hypothèse à laquelle le chercheur s'intéresse réellement :

> "… the hypothesis that the phenomenon to be demonstrated is in fact absent" (Fisher, 1990b/1935, p. 13)

Par exemple, si l'on souhaite montrer qu'il existe une différence non nulle entre les moyennes de deux populations, l'hypothèse nulle sera que cette différence est égale à zéro.

- Le résultat de la procédure de test est :
  - soit le rejet de l'hypothèse nulle,
  - soit la suspension du jugement : on ne rejette ni n'accepte l'hypothèse nulle.

- Il n'y a, par conséquent, qu'une seule possibilité d'erreur : rejeter l'hypothèse nulle alors qu'elle est vraie.

---

[1] Par exemple, dans Abdi (1987, p. 74) on trouve : "…l'effet est nul dans la Population, de là le terme d'Hypothèse Nulle."

- Les données de l'échantillon étant recueillies, on calcule pour la statistique de test (dont le choix repose sur l'intuition du statisticien), en supposant vraie l'hypothèse nulle, la probabilité d'obtenir un résultat au moins aussi extrême que celui observé. Cette probabilité, habituellement notée *p*, *conditionnelle à l'hypothèse nulle*, est le niveau (ou seuil) observé de signification (*significance level*). Elle est, selon Fisher, caractéristique des données observées et indicatrice du degré de réfutation de l'hypothèse nulle auquel ces données conduisent :

> "The actual value of *p* […] indicates the strength of the evidence against the hypothesis." (Fisher, 1990a/1925, p. 80)

Pour Fisher, ce seuil de signification, caractéristique d'une expérience *unique*, ne doit pas être confondu avec un taux d'erreur que l'on obtiendrait par échantillonnage répété dans *une même* population, même s'ils peuvent coïncider dans certains cas (voir, en particulier, 1955 et 1990c/1956, pp. 81-82).

Si *p* est jugé suffisamment faible, on rejette l'hypothèse nulle, on considère qu'on a réussi à en montrer la fausseté et le résultat est déclaré *significatif*. Pour juger de l'importance (ou plutôt de la petitesse) de *p*, Fisher fait souvent référence dans ses premiers écrits à un seuil de 0.05[2] (*cf.*, par exemple, 1990a/1925 p. 44, 1990b/1935, p. 13). Cependant il en vient, ultérieurement, à rejeter la notion d'un seuil (5% ou tout autre) absolu, intangible :

> "… for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of evidence and his ideas." (Fisher, 1990c/1956, p. 45)

Une des raisons de cette évolution est sans doute son opposition à la théorie de Neyman et Pearson (décrite plus loin en 1.2.) dans laquelle la notion d'un seuil fixé *a priori* est fondamentale.

Si *p* est trop élevé, on suspend le jugement. En effet, pour Fisher, l'hypothèse nulle ne peut être acceptée :

> "… and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist in order to give the facts a chance of disproving the null hypothesis." (Fisher, 1990b/1935, p. 16)

On voit bien là, dans l'idée de Fisher, le parallèle avec la logique. Pour démontrer la fausseté de la proposition "tous les chats sont gris", il suffit de trouver *un seul* contre-exemple; en revanche, en prouver la véracité nécessiterait d'examiner *tous* les chats, ce qui est impossible; on ne peut donc, sur la base d'une observation, que rejeter une telle proposition (si l'on observe que le chat n'est pas gris) ou rester dans le doute. On est proche des idées de Popper (1973/1939) sur le caractère *réfutable* des hypothèses et théories scientifiques. Sans doute faut-il voir là, d'ailleurs, une des raisons du succès des tests.

Une extension du raisonnement par l'absurde

Le raisonnement sous-jacent correspond à une extension probabiliste du raisonnement par l'absurde, où, pour démontrer la fausseté d'une proposition A, on procède ainsi :

- Supposons A vraie.
- Alors, et sans utiliser d'autres propriétés et théorèmes que ceux tenus pour vrais, on en déduit quelque chose de contradictoire sur une certaine proposition B, elle-même déjà connue par ailleurs (c'est-à-dire indépendamment de A) : que B est vraie alors qu'on la sait fausse, ou inversement.
- Si donc la véracité de A entraîne quelque chose d'impossible, c'est que A est nécessairement fausse (par application de *modus tollens*).

Cela devient, dans le cas du test de l'hypothèse nulle :

- Supposons vraie l'hypothèse nulle.
- Calculons, dans ce cas, la probabilité associée à un résultat au moins aussi extrême que celui observé. Si celle-ci est faible cela signifie qu'un événement rare, improbable, sous cette hypothèse, a été observé.
- Posons comme *principe* supplémentaire que *ce qui se produit effectivement (ce qui est observé) n'est pas rare.*
- Nous sommes alors en présence d'une contradiction, et nous sommes conduits à rejeter l'hypothèse nulle, la considérant comme fausse (par application de *modus tollens*).

[2] L'absence, à l'époque, des moyens de calculs que nous connaissons maintenant (calculatrices électroniques et surtout ordinateurs) a certainement joué un grand rôle dans l'uniformisation des seuils utilisés du fait qu'il fallait recourir à l'usage de tables et que celles-ci, très longues à établir, ne pouvaient que présenter un nombre limité d'entrées. Quant au choix précis du fameux 5%, Cowles et Davis (1982) ne le font remonter à Fisher que pour ce qui concerne un premier énoncé formel, et insistent sur l'antériorité de l'usage de ce seuil qui remonte, entre autres, à K. Pearson et Student.

Autrement dit, face à un résultat significatif, le chercheur doit choisir entre : admettre que l'hypothèse nulle est vraie et qu'un événement rare s'est produit, ou admettre que l'hypothèse nulle est fausse.

À titre anecdotique, on remarquera que si ce raisonnement, dans son principe, ne pose pas de problème en général (nous verrons plus loin qu'il n'en va pas de même dans son application), il est au moins un exemple où il est l'objet d'un contresens. Rogers *et al.* (1993), dans un article méthodologique, assurent en effet (p. 560) que "devant un *p* faible le chercheur *décide* que quelque chose d'inhabituel s'est produit *et* rejette l'hypothèse nulle". Mais si le chercheur opte pour l'inhabituel cela est tout à fait compatible avec l'hypothèse nulle, puisque précisément cette mesure de l'inhabituel est calculée sous cette hypothèse.

Il existe une variante de cette extension probabiliste du raisonnement par l'absurde, qu'on trouve présentée et dénoncée, par exemple, par Cohen (1994), et que Falk et Greenbaum (1995) qualifient de "version affaiblie". Cette variante se distingue par l'omission du principe supplémentaire et par le dernier point du raisonnement qui devient :

• Puisqu'un événement *improbable*, sous l'hypothèse nulle, s'est produit, c'est que l'hypothèse nulle est *probablement* fausse.

Clairement, ceci est le renversement des probabilités conditionnelles : on passe de *Pr*(*Données*|*Hypothèse*) à *Pr*(*Hypothèse*|*Données*). Mais ce passage n'est en rien justifié du point de vue logique (alors que dans la version précédente c'est, conjointement, l'adoption du principe supplémentaire et l'utilisation de *modus tollens* qui assurent la cohérence du raisonnement; ce qui a justement pour avantage de mettre en évidence la nécessité de ce principe supplémentaire dans le raisonnement sous-tendant le test de signification). Pour assurer la cohérence logique de cette variante du raisonnement, il faut donc poser comme postulat, ou principe, le renversement des probabilités conditionnelles, une sorte d'équivalent probabiliste de *modus tollens*. Mais alors il n'est plus besoin de faire référence au raisonnement par l'absurde, la proposition se réduisant au postulat lui-même! Cette variante n'est donc qu'illusoire.

1.2. LA THÉORIE DE NEYMAN ET PEARSON

C'est en 1928 et 1933 que paraissent leurs articles fondateurs. Bien que s'inspirant, au départ, des travaux de Fisher, leur approche participe davantage d'une théorie statistique de la décision et ils ne la présentent pas, au contraire de Fisher, comme le fondement d'une nouvelle logique inductive.

- Pour eux il s'agit bien moins de pouvoir conclure sur la véracité d'une hypothèse que d'adopter, à son égard, un comportement rationnel selon un critère de contrôle des erreurs de décision :

> "Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our *behaviour* with regard to them, in following which we insure that, *in the long run* of experience, we shall not be too often wrong." (Neyman et Pearson, 1933a, p. 291) (Italiques ajoutés.)
>
> Ou encore :
>
> "We have suggested that a statistical test may be regarded as a *rule of behaviour* to be applied *repeatedly* in our experience when faced with the *same set* of alternative hypotheses." (Neyman et Pearson, 1933b, p. 509) (Italiques ajoutés.)

C'est explicitement aux propriétés *à long terme* d'une telle procédure que Neyman et Pearson se sont intéressés, et non pas à ce qu'une expérience particulière peut permettre de conclure. En envisageant une répétition à l'infini et dans des circonstances identiques, la procédure doit être telle qu'on contrôle le nombre de décisions erronées. Bien qu'ils disent s'intéresser en premier lieu à la pratique scientifique (et E. Pearson y insiste encore en 1955), il n'est donc pas étonnant que beaucoup de leurs exemples traitent de cas relevant du contrôle de qualité (comme le problème de l'acceptation ou du rejet d'un lot de pièces manufacturées).

- La population parente, dont l'échantillon est censé être tiré au hasard, est envisagée comme l'ensemble des répétitions à l'infini de l'expérience réalisée.

- Comme chez Fisher il s'agit de mettre à l'épreuve *une* hypothèse particulière, qu'ils notent H_0.

Cependant cette hypothèse est envisagée d'un point de vue plutôt opposé à celui de Fisher. En effet, alors que chez ce dernier l'hypothèse effectivement testée est la négation de l'hypothèse d'intérêt, Neyman et Pearson précisent que H_0 est l'hypothèse à laquelle on s'intéresse *particulièrement* (1933a, p. 294), et qu'elle est souvent celle qui semble la plus probable *a priori* :

> "… in practice there are often strong *a priori* grounds for believing that this is the population sampled [c'est-à-dire que H_0 est vraie] …" (Neyman et Pearson, 1928a, p. 176).

Par exemple elle résulte d'une théorie assez bien établie qu'on ne voudrait pas remettre en cause sans de forts arguments. Ou bien c'est une “hypothèse nouvelle et importante qu'on ne souhaitera pas écarter à la légère” (1933b, p. 497). Dans un article en français de 1935, il arrive d'ailleurs à Neyman de parler de H_0 comme de l'hypothèse *à vérifier*, ce qui est un terme fort (même si, auparavant, il a pris soin de placer entre guillemets le verbe “accepter” à propos de cette hypothèse) et confirme le caractère privilégié de H_0.

Mais cette fois H_0 n'est plus suffisante pour construire le test. Il convient de prendre également en compte les hypothèses alternatives admissibles, notées H_i, et qu'on suppose toujours pouvoir être définies :

> "It is indeed obvious, upon a little consideration, that the mere fact that a particular sample may be expected to occur very rarely in sampling from [the population] would not in itself justify the rejection of the hypothesis that it has been so drawn, if there were no other more probable hypotheses conceivable." (Neyman et Pearson, 1928a, p. 178)

Autrement dit, il est vain de vouloir tester une hypothèse si on ne lui en reconnaît pas de concurrente.

L'ensemble des hypothèses doit réaliser une partition des valeurs admissibles; on admet qu'il n'existe pas d'autres possibilités que celles décrites par H_0 et les H_i. Le plus souvent une seule hypothèse alternative, H_1, est spécifiée. Cela peut correspondre au cas de Fisher, avec l'hypothèse nulle et sa négation, hypothèse complémentaire composite, mais il peut aussi s'agir de deux hypothèses ponctuelles.

- La procédure de test doit permettre de décider entre :
  - soit rejeter H_0, mais sans que cela soit au profit d'une hypothèse H_i particulière, sauf, bien sûr, s'il n'y a qu'une hypothèse alternative, auquel cas on l'accepte,
  - soit accepter H_0,
  - soit éventuellement rester dans le doute (cas où les données sont insuffisantes). Cette dernière possibilité est présentée comme un cas particulier de la deuxième, mais elle est rarement, pour ne pas dire jamais, illustrée par les auteurs qui, de ce fait, mettent l'accent sur la dichotomie acceptation/rejet de l'hypothèse H_0.

Il peut être utile de rappeler que “rejeter” ou “accepter” signifient pour eux “choisir telle ou telle action” et non pas croire ou non (ou plus ou moins) en l'hypothèse.

- Il y a donc deux possibilités d'erreurs, quand une décision est prise :
  - L'erreur dite de première espèce ou de type I consiste à rejeter H_0 alors que H_0 est vraie. La probabilité conditionnelle (à la véracité de H_0) correspondante est appelée risque de première espèce, et est notée α le plus souvent.
  - L'erreur dite de deuxième espèce ou de type II consiste à accepter H_0 alors qu'une hypothèse alternative H_i est vraie. Le risque de deuxième espèce est la probabilité conditionnelle (à la véracité de H_i) correspondante, souvent notée β_i (mais aussi parfois 1-β_i). Son complément 1-β_i (ou bien β_i), probabilité de choisir H_i alors que H_i est vraie, est appelée la *puissance* du test par rapport à H_i. Dans le cas où il existe plus d'une hypothèse alternative on définit la *puissance résultante* du test comme étant la probabilité, non conditionnelle cette fois, de correctement rejeter H_0. Mais cette puissance résultante fait intervenir, en plus des β_i propres à chaque H_i, les probabilités *a priori* de ces H_i et ne peut donc être déterminée, en général. Toujours dans ce cas, on parlera également de la *fonction* de puissance du test pour désigner la fonction qui associe la valeur de la puissance à la valeur du paramètre correspondant à chacune des hypothèses H_i. Cette fonction caractérise le test.

Toutes choses égales par ailleurs, les risques α et β varient en sens inverse : diminuer l'un fait augmenter l'autre. La puissance dépend du risque α, de la valeur du paramètre sous l'hypothèse alternative, et de la taille N de l'échantillon. En particulier, les autres termes étant fixés, la puissance augmente avec N.

- Le risque de première espèce α est choisi *a priori*. Selon Neyman et Pearson il existe une dissymétrie fondamentale entre les deux risques d'erreurs (1933b, p. 497). Rejeter une hypothèse H_0 vraie aura les mêmes conséquences quelles que soient les données (l'échantillon) ayant amené ce rejet, à savoir exclure cette hypothèse et se fermer une voie de recherche, mais sans savoir nécessairement vers laquelle se tourner (rappelons qu'il ne s'agit pas de discriminer entre les diverses H_i). Au contraire, accepter H_0 à tort sera de plus ou moins grande importance selon que l'hypothèse vraie sera plus ou moins éloignée de H_0. Cela équivaut à considérer que l'erreur de première espèce est, en général, la plus importante; ce qui correspond à leur point de vue, énoncé plus haut, que l'hypothèse à tester est celle à laquelle on s'intéresse particulièrement. En effet, on ne peut comprendre cette primauté accordée au risque de première espèce, dans le sens où son contrôle est des plus aisé (il suffit d'en fixer la valeur), que par l'importance de l'hypothèse correspondante. Bien sûr, ils reconnaissent qu'il faut, dans certains cas, moduler les choix en fonction du problème :

> "It is also evident that in certain cases we attempt to adjust the balance between the risks P_I and P_{II} to meet the type of problem before us." (Neyman et Pearson, 1933b, p. 497)

Quant au choix de la valeur précise à fixer pour α, les auteurs considèrent qu'il est du seul ressort du chercheur. Ils se réfèrent à Fisher pour les valeurs 0.05 et 0.01 qui sont, de loin, les plus courantes dans la pratique.

- La taille de l'échantillon est également fixée. En effet tous les résultats démontrés par les auteurs font appel à un espace à N dimensions où N est la taille de l'échantillon envisagé. La procédure ne conserve donc ses propriétés, sa validité, que dans la mesure où N est constant au regard de toutes les répétitions envisageables, donc fixé *a priori*. On notera que dans le cas d'une hypothèse alternative simple (ponctuelle), la procédure permet de déterminer l'effectif nécessaire une fois choisis les risques α et β, et la valeur du paramètre d'intérêt sous l'hypothèse alternative; ce que ne permet pas la théorie du test selon Fisher.

- Les hypothèses possibles, H_0 d'une part, les hypothèses alternatives H_i de l'autre, permettent de déterminer ce que l'on nomme la région critique, souvent notée ω : le sous-espace de l'espace des échantillons qui conduira à rejeter H_0 si l'échantillon observé s'y situe. Si, au contraire, l'échantillon appartient à une autre région ω', on acceptera H_0 (éventuellement s'il appartient à une troisième région ω'', incluse dans ω', on ne conclura pas). Ces régions sont calculées de façon que, α étant fixé, le risque de deuxième espèce soit minimum[3]. Signalons que la méthode de Neyman et Pearson permet également, en principe, de déterminer la statistique de test, ce qui n'est pas le cas chez Fisher.

- Au terme de l'expérience, on applique la règle précédemment définie et l'on conclut selon la région à laquelle l'échantillon (en pratique, la statistique de test) appartient. On remarquera que la région critique est construite tout à fait indépendamment des données; seules sont nécessaires les connaissances de α, N et des valeurs spécifiées par les hypothèses. Les données ne servent que pour conclure, et peu importe à quel point précis de l'espace des échantillons elles correspondent, la seule question étant de savoir si ce point appartient à ω ou non. Autrement dit encore, des échantillons différents, avec des seuils observés (p) différents, mais qui tous conduisent à un résultat significatif en fonction du risque α choisi, sont équivalents du point de vue de cette approche (d'ailleurs Neyman et Pearson définissent l'équivalence entre deux tests par l'égalité de leur "taille", c'est-à-dire de leur risque de première espèce).

Du point de vue méthodologique, il est utile et particulièrement éclairant de considérer comment Neyman présente le rôle des hypothèses et des risques dans son ouvrage de 1950 intitulé *First Course in Probability and Statistics*. Ce livre s'adresse à des étudiants qui n'ont besoin que d'une première approche des statistiques, aussi bien qu'à ceux qui pourraient ensuite se spécialiser en statistique mathématique, ou qui envisagent de s'engager dans un domaine utilisant les méthodes de la statistique. Il est donc moins mathématique que les articles cités précédemment, mais plus explicite quant à la méthodologie. D'autre part, écrit une quinzaine d'années après les articles fondamentaux des années 30, on peut supposer que l'auteur y exprime un point de vue mûrement pesé (Neyman considère lui-même que la théorie des tests présentée dans ce livre est élémentaire mais systématique, ainsi qu'il le mentionne dans une note au bas de la page 58 de son recueil de 1952).

Or, Neyman est très clair dans cet ouvrage et il apparaît que la prise en compte de l'importance relative des risques d'erreurs selon leur nature est le point fondamental qui doit guider la démarche du chercheur. Nous serions enclin à parler, à ce propos, du "Principe Méthodologique de Neyman". C'est le risque contre lequel il est le plus important de se prémunir (du point de vue de l'utilisateur) qui devra être posé comme étant, par définition, le risque de première espèce, et par conséquent l'hypothèse correspondante deviendra l'hypothèse testée :

> "Postulating this is to be the ordinary case we will use the expression *error of the first kind* to describe that particular error in testing hypotheses which is considered more important to avoid. The less important error will be called the *error of the second kind*. In the rare cases where the two kinds of error are of exactly the same importance, it is immaterial which of them is called error of the first kind and which the error of the second kind."[…]
> "This convention of labeling the two kinds of error is supplemented by a parallel convention concerning the use of the term *hypothesis tested. The term hypothesis tested is attached to H or to H'* [H et H' sont complémentaires] *in such a way that the rejection of the hypothesis tested when it is*

[3] Quand il existe plusieurs hypothèses alternatives, s'il est possible de trouver une même région critique telle que le risque β soit minimum pour toutes ces hypothèses alternatives, cette région est appelée la *meilleure* région critique et le test correspondant est dit *uniformément plus puissant* (par rapport à la classe des alternatives considérée). Mais Neyman reconnaît que de tels tests existent rarement (1935, p. 250). Par ailleurs, si la puissance du test est supérieure au risque α pour toutes les hypothèses alternatives, le test est dit *non biaisé*.

> *true is an error of the first kind*. It is usual to adjust the labels H and H' so that the hypothesis tested is labeled by H." (Neyman, 1950, pp. 263-264) (Les italiques sont de Neyman.)

L'évidence de ce principe n'a d'égal que son manque d'application dans la pratique, comme nous en reparlerons plus loin.

1.3. LE CADRE FRÉQUENTISTE

Bien qu'il soit habituel de classer les deux approches précédentes dans le cadre d'une conception *fréquentiste* de la probabilité, c'est-à-dire où celle-ci est exclusivement définie comme la limite d'une fréquence, ce qualificatif s'applique surtout à la théorie de Neyman et Pearson. À ce propos, Rouanet (1997) parle de "fréquentiste modéré" pour Fisher et de "fréquentistes radicaux" pour Neyman et Pearson. En effet, α et β prennent leur sens seulement quand on envisage que le chercheur réplique à l'infini son expérience (ce qui signifie dans des conditions *identiques* non seulement quant aux conditions expérimentales mais aussi quant aux règles de décision) : 100α % des fois le chercheur commettrait une erreur de type I si H_0 est vraie. D'ailleurs Neyman introduit explicitement cette conception fréquentiste de la probabilité dans ses ouvrages de 1950 et 1952.

Chez Fisher cette interprétation fréquentiste est implicitement évoquée quand il écrit :

> "Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion we should be led to follow up a false indication only *once in 22 trials*, even if the statistics were the only guide available. " (Fisher, 1990a/1925, p. 44) (Italiques ajoutés.)

ou encore :

> "In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will *rarely* fail to give us statistically significant results." (Fisher, 1990b/1935, p 14) (Italiques ajoutés.)

"Une fois sur 22 essais" et "rarement" renvoient à la notion de répétition, donc de fréquence. Cependant, comme nous l'avons déjà précisé lors de la présentation de sa théorie, Fisher s'est toujours violemment défendu d'une interprétation fréquentiste simple du seuil *p* (considérer ce seuil comme un taux d'erreur, requiert d'envisager, au moins dans certains cas, un échantillonnage répété non pas dans *une seule* population, mais dans *plusieurs*; *cf.*, par exemple, 1948, p. 201, où Fisher fait appel à une telle notion dans sa présentation du *t* de Student). Et plus tard, en 1959, il est particulièrement clair sur la signification qu'il accorde à la probabilité : c'est une mesure du degré d'incertitude, comme la conçoivent les tenants de l'approche subjectiviste[4].

> "The subject of a probability statement if we know what we are talking about, is singular and unique; we have some degree of uncertainty about its value, and it so happens that we can specify the exact nature and extent of our uncertainty by means of the concept of Mathematical Probability as developed by the great mathematicians of the 17th century Fermat, Pascal, Leibnitz, Bernoulli and their immediate followers." […]
>
> "The probability statements refer to the particular throw or to the particular result of shuffling the cards, on which the gambler lays his stake. The state of rational uncertainty in which he is placed may be *equated* to that of the different situation which can be imagined in which his throw is chosen at random out of an aggregate of throws, or of shufflings, which might equally well have occurred, though such aggregates exist only in imagination." (Fisher, 1959, p. 22)

Pour notre part, nous nous conformerons à l'usage le plus courant qui consiste à parler de *cadre fréquentiste* pour désigner les procédures traditionnelles de test, à la fois de Fisher et de Neyman-Pearson, et d'intervalle de confiance (de Neyman-Pearson). Cette appellation s'oppose au cadre bayésien et renvoie à la question de la probabilité des hypothèses dont nous examinons ci-après comment elle a été abordée par Fisher et Neyman et Pearson.

---

[4] Il n'est peut être pas inutile de prendre un exemple pour illustrer les différences de conception de la probabilité chez Neyman et Fisher. Considérons l'expérience qui consiste à tirer au hasard une carte dans un jeu de 52 cartes, à ne pas la regarder, et à se demander si cette carte est l'as de cœur. Si cette expérience est répétée indéfiniment, la fréquence de tirage de l'as de cœur tendra vers 1/52. Selon Neyman, c'est tout ce que peut signifier "la probabilité de tirer l'as de cœur vaut 1/52". Maintenant considérons un tirage particulier. Pour Neyman, il n'est pas pertinent de se demander quelle est la probabilité de tirer l'as de cœur, ou, plus précisément, cette probabilité ne peut être que 1 ou 0 (selon que c'est ou non l'as de cœur); alors que pour Fisher l'essence même de la probabilité est bien de caractériser l'incertitude sur *ce* tirage, qu'il égale à cette fréquence limite de 1/52.

1.4. LA QUESTION DE LA PROBABILITÉ DES HYPOTHÈSES

Pour Fisher

Pour Fisher, établir des probabilités sur les valeurs possibles du paramètre auquel on s'intéresse apparaît non seulement souhaitable (et souhaité par le chercheur), mais surtout comme la *seule* façon d'envisager la probabilité mathématique (voir, par exemple, l'article de 1959, où il rappelle que c'était aussi la conception des grands auteurs du XVII^{ème} siècle, en particulier de Gauss). Ce n'est que dans les cas, qu'il considère tout de même comme étant les plus fréquents, où l'on ne possède pas d'information fiable extérieure aux données permettant de fixer raisonnablement les probabilités *a priori* que Fisher rejette catégoriquement l'utilisation du théorème de Bayes comme outil pour obtenir les probabilités *a posteriori* (voir, par exemple, 1990b/1935, pp. 6-7; 1990c/1956, p. 40). Pour lui les probabilités *a priori* ne peuvent être fixées de manière arbitraire ("axiomatique" selon ses propres termes). Il rejette le choix d'une distribution uniforme comme distribution *a priori* pertinente dans ces cas où l'on ne possède pas d'autre information que les données. Pour lui une distribution, même uniforme, est en elle-même mathématiquement précise, et ne saurait être équivalente à une ignorance.

C'est dans cet esprit qu'il a développé la notion de probabilité fiduciaire (du latin *fiducia*, confiance), qui permet une estimation, sous forme d'intervalle, du paramètre considéré. Dans certains cas favorables où l'estimateur du paramètre possède la propriété d'exhaustivité (c'est-à-dire que l'estimateur exprime toute l'information sur le paramètre contenue dans les données), cela permet d'établir de tels jugements de probabilité sans faire appel aux probabilités *a priori*.

> "When knowledge *a priori* in the form of mathematically exact probability statements is available, the fiducial argument is not used, but that of Bayes. Usually exact knowledge is absent, and, when the experiment can be so designed that estimation can be exhaustive, similar probability statements *a posteriori* may be inferred by the fiducial argument." (Fisher, 1990b/1935 p. 198)

Alors que le test ne met en jeu que des probabilités *hypothétiques* (dans le sens où elles sont conditionnelles à une hypothèse), un énoncé fiduciaire est un jugement sur le monde *réel* (conditionnellement aux données observées). Et Fisher va même jusqu'à proposer la méthode fiduciaire comme la méthode ultime :

> "… for there is no other method ordinarily available for making correct statements of probability about the real world." (Fisher, 1990b/1935, pp. 198-199)

On trouvera un exposé technique de la méthode fiduciaire dans Fisher (1948) et une présentation rapide et claire dans Lépine et Rouanet (1975).

Par ailleurs, sur la fin de sa vie, Fisher se consacre à "réhabiliter" le théorème de Bayes en prônant ce qu'il considère comme étant la méthode de Bayes lui-même, c'est-à-dire une détermination expérimentale des probabilités *a priori*, de manière à éviter de les introduire de façon axiomatique (*cf.* Fisher, 1962).

Pour Neyman et Pearson

Pour Neyman et Pearson il semble de prime abord, c'est-à-dire à la lecture de leurs articles de 1928 et 1933, qu'il soit tout à fait naturel de probabiliser des hypothèses. Ainsi ils consacrent un long article (1933b) intitulé "The Testing of Statistical Hypotheses in Relation to Probabilities *a priori*" au rapport des probabilités *a priori* des hypothèses (c'est-à-dire avant le recueil des données) avec leur procédure de test. Par exemple, comme nous l'avons déjà mentionné, la notion de puissance résultante d'un test fait appel à ces probabilités. Mais comme ils estiment que ces probabilités sont inconnues en général, ils indiquent que leur motivation est de rechercher des procédures de test qui en soient indépendantes, en un certain sens qu'ils précisent. Cette motivation apparaît donc plutôt d'ordre pratique que d'ordre philosophique. Eux-mêmes déclarent :

> "It is evident at once not only that additional information beyond that supplied by the observations is necessary in order to define $C(H)$ [la classe des hypothèses admissibles], but also that the probabilities *a priori* ϕ_i of the admissible hypotheses H_i (i=0,1,2, …) falling within $C(H)$ must enter into the problem. Yet if it is important to take into account probabilities *a priori* in drawing a final inference from the observations, the practical statistician is nevertheless forced to recognize that the values of ϕ_i can only rarely be expressed in precise numerical form. It is therefore inevitable from the practical point of view that he should consider in what sense, if any, tests can be employed which are independent of probabilities *a priori*." (Neyman et Pearson, 1933b, p. 493)

En revanche, ils ne se préoccupent pas de la question des probabilités des hypothèses au regard des observations recueillies, probabilités *a posteriori* [5] (du moins pour ce qui concerne les tests), ce qui est cohérent

[5] Il est facile de démontrer que la probabilité *a posteriori* de H_0, étant donné un résultat significatif, est diminuée par rapport à ce qu'elle était *a priori*, dès lors que la puissance (puissance résultante dans le cas d'une hypothèse alternative composée) est supérieure au risque de première espèce.

avec leur approche de type décisionnel : c'est un comportement, une action, suite à une décision qui les intéresse et non une connaissance (probable) du monde.

Mais cette première impression doit être tempérée. En effet, dans son manuel de 1950, ainsi que dans le recueil de conférences paru en 1952 (une première édition, ronéotypée, date de 1938), Neyman indique très clairement que la seule conception de la probabilité qu'il admette est la conception fréquentiste. Par exemple :

> "For Fisher, probability appears as a measure of uncertainty applicable in certain cases but, regretfully, not in all cases. For me, it is solely the answer to the question "How frequently this or that happens."" (Neyman, 1952, p. 187)

Il apparaît alors que pour Neyman la probabilité sur le paramètre (l'hypothèse) ne saurait être envisagée que lorsqu'elle peut recevoir une interprétation fréquentiste, que lorsque le paramètre est vu lui-même comme une variable aléatoire; ce qu'il pense être très rarement le cas dans la pratique (comme exception il fournit l'exemple de la génétique mendélienne). Autrement, cette probabilité n'a plus de sens :

> "[…] Isn't this equivalent to discussing the probabilities of hypotheses themselves, which would be useless ? E.g., it would be useless to discuss the probability of Student's hypothesis because this would be the same as the probability of $\mu = 0$. As μ is an unknown constant, the probability of μ being equal to zero must be either $P\{\mu = 0\} = 0$ or $P\{\mu = 0\} = 1$ and, without obtaining precise information as to whether μ is equal to zero or not, it would be impossible to decide what is the value of $P\{\mu = 0\}$. […]
>
> Undoubtedly, μ is an unknown constant and, as far as we deal with the theory of probability as described in my first two lectures, it is useless to consider $P\{\mu = 0\}$".(Neyman, 1952, p. 56)

On peut se demander s'il s'agit là d'une évolution de la conception des auteurs ou simplement d'un changement dans la forme de la présentation (dans les articles de 1928 et 1933, très mathématiques, il n'est pas fait mention de la question de la nature des probabilités). Bien évidemment cela n'a aucune conséquence sur la théorie elle-même. Nous avons tendance à penser qu'il y a plus qu'un simple changement de forme car il nous semble peu crédible que les auteurs aient consacré un long article (1933b) à un problème qui, dans le cadre d'une conception purement fréquentiste, ne leur paraîtrait que très marginal (dans le sens où il ne se poserait pratiquement pas en pratique). Ce qui nous semble le plus vraisemblable c'est qu'au début Neyman ait été surtout intéressé par les aspects mathématiques de la théorie des tests puis, qu'avec le succès grandissant de cette théorie chez les chercheurs, il se soit de plus en plus intéressé aux problèmes posés par les applications, pour lesquels la question de l'interprétation de la probabilité est centrale. Quoiqu'il en soit, cette évolution est particulièrement bien illustrée par la manière dont est présentée la probabilité de commettre l'erreur de première espèce. En 1933 (1933b, p. 495, par exemple) on trouve que cette probabilité est égale à $\phi_0.P(\omega|H_0) = \phi_0.\alpha$ (où ϕ_0 est la probabilité *a priori* de l'hypothèse H_0); il s'agit de la probabilité *inconditionnelle* de commettre l'erreur, c'est-à-dire de la probabilité conjointe que H_0 soit vraie *et* qu'on la rejette. Alors qu'en 1952, au contraire, Neyman indique que cette probabilité d'erreur doit être comprise comme *conditionnelle* à la véracité de H_0 et se réduit donc à α (p. 57), toute considération de probabilité de l'hypothèse a disparue.

Tout de même, il est symptomatique de remarquer, comme Good (1984, p. 159) le rapporte, qu'il arrive à Neyman, dans le cas d'une application pratique, d'oublier ses principes fréquentistes et de succomber à la tendance (naturelle ?) à probabiliser les hypothèses (et, de surcroît, de commettre l'erreur d'égaler cette probabilité au seuil observé!) :

> "In these conditions [a *p*-value of 1/15], the odds of 14 to 1 that this loss was caused by seeding [of clouds] do not appear negligible to us." (Neyman *et al.*, 1969[6])

1.5. UN AMALGAME DES DEUX THÉORIES

Le plus souvent les chercheurs, et même les statisticiens, ne distinguent pas clairement les deux théories, mais utilisent divers amalgames. Par exemple, les chercheurs se basent le plus souvent sur le seuil observé *p* (Fisher), mais ils font aussi appel, occasionnellement, à la notion de risque de deuxième espèce et/ou de puissance (Neyman et Pearson). Ou bien les deux risques, de première et deuxième espèces, sont mis en avant (Neyman et Pearson), ou même la référence à Neyman et Pearson est explicite, mais le choix de H_0 se fait par négation de l'hypothèse d'intérêt, et sans considération aucune de l'importance relative des risques d'erreur, ce qui relève de l'approche fishérienne. Nous avancerons même l'opinion qu'il n'existe pas, dans la pratique, de "purs" neymaniens dans la mesure où nous n'avons jamais vu appliqué à des données réelles un test répondant au principe méthodologique de Neyman mentionné plus haut. Enfin, nous évoquerons encore ici les interprétations bayésiennes "sauvages" des résultats du test (*cf.* la section 2.2.) qui sont tout simplement hors du cadre des théories fréquentistes.

---

[6] Neyman, J., Scott, E. L., Smith, J. A. (1969) - Letter in *Science*, 165, p. 618, concernant une expérience d'ensemencement de nuages.

Ces amalgames sont d'ailleurs révélés par la terminologie et les notations en sont des exemples flagrants. Il est maintenant passé dans l'usage de parler “d'hypothèse nulle” (à la suite de Fisher) et de la noter “ H_0” (à la suite de Neyman et Pearson[7]), alors même que ces auteurs envisageaient très différemment, pour ne pas dire de façon opposée, le rôle de cette hypothèse. Le seuil observé (Fisher) est parfois qualifié de “risque” (Neyman et Pearson).

Ces amalgames ont été constatés, et souvent dénoncés, par de nombreux auteurs, par exemple par Morrison et Henkel (1970, p. 7). Gigerenzer parle à ce propos, en un sens péjoratif, de la “logique hybride de l'inférence statistique” (1993, p. 314). Mais cette appellation, même si l'on n'y voit qu'une tournure ironique, ne nous paraît pas satisfaisante. Hors de son contexte elle risque en effet d'accréditer l'idée qu'il existe effectivement une théorie hybride qui aurait sa propre logique et pourrait donc être justifiée formellement.

Même des auteurs qui entendent distinguer les théories de Fisher et de Neyman et Pearson paraissent victimes de ces amalgames. Ainsi Oakes (1986), s'il consacre un chapitre théorique aux grandes écoles de l'inférence statistique dans lequel les théories de Fisher et de Neyman et Pearson sont décrites, n'en mélange pas moins les idées de ces auteurs dans sa présentation générale du principe des tests en début d'ouvrage. Le choix de l'hypothèse testée est présenté à la manière de Fisher, et il est en même temps question de la possibilité d'accepter l'hypothèse nulle et de l'existence de deux risques d'erreurs.

Un autre exemple peut être trouvé dans l'introduction d'un article de Clark-Carter (1997). Le comble est que l'auteur y fait justement référence à la “théorie” hybride des tests avant de décrire succinctement chacune des deux théories. Bien qu'il soit précisé que pour Fisher l'hypothèse nulle ne peut être acceptée, la théorie de celui-ci s'y trouve décrite à travers les idées de Neyman et Pearson puisqu'il est question de deux types d'erreurs. Si l'auteur prend soin de préciser que le terme “erreur” n'était pas utilisé par Fisher, il laisse tout de même entendre que celui-ci prenait bien en compte ces deux cas. Le problème n'est pas de jouer sur les mots, on peut admettre qu'on court, en ne concluant pas, le risque de ne pas détecter un effet qui serait réel (Schwartz, 1984, p. 76, préfère évoquer à ce propos le risque d'un “manque à gagner” pour bien le différencier d'un risque d'erreur), mais bien de se rendre compte que la notion d'erreur de deuxième espèce n'est pas pertinente dans le cas de Fisher, dans le sens où elle n'intervient tout simplement pas dans la construction du test. Les idées de Neyman et Pearson sont également malmenées, bien qu'un de leurs articles soit cité, puisqu'il est sous-entendu que l'hypothèse alternative correspond à l'hypothèse de recherche, ce qui est à l'opposé de leur conception. Quant à la distinction jugement/décision, caractéristique des différences entre ces deux théories, elle est complètement passée sous silence.

Enfin, dernier exemple de ces confusions, Chow (1996), dans son ouvrage consacré à la défense du test de signification, distingue on ne peut plus clairement les théories de Fisher et de Neyman et Pearson en les mettant en parallèle dans son tableau 2.3 (page 21). Mais sa présentation comporte malencontreusement des inexactitudes, par exemple en affirmant que pour Neyman et Pearson la probabilité d’intérêt est “la probabilité inverse $p(H|D)$.” (Poitevineau et Lecoutre, 1997).

1.6. UN PEU DE TERMINOLOGIE

Pour désigner les tests statistiques de conception fréquentiste, on trouve dans la littérature les termes de :

- “test de signification” (*significance test* ou *test of significance*),
- “test de l'hypothèse nulle” (*null hypothesis testing, null hypothesis decision procedure*),
- “procédure de test de signification de l'hypothèse nulle” (*Null Hypothesis Significance Testing Procedure* ou NHSTP),
- “test d'hypothèses” (*test of hypotheses*),
- “test de décision” (*decision test*), (utilisé par Wald, cité par Hogben, 1957).

En général, les deux premiers termes sont plutôt utilisés pour désigner la théorie de Fisher (lui-même parle de *test of significance*), alors que “test d'hypothèses” ou “test de décision” s’emploient plutôt pour la théorie de Neyman et Pearson. Mais cette règle n'est que toute relative. Le troisième terme, qui semble s'imposer en psychologie à l'heure actuelle chez les méthodologistes, renvoie le plus souvent, de fait, à un amalgame des deux théories.

[7] Le terme “hypothèse nulle” est étranger à Neyman et Pearson. Neyman y fait simplement allusion dans une note au bas de la page 259 de son ouvrage de 1950, pour préciser que c'est un terme parfois utilisé pour désigner ce qu'il préfère appeler “l'hypothèse testée”, cette dénomination ayant, pour lui, l'avantage d'être plus descriptive.

Il faut encore signaler qu'en statistique mathématique c'est la théorie de Neyman et Pearson qui est "la théorie classique (i.e. fréquentiste) des tests statistiques", et qui s'oppose à ce titre à la théorie bayésienne.

Enfin, dans la suite, nous nous conformerons à l'usage et appellerons hypothèse nulle l'hypothèse testée, la notant H_0, quelque soit la théorie envisagée.

* * *

Les deux grandes théories fréquentistes des tests statistiques sont opposées dans leur conception, comme le résume bien l'opposition des notions de "raisonnement inductif", défendue par Fisher, et de "comportement inductif", défendue par Neyman et Pearson. Alors que pour Fisher le test est un élément, parmi d'autres, de jugement sur les résultats de l'expérience particulière menée par un chercheur, pour Neyman et Pearson le test est un guide pour une action et renvoie aux propriétés à long terme d'un comportement (pour être exact, il faut cependant remarquer que E. Pearson, en 1955, a déclaré que la notion de comportement inductif n'était pas de son fait; mais ceci ne change rien aux propriétés de la théorie dont il est coauteur).

Cette opposition, qui a été marquée par les vives controverses qui ont opposé Fisher à Neyman et Pearson, trouve ses racines dans les conceptions différentes de la probabilité qu'ont ces auteurs. Leurs théories diffèrent également par le statut de l'hypothèse testée, ainsi que par le rôle joué par l'hypothèse alternative à l'hypothèse testée et qui est crucial dans la théorie de Neyman et Pearson.

Dans les faits, c'est l'utilisation d'amalgames non fondés logiquement, auxquels peuvent même s'ajouter des interprétations bayésiennes, qui est dominante.

CHAPITRE 2

CRITIQUES ET ABUS DES TESTS

2.1. CRITIQUE DES TESTS

La très grande utilisation des tests statistiques dans le milieu scientifique ne doit pas masquer le fait qu'ils posent des problèmes qui sont essentiellement de deux ordres. D'une part ils ont été, et sont toujours, sévèrement critiqués, à la fois sur un plan théorique et sur un plan méthodologique; d'autre part, ils donnent lieu, dans leur utilisation, à des mésusages, des abus, des erreurs d'interprétation.

Puisque c'est à la *pratique* des chercheurs que nous nous intéressons, il nous faut préciser que nous ne considérons ici que les tests *usuels*. Il s'agit typiquement du test qu'un effet est égal à zéro, au moyen du *t* de Student, du *F* de l'analyse de variance, ou du test de l'indépendance ou de l'homogénéité de deux variables au moyen du χ^2. Des aménagements *ad hoc* ont pu être proposés pour répondre à certaines critiques; ainsi Hodges et Lehmann ont proposé dès 1954 une procédure pour tester un intervalle au lieu d'une hypothèse ponctuelle. Mais ils ne sont guère utilisés ou même connus.

Les critiques sont apparues très tôt (voir par exemple, Berkson, 1938, 1941, 1942, et même dès 1931 avec Tyler, cité par Carver, 1978), et de vives controverses ont constamment opposé Fisher à Neyman et Pearson. Dans le domaine de la psychologie et de la sociologie, elles se sont surtout multipliées dans les années soixante. Certaines s'appliquent seulement à l'une des approches, mais la plupart s'appliquent aux deux.

Nous allons maintenant passer en revue et discuter ces critiques.

2.1.1. Quelle population ?

Le problème de définir à quelle population on veut étendre les résultats observés est peut-être le premier à considérer, dans le sens où il est fondamental. Cependant il n'est que marginal pour notre propos dans la mesure où il n'est en rien spécifique de la méthodes des tests et se pose pour toute inférence, qu'elle soit formalisée ou non (comme quand on généralise “sauvagement” les conclusions d'une analyse descriptive).

L'idée, selon Fisher, d'une population hypothétique infinie à laquelle on généralise les résultats d'une expérience unique, par opposition à une population clairement définie dans laquelle on pourrait facilement (du moins en principe) échantillonner au hasard de façon répétée, a parfois été remise en cause (voir Hogben, 1957; Camilleri, 1962; Morrison et Henkel, 1969) en raison, entre autres, de son caractère bien peu opérationnel. Mais la pertinence de cette idée a aussi été reconnue (par exemple par Hagood, 1941). Par ailleurs on comprend qu'une telle conception qui laisse au chercheur une grande part de liberté, de choix, et donc de responsabilité dans la formulation de son inférence, puisse par là même apparaître inconfortable.

Il faut bien voir que la conception de Neyman et Pearson est exposée au même problème, dans la mesure où la notion de répétition à l'identique peut être illusoire comme on va le voir ci-dessous.

2.1.2. *Bis repetita*

La théorie de Neyman et Pearson conçoit la procédure de test sur le long terme, comme ces auteurs l'on bien précisé. Donner un sens fréquentiste à la probabilité α requiert de considérer une répétition à l'infini de tests *identiques* (les hypothèses, α, *N* et la région critique doivent rester inchangés), ce qui implique que le résultat d'un test particulier, c'est-à-dire la décision prise, ne doit pas jouer sur la construction des autres tests, ne doit pas modifier les expériences futures (dans le cadre d'un même problème, bien sûr). Dans le domaine du contrôle de qualité, auquel se réfèrent souvent Neyman et Pearson, on peut facilement imaginer la situation suivante. Dans une usine une machine produit un très grand nombre de pièces. Chaque jour un échantillon de pièces est prélevé, examiné, et l'on décide sur la base d'un test statistique d'accepter la production journalière ou de la rejeter et de réparer la machine. Réparer la machine est bien sûr coûteux pour l'entreprise (coût de la réparation auquel s'ajoute celui relatif à la perte de production), de même que la livraison de lots défectueux (entraînant remboursement ou remplacement). La même procédure est répétée chaque jour, et peu importe finalement, du

point de vue de l'entreprise, qu'un jour particulier la décision soit erronée ou correcte; ce qui compte c'est qu'à la longue (sur un an, dix ans,…) les risques soient contrôlés. On est bien loin de la situation dans la recherche scientifique où, même si le critère de reproductibilité est fondamental, une expérience n'est pratiquement jamais répétée à l'identique un très grand nombre de fois (surtout par un même chercheur). Pour reprendre les termes de Camilleri (1962) :

> "[this] is patently absurd and not in fact what scientists do. They do not test the same hypothesis over and over again." (Camilleri, 1962)

D'autant que les résultats d'une expérience scientifique entraînent, en général, une modification, même minime, des connaissances; d'où également une modification des expériences ultérieures. Fisher lui-même a insisté sur le fait que le chercheur s'intéresse à l'expérience particulière à laquelle il est confronté, et non à une quelconque collection de répétitions virtuelles (voir, par exemple, Fisher, 1955, p. 74, 1990c/1956, pp. 103-104).

Même en se plaçant sur le terrain (apparemment) favorable de l'exemple précédent des lots de pièces manufacturées, un problème se pose. Si le test réalisé se révèle non significatif, l'hypothèse nulle correspondant au bon fonctionnement de la machine sera acceptée (même provisoirement), rien ne sera modifié et le processus pourra se perpétuer. Mais si l'on observe un résultat significatif, il est alors évident que cela va entraîner des modifications, à savoir la mise au rebut du lot produit et la réparation de la machine (il serait évidemment absurde de laisser la machine en l'état : si le test est réalisé c'est bien pour détecter ce cas). Or cette modification de la machine signifie, du point de vue du modèle statistique, un changement de la valeur du paramètre considéré. Autrement dit, à partir de ce moment, la population a changé, et tous les échantillons tirés ensuite seront issus d'une nouvelle population. La conception de répétitions d'échantillonnages dans une *même* population est donc ici irréaliste. Elle n'est vraisemblable que dans la mesure où le paramètre testé est indépendant de la décision prise à la suite du test, ce qui en restreint considérablement la portée. (Pour trouver un exemple d'un tel cas, il suffit de reprendre l'exemple précédent mais en envisageant les choses du point de vue du *client*. Celui-ci réalise un test sur un échantillon prélevé dans le lot qu'il reçoit, et, selon le résultat, il accepte ou rejette le lot, mais il n'intervient pas dans le processus de fabrication.)

Barnard (1947) et Fisher (1955) font remarquer que l'espace des échantillons utilisé pour construire le test ne correspond pas forcément à une répétition infinie. Par exemple la répétition de l'expérience peut entraîner la variation des tailles des échantillons : Barnard prend l'exemple de graines de fleurs où, bien que le nombre total de graines semées d'une répétition à l'autre soit constant, le nombre de fleurs examinées varie parce que certaines graines n'arrivent pas à germination. En fait, c'est le problème général des données manquantes qu'il pose au travers de cet exemple. Ou encore, le test usuel de la pente d'une droite de régression est conditionnel aux échantillons ayant la même variance que celle de l'échantillon analysé, alors que des tirages répétés de N couples de valeurs dans une même population produiraient des échantillons de variances différentes. Il en résulte que la détermination de l'espace des échantillons approprié à une répétition réelle n'est pas aussi simple qu'il peut paraître.

Dans le même ordre d'idées, Cox présente en 1958 un exemple souvent utilisé depuis et qui renvoie au *principe de conditionnement* (si plusieurs expériences sont possibles, l'inférence ne doit dépendre que de l'expérience effectivement réalisée) :

On s'intéresse à la moyenne θ d'une population normalement distribuée. On choisit d'échantillonner une valeur, l'échantillonnage pouvant se faire dans deux populations de même moyenne θ et d'écart-type connu, σ_1 pour la première, σ_2 pour la seconde, avec $\sigma_1 \ll \sigma_2$ (par exemple $\sigma_1 = 10\sigma_2$). Le choix de la population se fait par tirage au hasard avec une probabilité de ½ pour chacune d'elles, le résultat de ce tirage étant connu de l'expérimentateur. Après Cox, on a souvent présenté cet argument en prenant pour exemple le cas d'un chercheur qui peut effectuer une mesure avec un appareil précis mais fragile ou bien avec un appareil moins précis mais robuste, le choix se faisant au hasard, par exemple parce que l'appareil précis tombe aléatoirement en panne un jour sur deux en moyenne. Pour tester avec une puissance maximum $\theta = 0$ contre $\theta = \theta_1$ ($\cong \sigma_1$) au seuil $\alpha = 0.05$, on a le choix entre :

[a] un test *conditionnel* à la population sélectionnée, pour ce qui regarde le contrôle des risques (c'est-à-dire qu'on pose $\alpha = 0.05$ pour l'expérience *présente*), qui amène à rejeter l'hypothèse nulle pour une valeur observée supérieure à $1.64\sigma_1$ ou à $1.64\sigma_2$, selon la population concernée,

[b] un test *inconditionnel*, amenant à rejeter l'hypothèse nulle pour une valeur observée supérieure à $1.28\sigma_1$ ou à $5\sigma_2$, selon la population concernée (pour l'expérience présente on a respectivement $\alpha = 0.10$ ou $\alpha \cong 0.00$, et en moyenne, pour toutes les expériences possibles, $\alpha = 0.05$).

Le test en [b] assure que l'on a bien une puissance maximum sur l'*ensemble* des répétitions, mais il ne paraîtra sans doute pas le meilleur au chercheur qui s'intéresse à ce qu'il peut tirer des données qu'il a effectivement recueillies. La transposition du problème en termes d'intervalle de confiance, comme le fait Ch. Robert, 1992, est encore plus parlante. La position en [a] amène à construire un intervalle étroit ou large selon la population choisie, alors que la position en [b] amène à un intervalle de largeur “moyenne”. Dans les termes de

l'exemple évoqué, cela correspond à un chercheur qui, soit adapte la précision de son estimation à l'instrument qu'il a utilisé (obtenant tantôt un intervalle précis, tantôt un intervalle large), soit utilise systématiquement une estimation de précision "moyenne".

Ce principe de conditionnement illustre parfaitement la différence d'approche entre Fisher (et les bayésiens) d'une part, et les fréquentistes comme Neyman d'autre part. Pour un fishérien, il s'agit de tirer le maximum d'information de l'expérience présente et il choisira le test (ou l'intervalle) présentant, pour celle-ci, la plus grande sensibilité, c'est-à-dire le test conditionnel. Au contraire, un neymanien choisira le test non conditionnel car il vise à maximiser la puissance sur le long terme (sur un grand nombre d'expériences de même type), peu lui important un résultat particulier. On peut se demander, pour reprendre l'exemple précédent, s'il y aurait beaucoup de chercheurs, même parmi ceux se réclamant de Neyman et Pearson, qui, ayant disposé de l'appareil le plus précis, calculeraient un autre intervalle que le plus étroit (la réponse nous semble évidente).

2.1.3. De l'arbitraire…

Dans la théorie de Fisher, puisqu'il n'existe pas d'erreur de type II, donc pas de risque β à minimiser, aucune justification formelle n'existe quant au choix de la région de rejet de l'hypothèse nulle. Dès lors, comme le remarque Rozeboom (1960), on pourrait tout aussi bien la choisir au centre de la distribution d'échantillonnage. La pratique qui consiste à choisir comme région de rejet l'extrémité de la distribution d'échantillonnage, bien que raisonnable, est donc arbitraire.

La théorie de Neyman et Pearson, qui justifie le choix de la région critique en fonction des hypothèses alternatives, pourrait être vue comme une réponse à cette critique, mais des auteurs comme Lindley (1957) ou Salsburg (1994) ont fait remarquer que le choix de fixer α (et de minimiser β) est également arbitraire, d'autres approches étant possibles (par exemple la minimisation d'une combinaison linéaire des deux risques). Salsburg considère qu'il s'est agi d'une option purement mathématique, sans rapport aucun avec la nature de la recherche scientifique. Mais en fait il s'est agi d'un choix méthodologique dans la mesure où l'un des deux risques est, par nature, pratiquement toujours plus important, que l'autre. Par ailleurs, selon le principe méthodologique de Neyman (1950) de choisir l'hypothèse nulle de façon à ce qu'elle corresponde au risque jugé le plus grave (*cf.* la section 1.2.), ce choix pourra varier avec des utilisateurs qui n'auraient pas les mêmes intérêts, comme Neyman lui-même le reconnaît volontiers, et il est donc éminemment subjectif. Ainsi il reprend l'exemple fameux donné par Fisher (1935) d'une femme qui affirme pouvoir déterminer, simplement en goûtant une tasse de thé au lait, si le lait a été versé avant ou après le thé. Si c'est un jury chargé de se prononcer sur cette affirmation qui réalise le test, l'hypothèse nulle sera plutôt que les jugements sont émis au hasard, donc que la proportion de succès est de ½ (en considérant que pour ce jury le cas le plus grave est celui où on admettrait l'existence de cette faculté de discrimination alors qu'il n'en est rien). Au contraire, si c'est la femme qui effectue le test, le plus grave pour elle est de rejeter à tort son affirmation et son hypothèse nulle sera une proportion de succès supérieure à ½ (par exemple 2/3).

Dans les deux théories le choix de la valeur du seuil de référence, qui marque la frontière entre significatif et non significatif, est aussi éminemment arbitraire ou subjectif et introduire des fonctions de coûts des erreurs ne fait que déplacer le problème (Rozeboom, 1960; Camilleri, 1962; Winer, 1971, p. 14; Skipper *et al.*, 1967, par exemple). Cela est encore plus sensible dans la théorie de Neyman et Pearson puisque les conclusions pourront être diamétralement opposées selon que l'on se trouve en deçà ou au delà du seuil de référence (à partir de mêmes données, deux chercheurs pourraient prendre des décisions différentes sur la base de risques α différents); alors que dans le cadre fishérien, dans la mesure où un résultat non significatif n'amène qu'à suspendre le jugement, il n'y aura pas de véritable contradiction.

Et encore, dans un même cadre théorique, pour un même seuil et à partir des mêmes données, deux chercheurs pourront ou non obtenir un résultat significatif selon qu'ils choisissent de mettre en œuvre un test bilatéral ou unilatéral (*cf.* la polémique à propos de la légitimité des test unilatéraux que nous évoquons dans la sous-section 2.3.2.).

Enfin Stevens (1968) dénonce "l'illusion d'objectivité" des tests dès lors que des tests tels le t ou le F sont utilisés pour des variables correspondant seulement à une échelle ordinale.

2.1.4. Critique du "postulat α"

Chez Fisher, il existe le postulat implicite que des hypothèses qui sont rejetées avec un même seuil observé p sont considérées comme présentant un même degré d'évidence contre elles; ce que Cornfield (1966) appelle le "postulat α" (de façon plutôt malheureuse, d'ailleurs; il vaudrait mieux parler de "postulat p"). Mais Lindley (1957) a présenté le paradoxe suivant. Dans le cadre du test d'une hypothèse simple H_0, il est toujours

possible de trouver une taille d'échantillon N telle que H_0 soit rejetée à un seuil de 5%, par exemple, et, simultanément, d'obtenir une probabilité *a posteriori* pour H_0 de 95%, alors même que la probabilité *a priori* de H_0 est aussi faible qu'on veut (mais non nulle). Cela signifie que le postulat ne peut être accepté : le seuil p, à lui seul, est insuffisant pour "mesurer" le degré de désaccord avec l'hypothèse testée. En effet, tel que le dit Lindley :

> "[…] the degree of conviction is not even approximately the same in two situations with equal significance levels. 5% in to-day's small sample does not mean the same as 5% in to-morrow's large one." (Lindley, 1957, p. 189)

Pour Lindley, comme pour Cornfield, le rapport de vraisemblance se révèle beaucoup plus approprié pour mesurer l'importance de l'évidence (du moins dans le cas où l'on teste deux hypothèses ponctuelles).

Cornfield a étendu cette critique à l'approche de Neyman et Pearson (et sans doute est-ce pour cela qu'il parle de "postulat α"). Mais à tort, car Neyman et Pearson, intéressés seulement par une *règle de comportement* et non par *une* décision particulière, n'ont jamais eu besoin de ce postulat. Et quand ils mentionnent que deux tests sont équivalents lorsque la probabilité de commettre une erreur de type I est la même (ce à quoi renvoie Cornfield), il s'agit seulement d'une *définition* qui ne préjuge en rien de la confiance à accorder à une hypothèse mais qui caractérise les procédures de test. La non pertinence de ce postulat, dans ce cas, est particulièrement claire quand on sait que les tests peuvent être construits, dans la théorie de Neyman et Pearson, indépendamment de tout résultat observé.

Nous reviendrons sur ce problème du lien entre p et N dans la section 6.4. à propos de l'expérience de Rosenthal et Gaito (1963), pour laquelle il est fondamental.

2.1.5. Et le principe de vraisemblance ?

Il a été reproché aux tests usuels de ne pas toujours respecter le principe de vraisemblance. Selon ce principe, émis la première fois par Barnard (1947), puis par Fisher (1955), deux résultats qui ont même vraisemblance doivent amener aux mêmes conclusions. En particulier, les procédures séquentielles ont été critiquées de ce point de vue (*cf.* Lindley, 1957; Cornfield, 1966). Mais cela se produit aussi dans d'autres cas. Ainsi, comme l'illustrent Lindley et Phillips (1976), pour tester une certaine hypothèse sur la fréquence parente d'un variable dichotomique, on peut choisir entre deux règles d'arrêt : soit décider à l'avance d'observer N réalisations indépendantes et compter le nombre de "succès" qui sera la variable aléatoire, soit observer jusqu'à recueillir un nombre de "succès" prédéfini, le nombre de réalisations N étant cette fois la variable aléatoire. Il se peut très bien que deux chercheurs, opérant avec des règles différentes, observent un même N et un même nombre de "succès" mais aboutissent à des conclusions différentes, rejeter l'hypothèse testée / ne pas la rejeter, alors que la vraisemblance de l'hypothèse est identique dans les deux cas. Edwards *et al.* (1963) considèrent ainsi que les tests traditionnels sont moins objectifs que les procédures bayésiennes (qui ne dépendent pas de la règle d'arrêt), tandis que Bernard (1996) estime que ces deux approches sont aussi arbitraires l'une que l'autre.

Cornfield (1966) a montré que si Neyman et Pearson avaient choisi de minimiser une combinaison linéaire des deux risques ($\lambda\alpha+\beta$, λ étant le coût relatif de l'erreur de première espèce par rapport à celle de seconde espèce) au lieu de minimiser β pour α fixé, les tests ainsi construits auraient respecté le principe de vraisemblance.

2.1.6. Être ou ne pas être… observé

La définition du seuil observé p, ou du risque α, repose sur la prise en compte d'événements qui ne se sont *pas* produits puisqu'on calcule la probabilité d'observer un événement *au moins* aussi extrême que celui observé. Baser une inférence sur quelque chose qui ne s'est pas produit peut paraître pour le moins étrange, sinon non pertinent. Cette critique, qui concerne aussi bien la théorie de Fisher que celle de Neyman et Pearson, a surtout été mise en avant par des tenants de l'approche bayésienne (voir, par exemple, Lindley, 1957; Lindley et Philipps, 1976), dans laquelle le conditionnement de la probabilité *a posteriori* ne fait intervenir que les données effectivement observées. En particulier, Jeffreys (1961) l'a énoncée sous forme de paradoxe :

> "*What the use of P implies, therefore, is that a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred.*" (Jeffreys, 1961, p. 385) (Les italiques sont de Jeffreys.)

Sur cette base et en fonction du fait que la fonction de vraisemblance de l'hypothèse nulle est croissante selon N pour un seuil fixé, Lindley (1957) remet en cause le fondement même du raisonnement du test : il rejette l'alternative "ou bien un événement qui avait peu de chances de se produire s'est tout de même produit, ou bien

l'hypothèse est fausse". Car la "chance" de l'événement observé est pour lui mesurée par la fonction de vraisemblance, alors que celle en jeu dans le test concerne l'événement observé *et* les événements plus extrêmes.

2.1.7. Chassez le naturel…

Le raisonnement manque de naturel, il est contre-intuitif. On calcule une probabilité correspondant à l'événement observé, *conditionnellement à une hypothèse* sur le paramètre d'intérêt (*Pr*(*Données*|*Hypothèse*)), alors qu'il semblerait plus naturel, à l'inverse, comme on le fait dans le cadre bayésien, de calculer une probabilité sur les valeurs possibles du paramètre, *conditionnellement à l'événement observé* (*Pr*(*Hypothèse*|*Données*)). Cette dernière probabilité est bien celle à laquelle s'intéresse le chercheur, comme l'exprime clairement Carver, par exemple :

> "It [statistical significance testing] still gives us an estimate of $p(D'|H_0)$, when what we want is $p(H_0|D')$, $p(R|D')$ and $p(H_1|D')$." (Carver, 1978, p. 392).

Les abus d'interprétation que commettent les chercheurs, et que nous verrons dans la section 2.2., en sont d'ailleurs une parfaite illustration. À ce propos également, il est intéressant de signaler l'étude menée auprès de chercheurs en psychologie par M.-P. Lecoutre (1983, 1991). Parmi les 23 chercheurs interrogés, dont la tâche était de conclure statistiquement à propos d'un jeu de données, un tiers, environ, ont déclaré souhaiter voir se développer l'usage de méthodes inférentielles qui constituent un prolongement des jugements naturels.

Ce manque de naturel est une des raisons invoquées par Albert (1995) pour préconiser une approche bayésienne, plutôt que fréquentiste, dans le cadre de l'enseignement des notions de statistique inférentielle.

C'est encore ce passage, illicite mais si naturel pour l'utilisateur, de l'une à l'autre de ces deux probabilités conditionnelles (de *Pr*(*D*|*H*), celle utilisée dans le test, à *Pr*(*H*|*D*), celle à laquelle s'intéresse le chercheur) qui est au cœur de la récente critique des tests de signification par Falk et Greenbaum (1995). Cependant leur argument central est quelque peu maladroit, à double titre. En effet, Falk et Greenbaum prennent pour exemple un cas où $Pr(H_0|Données) = 0.82$, alors que $p = Pr(Données|H_0) = 0.005$, et concluent donc à l'absurdité du rejet, par le test, de H_0. Mais cette forte probabilité provient d'un probabilité initiale élevée pour H_0 ($Pr(H_0) = 0.9989$), et si l'on ne prend pas en compte la valeur même de la probabilité initiale $Pr(H_0)$ pour seulement s'intéresser à son évolution, on constate que les données ont pour effet de la *diminuer*, signe que ces données sont plutôt en *contradiction* avec l'hypothèse. Et surtout, dans un tel cas où la probabilité initiale de l'hypothèse est connue, Fisher, comme Neyman et Pearson (du moins, pour ces derniers, si la probabilité initiale peut recevoir une interprétation fréquentiste), préconisent d'utiliser le théorème de Bayes, et non un test de signification (*cf.* 1.4.). Enfin, remarquons que l'argumentation de Falk et Greenbaum tend à faire penser que l'on rejette trop souvent l'hypothèse nulle (H_0), or, ainsi qu'on va le voir un peu plus loin, une des critiques quasi unanime des tests est que cette hypothèse est (pratiquement) toujours fausse.

2.1.8. Décision ou jugement

Dans la recherche scientifique, le test d'une hypothèse ne correspond pas à un processus de décision où il faut choisir, de façon irréversible, entre deux actions, au contraire du problème, pour un client ou un producteur, de l'acceptation ou du rejet d'un lot de marchandises. Une hypothèse scientifique n'est jamais définitivement acceptée; elle peut toujours être remise en cause en fonction de découvertes ultérieures. De même un résultat significatif n'implique pas une confiance absolue, totale, de la part du chercheur, dans l'hypothèse alternative; ce n'est qu'un élément d'information supplémentaire qui aide le chercheur dans son jugement sur l'hypothèse en question. C'est ce que Fisher, le premier, a toujours violemment objecté à Neyman et Pearson qu'il considérait, d'ailleurs, comme n'étant pas familiers de la pratique et des problèmes réels des chercheurs :

> "… for the tests of significance are used as an aid to judgement, and should not be confused with automatic acceptance tests, or "decision functions"." (Fisher, 1990a/1925, p. 128; voir aussi 1990b/1935, pp. 25-26; 1955; 1990c/1956, chapitre IV)

Cet argument a souvent été repris; par exemple par Rozeboom (1960) :

> "The null-hypothesis significance test treats acceptance or rejection of a hypothesis as though these were *decisions* one makes on the basis of the experimental data — i.e., that we elect to adopt one belief, rather than another, as a result of an experimental outcome. *But the primary aim of a scientific experiment is not to precipitate decisions, but to make an appropriate adjustment in the degree to which one accepts, or believes, the hypothesis or hypotheses being tested*." (Rozeboom, 1960) (Les italiques sont de Rozeboom.)

Pour Cox (1958) seule l'approche de Fisher est à considérer comme une réelle inférence statistique, celle de Neyman et Pearson relevant d'un problème très différent, celui de la "décision statistique". Mais certains auteurs, comme Bakan (1966), englobent l'approche fishérienne dans la critique, considérant que la méthode de

Neyman et Pearson n'a fait que révéler ce qui était implicite chez Fisher. Cet avis, mais pas les critiques, semble partagé par Lehmann (1993) qui voit ces deux théories plutôt comme complémentaires que comme réellement opposées.

2.1.9. Discontinu ou continu

Le passage d'un résultat significatif à un résultat non significatif est discontinu alors que la statistique de test est continue (Rozeboom, 1960). Ceci explique les discussions sur le choix de la valeur du seuil de signification α : ce choix n'aurait pas tant d'importance s'il n'y avait intériorisation, de la part des chercheurs, de la différence entre 0.05 et 0.06 comme différence entre le "vrai" et le "faux", la "réussite" et l'"échec". Skipper *et al*. (1967) évoquent même la joie ou l'horreur du chercheur selon que "son" *F* atteint 0.05 ou ne donne que 0.06. Ils regrettent particulièrement à ce propos que le choix du seuil se fasse, presque toujours, sans considération de la nature et du type de l'étude (contrairement à ce que préconisaient Fisher, *cf.* 1.1.).

Ce problème est également lié à celui portant sur l'alternative décision/jugement : une décision a un caractère discontinu, tandis que l'évolution d'un niveau de confiance en une hypothèse apparaît plutôt comme continu.

2.1.10. Un problème dérivé : un biais de publication

Comme l'on fait remarquer Sterling (1959), McNemar (1960) puis Bakan (1966), le fait que le test soit un critère important dans la sélection des articles soumis à publication, dans le sens où un résultat non significatif a peu de chances d'être publié, risque d'entraîner un biais de publication. Sterling a calculé en effet, pour quatre grands journaux de psychologie et pour l'année 1955 ou 1956, que 81% des articles mentionnaient l'usage de tests de signification, et que parmi ceux-ci 97% rejetaient l'hypothèse nulle. Ce faible nombre de résultats non significatifs publiés peut aussi bien résulter d'une sélection opérée par les chercheurs eux-mêmes, comme le pense McNemar (1960), que d'une politique éditoriale délibérée. Cette dernière est très bien illustrée par Melton, éditeur du *Journal of Experimental Psychology* pendant 12 ans, qui rappelle dans son éditorial de 1962 que les résultats non significatifs n'étaient acceptés que si la puissance était forte (comme cela est rarement le cas, il s'ensuit un très faible taux de publication des ces résultats).

Selon Sterling le biais résulterait du fait que, si plusieurs chercheurs testent, indépendamment les uns des autres, une même hypothèse nulle vraie (ou approximativement vraie), environ 5% d'entre eux trouveront un résultat significatif (au seuil de 5%) et seront pratiquement les seuls à même de publier, laissant ainsi croire à la réalité du phénomène étudié. Il s'ensuit une augmentation de "faux résultats" dans la littérature, et l'erreur de première espèce réelle atteint alors une valeur bien au delà de sa valeur nominale. Pour Bakan, comme pour Cohen lui-même, l'étude de Cohen (1962) (*cf.* la sous-section 5.1.1.) montrant la faible puissance des tests utilisés dans les articles publiés illustre ce biais.

L'argument de Sterling, plusieurs chercheurs testant la même hypothèse, n'apparaît pas très vraisemblable à Tullock (1959) car celui-ci suppose que les chercheurs n'ayant pas trouvé de résultats significatifs se manifesteraient après avoir lu le rapport d'un résultat significatif par un collègue. Pour Tullock il s'agirait plutôt de plusieurs chercheurs testant *plusieurs* hypothèses sans mérite, selon ses propres termes, ce qui aboutit au même résultat; il reste donc en accord, sur le fond, avec Sterling, quant au biais de publication.

En une trentaine d'années, la situation ne semble guère avoir évoluée. C'est ce que constatent Sterling *et al.* (1995) en s'appuyant sur le fait que le rejet de l'hypothèse nulle concerne 95% des articles utilisant un test et publiés en 1986-87 dans d'importants journaux de psychologie.

Mais cette critique peut être relativisée par la prise en compte de la critique évoquée dans la section qui suit : si, comme on va le voir, l'hypothèse nulle est presque toujours fausse, alors tout résultat est potentiellement significatif (s'il ne l'est pas encore c'est simplement par manque de sujets) et le biais de publication précédent est un faux problème. Il n'en reste pas moins que la sélection des articles qui en résulte peut avoir des conséquences sur la grandeur des effets publiés. Ainsi Sterling *et al.* (1995) arguent que les méta-analyses seraient faussées par le fait que les études publiées correspondraient plutôt à des effets observés suffisamment élevés pour entraîner un résultat significatif, et les études écartées (pour cause de non significativité) à des effets plutôt faibles, aboutissant au total à une surestimation.

2.1.11. Une hypothèse inutile

Une hypothèse nulle ponctuelle (c'est-à-dire attribuant au paramètre une valeur précise et non un intervalle) est pratiquement toujours fausse, ne serait-ce qu'à une quantité infinitésimale près. Aussi, pour autant qu'on y mette le prix en termes de taille de l'échantillon, tout résultat sera significatif et l'information apportée par le test est donc quasi-nulle (Berkson, 1942; Savage, 1957; Reuchlin, 1962, p. 371; Bakan, 1966; Meehl, 1967; Wilson *et al*., 1967; Lykken, 1968). Dans cet esprit, Carver (1978) qualifie, de façon imagée, l'hypothèse nulle d'"homme de paille". Nunnally (1960), conséquent, pousse l'argument à la limite pour conclure à l'inutilité de recueillir des données :

> "… if the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected. If rejection of the null hypothesis were the real intention in psychological experiments, there usually would be no need to gather data.". (Nunnally, 1960)

Cette critique n'est pas seulement le fait d'expérimentalistes; elle est aussi soulevée par des statisticiens, comme Berger (1985, pp. 20-22) ou Yates (1964) :

> "In many experiments […] it is known that the null hypothesis customarily tested, i.e. that the treatments produce no effects, is certainly untrue; such experiments are in fact undertaken with the different purpose of assessing the magnitude of the effects. Fisher himself was of course well aware of this, as is evinced in many of his own analyses of experimental data, but he did not, I think, sufficiently emphasise the point in his expository writings. Scientists were thus encouraged to expect final and definite answers from their experiments in situations in which only slow and careful accumulation of information could be hoped for. And some of them, indeed, came to regard the achievement of a significant result as an end in itself." (Yates, 1964, p. 320)

Même un défenseur "orthodoxe" des tests comme Binder (1963) admet cet argument et reconnaît qu'il serait plus judicieux de tester une hypothèse "approchée" (c'est-à-dire un intervalle autour de la valeur hypothétique du paramètre), mais il ne propose aucune méthode spécifique pour ce faire.

Pour illustrer le cas d'hypothèses nulles plausibles, ne reviennent bien souvent que les exemples de la génétique mendélienne et de la parapsychologie (voir, par exemple, Lindley, 1957).

Quant au cas d'une hypothèse nulle orientée (composée), Meehl (1967) relève que même dans le cas où la théorie sous-tendant l'hypothèse alternative est absolument sans intérêt (en fait un simple tirage au hasard de l'une des deux orientations possibles), l'accroissement de précision, par une taille d'échantillon élevée par exemple, entraîne que la probabilité de rejeter l'hypothèse nulle tend vers ½. On admettra donc, dans ces conditions, que le simple fait de trouver un écart significatif dans le sens attendu est une épreuve assez faible pour une hypothèse scientifique.

2.1.12. Corruption de méthode

Pour Carver (1978), l'introduction des tests statistiques dans la pratique scientifique a conduit à corrompre la méthode scientifique. Selon lui cette dernière consiste, une fois les données recueillies, à examiner si celles-ci sont compatibles avec l'hypothèse de recherche, puis, quand elles ne le sont pas, à envisager d'autres hypothèses (dont celle du hasard, éventuellement). Au contraire, dans le cas du test de signification, l'hypothèse du hasard est mise en avant; elle est la première testée, quelque intérêt scientifique qu'elle ait, et en conséquence, l'hypothèse de recherche ne sera même pas examinée si le test est non significatif, alors même qu'elle pourrait présenter une bonne compatibilité avec les données. Si Carver met bien en évidence un problème réel, son argument doit tout de même être tempéré par le fait que, contrairement (et heureusement) à ce qu'il affirme, tout chercheur n'accepte pas automatiquement l'hypothèse nulle en cas de résultat non significatif, et que donc l'hypothèse de recherche n'est pas forcément écartée définitivement.

2.1.13. Le statut de l'hypothèse de recherche

Les problèmes soulevés par les tests sont particulièrement bien illustrés par la question du statut de l'hypothèse de recherche, celle à laquelle le chercheur s'intéresse réellement, et qu'il espère confirmer : doit-on identifier l'hypothèse de recherche à l'hypothèse nulle ou à l'hypothèse alternative ?

Cette question a fait l'objet de nombreuses discussions (voir, par exemple, Grant, 1962; Binder, 1963; Edwards, 1965; Wilson *et al*., 1967).

Déjà il faut remarquer que le choix est limité par le fait que l'hypothèse nulle doit être ponctuelle pour permettre les calculs, comme Fisher le précise[8] :

> "It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the “problem of distribution”, of which the test of significance is the solution." (Fisher, 1990b/1935, p. 16)

Mais il arrive que l'hypothèse de recherche conduise à une prédiction précise pour le paramètre statistique et qu'elle puisse donc être candidate au rôle d'hypothèse nulle; c'est entre autres le cas lorsqu'il s'agit de validation de modèles. Par exemple Piaget (1975, p. 117), pour expliquer certains phénomènes en perception visuelle, introduit l'hypothèse d'une distribution au hasard des points de centration.

Dans le cadre fishérien, le choix peut sembler immédiat : puisqu'on ne peut que rejeter une hypothèse, pour obtenir une conclusion en faveur de l'hypothèse de recherche il faut que celle-ci soit représentée par l'hypothèse alternative. Mais alors, que faire dans le cas, par exemple, de l'hypothèse de Piaget évoquée à l'instant ? Sa négation ne donne pas de valeur précise, bien au contraire, et donc pas d'hypothèse nulle praticable. Cet exemple met en lumière le fait que la conception de Fisher ne vaut que pour autant qu'on ne veuille pas tester une hypothèse précise.

Quant à ceux qui se réfèrent, explicitement ou non, à la théorie de Neyman et Pearson, ils ne choisissent jamais, en pratique, l'hypothèse nulle selon le principe de Neyman (*cf.* section 1.2.) de prendre d'abord en compte la gravité relative des risques. Sans doute parce que ce principe est rarement exposé dans les manuels, mais aussi certainement parce qu'il n'est facile à appliquer que dans le cas du test de deux hypothèses ponctuelles qui ne se rencontre guère en pratique.

Contrairement à ce que Neyman et Pearson envisageaient plutôt, l'identification de l'hypothèse de recherche à l'hypothèse nulle est une position minoritaire en pratique, entre autres en raison des critiques formulées ci-après, mais sans doute aussi parce que cette théorie est, comme nous l'avons vu, souvent utilisée sous une forme “hybride”, le choix des hypothèses relevant du cadre fishérien.

Si l'on identifie l'hypothèse d'intérêt avec l'hypothèse alternative, alors, comme nous l'avons rappelé précédemment, il suffit de choisir un effectif assez élevé pour être pratiquement sûr d'un résultat favorable (significatif), quelle que soit la valeur scientifique de cette hypothèse, au point que Wilson *et al.*, 1967, soulignent à ce propos :

> "The present writers think that the indiscriminate cataloguing of trivial effects is, in fact, a major problem in psychology today…". (Wilson *et al.*, 1967)

Si l'on identifie l'hypothèse d'intérêt avec l'hypothèse nulle, on se trouve confronté à l'alternative suivante :

- Monter une expérience très sensible (augmenter la précision expérimentale), par exemple en prenant un grand effectif, c'est se ramener au cas précédent de rejet de l'hypothèse, alors même que celle-ci peut apparaître comme une très bonne approximation de la réalité, sinon même la meilleure disponible. Ce que Grant (1962) traduit ainsi :

  > "When theory or other circumstances permit the prediction of differences of specified size, using these predictions as the value in H_0 is tactically inappropriate, frustrating and self-defeating." (Grant, 1962)

- Monter une expérience très peu sensible (de très faible précision expérimentale), c'est permettre une corroboration à bon compte de l'hypothèse de recherche (dans le cadre de la théorie de Neyman et Pearson), un paradoxe dénoncé par Rouanet (1967, p. 12). Il suffirait, par exemple, de ne prendre que deux ou trois sujets pour être presque assuré de ne pas la rejeter. Là encore, en poussant à la limite, il n'est même plus besoin d'expérimenter.

Dans les faits, la plupart du temps l'hypothèse nulle est celle d'une absence d'effet (d'où, justement, la confusion entre hypothèse nulle et hypothèse de la valeur zéro), sans prise en compte des problèmes évoqués ci-dessus.

2.1.14. La question de l'intensité de l'effet

Le test ne dit rien quant à l'intensité, l'importance de l'effet parent (*cf.*, par exemple, O'Brien et Shapiro, 1968; Rouanet *et al.*, 1976). Un résultat significatif n'est qu'une indication de l'*existence* de l'effet supposé; un résultat non significatif un constat d'ignorance selon Fisher. Aller plus loin sur la seule base du test renvoie à l'erreur (exposée plus loin en 2.2.3.) d'assimiler significativité statistique et importance de l'effet. Or, presque tous les auteurs reconnaissent que la question de l'intensité de l'effet est essentielle. Dans un article de 1990, jetant un regard rétrospectif sur les pratiques statisticiennes en psychologie, Cohen se fait acerbe :

> "A salutary effect of power analysis is that it draws one forcibly to consider the magnitude of effects. In psychology, and especially in soft psychology, under the sway of the Fisherian scheme, there has

[8] Comme dit plus haut, il existe cependant des solutions fréquentistes pour le test d'un intervalle, mais elles sont assez peu connues des psychologues et ne sont que rarement utilisées dans la pratique.

> been little consciousness of how big things are. […] Because science is inevitably about magnitudes, it is not surprising how frequently *p* values are treated as surrogates for effect sizes. […] In retrospect, it seems to me simultaneously quite understandable yet also ridiculous to try to develop theories about human behavior with *p* values from Fisherian hypothesis testing and no more than a primitive sense of effect size. And I wish I were talking about the long, long ago." (Cohen, 1990, p. 1309).

Lykken (1968) présente un exemple particulièrement démonstratif où un effet moyen serait compatible avec certaines hypothèses psychologiques alors qu'un effet trop grand, ou trop faible, serait embarrassant pour la théorie.

Là encore les psychologues sont rejoints par des statisticiens; tel Cox qui, bien que défendant les tests (selon Fisher), reconnaît que la significativité statistique est très différente de la significativité scientifique :

> "statistical significance is quite different from scientific significance. Therefore estimation, at least roughly, of the magnitude of effects is in general essential regardless of whether statistically significant departure from the null hypothesis is achieved." (Cox, 1977, p. 61)

En fait il est très difficile de trouver un auteur qui s'inscrive contre l'intérêt d'estimer la grandeur de l'effet. On ne trouve guère que Chow (1988, 1989, 1996) qui considère que c'est inutile, du moins dans le cas où il s'agit de confirmer ou infirmer une assertion théorique, et qui défend le caractère binaire des décisions basées sur le test (au sens de Neyman et Pearson). Son argumentation est que l'acceptation ou le rejet d'une théorie est de type syllogistique et qu'un syllogisme est soit vrai soit faux. Il est d'ailleurs symptomatique que soit présenté, dans son article de 1989, un véritable diagramme de programmation, avec une succession de décisions binaires pour représenter le processus de corroboration d'une théorie. Il en deviendrait facile d'automatiser le comportement d'un chercheur ! Folger (1989) a facilement objecté que la validité (ou l'invalidité) d'une théorie ne relève pas d'une certitude comme il en existe dans le processus de déduction, et Falk et Greenbaum (1995) relèvent que le raisonnement de Chow est encore un exemple du renversement (abusif) des probabilités conditionnelles (de $Pr(D|H)$ à $Pr(H|D)$, *cf.* 2.1.7.).

Ce problème de la grandeur de l'effet peut encore se définir comme la possibilité de mettre en évidence, pour l'effet testé :

- Soit une intensité, ou grandeur, dite *négligeable*. C'est-à-dire une valeur qui, si elle n'est pas strictement nulle, pourra être tenue pour suffisamment faible pour constituer une bonne approximation du zéro (au moins à un certain stade de la recherche).
- Soit une intensité dite *notable*; c'est-à-dire, au contraire du cas précédent, importante, ou tout au moins impossible à négliger.

Nous reprenons ici la terminologie utilisée par B. Lecoutre, Lépine et Rouanet (Lépine et Rouanet, 1975; B. Lecoutre, 1984a). Bien entendu, il peut se faire que l'intensité d'un effet ne soit ni négligeable, ni notable, c'est-à-dire qu'elle soit intermédiaire, moyenne. Cohen (1962, 1969) parle lui d'effet “petit” (*small*), “moyen” (*medium*) ou “fort” (*large*). Si les deux terminologies recouvrent des notions voisines, voire équivalentes en ce qui concerne les deux dernières catégories (effet moyen et effet fort), il n'en va pas forcément de même pour la première. Pour Cohen, un effet petit n'est pas nécessairement négligeable, ni nécessairement non négligeable d'ailleurs. C'est un effet difficile à déceler mais qui existe, alors que la notion de négligeabilité englobe celle d'un effet existant mais d'intensité inférieure à une certaine limite, aussi bien que celle d'un effet nul (inexistant).

La question de l'intensité de l'effet est intiment liée à celle du statut de l'hypothèse de recherche, évoquée précédemment. En effet, ce qui est sous-jacent dans cette dernière question, étant entendu qu'un modèle n'est jamais parfait (autrement dit que l'hypothèse nulle *ponctuelle* n'est pas crédible), c'est bien :

- Soit la possibilité de mettre en évidence une différence négligeable entre la vraie valeur (inconnue) du paramètre et la valeur fixée par l'hypothèse (le modèle). C'est-à-dire de pouvoir stipuler une zone d'équivalence, ou d'approximation, autour de l'hypothèse nulle, de façon à ne pas systématiquement “jeter le bébé avec l'eau du bain” et de pouvoir admettre, au moins provisoirement, une hypothèse qui décrit suffisamment bien le phénomène étudié (à un moment donné du développement d'un domaine).
- Soit la possibilité de mettre en évidence un écart notable, ou au moins non négligeable, trop important par rapport à l'hypothèse pour pouvoir continuer à utiliser celle-ci.

Cette incapacité à traiter le problème de la grandeur de l'effet ainsi que les raisons présentées dans le paragraphe précédent font que les tests usuels sont inadaptés à la validation de modèles, comme Rouanet, notamment, l'a bien souligné et illustré (Rouanet, 1967, 1986; Rouanet *et al*., 1978) et comme le remarque encore récemment Bacher (1998) à propos des modèles structuraux. Même un auteur comme Frick (1996), qui défend l'intérêt des tests dans certaines conditions, partage ce point de vue.

2.1.15. Le paradoxe fondamental

Que le plus souvent on identifie l'hypothèse d'intérêt à l'hypothèse alternative amène par ailleurs à mettre l'accent sur un paradoxe fondamental mis en avant par Berkson dès 1942, repris par Morrison et Henkel (1970, p. 309), et à propos duquel Cohen (1990, p. 1307) évoque sa surprise quand il y fut confronté la première fois.

Un résultat significatif n'est pas une information positive *en faveur* de l'hypothèse alternative, il n'est qu'une évidence *contre* l'hypothèse nulle; alors que le but de l'inférence scientifique est au contraire d'apporter des éléments en faveur d'une hypothèse, d'une théorie :

> "Nor do you find experimentalists typically engaged in disproving things. They are looking for appropriate evidence for affirmative conclusions. Even if the mediate purpose is the disestablishment of some current idea, the immediate objective of a working scientist is likely to be gain affirmative evidence in favor of something that will refute the allegation which is under attack." (Berkson, 1942)

Ce paradoxe est au cœur de l'article de Meehl (1967) qui compare, en les opposant, les démarches expérimentales en psychologie et en physique.

La meilleure illustration est en fait fournie par Fisher lui-même qui décrit la démarche statistique comme visant à corroborer une hypothèse par la vérification des conséquences qui en sont tirées :

> "The statistical examination of a body of data is thus logically similar to the general alternation of inductive and deductive methods throughout the sciences. A hypothesis is conceived and defined with all necessary exactitude; its logical consequences are ascertained by a deductive argument; these consequences are compared with the available observations; if these are completely *in accord* with the deductions, *the hypothesis is justified* at least until fresh and more stringent observations are available." (Fisher, 1990a/1925, p. 8) (Italiques ajoutés.)

Ce qui ne l'empêche pas de développer une méthode qui fonctionne en sens contraire (!), méthode sur laquelle Meehl (1978) émet un jugement définitif lorsqu'il accuse Fisher d'avoir détourné les psychologues du droit chemin méthodologique et d'avoir porté un très mauvais coup à la psychologie en introduisant la méthodologie du rejet de l'hypothèse nulle :

> "I suggest to you that Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology." (Meehl, 1978, p. 817)

2.2. LES ABUS OU ERREURS D'INTERPRÉTATION

Si les tests font l'objet, en psychologie et ailleurs, d'une utilisation intensive, ils sont tout aussi intensivement l'objet de contresens, d'abus et de mésusages. Dans la mesure où nous adoptons ici une approche normative, nous caractériserons d'abord ces dérives comme des "erreurs", en insistant sur le fait que ce terme renvoie à un cadre de justification particulier (ici, Fisher ou Neyman et Pearson) : ce qui est erreur dans un cadre peut éventuellement ne pas l'être dans un autre. Bien entendu, nous exclurons les erreurs d'utilisation (par exemple l'usage d'une procédure non appropriée), qu'il arrive de rencontrer dans les publications, mais qui sont hors de notre propos. Un autre terme possible serait celui de "biais", qui est effectivement utilisé pour désigner des distorsions, des erreurs, dans le cadre plus général des études sur les jugements probabilistes (*cf.* Kahneman et Tversky, 1972), mais nous en réserverons l'usage à l'approche descriptive (chapitre 6).

À partir des pratiques, cinq grands types d'erreurs peuvent être distingués.

2.2.1. Le renversement des conditions : de *Pr*(*Données*|*Hypothèse*) à *Pr*(*Hypothèse*|*Données*)

La première erreur, la plus typique, et que nous avons déjà évoquée (*cf.* 2.1.7.), est de considérer la probabilité p (seuil observé) ou α comme une probabilité *concernant* l'hypothèse et non plus *conditionnelle* à celle-ci, et en plus d'omettre la proposition "conditionnellement aux données observées" (mais nous pouvons admettre que cette dernière proposition est implicitement entendue par les chercheurs). Ce qui donne :

La probabilité que l'hypothèse nulle soit vraie est p (ou α)

Koehler (1996) note, dans un cadre beaucoup plus général que celui de l'interprétation des tests statistiques, que la confusion de telles probabilités conditionnelles est courante et la dénomme “erreur inverse” (*inverse fallacy*).

En réalité on trouve plutôt cette erreur sous la forme suivante, que Carver (1978) appelle “fantasme des cotes contre le hasard” (*Odds-Against-Chance Fantasy*) :

La probabilité que les résultats soient dus au seul hasard est p (ou α)

et plus fréquemment encore sous une formulation (mathématiquement) équivalente, et que Carver (1978) appelle “fantasme de la validation de l'hypothèse de recherche” (*Valid Research Hypothesis Fantasy*) :

La probabilité que l'hypothèse alternative soit vraie est 1-*p* (ou 1-α)

Un exemple particulièrement frappant est offert par Rosenthal et Gaito (1964) qui concluent en faveur de leur hypothèse d'existence d'un certain effet sur la base des valeurs 1 - *p* correspondant à diverses études (en quelque sorte ils en considèrent une moyenne) :

> "In summary, the probability [of the effect] was established for several samples of psychologists. For one *N* of 20, *p* = .88; for one *N* of 19, *p* = .996; for a smaller *N* of 2, *p* = “1.00”; and for another *N* of 2, *p* = “0.00”." (Rosenthal et Gaito, 1964)

Il arrive aussi que la confusion soit telle qu’il devient difficile de préciser de quelle(s) forme(s) d'erreur il s’agit :

> "La majorité des chercheurs en psychologie ont recours à une épreuve de *signification statistique* pour décider si les résultats obtenus confirment ou infirment leur hypothèse. Cette épreuve permet d’établir quelle est la probabilité d’obtenir de tels résultats plutôt que ceux correspondant à l’hypothèse nulle, soit un postulat statistique attribuant les variations comportementales à des erreurs d’échantillonnage et de mesure, ainsi qu’au hasard." (M. Robert, 1995, p. 66) (Les italiques sont de l'auteur.)

Il faut reconnaître que Fisher lui-même prépare on ne peut mieux le terrain lorsqu'il explique que plus *p* est petit, plus l'évidence contre l'hypothèse nulle est grande :

> "The actual value of *p* […] indicates the strength of the evidence against the hypothesis." (Fisher, 1990a/1925, p. 80)

D'autant qu'il maintient encore cette conception en 1955, évoquant une “véritable mesure de la confiance” avec laquelle on peut soutenir une opinion (1955, p. 74). Il est alors bien tentant de formaliser simplement cette idée en identifiant “*evidence”*, confiance et probabilité… Même si Fisher se fait plus explicite ultérieurement en précisant qu'il n'est pas question d'assimiler cette confiance à une probabilité :

> "The level of significance in such cases fulfils the conditions of a measure of the rational grounds for the disbelief it engenders. It is more primitive, or elemental than, and does not justify, any exact probability statement about the proposition." (Fisher, 1990c/1956, p. 46)

Cette erreur est à ce point prégnante que Neyman lui-même l'a commise, comme nous l'avons déjà signalé dans la section 1.4. (voir Good, 1984, p. 159). Bien mieux, si l'on remonte aux débuts des tests avec Student (*alias* W. S. Gosset), on s'aperçoit, à la lecture de son fameux article de 1908, que le test est présenté comme répondant naturellement à la question de la probabilité de l'hypothèse. Ainsi, illustrant sa méthode par l'étude de la comparaison de deux soporifiques, et trouvant que ce qu'on noterait maintenant 1 - *p* vaut 0.9985, Student écrit :

> "From the table the probability is .9985 or the odds are about 666 to 1 that 2 [le second soporifique étudié] is the better soporific." (Student, 1908, p. 21)

Ceux qui commettent cette erreur sont donc en bonne compagnie.

2.2.2. 1 - *p* (ou 1 - α) considéré comme probabilité de reproduction du résultat

La deuxième erreur est de considérer 1 - *p* (ou 1 - α) comme la probabilité de reproduire le résultat observé (*Replicability or Reliability Fantasy*, dans le langage de Carver, 1978). Les théories traditionnelles des tests ne fournissent aucune indication de la probabilité de reproduire le résultat observé. Tout au plus, dans le cadre neyman-pearsonien, peut-on fournir une probabilité *conditionnelle* à une valeur spécifiée sous l'hypothèse alternative, similaire, voire identique, à la puissance du test.

Cette reproductibilité peut être envisagée de deux manières :

• Soit elle ne concerne que la significativité du résultat, c'est le cas le plus courant; ce qui donne un énoncé du type :

la probabilité qu'une réplique soit significative est 1 - p (ou 1 - α),

Goodman (1992), pour dénoncer cette erreur, produit un tableau de la probabilité qu'une réplique soit significative dans le même sens (au seuil de 0.05), en fonction du *p* obtenu dans une première expérimentation, à la fois dans le cadre neyman-pearsonien (en prenant la valeur observée pour hypothèse alternative) et dans le cadre bayésien standard (où la distribution *a priori* est uniforme). Les valeurs s'écartent considérablement de 1 - *p* :

p	Neyman-Pearson	Bayésien standard
0.10	0.37	0.41
0.05	0.50	0.50
0.03	0.58	0.56
0.01	0.73	0.66
0.005	0.80	0.71
0.001	0.91	0.78

Tableau 1

Probabilité d'une réplique significative dans le même sens (au seuil de 0.05) en fonction du seuil p du résultat

• Soit la reproductibilité concerne la valeur même de l'effet.

Un exemple typique de ce dernier cas est fourni par Nunnally (1975) qui, à propos d'un résultat significatif à 5%, précise que la différence observée correspondante a 95 chances sur 100 d'être observée de nouveau :

"… 95 [chances] out of 100 that the observed difference will hold up in future investigations." (Nunnally, 1975, p. 195; cité par Carver, 1978)

Il a été souvent reproché à Melton (1962) de commettre cette erreur (voir, par exemple, Bakan, 1966). Or, celui-ci n'a jamais affirmé que la probabilité de reproduire les résultats était égale à 1 - *p*; simplement, il a considéré que plus le seuil est faible et plus la reproductibilité est assurée (ce qui est justifié, quel que soit le cadre théorique qu'on adopte, comme on peut le constater dans le tableau précédent). Par ailleurs, il a insisté sur le fait que pour mériter la publication, le critère du test statistique était très insuffisant et qu'il était souhaitable d'élucider la nature de la relation fonctionnelle entre les variables, de répéter les expériences, de déterminer les conditions d'obtention des effets, *etc.* Autrement dit, il a adopté, pour l'essentiel, une attitude conforme à ce que souhaitaient beaucoup de ceux qui critiquaient les tests. Les reproches qui lui ont été adressés ne sont donc pas mérités, bien au contraire.

2.2.3. Significativité statistique et significativité substantielle

La troisième erreur est de confondre la significativité statistique avec la significativité substantielle (*substantive significance*). C'est considérer que plus un résultat est significatif, plus il est scientifiquement intéressant, et/ou que plus l'effet correspondant dans la population parente est grand (*substantive importance*).

Une illustration récente est fournie, à plusieurs reprises, par un article de Bassock *et al.* (1995), où l'on trouve, par exemple :

"In addition to the overall interpretative bias there was a *very strong* interaction between the training and the transfer problems [$\chi^2(1) = 14.71$, $p < 0.001$]." (Bassock *et al.*, 1995) (Italiques ajoutés.)

Cette erreur a été dénoncée très souvent (voir, par exemple, Selvin, 1957; Kish, 1959; Bolles, 1962; Bakan, 1966; O'Brien et Shapiro, 1968; Gold, 1969; Morrison et Henkel, 1969; Winch et Campbell, 1969). D'une manière implicite c'est contre elle que Reuchlin (1962, p. 370) met en garde le psychologue, lorsqu'il insiste sur le fait que c'est à celui-ci, et non au statisticien, de décider des hypothèses statistiques à tester. C'est au psychologue de savoir si, du point de vue de la signification psychologique, il ne vaut pas mieux choisir pour hypothèse nulle qu'entre les moyennes de deux groupes la différence est inférieure à un point (pour une certaine échelle), plutôt qu'une différence exactement égale à zéro.

2.2.4. L'acceptation de l'hypothèse nulle

La quatrième erreur est de conclure à la véracité de l'hypothèse nulle en cas de résultat non significatif. Ceci n'est une erreur, à strictement parler, que dans le cadre fishérien. Et dans le cadre neyman-pearsonien, l'acceptation de l'hypothèse nulle ne se comprend que par référence à une puissance, ou, à défaut, une courbe de puissance, clairement indiquée. Peu de chercheurs mentionnent explicitement à quelle théorie ils se réfèrent, mais on peut considérer que celle de Fisher est en jeu dès lors que le seuil observé est le seul mentionné, sans référence à un seuil fixé à l'avance et qu'il n'est pas question non plus de la puissance du test utilisé.

A nouveau chez Bassock *et al.* (1995) on peut lire :

> "Subject's performance *was not affected* by differences in the size of the assigned and the receiving sets [$\chi^2(1) = 0.08$, n.s.], so we combined the results of subjects…" (Bassock *et al.*, 1995) (Italiques ajoutés.)

Ceci montre bien que cet auteur conclut inférentiellement (ici à l'absence d'effet *parent*) et ne se contente pas de jugements descriptifs qui seraient valides, car la combinaison des résultats n'est légitime que si l'effet *parent* est nul.

Harcum (1990) donne d'autres exemples d'acceptations "désinvoltes" d'hypothèses nulles, dont certains qui sont apparus dans le *Journal of Experimental Psychology*.

Cette erreur peut même être le fait de ceux qui en sont parfaitement avertis. C'est par exemple le cas de Mialaret (1996) qui consacre un paragraphe intitulé "Remarque importante sur le type de conclusion" pour mettre en garde contre l'acceptation de l'hypothèse nulle :

> "La consultation des tables permet simplement de dire que l'on ne peut pas refuser l'hypothèse posée au début. Il est vrai que, dans la pratique, beaucoup diront, et cela par un abus de langage strict, que les 3 groupes ne présentent pas de différence significative entre eux, qu'ils appartiennent à la même population. L'interprétation correcte est bien : « on ne peut pas refuser l'hypothèse posée au départ »." (Mialaret, 1996, p. 127)

L'auteur dénonce ici avec raison l'abus consistant à conclure à l'existence d'une même population d'origine en cas de résultat non significatif[9]. Mais il commet lui-même cet abus dans les pages précédentes, à l'occasion de la présentation de l'intervalle de confiance, qu'il utilise pour introduire les tests "en douceur". Dans la partie intitulée "Comparaison des moyennes de deux échantillons appariés", il donne en effet un exemple (portant sur l'influence de l'ensemencement sur le nombre de poissons pêchés) dont la conclusion est précisément une absence d'effet :

> "La valeur 0 étant comprise dans l'intervalle de confiance on ne peut pas refuser l'hypothèse nulle selon laquelle les deux séries de valeurs ont la même moyenne. On dira, en d'autres termes, que l'ensemencement *n'a pas eu d'effet* sur la prise des pêcheurs." (Mialaret, 1996, p. 112) (Italiques ajoutés.)

2.2.5. L'omission de la condition

La dernière erreur consiste à considérer la probabilité p (seuil observé) ou α comme une probabilité *non conditionnelle*, c'est-à-dire d'omettre la proposition "si l'hypothèse nulle est vraie", ce qui donne :

La probabilité de commettre l'erreur de première espèce est p (ou α).

En effet la probabilité de commettre l'erreur de première espèce c'est, à strictement parler, la probabilité *conjointe* que H_0 soit vraie *et* qu'on rejette H_0. Soit, avec $Pr(H_0)$ désignant la probabilité *a priori* de H_0 :

$$Pr(H_0 \cap \text{rejeter } H_0) = Pr(H_0).Pr(\text{rejeter } H_0|H_0) = Pr(H_0).p \quad (\text{ou } Pr(H_0).\alpha)$$

(En 1933, c'est bien ainsi que la définissent Neyman et Pearson; voir, par exemple 1933b, p. 495.)

Si l'on refuse l'idée de probabiliser une hypothèse, alors $Pr(H_0)$ ne peut prendre que les valeurs 0 (H_0 est fausse) ou 1 (H_0 est vraie) et la probabilité de commettre l'erreur de première espèce se réduit à l'alternative 0 ou bien p (ou α), mais non à un seul de ses termes.

Cette erreur se rencontre fréquemment, en particulier dans des ouvrages ou articles spécialisés, comme par exemple dans Kirk (1982, pp. 36-37), ou encore dans Rogers *et al.* (1993).

Assurément ceci est de peu d'importance dans la mesure où l'on peut avancer que la proposition "quand l'hypothèse nulle est vraie" est sous-entendue car allant de soi dès que l'on évoque le risque de première espèce, et que le langage courant favorise de telles ambiguïtés. Neyman lui-même, dans son recueil de 1952 (p. 57), évoluant par rapport à la présentation de 1933, identifie cette probabilité de commettre l'erreur de première

[9] En revanche, contrairement à ce qu'affirme Mialaret, il n'y a pas d'abus de langage à dire qu'il n'y a pas de différence significative; cette expression n'est, par définition, que le compte-rendu du résultat du test, et ne préjuge pas de l'interprétation de ce résultat.

espèce à α, arguant qu'elle doit être comprise, et calculée, sous l'hypothèse de la véracité de H_0 (il énonce pourtant, dans le même passage, que les circonstances dans lesquelles on commet une erreur de ce type correspondent à la *conjonction* de la véracité de H_0 et de son rejet).

De la même façon on trouve parfois que :

La puissance est la probabilité de correctement rejeter l'hypothèse nulle.

Alors qu'il faudrait encore préciser “si l'hypothèse nulle est fausse”. Là aussi ce peut être le signe d'une erreur réelle ou d'une simple ambiguïté de langage, selon que l’on comprend “correctement rejeter l'hypothèse nulle” comme “l'hypothèse nulle est fausse *et* on la rejette” (erroné) ou comme “rejeter l'hypothèse nulle *sachant* qu′elle est fausse” (correct).

2.2.6. Tel est pris…

Ces erreurs sont fréquentes, même parmi des chercheurs confirmés, comme le soulignent Wilson (1961) et Bakan (1966) :

> "Most psychologists and other users of statistics believe that this minimum significance level is the ‘probability that the results are due to chance’ and many applied statistics texts support this belief." (Wilson, 1961, p. 230)
>
> "The psychological literature is filled with misinterpretations of the nature of the tests of significance." (Bakan, 1966)

Même ceux qui critiquent ne sont pas irréprochables.

Ainsi Bolles (1962), alors même qu'il vise à mettre en garde contre la confusion entre significativité statistique et significativité substantielle, commet la première erreur décrite ci-dessus (dans sa 3ème formulation) en introduction de son article :

> "… when a statistician rejects the null hypothesis at a certain level of confidence, say .05, he may be fairly well assured (p = .95) that the alternative statistical hypothesis is correct." (Bolles, 1962, p. 639)

Quant à Cohen, grand pourfendeur de la théorie de Fisher, c'est à propos de la notion qu'il a le plus à cœur de promouvoir, celle de puissance, qu'il lui arrive d'écrire :

> "…the question of the probability that his investigation would lead to statistically significant results, i.e., its power ?" (Cohen, 1969, p. vii)

Là encore il manque la condition “si l'hypothèse nulle est fausse”.

Oakes (1986) fournit plusieurs exemples d'abus, dont celui-ci, commis par Morrison et Henkel, éditeurs du célèbre recueil d'articles critiques, qui omettent, eux aussi, la condition :

> "… Thus, any difference in the groups on a particular variable in a given assignment will have some calculable probability of being due to errors in the assignment procedure…" (Morrison et Henkel, 1970, pp. 195-196)

Cette liste n'est bien sûr pas exhaustive.

2.3. DES RAISONS DES ABUS ET DE LEUR PERSISTANCE

Les erreurs précédentes peuvent apparaître comme une critique supplémentaire des tests. En effet, une méthode qui donne lieu à tant d'erreurs dans son application, même chez des usagers avertis, pose, pour le moins, le problème de son adéquation aux besoins des chercheurs. Parvenu à ce point, il nous paraît donc approprié de renommer ces erreurs en “abus d’utilisation” et de rechercher dans leur existence même des raisons de la popularité des tests statistiques.

Il s’agit en effet maintenant de chercher à expliquer pourquoi les critiques des tests n'ont guère influé sur la pratique des chercheurs en psychologie, comme en témoignent les articles encore récemment publiés sur le sujet (voir, par exemple, Cohen, 1990, 1994; Schmidt, 1996). Le titre de l'article de Falk et Greenbaum (1995) est d'ailleurs particulièrement explicite : “Significance tests die hard”.

2.3.1. La popularité des tests et l'existence des abus

L'ambiguïté de la terminologie

Le choix du terme même de “test de signification” porte à confusion. Ce choix est malheureux; “significatif” renvoyant, dans le discours commun, à quelque chose qui donne du sens et qui a de l'importance, il favorise la confusion entre significativités statistique et substantielle. Cependant cela ne peut être une raison profonde, les scientifiques étant habitués à redéfinir pour leur usage des mots usuels dans un sens spécifique.

L'objectivité

Les tests confèrent aux conclusions tirées une impression d'objectivité (Bakan, 1966; Stevens, 1968; Carver; 1978; Efron, 1986).

Sans aucun doute ceci est une raison fondamentale; d'ailleurs Fisher lui-même clamait haut cette objectivité. Il est évident que l'objectivité est un souci crucial des chercheurs scientifiques et une méthode d'inférence qui s'en réclame ne pouvait être que bienvenue. La théorie de Neyman et Pearson, avec l'accentuation de la formalisation, semblait également répondre à cette attente. Winch et Campbell (1969), par exemple, illustrent bien ce point :

> "We reason that it is very important to have a *formal* and *nonsubjective* way of deciding whether a given set of data shows haphazard or systematic variation. […] And we believe it is important not to leave the determination of what is systematic or haphazard arrangement of data to the *intuition* of the investigator." (Winch et Campbell, 1969) (Italiques ajoutés.)

La scientificité

L'appareillage mathématique des tests fournit une apparence de scientificité (*cf.*, par exemple, Carver, 1978; Dar, 1987). La rigueur des mathématiques et l'aura dont elles jouissent sont sensées rejaillir sur l'ensemble de la recherche (un effet de halo en quelque sorte), assurant sa validité. En psychologie en particulier, elle permet au chercheur de pouvoir se défendre contre ceux qui accusent la psychologie de ne pas être une science. Carver évoque encore une autre attitude défensive, dans les cas où l'effectif est faible, pour rendre compte de l’attachement au test : un résultat significatif est vu comme une parade à une critique qui porterait sur la faible taille de l'échantillon.

Falk et Greenbaum (1995) avancent encore l'argument qu'il existe chez les chercheurs un manque de différenciation entre le raisonnement déductif pur et la logique du test.

Le renfort de Popper

L'approche de Fisher offre une ressemblance avec les idées de Popper (1973/1939). Pour ce dernier la démarcation entre énoncés scientifiques et non scientifiques est réalisée sur la base du caractère réfutable, ou non, de ces énoncés : une hypothèse scientifique est une hypothèse qui peut être empiriquement contredite; en revanche, jamais elle ne pourra être vérifiée (au sens de “logiquement prouvée”). On retrouve la même asymétrie entre réfutation et vérification que chez Fisher. La théorie des tests de Fisher, comme le pense Oakes (1986), a ainsi pu bénéficier du succès certain des idées de Popper. Encore récemment Serlin et Lapsley (1993) invoquent ces idées, et Reuchlin leur consacre un paragraphe dans son ouvrage de 1992. Bien sûr cet argument ne s'applique pas à la théorie de Neyman et Pearson puisque celle-ci vise à des décisions, des actions, ce qui ne concerne pas Popper, et que pour lui il est impossible d'accepter une hypothèse.

Un confort assuré

Les tests assurent un confort certain à leurs utilisateurs. En semblant fournir une procédure automatique, les tests dispensent d’une réflexion supplémentaire (Bakan, 1966; Carver, 1978; Falk and Greenbaum, 1995); de plus l'existence de conventions assurant un style “scientifique” d'analyse est des plus rassurantes (Johnstone, 1988).

Une économie

Les tests déchargent le chercheur de la tâche d'interprétation. Carver (1978) évoque le fait particulier que bien souvent le chercheur se trouve confronté à des unités de mesure qu'il ne sait interpréter (il prend l'exemple d'une échelle de “qualité d'enseignement” : que représente une différence de 10 points sur cette échelle ?). Dans ces conditions, le test, avec sa possibilité de déclarer “significatif” un effet, est vu comme une

solution, déchargeant le chercheur de la tâche d'interprétation, comme si la significativité statistique se suffisait à elle-même. On retrouve la confusion entre significativité statistique et significativité substantielle.

Des réponse aux attentes

Les tests semblent répondre exactement à l'attente des chercheurs, à propos de la probabilité des hypothèses, de la réplicabilité des résultats, de l'importance des effets, comme nous en avons déjà parlé dans la section 2.2., et cela, en raison justement des abus dont ils font l'objet. Selon beaucoup d'auteurs (voir, par exemple, Carver, 1978; Oakes, 1986; Cohen, 1990; Falk et Greenbaum, 1995; Schmidt, 1996), il faut voir là la principale raison de la popularité des tests. Ce sont les abus mêmes qui conduiraient à une illusion d'adaptation. Et ces abus seraient favorisés par le manque de naturel du raisonnement sur lequel est basé le test (*cf.* section 2.1.) qui est une difficulté à comprendre la portée exacte du test.

2.3.2. La persistance de la popularité des tests et des abus

Il reste à expliquer la persistance de cette popularité et de ces abus malgré une vague ininterrompue de critiques. Plusieurs hypothèses peuvent être avancées, mais sans doute aucune n'est à elle seule suffisante.

Une question de temps

Une première explication est que les critiques n'ont pas encore eu le temps de se diffuser.

Mais deux arguments viennent à l'encontre de cette raison.

1) D'une part, comme on l'a vu, les critiques apparaissent régulièrement depuis les débuts des tests (dès les années 30), ce qui a laissé tout de même un temps respectable pour leur diffusion, d'autant qu'elles sont parues, en général, dans des revues de renom.

2) D'autre part, l'exemple de la controverse sur les tests unilatéraux et bilatéraux montre que les recommandations statistiques peuvent être rapidement adoptées, comme en témoigne, par exemple, Burke (1953) :

> "While the *popularity* of one-tailed tests is undoubtedly attributable in part to the overwillingness of psychologists as a group *to make use of the statistical recommendations they have most recently read*,…". (Burke, 1953) (Italiques ajoutés.)

Cette controverse, qui date des années 50 et concerne la question de la légitimité des tests unilatéraux, les tests bilatéraux étant jusqu'alors la norme (voir, par exemple, Marks, 1951, 1953; Jones, 1952, 1954; Hick, 1952; Burke, 1953, 1954) a, de plus, eu pour support les mêmes journaux scientifiques que la controverse qui nous occupe.

Nous serions alors plutôt conduit à rejeter cette hypothèse.

Une ignorance du domaine

Dans le même ordre d'idées, on pourrait penser que les chercheurs, à part ceux travaillant dans le domaine ou particulièrement intéressés par ces questions, ne lisent pas, en général, les articles méthodologiques. Mais on se heurte, comme précédemment, au même argument de l'impact de la controverse sur les tests unilatéraux et bilatéraux, et cela apparaît donc peu plausible; sauf à supposer que l'attitude des psychologues a changé après les années 50 et qu'ils ont alors délaissé ce qui ne concernait pas directement leur sujet. Ce n'est pas impossible, étant donné la tendance à la spécialisation, en psychologie comme dans les autres disciplines scientifiques, et en raison du nombre croissant de publications.

Par ailleurs, même si les critiques sont lues, on peut parfois se demander comment elles le sont, ou comment elles sont utilisées par la suite. Ainsi Cohen profère une critique violente des tests d'inspiration fishérienne (le ton de l'article de 1994 est particulièrement violent), mais ne trouve rien à reprocher à la méthode de Neyman et Pearson, bien au contraire puisqu'il en est un fervent défenseur au travers de l'étude de la puissance. Et pourtant, à l'appui de sa critique il cite des auteurs (tels Rozeboom, 1960, et Bakan, 1966) qui s'avèrent au moins aussi durs envers la théorie de Neyman et Pearson qu'envers celle de Fisher; mais cela semble scotomisé, pour ainsi dire, par Cohen. Schmidt (1996) fait de même.

Une légitimité de fait

Le nombre considérable de domaines de l'activité scientifique touchés par les tests et le temps passé depuis l'apparition des premières critiques peuvent en eux-mêmes apparaître comme des arguments en faveur des pratiques actuelles : qui irait imaginer qu'une méthode aussi universelle et qui a survécu à de telles attaques

depuis si longtemps ne soit pas appropriée ? Frick (1996) donne ainsi l'exemple d'un “reviewer” anonyme qui avance ces raisons :

> "A way of thinking that has survived decades of ferocious attacks is likely to have some value." (Frick, 1996, p. 379)

Une résistance au changement

Les problèmes soulevés par les critiques sont beaucoup plus fondamentaux que dans le cas de la controverse test bilatéral *vs* test unilatéral où il n'y a pas de remise en cause du cadre conceptuel général. On peut penser que les résistances des chercheurs à modifier leur pratique sont alors beaucoup plus grandes. C'est ce qu'exprime Bakan (1966) sous une forme imagée en considérant le test comme un des “fils de la tapisserie culturalo-scientifique” de la psychologie, profondément emmêlé avec les autres : tirer dessus (reconnaître son inadéquation) risquerait de démanteler l'ensemble de la tapisserie; aussi les psychologues préféreraient-ils effectuer des ajustements (pourvoir le test de qualités qu'il n'a pas). Et comme le test n'est pas totalement sans mérite, Bakan se demande si nous ne sommes pas confrontés au phénomène de la grande résistance à l'extinction d'un apprentissage avec renforcement partiel.

Une soumission aux critères de publication

Le fait que le test statistique soit devenu un critère très important, une norme pour la publication joue certainement un grand rôle dans la perpétuation des attitudes. Winkler (1974) relève la grande uniformité des plans d’expérience et des analyses statistiques afférentes dans les articles publiés dans les revues relevant de l’*American Psychological Association*. Aussi un chercheur, même convaincu du peu d'intérêt d'un test, hésitera à ne pas en produire, craignant un refus de publication. Cela est clairement exprimé dans M.-P. Lecoutre (1991) où certains chercheurs justifient ainsi l'utilisation du test : “Si on veut se donner le maximum de chances de publier, il faut présenter un résultat significatif”, “Je fais des tests seulement pour satisfaire les journaux dans lesquels je publie”. Carver, en 1978, était pessimiste quant aux chances que les tests soient abandonnés tant que des éditeurs de journaux scientifiques n'auraient pas pris explicitement position contre leur utilisation.

Une soumission aux statisticiens

Oakes (1986) invoque la soumission des chercheurs à l'autorité des statisticiens. Mais cela ne tient pas compte du fait que ces derniers sont eux-mêmes divisés quant aux théories statistiques, ce dont les psychologues sont parfaitement ignorants pour la plupart (comme le rappelle Gigerenzer, 1993).

Un enseignement qui n'évolue pas

Un facteur probablement décisif, bien qu'assez peu mentionné en général, est le rôle de l'enseignement des statistiques en psychologie. D'une part cet enseignement est muet sur les controverses parmi les statisticiens (voir la préface de Oakes, 1986, par exemple, ou Gigerenzer, 1993). Et d'autre part, comme le souligne Schmidt (1996), malgré toutes les critiques cet enseignement n'a pratiquement pas évolué durant les dernières décennies et les tests y occupent une place bien trop importante, perpétuant ainsi l'état de fait. Pour lui, il n'y aura pas de changement sensible qui ne passe par une modification profonde de l'enseignement; ce à quoi il s'est d'ailleurs engagé.

Un moindre mal

Il n'y a pas de solution immédiate et praticable qui fasse l'unanimité. Les méthodes de rechange (d'intervalle de confiance, de rapport de vraisemblance, bayésiennes, *etc.*) ne sont pas disponibles facilement en dehors des situations élémentaires et/ou font l'objet d'autres critiques (*cf.* le chapitre 3). Et les chercheurs ne sont pas prêts, en général, à abandonner purement et simplement l'inférence dont ils sentent bien la nécessité, comme ils l'expriment, par exemple, dans les expériences menées par M.-P. Lecoutre (1991). Aussi l'utilisation des tests pourrait être considérée par ces chercheurs comme un moindre mal.

Un comportement névrotique

Gigerenzer (1993) propose, lui, une métaphore psychanalytique dans laquelle les idées bayésiennes jouent le rôle du ça, la théorie de Neyman et Pearson le rôle du surmoi et celle de Fisher le rôle du moi. Il existerait un conflit entre le ça et le surmoi, d’où l'anxiété, et il en résulterait un aveuglement dogmatique aboutissant à la négation du problème, ce qui expliquerait que les choses ne changent pas, ou si peu (1993,

p. 325). Mais quelque original et intéressant que soit cet argument, Falk et Greenbaum (1995) l'ont écarté en faisant fort justement remarquer que les psychologues ne connaissent pas d'angoisse (du moins à ce propos). Il nous semble effectivement, au travers de notre expérience de “conseiller en statistiques”, que l'angoisse des psychologues ne porte pas sur la question “dois-je faire un test ?”, mais seulement sur la question “ai-je fait le *bon* test ?”.

Un faux problème

Rozeboom (1960) suggère que si les critiques n'ont pas perturbé les psychologues, c'est peut-être tout simplement que ceux-ci n'ont jamais pris les tests au sérieux (au moins pour ce qui concerne la théorie de Neyman et Pearson). Mais on peut craindre que cette vision soit optimiste.

Une religion

En fait, souvent les tests font l'objet d'une véritable religion et sont pratiqués comme un rituel (Pitz, 1972, cité par Winkler 1974; Salsburg, 1985). Guttman (1979, 1983) a d'ailleurs dénoncé à ce propos les “adorateurs des étoiles”. En référence à cette importance accordée aux étoiles, Rouanet (1991a) évoque, ironiquement, une “respectabilité hôtelière”. Il n'est pas facile de s'opposer à une religion, surtout quand elle est si répandue, et les arguments logiques ont peu de prise en ce domaine.

* * *

Si dans l'ensemble les chercheurs ont accueilli ce nouvel outil statistique qu'a été le test de signification de façon très favorable, il est frappant de constater que les critiques à son encontre ont été immédiates et n'ont pas cessé; sans trop évoluer d'ailleurs : des critiques fondamentales, qui dénoncent l'inadéquation du test à la démarche expérimentale, ont été formulées très tôt et sont périodiquement réitérées ou redécouvertes.

Si les tests ont été constamment critiqués, ce n'est pas seulement en raison de leurs défauts intrinsèques mais également parce qu'il donnent massivement lieu à des erreurs ou abus d'interprétation de la part des chercheurs qui les utilisent. La critique est donc double : critique de l'instrument et de son usage.

Tout un faisceau de raisons concourt à justifier la popularité des tests et à perpétuer leur usage. Des raisons importantes de cette perpétuation sont les motifs mêmes de leur succès dans la communauté scientifique, et notamment l’apparence d’objectivité et de scientificité qu'ils apportent aux conclusions tirées, ainsi que l’illusion d’adaptation que leur confèrent leurs abus d’utilisation. En retour, le rôle de norme, de rituel que joue maintenant le test dans les publications légitime de la part des chercheurs une résistance considérable au changement.

2ème PARTIE

APPROCHE PRESCRIPTIVE

Même s'ils ont une visée d'application, la plupart des travaux fondateurs des deux grandes théories des tests que nous avons considérées dans la partie précédente sont avant tout des travaux de statistique destinés à être lus par des statisticiens. C'est par l'intermédiaire d'autres auteurs, psychologues et/ou statisticiens que les tests statistiques sont rendus accessibles au chercheur en psychologie[10]. Ces exégètes traduisent la norme statistique en une prescription, c'est-à-dire en une recommandation expresse accompagnée de toutes les instructions utiles pour la mise en œuvre effective des tests dans un domaine donné. Ce sont les manuels de statistique appliquée qui permettent une diffusion accréditée de cette prescription auprès des étudiants et des chercheurs, et qui ont été généralement à l'origine des logiciels d'analyse statistique des données. Ces manuels servent également de référence (d'autorité) aux chercheurs pour justifier, s'il y a lieu, l'utilisation de telle ou telle méthode.

Mais les prescripteurs ont également un autre rôle à jouer, qui est au contraire, en s'appuyant sur les critiques des tests, d'en contre-indiquer éventuellement l'usage et dans ce cas de recommander d'autres méthodes. La plupart des auteurs qui ont critiqué l'usage des tests de signification ont d'ailleurs tenu ce rôle. Si certains, tout en reconnaissant les limites des tests de signification, en défendent néanmoins un usage restreint (Fowler, 1985; Frick, 1995b, 1996; Greenwald *et al.*, 1996), d'autres, tel Hogben (1957), en arrivent jusqu'à préconiser l'abandon pur et simple de toute méthode d'inférence statistique (du moins dans le contexte de la recherche scientifique) pour ne plus se fier qu'au "bon sens". Cependant l'attitude majoritaire est de chercher à prescrire des solutions de rechange.

La plupart de ces solutions renvoient à d'autres méthodes d'inférence, notamment l'intervalle de confiance et les procédures bayésiennes. Une étude normative de ces méthodes dépasserait le cadre de ce travail, et nous nous contenterons de passer brièvement en revue les principales approches proposées et d'examiner leurs implications méthodologiques ainsi que les difficultés qu'elles soulèvent. Cet examen, qui fera l'objet du chapitre 3, sera un élément important pour savoir si l'usage des tests peut véritablement être remis en cause et pour expliquer quelle évolution on peut attendre des pratiques des chercheurs dans les années à venir.

Dans le chapitre 4 nous analyserons quelques uns des manuels d'inférence statistique parmi les plus connus à l'usage des psychologues. Nous examinerons comment y sont présentés les tests de signification, en référence aux théories statistiques, et dans quelle mesure ils tiennent compte des critiques formulées et intègrent certaines solutions de rechange.

[10] Fisher a également joué le rôle d'un prescripteur avec ses deux grands ouvrages de 1925 et 1935 (*Statistical Methods for Research Workers* et *The Design of Experiments*); mais, de nos jours, ces livres ne sont plus vraiment des ouvrages de base pour l'étudiant ou le chercheur en psychologie.

CHAPITRE 3

LES SOLUTIONS DE RECHANGE PRÉCONISÉES

3.1. LA MESURE DE LA GRANDEUR DE L'EFFET

La suggestion de loin la plus fréquente est d'étudier l'intensité des effets (voir, par exemple : Nunnally, 1960; Cohen, 1962, 1969, 1990; Hays, 1963; Bakan, 1966; Vaughan et Corballis, 1969; Dwyer 1974; Craig *et al.*, 1976; Cox, 1977; Carver, 1978; Guttman, 1983; B. Lecoutre, 1984a; Harris, 1991; Rogers *et al.*, 1993; Rouanet, 1996; Schmidt, 1996), sans pour autant, bien sûr, assimiler l'intérêt d'un effet à sa grandeur (voir, par exemple, O'Grady, 1982). Cette étude de l'effet est vue soit comme un prolongement, soit comme un remplacement de la procédure de test, et elle a été abordée de façons très différentes. Loin de s'opposer aux approches décrites dans les sections suivantes, elle en constitue au contraire le plus souvent un préalable.

Nous traiterons comme équivalents les différents termes que l'on rencontre, à savoir : "intensité", "taille", "magnitude", "grandeur", "importance".

Nous présentons ici brièvement et sans prétention d'exhaustivité quelques unes des statistiques les plus souvent proposées pour mesurer la grandeur de l'effet.

3.1.1. Le plus simple

Sans aucun doute, dans le cas de la comparaison de deux groupes pour une variable numérique, l'indicateur le plus simple et le plus naturel est la différence des deux moyennes (*cf.* Vaughan et Corballis, 1969; B. Lecoutre, 1984a). Cet indicateur est cependant critiqué (voir, par exemple, Richardson, 1996), sous le prétexte qu'il dépend de la procédure particulière utilisée pour le recueil des données (choix de la variable, essentiellement). Pour cette raison on lui préfère souvent un indicateur relatif, calibré par un écart-type (*cf.* ci-dessous). Mais on peut alors objecter qu'une mesure relative, sans unité, ne donne que l'illusion d'une mesure invariante selon les procédures expérimentales, la question de savoir si ces dernières ont ou non une influence étant de nature *empirique* et non pas statistique.

3.1.2. Les indicateurs de grandeur d'effet de Cohen (1962, 1969)

Il s'agit d'indicateurs de la grandeur de l'effet dans la population parente et ils couvrent l'essentiel des situations rencontrées en psychologie (variables numériques et catégorielles).

Dans le cas de l'effet d'un facteur expérimental sur une variable numérique, Cohen propose une mesure relative, le coefficient *f*, qui est le rapport de l'effet brut (défini comme l'écart-type des moyennes théoriques des niveaux du facteur) sur l'écart-type "d'erreur". Dans le cas particulier de la comparaison de deux moyennes, il définit le coefficient *d* où seul le numérateur change et devient la valeur absolue de la différence des moyennes. De ce fait ces deux indicateurs *d* et *f* ne sont pas homogènes (on a : $d = 2f$), et il doit donc donner des définitions différentes des effets "faible", "moyen" et "fort" pour chacun des deux cas. Au contraire, B. Lecoutre (1984a) choisit, pour représenter l'effet brut, un indicateur homogène au cas de deux moyennes, la moyenne quadratique des différences deux à deux des moyennes, cet indicateur étant calibré par un écart-type ("standardisé") pour obtenir un effet relatif. Essentiellement préoccupé par la question de la puissance, Cohen ne traite pas du problème de l'estimation de ses indicateurs à partir des données d'échantillons. L'application directe de ses formules au niveau de l'échantillon amène à des estimateurs qui peuvent être fortement biaisés; ce problème de l'estimation est notamment évoqué par Richardson (1996).

Dans le cas de la mesure du degré d'association linéaire de deux variables numériques, Cohen préconise l'emploi de la valeur absolue du coefficient de corrélation usuel de Bravais-Pearson, mais en général on préfère utiliser son carré (souvent appelé coefficient de détermination) qui s'interprète comme la part de variance "expliquée" par le modèle linéaire par rapport à la variance totale.

Dans le cas de tableaux de contingence, le coefficient *l* de Cohen n'est autre que le carré moyen de contingence ϕ^2 (analogue au χ^2 mais calculé à partir des fréquences, soit $\phi^2 = \chi^2/N$), défini au niveau de la population.

L'ensemble des indicateurs proposés par Cohen sont donnés dans le Tableau 2 de la sous-section 5.1.1.

3.1.3. Les indicateurs en "part de variance expliquée"

Il s'agit d'exprimer l'effet d'une variable indépendante sur une variable dépendante numérique. Le plus souvent la variable indépendante est un facteur expérimental ou, d'une manière plus générale, une comparaison, mais ce peut être aussi une variable numérique. L'idée est de mesurer l'effet comme la proportion de variance imputable à celui-ci par rapport à la variance totale; ou, ce qui est équivalent, comme la réduction relative de l'incertitude sur la variable dépendante, exprimée par sa variance, apportée par la connaissance de la variable indépendante.

À partir de cette idée, divers indicateurs sont possibles. Dans la pratique, essentiellement trois coefficients sont utilisés, les deux premiers portant l'accent sur l'effet au niveau de l'échantillon et le troisième sur l'effet au niveau de la population parente.

• Le coefficient r^2 (coefficient de détermination de K. Pearson).

Il s'agit du carré du coefficient de corrélation usuel. Il est particulièrement adapté au cas d'une variable indépendante numérique, mais convient également dans le cas où celle-ci est dichotomique (un facteur à deux modalités). Ce coefficient présente l'avantage de pouvoir être très facilement calculé à partir du t de Student (ou du F) et du degré de liberté (q) : $r^2 = t^2 / (t^2+q)$. Il a également un rapport direct avec l'indicateur d de Cohen : $\rho^2 = d^2 / [d^2+(1/pq)]$ (ρ étant le coefficient de corrélation parent, et p et q les proportions des deux populations).

Quand le facteur a plus de deux modalités, ou quand on s'intéresse à une relation non linéaire, il y a lieu de le généraliser par le coefficient suivant.

• Le coefficient η^2 (coefficient de différenciation de K. Pearson).

Il est défini comme le rapport de la somme des carrés associée à la comparaison étudiée (par exemple la comparaison globale sur tel facteur, ou d'interaction de tels et tels facteurs, etc.) à la somme des carrés totale. C'est également le carré du coefficient de corrélation multiple entre les valeurs de la variable dépendante et les k - 1 (pour k groupes) variables dichotomiques que l'on peut construire pour coder l'appartenance des sujets aux groupes (dans le cas de deux groupes, η^2 et r^2 sont équivalents).

Ce coefficient est une statistique descriptive, au sens défini par Rouanet *et al.* (1987). C'est-à-dire qu'il ne dépend que de la distribution des fréquences et donc reste inchangé si, les valeurs de la variable dépendante restant les mêmes, tous les effectifs sont multipliés par une même constante.

En tant qu'estimateur de l'effet parent, il présente la propriété d'être biaisé, surestimant systématiquement l'effet parent. Diverses formules de correction ont été proposées (voir, par exemple, Abdi, 1987; Richardson, 1996), mais elles ne sont pratiquement pas utilisées.

• Le coefficient ω^2 de Hays (1963).

Ce coefficient ω^2 est approprié quand le facteur est systématique. Dans le cas d'un facteur aléatoire on parlera du coefficient ρ^2, de Fisher, défini de manière analogue et encore appelé coefficient de corrélation intra-classe.

Il caractérise l'effet relatif théorique de la comparaison étudiée, c'est-à-dire l'effet dans la population parente considérée. Il est défini comme le rapport de la variance (théorique) associée à la comparaison étudiée à la variance (théorique) totale.

Au niveau d'un échantillon, on calcule un estimateur en formant le rapport des estimateurs des variances en question, ce qui donne un estimateur biaisé de l'effet parent (le rapport de deux estimateurs non biaisés n'étant pas lui-même non biaisé). Cet estimateur n'est pas une statistique descriptive, là n'est d'ailleurs pas son but, car il est sensible aux variations d'effectifs, toutes choses égales par ailleurs. De plus, il peut fournir une valeur négative, ce qui est le cas quand le F correspondant de l'analyse de variance est inférieur à 1. La pratique usuelle, et adoptée par Hays, de remplacer alors cette valeur par zéro pose problème (Vaughan et Corballis, 1969). Par ailleurs, l'extension au cas de plans déséquilibrés (groupes de tailles inégales) soulève également des difficultés (Vaughan et Corballis, 1969).

L'estimateur de ω^2 fournit en général une valeur plus faible que celle de η^2. Dans le cas où l'on échantillonne toute la population, il n'y a pas de différence entre η^2 et ω^2. On notera encore que le carré du f de Cohen et le ω^2 de Hays ont même numérateur et ne diffèrent que par la référence à laquelle ils sont comparés : la variance "d'erreur" uniquement pour f^2 et la variance totale pour ω^2.

3.1.4. L'étude de tableaux de contingence

Dès la situation la plus simple de la comparaison de deux proportions de groupes indépendants, en plus des indicateurs naturels de la grandeur de l'effet (la différence ou le rapport des deux proportions) de nombreux

indicateurs plus élaborés ont été définis : rapport des cotes (*odds ratio*), coefficients *Q* de Yule, coefficient de corrélation pondéré ou équipondéré, *etc.*

Généralisant cette situation de base, un grand nombre de coefficients généraux ont été proposés pour mesurer le degré d'association entre deux variables catégorielles (on les trouve dans la plupart des grands logiciels statistiques), les plus connus étant le ϕ^2 et le coefficient de contingence de Cramér ($\phi/\sqrt{k}$, k étant le plus petit des deux degrés de liberté).

Dans le cas particulier de la relation entre deux variables binaires, Rosenthal et Rubin (1982) proposent de calculer le BESD, *Binomial Effect Size Display* (en fait ils le proposent plus généralement pour tous les cas où l'on compare deux groupes – typiquement un groupe expérimental et un groupe contrôle – la variable dépendante étant dichotomisée quand elle est numérique). L'effet est défini comme la différence des taux de "réussite" entre les deux groupes mais sur un tableau transformé de telle façon que les quatre effectifs marginaux soient tous égaux à 100, sous la contrainte que le r^2 (égal au ϕ^2 dans le cas d'un tableau 2×2) reste égal à celui d'origine. Cela réalise une sorte de calibrage. Mais si l'on peut admettre aisément une transformation des données consistant à donner autant de poids aux deux groupes (au moins quand il s'agit de groupes expérimentaux et contrôles), on ne voit pas comment légitimer une transformation identique de la variable dépendante. Aussi cette solution a-t-elle été fortement critiquée (Crow, 1991; McGraw, 1991; Strahan, 1991).

3.1.5. Effets bruts ou effets relatifs ?

Ces coefficients *f*, η^2, ω^2 sont des indicateurs relatifs; d'ailleurs dans la littérature anglo-saxonne le terme *effect size* (grandeur de l'effet) est presque toujours entendu comme "grandeur relative". En tant que tels, ils peuvent présenter le grand avantage de pouvoir comparer des effets portant sur des variables différentes et d'être parlants même quand on ne connaît pas grand-chose d'un domaine ou d'une variable; mais ils présentent aussi de sérieux désavantages.

Ainsi, dans le cas du *f*, l'écart-type d'"erreur" apparaissant seul au dénominateur, l'effet relatif augmente dès que, ce qui est tout de même souhaitable, cet écart-type d'"erreur" diminue, même si l'effet absolu reste très faible. De même, pour les "pourcentages de variance expliquée" (η^2, ω^2), un même facteur peut voir son importance augmenter d'une expérience à l'autre, simplement parce que la variabilité intra-groupe est mieux contrôlée. Par ailleurs, pour une même variable dépendante, la nature des facteurs retenus dans le plan d'analyse ou contrôlés influence le résultat et la part de variance expliquée par tel facteur n'existe pas dans l'absolu (Oakes, 1986, p. 64). Par exemple, admettons que les facteurs A et B aient des effets additifs. On s'intéresse à l'effet de A, mais dans un cas on fait varier simultanément les deux facteurs alors que dans un second cas on opère à un niveau fixé du facteur B. Toutes choses égales par ailleurs, la variance totale sera plus grande dans le premier cas et le coefficient (η^2 ou ω^2) sera plus faible. Le résultat peut encore être fortement affecté par le choix des niveaux des facteurs (Levin, 1967), la fidélité des mesures (O'Grady, 1982). Le seul fait que ces coefficients puissent s'exprimer comme un pourcentage est donc loin d'assurer leur comparabilité d'une étude à l'autre.

Par ailleurs, le calibrage de l'effet brut est le plus souvent réalisé par un écart-type "moyen", sous l'hypothèse d'homogénéité des variances des différents groupes. Mais si cette hypothèse n'est pas réaliste, Richardson (1996) considère qu'il devient alors très difficile de définir un effet relatif.

Rosenthal et Rubin (1982) reprochent au coefficient r^2, et par conséquent aux indicateurs relatifs en général, de ne pas donner une bonne image de l'importance "réelle" de l'effet. Ainsi ils indiquent qu'un r^2 de seulement 0.10 peut très bien correspondre à la diminution d'un taux de mortalité de 66% à 34%, diminution effectivement spectaculaire.

B. Lecoutre (1996, pp. 51-53) montre que dans le cas d'une comparaison intra-sujets l'inférence sur l'effet calibré par un écart type d'erreur ne répond pas à la question de l'importance de l'effet moyen, mais renvoie plus fondamentalement à la distribution des effets individuels. Ainsi, par exemple, dans la situation de base de la comparaison de deux traitements dans un plan $S\times T_2$, l'inférence sur le rapport δ/σ (où δ et σ sont respectivement la moyenne et l'écart-type de la variable "différence entre les deux traitements") permet en fait de répondre à la question de savoir si *dans la plupart des cas* la différence est de même sens que la différence moyenne (par exemple est positive). Pour associer la problématique de l'importance de l'effet à celle de la prise en compte des différences individuelles, il faut considérer un indicateur plus général, du type $(\delta\text{-}x)/\sigma$, qui se relie directement à la proportion des différences parentes supérieures à x.

D'autres auteurs insistent sur l'intérêt de rapporter tout simplement l'effet brut, au moins dans le cas d'un contraste (voir, par exemple, Vaughan et Corballis, 1969). De manière plus radicale, Oakes (1986, pp. 62-63) critique l'utilisation d'indicateurs relatifs car ils incitent le psychologue à ne pas "prendre au sérieux" les variables utilisées (et leurs unités), alors même que pour lui une tâche primordiale est justement de donner sens à ces variables.

3.1.6. Une étape indispensable mais insuffisante

La description de la grandeur des effets observés est certainement une étape indispensable, et il est utile de pouvoir disposer à la fois d'indicateurs des effets bruts et des effets relatifs, tout en restant conscient de leurs avantages et inconvénients respectifs. En ce qui concerne l'inférence, la simple estimation ponctuelle est incontestablement insuffisante. Ajouter une telle estimation au test de signification usuel d'un effet nul est certes un progrès, mais suggère fortement une généralisation qui reste impressionniste et présente des dangers réels. En particulier, on sait bien qu'un effet observé faible associé à un résultat non significatif est souvent perçu par le chercheur comme un cas "en faveur de l'absence d'effet vrai" (M.-P. Lecoutre, 1991), alors qu'il n'est souvent qu'un constat d'ignorance.

3.2. L'ÉTUDE DE LA PUISSANCE

L'étude de la puissance, ou de son complément, le risque de deuxième espèce, a souvent été recommandée : voir, par exemple, Binder 1963; Skipper *et al.*, 1967; LaForge, 1967; O'Brien et Shapiro 1968. Son défenseur le plus ardent en psychologie est Cohen (1962, 1969, 1990, 1992).

Rappelons que le concept de puissance statistique a été introduit par Neyman et Pearson (1933b) comme le complément de l'erreur de deuxième espèce β, et a un rôle explicite fondamental dans leur théorie (un test de Neyman-Pearson consiste à construire une région critique qui minimise β, ou de manière équivalente qui maximise la puissance, pour un α fixé). La puissance, pour une hypothèse simple H_i, est la probabilité que la statistique de test tombe dans la région critique, si H_i est vraie, et en conséquence est équivalente à $1-\beta_i$ (β_i désignant la probabilité de commettre une erreur de deuxième espèce étant donné H_i). Dans le cas d'hypothèses multiples, la puissance peut être calculée pour chacune des hypothèses simples, ce qui conduit à la *fonction de puissance*.

La puissance n'est donc bien entendu pas envisagée contre la théorie de Neyman et Pearson, puisqu'elle en est le cœur, mais elle est invoquée pour pallier les insuffisances de la théorie de Fisher au regard de problèmes de sensibilité. Ainsi Cohen est très critique à l'égard de Fisher, mais pas du tout à l'égard de la théorie de Neyman et Pearson, à laquelle, selon lui, rien n'est à reprocher fondamentalement : seuls les chercheurs sont à blâmer pour leurs abus d'utilisation.

3.2.1. La puissance peut être un guide utile pour la planification des expériences

La puissance peut apparaître comme un guide utile pour planifier des expériences de sensibilité suffisante et donner une estimation des effectifs nécessaires sous des hypothèses raisonnables, évitant ainsi une perte de temps et de moyens. Certes, pour arriver à ce résultat, il est souvent nécessaire d'utiliser des effectifs élevés, plus élevés que les effectifs habituels (voir, dans le chapitre 5, les études sur les publications réalisées en psychologie ou en médecine). Dans certains cas, ces effectifs peuvent même être impraticables, comme le rappelle Schmidt (1996). Ceci est d'autant plus le cas dans un plan d'expérience complexe où il faut assurer simultanément une puissance suffisante pour plusieurs comparaisons. C'est dans de tel cas que les solutions complémentaires que nous rappelons en 3.6. peuvent être un complément ou un substitut.

S'en préoccuper avant le recueil des données présente en outre l'avantage d'obliger le chercheur à réfléchir sur l'intensité des effets (et sur les variances) qu'il s'attend à observer :

> "A salutary effect of power analysis is that it draws one forcibly to consider the magnitude of effects." (Cohen, 1990, p. 1309).

Il est paradoxal à ce propos de constater que Chow (1991, 1996) critique l'importance accordée à la puissance, en particulier chez Cohen, justement en raison de l'accent porté sur la grandeur de l'effet qu'il juge non pertinent; paradoxal car cette attaque provient d'un défenseur acharné des tests traditionnels (*cf.*, par exemple, Chow, 1988, 1996).

3.2.2. Les problèmes posés par l'utilisation de la puissance pour interpréter les données

Il est indéniable que prendre en compte la puissance du test utilisé peut effectivement constituer un garde-fou pour éviter notamment des conclusions hasardeuses d'absence d'effet en cas de résultat non significatif. Mais, dès que l'on tente d'utiliser la puissance pour interpréter les données, des critiques

fondamentales s'imposent. Par définition, toutes les critiques développées contre la théorie neyman-pearsonienne, déjà évoquées dans la section 2.1., sont pertinentes ici et nous n'y reviendrons pas; simplement nous allons souligner quelques points particuliers.

- Déjà, remarquons que la notion même d'interprétation des données, sauf à lui donner un sens des plus pauvre, n'est pas pertinente dans le cadre neyman-pearsonien pour lequel seule l'action (résultant d'une décision) importe, les données ne jouant qu'un rôle d'intermédiaire pour la prise de décision.

- L'étude de la puissance, qui suppose que α et N sont fixés, est incompatible avec l'usage du seuil observé p (Fisher).

- On notera que le calcul de la puissance requiert, en plus de la connaissance de la valeur du paramètre sous l'hypothèse alternative, la connaissance de la variance, ce qui, en dehors des cas d'école, n'est pas réalisé dans la pratique. On a alors souvent recours à une estimation de la variance, et il ne s'agit donc plus que de la puissance estimée. Cette difficulté n'est en rien fondamentale, mais il en va autrement des suivantes.

- L'utilisation de la puissance a été souvent proposée pour interpréter un résultat non significatif : si elle paraît insuffisante au chercheur, celui-ci suspendra son jugement, sinon il acceptera éventuellement l'hypothèse nulle. D'une manière plus formelle, si l'on veut montrer qu'un effet, disons une différence de deux moyennes, est négligeable, c'est-à-dire inférieur en valeur absolue à une limite donnée, la "méthode de la puissance" peut être ainsi décrite : on utilise cette limite comme valeur du paramètre sous l'hypothèse alternative pour le calcul de puissance et, en cas de résultat non significatif, on conclut à la négligeabilité de la différence, arguant qu'on avait toute chance de la rejeter si elle ne correspondait pas à la réalité. Cohen (1990) qualifie ce raisonnement d'"'impeccable", et même un auteur critique à l'égard des tests comme Oakes admet que l'on recherche un résultat non significatif associé à une puissance élevée (en général supérieure à 80% ou 90%) pour "valider" l'hypothèse nulle (Oakes, 1986, p. 11).

Or le calcul de puissance ne dépend pas des données, de la valeur observée de la statistique de test, et il n'est pas pertinent une fois les données recueillies pour interpréter un résultat particulier (Goodman et Berlin, 1994). Ainsi Falk et Greenbaum (1995) rappellent qu'un résultat non significatif n'implique nullement que la probabilité de H_0 conditionnellement aux données soit forte (simplement cette probabilité est augmentée si β, le risque de deuxième espèce, diminue). De fait, la seule chose qui importe, dans la théorie de Neyman et Pearson, est de savoir si le résultat est significatif ou non; de la même façon que tous les résultats significatifs sont équivalents entre eux (du point de vue décisionnel), tous les résultats non significatifs le sont également. Si l'on a fixé $\alpha = 0.05$, peu importe que l'on ait, par exemple, $p = 0.06$ ou $p = 0.48$. On remarquera à ce propos que la "puissance observée" (celle calculée sous l'hypothèse que la valeur vraie du paramètre est précisément la valeur observée), et qui est parfois évoquée par des utilisateurs du seuil observé, est nécessairement inférieure ou égale à ½, donc peu informative.

Il en résulte que cette méthode de la puissance possède des propriétés formelles indésirables, qui font qu'elle n'est pas recommandable. Ainsi Schuirmann (1987) l'a discutée dans le cadre du "problème de l'équivalence" en pharmacologie, où par définition deux traitements sont équivalents (en moyenne) si la différence des moyennes est inférieure en valeur absolue à une limite donnée (ce qui est vu comme l'opérationalisation du problème d'acceptation de l'hypothèse d'égalité des moyennes, la valeur zéro de l'hypothèse étant remplacée par un intervalle). Schuirmann arrive à la conclusion que l'étude de la puissance est inadéquate pour traiter du problème de l'équivalence :

i) À degrés de liberté (effectifs) constants, plus la précision expérimentale augmente et plus la valeur observée doit être faible (et tendre vers 0, à la limite) pour pouvoir conclure à l'équivalence. On aboutit ainsi au paradoxe qu'une même valeur observée, qui permet de conclure à l'équivalence à un niveau de précision donné, peut très bien amener à rejeter l'équivalence si la précision expérimentale est améliorée.
ii) Comme l'hypothèse nulle réellement testée est celle d'une différence strictement nulle et non celle d'un intervalle négligeable, le risque α nominal ne correspond au risque réel que si la vraie différence est strictement égale à zéro. Si cette différence vraie, tout en restant négligeable, tend vers la limite de l'intervalle d'équivalence, le risque de rejeter à tort l'équivalence tend, par construction, vers la puissance, donc s'accroît considérablement[11]. Par ailleurs, si cette différence vraie s'écarte de zéro, toujours en restant négligeable, le risque de rejeter l'équivalence augmente fortement quand la largeur de l'intervalle d'équivalence, relativement à l'écart-type, devient très grande (c'est-à-dire quand la précision est très bonne).

[11] La vitesse à laquelle le risque α réel tend vers la puissance est alors déterminante pour la qualité du test; mais cet aspect n'est pas étudié par les défenseurs de l'approche de la puissance.

iii) De plus, quand l'effectif augmente, la probabilité de conclure à l'équivalence alors qu'elle est fausse augmente également.

3.3. L'INTERVALLE DE CONFIANCE

De très loin, la solution la plus souvent mentionnée est l'utilisation de l'intervalle de confiance, au sens de Neyman et Pearson (voir, par exemple, Natrella, 1960; Nunnally, 1960; Rozeboom, 1960; Grant, 1962; LaForge, 1967; Carver, 1978; Oakes, 1986, Casella et Berger, 1987; Evans *et al.*, 1988; Cohen, 1994; Loftus et Masson, 1994; Schmidt, 1996). On ne trouve guère que Cornfield (1966), Johnstone (1988), Frick (1995a) ou Kadane (1995) pour englober l'intervalle de confiance dans la même critique que le test. Il est symptomatique qu'apparaissent maintenant dans des journaux scientifiques des éditoriaux mettant en cause la pratique des tests et/ou prônant un développement de l'utilisation des intervalles de confiance, même si, pour l'instant, cela concerne essentiellement le domaine médical (par exemple, Lutz et Nimmo, 1977; Rothman, 1978; Berry, 1986; Evans *et al.*, 1988; Braitman, 1988, 1991; Loftus, 1993; Falissard et Landais, 1995).

Malgré ce consensus des méthodologistes en faveur des méthodes de confiance, elles sont encore largement sous-employée comme le note Rouanet (1991b, p. 33), même s'il semble y avoir quelque évolution. Ainsi il n'est plus rare de voir mentionnés des intervalles de confiance, en appoint des tests, dans des journaux comme le *Journal of Abnormal Psychology*.

En premier lieu, rappelons qu'il existe une correspondance directe entre l'intervalle de confiance et le test (voir, par exemple, Lehmann, 1959, pp. 78-83) : un intervalle de confiance $1 - \alpha$ est généralement construit comme l'ensemble des valeurs qui, si elles étaient choisies pour hypothèse nulle, conduiraient à un résultat de test non significatif au seuil (bilatéral) α.

L'interprétation correcte, dans le cadre fréquentiste, d'un intervalle de confiance $1 - \alpha$ est la suivante : si l'on répétait l'expérience un nombre infini de fois, $100 \times (1 - \alpha)$ % des intervalles calculés contiendraient la vraie valeur du paramètre. Le paramètre d'intérêt n'est pas probabilisé, ce sont les bornes de l'intervalle qui varient d'une répétition à l'autre. Pour une expérience particulière, l'utilisateur affirme que l'intervalle calculé contient effectivement le paramètre, ou décide d'agir comme si c'était le cas; ainsi, sur un très grand nombre de répétitions, cet utilisateur ne se trompera que $100 \times (1 - \alpha)$ % des fois (voir, par exemple, Neyman, 1952, pp. 196, 209-210).

En tant que méthode d'estimation de l'effet, il présente déjà l'avantage de recentrer la discussion sur celui-ci. En tant qu'intervalle, par rapport au test usuel d'une hypothèse ponctuelle, il invite à prendre en considération la possibilité d'autres valeurs que celle choisie pour hypothèse nulle. Un autre avantage souvent mis en avant, est qu'un intervalle apporte une information directe sur la variabilité en jeu, sur la précision des résultats, alors que dans une statistique de test, comme le *t* de Student, par exemple, cette variabilité reste cachée dans le dénominateur.

On peut également calculer l'effectif nécessaire pour arriver à un intervalle de largeur donnée (pour autant qu'on ait une idée de la variance en jeu).

Cependant, l'intervalle de confiance pose, lui aussi, bien des problèmes.

3.3.1. Les mêmes critiques que les tests de signification

La correspondance avec le test (neyman-pearsonien) a pour conséquence que la plupart des critiques développées contre ce dernier s'appliquent autant à l'intervalle de confiance : manque de naturel, recours à la notion de répétition à l'infini, caractère décisionnel, rôle des valeurs non observées (en énonçant que 95% des intervalles *possibles* contiennent le paramètre, on fait référence à des intervalles qui ne sont pas observés), etc.

3.3.2. Des conclusions surprenantes

Il est des cas où l'intervalle de confiance amène à des conclusions qui peuvent sembler surprenantes. Citons en quelques exemples :

- Il est possible de trouver une borne négative pour un paramètre positif, telle une variance (Scheffé 1959, p. 229).

• Dans le cas du rapport de deux moyennes normales (régression, par exemple), pour un coefficient $1 - \alpha < 1$, il est possible de trouver comme solution $]-\infty , +\infty[$ (Lehmann, 1959, pp. 182-183). Associer un coefficient de confiance d'au moins 95% à un intervalle certain, c'est montrer une prudence pour le moins excessive.

• Supposons que l'on dispose pour le paramètre à estimer θ de deux observations aléatoires tirées d'une distribution uniforme sur l'intervalle $[\theta - 0.5 , \theta + 0.5]$. Soit y_1 la plus petite valeur tirée et y_2 la plus grande. On montre que $[y_1 , y_2]$ est un intervalle de confiance 50% pour θ. Mais il peut très bien arriver que l'on observe $y_2 - y_1 \geq 0.5$, auquel cas il est *certain* que $y_1 < \theta < y_2$, bien que la confiance ne soit que de 50% (Bernardo et Smith, 1994, pp. 468-469).

Cornfield (1966), comme Bernardo et Smith (1994), considèrent ces critiques comme dévastatrices pour la méthode. En fait, ces situations ne font que mettre clairement en évidence les différences d'approche entre la théorie de Neyman et Pearson et celle des bayésiens ou de Fisher, car ces paradoxes apparents proviennent du fait que l'intervalle de confiance, tout comme le test, n'est pas conditionnel aux valeurs observées mais conditionnel au paramètre théorique (quelle que soit sa valeur).

3.3.3. Des intervalles qui ne sont pas toujours disponibles

D'un point de vue pratique, en dehors des situations élémentaires, les intervalles de confiance ne sont pas toujours disponibles; il existe même des tests qui ne peuvent être transposés sous forme d'intervalle de confiance; en particulier dès que la situation est un peu complexe (Natrella, 1960; Bernardo et Smith, 1994). À ce propos, Rouanet (1991b, p. 7) évoque la "rouerie des conseillers statistiques" qui recommandent d'user davantage de méthodes qu'ils savent non forcément disponibles.

3.3.4. Est-ce le bon intervalle ?

Une critique méthodologique concerne les intervalles de confiance usuels qui sont construits autour de la valeur estimée du paramètre. Ainsi l'intervalle de confiance usuel pour une différence moyenne δ est un intervalle symétrique centré sur la différence observé d. Il n'est donc pas directement adapté dans le cas où l'on veut montrer que cette différence est négligeable. S'agissant de montrer que la différence est, sinon nulle, du moins proche de zéro, ceci requiert un intervalle de confiance centré sur zéro (du type $[-x , +x]$ avec $x > 0$), et non sur la valeur particulière d observée (soit encore un intervalle pour la valeur absolue de δ). La construction d'un tel intervalle, ou, ce qui revient au même, la construction d'un test de l'hypothèse nulle $H_0 : |\delta| > x$ (que l'on veut rejeter) contre $H_1 : |\delta| \leq x$, est possible. Mais elle a une longue histoire[12] qui révèle de nouvelles difficultés.

Dans une courte note, Bartko (1991) fournit aux psychologues des références de biostatistique, domaine où l'on s'est préoccupé du problème depuis longtemps, dans le cadre des essais cliniques de "bioéquivalence" en pharmacologie. Serlin et Lapsley (1993) traitent de la validité approximative des hypothèses (le principe du "good enough") et proposent au psychologue une procédure de test. Cette solution paraît s'imposer dans le cadre de la théorie de Neyman et Pearson puisqu'il s'agit du test uniformément plus puissant (du moins si la variance σ est connue ou si, comme le font Serlin et Lapsey, on teste $H_0 : |\delta/\sigma| > x$ contre $H_1 : |\delta/\sigma| \leq x$). Mais les auteurs semblent ignorer qu'elle a été proposée par ailleurs à différentes reprises (notamment Bondy, 1969; Anderson et Hauck, 1983; Rocke, 1984; Fowler, 1984, 1985; Wellek et Michaelis, 1991), et qu'elle a toujours été abandonnée, en raison de propriétés "indésirables" qui la rendent inacceptable. En particulier, avec cette procédure, quand la différence observée d est nulle, le seuil de signification observé p est toujours nul, et on conclut donc à un effet négligeable à n'importe quel seuil α, quelles que soient la valeur x et la précision expérimentale. De plus la région de rejet de l'hypothèse nulle n'est pas convexe, et en conséquence il est possible de conclure $|\delta| \leq x$ alors même que la différence observée est extérieure à cet intervalle (Schuirmann, 1987). Plus encore, cette procédure est encore applicable dans le cas d'un intervalle centré sur n'importe quelle valeur Δ fixée à l'avance (c'est-à-dire qui correspond à une hypothèse nulle $H_0 : |\delta - \Delta| > x$). Mais différents tests peuvent alors conduire à des résultats incompatibles, et aboutir à des paradoxes tels que pouvoir conclure $|\delta| \leq 0.5$, alors que pour les mêmes données on ne peut pas conclure $-0.7 \leq \delta \leq 0.51$ (Schervish, 1995, p. 252). Bien entendu, si on construit l'intervalle de confiance associé à cette procédure de test, ces propriétés indésirables se retrouvent,

[12] Hodges et Lehmann ont abordé cette question dès 1954, mais en considérant le problème inverse : tester l'hypothèse nulle $|\delta| < x$. À cette occasion, remarquons que les procédures de test fréquentistes dont il est question dans cette sous-section sont en fait des amalgames : le cadre technique est totalement neyman-pearsonien, mais le statut des hypothèses relève de Fisher (on ne peut que rejeter une hypothèse, l'hypothèse intéressante est identifiée à l'hypothèse alternative).

et on constate par exemple que la largeur de l'intervalle varie de façon non monotone avec les effectifs (par exemple on peut avoir un intervalle plus court pour n=4 que pour n=16, *cf.* B. Lecoutre, 1996, p. 225).

Cependant des solutions plus raisonnables ont pu être proposées. Ainsi Rogers *et al.* (1993) recommandent aux psychologues une procédure d'usage courant dans le domaine de la bioéquivalence. Dans ce cas on accepte l'hypothèse alternative $H_1 : |\delta| < x$ (donc l'équivalence) si l'intervalle de confiance usuel à 1-2α (et non 1-α) est inclus dans l'intervalle [-x , +x]. Sous cette forme, elle a notamment été proposée par Westlake (1981) et justifiée par Deheuvels (1984). Elle revient encore à conclure $|\delta| < x$ si chacun des deux tests unilatéraux $H_{01} : \delta = -x$ *vs* $H_{11} : \delta > -x$ et $H_{02} : \delta = x$ *vs* $H_{12} : \delta < x$, est significatif au seuil α). Cette procédure a été discutée (sous l'appellation de "*two one-sided tests*"), notamment par Schuirmann (1987).

Bien entendu les intervalles de confiance associés à cette procédure sont plus larges que dans le cas précédent du test uniformément plus puissant, et ont une probabilité de couverture plus grande que 1-α. La recherche de solutions "intermédiaires" est toujours une question d'actualité (Berger et Hsu, 1996).

3.3.5. Les abus d'interprétation : une situation paradoxale

Peut-être plus encore que les test traditionnels, les intervalles de confiance donnent lieu à des erreurs ou abus d'interprétation, conduisant à une situation paradoxale.

Rappelons en effet que, dans la conception fréquentiste, les bornes observées de l'intervalle de confiance pour l'échantillon (unique) dont on dispose ne sont interprétables qu'en référence à l'ensemble de tous les intervalles qu'on aurait pu observer (mais que l'on n'a pas observé). Supposons pour fixer les idées que l'on ait obtenu l'intervalle de confiance 0.95 [1.58 , 2.64] pour une différence de moyenne δ. Formellement, les bornes de l'intervalle de confiance pour le paramètre δ sont des grandeurs *aléatoires*, qui varient d'un échantillon à un autre. L'interprétation *correcte* de l'intervalle de confiance 0.95 est alors : "95% *des intervalles calculés sur l'ensemble des échantillons possibles (tous ceux qu'il est possible de tirer) contiennent la vraie valeur* δ". Mais cet énoncé est *conditionnel à* δ : il ne dépend pas des observations et est déterminé avant leur recueil. En fait dans ce cadre de justification, les seules probabilités envisagées sont les probabilités d'échantillonnage conditionnelles à δ. En revanche les valeurs possibles du paramètre *ne peuvent pas être probabilisées* : les bornes obtenues pour l'échantillon observé étant ici [1.58 , 2.64], l'événement "1.58<δ<2.64" est vrai ou faux (puisque δ est fixé), et nous ne pouvons pas lui attribuer de probabilité (sinon 1 ou 0). Il est donc *illégitime* d'écrire "*Pr*(1.58<δ<2.64)=0.95" ou d'énoncer que "*il y a* 95% *de chances que la différence inconnue* δ *soit comprise entre* 1.58 *et* 2.64".

Or, il est très courant de trouver des énoncés du type suivant :

> *La probabilité que la vraie valeur du paramètre soit comprise entre x et y (les bornes particulières calculées à partir des données observées) est 1 -* α.

qui est un abus, sinon une erreur, d'interprétation dans le cadre fréquentiste, et ceci même chez des utilisateurs avertis ou chez des statisticiens (*cf.* B. Lecoutre, 1997). C'est de cette manière que Cl. Robert, par exemple, présente l'intervalle de confiance dans un manuel récent :

> "Par exemple, si dans un sondage de taille 1000, on trouve [fréquence] = 0,613, la proportion π_1 à estimer a une probabilité 0,95 de se trouver dans la fourchette : […] ≈ [0,58 ; 0,64]" (Cl. Robert, 1995, pp. 221-222)

Si cet abus est si fréquent, c'est tout simplement qu'il est parfaitement naturel, comme Kadane (1995), par exemple, le relève :

> "Again it is not clear why such a set [l'intervalle] should be of interest unless one makes the *natural error* of thinking of the parameter as random and the confidence set as containing the parameter with a specified probability. Again, this is a statement only a Bayesian can make, although confidence intervals are *often so misinterpreted*. I find the classical quantities useless for making decisions and believe that they are widely misinterpreted as Bayesian because the Bayesian quantities are more natural." (Kadane, 1995). (Italiques ajoutés.)

Il n'est donc pas étonnant qu'un bayésien comme Phillips succombe à cet abus, qu'il commet plusieurs fois (par exemple, Phillips, 1973, pp. 319, 323). Il en vient à résumer la différence entre l'intervalle de crédibilité (méthode bayésienne dans laquelle le paramètre à estimer est probabilisé) et l'intervalle de confiance de la façon suivante :

> "A non-Bayesian states that there is a 95% chance that the [obtained] confidence interval contains the true value of the population mean. A Bayesian would say there is a 95% chance that the

> population mean falls between the obtained limits. One is a probability statement about the interval, the other about the population parameter." (Phillips, 1973, p. 335)

Autrement dit, la différence entre les méthodes se ramènerait à la différence entre “l'intervalle contient la moyenne” et “la moyenne est contenue dans l'intervalle” qui ne seraient pas des énoncés sémantiquement équivalents ! Commettant lui-même l'abus, Phillips gomme toute différence entre les deux méthodes et il n'a plus comme solution pour essayer de justifier une différence, que de poser la distinction entre les formes actives et passives de la phrase en lieu et place de la distinction relative à l'objet de la probabilité. On se demande ce que peut y comprendre le lecteur non averti, et l'on conçoit qu'il se dise que s'il n'y a pas de réelle différence entre intervalle de confiance et intervalle de crédibilité, alors, autant en rester au premier.

On retrouve cette même opposition entre les formes actives et passives de l'énoncé chez Nunnally (1960, p. 647 et note de bas de page), mais cette fois pour différencier (dans son esprit) un énoncé fréquentiste correct d'un énoncé fréquentiste abusif : “la probabilité est de 0.99 que la vraie valeur soit entre 0.30 et 1.00” signifierait autre chose que “la probabilité est de 0.99 que l'intervalle de 0.30 à 1.00 contienne le paramètre” !

C'est encore certainement cet abus qui explique que la théorie de l'intervalle de confiance soit invoquée par des auteurs aussi critiques à l'égard de la théorie de Neyman et Pearson que Rozeboom (1960) par exemple.

Dans la pratique, l’utilisateur a donc le choix entre trois attitudes : 1) conserver le cadre de justification fréquentiste de l’intervalle de confiance et se satisfaire de l’interprétation “correcte”; 2) conserver ce cadre tout en adoptant l’interprétation bayésienne alors “erronée”; 3) adopter explicitement le cadre de justification bayésien. Comme nous l’avons dit, tout montre que la majorité des utilisateurs adoptent actuellement la deuxième attitude. On notera d’ailleurs que la même situation apparaît dans le cas du test de signification, pour lequel le seuil de signification est souvent interprété comme la probabilité “*que l'hypothèse nulle soit vraie*” alors qu'il est la probabilité *conditionnelle* “de rejeter [à tort] l'hypothèse nulle si cette hypothèse est vraie”. On peut donc penser que ce sont paradoxalement leurs interprétations bayésiennes sauvages qui rendent ces procédures populaires. Et, comme le dit Rouanet (Rouanet *et al.*, 1991, p. 43) à ce propos, “tolérer l’erreur, quelle perspective peu exaltante !”.

3.4. LES MÉTHODES DE VRAISEMBLANCE

Un certain nombre d'auteurs ont prôné l'utilisation de la fonction de vraisemblance, et particulièrement de rapports de vraisemblance (ainsi Rozeboom, 1960; Edwards *et al*., 1963; Edwards, 1965; Cornfield, 1966; Wilson *et al*., 1967; Winkler 1974; Oakes, 1986).

Si Winkler (1974) conseille de rapporter la fonction de vraisemblance toute entière, le plus souvent il est question de présenter plutôt le rapport de vraisemblance. Dans le cas simple où deux hypothèses ponctuelles H_0 et H_1 sont en cause, la méthode du rapport de vraisemblance consiste à calculer le rapport des densités de probabilité de la statistique observée (x) sous H_0 et sous H_1, $f(x|H_0) / f(x|H_1)$. On peut éventuellement retenir H_0 ou H_1 selon que ce rapport est supérieur ou inférieur à une constante choisie arbitrairement (un, par exemple, si l'on ne privilégie aucune hypothèse), mais plus simplement s'en tenir à ce rapport qui exprime, au vu des résultats, les “chances” d'une hypothèse relativement à l'autre (et qui peut être considéré comme une solution de rechange à la probabilisation des hypothèses). Ce rapport apparaît comme objectif car il présente l'avantage de ne faire intervenir ni probabilités *a priori*, ni éléments non observés (il n'est plus question de la probabilité d'obtenir une statistique *au moins* aussi grande que celle obtenue) et d'être symétrique par rapport aux hypothèses (on peut permuter le rôle de celles-ci sans modifier la conclusion; ce qui n'est pas le cas dans la théorie de Neyman et Pearson où la permutation ne s'applique pas aux risques). C'est pourquoi Cornfield (1966), par exemple, soutient que le rapport de vraisemblance est approprié pour juger de la force d'évidence des données, au moins dans ce cas simple. Seulement, il est bien rare, dans la pratique, que le psychologue soit confronté au test de deux hypothèses ponctuelles (Wilson *et al*., 1967). La méthode peut être étendue au cas d'hypothèses composées, mais nécessite alors l'introduction des probabilités *a priori* (comme le souligne d'ailleurs l'appellation “facteur de Bayes” par laquelle on désigne la généralisation du rapport de vraisemblance).

Ces propositions sont souvent associées aux méthodes bayésiennes; elles sont d’ailleurs souvent préconisées par les mêmes auteurs.

3.5. LES MÉTHODES BAYÉSIENNES

Si la théorie bayésienne est très utilisée comme modèle de comportement du sujet dans des situations d'incertitude (comme le soulignent par exemple Winkler, 1974, ou B. Lecoutre, 1994) et apparaît également comme modèle dans d'autres situations (par exemple à propos de la mémoire ou de la catégorisation, *cf.* Anderson, 1991), elle connaît beaucoup moins de succès auprès des psychologues en tant que méthode

d'inférence statistique. Un certain nombre d'auteurs conseillent pourtant l'utilisation de méthodes bayésiennes en psychologie (Rozeboom, 1960; Edwards *et al*. ,1963; Edwards, 1965; Bakan 1966; Cornfield, 1966; Wilson *et al*., 1967; Winkler 1974; Rouanet *et al.*, 1976, 1978; Hoc, 1983; Rouanet et Lecoutre, 1983; B. Lecoutre, 1984a; et l'ensemble de l'ouvrage édité par Rouanet *et al.*, 1991/1997).

3.5.1. Un changement de l'objet de la probabilité

La méthode bayésienne consiste à calculer, au moyen du théorème de Bayes, la distribution *a posteriori* pour le paramètre auquel on s'intéresse, à partir des données observées (et d'un modèle d'échantillonnage associé) et des probabilités *a priori* sur le paramètre. Elle suppose donc, en général, de considérer la probabilité comme une mesure de l'incertitude relative à la valeur du paramètre inconnu. Portant directement sur l'effet d'intérêt (la valeur du paramètre retenu), elle fournit une solution directe aux questions de négligeabilité ou de notabilité de la grandeur de l'effet (voir, en particulier, B. Lecoutre, 1984a; Rouanet, 1996).

3.5.2. La question de l'objectivité

Bien sûr la nécessité, dans l'approche bayésienne, de spécifier des probabilités *a priori* a été, et est encore, beaucoup critiquée comme introduisant un élément éminemment subjectif. En particulier, il est certain que l'hostilité, même si elle doit être nuancée, de personnages aussi prestigieux que Fisher et Neyman (*cf.* 1.4.) a été d'un poids considérable. L'approche non informative, dite aussi standard ou encore fiducio-bayésienne selon la terminologie forgée par Rouanet *et al.* (1976), qui consiste à se placer dans un état d'ignorance *a priori* sur la valeur du paramètre peut être vue comme une réponse à cette critique (Rozeboom exprime une idée voisine dans l'importante note technique de son article de 1960). Par ailleurs, le poids de la distribution *a priori* dans la distribution *a posteriori* diminue d'autant que la masse des données s'accroît (du moins si l'on ne prend pas des cas extrêmes); deux chercheurs partant de distributions *a priori* différentes s'accorderont donc sur leurs conclusions si les données sont suffisantes. Enfin, si l'objectivité des méthodes traditionnelles de test se réduit finalement à leur pouvoir de communication, comme on l'a vu plus haut, on ne voit plus bien ce qui peut être reproché de plus, sur ce plan, aux tenants de l'approche bayésienne, les caractéristiques (famille, paramètres) d'une distribution *a priori*, même subjective, pouvant être aussi facilement communiquées que le choix d'un seuil de signification.

3.5.3. D'autres arguments à leur encontre

Mais, au delà de cette question de l'objectivité, il existe d'autres arguments qui expliquent la désaffection pour les méthodes bayésiennes.

Ainsi, l'important article d'Edwards *et al*. (1963), qui est en fait un véritable manuel à l'intention des psychologues, nécessite déjà, de par sa technicité, une bonne volonté certaine de la part du lecteur et l'invite à réfléchir plus qu'il ne lui fournit une technique “toute faite” d'application immédiate. Par ailleurs il se place sur le terrain des tests traditionnels en envisageant, à la manière de Jeffreys (1961), le test d'un hypothèse nulle approximative (dans le sens d'un intervalle autour de cette hypothèse) et la méthode aboutit à moins souvent rejeter cette hypothèse nulle que ne le font les méthodes traditionnelles. On ne s'étonnera donc pas que le psychologue ne voit pas là un vrai changement de problématique et ne soit pas attiré par une approche qui lui fournit moins fréquemment les résultats significatifs dont il a besoin pour pouvoir être publié.

Par ailleurs, il a été reproché aux théoriciens bayésiens de trop se tourner vers des problèmes de théorie de la décision, au détriment des problèmes de l'inférence statistique (voir, par exemple, Winkler, 1974). On pourra d’ailleurs regretter que cette tendance aille en s'accentuant (voir, par exemple, les ouvrages de Berger, 1985, et Ch. Robert, 1992) et puisse donner l'impression que les méthodes bayésiennes se réduisent au champ de la théorie de la décision. Ainsi Ch. Robert définit l’inférence statistique comme devant fournir

> "un critère d’évaluation, qui décrit les conséquences de chaque décision en fonction des paramètres du modèle" (Ch. Robert, 1992, p. 41).

En 1974 Winkler regrettait également l'absence d'un grand traité bayésien. Les manuels disponibles et s'adressant explicitement aux psychologues, tel celui de Philipps (1973) (ou même l'article d'Edwards *et al*. précédemment évoqué) sont maintenant anciens. De plus ils sont souvent limités à la présentation de cas élémentaires (inférence sur une moyenne ou sur la différence de deux moyennes) et n'apportent donc pas de solutions pour les plans d'expérience un peu complexe généralement utilisés par les psychologues. Cette situation est d'autant plus paradoxale que les méthodes bayésiennes sont plutôt associées par les chercheurs aux

cas complexes; ainsi on entend souvent des commentaires du type “je n'ai pas besoin de [fiducio-]bayésien parce que je ne fais que des choses simples comme la comparaison de deux moyennes”. Pourtant, des solutions puissantes et pratiques sont possibles, comme on le verra plus loin (*cf.* 3.5.5.).

3.5.4. Un objet de rejet

En fait, jusqu’à présent les méthodes bayésiennes se sont heurtées à la méfiance, sinon à l’opposition de principe, des chercheurs, qui pensent qu’elles sont trop compliquées à utiliser et trop subjectives pour être scientifiquement acceptables. Cette attitude est illustrée par le commentaire de Falk et Greenbaum (1995) qui commencent par reconnaître l’intérêt potentiel de l’inférence bayésienne :

> "Bayesian inference might, in principle, fill the void created by abandoning significance-testing"

mais ajoutent aussitôt la restriction :

> "Implementation of Bayesian analysis, however, requires subjective assessments of prior distributions, and often involves technical problems." (Falk et Greenbaum, 1995, p. 92)

En conséquence, les méthodes bayésiennes pour l’analyse des données expérimentales ont été constamment écartées (par exemple : Wilson, Miller, et Lower, 1967; Frick, 1996) pour des raisons *a priori* qui sont de plus en plus injustifiées. Certains critiques, tel Chow (1996, chapitre 7) ignorent même les fondations de la statistique bayésienne et ses développements théoriques considérables qui remettent de plus en plus en question l’approche fréquentiste dans tous les domaines, tout autant que les débats méthodologiques déclenchés par ces développements (voir par exemple les numéros spéciaux de *Statistics in Medicine* 1993, vol. 12) et de *The Statistician* (1993, vol. 42).

3.5.5. Une disponibilité nouvelle

Des méthodes bayésiennes standard, objectives, bien adaptées à la spécificité de l’analyse des données expérimentales, sont maintenant disponibles et apparaissent être une solution de rechange prometteuse, même pour les chercheurs qui insistent sur la nécessité d’objectivité. Elles sont basées sur des définitions opérationnelles plus utiles que les procédures fréquentistes, et sont pleinement justifiées, au moins aussi objectives que les autres techniques statistiques (Berger, 1985, p. 110).

De plus, comme l'ont montré depuis plus de vingt ans en France les travaux initiés par Rouanet et Lépine, elles peuvent être utilisées aussi facilement que les tests de signification traditionnels (*cf.* Rouanet *et al.*, 1976; Rouanet *et al.*, 1978; Hoc, 1983; Rouanet et Lecoutre, 1983; B. Lecoutre, 1984a, 1985, 1996; Bernard *et al.*, 1985; Rouanet *et al.*, 1991/1997; Rouanet, 1996). Ces travaux sont maintenant rendus accessibles par la mise au point de méthodes de calcul et la réalisation de logiciels informatiques pour les situations complexes (*cf.* Poitevineau et Bernard, 1986; B. Lecoutre *et al.*, 1992; B. Lecoutre et Poitevineau, 1992; B. Lecoutre *et al.*, 1995; Grouin et Lecoutre, 1996; Bernard, 1997; B. Lecoutre et Charron, 1997; B. Lecoutre *et al.*, 1997a). L’utilisation de ces méthodes est d’ailleurs acceptée par les revues psychologiques (par exemple : Ciancia *et al.*, 1988; M.-P. Lecoutre, 1992; Clément et Richard, 1997, sans compter de très nombreux articles publiés en français).

Dans d’autres domaines, et tout particulièrement dans celui des essais cliniques en médecine et en pharmacologie, les méthodes bayésiennes sont de plus en plus développées et connues. Des propositions méthodologiques voisines de celles proposées dans les références précédentes ont été avancées pour leur utilisation de routine, par exemple par Spiegelhalter, Freedman et Parmar (1994). Ces auteurs suggèrent également d’utiliser des distributions *a priori* “sceptiques” ou “enthousiastes” pour éprouver la robustesse des conclusions (voir également B. Lecoutre, 1996, chap. 3; B. Lecoutre *et al.*, 1997b). Ils insistent sur le fait que la motivation pour utiliser la méthodologie bayésienne est davantage pratique qu’idéologique.

3.5.6. Les abus d'interprétation revisités à la lumière du cadre bayésien

Nous avons dit plus haut que les erreurs d'interprétation renvoyaient au cadre de justification adopté. Il est possible d'en justifier certaines en passant du cadre théorique traditionnel au cadre bayésien, et plus particulièrement au cadre fiducio-bayésien (ou bayésien standard).

Les tests de signification

Le test de signification trouve une réinterprétation naturelle si l'on choisit une distribution *a priori* non informative (cadre fiducio-bayésien).

Le cas privilégié est celui du test unilatéral pour une comparaison de moyennes (Edwards *et al.*, 1963; Pratt, 1965; B. Lecoutre, 1984a, 1984b, 1991; Casella et Berger, 1987) :

le seuil observé p est la probabilité que l'effet vrai soit du signe inverse de celui de l'effet observé.

Il devient alors licite d'interpréter p ou $1 - p$ respectivement comme la probabilité que l'hypothèse nulle (l'hypothèse composée $H_0 : \delta \leq 0$, par exemple) ou que l'hypothèse alternative (l'hypothèse composée $H_1 : \delta > 0$, par exemple) soit vraie, conditionnellement aux données observées. Mais, comme le remarque par exemple B. Lecoutre (1984a), la pauvreté de cette information devient flagrante car on ne fait que préciser la direction, positive ou négative, de l'effet.

Dans le cas d'un test bilatéral (de l'hypothèse nulle ponctuelle $H_0 : \delta = 0$), même si Berger et Sellke (1987) contestent l'utilité du seuil p qui surestime généralement la probabilité *a posteriori* $Pr(H_0|données)$, il est encore possible de réinterpréter ce seuil (B. Lecoutre, 1984a, 1984b, 1991) :

le seuil observé p est la probabilité que l'effet vrai soit extérieur à l'intervalle [0 ; 2×effet observé] (pour un effet observé positif).

Mais si cette réinterprétation reste simple, fournissant un intervalle de probabilité $1 - p$ pour le paramètre, il n'est plus justifié de parler de p comme de la probabilité *a posteriori* de H_0 (celle-ci étant d'ailleurs nulle dans un cadre bayésien non informatif, puisque H_0 est ponctuelle).

Dans le cas d'une comparaison à plusieurs degré de liberté, il existe encore une réinterprétation fiducio-bayésienne du seuil observé, généralisation du cas précédent : $1 - p$ est la probabilité que l'effet vrai se situe dans une hypersphère centrée sur l'effet observé et de rayon égal à la grandeur de celui-ci (*cf.* B. Lecoutre, 1984a, 1984b). Comme précédemment, p n'est pas la probabilité *a posteriori* de H_0.

L'intervalle de confiance

Dans la section 3.3. nous avons présenté l'abus fréquent consistant à attribuer une probabilité (le "niveau de confiance" $1 - \alpha$) à l'intervalle de confiance *particulier* calculé à partir des résultats. Le pendant bayésien de l'intervalle de confiance est l'intervalle de crédibilité (l'intervalle qui a une probabilité déterminée de contenir le paramètre, compte tenu des données). Or, dans des cas courants (notamment celui des intervalles usuels sur une moyenne ou la différence de deux moyennes) les limites de l'intervalle de crédibilité fiducio-bayésien de probabilité $1 - \alpha$ coïncident avec celles de l'intervalle de confiance traditionnel (neyman-pearsonien). Dans de tels cas l'abus précédent n'en est donc plus un si l'on adopte explicitement le cadre bayésien.

Par suite il est possible de considérer , comme le font M.-P. Lecoutre (1991) et B. Lecoutre (1991), que la plupart des chercheurs se comportent comme des "fiducio-bayésiens" naturels ou naïfs.

3.6. LES AUTRES PROPOSITIONS

Les propositions suivantes ont un statut différent. Ce ne sont pas, sauf les méta-analyses, des méthodes statistiques mais des moyens méthodologiques.

C'est en particulier dans les cas où l'obtention d'une conclusion recherchée à partir d'une seule expérience nécessiterait des effectifs impraticables qu'elles peuvent apporter des compléments, voire des substituts intéressants.

Notons que dans ce dernier cas les méthodes bayésiennes, en permettant d'intégrer explicitement des informations extérieures aux données dans l'analyse, constituent également une solution privilégiée pour pallier l'insuffisance de l'information expérimentale.

3.6.1. La répétition des expériences

Parmi les propositions souvent énoncées, on trouve celle de répliquer les expériences (voir, par exemple, Tullock, 1959; Bolles, 1962; Carver, 1978; Johnstone, 1988; Cohen, 1994). La reproductibilité des résultats étant un critère essentiel de la méthode expérimentale, il n'est pas étonnant que cette proposition ne soit pas critiquée.

3.6.2. La diminution des erreurs de mesure

Cohen (1990), entre autres, propose d'améliorer le contrôle des erreurs de mesure, ce qui permettrait de diminuer la variabilité associée et donc d'augmenter la précision des estimations. Mais ceci, bien qu'évidemment souhaitable, ne change rien au problème de fond (sauf à considérer qu'on atteint une précision totale et qu'il n'existe pas d'autre source de variabilité que celle examinée, ce qui est bien sûr irréaliste).

3.6.3. La manipulation des variables

Au delà de l'utilisation des tests statistiques, Prentice et Miller (1992) critiquent l'accent porté sur la grandeur de l'effet et défendent l'idée que l'importance d'un effet doit être montrée par sa robustesse au travers de manipulations des variables indépendantes ou dépendantes plutôt que par un critère statistique. Ainsi un effet sera important s'il se manifeste, même très faiblement, quand on opère une "manipulation minimale de la variable indépendante" ou quand on choisit une "variable dépendante difficile à influencer". Si cette approche méthodologique est intéressante, il faut néanmoins signaler que les auteurs fondent leur critique de l'étude de l'effet sur le fait que dans celle-ci on assimilerait importance (scientifique) d'un effet et valeur élevée, ce qui n'est pas le cas (nous avons déjà signalé que Lykken, 1968, a fourni un exemple dans lequel l'effet intéressant ne devait être ni élevé ni faible mais moyen). Par ailleurs leurs exemples sont loin d'être convaincants, le caractère "difficile à influencer" des variables dépendantes n'étant pas justifié et semblant posé *a priori*.

3.6.4. Les méta-analyses

La possibilité de combiner statistiquement les résultats de plusieurs études indépendantes par des méta-analyses est évidemment une perspective prometteuse. Un intérêt supplémentaire est que l'accent est souvent mis, dans ces méta-analyses, sur la grandeur des effets (cf. Glass, 1976; Smith et Glass, 1977; Schmidt, 1996). Mais les solutions proposées sont parfois caricaturales, comme celle de Rosenthal et Rubin (1979) qui ne fait que combiner les seuils p (en les transformant en "score z", valeur d'une distribution normale centrée-réduite, quels que soient le test original et les effectifs !) pour aboutir simplement à un nouveau test de signification, sous le prétexte que grandeur d'effet et seuils p sont corrélés et que les tailles d'échantillons sont supposés homogènes dans les études concernant un même domaine de recherche. De même, les "pourcentages de variance expliquée" étant non directionnels, en combiner plusieurs, lors d'une méta-analyse, par exemple, peut donner une fausse idée de la situation si les profils des résultats sont très différents d'une étude à l'autre (Richardson, 1996), le cas limite étant celui où deux effets à un degré de liberté sont de même intensité mais de signes opposés.

Mais la difficulté principale rencontrée est celle de la sélection des études incluses dans les méta-analyses, puisqu'il faut que les conditions expérimentales d'une étude à l'autre soient, sinon identiques, du moins suffisamment proches (Schmidt, 1996). Il faut en outre tenir compte des biais de publication des articles en faveur des résultats statistiquement significatifs, un sujet largement traité par ceux qui critiquent les test (*cf.* 2.1.10.) et évoqué encore récemment par Cohen (1994), mais curieusement ignoré par Schmidt. Il semble malheureusement très difficile, sinon impossible, de corriger ce biais, et la tentative de Rosenthal (1979) a d'ailleurs été fortement critiquée par Oakes (1986). En outre, l'effet pris en compte est la plupart du temps un effet relatif (ne serait-ce que parce que les variables dépendantes ne sont pas toujours exactement les mêmes), et le moyennage de tels effets pose des problèmes, la moyenne de rapports n'étant évidemment pas le rapport des moyennes.

Soulignons enfin que ces méta-analyses ont un statut particulier. Par définition, elles ne peuvent opérer que lorsqu'un corpus de résultats a déjà été constitué et n'ont aucun sens dans le cas d'une expérience prise isolément. En ce sens elles ne peuvent constituer réellement une solution de rechange au test; bien plutôt on retrouve à leur niveau la même problématique : soit réaliser les méta-analyses au moyen de tests traditionnels, soit au moyen d'autres solutions.

3.6.5. L'analyse fiduciaire

L'analyse fiduciaire de Fisher (évoquée en 1.4.) aurait sa place parmi les solutions de rechange, mais à quelques exceptions près (comme Lépine et Rouanet, 1975, B. Lecoutre et M.-P. Lecoutre, 1979, ou Hoc, 1983), elle a été purement et simplement ignorée. Ce sur quoi nous voulons insister ici, c'est sur l'injustice de certains critiques comme Cohen (1994) et Frick (1996) (qui sont par ailleurs incohérents dans leur position vis-à-vis des

théories de Fisher et de Neyman et Pearson, ainsi que nous l'avons signalé en 2.3.2.). Ils ont reproché à Fisher sa méthode de test pour ce qu'elle détourne l'utilisateur de la probabilité intéressante (*Pr*(*Hypothèse*|*Données*)), mais ont ignoré la solution de Fisher pour estimer la probabilité du paramètre (conditionnellement aux données) par sa méthode fiduciaire.

À titre anecdotique, on remarquera que le logiciel SAS, dont les auteurs ont pour politique de ne retenir que des méthodes statistiques incontestées, fournit pourtant explicitement, dans la procédure “Probit”, non pas des intervalles de confiance des paramètres mais bien des intervalles fiduciaires (suivant en cela Finney, D. J., 1971 : *Probit Analysis*. London: Cambridge University Press).

* * *

Si l’on s’en tient à l’intérêt méthodologique, seules les méthodes d’intervalle de confiance et, peut-être plus encore les méthodes bayésiennes paraissent devoir s’imposer comme véritables “challengers” des tests traditionnels. Elles incluent d’ailleurs dans leur application la mesure de la grandeur de l’effet observé qui est une étape indispensable, mais insuffisante, dans l’analyse des données. On ne niera pas pour autant l’intérêt d’autres solutions évoquées, et notamment la réplication des expériences dont l’intérêt est incontestable, ainsi que les méta-analyses qui sont d’ailleurs déjà utilisées mais qui ne sont qu’une procédure supplémentaire ajoutée à des analyses existantes.

CHAPITRE 4

QUELQUES OUVRAGES DE RÉFÉRENCE

Nous allons maintenant examiner, au travers de quelques ouvrages, comment les travaux normatifs des fondateurs sont traduits par des méthodologistes, psychologues et/ou statisticiens, en une “prescription” à l'usage directe de l'étudiant et/ou du chercheur en psychologie.

Cet examen ne peut être exhaustif (le nombre de ce type d'ouvrages est impressionnant) et le choix des livres sélectionnés est bien sûr arbitraire. Nous avons choisi de ne retenir que des ouvrages “incontournables”, dont les succès en librairie ont été évidents (à l'échelle des ouvrages universitaires, bien entendu). Il sera donc question des livres de Siegel, Winer, Hays et Kirk, en ce qui concerne les auteurs anglo-saxons, et de Faverge et Reuchlin en ce qui concerne les auteurs de langue française. Les ouvrages sont présentés dans l'ordre chronologique de leur parution. Quand deux dates sont indiquées, elles correspondent respectivement aux parutions de la première édition et de celle que nous avons consultée.

Chaque ouvrage est analysé selon la même grille :

La notion de probabilité.
Une conception particulière de la probabilité est-elle évoquée, implicitement ou explicitement ?

La présentation des tests.
S'agit-il de la présentation de la théorie de Fisher, de Neyman et Pearson, ou d'une présentation hybride ? Donne-t-on une règle pour la formulation des hypothèses et quel est le statut de l'hypothèse de recherche ?

La grandeur de l'effet.
Est-elle évoquée, et si oui comment ?

La présentation de l'intervalle de confiance.
Comment est interprété cet intervalle ? Le lien avec le test est-il précisé ?

Le calcul de N.
Est-il indiqué comment déterminer la taille d'un échantillon ?

Les exemples.
Comment, dans les exemples, sont appliqués les principes d'interprétation énoncés par l'auteur (particulièrement dans le cas de résultat non significatif) ?

Les abus.
En relève-t-on ?

4.1. J.-M. FAVERGE : MÉTHODES STATISTIQUES EN PSYCHOLOGIE APPLIQUÉE. (1950/1975)

L'ouvrage comprend trois tomes, sans références bibliographiques. La première édition du premier tome date de 1950, la septième, que nous avons consultée, date de 1975. Ce livre est plutôt destiné aux étudiants en psychologie et vise à leur présenter, davantage d'un point de vue appliqué que théorique, un ensemble de méthodes statistiques dont ils pourront avoir besoin. Il a donc un caractère généraliste et couvre la statistique descriptive aussi bien qu'inférentielle. Les procédures de test présentées sont réparties dans les tomes 1 et 2, le tome 3 étant surtout consacré à la psychologie mathématique.

La notion de probabilité

Il est remarquable que dans les deux premiers tomes l'auteur n'emploie jamais le terme *probabilité*, pour faire constamment référence à la notion de *fréquence* (on pourrait dire qu'il s'agit là par défaut d'une interprétation fréquentiste de la probabilité).

La présentation des tests

Les tests statistiques sont introduits dans le premier tome, après la présentation de méthodes descriptives (tracé d'histogrammes, *etc.*) dans un chapitre (VI) qui traite également de l'intervalle de confiance. Ils apparaissent sous la dénomination "épreuve de signification" ou "épreuve de nullité" (de la différence entre un paramètre et une valeur donnée, …), le terme de "nullité" renvoyant clairement, par les exemples donnés, à la notion de réfutation et non nécessairement à la valeur zéro d'un paramètre.

La présentation des tests peut être qualifiée d'hybride. En effet, à la manière de Fisher, qui n'est pas cité, on trouve :

- La détermination de l'hypothèse nulle par négation de l'hypothèse d'intérêt.
- La mise en garde (tome 1, p. 80) contre l'interprétation d'un résultat non significatif comme étant une démonstration de la véracité de l'hypothèse nulle. Une telle interprétation est explicitement qualifiée d'"erreur", l'hypothèse nulle ne pouvant qu'être, éventuellement, infirmée.
- L'absence de référence à un risque de 2ème espèce. Le seul risque qui soit mentionné est celui encouru quand on affirme que l'hypothèse nulle est fausse, et il n'est d'ailleurs pas qualifié de "1ère espèce".
- La référence à la notion de *sensibilité* du test à détecter un écart à l'hypothèse nulle, plutôt qu'à celle de puissance.

En revanche :

- Il semble que l'on travaille avec un seuil fixé (noté P), puisque l'on compare la statistique de test à une borne déterminée en fonction de ce seuil, sans qu'aucune allusion soit faite au seuil observé (il n'est pas calculé). Cependant il n'est pas dit clairement, ni n'apparaît dans les exemples, que ce seuil doive être choisi *a priori*.
- La référence à la sensibilité du test est complétée par un renvoi aux travaux de Neyman et Pearson (mais sans que soit donnée une référence précise), avec la remarque qu'un résultat non significatif dans le cas d'un test "sensible" permet pratiquement d'admettre l'hypothèse nulle, ce qui est quelque peu contradictoire avec la mise en garde évoquée précédemment.

L'auteur termine la présentation des méthodes de test et d'intervalle de confiance par un commentaire dans lequel il marque nettement sa préférence pour les problèmes d'estimation, réservant aux tests un rôle de garde-fou dans les cas où les échantillons sont faibles, tout en regrettant qu'ils puissent être, en psychologie, une excuse à des données peu nombreuses. Il évoque succinctement le problème de la fausseté, en psychologie, des hypothèses nulles ponctuelles pour encore limiter l'intérêt des tests.

La grandeur de l'effet

Aucune attention particulière n'est portée sur la grandeur de l'effet qui n'est pas évoquée dans le chapitre de présentation des tests et qui n'est pas spécialement commentée dans les exemples. Dans le deuxième volume, deux paragraphes sont dédiés respectivement au coefficient η^2 et au coefficient intraclasse de Fisher, mais dans des contextes spécifiques qui masquent la généralité de ces coefficients. Ainsi, le coefficient η^2, appelé improprement *rapport de corrélation* (celui-ci étant égal à η et non à son carré), n'est abordé que dans le cadre du pronostic, et le coefficient intraclasse, bien qu'évoqué dans un chapitre général sur l'analyse de variance, est réduit par sa présentation au problème de la fidélité d'une mesure. On notera par ailleurs que l'auteur, bien que distinguant facteurs aléatoires et systématiques, ne différencie pas le coefficient intraclasse du ω^2 de Hays (non mentionné), le premier étant supposé adapté aux deux types de facteurs.

La présentation de l'intervalle de confiance

La présentation de l'intervalle de confiance, qui précède celle du test, est conforme à la théorie fréquentiste et ne donne lieu à aucun abus d'interprétation. L'interprétation consiste à admettre que l'échantillon fait bien partie des 95% (par exemple) contenant effectivement le paramètre et à affirmer que celui-ci est compris dans l'intervalle observé. Il n'est pas fait mention des rapports entre test et intervalle de confiance.

Le calcul de N

La puissance des tests n'étant pas abordée, il n'est pas étonnant que le calcul *a priori* d'une taille d'échantillon ne soit pas présenté. Ce calcul n'est pas non plus présenté dans le cadre de l'intervalle de confiance.

Les exemples

Les exemples proposés, illustrations des procédures aussi bien que corrigés d'exercices, ne donnent pas lieu à une interprétation psychologique des résultats et l'on ne trouve que des conclusions du type “la différence est significative à P = 0.01” ou “la différence est non significative”.

Les abus

Le corrigé du problème n° 22 met à mal la mise en garde de l'auteur contre l'acceptation de l'hypothèse nulle. La question est de savoir si 12 coefficients de validité provenant de deux batteries de tests psychologiques sont homogènes. Or, ayant trouvé $\chi^2_{11} = 17.5838$, non significatif, l'auteur conclut (p. 158) que les 12 coefficients peuvent être considérés comme homogènes, c'est-à-dire qu'il accepte l'hypothèse nulle.

Plus grave est le fait que cette attitude se trouve implicitement légitimée dans un des paragraphes traitant de la corrélation intragroupe, où il est écrit :

> "On fait ainsi la seule hypothèse d'*homogénéité de la corrélation* que l'on a d'ailleurs *toujours* la précaution d'éprouver en calculant χ^2." (Faverge, 1975, tome 2, p. 182) (Seconds italiques ajoutés.)

Il est clairement sous-entendu ici que l'hypothèse d'homogénéité sera confirmée par un χ^2 non significatif. “Toujours” suppose une position de principe de la part de l'auteur (il n'est pas question de condition d'effectif, de puissance ou de sensibilité suffisante du test), en totale opposition avec la mise en garde précitée.

Cette interprétation est ultérieurement confirmée et rendue explicite, quand, à propos de la régression linéaire, Faverge écrit :

> "Comme nous l'avons dit, on a avantage à rechercher si une transformation de l'échelle des *x* peut conduire à un schéma linéaire, *c'est-à-dire à un F_2 non significatif*." (Faverge, 1975, tome 2, p. 268) (Italiques ajoutés.)

4.2. S. SIEGEL : NONPARAMETRIC STATISTICS FOR THE BEHAVIORAL SCIENCES. (1956)

Dans ce livre, Siegel cherche à promouvoir les méthodes dites “non paramétriques” qu'il estime particulièrement appropriées à l'analyse des données recueillies en sciences humaines. Le livre a connu un très large succès, bien au delà du seul lectorat étudiant, et l'on peut dire qu'à travers lui l'auteur a formé aux tests non paramétriques plusieurs générations de chercheurs en psychologie. La lecture du livre ne demande qu'un minimum de connaissances en statistiques et en mathématiques et en particulier ne suppose pas nécessairement de connaître les tests statistiques. Mais si aucune démonstration n'apparaît, l'auteur s'efforce de présenter le raisonnement sous-tendant les tests présentés. Ainsi qu'il est précisé dans l'introduction, l'ouvrage ne concerne que les méthodes de test et exclut les méthodes d'estimation.

L'auteur fournit une bibliographie assez importante et n'hésite pas à renvoyer à des publications purement statistiques. Ainsi deux ouvrages de Fisher sont cités (*Statistical Methods for Research Workers* et *The Design of Experiments*). En revanche, Neyman et Pearson ne sont cités que dans une remarque (p. 104) concernant la puissance du “test exact” de Fisher (remarque précisant “puissance au sens de Neyman et Pearson”), sans aucune référence. Aucune mention n'est faite du rôle de ces auteurs dans l'élaboration des théories des tests statistiques.

La notion de probabilité

La notion de probabilité, supposée connue, ne fait l'objet ici d'aucune définition ni d'aucun commentaire. Cependant, de par l'insistance de l'auteur sur l'importance de l'objectivité, en début du chapitre 2, il semblerait plutôt qu'il s'en tienne à la conception fréquentiste. Un autre indice en ce sens tient au fait que la modification de Tocher du “test exact” de Fisher (qui introduit une procédure “très fréquentiste” de décision par tirage au hasard pour obtenir un seuil α juste à la valeur souhaitée) est décrite sans remarque particulière (p. 103).

La présentation des tests

La procédure générale de tout test statistique est explicitée dans le deuxième chapitre qui suit immédiatement l'introduction. Le test y est présenté comme un moyen de prendre une décision objective quant à la question de savoir si les données confirment l'hypothèse de recherche. La procédure de test est alors décrite comme une succession de six étapes, allant du choix de l'hypothèse nulle à la prise de décision. La théorie statistique de la décision et les fonctions de coût sont évoquées dans une note de bas de page (renvoyant à

Savage et Wald, entre autres), mais l'auteur doute de son applicabilité en sciences du comportement en raison de la difficulté à définir ces fonctions.

La plupart des éléments de cette présentation sont caractéristiques de l'approche de Neyman et Pearson, comme l'est déjà le choix du terme “décision” :

- Les deux risques d'erreurs sont mentionnés et la notion de puissance est largement évoquée, mais de façon assez confuse et peu rigoureuse, sans que soit précisé son calcul.
- Les choix du risque de première espèce α et de la taille N de l'échantillon doivent être faits avant le recueil des donnés. Cela découle de l'ordre des étapes indiquées, et est clairement explicité ensuite (“In advance of the data collection”, p. 8).
- Le seuil observé (appelé “probabilité associée”) ne sert que par comparaison à α et ne reçoit pas d'interprétation.
- Il est constamment fait référence à la puissance pour ce qui concerne le choix des divers tests.

Mais d'autres éléments relèvent manifestement de l'approche de Fisher :

- Il est dit que l'hypothèse nulle est formulée avec la volonté expresse, en général, de la rejeter. Il est également indiqué que l'hypothèse nulle est systématiquement celle d'une *absence* de différence (sans toutefois justifier ainsi son appellation). L'hypothèse de recherche est identifiée à ce qui est appelé l'hypothèse alternative, notée H_1. Si ce terme est d'origine neyman-pearsonienne, il est utilisé ici d'une manière si pauvre qu'il ne s'agit plus que d'une étiquette; en particulier aucune distinction n'est faite entre hypothèses ponctuelle et composée, d'où un manque de rigueur dans la présentation de la puissance.

L'hypothèse nulle apparaît implicitement comme la négation de l'hypothèse alternative dans le cas où elle n'est pas orientée. Mais dans le cas où elle est orientée, Siegel ne justifie pas le choix d'une hypothèse nulle ponctuelle. Ainsi, dans le premier exemple présenté, on trouve H_0: p(face) = p(pile) = ½, H_1: p(face) > p(pile), et on se demande ce que peut penser un lecteur qui découvre les tests de la disparition du cas p(face) < p(pile).

- Quand le résultat est significatif, la décision est de rejeter l'hypothèse nulle et d'accepter l'hypothèse alternative. Quand le résultat est non significatif, l'auteur s'en tient à dire qu'on ne peut pas rejeter l'hypothèse nulle au seuil fixé (il est intéressant de remarquer que ceci est formulé rapidement en début de chapitre, mais qu'ensuite, dans le paragraphe détaillant l'étape de décision, rien n'est dit du cas non significatif).

Il s'agit donc bien là d'une présentation hybride des tests. Hybride et contradictoire car Siegel définit, à juste titre, l'erreur de type II comme celle consistant à *accepter* H_0 alors qu'elle est fausse; mais… *jamais* il n'accepte H_0, comme nous venons de le signaler, et ne peut donc jamais commettre cette erreur.

La grandeur de l'effet

La question de l'intensité d'un effet ne retient pas l'attention de l'auteur (il faut dire qu'elle pose problème dans le cadre des méthodes non paramétriques) et n'est abordée que dans le cadre des mesures de corrélation (dernier chapitre). Et encore, à cette occasion, l'accent est davantage porté sur le test de signification associé à la mesure de corrélation qu'à la mesure elle-même. L'auteur donne l'impression que seule la question de la simple existence d'une corrélation au niveau d'une population est digne d'intérêt, et que peu importe l'intensité de la liaison :

> "Establishing that a correlation exists between two variables may be the ultimate aim of a research, […]."
>
> "It is, of course, of some interest to be able to state the degree of association between two sets of scores from a given group of subjects. But it is perhaps of greater interest to be able to say whether or not some observed association in a *sample* of scores indicates that the variables under study are most probably associated in the *population* from which the sample was drawn." (Siegel, 1956, p. 195) (Les italiques sont de l'auteur.)

On note que le problème de l'estimation est évacué, et également que la formulation de l'inférence statistique est abusive, laissant entendre qu'elle fournit la probabilité que l'association existe dans la population.

La présentation de l'intervalle de confiance

Comme nous l'avons dit au début, il n'est aucunement question d'intervalle de confiance puisque le problème de l'estimation est exclu du champ du livre.

Le calcul de N

Siegel précise qu'idéalement α et β (risque de deuxième espèce) devraient être choisis *a priori*, avant le recueil des données, et que cela fixe alors la taille de l'échantillon. Il n'indique pas pour autant comment effectuer le calcul de N et se borne à préciser qu'en pratique on fixe plutôt α et N et que cela détermine β

(toujours sans mentionner que β dépend également de la valeur du paramètre sous l'hypothèse alternative et sans expliciter le calcul).

Les exemples

Les exemples sont le plus souvent issus du champ de la psychologie sociale (les références de l'article d'origine sont citées quand les données sont réelles). Il sont systématiquement analysés selon les six étapes évoquées dans le chapitre 2. Dans presque tous les cas le test est significatif et la conclusion est l'acceptation de l'hypothèse de recherche. Quand, exceptionnellement, il ne l'est pas, l'hypothèse nulle n'est pas acceptée mais simplement non rejetée, sans autre interprétation : à aucun moment l'hypothèse nulle n'est considérée validée par un résultat non significatif.

Les abus

La présentation première de l'inférence statistique traditionnelle, dans l'introduction, apparaît ambiguë. Siegel laisse en effet entendre, bien qu'en termes vagues, qu'un test produit une probabilité concernant une hypothèse :

> "A common problem for statistical inference is to determine, *in terms of probability*, whether observed differences between two samples signify that the populations sampled are themselves really different." (Siegel, 1956, p. 2) (Italiques ajoutés.)

Cette ambiguïté est encore plus nette dans la citation rapportée précédemment (*cf. La grandeur de l'effet*).

Sans équivoque, on trouve en page 9 l'abus mineur évoqué en 2.2.5. consistant à définir α (et β) comme la probabilité, non conditionnelle, de commettre l'erreur de première (et deuxième) espèce. Cet abus est répété en page 14.

4.3. B. J. WINER : STATISTICAL PRINCIPLES IN EXPERIMENTAL DESIGN. (1962/1971)

Ce livre est destiné aux étudiants et chercheurs de sciences humaines et de biologie. C'est un ouvrage spécialisé puisqu'il ne traite que de l'inférence statistique et plus spécifiquement encore, comme son titre l'indique, que de l'analyse de variance et des plans d'expérience. Il suppose des connaissances préalables en statistique, et notamment celles des bases de l'inférence statistique. Ces bases sont cependant rappelées dans le premier chapitre; les autres chapitres étant consacrés à l'étude de divers plans d'expérience. Une double bibliographie est proposée, la première partie renvoyant aux travaux statistiques et la seconde renvoyant aux travaux expérimentaux dont sont issus certains des exemples.

L'introduction se termine par un bref hommage aux travaux fondateurs de Fisher en matière de plans d'expérience. Neyman et Pearson, quant à eux, ne sont pas cités dans l'ouvrage.

La notion de probabilité

L'auteur ne parle pas des différentes conceptions de la probabilité, ni de la sienne en particulier. Mais comme il écrit (p. 8), à propos des distributions d'échantillonnage, "… the relative frequency (probability) of statistics …", une conception fréquentiste est une hypothèse des plus plausibles.

La présentation des tests

Les notions de base de l'inférence statistique sont exposées dans le premier chapitre. Elles ont trait à l'estimation, ponctuelle et par intervalle, et aux tests. La plupart des éléments de présentation du test de signification renvoient à l'approche de Neyman et Pearson. Ainsi :

- Le test est présenté comme "un ensemble de règles permettant d'aboutir à une *décision* concernant l'hypothèse".
- Deux hypothèses sont introduites, l'hypothèse testée et l'hypothèse alternative (elles sont respectivement notées H_1 et H_2 au lieu des traditionnelles H_0 et H_1, ce qui ne facilite pas particulièrement la lecture; ici nous continuerons d'utiliser les notations usuelles), et les erreurs de types I et II sont mentionnées. Le terme d'hypothèse nulle n'est pas utilisé; simplement l'expression *to nullify* (une hypothèse) est employée dans les généralités sur les tests comme synonyme de rejeter.
- La notion de puissance d'un test est exposée (mais il n'est pas question de la fonction de puissance). Il est même précisé que les utilisateurs n'y prêtent pas suffisamment attention, accordant trop d'importance au seul risque de première espèce (p. 13). À cette occasion, Winer insiste sur le caractère purement conventionnel des

habituels seuils $\alpha = 0.05$ et $\alpha = 0.01$, remarquant que dans bien des cas 0.30 et 0.20 seraient des valeurs beaucoup plus adaptées.
- Le seuil observé ne joue pas d'autre rôle que celui de repère par rapport au seuil fixé.
- Le choix du risque α devrait se faire avant le recueil des donnée (p. 160). Winer remarque cependant que la convention est plutôt de choisir α au vu des résultats (on annonce "significatif à 0.05" ou 0.01 ou 0.001, *etc.*, selon le résultat).
- Si l'auteur reconnaît que le plus généralement l'hypothèse de recherche correspond à l'hypothèse alternative, il n'en préconise pas moins (p. 12) que le choix de l'hypothèse nulle soit effectué, si possible, de telle façon que l'erreur la plus coûteuse soit celle de première espèce puisque le risque correspondant est directement sous le contrôle de l'utilisateur (on retrouve le principe de Neyman). Ce principe est effectivement appliqué lors de la première illustration d'une procédure de test (p. 14).

Cependant, la formulation des règles de décision d'un test est assez particulière. L'auteur énonce que ces règles ne concernent que le rejet ou non de H_0 et que "le rejet de H_0 peut être regardé comme la décision d'accepter H_1, [tandis que] le non rejet de H_0 peut être regardé comme une décision contre l'acceptation de H_1" (p. 11). Mais en toute logique, rejeter H_1 c'est accepter H_0, ces hypothèses étant complémentaires, et c'est d'ailleurs bien ce qu'on trouve dès le premier exemple fourni (pp. 15 et 18)[13]. Nous pensons qu'il faut voir dans ce choix des termes (éviter de parler de l'acceptation de H_0) l'influence de Fisher (dont les idées sont bien différentes, un résultat non significatif n'étant pour lui qu'un constat d'ignorance et non une décision contre H_1), d'autant que la description de la logique du test, quelques lignes plus haut, s'inscrit directement dans la tradition de Fisher : il n'est question que d'une seule hypothèse dont on cherche à savoir si elle est contredite ou non par les données, l'hypothèse alternative n'étant introduite que plus tard. On ne peut pour autant parler d'une logique hybride, il s'agit davantage d'une "teinture" fishérienne sur une construction neyman-pearsonienne.

La grandeur de l'effet

La grandeur de l'effet en tant que telle ne fait l'objet que d'un court paragraphe dans lequel il n'est question que du coefficient ω^2 de Hays (qui n'est pas cité). Le coefficient intraclasse est rapidement présenté dans le cadre de la régression polynomiale puis dans celui des facteurs aléatoires. Aucun de ces deux coefficients n'est très commenté.

La présentation de l'intervalle de confiance

L'estimation en général (ponctuelle et par intervalle) est présentée avant la théorie des tests. L'interprétation fréquentiste de l'intervalle de confiance est tout à fait correcte, et il est bien précisé que pour un cas donné l'énoncé selon lequel le paramètre appartient à l'intervalle calculé est simplement soit vrai, soit faux (pp. 10 et 24). La relation entre l'intervalle de confiance et le test de signification est bien indiquée.

Le calcul de N

Le calcul de la taille d'échantillon nécessaire pour atteindre une puissance donnée est détaillé et n'est pas limité au seul cas d'une comparaison à un degré de liberté.

Les exemples

Les exemples, nombreux, servent essentiellement à illustrer le calcul numérique. Le plus souvent l'auteur se borne à conclure que l'hypothèse testée est rejetée (ou non), que les données contredisent (ou non) l'hypothèse, sans interpréter plus avant les résultats. Ainsi que nous l'avons noté précédemment, le non rejet de H_0 signifie en réalité, dans l'esprit de l'auteur, son acceptation. Cela devient manifeste dès le premier exemple, artificiel, d'un test. En effet, étant donnée l'hypothèse $H_0 : \mu \leq 50$, il est dit (à deux reprises) que son non rejet implique que $\mu \leq 50$: "Thus, if [H_0] is not rejected, the evidence would indicate that μ is equal to or less than 50." (p. 18). On retrouve cette attitude nette dès que des résultats sont interprétés plus complètement, en particulier lorsqu'il s'agit de tester une interaction. Par exemple, à partir d'un test d'interaction non significatif il est affirmé que les effets de deux facteurs sont indépendants (pp. 633-634), ou que les effets principaux sont additifs (p. 638). Ces conclusions sont énoncées sans référence à la puissance du test.

[13] Si l'on voulait discerner l'acceptation de H_0 de son non rejet (suspension de la décision), encore faudrait-il le faire clairement, ce qui conduirait à construire trois régions : la région critique, la région d'acceptation et celle de non décision. C'est d'ailleurs ainsi que Neyman et Pearson (1933b) présentent le cas général, mais il semble qu'ils n'ont pas poursuivi plus avant dans cette distinction. Et si l'on tient absolument à poser que ne pas rejeter H_0 alors qu'elle est fausse est une erreur, un souci de cohérence commanderait d'introduire une erreur de troisième espèce : suspendre la décision alors que H_0 est vraie.

Les abus

En ce qui concerne l'acceptation de H_0, on ne peut parler d'abus si l'on admet que l'auteur adopte la théorie de Neyman et Pearson, mais on regrettera qu'il ne soit pas plus clair quant au fait qu'il s'agit bien d'une acceptation.

Signalons à ce propos que Winer utilise le test usuel pour tester une hypothèse de négligeabilité (p. 700), cette dernière étant rejetée en cas de résultat significatif (ce qui se passe dans l'exemple choisi) et apparemment acceptée sinon, sans référence à la puissance en jeu.

Bien que Winer interprète toujours correctement α et β comme des probabilités conditionnelles, le chapeau introductif du paragraphe sur les tests pourrait laisser penser le contraire en raison d'expressions vagues et ambiguës, telle "l'expérimentateur attache un énoncé probabiliste à ses décisions" (p. 10). De même, il est très discutable de présenter ces mêmes probabilités comme fournissant une mesure de la "précision" des décisions prises. Il est encore plus regrettable de confondre α et β avec la grandeur (*magnitude*) des erreurs de type I et II (p. 11) qu'il est beaucoup plus naturel d'exprimer par l'écart entre les valeurs du paramètre sous chacune des hypothèses[14].

En revanche, il est abusif de laisser entendre que l'intervalle de confiance est la seule méthode d'estimation par intervalle :

> "An interval estimate is frequently referred to as a *confidence interval* for a parameter." (Winer, 1971, p. 10) (Les italiques sont de l'auteur.)

4.4. W. L. HAYS : STATISTICS FOR THE SOCIAL SCIENCES. (1963/1973)

La première édition a été publiée en 1963 sous le titre *Statistics for Psychologists*. La seconde (1973) a été assez fortement augmentée. Les ajouts ne concernent pas la présentation de l'inférence statistique, à la notable exception du chapitre 19, entièrement nouveau et consacré à des méthodes bayésiennes.

L'ouvrage s'adresse à des étudiants et porte à la fois sur la statistique descriptive et la statistique inférentielle (cette dernière étant tout de même davantage développée, en particulier à travers l'analyse de variance), en se limitant à des procédures univariées. À la différence de bien des manuels de ce genre, l'auteur a choisi de privilégier la théorie plutôt que l'application, et les fondements des méthodes sont relativement détaillés (bien qu'il ne s'agisse pas, comme il est souligné dans la préface, d'un ouvrage de statistique mathématique).

Hays est assez critique à l'égard des chercheurs en psychologie, puisqu'il déclare dans la préface :

> "Statistics has unfortunately achieved almost the status of a superstition in some quarters in psychology, and I hope, in all humility, that this text sets a slightly more liberal and rational one." (Hays, 1963, pp. vi-vii)

La notion de probabilité

L'interprétation fréquentiste de la probabilité est présentée en premier, mais l'interprétation subjectiviste est également citée. L'auteur insiste sur le fait que ce ne sont que diverses interprétations d'un modèle mathématique abstrait et qu'en tant que telles, aucune n'est plus vraie que l'autre.

La présentation des tests

Les tests traditionnels sont présentés dans un chapitre intitulé "Hypothesis Testing and Interval Estimation". Cette présentation est originale dans la mesure où les tests sont abordés sous l'angle de la théorie de la décision, dans la lignée des travaux de Wald à partir de la théorie de Neyman et Pearson. Ainsi il est fait référence à la fonction de coût des diverses décisions possibles (accepter ou rejeter telle hypothèse), et au problème de la recherche d'une règle de décision optimale. Mais cette approche n'est choisie par Hays qu'en raison de son caractère pédagogique, et il ne la considère pas applicable à la recherche en psychologie où il est impossible, en général, de définir des fonctions de coût. Suit alors une présentation plus traditionnelle de la théorie de Neyman et Pearson, où, justement, l'auteur insiste sur le caractère conventionnel du choix du risque α en l'absence de fonction de coût. Dans cette présentation nous relèverons les points suivants :

- On ne teste pas *une* hypothèse; il s'agit en réalité toujours de décider entre *deux* hypothèses.

---

[14] À la limite, on pourrait l'accepter pour β car celui-ci, N et α étant fixés, est une fonction monotone de l'écart $|\delta_0-\delta_1|$ entre les valeurs du paramètre sous les différentes hypothèses. En revanche, c'est tout à fait inadmissible pour α puisqu'il est fixé, constant, quel que soit cet écart.

- Le risque α, ainsi que la taille N de l'échantillon sont spécifiés avant le recueil des données (ou, du moins, indépendamment de celui-ci). Ceci résulte implicitement de l'ordre des opérations décrites pour la mise en œuvre d'un test (pp. 336-337) et des exemples donnés.
- En cas de résultat non significatif, accepter H_0 ou suspendre le jugement dépendra des circonstances particulières. En particulier il n'y aura pas d'acceptation ferme quand le risque β n'est pas connu (c'est-à-dire en dehors du cas du test de deux hypothèses ponctuelles).
- Le choix de l'hypothèse H_0 doit reposer sur le principe de Neyman (sans que celui-ci soit cité) : H_0 doit correspondre, en général, au risque le plus grave.
- Il n'est pas fait mention du seuil observé p.

Il s'agit donc d'une présentation fidèle de la théorie de Neyman et Pearson et non d'un amalgame. La seule dérogation notable concerne l'acceptation de H_0 en cas de résultat non significatif, pour laquelle Hays est beaucoup plus réservé que Neyman et Pearson.

On remarquera que Fisher et Neyman et Pearson ne sont cités qu'une fois, en fin de chapitre, sans aucune allusion aux différents qui les ont opposés. Bien au contraire, on pourrait croire à une seule théorie des tests, progressivement élaborée par ces auteurs :

> "The general theory of hypothesis testing first took form under the hand of Sir Ronald Fisher in the nineteen-twenties, but it was carried to a high state of development in the work of J. Neyman and E. S. Pearson, beginning about 1928." (Hays, 1973, p. 375)

Le chapitre se termine en relativisant le rôle des tests par des mises en garde utiles contre la confusion entre significativité statistique et substantielle, contre une tendance à juger de la qualité ou de la scientificité d'une recherche sur la seule base des résultats des tests statistiques ou de leur degré de sophistication :

> " However, conventions about significant result should not be turned into canons of good scientific practice. Even more emphatically, a convention must not be made a superstition. […] It is a grave error to evaluate the “goodness” of an experiment only in terms of the significance level of its results." (Hays, 1973, p. 385)

Dans sa conclusion, l'auteur souligne, peut-être de façon trop générale, que le bon sens doit guider le chercheur dans son application d'une technique statistique :

> "But if there is ever a conflict between the use of a statistical technique and common sense, then common sense comes first." (Hays, 1973, p. 386)

La théorie de Neyman et Pearson n'est pas la seule présentée, puisque le dernier chapitre (le 19, nouveau par rapport à la première édition) est consacré à l'introduction de certaines méthodes bayésiennes (“Some elementary Bayesian methods”) que Hays pense être appelées à se développer et à l'égard desquelles il se montre plutôt favorable. Mais Rouanet (1997) note que ce chapitre a disparu dans les éditions postérieures à 1980, les méthodes bayésiennes n'étant toujours pas utilisées en psychologie.

La grandeur de l'effet

La grandeur de l'effet tient une place importante, particulièrement (mais pas exclusivement) dans le cadre de l'analyse de variance où est présenté, entre autres, le coefficient ω^2 défini par l'auteur (*cf.* 3.1.).

La présentation de l'intervalle de confiance

L'intervalle de confiance est présenté à la suite des tests de signification, dans le même chapitre. L'interprétation qui en est faite est “orthodoxe” et ne comporte pas d'abus (l'accent est mis sur le caractère “délicat” de cette interprétation). Par ailleurs, la relation entre l'intervalle de confiance et le test (neyman-pearsonien) est mentionnée à deux reprises (pp. 376 et 378). Pour Hays, le calcul et la présentation des intervalles de confiance devraient faire partie des habitudes.

Le calcul de N

L'auteur indique la possibilité de calculer la taille d'un échantillon à partir d'une puissance requise pour un test, ou d'une largeur fixée pour un intervalle de confiance, et cela fait l'objet de quelques exercices. Cependant ce point est davantage développé dans le cas de l'intervalle de confiance que dans celui du test, pour lequel il ne donne pas lieu à un paragraphe spécifique.

Les exemples

Dans les exemples proposés, les conclusions vont un peu au delà du simple verdict significatif / non significatif, mais restent empreintes de prudence : “le facteur semble avoir un effet”, “l'évidence est insuffisante pour déterminer si le facteur a un effet”, “Il n'y a virtuellement aucune évidence d'une interaction” (ce qui n'est

pas équivalent à “virtuellement, il y a évidence d'une absence d'interaction”). On ne trouve donc pas d'abus dans les illustrations des méthodes (encore qu'une acceptation de H_0 n'aurait pas été véritablement un abus, étant donné le cadre neyman-pearsonien choisi par l'auteur).

Les abus

Tout de même, la présentation des tests est parfois empreinte de maladresse, d'abus de langage. Ainsi, à propos de la puissance, il est écrit :

"H_0, the hypothesis actually being tested
H_1, the hypothesis, whatever it may be, that is *true*." (Hays, 1973, p. 357) (Italiques ajoutés)

Il ne s'agit pas d'un lapsus puisque l'adjectif “true” est systématiquement utilisé dans ce contexte dans les pages suivantes (et même en titre de paragraphe : “Power of tests against various true alternatives”). Si l'hypothèse alternative spécifie la *vraie* valeur du paramètre, est-il encore besoin d'un test ? Évidemment, Hays veut ici parler d'hypothèses ponctuelles, ou “exactes” selon la définition qu'il en donne lui-même page 336. C'est vraisemblablement dans l'intitulé de l'axe des abscisses de la fonction de puissance, qui, certes peut être “*vraie* valeur du paramètre”, qu'il faut voir l'origine de cet abus de langage. À ce propos, Hays précise que la puissance ne peut être calculée tant que l'hypothèse alternative n'est pas ponctuelle. En réalité ce qu'il veut dire, ainsi qu'il l'expose d'ailleurs clairement par la suite, c'est qu'en cas d'hypothèse composée on ne peut plus parler d'une seule valeur de la puissance[15] mais d'une fonction, qui reste parfaitement calculable. (Cette dernière maladresse a été portée à un haut degré d'achèvement et agrémentée de contresens par Chow, 1996; *cf.* Poitevineau et Lecoutre, 1997.)

En revanche, ce n'est plus un abus de langage quand Hays, au cours d'un exemple (p. 366), utilise comme argument en faveur de l'acceptation de H_0 une erreur de seconde espèce dépendant non seulement de la valeur théorique du paramètre mais aussi de la valeur de la statistique de test observée, autrement dit un “β observé”, pendant du seuil observé p. Il s'agit là d'une incohérence dans la mesure où l'auteur s'inscrit dans le cadre strict de la théorie de Neyman et Pearson, auquel il se tient parfaitement dans tout le reste du chapitre.

4.5. M. REUCHLIN : PRÉCIS DE STATISTIQUE. (1976)

Il s'agit d'un ouvrage destiné aux étudiants du premier cycle de sciences humaines et tout particulièrement aux étudiants en psychologie. Sa principale caractéristique est d'aborder les statistiques d'un point de vue notionnel et il est voulu complémentaire plutôt que concurrent d'ouvrages plus formels destinés au même public.

La place accordée à l'inférence statistique peut sembler assez réduite puisque celle-ci n'est véritablement traitée qu'au chapitre VII, le dernier, et sans être particulièrement développée. Cependant de nombreuses allusions aux tests, ou présentations partielles, sont faites dans les chapitres précédents, notamment dans le premier, et en particulier dans les exemples.

Aucune bibliographie n'est donnée, sauf en ce qui concerne les travaux d'où sont tirés les exemples réels. Ni Fisher, ni Neyman et Pearson, ni d'autres ne sont crédités d'un rôle quelconque dans la genèse des méthodes de test statistique puisqu'elle n'est pas évoquée (si Fisher est cité, ce n'est que pour des raisons très “techniques”, concernant la distribution du F par exemple).

La notion de probabilité

Le terme de probabilité ne figure pas dans l'index et aucune définition n'en est donnée (ce dont on peut s'étonner dans un ouvrage justement à caractère notionnel). Comme le terme est tout de même utilisé de temps en temps, c'est donc une notion que l'auteur suppose connue. Le premier chapitre, bien que traitant du “Caractère variable des conduites” ne comporte pas le mot probabilité (en réalité “probable” figure une fois, en introduction), en revanche, il y est fait référence à la notion de variations fortuites, imprévisibles, par opposition aux variations systématiques, prévisibles. En fait, le terme de probabilité n'apparaît que dans le dernier chapitre (celui sur l'inférence), d'abord lorsque sont évoquées les méthodes d'extraction d'échantillons représentatifs d'une population, puis à propos de l'intervalle de confiance et des tests. Il est difficile, par manque d'indications de l'auteur, d'inférer quelle est sa conception de la probabilité.

[15] Sauf à reprendre le concept de “puissance résultante” introduit par Neyman et Pearson (1933b), mais qui fait intervenir les probabilités *a priori* (moyenne des puissances possibles pondérées par ces probabilités, c'est la probabilité non conditionnelle de détecter un écart réel à H_0), et qu'ils n'ont plus considéré ensuite.

La présentation des tests

La procédure de test est présentée à partir d'exemples, sans que soit donné au préalable (ni ensuite) de réel "schéma directeur". Le premier exemple est quelque peu déroutant et pour le moins maladroit. Le test est en effet introduit pour répondre à la question de savoir, le paramètre parent étant connu, si un échantillon, pour lequel on dispose d'une estimation de ce paramètre, peut être considéré comme représentatif de la population. Mais la réponse à cette question est immédiate si la procédure d'extraction de l'échantillon est connue (l'auteur a vanté les mérites, dans les pages précédentes, des échantillons tirés au hasard); ou, si l'on veut définir la représentativité par l'écart entre l'estimation et la valeur du paramètre, il n'est nul besoin de test puisqu'on dispose de toute l'information nécessaire. Reuchlin veut sans doute évoquer ici le cas où, la méthode d'échantillonnage étant inconnue, on souhaite tester l'hypothèse qu'il s'agit d'un tirage au hasard; mais la formulation est ambiguë.

D'une manière générale, on relève que :

- Le mot "décision" est souvent employé (par exemple il s'agit de décider que les différences sont significatives), ce qui rappelle plutôt l'approche de Neyman et Pearson.
- Le seuil fixé (noté P) ne l'est pas forcément à l'avance (mais alors peut-on encore parler de seuil fixé ?) si l'on se réfère à certains exemples où la question du choix du seuil est posée une fois calculée la statistique de test (plus que Neyman et Pearson, cela évoque Fisher dans sa première époque). Le seuil observé ne sert que par comparaison au seuil fixé, pour décider si le résultat est significatif ou non.
- La notion de risque est introduite avec l'intervalle de confiance. L'existence de deux types de risques n'est pas mentionnée, ce qui suppose alors, conformément à la théorie de Fisher, qu'on n'accepte jamais l'hypothèse nulle.
- Aucun principe de choix de l'hypothèse nulle n'est véritablement dégagé; le terme même d'hypothèse nulle n'est présenté "qu'en passant" et celle-ci est identifiée à l'hypothèse de la valeur zéro du paramètre (p. 205). Selon les exemples, le lecteur peut aussi bien penser que l'hypothèse nulle est celle qu'on espère rejeter (inexistence d'une corrélation) ou au contraire confirmer (exemple d'une théorie génétique prédisant une certaine proportion d'un caractère donné). Ce dernier cas pourrait même être considéré comme privilégié puisqu'il est écrit :

> "*Dans tous les cas*, il s'agit de savoir si la fréquence (ou la moyenne) observée prend une valeur qui permette de décider, en acceptant un risque d'erreur déterminé, que les observations sont *compatibles* avec ce que l'on sait de la population ou avec la théorie, le raisonnement que l'on a utilisés pour déterminer la valeur *attendue* du paramètre." (Reuchlin, 1976, p. 201) (Italiques ajoutés.)

- S'il apparaît clairement qu'en cas de résultat significatif l'hypothèse nulle est rejetée, le doute peut subsister quant à l'attitude à adopter en cas de résultat non significatif. Ou plutôt, au travers des exemples proposés, on trouve les deux attitudes : simplement ne pas rejeter l'hypothèse, ou l'accepter (ce dernier cas étant contradictoire avec la mention d'un seul risque, comme on l'a signalé précédemment).
- À la fin du chapitre sur l'inférence, l'auteur met en garde contre le danger de confondre significativité statistique et significativité psychologique.

Il en ressort une conception encore hybride du test de signification, qu'on pourrait aussi qualifiée de "floue".

La grandeur de l'effet

La grandeur de l'effet est évoquée dès le premier chapitre, par des allusions à l'analyse de variance. En fait, Reuchlin laisse entendre que le but premier de l'analyse de variance est d'indiquer les importances relatives des diverses sources de variation, mais il n'y a pas de discussion systématique du problème de la mesure d'un effet. Le coefficient η^2 est introduit dans un paragraphe traitant de la "Relation entre une variable nominale et une variable d'intervalles", mais n'est pas repris ensuite lorsqu'il est question de l'analyse de variance. L'auteur insiste pourtant, en conclusion du dernier chapitre, sur l'insuffisance du test s'il n'est pas prolongé par des informations portant sur la grandeur de l'effet.

La présentation de l'intervalle de confiance

La présentation de l'intervalle de confiance précède celle du test. Elle consiste à introduire d'abord l'intervalle d'échantillonnage puis à permuter les rôles du paramètre et de son estimateur. Cette façon de procéder pourrait évoquer la conception fiduciaire de Fisher mais, outre que celui-ci et sa méthode ne sont pas mentionnés, elle est immédiatement suivie d'une interprétation fréquentiste correcte. La relation avec le test est indiquée, c'est même ainsi que les procédures de test sont abordées (pour tester une hypothèse sur la valeur du paramètre, on calcule l'intervalle de confiance et on regarde si cette valeur appartient à l'intervalle), mais toujours dans le cadre d'un exemple, sans en expliciter la généralité.

Le calcul de N

La puissance des tests n'est pas présentée, et le calcul d'une taille d'échantillon non plus.

Les exemples

Les exemples et les exercices proviennent le plus souvent du champ de la psychologie. Les corrigés d'exercices donnent lieu à une interprétation psychologique des résultats qui va au delà de la simple "différence significative à P = 0.05" et resitue ces résultats par rapport aux hypothèses psychologiques de l'étude.

Les abus

En ce qui concerne l'intervalle de confiance, si, au début de la page 99, une première interprétation est correcte :

> "On peut dire que si, l'intervalle ± .10 est délimité autour de chacune des valeurs prises par F [la variable aléatoire Fréquence] au cours des estimations successives, cet intervalle contiendra φ [la fréquence parente] pour 95 estimations sur 100."

quelques lignes plus loin, une seconde formulation, relativement ambiguë, nous semble pencher vers l'abus :

> "Dès lors, le psychologue ne disposant que d'une seule valeur *f* de F pourra délimiter les bornes *(limites de confiance)* d'un intervalle *(intervalle de confiance)* entre lesquelles il a 95 chances sur 100 de trouver φ (seuil de .05). Dans l'exemple, cette valeur de *f* de F estimée sur un échantillon de 100 enfants est de .54." (Reuchlin, 1976, p. 199) (Les italiques sont de l'auteur.)

En effet, il nous paraît plutôt que "95 chances sur 100" s'adresse à l'intervalle particulier construit autour de la valeur observée *f*.

En ce qui concerne le test, nous avons déjà signalé que l'hypothèse nulle pouvait apparaître comme étant celle que l'on cherche à vérifier, laissant entendre alors qu'on pouvait se satisfaire d'un résultat non significatif. C'est par exemple le cas page 191 où, pour vérifier qu'un échantillon a été correctement tiré, on attendra une différence non significative entre la valeur observée et celle du paramètre (connu) de la population. Même dans le cas où l'hypothèse de recherche est identifiée à l'hypothèse alternative (terme qui n'est pas employé par Reuchlin), on trouve encore que l'hypothèse nulle peut être acceptée; ainsi, dans un exemple, il est conclu que le hasard suffit à expliquer les faibles variations observées entre les moyennes. Ce ne peut être seulement une formulation maladroite, car il est bien précisé :

> "Une telle issue de l'expérience n'est pas un « échec » : elle apporte un élément d'information." (Reuchlin, 1976, p. 26)

Une incorrection apparaissant dans la conclusion, est l'affirmation que la signification statistique est une condition nécessaire pour que le résultat présente un intérêt psychologique (p. 215). En fait, il peut très bien arriver qu'un résultat soit non significatif et qu'on puisse, de plus, conclure à la négligeabilité de l'effet parent. Et cette conclusion peut être intéressante sur le plan psychologique (une différence négligeable entre les effets d'un apprentissage massé et d'un apprentissage distribué, pour reprendre un des exemples de l'auteur, constituerait un résultat intéressant). Un test significatif ne peut donc être posé comme une condition nécessaire.

4.6. R. E. KIRK : EXPERIMENTAL DESIGN: PROCEDURES FOR THE BEHAVIORAL SCIENCES. (1982)

Ce livre est très proche de celui de Winer, de par le public visé et de par le contenu puisqu'il présente un ensemble de plans d'expériences dans le cadre de l'analyse de variance. Comme pour Winer, la lecture requiert d'être déjà introduit aux statistiques. Cependant le rappel de ce qu'est l'inférence statistique figurant dans le premier chapitre est plus détaillé que chez Winer et pourrait même convenir pour une première exposition de ce sujet. La bibliographie, importante, ne concerne pratiquement que le domaine des statistiques; Fisher y apparaît, mais non Neyman et Pearson. Ces derniers sont toutefois nommés, avec Fisher, pour leurs travaux sur la théorie de l'inférence statistique dans un bref passage du chapitre 1 (p. 9). Toujours comme Winer, Kirk cite les travaux pionniers de Fisher en matière de plans d'expérience.

La notion de probabilité

Les probabilités, supposées connues, ne donnent pas lieu à commentaires et l'auteur ne mentionne ni les diverses conceptions de la probabilité, ni celle qu'il adopte.

La présentation des tests

La présentation des tests précède celle des estimations par intervalle. Nous allons voir qu'elle est très voisine de celle de Winer. Là encore, la plupart des éléments du test s'inscrivent dans la lignée de la théorie de Neyman et Pearson :

- Il s'agit de prendre une *décision* quant au rejet ou non de H_0, l'hypothèse nulle. Ce terme est abondamment utilisé, et l'auteur n'hésite pas à parler de "décision finale à propos de l'hypothèse scientifique" (p. 27).
- Deux hypothèses sont en jeu, l'hypothèse nulle et l'hypothèse alternative (notées, comme à l'habitude, H_0 et H_1), qui sont exclusives et exhaustives. Les erreurs de types I et II sont décrites. Les valeurs usuelles du risque de première espèce α (0.05, 0.01) sont données comme de simples conventions arbitraires. Il est précisé qu'en général l'erreur de type I est jugée plus sérieuse que celle de type II (p. 38), mais sans que la conséquence en soit tirée puisque cela ne retentit pas sur le choix des hypothèses, comme on le verra plus loin.
- L'attention est portée sur la notion de puissance d'un test et le calcul de la puissance est explicité dans le cas simple d'hypothèses ponctuelles. Mais il n'est pas fait allusion à la fonction de puissance.
- Le risque α et N doivent être choisis avant le recueil des données (comme en témoigne l'ordre des étapes d'un test, p. 32).
- Le seuil observé n'intervient que pour comparaison au seuil fixé α, et s'il est bon de le communiquer dans un article, c'est uniquement pour permettre au lecteur d'appliquer son propre critère.

Mais les points suivants évoquent la théorie de Fisher :

- L'hypothèse de recherche est associée à l'hypothèse alternative H_1 et l'hypothèse nulle H_0 est sa négation (p. 26).
- Le résultat d'un test est le rejet de H_0 si la statistique de test appartient à la région critique, son non rejet sinon (p. 32). À la différence (apparente) de Winer, il n'est pas question de décision contre l'acceptation de H_1.

Mais que peut signifier "non rejet de H_0" d'autre que "acceptation de H_0" quand on se place, comme l'auteur, dans le cadre d'une approche décisionnelle et que les deux hypothèses en jeu sont complémentaires ? D'ailleurs il est bien dit que le processus de test est un processus visant à choisir entre H_0 et H_1 (p. 26). Kirk ébauche bien une justification en précisant qu'en cas de résultat non significatif "on n'a pas prouvé que l'hypothèse nulle est vraie — seulement que les faits ne garantissent pas son rejet" (p. 33). Mais il peut être dit la même chose du cas inverse (un résultat significatif n'est pas la preuve de la véracité de l'hypothèse alternative), car il n'est aucunement question de preuve, de certitude, en matière d'inférence statistique; et d'ailleurs, du point de vue de Neyman et Pearson il ne s'agit pas de se prononcer sur la véracité d'une l'hypothèse mais simplement de décider, de "faire comme si". Finalement, c'est quelques pages plus loin, quand on en vient à l'explicitation des types d'erreurs I et II, que l'identité du non rejet et de l'acceptation est reconnue :

> "If the null hypothesis is true and the experimenter does not reject it, a *correct acceptance* has been made; if the null hypothesis is false and the experimenter rejects it, a *correct rejection* has been made " (Kirk, 1982, p. 36) (Les italiques sont de l'auteur.)

Tout ceci est donc très similaire à ce qu'on trouve chez Winer, seulement on est plus proche ici de la "logique hybride" dans la mesure où le choix de l'hypothèse nulle correspond à ce que l'on trouve chez Fisher.

Évoquant, très brièvement, le fait qu'il existe des critiques des tests de signification, l'auteur se contente de mettre en garde contre la confusion entre significativité statistique et significativité pratique (p. 41).

La grandeur de l'effet

La grandeur de l'effet est abordée, bien qu'assez rapidement. Seuls des coefficients de grandeur relative sont présentés : d de Cohen, η, ω^2 et coefficient de corrélation intraclasse. L'auteur insiste sur le fait qu'un résultat significatif ne renseigne que sur l'existence d'un effet et non sur son intensité, et que cet effet peut très bien s'avérer trivial.

La présentation de l'intervalle de confiance

Le premier chapitre se termine par un paragraphe sur l'estimation par intervalle intitulé d'une façon très générale "Interval estimation". Or seul l'intervalle de confiance y figure et l'on pourrait croire, à la lecture, qu'il n'existe pas d'autre méthode d'estimation par intervalle. Le lien entre le test et l'intervalle de confiance est indiqué; en revanche, on cherchera vainement une interprétation fréquentiste rigoureuse de ce dernier, celle qui est donnée étant plutôt abusive.

Le calcul de N

Selon l'auteur, le calcul de la taille de l'échantillon préalablement au recueil des données devrait être une routine. Le principe de ce calcul est détaillé, à la fois pour le cas où l'écart-type est connu et pour celui, plus usuel, où il ne l'est pas.

Les exemples

Comme chez Winer, beaucoup d'exemples sont destinés essentiellement à illustrer les étapes du calcul numérique et sont donc peu interprétés. Il en va partiellement de même en ce qui concerne les très nombreux exercices proposés qui ont aussi souvent trait à des aspects purement mathématiques. Même dans les cas où les exemples proviennent, ou sont inspirés, de situations réelles (souvent issues d'études de psychologie), les conclusions que l'on trouve dans les corrigés des exercices sont le plus souvent limitées à "on rejette H_0" ou à "on ne rejette pas H_0". Ainsi "on ne rejette pas l'hypothèse nulle d'homogénéité des variances" (p. 882). Tout de même, par deux fois au moins, le résultat non significatif donne clairement lieu à l'acceptation de l'hypothèse nulle d'absence d'effet (pp. 856 et 887).

Les abus

Si la présentation du test suit, pour l'essentiel, l'approche de Neyman et Pearson, certaines formulations évoquent très nettement l'abus, par rapport à cette approche, consistant à voir dans le test un moyen de donner une probabilité de la véracité de l'hypothèse testée. Ainsi, il est affirmé (p. 27) que le résultat du test fournit une base pour inférer la véracité ou la fausseté probable de l'hypothèse scientifique. Ou encore, que l'hypothèse "$\mu = 115$ est considérée improbable" (p. 30). Cependant l'auteur n'en vient jamais à écrire explicitement que la probabilité que H_1 soit vraie est égale à $1-\alpha$.

Dans les pages 36 à 38, les probabilités (inconditionnelles) de commettre chacune des deux erreurs possibles sont systématiquement identifiées à α et β.

Quant à l'intervalle de confiance, on peut dire qu'il est constamment interprété abusivement. Il en est ainsi dès son introduction :

> "…a random sample can be used to specify a segment or interval on the number line such that the parameter has a high probability of lying on the segment. The segment is called a *confidence interval*." (Kirk, 1982, p. 42) (Les italiques sont de l'auteur.)

Et c'est bien à un intervalle particulier, et non à la variable aléatoire intervalle, que l'auteur se réfère toujours; par exemple :

> "We can be 95% confident that the population mean is between 114.06 and 119.94." (Kirk, 1982, p. 43)

Le fait de parler de "confiance" au lieu de probabilité ne change fondamentalement rien, car nulle part il n'est précisé en quoi cette confiance serait à distinguer d'une probabilité.

Nous avons déjà mentionné qu'un autre abus est de sous-entendre que l'intervalle de confiance est la seule méthode d'estimation par intervalle.

On notera enfin un abus d'une autre nature. Kirk affirme (p. 36) que lorsque l'hypothèse nulle est composée (par exemple, $\mu \leq 100$), l'hypothèse alternative doit l'être également. Si cela est pratiquement toujours le cas en pratique, il n'y a pas pour autant de raison théorique de l'imposer.

* * *

Il apparaît de nettes distorsions entre les deux théories des tests de signification et leur présentation dans ces six ouvrages. Des ambiguïtés, sinon des abus, apparaissent le plus souvent dans les interprétations des résultats; de plus ces interprétations révèlent des divergences entre la présentation des principes et l'exposé de la pratique à partir des exemples.

Sur le seul critère de la présentation des tests, les six livres se répartissent en :

- un neyman-pearsonien pur (Hays), qui se distingue nettement par sa rigueur, sa conformité à la théorie statistique,
- deux neyman-pearsoniens plus "approximatifs" (Winer et Kirk),
- et trois "hybrides" (Siegel, Faverge et Reuchlin).

L'approche de Neyman et Pearson l'emporte donc, en fréquence, sur celle de Fisher qui n'est jamais présentée en tant que telle et n'apparaît que comme élément d'un amalgame des deux théories.

Il est nécessaire d'ajouter que la prolifération actuelle de logiciels statistiques de grande diffusion est susceptible de modifier quelque peu la situation. Ces logiciels ont de plus en plus tendance à se substituer aux manuels de statistiques dans un rôle prescriptif qui peut se retrouver à différents niveaux :
i) dans le manuel d'utilisation qui, par-delà sa fonction de mode d'emploi du logiciel, peut se présenter comme un véritable manuel de statistique (dès les années 70, le manuel du logiciel SPSS a pu jouer un rôle prescriptif auprès de chercheurs en sociologie et en psychologie);
ii) dans l'aide en ligne qui peut être une simple reprise du manuel ou apporter d'autres informations;
iii) dans les sorties présentées, où les résultats numériques sont de plus en plus souvent agrémentés de commentaires ou de conclusions *en clair*.

On peut à l'évidence craindre que ce nouveau rôle entraîne des simplifications abusives (nécessité commerciale de diffusion auprès du plus grand nombre d'utilisateurs oblige), ainsi que des distorsions supplémentaires tant dans la justification que dans l'interprétation des procédures. Il nous était impossible de faire l'étude de ces logiciels, en raison de leur nombre, mais surtout en raison de leur coût (leur étude nécessitant leur utilisation et donc leur achat). Il nous est cependant apparu que certains logiciels, diffusés à grand renfort de publicité, avaient en ce qui concerne leur rôle prescriptif une qualité pour le moins discutable (ce qui ne préjuge en rien ici de la qualité des procédures et des calculs). À titre d'illustration, nous rapporterons simplement les deux exemples suivants.
- Dans le logiciel STAT*lab* version 2.0 (CNET - France Télécom - SLP, 1994), pour le test d'indépendance du χ^2, après avoir entré les données et choisi un seuil de signification (0.05), on obtient pour seul résultat :

> "Meurtre et Agression [les variables de l'exemple] sont indépendantes au seuil de 5%"

L'assimilation d'un résultat non significatif (terme qui n'apparaît même plus) et de la conclusion d'indépendance se trouve légitimée.
- À la page 201 du manuel du logiciel de traitement d'enquête et d'analyse de données *QUESTION* (Grimmer logiciels, 1993) on peut lire les propos suivants qui se passent de commentaire :

> "En fonction de la valeur du Khi-deux et du nombre de degrés de liberté, le logiciel calcule la probabilité exacte. Si l'on se donne un seuil de 5% de risques, une probabilité inférieure à ce seuil signifie que l'erreur d'échantillonnage est faible, on suppose qu'il existe une dépendance entre les 2 variables ligne et colonne. Le hasard intervient seulement dans moins de 5 chances sur 100, dans la répartition observée des effectifs dans le tableau. Le hasard, l'erreur d'échantillonnage sont considérés comme négligeables. L'hypothèse d'indépendance est rejetée."

Malheureusement de graves dégradations peuvent aussi se retrouver dans certains ouvrages de statistique destinés aux étudiants, comme le constate J.-L. Durand dans son analyse d'un manuel statistique pour les psychologues récemment publié (Durand, 1997).

3ème PARTIE

APPROCHE DESCRIPTIVE

Nous en venons maintenant à étudier les attitudes des chercheurs en psychologie à l'égard des tests de signification. Les études réalisées dans cette optique peuvent être classées en deux groupes.

• Les études du premier groupe, qui feront l'objet du chapitre 5, consistent en des réanalyses statistiques des résultats publiés dans des revues scientifiques. Leur objectif premier est d'ordre méthodologique, ou encore prescriptif : il s'agit de mettre en évidence certains problèmes soulevés par l'usage des tests de signification, et éventuellement d'en tirer les conséquences pour une meilleure pratique.

Dans le cadre d'une approche descriptive, elles concernent une situation particulière, celle de publication, qui d'une part n'est que le résultat final de l'activité du chercheur et qui d'autre part doit s'accorder avec toutes les normes et contraintes en vigueur, explicites et implicites.

Nous compléterons la présentation d'un certain nombre de ces études par celle d'une réanalyse que nous avons effectuée dans une perspective plus descriptive : il s'agira d'une part de recenser quels sont les abus d'interprétation des tests effectivement commis, et d'autre part de chercher à préciser quelle est la portée réelle des conclusions autorisées en ce qui concerne l'importance des effets, en relation précisément avec les abus (ou les insuffisances) des interprétations fournies par les auteurs.

• Les études du deuxième groupe, auquel sera consacré le chapitre 6, consistent en des expériences menées auprès de chercheurs (éventuellement d'étudiants). Elles peuvent elles-mêmes être subdivisées en deux sous-groupes.

- Un premier sous-groupe comprend des études par questionnaires qui portent directement sur la signification des probabilités associées aux tests de signification et qui font directement appel à des connaissances "livresques". On pourrait même parler à leur propos de "questions de cours". Ici il ne s'agit que de confirmer l'existence des erreurs d'interprétation de ces probabilités par rapport à la référence normative; certaines de ces études ont d'ailleurs été réalisées par des statisticiens.
- Un second sous-groupe comprend des études beaucoup plus proches d'une situation réelle, où il est davantage fait appel à des jugements "intuitifs", "spontanés", pour lesquels le sujet n'a pas forcément à sa disposition une réponse stéréotypée, "toute prête".

Tversky et Kahneman (1971) ont ainsi été les instigateurs d'une série d'expériences sur les représentations des chercheurs dans diverses situations d'inférence statistique. Toutefois leur perspective, qui s'inscrit dans le cadre beaucoup plus général de l'étude des jugements probabilistes, reste relativement normative : il s'agit ici encore avant tout de répertorier des distorsions, des *biais* dans leur terminologie, par rapport à une référence normative (le cadre privilégié étant le cadre bayésien).

M.-P. Lecoutre (1991) adopte au contraire explicitement une approche beaucoup plus descriptive : il s'agit, à partir de situations expérimentales réalistes, de "recueillir des informations sur l'ensemble de la démarche du chercheur dans l'analyse statistique des données", et d'en dégager "un certain nombre d'intuitions probabilistes fondamentales, de lignes de cohérence dans les jugements émis, avec pour finalité la formalisation de modèles descriptifs du fonctionnement cognitif spontanément développés dans ce type de situations". Cette approche n'exclut pas pour autant une perspective normative et s'articule avec un objectif prescriptif, puisqu'il s'agit également de "faire mieux connaître les besoins et motivations réels des utilisateurs de la statistique, afin d'aboutir à une présentation des différentes méthodes d'inférence statistique existantes dégageant clairement leurs apports respectifs et leurs implications méthodologiques". (voir également M.-P. Lecoutre, 1982, 1983, 1988; M.-P. Lecoutre *et al.*, 1990; M.-P. Lecoutre et Rouanet, 1993)

C'est bien entendu cette dernière conception, permettant l'articulation des différentes approches, que nous privilégierons ici. La présentation des travaux effectués inclura une expérience que nous avons réalisée en collaboration avec M.-P. Lecoutre et qui permettra notamment, par la comparaison de résultats obtenus chez des psychologues et chez des statisticiens, d'étudier le rôle de la "culture statistique" dans les interprétations des résultats des tests de signification.

Enfin nous détaillerons et reprendrons une expérience de Rosenthal et Gaito (1963) qui a été la première étude portant sur la manière dont les chercheurs en psychologie expérimentale interprètent les seuils de signification associés à un test.

CHAPITRE 5

RÉANALYSES D'ARTICLES PUBLIÉS

5.1. RÉANALYSES ANTÉRIEURES

Les réanalyses présentées ici ont différents objectifs méthodologiques. Elles s'appliquent à des études très diverses par leurs sujets, sans qu'il s'agisse de chercher à synthétiser des résultats obtenus, ce qui les distingue des méta-analyses au sens habituel du terme.

5.1.1. L'étude de Cohen (1962)

En 1962 Cohen publie une recherche qui servira de modèle à beaucoup d'autres.

Il s'agit pour lui de voir comment les psychologues, si soucieux de se prémunir contre l'erreur de première espèce, se gardent de l'erreur de seconde espèce. Autrement dit, voir si la puissance des tests utilisés par les psychologues est suffisante pour que l'hypothèse nulle ait de bonnes chances d'être rejetée quand elle est fausse. Il s'inscrit donc clairement, et explicitement d'ailleurs, dans la lignée de Neyman et Pearson.

A cette fin il analyse tous les articles parus dans le volume 61 (1960) du *Journal of Abnormal and Social Psychology*.

Pour les besoins de l'analyse, il considère que le test est bilatéral avec un risque α de 0.05 pour tous les articles, et surtout il met au point une grille pour la classification des effets.

Définition des effets

Une première difficulté, propre à ce type d'études, est la grande diversité des variables dépendantes utilisées, et donc des unités de mesure lorsqu'il s'agit de variables numériques; ce qui gêne la comparaison d'un article à l'autre, voire à l'intérieur d'un même article. Aussi Cohen choisit de ne prendre en compte, pour les variables numériques, que des effets *relatifs* en calibrant la grandeur de l'effet absolu par un écart-type (celui correspondant à la source dite "d'erreur").

Une deuxième difficulté est que l'intensité des effets attendus, nécessaire aux calculs de puissance, est rarement spécifiée. Il définit alors, plus ou moins arbitrairement, trois niveaux de grandeur d'effet : faible, moyen, fort.

Enfin comme différents tests sont en jeu, la définition de la grandeur de l'effet est adaptée à chaque cas.

Cela se traduit par le tableau suivant :

Test	Paramètre de la population	Effet vrai		
		Faible	Moyen	Fort
[1] t (comparaison de 2 moyennes)	$\lvert\mu_1 - \mu_2\rvert / \sigma = d$	0.25	0.50	1.00
[2] F (comparaison de k moyennes)	$\sigma_{effet} / \sigma = f$	0.125	0.25	0.50
[3] t (coeff. corrélation $\rho = 0$)	$\lvert\rho\rvert$	0.20	0.40	0.60
[4] comparaison de 2 ρ	$\lvert\rho_1 - \rho_2\rvert$	0.10	0.20	0.30
[5] test du signe	$\lvert P - 0.50\rvert$	0.10	0.20	0.30
[6] comparaison de 2 proportions	$\lvert P_1 - P_2\rvert$	0.10	0.20	0.30
[7a] χ^2 (égalité de k proportions)	P_{max} / P_{min}	3 / 2	2 / 1	4 / 1
[7b] χ^2 (tableau de contingence)	$\Sigma[(P_{1i} - P_{0i})^2/P_{0i}] = l$	variable en fonction de la taille du tableau		

Tableau 2

Critères de grandeur d'effet relatif pour différents problèmes (Cohen, 1962)

Remarques

- Dans le cas [2] de la comparaison de plusieurs moyennes, σ_{effet} est l'écart-type calculé sur les moyennes parentes. C'est une mesure usuelle de la grandeur de l'effet en analyse de variance. Dans le cas particulier de deux moyennes, μ_1 et μ_2, on obtient $\sigma_{effet} = |\mu_1 - \mu_2| / 2$, d'où, par souci de cohérence, le choix des limites comme moitié de celles du cas [1]. Comme dans [1], σ est l'écart-type résiduel ou "intra-classe".
- En 1969 Cohen a modifié les définitions des cas [4] et [6] qui deviennent respectivement : $|z_1-z_2|$ (avec $z_i = ½ \log[(1+\rho_i)/(1-\rho_i)]$) et $|\phi_1-\phi_2|$ (avec $\phi_i = 2\arcsin\sqrt{P_i}$). Il avait déjà envisagé ces définitions en 1962, car elles ont l'avantage que la puissance ne dépend que de la différence des termes et non des valeurs de chacun, propriété que n'ont pas les définitions originelles. Il les avait rejetées, à l'époque, en raison de leur plus grande complexité.
Par ailleurs il a regroupé les cas [7a] et [7b] en un seul, en assimilant le cas [7a] au cas [7b] et a noté l'indicateur *e* au lieu de *l*.
Il a également modifié toutes les limites en les abaissant. Par exemple pour le cas [1] elles deviennent : 0.20, 0.50, 0.80.
- Notons enfin, comme nous l'avons déjà dit (*cf.* 2.1.14.), qu'il faut différencier ici la notion d'effet "faible" de celle d'effet négligeable ou de limite "d'indifférence". Cohen considère qu'un effet parent mérite d'être détecté à partir du moment où il existe réellement (c'est-à-dire où il n'est pas strictement nul). Pour lui un effet "faible" est un effet qui reste intéressant mais qu'on ne peut observer "à l'œil nu", autrement dit qui requiert la mise en œuvre d'un appareillage (statistique) pour sa mise en évidence. La problématique de la négligeabilité de l'effet n'est pas posée, la seule contrainte que retient Cohen a trait à la précision expérimentale :

> ""Small" effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity is a bootless task, yet not so large as to make them fairly perceptible to the naked observational eye." (Cohen, 1969, p. 13).

De même, le critère d'un effet "fort" ne doit pas être si élevé que tout calcul statistique serait sans intérêt.

Résultats

Sur les 78 articles constituant le volume 61, 70 font état de tests. Cohen caractérise chaque article par une seule valeur, la puissance moyenne des tests relatifs aux hypothèses principales, calculée pour chacun des cas d'un effet vrai supposé "faible", "moyen" ou "fort". Ainsi chaque article compte autant qu'un autre quel que soit le nombre de tests pratiqués, ce qui évite d'éventuels biais. Cohen rapporte la moyenne, la médiane et l'écart-type des 70 valeurs obtenues :

	Effet vrai		
	faible	moyen	fort
Moyenne	0.18	0.48	0.83
Médiane	0.17	0.46	0.89
Ecart-type	0.08	0.20	0.16

Tableau 3

Statistiques des puissances calculées en fonction de l'effet vrai supposé pour 70 articles du Journal of Abnormal and Social Psychology (1960)

Les résultats sont pratiquement inchangés quand on prend en compte l'ensemble des tests (4829, relatifs aux hypothèses annexes aussi bien que principales), les moyennes devenant alors respectivement : 0.20 (effet faible), 0.50 (effet moyen) et 0.83 (effet fort).

Sauf à supposer un effet vrai fort, en moyenne la puissance est peu élevée (relativement à la convention qui veut que la puissance soit d'au moins 0.80 pour être convenable). Dans ces conditions un chercheur a donc assez peu de chances de rejeter l'hypothèse nulle alors même qu'elle est fausse (une chance sur deux pour un effet moyen). Cohen en conclut que les études devraient porter sur des échantillons plus grands que ce qui se pratique habituellement, méthode la plus simple pour accroître la puissance. Il préconise de doubler ou tripler les tailles d'échantillon. À ce sujet, il a constaté que l'effectif moyen des études rapportées dans les 70 articles est de 68 (mais avec un écart-type de 55, ce qui indique que la distribution est fortement asymétrique et que la médiane est certainement beaucoup plus faible que 68).

Et cependant, comme Cohen le remarque, presque tous ces articles font état de résultats significatifs. Ceci semble contradictoire avec les résultats de puissance, à moins d'admettre l'une des deux explications suivantes :

(i) Les effets parents sont forts, option que Cohen n'a pas l'air de soutenir vraiment. À ce propos on regrettera que, parallèlement aux résultats sur la puissance, Cohen n'ait pas fourni de statistiques sur la grandeur des effets *observés*, ce qui aurait apporté des informations. Ces informations peuvent être en partie trouvées dans l'article de 1989 de Seldmeier et Gigerenzer (*cf.* ci-dessous) qui ont estimé les grandeurs d'effet, sous forme d'un coefficient de corrélation *r*, à partir d'un échantillon aléatoire de 20 articles parmi les 70. Ils font état d'une médiane de 0.31 (pour des valeurs allant de 0.12 à 0.69), ce qui correspond plutôt à un effet moyen à faible selon le critère de Cohen pour $|\rho|$. Bien entendu c'est à la grandeur de l'effet parent que Cohen se réfère, mais compte tenu de l'observation 0.31, l'hypothèse d'un effet parent élevé n'est pas celle qui semble la plus plausible.

(ii) Il existe un biais dans la sélection pour publication des articles, et c'est plutôt l'avis de Cohen. On retrouve là l'argument déjà évoqué dans la sous-section 2.1.10.

Pour Cohen, les études de puissance devraient être systématiquement pratiquées en préalable à l'expérimentation. Cela permettrait de calculer l'effectif nécessaire en fonction des buts de l'étude (taille de l'effet qu'on souhaite mettre en évidence) et éviterait ainsi une perte de temps et de moyens. C'est d'ailleurs ce qui est couramment pratiqué dans le cadre des essais cliniques en médecine et en pharmacologie.

5.1.2. La réplique de Seldmeier et Gigerenzer (1989)

En 1989, Seldmeier et Gigerenzer effectuent une réplique de l'étude de Cohen, dans le but de déterminer l'impact des études de puissance sur les articles publiés.

Ils commencent par rapporter les résultats d'études semblables à celle de Cohen dans les domaines de la psychologie appliquée, de la sociologie, des sciences de l'éducation, de la communication et du marketing. Ces résultats sont assez voisins, dans la majorité des cas, de ceux obtenus par Cohen.

Puis ils réanalysent tous les articles de l'année 1984 du *Journal of Abnormal Psychology* (Le *Journal of Abnormal and Social Psychology* s'étant scindé en celui-ci d'une part et en le *Journal of Personality and Social Psychology* d'autre part) à la manière de Cohen et en conservant, autant que possible, les mêmes critères de grandeur d'effet que dans l'étude originelle de 1962, pour des raisons évidentes de comparabilité. Sur les 64 expériences considérées, ils obtiennent respectivement 0.21, 0.50 et 0.84 comme moyennes des puissances pour un effet vrai supposé "faible", "moyen" ou "fort" (les médianes sont respectivement 0.14, 0.44 et 0.90). La médiane des effets observés (calculée uniquement sur un échantillon aléatoire de 20 articles), en termes de *r*, est de 0.27 (pour 0.31 en 1960). Ainsi, en 24 ans, les choses n'ont pratiquement pas changé; ce qui est cohérent avec la constatation, par Seldmeier et Gigerenzer, que la puissance n'est pratiquement jamais mentionnée dans les articles analysés.

Par ailleurs les auteurs mentionnent sept expériences (soit 11% des cas) où l'hypothèse de recherche, identifiée à l'hypothèse nulle, est considérée confirmée par un résultat de test non significatif (pour l'année 1960, Cohen ne fait pas mention de tels cas). La médiane de la puissance des tests en question, sous l'hypothèse d'un effet parent moyen, est seulement de 0.25 (les valeurs s'étendant de 0.13 à 0.67), ce qui fait dire aux auteurs que ces expériences sont sans valeur empirique. Ce jugement est d'autant plus justifié qu'on remarquera que cette valeur (0.25) est nettement inférieure à celle de la médiane portant sur l'ensemble des résultats (0.44).

Pour Seldmeier et Gigerenzer, les raisons de cette "imperméabilité" des psychologues aux analyses de puissance sont de deux ordres. D'une part, la théorie de Fisher étant apparue la première, les psychologues ont été habitués à celle-ci, et les conceptions de Neyman et Pearson ne leur ont été enseignées que tardivement, et encore, à travers le prisme de ce qu'ils appellent "la logique hybride" (*cf.* 1.5.). Ceci fait, entre autres, que l'accent n'est pas mis sur le choix de la valeur de l'hypothèse alternative, pourtant nécessaire au calcul de la puissance. D'autre part, les erreurs d'interprétations (*cf.* 2.2.) font que les psychologues pensent que le *p* (ou l'α) suffit à répondre à leurs questions, et la notion de puissance leur apparaît donc superflue.

5.1.3. L'étude de Haase *et al.* (1982)

Alors que pour juger de l'intensité d'un effet Cohen (1962) établit des conventions (*cf.* plus haut sa définition d'un effet faible, moyen ou grand), l'intention de Haase *et al.* (1982) est au contraire de fournir à cette fin une base empirique par la compilation d'un très grand nombre de résultats. Pour cela ils recensent la grandeur des effets, exprimée par le coefficient η^2 (part de variance "expliquée" par le facteur expérimental), calculé à partir des tests statistiques rapportés dans les articles parus dans le *Journal of Counseling Psychology* durant dix ans, de 1970 à 1979. Leur choix du coefficient η^2 repose sur le fait qu'il est descriptif, donc indépendant des effectifs, normé (entre 0 et 1), ce qui permet des comparaisons entre situations différentes (calculs à partir de *t* et

de *F* aussi bien que de χ^2), et qu'il peut être calculé à partir du minimum d'information habituellement présent dans les articles. En fait, dans le cas de plans à plusieurs facteurs toute l'information nécessaire est rarement présentée et ils ne calculent alors que le η^2 partiel – limité aux seuls facteurs considérés dans la comparaison – ce qui implique une surestimation par rapport à η^2, qu'ils estiment cependant légère. Les tests multivariés sont exclus de l'étude, de même que les articles portant essentiellement sur des problèmes de validité, de test-retest d'échelles, où l'on peut s'attendre à des effets beaucoup plus grands que dans d'autres situations expérimentales.

Au total 11044 valeurs de η^2 sont calculées sur l'ensemble des 10 années. Les résultats sont les suivants :

η^2	Fréquence	%
.00-.09	5988	54.2
.10-.19	2002	18.2
.20-.29	965	8.7
.30-.39	641	5.8
.40-.49	454	4.1
.50-.59	336	3.0
.60-.69	258	2.4
.70-.79	196	1.8
.80-.89	125	1.1
.90-1.0	79	0.7

Tableau 4

Distribution des 11044 valeurs de η^2 *calculées à partir des résultats parus dans le* Journal of Counseling Psychology *de 1970 à 1979*

La distribution est très asymétrique. La médiane est de 0.0830 (1er quartile = 0.0428, 3ème quartile = 0.2682), la moyenne de 0.1589. Pour autant que ces données soient représentatives des recherches effectuées dans le domaine, un chercheur a donc toutes les chances (75%) d'obtenir moins de 27% de variance expliquée, et plus d'une chance sur deux d'en obtenir moins de 10%. Les auteurs notent que ces valeurs semblent faibles, mais que cela n'a rien d'extraordinaire dans le cadre des recherches sur le comportement. Ils remarquent que pour chacune des 10 années prises isolément les résultats sont relativement homogènes, la médiane variant entre 0.0713 (pour 1971) et 0.1356 (pour 1979), mais montrent une très légère tendance à l'accroissement.

Ces valeurs sont offertes par les auteurs comme base de comparaison pour évaluer grossièrement de nouveaux résultats dans un domaine comparable à celui de la psychologie de “counseling” (avec toutes les précautions d'usage sur la représentativité des présents résultats, la prise en compte de la nature des variables, des facteurs, *etc.*).

Par ailleurs les auteurs préconisent fortement de se baser sur ces résultats pour choisir la taille de l'effet entrant dans un calcul de puissance *a priori* quand il n'y a pas d'autre information disponible. Ils indiquent par exemple que dans le cas de la comparaison de la moyenne de deux groupes, un effectif par groupe de 88 est nécessaire pour atteindre une puissance de 0.80 en posant $\alpha = 0.05$ et $\eta^2 = 0.0830$.

Par rapport à Cohen (1962), Haase *et al.* remarquent que la médiane (0.0830) est un peu supérieure à ce qui correspond à un effet “moyen” chez Cohen (correspondant à $\eta^2 = 0.0588$), ce qu'ils attribuent essentiellement à leur utilisation du η^2 partiel.

5.1.4. L'étude de Clark-Carter (1997)

Clark-Carter (1997) applique la démarche de Cohen à l'analyse d'une revue britannique, le *British Journal of Psychology*. La méthode est semblable à celle de Cohen et utilise, pour les critères d'effet faible, moyen ou fort, ses définitions de 1969 (un peu moins sévères que celles de 1962). L'analyse des années 1993 et 1994 du *British Journal of Psychology* conduit à conserver 54 articles, soit 96 études et 1243 énoncés inférentiels. Les résultats sont très semblables à ceux des études précédentes, comme le montre le tableau de la puissance des tests utilisés :

	Effet vrai		
	faible	moyen	fort
Moyenne	0.17	0.59	0.82
Médiane	0.13	0.59	0.94
Ecart-type	0.13	0.29	0.23

Tableau 5

Statistiques des puissances calculées en fonction de l'effet vrai supposé pour 54 articles du British Journal of Psychology *(1993 et 1994)*

On regrettera ici encore l'absence d'analyse de la grandeur des effets observés, de même que le manque d'information sur l'existence de résultats non significatifs et leur statut.

Un seul article fait explicitement référence à la notion de puissance (alors même que celle-ci est faible, 0.40, sous l'hypothèse d'un effet vrai moyen) et 10 articles mentionnent la taille de l'effet relatif. Le choix de la taille de l'échantillon ne semble jamais avoir été guidé par des considérations de puissance des tests.

Clark-Carter montre également que les études réalisées dans des domaines où le contrôle expérimental est supposé plus strict et la validité des mesures plus grande (la mémoire, la perception, la lecture, par opposition à la personnalité, aux croyances, *etc.*) ne donnent pas lieu à une puissance plus grande, voire même au contraire. Ainsi, le tableau suivant rapporte la puissance moyenne pour deux grands types de domaines d'étude :

	Effet vrai		
	faible	moyen	fort
Mémoire/Perception/Lecture	0.17	0.59	0.82
Croyances/Personnalité/Anormal	0.21	0.66	0.88

Tableau 6

Puissance moyenne en fonction de l'effet vrai supposé et du type de domaines d'étude

Finalement, l'auteur constate que la situation par rapport à la puissance n'a toujours pas évolué en 1993-1994. Pour lui, deux raisons principales expliquent cet état de fait : l'étude de la puissance requiert, d'une part une connaissance de l'effet, or celui-ci n'est connu qu'après-coup, d'autre part un niveau de "sophistication mathématique" que ne posséderaient pas nombre de chercheurs.

5.1.5. L'étude de Freiman *et al.* (1978)

Cohen, comme Seldmeier et Gigerenzer (et très probablement Clark-Carter, bien qu'il ne le précise pas), ont essentiellement observé des résultats significatifs.

Dans le domaine médical, on trouve une étude portant spécifiquement sur les résultats "négatifs", c'est-à-dire non significatifs; il s'agit de l'étude de Freiman *et al.* (1978).

Ces auteurs ont analysé les articles publiés dans 20 journaux de médecine (principalement le *Lancet*, le *New England Journal of Medicine*, et le *Journal of the American Medical Association*) au cours des années 1960 à 1977. Ils ont retenu les résultats non significatifs, au seuil unilatéral de 0.05, d'études comparatives (un groupe traité et un groupe contrôle) dont la variable dépendante était dichotomique (taux de mortalité, taux de complication ou taux de "non amélioration") et ceci quel que soit le statut de l'hypothèse de recherche. 71 études ont ainsi été sélectionnées. La plupart de ces études concluent à l'inexistence d'un effet clinique intéressant. Une seule mentionne la prise en compte préalable des deux risques α et β.

Le calcul de puissance est effectué en prenant pour hypothèse alternative que la vraie différence entre le groupe traité et le groupe contrôle est tour à tour soit de 25%, soit de 50% du taux observé dans le groupe contrôle. Les résultats sont donnés dans le tableau suivant :

Puissance	Effet vrai 25%		Effet vrai 50%	
	Effectif	% cumulé	Effectif	% cumulé
1.00-.90	4	5.63	21	29.58
.89-.80	1	7.04	1	30.99
.79-.70	2	9.89	4	36.62
.69-.60	2	12.68	9	49.30
.59-.50	5	19.72	5	56.34
.49-.40	7	29.58	4	61.97
.39-.30	2	32.39	8	73.24
.29-.20	16	54.93	9	85.92
.19-.10	25	90.14	9	98.59
.09-.00	7	100.00	1	100.00

Tableau 7

Distribution de la puissance en fonction de l'effet vrai supposé pour 71 études "négatives" parues dans 20 journaux de médecine de 1960 à 1977

Dans l'hypothèse d'une réduction de 25%, la médiane de la puissance des tests pratiqués est seulement de 0.26 (résultat identique à celui de Seldmeier et Gigerenzer sous l'hypothèse d'un effet vrai moyen, *cf.* plus haut), et seulement 7% ont une puissance d'au moins 0.80. Dans l'hypothèse d'une réduction de 50%, la médiane est de 0.60 et 31% des tests ont une puissance d'au moins 0.80.

Freiman *et al.* en concluent que les effectifs utilisés sont le plus souvent trop faibles pour avoir une chance raisonnable de mettre en évidence une différence cliniquement intéressante et que les conclusions négatives de ces articles sont donc sujettes à caution.

5.1.6. La réplique de Moher *et al.* (1994)

Moher *et al.* (1994) ont constaté que l'article de Freiman *et al.* a été cité plus de 700 fois depuis sa parution, ce qui les laissent présumer que les chercheurs ont porté un intérêt certain aux résultats. Afin de déterminer s'il y avait une évolution de la pratique des chercheurs en médecine concernant la puissance des tests et le calcul préalable des effectifs, ce que ce nombre élevé de citations pouvait laisser entrevoir, Moher *et al.* ont analysé les années 1975, 1980, 1985, 1990 du *Lancet*, du *New England Journal of Medicine*, et du *Journal of the American Medical Association* (les principales publications étudiées par Freiman *et al.*). Les critères de sélection des articles et d'analyses sont semblables à ceux de Freiman *et al.*, à ceci près qu'en plus des études dont la variable dépendante est dichotomique, celles dont la variable dépendante est continue sont également incluses.

De 1975 à 1990, le nombre d'essais cliniques (*Randomized Controlled Trials*) rapportés dans les revues a doublé, mais le taux de résultats "négatifs" est resté stable, de l'ordre de 30%.

Au total les auteurs ont répertorié 102 études "négatives". Seulement 20% évoquent, au delà de la non significativité statistique, l'intérêt clinique de l'étude (quel que soit le sens de la conclusion). Parmi ces 102 études, 70 satisfont aux critères retenus pour le calcul de puissance et les auteurs présentent les pourcentages d'études dont la puissance est d'au moins 0.80 :

Année	Nbre articles	Effet vrai 25%	Effet vrai 50%
1975	16	12%	25%
1980	15	13%	47%
1985	15	7%	27%
1990	24	25%	42%
Total	70	16%	36%

Tableau 8

Pourcentages d'études "négatives" dont la puissance est d'au moins 0.80 en fonction de l'effet vrai supposé pour 70 études "négatives" parues dans 3 journaux de médecine de 1975 à 1990

Le nombre d'études possédant une puissance suffisante selon le critère habituel (≥ 0.80) reste faible, bien que légèrement supérieur à celui trouvé par Freiman *et al.* : sur l'ensemble des années, 16% et 36%, respectivement pour les cas d'un effet relatif moyen et fort (25% et 50%) et il ne semble pas y avoir d'amélioration au cours du temps.

Un calcul préalable de l'effectif n'a été effectué que dans le tiers des 102 études "négatives"; mais les auteurs notent un accroissement depuis 1980.

La conclusion des auteurs est que, malgré l'abondance des citations du travail de Freiman *et al.*, le comportement des chercheurs en médecine ne semble pas avoir été modifié. Ils voient deux explications à cette attitude : d'une part une difficulté des chercheurs à déterminer la valeur d'un effet cliniquement intéressant (indispensable au calcul de puissance), d'autre part la croyance que, quel que soit son résultat, un essai clinique pourra être utilisé ultérieurement dans le cadre de méta-analyses.

Nous retrouvons donc, dans le domaine des essais cliniques, sensiblement la même situation qu'en psychologie.

5.2. UNE RÉANALYSE FIDUCIO-BAYÉSIENNE

Nous présentons maintenant une réanalyse que nous avons effectuée dans une perspective plus descriptive que celle en jeu dans les études de puissance : il s'agit d'une part de recenser quels sont les abus d'interprétation des tests explicitement commis, et d'autre part de chercher à préciser quelle est la portée réelle des conclusions autorisées en ce qui concerne l'importance des effets, en relation précisément avec les abus (ou les insuffisances) des interprétations fournies par les auteurs. En retour cela permettra d'examiner si les tailles d'échantillon sont suffisantes pour obtenir des conclusions satisfaisantes sur l'importance des effets (relativement à un certain critère).

La méthode fiducio-bayésienne (qui utilise une distribution *a priori* non informative) nous servira de norme : c'est dans le cadre de cette méthode, et donc par rapport à elle, que nous tâcherons de répondre en examinant comment les conclusions tirées par les chercheurs à partir de tests statistiques usuels pourraient être prolongées ou modifiées. Cette méthode permet de choisir le type d'inférence *a posteriori*, ce que ne permettent pas, en toute rigueur, les méthodes fréquentistes (de test ou d'intervalle de confiance) de recherche de conclusion d'effet négligeable ou notable. En effet, dans le cadre fréquentiste, le type de conclusion recherché doit être fixé indépendamment des données recueillies pour assurer la validité des interprétations fréquentistes des risques de première et deuxième espèces du test ou de la confiance dans le cas de l'intervalle[16].

Dans cette réanalyse nous étudierons la grandeur des effets *relatifs*. Malgré ses inconvénients (*cf.* 3.1.), cette solution s'impose puisqu'elle permet de comparer des résultats portant sur des variables différentes et de les moyenner. Par ailleurs, il serait difficile de travailler systématiquement sur les mesures brutes, les critères de jugement n'étant pas fournis par les auteurs.

Pour faciliter la comparaison avec les études antérieures, nous avons choisi de réanalyser des articles parus dans le *Journal of Abnormal Psychology*. Nous avons retenu le volume 103 (année 1994), c'est-à-dire le plus récent disponible au moment de ce travail.

5.2.1. Méthode

Sélection des articles et des analyses

L'année 1994 du *Journal of Abnormal Psychology* comprend quatre numéros dont le premier a été éliminé de l'étude car il était entièrement consacré à une rétrospective des résultats de l'analyse factorielle. Parmi les 75 articles originaux que contiennent les autres numéros, nous en avons tiré 20 au hasard pour l'étude. Ce nombre peut paraître faible, mais il tient compte du nombre élevé de tests, plus de 35 en moyenne par article retenu, qui conduit à un nombre important de réanalyses. Nous n'avons pas opéré de distinction entre les travaux purement expérimentaux (dans lesquels il y a effectivement manipulation d'une ou plusieurs variables indépendantes) et ceux davantage portés sur l'observation (visant à caractériser une certaine population, par exemple).

[16] Remarquons cependant que nous serions sensiblement arrivé aux mêmes résultats en utilisant des intervalles de confiance au lieu de méthodes fiducio-bayésiennes.

Nous n'avons retenu que les tests univariés du *t* de Student, du *F* de Snédécor (analyse de variance) et du coefficient de corrélation[17], ce qui représente la très grande majorité des cas, et avons éliminé les tests sur les variables qualitatives (essentiellement des tests du χ^2), sauf quand il s'agit de *t* ou des *F* (cas de variables binaires). Nous avons également éliminé les analyses multivariées (analyses discriminantes, régression multiples, …)[18]. Dans les cas d'analyse de variance multivariée (MANOVA), seuls les résultats univariés (toujours présentés) ont été pris en compte. Les tests avec ajustement du risque de première espèce en cas de comparaisons multiples (correction de Bonferroni, par exemple) ont été inclus puisque cela n'influe pas sur le calcul de l'effet mais seulement sur le test.

Pour simplifier, dans la suite nous parlerons de test *F* pour signifier aussi bien un *t* qu'un *F*.

Conventions

Nous avons adopté les conventions suivantes :

- Dans chaque article, les analyses statistiques ont été classées en deux groupes, “analyses principales” ou “analyses secondaires”, selon le statut des hypothèses correspondantes (seul un article ne contient que des analyses principales). Les hypothèses principales sont celles qui se réfèrent directement aux objectifs principaux de la recherche et les hypothèses secondaires sont toutes les autres. Ainsi les analyses secondaires recouvrent aussi bien les analyses cherchant à vérifier que les groupes étudiés sont bien équivalents avant traitement ou que la manipulation expérimentale est effective, que celles qui portent sur des facteurs secondaires du plan d'expérience ou qui sont d'ordre exploratoire.

- Dans le cas du *t* de Student, nous en avons toujours pris la valeur absolue; c'est-à-dire que nous avons considéré l'effet du facteur comme toujours positif. Ce qui est légitime dans la mesure où le sens de la comparaison est arbitraire; seule l'interprétation devant en tenir compte. Nous avons procédé de même pour les coefficients de corrélation. Dans le cas de coefficients issus d'une matrice d'intercorrélations cela ne serait pas défendable si la matrice devait être traitée dans son ensemble, mais ici les coefficients ont toujours été considérés isolément.

- Comme il est habituel dans ce genre d'étude (*cf.* Haase *et al.*, 1982, par exemple), les corrélations du type “test-retest” d'échelles ou coefficients de validité ont été éliminées.

- Dans le cas de comparaisons à plusieurs degrés de liberté sur des occasions (“mesures répétées”), il n’a pas été tenu compte de l’utilisation éventuelle de la correction de Greenhouse et Geisser pour pallier le non respect de la condition de circularité et l'effet a été calculé sous l'hypothèse de circularité. Ces comparaisons étant peu fréquentes, le biais introduit est très limité.

- Dans les cas ambigus, sans indication fournie par les auteurs, nous avons considéré que le seuil indiqué était le seuil bilatéral.

- Quand était fourni le seuil observé, mais pas la statistique de test, celle-ci a été recalculée par programme.

- Dans les cas de résultats significatifs, quand était fournie seulement une indication du type “$p < 0.03$”, nous avons pris comme seuil observé la valeur indiquée (dans cet exemple 0.03) en considérant que vraisemblablement *p* devait se situer aux alentours (vraisemblablement ici entre 0.03 et 0.02). Dans les cas, très rares, où était simplement mentionné un effet significatif, sans aucune autre indication, nous avons pris comme seuil observé 0.05 (ce qui aboutit à sous-estimer la valeur de l'effet).

- Dans les cas de résultats non significatifs, quand seule était fournie une indication du type “$F < 1$” ou “$p > 0.10$” , les calculs ont été effectués en prenant la limite indiquée ($F = 1$ ou $p = 0.10$), ce qui aboutit à surestimer la valeur de l'effet. Le biais ainsi introduit peut éventuellement être important lorsqu'il s'agit de la question de la négligeabilité d'un effet parent; ainsi la négligeabilité pourrait être assurée pour le seuil *p* inconnu mais pas pour 0.10. En revanche, quand seule était fournie la mention “non significatif”, sans aucune autre

[17] Quelques corrélations partielles ont été traitées comme des corrélations simples (cela ne concerne qu’un seul article). D’autre part les coefficients de corrélations de Spearman (un seul article également) ont été traités comme des coefficients usuels de Bravais-Pearson.

[18] Parmi les 20 articles sélectionnés au départ, trois ne contenaient pratiquement que des analyses multivariées. Aussi nous les avons remplacés par trois autres.

indication, nous avons préféré ne pas effectuer la réanalyse, le biais pouvant être considérable. En définitive, cela aboutit à un biais qui s'ajoute au biais éventuel de sélection des articles publiés déjà mentionné (*cf.* 2.1.10.).

- Dans les cas de données manquantes, si leur répartition dans les différents groupes de sujets n'était pas précisée mais que leur proportion était très faible, nous avons procédé comme si les données étaient complètes (nous avons par ailleurs vérifié que la variation des degrés de liberté ne retentissait pratiquement pas sur le résultat).

- Quand la répartition d'un sous-ensemble de sujets dans les modalités d'un facteur n'était pas précisée, nous avons considéré qu'elle était proportionnelle à la répartition de l'ensemble des sujets.

- Dans le cas de comparaisons impliquant une réunion de groupes, nous avons retenu l'option d'équipondération (un même poids est accordé à chacun des groupes, plutôt qu'un poids proportionnel à l'effectif), en l'absence d'indication des auteurs (ce qui est la règle).

Outils

Pour une comparaison à un degré de liberté, nous avons utilisé les indicateurs d'effet privilégiés, associés à un contraste entre moyennes : différence de moyennes, différence de différences de moyennes (pour une interaction), *etc*.

Pour caractériser la grandeur de l'effet d'un facteur à plusieurs modalités, et plus généralement d'une comparaison à plusieurs degrés de liberté, nous avons utilisé l'indicateur proposé par B. Lecoutre (1984a, pp. 70-76), c'est-à-dire la moyenne quadratique (pondérée) des effets partiels. Par exemple, dans le cas de l'effet global d'un facteur, il s'agit de la moyenne quadratique des différences de moyennes deux à deux entre toutes les modalités du facteur. Cet indicateur présente les propriétés intéressantes d'être, par définition, homogène au cas privilégié de la comparaison de deux moyennes, et de rester inchangé si tous les effectifs sont multipliés par une même valeur.

Pour obtenir un indicateur relatif, on calibre par l'écart-type de la comparaison adjointe à celle étudiée (écart-type "intragroupe" de la distribution des mesures individuelles, représentant la dispersion expérimentale, l'"erreur" de mesure). Nous noterons E la valeur observée de cet indicateur calibré (qui inclut et généralise le coefficient d de Cohen) et ε la valeur vraie (parente).

E^2 est proportionnel au rapport des carrés moyens de la statistique F, d'où les relations suivantes avec F (et donc t^2 pour un degré de liberté) et le coefficient η^2 (partiel, dans le cas où le plan comporte plus d'un facteur) :

$$E^2 = b^2 F$$

$$E^2 = (b^2 q/v) \times \eta^2 / (1-\eta^2) \text{ soit : } \eta^2 = E^2 / (E^2 + b^2 q/v)$$

Les constantes b, q et v renvoient uniquement aux groupes : b se calcule à partir des effectifs et des coefficients des contrastes définissant la comparaison (formellement, c'est la composante de norme associée aux groupes), q est la somme des (effectif - 1) pour les groupes concernés; et v est le nombre de degrés de liberté associé aux groupes (q/v est le rapport des degrés de liberté du dénominateur et du numérateur du F).

La valeur de E peut donc être retrouvée à partir de l'indication du F (ou t), de ses degrés de liberté et des effectifs des groupes. Nous avons utilisé cette propriété dans les cas où nous ne disposions ni d'indication directe de grandeur de l'effet (de loin le cas le plus fréquent), ni de tableaux de moyennes et d'écart-types. La seule difficulté pour retrouver la grandeur de l'effet dans ce cas concerne le calcul de b. Dans le cas simple, mais fréquent, de la comparaison des moyennes de deux groupes indépendants de tailles n_1 et n_2 on a : $b^2 = 1/n_1 + 1/n_2$; dans des cas complexes, comme celui de certaines interactions, l'écriture d'un programme informatique a été nécessaire.

Dans le cas de la corrélation entre deux variables numériques, nous retenons le coefficient usuel r de Bravais-Pearson (nous notons ρ le coefficient parent).

Pour faciliter la comparaison avec les études antérieures, nous reprenons, pour définir un effet négligeable et un effet notable, les critères proposés par Cohen (1969) :

pour ε : $\varepsilon \leq 0.20$ pour un effet négligeable, $\varepsilon \geq 0.80$ pour un effet notable;

pour ρ : $\rho \leq 0.10$ pour un effet négligeable, $\rho \geq 0.50$ pour un effet notable.

Comme nous l'avons déjà remarqué (en 2.1.14.), pour Cohen la limite inférieure choisie pour caractériser un effet "petit", ne répond pas forcément à la problématique de la négligeabilité. Cependant, les limites proposées par Cohen peuvent sembler suffisamment faibles pour conduire effectivement à une conclusion d'effet négligeable (voire même trop sévères; voir, par exemple, Corroyer et Rouanet, 1994).

Les tests *F* ont été réanalysés au moyen des programmes LeBayésien (B. Lecoutre et Poitevineau, *in* B. Lecoutre, 1996) et PAC (B. Lecoutre et Poitevineau, 1992). Pour ce qui concerne les coefficients de corrélation, nous avons écrit un programme spécifique calculant une approximation de la distribution *a posteriori* (pour une distribution *a priori* non informative) selon la solution proposée par Lee (1989, pp. 168-175).

Au total, 735 réanalyses ont été réalisées; 384 à partir de *F*, et 351 à partir de coefficients de corrélation. Sept articles contiennent uniquement des *F*, et 13 à la fois des *F* et des *r*. Les comparaisons à un degré de liberté représentent l'essentiel des tests *F* (294/384, soit 76.6%).

Le croisement de l'article (1 à 20) et du type d'analyses (principale/secondaire) correspondant au test n'est pas complet : il existe 34 combinaisons (ou groupes) article×type pour les *F*, et 16 pour les *r*.

5.2.2. Commentaires sur la présentation des tests

Nous allons d'abord commenter la manière dont les tests sont présentés dans les articles, en utilisant, dans la mesure du possible, la grille qui nous a servi lors de l'analyse des ouvrages de référence (chapitre 4.).

La présentation des tests

Aucun article ne fait référence à Fisher ou Neyman et Pearson, ce qui ne surprend pas.

La présentation des résultats des tests est relativement stéréotypée et fait apparaître une forte dissymétrie entre le cas d'un résultat significatif et celui d'un résultat non significatif.

- Pour un résultat significatif, dans tous les articles sauf un, les auteurs fournissent le seuil observé ou au moins une borne supérieure de celui-ci ("$p < 0.05$" par exemple), ainsi que souvent la statistique de test (*t* ou *F*). Dans l'article qui fait exception, les auteurs indiquent que le seuil de signification a été fixé à 0.05. Dans un seul article également, les auteurs font mention de l'erreur de seconde espèce, qu'ils disent considérer plus grave que celle de première espèce.

Comme il était prévisible, 0.05 est la limite extrême admise pour la significativité. Cependant, dans les cas où le seuil observé en est proche, bien que supérieur, il n'est pas rare que soit mentionnée une "tendance".

- Pour un résultat non significatif en revanche, beaucoup moins d'informations sont généralement fournies. Il est même fréquent que ne figure que la mention "non significatif", sans aucune autre indication.

La grandeur de l'effet

Si l'on excepte les coefficients de corrélation, les grandeurs des effets ne sont pratiquement jamais mentionnées. Elles n'apparaissent en fait, sous forme de coefficient η, que dans un article, où elle sont jugées moyennes à fortes mais ne sont pas autrement utilisées pour l'interprétation. La plupart du temps, les interprétations sont simplement faites en termes de "tout ou rien" : l'effet existe ou non, peu importe son intensité.

Même pour les coefficients de corrélation, en cas de résultat non significatif, la valeur n'est pas toujours indiquée, et dans les matrices de corrélations certaines valeurs sont remplacées par "n.s.".

La présentation de l'intervalle de confiance

On ne trouve d'intervalles de confiance que dans un seul article, et encore s'agit-il d'intervalles de confiance (à 99%) pour des *odds ratio* dans le cadre de régressions logistiques (que nous n'avons pas retenues dans notre étude).

Calcul de N et puissance

Sur les 20 articles, les tailles des échantillons varient de 13 à 1508, avec une moyenne de 190 et une médiane de 72. Seuls deux articles font référence à la notion de puissance. L'un d'eux mentionne que le calcul de puissance a été réalisé avant l'expérimentation, pour s'assurer que la puissance serait d'au moins 0.90 sous l'hypothèse d'un effet moyen avec $\alpha = .05$. Dans l'autre article les auteurs mentionnent que la puissance est insuffisante pour conclure lors d'un résultat non significatif concernant un point secondaire, mais ils n'en parlent plus et ne se privent pas de conclure à une absence d'effet à propos d'un autre résultat non significatif concernant un point plus important.

Les abus

Pour les résultats significatifs, on ne peut parler d'abus dans l'usage des seuils observés *p* puisqu'ils ne servent qu'à établir la significativité et ne sont pas autrement commentés. Dans un article, cependant, "significatif" est systématiquement remplacé par "fiable" (*reliable*), et "non significatif" par "non fiable", ce que l'on peut considérer comme un cas d'interprétation de 1 - *p* comme la probabilité de reproduire le résultat observé (*cf.* 2.2.2.).

Pour les résultats non significatifs (tous les articles en contiennent au moins un), dans exactement la moitié des articles (10), on constate des conclusions explicites d'absence d'effet parent, du genre "il n'y a pas d'effet du facteur *x*" ou "il n'y a pas de différence entre les groupes" (bien sûr ne sont pas comptés comme tels les énoncés du type "il n'y a pas d'effet *significatif* …"). Il est parfois difficile de distinguer entre un énoncé descriptif et un énoncé inférentiel, et le décompte précédent a été fait en laissant aux auteurs le bénéfice du doute. Comme nous le verrons plus loin, il est en fait très rare que de telles conclusions se justifient, même en comprenant "il n'y a pas d'effet" comme "l'effet est négligeable".

Deux de ces articles prennent, de manière plus ou moins explicite, pour hypothèse nulle l'hypothèse de recherche, soit sur l'ensemble des articles une proportion de 10% analogue à celle de 11% trouvée par Seldmeier et Gigerenzer en 1984. La non significativité du résultat y est directement interprétée comme une corroboration de cette hypothèse, sans recourir à une procédure statistique adaptée (il n'est même jamais fait mention de la puissance du test utilisé). Dans l'un des articles la conclusion est en outre "renforcée" en invoquant des corrélations elles-mêmes non significatives sensées montrer l'indépendance de deux composantes (la possibilité d'une dépendance non linéaire n'étant même pas évoquée)[19].

5.2.3. Résultats des réanalyses

Les effets observés

Nous présentons ci-après les résultats concernant, d'une part les grandeurs d'effet calibrés *E* associés aux tests *F* et d'autre part les coefficients de corrélation *r*. Le Tableau 9 fournit les moyennes et médianes calculées séparément pour les analyses principales et pour les analyses secondaires. Dans chaque cas nous donnons deux moyennes. La première, qualifiée d'"équipondérée", correspond à l'attribution à chaque test d'un poids identique; la seconde, qualifiée de "pondérée", correspond à l'attribution à chaque test d'un poids inversement proportionnel au nombre de tests du même type (analyse principale ou analyse secondaire) dans l'article auquel il appartient. Cela revient à considérer que l'"unité" d'analyse est soit le test soit l'article×type. On retrouve ces deux options dans les études précédentes; par exemple, Haase *et al.*, 1982, prennent le test pour unité, alors que Cohen, 1962, choisit l'article (une troisième possibilité serait de ramener l'unité à l'expérience, *cf.* par exemple Cohen, 1962). On retrouve la même distinction au niveau des médianes. La médiane "équipondérée" est la médiane calculée à partir de l'ensemble des tests, alors que la médiane "pondérée" est la médiane calculée uniquement sur les *moyennes* des groupes.

[19] Dans aucun de ces cas, l'analyse fiducio-bayésienne ne permet d'obtenir une conclusion d'effet parent négligeable avec une garantie d'au moins 0.90.

Effet calibré observé E	Toutes analyses (N = 384)	Analyses principales (N = 224)	Analyses secondaires (N = 160)
Moyenne équipondérée	0.674	0.715	0.618
Médiane “équipondérée”	0.528	0.619	0.437
Moyenne pondérée	0.794	0.834	0.750
Médiane “pondérée”	0.726	0.749	0.718

Coefficient de corrélation observé r	Toutes analyses (N = 351)	Analyses principales (N = 175)	Analyses secondaires (N = 176)
Moyenne équipondérée	0.273	0.268	0.278
Médiane “équipondérée”	0.220	0.230	0.220
Moyenne pondérée	0.288	0.240	0.325
Médiane “pondérée”	0.220	0.196	0.297

Tableau 9

Grandeurs des effets calibrés E et des coefficients de corrélation r : moyennes et médianes en fonction du type d'analyse et pour l'ensemble des analyses

Les Tableaux 10 et 11 et les Figures 1 et 2 donnent les distributions de *E* et de *r* sur l’ensemble des analyses; les résultats correspondants pour chacun des deux types d’analyses sont fournis en annexe B dans les Tableaux et Figures B1 à B4.

Effet calibré E

Ensemble des analyses, principales et secondaires ($N = 384$)

Effet calibré	*n*	%
[0.0-0.2]	47	12.2
]0.2-0.4]	91	23.7
]0.4-0.6]	72	18.8
]0.6-0.8]	50	13.0
]0.8-1.0]	39	10.2
]1.0-1.2]	42	10.9
]1.2-1.4]	16	4.2
]1.4-1.6]	9	2.3
]1.6-1.8]	5	1.3
]1.8-2.0]	3	0.8
]2.0-5.2]	10	2.6

Tableau 10

Distribution des effets calibrés E

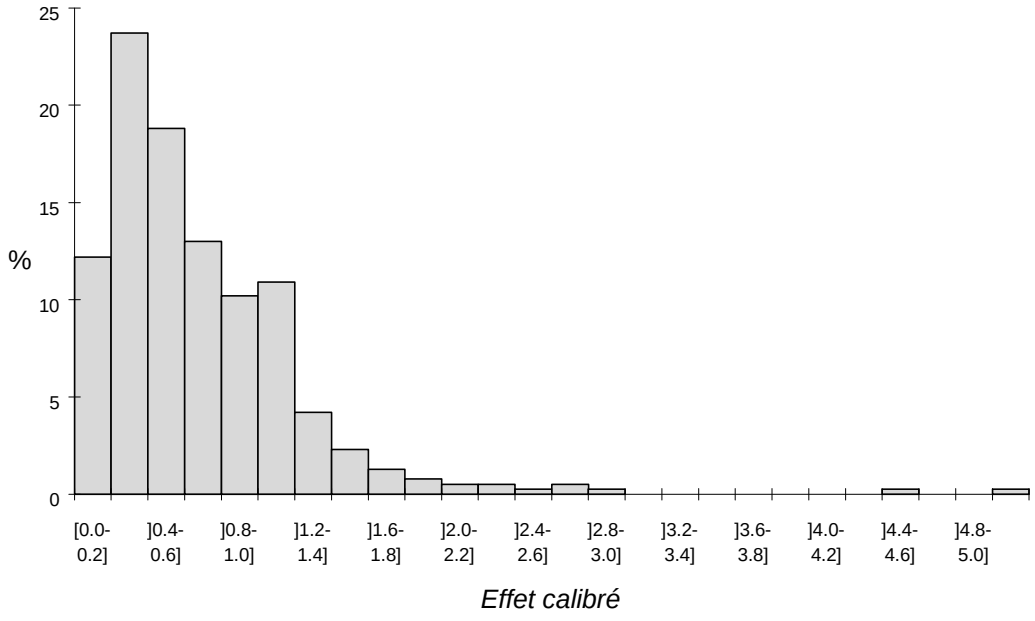


Figure 1

Histogramme des effets calibrés E

Coefficient de corrélation r

Ensemble des analyses, principales et secondaires (*N* = 351)

r	*n*	%
[0.0-0.1]	84	23.9
]0.1-0.2]	76	21.7
]0.2-0.3]	61	17.4
]0.3-0.4]	55	15.7
]0.4-0.5]	25	7.1
]0.5-0.6]	16	4.6
]0.6-0.7]	16	4.6
]0.7-0.8]	10	2.8
]0.8-0.9]	7	2.0
]0.9-1.0]	1	0.3

Tableau 11

Distribution des coefficients de corrélation r

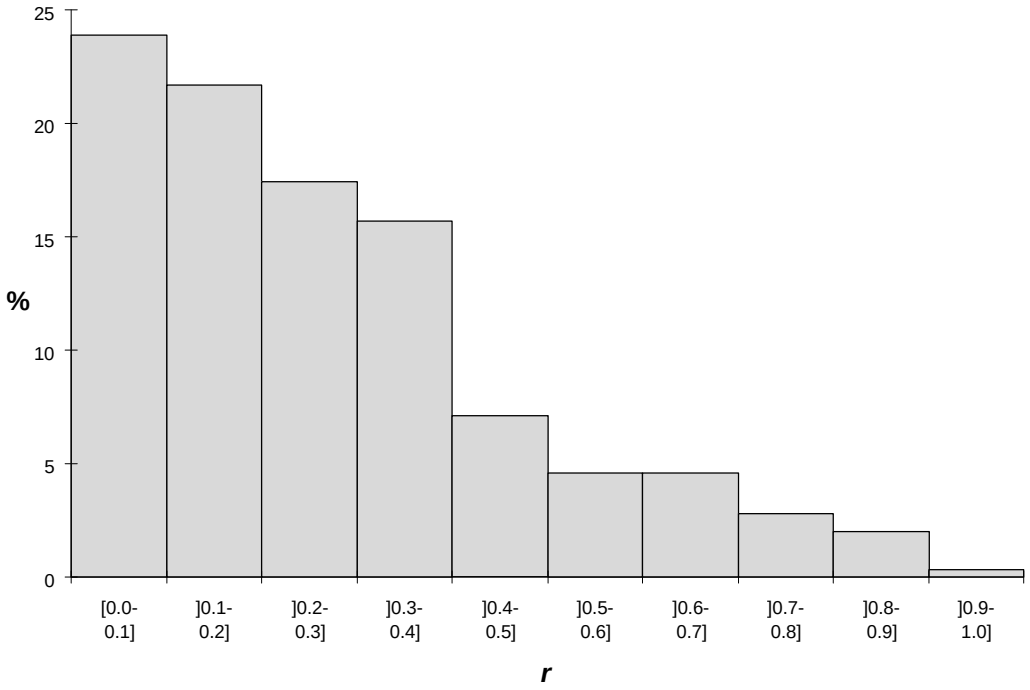


Figure 2

Histogramme des coefficients de corrélation r

Les effets calibrés *E* ont des distributions nettement asymétriques, avec une majorité d'effets moyens (entre 0.20 et 0.80). Les moyennes et médianes sont supérieures dans le cas des analyses principales, la différence moyenne entre les deux types d'analyses étant de l'ordre de 0.10. Ceci n'est pas étonnant, dans la mesure où, pour les analyses secondaires qui sont souvent destinées à montrer l'équivalence de groupes, on peut s'attendre à des effets plus faibles : on observe effectivement pour ces analyses davantage de valeurs très faibles et moins de valeurs élevées. Néanmoins les histogrammes des analyses principales et secondaires ont une forme similaire, la classe modale étant]0.2-0.4] dans les deux cas.

L'étendue des valeurs des effets calibrés est considérable : de 0 à 5.065 (de 0.010 à 4.504 si on se limite aux analyses principales). Les moyennes et les médianes pondérées sont relativement élevées; ainsi la moyenne pour les analyses principales dépasse le critère d'effet notable et la médiane en est proche. Si certains effets sont triviaux[20], l'importance des moyennes ne peut pas pour autant être réduite à quelques cas atypiques puisqu'on constate que 37% des effets observés des analyses principales sont notables (dépassent 0.80).

Les coefficients de corrélation *r* ont une distribution très fortement asymétrique et montrent des effets nettement moins élevés que pour les effets calibrés (dans la mesure où les critères sont comparables), avec des moyennes et des médianes égales ou légèrement supérieures au critère d'effet négligeable ($r \leq 0.10$). Près d'un quart des corrélations observées sont d'ailleurs négligeables. La différence entre les analyses principales et secondaires est faible, et, à l'inverse du cas précédent, c'est pour les analyses secondaires que les moyennes et médianes des corrélations sont généralement les plus grandes, surtout pour la solution pondérée.

Globalement, on constate donc dans tous les cas considérés que plus d'un effet observé sur deux est moyen, ce qui n'est pas surprenant au regard des études antérieures. Pour permettre de mieux situer nos résultats par rapport à ces études antérieures, le tableau suivant fournit les moyennes et médianes des coefficients η et η^2 :

Ensemble *F* et *r*

	Toutes analyses		Analyses principales		Analyses secondaires	
	η	η^2	η	η^2	η	η^2
moyenne	0.293	0.131	0.306	0.136	0.279	0.124
médiane	0.249	0.062	0.276	0.076	0.219	0.048

Tableau 12

*Moyennes et médianes des 735 grandeurs d'effet (*F *et* r *mélangés), en termes de coefficients* η *et* η^2*, en fonction du type d'analyses et pour l'ensemble des analyses*

Ces valeurs sont comparables à celles données par Seldmeier et Gigerenzer, 1989 (*r* médian = 0.31 pour 1960 et 0.27 pour 1984, aussi bien pour l'ensemble des analyses que pour les principales), Haase *et al.*, 1982 (η^2 moyen = 0.159, médian = 0.083), ou encore à celles trouvées par Fowler (1985) qui a effectué deux réanalyses d'articles du *Journal of Applied Psychology*, pour 1975 et pour 1980 (ω^2 médians = 0.050 et 0.078). La comparaison avec les résultats de Seldmeier et Gigerenzer est particulièrement intéressante puisqu'il s'agit du même journal. Il semble donc que la grandeur des effets relatifs n'ait pratiquement pas évolué depuis 1960.

En conséquence, sur la base des coefficients η et η^2, on s'attendrait donc pour les articles de notre étude à des grandeurs d'effet tout juste moyennes. Or, descriptivement, les effets observés sont plutôt forts, en moyenne, en ce qui concerne les hypothèses principales et les effets calibrés *E*. C'est que le calibrage de ces deux indicateurs est différent, η étant rapporté à la variabilité totale, et *E* uniquement rapporté à la variabilité "intra-groupes". On rejoint l'argument de Rosenthal et Rubin (1982), évoqué en 3.1., que les indicateurs en part de variance totale masquent la grandeur "réelle" de l'effet. Il y a tout lieu de supposer qu'il en va de même dans les études antérieures et que donc les effets observés en psychologie (dans les domaines couverts par les journaux étudiés, en tous cas) ne sont pas si faibles qu'il peut être dit.

[20] C'est par exemple le cas de celui de 5.065 qui est l'effet sur le masquage visuel du SOA (*Stimulus Onset Asynchrony*), durée séparant la présentation du stimulus cible de celle du masque.

Les résultats significatifs/non significatifs

Le Tableau 13 croise les conclusions descriptives sur l'effet observé avec le résultat du test de signification, significatif ou non significatif sur l'ensemble des analyses; les résultats séparés pour chacun des deux types d'analyses, qui sont fournis en annexe B dans les Tableaux B5 et B6, ne font apparaître que des différences relativement faibles entre ces deux types.

Effet calibré

Toutes analyses	*Conclusion sur l'effet observé E*			
test	négligeable	moyen	notable	
Significatif	4 *(1.5%)*	135 *(51.1%)*	125 *(47.3%)*	*N*=264 *(68.7%)*
N.S.	43 *(35.8%)*	77 *(64.2%)*	0	*N*=120 *(31.2%)*
Total	47 *(12.2%)*	212 *(55.2%)*	125 *(32.6%)*	*N*=384

Coefficient de corrélation

Toutes analyses	*Conclusion sur le coefficient observé r*			
test	négligeable	moyen	notable	
Significatif	4 *(2.2%)*	124 *(68.9%)*	52 *(28.9%)*	*N*=180 *(51.3%)*
N.S.	80 *(46.8%)*	91 *(53.2%)*	0	*N*=171 *(48.7%)*
Total	84 *(23.9%)*	215 *(61.2%)*	52 *(14.8%)*	*N*=351

Tableau 13

Croisement du résultat du test de signification et des conclusions descriptives, pour l'effet calibré (sous-tableau supérieur) et pour le coefficient de corrélation (sous-tableau inférieur)

Globalement les résultats significatifs sont les plus nombreux. Il n'y a cependant que 68.7% de tests significatifs pour les tests *F*, et surtout 51.3% "seulement" pour les tests sur *r* (en se limitant aux analyses principales, on trouve respectivement 76.3% et 45.7%). En raison des biais de sélection souvent évoqués, on aurait pu s'attendre à des proportions plus importantes. Diverses raisons peuvent expliquer ce constat. D'abord, en ce qui concerne les analyses secondaires, un bon nombre d'entre elles visent à vérifier les conditions d'équivalence entre groupes de sujets; le résultat non significatif est alors celui espéré par les auteurs. En ce qui concerne les analyses principales, il faut remarquer que les études concernent assez souvent des échantillons sur lesquels on mesure un grand nombre de variables et qu'il n'est pas rare qu'il s'agisse d'études exploratoires. Il est alors moins surprenant que sur la "masse" des mesures un nombre appréciable de celles-ci soient non significatives. Ceci vaut particulièrement pour les coefficients de corrélation. Il faut également tenir compte de la manière dont les analyses ont été classées. Par exemple si les hypothèses importantes concernaient les effets principaux de deux facteurs, l'étude de leur interaction, bien que ne donnant pas lieu à une hypothèse particulière, était le plus souvent rapportée parmi les analyses principales. Ces interactions non significatives ont d'autant abaissé la proportion d'analyses principales significatives.

Il n'en reste pas moins que le nombre relativement élevé de résultats non significatifs met en évidence la nécessité d'une méthodologie adaptée à ce cas.

En ce qui concerne les effets calibrés, comme nous l'avons déjà remarqué précédemment, les effets observés notables sont assez nombreux. De fait, quand le résultat est significatif, les effets observés notables (47.3%) sont presque aussi nombreux que les effets moyens.

Pour les effets calibrés, comme pour les corrélations, tous les effets observés notables sont d'ailleurs significatifs, et les effets négligeables sont presque tous non significatifs (il n'y a qu'environ 2% d'effets négligeables significatifs). Par contre, comme on pouvait s'y attendre, les effets moyens sont partagés, avec

seulement une majorité limitée (respectivement 135/212 = 63.4% pour *E* et 124/215 = 57.8% pour *r*) de résultats significatifs.

On voit ainsi que, d'une certaine manière, compte tenu des effectifs utilisés, les informations apportées par les critères d'importance de l'effet observé considérés ici et par le test de signification sont redondantes. Il en résulte que, si l'on se contente dans les publications d'ajouter au résultat du test un indicateur de la grandeur de l'effet observé, l'utilisation des critères proposés par Cohen crée pour le chercheur des situations propices à la généralisation systématique du résultat [effet observé notable (resp. négligeable) et test significatif (resp. non significatif)] en une conclusion inférentielle d'effet vrai notable (resp. négligeable). Les situations qui pourraient paraître "conflictuelles" au chercheur, [effet observé notable (resp. négligeable) et test non significatif (resp. significatif)] (voir notre expérience 1 au chapitre 6) ne se rencontrent guère dans les publications au vu de notre réanalyse. Mais une telle généralisation peut être largement abusive et elle ne peut en tout état de cause être justifiée formellement que par l'utilisation de procédures inférentielles appropriées. Les réanalyses fiducio-bayésiennes permettront précisément d'examiner les conclusions sur l'importance des effets vrais qui peuvent être réellement obtenues.

Réanalyses fiducio-bayésiennes : types de conclusions recherchées

Rappelons d'abord que, pour une comparaison à un degré de liberté avec un seuil observé bilatéral *p*, 1-*p*/2 peut être réinterprété comme la probabilité fiducio-bayésienne que l'effet vrai soit de même signe que l'effet observé, ce qui fournit une première réanalyse évidente. Mais on peut évidemment aller au delà et rechercher des conclusions relatives à l'importance des effets. En nous fixant une garantie de 0.90, nous retiendrons les cinq types d'énoncés suivants, mutuellement exclusifs, pour un effet unidimensionnel orienté (cas d'un contraste entre moyennes et du coefficient de corrélation), que nous supposerons positif :

- Recherche d'une conclusion d'effet négligeable.
Il s'agit de montrer que la valeur absolue de l'effet vrai est plus petite que la borne de négligeabilité :
$Pr(|\varepsilon| < 0.20) \geq 0.90$ (dans le cas d'un *F*)
$Pr(|\rho| < 0.10) \geq 0.90$ (dans le cas d'un *r*)

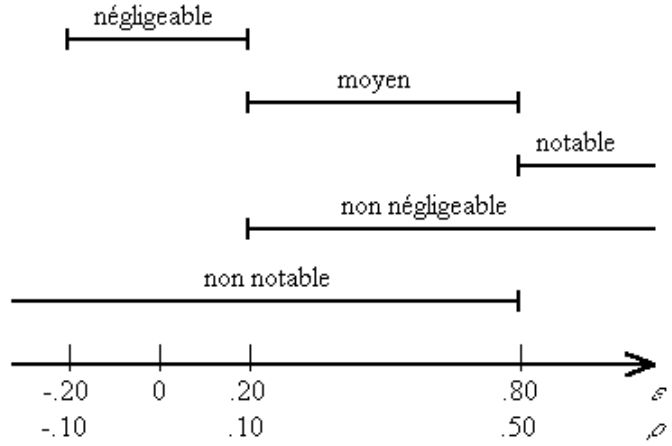
- Recherche d'une conclusion d'effet notable.
Il s'agit de montrer que l'effet est plus grand que la borne de notabilité :
$Pr(\varepsilon > 0.80) \geq 0.90$ (dans le cas d'un *F*)
$Pr(\rho > 0.50) \geq 0.90$ (dans le cas d'un *r*)

- Recherche d'une conclusion d'effet moyen.
Il s'agit de montrer que l'effet est compris entre les bornes de négligeabilité et de notabilité :
$Pr(0.20 < \varepsilon < 0.80) \geq 0.90$ (dans le cas d'un *F*)
$Pr(0.10 < \rho < 0.50) \geq 0.90$ (dans le cas d'un *r*)

- Recherche d'une conclusion d'effet non négligeable (ou au moins moyen) dans le sens de l'effet observé.
Il s'agit de montrer que l'effet est *au moins* moyen (non négligeable), sans que l'on puisse obtenir ni de conclusion d'effet moyen (c'est-à-dire limité aussi supérieurement), ni de conclusion d'effet notable :
$Pr(\varepsilon > 0.20) \geq 0.90$ (dans le cas d'un *F*)
mais $Pr(\varepsilon > 0.80) < 0.90$ et $Pr(0.20 < \varepsilon < 0.80) < 0.90$
$Pr(\rho > 0.10) \geq 0.90$ (dans le cas d'un *r*)
mais $Pr(\rho > 0.50) < 0.90$ et $Pr(0.10 < \rho < 0.50) < 0.90$
Cet énoncé correspond à l'étiquette "non négl." dans les tableaux suivants.

- Recherche d'une conclusion d'effet non notable.
Il s'agit des cas où aucune des conclusions précédentes ne peut être obtenue, mais où on peut cependant exclure la conclusion d'effet notable :
$Pr(\varepsilon < 0.80) \geq 0.90$ (dans le cas d'un *F*)
$Pr(\rho < 0.50) \geq 0.90$ (dans le cas d'un *r*)

Soit, résumé graphiquement :



Pour un effet multidimensionnel, dans la mesure où l'indicateur numérique ε est seulement un résumé grossier de l'effet, seuls les cas de recherche de conclusions d'effet négligeable et d'effet non notable se généralisent de manière univoque. Les autres cas soulèvent notamment la question de la définition de la notion d'effet multidimensionnel notable (ou non négligeable, la distinction entre ces deux cas étant seulement relative). Néanmoins les types d'énoncés précédents peuvent encore être utilisés pour l'indicateur ε (dans ce cas bien entendu toujours positif), en y ajoutant la condition supplémentaire que le résultat du test doit être significatif pour pouvoir rechercher une conclusion autre que celles d'effet négligeable et d'effet non notable[21] (pour plus de détails, voir B. Lecoutre, 1984a).

Pour préciser la possibilité d'obtenir chacun des types d'énoncés précédents et pour avoir alors un ordre de grandeur des effectifs minimaux nécessaires, nous donnons dans le Tableau 14 ces effectifs, d'une part pour différentes valeurs de la différence (calibrée) observée *E* (il s'agit dans ce cas de l'effectif total), dans le cas de la comparaison des moyennes de deux groupes indépendants équilibrés, et d'autre part pour différentes valeurs du coefficient de corrélation observé *r*. Les valeurs considérées correspondent au milieu des intervalles quand ceux-ci sont finis. Les cases grisées correspondent à des impossibilités.

[21] Notamment, dans le cas d'un effet notable, alors que pour un effet unidimensionnel signé, l'énoncé fiducio-bayésien $Pr(\varepsilon > 0.80) \geq 0.90$ implique que le résultat du test est significatif (du moins au seuil unilatéral 0.05), il n'en va pas de même pour un effet multidimensionnel.

conclusion sur l'effet vrai	différence observée $E = 0$	$E = 0.5$	$E = 1$	$E = 1.5$
négligeable: $Pr(\lvert\varepsilon\rvert < 0.20) \geq 0.90$	270			
moyen: $Pr(0.20 < \varepsilon < 0.80) \geq 0.90$		124		
notable: $Pr(\varepsilon > 0.80) \geq 0.90$			188	20
non négligeable: $Pr(\varepsilon > 0.20) \geq 0.90$		76	14	8
non notable: $Pr(\varepsilon < 0.80) \geq 0.90$	12	76		
significatif à 0.05 (bilatéral)		64	14	10

conclusion sur la corrélation vraie	corrélation observée $r = 0$	$r = 0.30$	$r = 0.75$
négligeable: $Pr(\lvert\rho\rvert < 0.10) \geq 0.90$	268		
moyen: $Pr(0.10 < \rho < 0.50) \geq 0.90$		55	
notable: $Pr(\rho > 0.50) \geq 0.90$			10
non négligeable: $Pr(\rho > 0.10) \geq 0.90$		38	≤ 4
non notable: $Pr(\rho < 0.50) \geq 0.90$	6	29	
significatif à 0.05 (bilatéral)		41	5

Tableau 14

Effectif total minimum pour pouvoir conclure aux différents types d'énoncés en fonction de la valeur observée de l'effet calibré (sous-tableau supérieur) ou du coefficient de corrélation (sous-tableau inférieur). En grisé, les impossibilités

Il apparaît qu'un énoncé de négligeabilité est le plus difficile à obtenir puisque même dans le cas le plus favorable, un effet observé égal à zéro, cela requiert de disposer d'au moins 270 sujets en tout et cette exigence s'accroît assez rapidement (pour un effet observé égal à 0.05, il faut un minimum de 480 sujets au total pour pouvoir conclure, et 756 pour 0.10). Même une conclusion d'effet moyen est encore relativement exigeante en termes d'effectif. Évidemment, des énoncés plus vagues (“non négligeable”, “non notable”) sont plus faciles à obtenir. Ces effectifs sont bien sûr relatifs au choix des bornes dont dépend la largeur des intervalles correspondant aux énoncés.

Dans les résultats qui suivent, nous examinerons d’abord la question de la généralisation du résultat d’un test de signification (significatif ou non) en une conclusion sur l’importance de l’effet vrai, puis ensuite la question de la généralisation d’une conclusion descriptive sur l’importance de l’effet en une conclusion inductive du même type. Bien que les résultats relatifs à ces deux questions soient en partie redondants, il est utile de les séparer, dans la mesure où la première question renvoie aux abus d’interprétation des tests, tandis que la seconde correspond à une démarche naturelle.

Résultats significatifs/non significatifs et conclusions fiducio-bayésiennes

Le Tableau 15 croise la significativité des résultats et les conclusions fiducio-bayésiennes pour les deux types d'effets considérés (les résultats selon le type d'analyses sont donnés en annexe B dans les Tableaux B7 et B8). Les cases grisées correspondent à des impossibilités.

Effet calibré

Toutes analyses — *Conclusion sur l'effet vrai ε*

test	négligeable	moyen	notable	non négl.	non notable	?
Significatif (*N*=264)	1 *(0.4%)*	37 *(14.0%)*	49 *(18.6%)*	134 *(50.8%)*	40 *(15.2%)*	3 *(1.1%)*
N.S. (*N*=120)	6 *(5.0%)*				93 *(77.5%)*	21 *(17.5%)*
Total (*N*=384)	7 *(1.8%)*	37 *(9.6%)*	49 *(12.8%)*	134 *(34.9%)*	133 *(34.6%)*	24 *(6.2%)*

Coefficient de corrélation

Toutes analyses — *Conclusion sur le coefficient vrai ρ*

test	négligeable	moyen	notable	non négl.	non notable	?
Significatif (*N*=180)	0	72 *(40%)*	35 *(19.4%)*	49 *(27.2%)*	21 *(11.7%)*	3 *(1.7%)*
N.S. (*N*=171)	0				167 *(97.7%)*	4 *(2.3%)*
Total (*N*=351)	0	72 *(20.5%)*	35 *(10.0%)*	49 *(14.0%)*	188 *(53.6%)*	7 *(2.0%)*

Tableau 15

Croisement du résultat du test de signification et des inférences, pour l'effet calibré (sous-tableau supérieur) et pour le coefficient de corrélation (sous-tableau inférieur). En grisé, les impossibilités

Pour les effets calibrés, un résultat significatif ne permet une conclusion d'effet vrai notable que dans 18.6% des cas (17.5% et 20.4% respectivement pour les analyses principales et secondaires). La majorité des effets significatifs (64.8%) autorisent en fait la conclusion plus faible d'un effet supérieur à 0.20, c'est-à-dire non négligeable (50.8%) ou moyen (14.0%). Il existe un seul cas où un résultat significatif permet une conclusion d'effet vrai négligeable.

Un résultat non significatif ne peut être prolongé en une conclusion d'effet vrai négligeable que dans 5% des cas. Ceci n'est pas surprenant puisque pour le critère de négligeabilité retenu un tel effet est, comme il apparaît dans le Tableau 14, très difficile à mettre en évidence. Effectivement, les six cas permettant cette conclusion proviennent de deux expériences dont les effectifs sont de 508 et 1508. La conclusion majoritaire (77.5%) est donc beaucoup plus modeste et garantit seulement que l'effet n'est pas notable; il existe même une proportion appréciable de cas (17.5%) pour lesquels cette dernière conclusion n'est même pas garantie.

Les résultats sont comparables pour les coefficients de corrélation. Un résultat significatif ne permet une conclusion d'une corrélation vraie notable que dans 19.4% des cas (23.8% et 16.4% respectivement pour les analyses principales et secondaires). La majorité des résultats significatifs (67.2%) autorisent en fait la conclusion plus faible d'une corrélation supérieure à 0.10, c'est-à-dire non négligeable (27.2%) ou moyenne qui est ici plus facile à obtenir (40.0%).

Les résultats non significatifs ne permettent d'obtenir aucune conclusion d'effet négligeable, mais autorisent seulement, à de rares exceptions près, une conclusion d'effet non notable (97.7%).

En résumé, il est donc très largement abusif de prendre systématiquement les qualificatifs "significatif" et "non significatif" pour synonymes respectifs de "notable" et de "négligeable".

Conclusions descriptives et conclusions fiducio-bayésiennes

Le Tableau 16 croise les conclusions descriptives et les conclusions fiducio-bayésiennes pour les deux types d'effets considérés (les résultats selon le type d'analyses sont donnés en annexe B dans les Tableaux B9 et B10). L'étiquette "?" correspond à l'impossibilité d'obtenir aucun des types de conclusions. Les cases grisées correspondent à des impossibilités (par exemple, il est impossible de conclure à un effet notable si l'effet observé n'est pas lui-même notable).

Effet calibré

Toutes analyses

Conclusion sur l'effet vrai ε

Conclusion sur l'effet observé E	négligeable	moyen	notable	non négl.	non notable	?
négligeable (*N*=47)	7 *(14.9%)*				39 *(83.0%)*	1 *(2.1%)*
moyen (*N*=212)		37 *(17.4%)*		58 *(27.4%)*	94 *(44.3%)*	23 *(10.8%)*
notable (*N*=125)			49 *(39.2%)*	76 *(60.8%)*		0
Total (*N*=384)	7 *(1.8%)*	37 *(9.6%)*	49 *(12.8%)*	134 *(34.9%)*	133 *(34.6%)*	24 *(6.2%)*

Coefficient de corrélation

Toutes analyses

Conclusion sur le coefficient vrai ρ

Conclusion sur le coefficient r	négligeable	moyen	notable	non négl.	non notable	?
négligeable (*N*=84)	0				84 *(100.0%)*	0
moyen (*N*=215)		72 *(33.5%)*		32 *(14.9%)*	104 *(90.4%)*	7 *(3.2%)*
notable (*N*=52)			35 *(67.3%)*	17 *(32.7%)*		0
Total (*N*=351)	0	72 *(20.5%)*	35 *(10.0%)*	49 *(14.0%)*	188 *(53.6%)*	7 *(2.0%)*

Tableau 16

Croisement des conclusions descriptives et des inférences, pour l'effet calibré (sous-tableau supérieur) et pour le coefficient de corrélation (sous-tableau inférieur). En grisé, les impossibilités

Pour les effets calibrés, les effets observés notables (qui sont tous significatifs) permettent une généralisation à un effet *vrai* du même type dans 39.2% des cas (36.1% et 45.2% respectivement pour les analyses principales et secondaires). Pour les autres cas il est possible de conclure que l'effet vrai est au moins non négligeable, c'est-à-dire supérieur à 0.20.

Seuls 14.9% (respectivement 17.6% et 13.3%) des effets observés négligeables (dont presque 91% sont non significatifs) permettent une conclusion semblable pour l'effet *vrai*. Pour les effets moyens la proportion correspondante n'est que légèrement supérieure (17.4%, respectivement 21.8% et 11.4%).

Les coefficients de corrélation observés notables (qui sont tous significatifs) permettent une généralisation dans la majorité des cas (67.3%); mais il y a ici une différence importante entre les analyses principales pour lesquelles cette proportion atteint 86.4% et les analyses secondaires pour lesquelles elle n'est que de 53.3%.

Aucune des corrélations négligeables observées (dont 95% sont non significatives) ne permet d'obtenir une conclusion d'effet négligeable pour l'effet vrai. Enfin, pour une corrélation observée moyenne, une généralisation à une conclusion du même type peut être obtenue dans environ 33.5% des cas.

Même s'il est plus facile d'obtenir une conclusion d'effet notable pour une corrélation, les résultats sur l'effet calibré et sur le coefficient de corrélation vont en grande partie dans le même sens. Il faut en effet tenir compte du fait que les critères de négligeabilité et de notabilité ne sont pas directement comparables, et également du constat que nous avons fait que les coefficients de corrélation sont beaucoup plus souvent associés aux études exploratoires, même lorsqu'il s'agit d'hypothèses principales.

5.2.3. Conclusion

L'analyse bayésienne a mis en évidence les dangers de conclure à un effet notable (resp. négligeable) sur la seule base d'un résultat significatif (resp. non significatif). On constate cependant, à l'examen des articles, que si l'abus consistant à conclure à une absence d'effet en cas de résultat non significatif est particulièrement fréquent, en revanche, l'abus symétrique (conclure à tort à un effet fort en cas de résultat significatif) n'est pas réellement commis. Il existe, en cas de résultat significatif, une sorte de "tabou", puisqu'alors les auteurs ne font généralement aucun commentaire sur l'importance de l'effet vrai. On peut y voir le rôle des rapporteurs, mais peut être encore plus la crainte des chercheurs de s'exposer à des reproches. Sans doute faut-il voir là l'influence des mises en garde contre la confusion entre la significativité statistique et la significativité substantielle. Pour expliquer alors l'effet différent des mises en garde selon qu'il s'agit d'un résultat significatif ou non significatif, nous avancerons l'hypothèse suivante. Le cas d'un résultat significatif est vécu comme un "succès" et il n'est pas nécessaire d'aller plus loin dans l'interprétation puisqu'il est suffisant pour accéder à la publication. En revanche, le cas d'un résultat non significatif serait un "échec" pour le chercheur s'il s'en tenait à un constat d'ignorance. Dans le cadre des pratiques en cours, la seule issue pour éviter cet échec est de le transformer en "succès" (un énoncé informatif) en affirmant qu'il n'y a pas d'effet. La raison principale pour laquelle les rapporteurs tolèrent cet abus (au moins dans les articles que nous avons analysés) est peut être simplement qu'ils sont eux-mêmes des utilisateurs des tests soumis aux mêmes difficultés.

L'analyse fiducio-bayésienne apparaît clairement comme un garde-fou contre les abus, en particulier dans les cas de résultats non significatifs. Même les absences de conclusions peuvent être regardées comme positives dans le sens où elles mettent clairement en évidence la nécessité d'améliorer la précision expérimentale et ne peuvent être confondues avec l'absence d'effet.

* * *

En conclusion de ce chapitre, nous reviendrons sur les conséquences qui peuvent être tirées des réanalyses statistiques d'articles publiés sur la nécessité de modifier les pratiques actuelles.

Ces réanalyses statistiques, inaugurées par Cohen (1962), sont devenues maintenant relativement courante, en psychologie mais aussi dans d'autres domaines comme la médecine. La plupart du temps, il s'agit, dans une visée clairement prescriptive, d'étudier la puissance des tests utilisés par les chercheurs. Dans ce cas, les données (les effets, les statistiques de test) apparaissant dans les articles n'ont aucun rôle : pour calculer la puissance du test utilisé il suffit de connaître la structure du plan d'analyse, les effectifs, et la valeur de l'effet vrai qui est fixée par hypothèse. Les résultats de ces réanalyses sont convergents en ce qu'ils conduisent à conclure que la puissance est plutôt faible, en supposant les effets vrais faibles ou moyens.

Des études à caractère plus descriptif il semblerait ressortir que les effets observés sont relativement petits. Cependant, dans ces études, les grandeurs d'effet sont le plus souvent mesurées par des indicateurs en part de variance expliquée, alors que notre réanalyse tend à montrer que l'effet apparaît plus important si l'on utilise un autre critère pour lequel la calibration se fait par rapport à la variabilité inter sujets. Notre étude peut donc apparaître comme une illustration, une confirmation de la critique de Rosenthal et Rubin (1982) sur le "biais" des indicateurs en part de variance expliquée (*cf.* 3.1.5.). Il s'ensuit qu'une conduite raisonnable consisterait à rapporter différentes mesures d'effet, à commencer par celles portant sur les effets bruts. À cette occasion il faut souligner que la pratique de rapporter la valeur du F (et non un simple "$p < 0.05$") a au moins le mérite de permettre le calcul de l'effet calibré (à condition de fournir également l'information sur les effectifs des groupes, ce qui est pratiquement toujours le cas). Au vu de nos résultats également, on peut douter de la pertinence des résultats des études de puissance qui reposent sur l'hypothèse d'effets vrais assez faibles, d'autant qu'en général ces effets sont justement mesurés par un indicateur rapporté à la variabilité inter sujets (comme le f de Cohen).

Cependant, nous avons constaté également, surtout pour ce qui concerne les effets calibrés, qu'en général les auteurs ne se donnent pas vraiment les moyens, en termes d'effectif, de généraliser précisément les résultats observés au delà de la simple significativité. Ceci apparaît même, pour l'effet calibré, dans le cas favorable d'un effet observé notable. Cela montre que les calculs préalables d'effectif (quasiment ignorés jusqu'à présent) seraient profitables, sinon nécessaires, aux psychologues. On rejoint alors les résultats des études de puissance sur la modestie de la puissance des tests pratiqués en psychologie, seulement la conclusion est ici basée sur les effets réellement observés et non sur des valeurs hypothétiques.

CHAPITRE 6

EXPÉRIENCES AUPRÈS DES CHERCHEURS

6.1. QUESTIONNAIRES SUR L'INTERPRÉTATION DES TESTS

Oakes (1986) a proposé à 70 psychologues universitaires six assertions concernant un résultat significatif à 0.01. Ils devaient les caractériser comme vraies ou fausses (en réalité elles sont toutes fausses) puis dire si elles correspondaient à leur façon usuelle d'interpréter un test. Les résultats sont les suivant (les pourcentages sont ceux des réponses “vraie” et “oui”, respectivement aux deux questions; le total est supérieur à 100% du fait de réponses multiples) :

[1] L'hypothèse nulle est catégoriquement réfutée. (1.4% - 1.4%)
[2] La probabilité de l'hypothèse nulle est déterminée. (35.7% - 45.7%)
[3] L'hypothèse expérimentale est catégoriquement prouvée. (5.7% - 2.9%)
[4] La probabilité de l'hypothèse expérimentale peut être déduite. (65.7% - 42.9%)
[5] La probabilité que la décision prise soit fausse est connue. (85.7% - 68.6%)
[6] Une réplique a une probabilité de 0.99 d'être significative. (60.0% - 34.3%)

Seuls trois sujets ont correctement répondu “faux” à toutes les propositions (soit 4.3%).

On retrouve bien ici les erreurs d'interprétation dont il a été fait état dans la section 2.2. (le seuil comme probabilité de la vérité de l'hypothèse nulle, le complément du seuil comme probabilité de la vérité de l'hypothèse alternative ou comme probabilité d'obtenir une réplique significative). Il est par ailleurs manifeste que les sujets manquent de cohérence puisque les options [2], [4] et [5] ne donnent pas lieu au même nombre de réponses “vraie” alors qu'elles ne sont en fait que des formulations différentes d'une même proposition; seuls 23 sujets (sur les 70) donnent une réponse identiques à ces trois questions (17 répondant “vraie” et 6 répondant “faux”). Cela met en évidence l'importance de la formulation verbale des questions.

Falk et Greenbaum (1995) ont testé la prégnance de ces erreurs en interrogeant 53 étudiants en psychologie, après qu'ils ont eu deux enseignements en statistique et qu'ils ont lu l'article de Bakan (1966) qui développe largement la nature et les dangers des erreurs d'interprétation. Là aussi la tâche est de caractériser comme “vraies” ou “fausses” un certain nombre de propositions. Les résultats font apparaître que la mise en garde a peu, sinon pas, d'effet, seuls sept sujets répondant correctement (seule l'option [5] est vraie). Bien que n'y étant pas obligés, les sujets n'ont donné qu'une seule réponse “vraie”.

[1] On a prouvé que H_0 n'est pas vraie. (3.8%)
[2] On a prouvé que H_1 est vraie. (0%)
[3] On a montré que H_0 est improbable. (79.2%)
[4] On a montré que H_1 est probable. (3.8%)
[5] Aucune des réponses précédentes n'est correcte. (13.2%)

Dans ces questionnaires la réponse correcte n'est pas présentée aux sujets (chez Falk et Greenbaum l'option [5], vraie, ne donne en fait aucun énoncé précis). Peut-être est-ce parce que l'énoncé de la réponse correcte se distinguerait manifestement de ceux présentés ici car il est nécessairement plus long étant donné qu'il faut énoncer la condition à laquelle se rapporte la probabilité. On pourrait penser que cette absence entraîne une surestimation des mauvaises réponses. Mais il semble bien, selon les résultats obtenus par Freeman (1993), que la présence d'un énoncé correct n'améliore pas vraiment les résultats. À la question de savoir comment conclure dans le cas où un traitement est significativement meilleur qu'un placebo (avec $p < 0.05$), Freeman obtient les réponses suivantes dans un échantillon de 397 dentistes, docteurs et étudiants en médecine (les réponses sont mutuellement exclusives) :

[1] On a prouvé que le traitement est meilleur que le placebo. (15%)
[2] Si le traitement est inefficace, il y a moins de 5 chances sur 100 d'obtenir un tel résultat. (19%)
[3] L'effet observé du traitement est si grand qu'il y a moins de 5 chances sur 100 que le traitement ne soit pas meilleur que le placebo. (52%)
[4] Je ne sais pas vraiment ce qu'est le p et je ne veux pas répondre au hasard. (15%)

La réponse correcte [2] est très peu choisie, surtout en regard de la réponse [3], et d'autant que, le questionnaire étant envoyé par la poste, Freeman pense qu'il en résulte une sous-estimation des erreurs (les non-

répondants lui paraissant probablement plutôt ignorants de la bonne réponse). De plus, l'auteur remarque que certains sujets (sans précision du nombre) venaient de suivre un rapide cours de statistique. Or, si chez ceux-ci la proportion de réponses [4] diminue au bénéfice de la réponse [2], l'option [3] se maintient autour de 50%.

Certainement, il existe d'autres études de ce genre. Ces questionnaires étant à réponses fermées, il est évidemment difficile de savoir si ces réponses correspondent vraiment à celles que donneraient les sujets dans une situation plus libre; en particulier, il est difficile de se faire une idée des erreurs majoritaires en situation naturelle. Par ailleurs, les réponses proposées n'étant pas identiques pour les trois questionnaires, il est délicat d'en comparer les résultats, d'autant que la formulation semble avoir une grande influence. Il en ressort tout de même que la fréquence des erreurs est très grande et que parmi celles-ci une conclusion en termes de certitude ("on a prouvé que…") est nettement minoritaire. En revanche, le faible nombre de réponses [4] ("H_1 probablement vraie") dans l'expérience de Falk et Greenbaum paraît étonnant et semble plutôt en contradiction avec les résultats de Oakes pour sa propre option [4], voisine de celle de Falk et Greenbaum. Peut-être est-ce dû à l'enseignement que les sujets de Falk et Greenbaum venaient de suivre : on peut supposer qu'ils en ont retiré que le seuil ne s'appliquait qu'à l'hypothèse nulle.

6.2. ÉTUDES EN SITUATION

Tversky et Kahneman (1971) ont été les instigateurs d'une série d'expériences sur les représentations des chercheurs dans diverses situations d'inférence statistique.

Un des problèmes type qu'ils ont posé aux chercheurs a trait à la réplication d'expériences. C'est le suivant : "Supposez que vous ayez réalisé une expérience sur 20 sujets et que vous ayez obtenu un résultat significatif qui confirme votre théorie ($z = 2.23$, $p < 0.05$ bilatéral). Vous avez maintenant décidé de prendre un groupe supplémentaire de 10 sujets. D'après vous, quelle est la probabilité que les résultats soient significatifs, avec un test unilatéral, pour ce seul groupe supplémentaire ?". (Remarquons au passage que les auteurs proposent aux sujets un cas d'école puisqu'ici l'écart-type parent est supposé connu, d'où l'utilisation du "z" de la loi normale.)

La médiane des réponses de 84 psychologues interrogés lors de rencontres du *Mathematical Psychology Group* et de l'*American Psychological Association* est de 0.85. Elle est très supérieure à la probabilité prédictive bayésienne standard obtenue par les auteurs en utilisant une distribution *a priori* non informative et qui vaut 0.478.

> Dans ce cas (distribution *a priori* non informative et écart-type connu), cette probabilité prédictive est une distribution normale dont la moyenne est l'effet observé, et dont la variance est égale à la somme de la variance de la distribution fiducio-bayésienne pour l'expérience passée et de la variance de la distribution d'échantillonnage pour l'expérience à venir.
>
> Soit ici : $D \sim \mathcal{N}(d, 3\sigma^2/20)$ (d étant l'effet observé dans l'expérience passée; ici $d = z.\sigma/\sqrt{20} = 2.23\sigma/\sqrt{20}$).
>
> La statistique de test de la nouvelle expérience sera $D\sqrt{10}/\sigma$.
>
> La probabilité prédictive est donc :
>
> $Pr(D\sqrt{10}/\sigma > 1.645) = Pr(D > 1.645\sigma/\sqrt{10})$
>
> $= Pr(Z > 1.645\sqrt{2}/\sqrt{3} - 2.23/\sqrt{3}) = 0.478$ (avec $Z \sim \mathcal{N}(0, 1)$).

On remarquera également que les auteurs ne donnent aucune indication chiffrée de la variabilité des réponses, pas même l'étendue des valeurs observées. Cette attitude ne se limite pas à cet article mais est, en fait, caractéristique de toutes leurs publications. Il est donc très difficile d'apprécier et de discuter les résultats. Il est tout de même paradoxal d'offrir si peu d'informations au lecteur alors même qu'il est question, dans ces expériences, de l'interprétation des statistiques.

Une variante mettant en jeu encore plus spécifiquement la notion de puissance des tests est également étudiée. Cette fois on demande au sujet, face à un résultat d'expérience inattendu mais significatif ($t = 2.70$, $N = 40$) et prometteur du point de vue théorique, s'il conseille de répliquer l'expérience avant publication, et alors avec quel effectif.

Soixante-six des 75 personnes ayant répondu (soit 88%) se prononcent en faveur d'une réplique et la médiane des effectifs préconisés est de 20. Enfin, en supposant que la réplique est effectuée avec $N = 20$ et qu'on trouve un résultat dans le même sens que précédemment mais non significatif ($t = 1.24$) on demande au sujet quel conseil il donnerait parmi :

[1] mélanger les résultats et présenter les conclusions comme un fait (0 réponses),

[2] rapporter les résultats comme provisoires (26 réponses),

[3] recommencer avec un nouveau groupe d'effectif (21 réponses, effectif médian : 20),
[4] trouver une explication qui expliquerait la différence entre les deux groupes (30 réponses).

Ces résultats peuvent paraître paradoxaux dans la mesure où la réplique, compte tenu du faible effectif, apparaît plutôt confirmer les premiers résultats. En particulier, la réponse [1] n'est jamais donnée et la réponse majoritaire [4] est injustifiée dans le sens où rien n'indique (en supposant les variances égales) une réelle différence parente (la différence calibrée des moyennes entre les deux groupes est de 0.145). Il est patent que la plupart des chercheurs ne se rendent pas compte de la faible puissance du second test quand, à la réponse [3], ils préconisent de recommencer avec le même effectif ($N = 20$).

L'explication avancée par Tversky et Kahneman pour rendre compte de l'ensemble des résultats présentés est une sous-estimation des fluctuations d'échantillonnage, autrement dit un biais induit par une "heuristique de représentativité" : des échantillons, même petits, tirés aléatoirement d'une même population sont attendus comme très semblables à la population parente (très représentatifs) et par là même très semblables entre eux. Cette explication est générale, dans le cadre des situations mettant en jeu les notions de hasard et d'échantillonnage, et n'est pas liée aux particularités de la situation expérimentale étudiée (le test de signification). Ce phénomène, étudié sur des sujets "naïfs" (*cf.*, par exemple, Kahneman et Tversky, 1972), n'épargnerait donc pas des sujets avertis dans le domaine des statistiques, comme les psychologues (*cf.* Nisbett et Ross, 1981). À ce propos il est frappant de noter qu'il n'y a pratiquement pas de différence entre les réponses médianes des membres de l'*American Psychological Association* et ceux du *Mathematical Psychology Group*, alors qu'on peut penser que ces derniers sont entraînés à la psychologie mathématique et plutôt experts en statistiques.

Tversky et Kahneman résument leurs conclusions en caractérisant ainsi un chercheur qui croit à la "loi des petits nombres" :

- Il joue (*gambles*) ses hypothèses de recherche sur un petit échantillon, sans réaliser que les chances contre lui sont très grandes. Il surestime la puissance du test.
- Il a une confiance excessive dans les premières tendances de ses données et dans la stabilité des patrons de réponses observés. Il surestime le degré de significativité.
- Il a une confiance excessive dans la réplicabilité des résultats significatifs. Il sous-estime la largeur des intervalles de confiance.
- Il attribue rarement aux fluctuations d'échantillonnage un écart des résultats par rapport aux attentes, car il trouve toujours une explication "causale" à ces écarts. Il a donc peu de chance de reconnaître l'intervention des fluctuations d'échantillonnage. Sa croyance en la loi des petits nombres restera donc intacte.

M.-P. Lecoutre (1983), pour sa part, s'est intéressée à des situations, qu'elle qualifie de "conflictuelles", où l'on fournit des informations qui semblent contradictoires. L'étude concerne également des psychologues, principalement des expérimentalistes. Quand la contradiction est entre une statistique descriptive élémentaire (par exemple une valeur observée importante pour un effet d'interaction) et le résultat du test correspondant (non significatif), la majorité des chercheurs suspendent leur jugement. Dans le cas où la contradiction porte sur deux tests de signification (un premier résultat est significatif, une réplique non) les résultats recoupent ceux obtenus par Tversky et Kahneman (1971) dans la mesure, d'une part, où cette situation est jugée conflictuelle par les sujets (ce qui montre qu'ils attendaient une réplique significative au vu du premier test), et où, d'autre part, beaucoup rejettent l'idée de combiner les deux résultats en un seul. Un questionnaire, combinant attente *a priori* sur les hypothèses, moyenne observée et résultat du test, permet aussi de constater que la plupart des chercheurs (20 sur 23), fondent essentiellement leurs conclusions sur les résultats du test mais que, s'ils concluent unanimement, et indépendamment des effets observés et des attentes, à l'existence d'un effet en cas de résultat significatif, leurs attitudes sont très diversifiées en cas de résultat non significatif.

Oakes (1986) a proposé une explication concurrente de l'hypothèse de "représentativité", l'hypothèse de "significativité" (*significance hypothesis*). Contrairement à l'hypothèse de Tversky et Kahneman, celle de Oakes est spécifique du type d'expériences évoqué ici, car elle implique directement le test de signification.

Selon l'hypothèse de "significativité" les chercheurs perçoivent le résultat du test comme une dichotomie : un effet significatif existe et a de grandes chances d'être répliqué / un effet non significatif n'existe pas et a très peu de chance d'apparaître significatif dans une réplique. Les sujets raisonneraient, face au résultat d'un test de signification, en termes de tout ou rien, plutôt qu'en se faisant une idée (une estimation) de l'intensité de l'effet et ils ont tendance à surestimer la similitude du *résultat du test* (significatif ou non).

Selon l'hypothèse de "représentativité" les sujets ont tendance a surestimer la similitude des *statistiques de test*.

Dans certaines situations, telle celle décrite plus haut, les deux hypothèses prévoient les mêmes résultats. C'est pourquoi Oakes (1986) a introduit, dans une de ses expériences, une situation lui permettant de trancher. Il s'agit toujours d'une situation de réplique.

Un premier résultat est significatif ($z = 1.64$, $p = 0.05$, $N = 20$). Quand il demande la probabilité pour que le résultat d'une nouvelle expérience, avec $N = 40$, soit significatif à $p < 0.01$, Oakes obtient, en moyenne, 0.747 (la probabilité prédictive bayésienne standard est de 0.50); ce qui va dans le sens de son hypothèse. En effet, si les sujets s'en tenaient à la "représentativité", ils auraient plutôt attendu une même statistique ($z = 1.64$), donc *non significative* à 0.01; au contraire, si le premier résultat les conforte dans l'idée que l'effet existe, alors ils s'attendent à le voir confirmé (en termes de signification), d'autant que la taille du deuxième échantillon est plus élevée. On notera par ailleurs qu'Oakes, pour conforter son hypothèse, invoque les résultats obtenus par Rosenthal et Gaito (1963), ce qui est très contestable, comme nous le verrons en 6.4.

M.-P. Lecoutre et Rouanet (1993) ont également cherché à départager les deux hypothèses. Leur étude met en jeu trois situations (effet observé important et significatif / effet faible et non significatif / effet très faible et non significatif). Le jugement des sujets (50 chercheurs en psychologie, chacun étant confronté aux trois situations) est une prédiction sur, à la fois, la probabilité que le signe de l'effet dans une réplique de même taille ($N = 20$) soit identique et la probabilité que le test soit significatif (situation 1) ou non (situations 2 et 3). Le fait que, pour le signe de l'effet, les auteurs constatent une sous-estimation (par rapport aux probabilités bayésiennes standard) va à l'encontre de l'hypothèse de "représentativité"; contrairement à ce que prévoit cette dernière, les sujets se sont montrés très prudents quant à la généralisation des résultats obtenus sur le premier échantillon. Cela ne fortifie pas pour autant l'hypothèse de Oakes, car celle-ci ne permet pas de prédiction particulière dans ce cas. En revanche, dans le cas du test, on observe une surestimation, compatible avec les deux hypothèses. Il est remarquable également que beaucoup de sujets ont tendance à faire des prédictions voisines pour le signe de l'effet et pour le test, ce qui, selon les auteurs, pourrait indiquer que les sujets, en absence d'intuition à propos des tests, utiliseraient celle se rapportant à la statistique naturelle (ici le signe). Cette attitude pourrait même suffire à expliquer les résultats pour le test. On notera enfin que dans presque tous les cas (5 des 6 questions) la réponse modale est 0.50, mais que la variabilité est importante et montre l'hétérogénéité des sujets.

Oakes (1986) a également cherché à montrer que les chercheurs en psychologie ont tendance à surestimer la grandeur d'un effet lorsqu'ils se basent sur le degré de significativité. Il présente à 60 sujets, psychologues universitaires, la situation suivante : 500 personnes, 250 psychologues et 250 psychiatres, ont répondu à un questionnaire de tendances psychopathologiques et les psychologues ont obtenu un score moyen significativement plus élevé que les psychiatres ($p = 0.05$ exactement, t de Student bilatéral). Il demande alors aux sujets d'estimer le nombre de psychologues parmi les 250 scores les plus élevés. Les sujets répondent, en moyenne, 163, alors que la réponse correcte, selon Oakes, est de 134; soit une erreur relative de plus de 20% (celle-ci atteint même 33% quand le seuil indiqué dans la consigne est de 0.01). Il faut cependant remarquer que la réponse "correcte" ne va pas de soi; la solution de Oakes n'est qu'approximative car il remplace un paramètre par son estimation et elle dépend par ailleurs de la forme des distributions.

6.3. EXPÉRIENCE 1 : "PSYCHOLOGUES ET STATISTICIENS"

6.3.1. Buts de l'expérience

Nous avons effectué cette étude en collaboration avec Marie-Paule Lecoutre (U.F.R. de psychologie, université de Rouen). Elle se situe dans la lignée des études initiées par Kahneman et Tversky concernant le comportement de sujets en situation d'incertitude, et plus particulièrement de chercheurs dans des situations familières d'analyse statistique.

Outre la confirmation des résultats antérieurs de M.-P. Lecoutre (1983, 1991) concernant des situations conflictuelles, cette expérience a deux objectifs principaux :

(1) Introduire une question portant explicitement sur une décision telle qu'il en existe dans les situations réelles d'expérimentation : au delà d'un jugement, bien souvent le chercheur doit décider s'il va continuer son expérience ou non, s'il va soumettre pour publication ses résultats ou non, …

(2) Comparer les comportements des chercheurs en psychologie avec ceux des statisticiens de l'industrie pharmaceutique qui utilisent les mêmes outils statistiques que les psychologues mais dans un contexte manifestement différent. Les différences de culture statistique entre psychologues et statisticiens recouvrent différents aspects, notamment :

la formation : davantage “pratique” pour les psychologues et “théorique” (et plus approfondie) pour les statisticiens;
l'exercice : occasionnel chez les psychologues et quotidien chez les statisticiens;
le degré d'implication : les psychologues travaillent sur leurs propres données et hypothèses, ce qui n'est pas le cas des statisticiens industriels qui ne sont souvent que l'un des éléments de la “chaîne de décision”;
le poids des enjeux économiques : souvent absents chez les psychologues, ils sont par contre incontournables chez les statisticiens;
la référence à un cadre théorique dans le domaine de recherche, qui ne concerne pas en principe les statisticiens, mais qui est une composante fondamentale pour tous les psychologues de notre expérience.

6.3.2. Matériel

Les sujets sont mis dans la situation de se prononcer sur les résultats d'une étude (fictive) de pharmacologie ayant pour but la mise au point d'un certain médicament. Cette étude aboutit à la comparaison des moyennes de deux groupes de 15 patients chacun, un groupe expérimental qui reçoit le médicament et un groupe contrôle, sous placebo. Un critère pour juger de l'importance de l'effet du médicament est fourni aux sujets sous la forme d'un avis d'experts selon lequel la différence doit être au moins de +3 pour que le médicament soit cliniquement intéressant.

Chacun des sujets est confronté à quatre situations possibles, variant selon la différence observée (notée *d*) entre les moyennes des deux groupes et le résultat d'un *t* de Student (test bilatéral). Ces situations ont été choisies, sur la base de résultats antérieurs (voir M.-P. Lecoutre, 1983, 1991), pour apparaître comme conflictuelles (“désaccord” entre *d* et le test, situations 1 et 4) ou non conflictuelles (“accord” entre le *d* et le test, situations 2 et 3). Soit :

		$p = 0.001$	$p = 0.50$
Effet observé	faible ($d = 1.52$)	*situation 1* (conflictuelle)	*situation 2* (non conflictuelle)
	fort ($d = 6.07$)	*situation 3* (non conflictuelle)	*situation 4* (conflictuelle)

Tableau 17

Caractéristiques, en termes d'effet observé et de seuil observé, des quatre situations proposées aux sujets

Les valeurs *p* ont été choisies de façon à ce que le problème de la variation de la limite de significativité d'un sujet à l'autre ne se pose pas : 0.001 est clairement significatif (par rapport aux pratiques habituelles), et 0.50 clairement non significatif. De même, les valeurs de la différence *d* (+1.52 et +6.07) sont nettement inférieure et supérieure au critère d'intérêt (+3).

6.3.3. Consigne

Le texte précis présenté aux sujets est le suivant :

Je vous demande de vous placer dans la situation suivante.

Soit une recherche en pharmacologie pour la mise au point d'un certain médicament.

On considère les données relatives à 30 sujets:
- 15 qui ont pris le médicament (groupe g1)
- 15 qui ont pris un placebo (groupe g2)

On considère la différence observée *d* entre les moyennes des deux groupes pour une certaine variable, le test *t* de Student usuel (pour groupes indépendants) correspondant, et son seuil observé (bilatéral) *p*.

Je vais vous demander d'étudier attentivement 4 situations possibles qui se différencient selon la valeur de la différence observée *d* et le résultat du *t* de Student, et je vais vous poser successivement 3 questions.

Voici les 4 situations que je vous demande d'examiner.

Situation 1	d=+1.52	t=+3.67 *significatif* (p=0.001)
Situation 2	d=+1.52	t=+0.68 *non significatif* (p=0.50)
Situation 3	d=+6.07	t=+3.67 *significatif* (p=0.001)
Situation 4	d=+6.07	t=+0.68 *non significatif* (p=0.50)

Les médecins experts consultés considèrent que le médicament testé a un effet cliniquement intéressant si la différence pour la variable considérée est au moins égale à +3.

Question 1
Ma première question est la suivante.
Pour chacune des 4 situations, quelle conclusion tireriez-vous quant à l'effet du médicament étudié ?
À chaque fois je vous demande de justifier votre réponse

Question 2
Initialement on avait planifié l'expérimentation avec en tout 60 sujets, 30 sujets par groupe. Les résultats présentés ici sont en fait des résultats intermédiaires.
Ma deuxième question est la suivante.
En vous appuyant sur les données disponibles, quelle prédiction feriez-vous sur les résultats de l'analyse statistique que l'on obtiendrait si on allait jusqu'au bout de l'expérimentation prévue, et donc si on ajoutait 15 sujets par groupe (soit donc en tout 30 sujets par groupe), d'abord sur d puis sur t, et par suite sur la conclusion relative à l'effet du médicament étudié ?
Comme précédemment, je vous demande de justifier votre réponse.

Question 3
Économiquement, il serait intéressant d'arrêter l'expérimentation avec seulement les 15 premiers sujets par groupe.
Ma troisième question est la suivante.
Pour quelle(s) situation(s), toujours parmi les 4 considérées, prendriez-vous la décision d'arrêter l'expérimentation sur la base des données disponibles, soit avec seulement 30 sujets en tout (15 par groupe).
Comme précédemment, je vous demande de justifier vos réponses.

Pour terminer, avez-vous des commentaires à faire sur l'expérience que vous venez de passer ? sur les situations considérées ? les questions posées ? *Etc.*
Merci de votre participation.

La question 1 sert de base aux questions suivantes et est une réplique partielle des expériences antérieures de M.-P. Lecoutre.

Les questions 2 et 3 renvoient au problème de l'interruption avant terme d'une expérience sur la base d'une analyse intermédiaire. La question 2 concerne une situation de prédiction qui n'est pas "standard" pour des chercheurs en psychologie et pour laquelle les sujets ne disposent pas de réponse stéréotypée. En même temps elle est le préalable de la question suivante. L'originalité par rapport aux questions considérées par Tversky et Kahneman (1971) et par M.-P. Lecoutre et Rouanet (1993) tient ici au fait que la prédiction porte sur l'ensemble des résultats (résultats actuels + réplique) et pas seulement sur les résultats de la réplique. Cette question 2, portant sur le mélange d'une information incertaine (les données futures) et d'une information connue (les données déjà recueillies), apparaît *a priori* plus difficile que les précédentes qui portaient uniquement sur des données futures.

La question 3 introduit un élément décisionnel explicite. Si elle est peut-être moins courante dans le domaine de la psychologie que dans celui de la pharmacologie, elle correspond néanmoins à une situation naturelle pour le chercheur qui doit parfois décider, au vu de résultats partiels, s'il doit ou non continuer à expérimenter (le "coût", par exemple en temps ou en sujets, intervenant de façon informelle). Elle permet en outre d'apporter une précision par rapport à la première question dans la mesure où, par l'introduction d'une contrainte de type économique, elle force davantage le sujet à se déterminer, et est une sorte d'indicateur grossier du degré de confiance du sujet en sa réponse à la question 1.

La question 3 en particulier est intéressante d'un point de vue normatif, car elle divise les méthodologistes. Si beaucoup y voient un problème analogue à celui des comparaisons multiples et pensent que les tests de signification intermédiaires doivent être effectués à un seuil plus petit que celui utilisé en cas d'analyse unique, d'autres objectent qu'une telle procédure est en contradiction avec le fait qu'un résultat significatif à un seuil fixé est d'autant moins intéressant que l'effectif augmente (ce qui suggère de diminuer le seuil quand l'effectif augmente), ce qui rejoint les critiques des tests présentées en 2.1.13. (sans parler de l'approche bayésienne du problème). On peut donc s'attendre à ce que cette question divise également les sujets.

6.3.4. Sujets

Deux groupes de sujets ont été interrogés :
- D'une part un groupe de 20 psychologues.
Il s'agit de 19 chercheurs ou enseignants-chercheurs en psychologie, appartenant à des laboratoires de l'université ou du C.N.R.S. situés dans la région parisienne et d'un enseignant-chercheur en psychologie de l'université de Rouen. Ce sont tous des chercheurs confirmés, exerçant en psychologie générale, psychologie de l'enfant ou psychologie sociale, et utilisant couramment des tests pour la plupart, ou sachant au moins en "lire" les résultats. Aucun n'a reçu de formation particulière en statistiques autre que celle proposée dans le cadre des études de psychologie, mais quelques uns ont approfondi leurs connaissances en ce domaine par des lectures personnelles.
- D'autre part un groupe de 25 statisticiens.
Il s'agit de statisticiens de l'industrie pharmaceutique travaillant dans le domaine des essais cliniques et appartenant à différentes entreprises.

6.3.5. Passation

La passation était individuelle. Le texte de présentation était lu au sujet et était reproduit sur une feuille qui restait devant lui tout au long de la passation. Les résultats pour les quatre situations lui étaient ensuite présentés sur une feuille qui restait toujours à sa vue. L'expérimentateur répondait alors aux questions éventuelles. Typiquement celles-ci ont concerné deux points : d'une part la pertinence de la variable dépendante utilisée dans l'étude pharmacologique (on assurait au sujet que le choix de la variable ne faisait pas problème), d'autre part la confiance à accorder au critère des experts (là encore on indiquait que les experts étaient bien qualifiés et que le critère n'était pas controversé). Puis les trois questions étaient posées successivement, la réponse étant notée par l'expérimentateur ou par le sujet lui-même. À tout moment il était répondu à toute demande d'information ou de précision de la part du sujet. En général ceci s'est limité à faire préciser, au moment de la question 2, que le tirage des patients était bien aléatoire, qu'il n'y avait pas de biais de sélection.

La passation n'a pas soulevé de problème et a duré de 15 à 20 minutes, en général.

6.3.6. Résultats

D'un point de vue normatif, les situations 1 et 3 permettent de conclure, respectivement "pas d'effet cliniquement intéressant" et "effet cliniquement intéressant", et sont donc en ce sens "favorables". Les situations 2 et 4, au contraire, ne permettent pas de conclure, en raison de la grande variabilité observée. Ces réponses normatives peuvent être justifiées dans le cadre fiducio-bayésien, une distribution *a priori* non informative étant appropriée ici puisque l'on ne fournit aux sujets aucune information extérieure aux données[22]. On obtient les énoncés suivants pour l'effet vrai :

Situation 1 :	$Pr(\lvert\text{effet vrai}\rvert < 3)$	= 0.999	pas d'effet cliniquement intéressant
Situation 2 :	$Pr(\lvert\text{effet vrai}\rvert < 3)$	= 0.719	pas de conclusion
Situation 3 :	$Pr(\text{effet vrai} > 3)$	= 0.963	effet cliniquement intéressant
Situation 4 :	$Pr(\text{effet vrai} > 3)$	= 0.634	pas de conclusion

Les trois tableaux suivants (Tableau 18 à Tableau 20) présentent les résultats, en pourcentage, pour chacune des questions. Les pourcentages des réponses sont rapportés respectivement pour le groupe des

[22] Les mêmes réponses normatives peuvent être obtenues avec des procédures fréquentistes, par exemple des intervalles de confiance.

psychologues ($N = 20$) et le groupe des statisticiens ($N = 25$). ("?" représente les non-réponses ou réponses d'ignorance.)

Question 1

	situation 1 *conflictuelle*		situation 2		situation 3		situation 4 *conflictuelle*	
	Psy	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*
Efficace	45%	12%	0	0	100%	96%	0	12%
Non efficace	40%	80%	85%	84%	0	0	35%	36%
?	15%	8%	15%	16%	0	4%	65%	52%

Tableau 18

Pourcentages des réponses à la question 1 selon la situation expérimentale et le groupe de sujets (Psychologues/Statisticiens). Les cases grisées représentent les différences marquantes entre les deux groupes de sujets

Question 2

		situation 1 *conflictuelle*		situation 2		situation 3		situation 4 *conflictuelle*	
		Psy	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*
d	Même chose	55%	64%	55%	52%	70%	68%	50	52%
	↘	5%	0	5%	0	10%	0	10%	0
	↗	5%	4%	0	4%	0	0	0	0
	?	35%	32%	40%	44%	20%	32%	40%	48%
t	Même chose	50%	48%	40%	16%	60%	52%	35%	16%
	↘	10%	8%	10%	16%	15%	8%	30%	28%
	↗	5%	0	5%	0	5%	0	5%	0
	?	35%	44%	45%	68%	20%	40%	30%	56%
conclusion	Même chose	60%	76%	60%	52%	75%	84%	45%	28%
	Efficace	0	4%	0	0	0	0	20%	20%
	Non efficace	5%	4%	0	8%	0	0	0	0
	?	35%	16%	40%	40%	25%	16%	35%	52%

Tableau 19

Pourcentages des réponses à la question 2 selon la situation expérimentale et le groupe de sujets (Psychologues/Statisticiens). Les cases grisées représentent les différences marquantes entre les deux groupes de sujets

Question 3

	situation 1 *conflictuelle*		situation 2		situation 3		situation 4 *conflictuelle*	
	Psy	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*	*Psy*	*Stat*
Arrêter	60%	52%	55%	52%	75%	88%	30%	4%
Continuer	40%	48%	40%	44%	25%	12%	60%	96%
?	0	0	5%	4%	0	0	10%	0

Tableau 20

Pourcentages des réponses à la question 3 selon la situation expérimentale et le groupe de sujets (Psychologues/Statisticiens). Les cases grisées représentent les différences marquantes entre les deux groupes de sujets

Dans l'analyse des résultats, certaines des conclusions descriptives seront généralisées en utilisant des procédures bayésiennes standard. Nous donnerons dans ce cas un énoncé bayésien, associé à une garantie (probabilité) 0.90, relatif aux proportions parentes de la réponse considérée, notées respectivement φ_s pour les statisticiens et φ_p pour les psychologues. Cet énoncé portera soit directement sur ces proportions soit sur leur différence ou leur rapport. Par exemple l'énoncé $[\varphi_s - \varphi_p > 0.22]$ pour la réponse "Non efficace" signifie qu'on peut inférer, avec la garantie 0.90, que la différence entre les proportions de cette réponse chez les statisticiens et les psychologues est d'au moins 0.22 (formellement, $Pr(\varphi_s - \varphi_p > 0.22) = 0.90$).

Question 1 "Quelle conclusion tireriez-vous quant à l'effet du médicament étudié ?"

- Situation 1 (*d* faible, test significatif : situation "conflictuelle")

Cette situation est celle pour laquelle les réponses des psychologues sont les plus dispersées. La réponse (légèrement) majoritaire (45%: 9/20) est de conclure à l'efficacité clinique du médicament, sur la seule base du test significatif qui est alors assimilé à la démonstration d'un effet important (en grandeur). On retrouve là un des abus de l'utilisation des tests, qui est de confondre significativité statistique et significativité substantielle. Cette attitude peut être extrêmement affirmée; ainsi un des sujets est allé jusqu'à déclarer: "Les experts ont tort. Ils doivent revoir leur critère." Les autres sujets ont explicitement reconnu qu'ils ne tenaient pas compte du critère des experts.

Cependant presque autant de psychologues (40%: 8/20) concluent à l'inefficacité du médicament, se fixant sur la valeur faible de *d*. L'argument est alors qu'il existe bien un effet non nul, puisque le test est significatif, mais que cet effet est trop faible pour être cliniquement intéressant. Cet argument est en accord avec la réponse normative.

Trois sujets enfin, ne peuvent conclure. Pour eux *d* et le test vont dans des sens opposés, il y a conflit, et ils souhaitent poursuivre les observations pour se faire une opinion définitive. Ces sujets considèrent aussi le test significatif comme indicateur d'un effet important (sinon il n'y aurait pas de conflit), mais prennent également en compte le critère des experts.

Les statisticiens, au contraire des psychologues, marquent nettement leur homogénéité, 80% (20/25) d'entre eux concluant à l'inefficacité du médicament. Seuls 12% (3/25) concluent abusivement à l'efficacité clinique.

La différence observée des choix entre les deux groupes est donc importante : les statisticiens donnent deux fois plus souvent la réponse "Non efficace" que les psychologues (80% contre 40%), et à l'opposé nettement moins souvent la réponse abusive "Efficace" (12% contre 45%).

On peut inférer que les deux populations se différencient bien, tant pour la proportion de réponse "Non efficace" : [φ_s-φ_p > 0.22 et φ_s/φ_p > 1.43] que pour celle de la réponse "Efficace" [φ_p-φ_s > 0.16 et φ_p/φ_s > 1.90].

- Situation 2 (*d* faible, test non significatif : situation "non conflictuelle")

Cette situation donne lieu à un large consensus parmi les sujets, quel que soit le groupe : 85% des psychologues et 84% des statisticiens concluent à l'inefficacité du médicament; elle est considérée par ces sujets comme "favorable", dans la mesure où la différence observée *d* et le test vont, selon eux, dans le même sens. On retrouve ici l'abus consistant à voir dans un résultat non significatif la démonstration d'une absence d'effet. Ceci est en accord avec les résultats que nous avons obtenus dans les réanalyses d'articles, où la moitié des articles présentaient des conclusions explicites d'absence d'effet parent en cas de résultat non significatif (*cf.* 5.2.2.). Cette attitude est particulièrement manifeste chez un des psychologues qui, ayant conclu à l'inefficacité dans la situation 1, conclut de même ici mais en étant "encore plus convaincu de l'absence d'intérêt du médicament du fait que le test est non significatif". Les statisticiens, non seulement n'évitent pas cet abus, mais le commettent tout autant que les psychologues. On peut inférer que la différence entre les deux populations est limitée [$|\varphi_p$-$\varphi_s| < 0.18$], mais les effectifs des groupes sont insuffisants pour pouvoir conclure qu'elle est négligeable.

- Situation 3 (*d* fort, test significatif : situation "non conflictuelle")

Cette situation conduit à une presque unanimité des sujets (à une exception près) pour déclarer le médicament efficace. Elle est généralement perçue comme particulièrement simple et favorable.

On peut inférer une très large prédominance de la conclusion d'efficacité dans chacune des deux populations [$\varphi_p > 0.95$ et $\varphi_s > 0.89$], et au plus une différence modérée entre celles-ci [$|\varphi_p$-$\varphi_s| < 0.11$].

- Situation 4 (*d* fort, test non significatif : situation "conflictuelle")

Cette situation donne lieu dans chacun des groupes à une majorité de non-réponses : 65% (13/20) des psychologues et 52% (13/25) des statisticiens la perçoivent comme conflictuelle et ne concluent pas (à raison).

Les autres psychologues (35% : 7/20) concluent à l'inefficacité en se fondant exclusivement sur le test non significatif (comme dans la situation 2). La proportion est semblable chez les statisticiens (mais cette conclusion ne peut être généralisée : [$|\varphi_p$-$\varphi_s| < 0.23$]). Enfin, la réponse "Efficace", absente chez les psychologues, est donnée par 3 (12%) statisticiens.

On notera, chez les psychologues, la dissymétrie entre les situations perçues comme conflictuelles (la 1 et la 4) : alors que dans la situation 1, les deux réponses "Efficace" et "Non efficace" apparaissent, seule "Non efficace" est donnée dans la situation 4. Cela illustre bien l'influence du test de signification : c'est le résultat

significatif qui “autorise” la conclusion d'efficacité dans la situation 1, comme nous l’avons souligné précédemment.

Les résultats des psychologues confirment ceux obtenus par M.-P. Lecoutre (1983, 1991) : ce sont les mêmes situations qui sont jugées favorables (ici la 2 et la 3) ou conflictuelles (la 1 et la 4), et la place accordée au test dans les critères du jugement est prépondérante.

Question 2 “Quelle prédiction feriez-vous si on allait jusqu'au bout de l'expérimentation ?”

C'est cette question qui a posé le plus de problèmes aux sujets, quelle que soit leur formation. Chez les psychologues, elle a donné lieu à des commentaires du genre “On n'est pas habitué à ce type de questions”, sans que ces commentaires soient négatifs car ils sont souvent associés à “Pourtant c'est une question très intéressante”. Les statisticiens de l'industrie pharmaceutique, bien que plus familiers de ce type de questions, ne paraissent pas mieux armés pour y faire face. En conséquence, le nombre de non-réponses (“Je ne sais pas”, “Je ne peux rien prédire dans ce cas”, …) est très supérieur à celui observé dans les deux autres questions, surtout dans les situations 2 et 4 où le test est non significatif.

On remarquera tout d'abord que les psychologues ont eu beaucoup plus de mal que les statisticiens à donner une prédiction sur la valeur de la statistique de test (t) et qu’ils ont spontanément plutôt fait porter leur réponse sur le seuil p, pour lequel le rôle de l’effectif leur paraissait plus clair. En fait, les résultats peuvent se résumer assez simplement. Quels que soient la situation et l'objet de la prédiction (d, le test, la conclusion) considérés, on obtient sur l'ensemble des sujets essentiellement deux réponses : une non-réponse et la réponse “globalement la même chose” qui est la réponse majoritaire en général et qui renvoie au “biais de représentativité” de Tversky et Kahneman (1971).

Il n'est pas surprenant de constater, au vu des résultats à la première question, que c'est la situation 3 qui donne lieu, pour les deux groupes, au taux de non-réponse le plus faible, et à la proportion de réponses “La même chose” la plus élevée.

Il faut tout de même noter, pour la situation 4, un nombre non négligeable (respectivement 30%: 6/20 et 28%: 7/25) de réponses indiquant que le seuil p associé au test va diminuer et va peut-être “atteindre la significativité”. Ceci est (correctement) justifié par les sujets en précisant que “toutes choses égales par ailleurs, l'augmentation du nombre de degrés de liberté entraîne un accroissement de la significativité”. En conséquence, on trouve 4 et 5 sujets (parmi les 6 et 7 en question) pour qui la conclusion passerait de l'incertitude à la reconnaissance d'un effet intéressant.

Le croisement des trois critères (d, test, conclusion), pour chacune des situations, ne fait que confirmer ce que l'on pouvait raisonnablement supposer, c'est-à-dire que les sujets donnent généralement une même réponse (toujours “La même chose” ou toujours “Je ne peux rien prédire”). Ceci est un peu moins vrai chez les statisticiens qui passent plus volontiers d'une réponse à l'autre.

Question 3 “Pour quelle(s) situation(s) prendriez-vous la décision d'arrêter l'expérimentation ?”

La décision d'arrêt est majoritaire dans les trois premières situations, mais on observe tout de même des variations selon ces situations. Dans les deux premières situations, la décision d'arrêter n'est que légèrement majoritaire et les deux groupes de sujets sont relativement proches. On ne peut cependant en inférer une différence négligeable : [$|\varphi_p-\varphi_s| < 0.26$] pour la situation 1 et [$|\varphi_p-\varphi_s| < 0.24$] pour la situation 2.

Dans la situation 3 les réponses sont plus tranchées, avec une forte majorité de réponses “Arrêter” dans chacun des deux groupes (respectivement 75%: 15/20 et 88%: 22/25). Ceci est cohérent avec le fait que cette situation apparaît aux sujets comme la plus simple. Il faut noter que certains psychologues ont considéré cette situation comme particulière et y ont répondu en adoptant un point de vue “scientifique” plutôt qu’un point de vue “économique” (comme le spécifiait la consigne). Typiquement cela est exprimé par : “Dans cette situation je continue, car on a mis le doigt sur quelque chose d'intéressant. Il y a quelque chose à creuser.”

Dans la situation 4, les deux groupes se distinguent beaucoup plus : alors que les statisticiens sont presque unanimes (96% : 24/25) à prendre la décision de continuer, les psychologues ne sont que 60% (12/20) à prendre cette décision [$\varphi_s-\varphi_p > 0.21$]. En général les sujets justifient la réponse “Continuer” par l'espoir que la différence observée, importante, devienne significative.

Question 1 × Question 3

Il est intéressant de croiser les questions 1 et 3, cette dernière pouvant être vue comme un indicateur du degré de certitude de la réponse à la première. Le Tableau 21 et le Tableau 22, ci-dessous, présentent les effectifs correspondants.

Psychologues :	*Question 3*											
	Situation 1			*Situation 2*			*Situation 3*			*Situation 4*		
Question 1	Arrêt	Cont.	?	Arrêt	Cont.	?	Arrêt	Cont	?	Arrêt	Cont.	?
Efficace	5	4	0	0	0	0	15	5	0	0	0	0
Non efficace	5	3	0	9	7	1	0	0	0	4	2	1
?	2	1	0	2	1	0	0	0	0	2	10	1

Tableau 21

Croisement des réponses des Psychologues (en effectif) aux questions 1 et 3 selon la situation expérimentale

Statisticiens :	*Question 3*											
	Situation 1			*Situation 2*			*Situation 3*			*Situation 4*		
Question 1	Arrêt	Cont.	?	Arrêt	Cont.	?	Arrêt	Cont	?	Arrêt	Cont.	?
Efficace	1	2	0	0	0	0	21	3	0	0	3	0
Non efficace	10	10	0	12	8	1	0	0	0	1	8	0
?	2	0	0	1	3	0	1	0	0	0	13	0

Tableau 22

Croisement des réponses des Statisticiens (en effectif) aux questions 1 et 3 selon la situation expérimentale

Pour ce qui concerne la conclusion d'efficacité, la situation 3 est celle qui est associée à la plus grande certitude des sujets puisque presque tous estiment que les premiers résultats sont suffisants pour conclure. Parmi les psychologues qui concluent à l'efficacité pour la situation 1, la moitié d'entre eux (4/9) sont enclins à poursuivre l'expérience, ce qui indique une certaine incertitude dans leur conclusion; ceci n'est pas étonnant, puisque cette situation est jugée conflictuelle.

Pour ce qui concerne la conclusion d'inefficacité, les deux groupes sont partagés quant à l'arrêt ou non de l'expérience, dans la situation 1. Il est plus surprenant de constater que ce partage vaut encore pour la situation 2, pourtant perçue comme favorable (différence observée et test allant dans le même sens). À l'examen des réponses, on constate en fait qu'une bonne partie des psychologues concernés (7/17) ne sont pas sûrs de leur conclusion, et que ce doute n'apparaît qu'à la question 3. Il s'exprime alors par des phrases du type "Ça peut changer; il faut voir si la tendance est confirmée ou infirmée", "Ça peut devenir significatif".

6.3.7. Conclusion

On constate finalement que psychologues et statisticiens se comportent globalement de façon assez similaire, et que ces derniers, malgré leur formation, ne sont pas à l'abri des abus d'interprétation des tests, et tout particulièrement de celui consistant à voir dans un résultat non significatif la démonstration d'une absence d'effet.

Il est cependant possible de distinguer les deux groupes sur le point suivant. Ce sont les psychologues qui tiennent le moins compte du critère d'intérêt de l'effet et qui privilégient le résultat du test, confondant significativité statistique et significativité substantielle. Cette confusion, qui est le fait de près de la moitié des psychologues, confirme le rôle important que ceux-ci accordent aux tests dans les critères de jugement : un résultat significatif peut être si prégnant que toute autre considération, notamment celle du critère des experts, est évacuée. Ceci apparaît caractéristique des psychologues[23], dans la mesure où une affectation des sujets aux

[23] On pourrait objecter que dans le cas présent l'avis des médecins experts pourrait paraître aux psychologues un critère extérieur, peu pertinent (ce qui ne serait pas le cas pour les statisticiens de l'industrie pharmaceutique). Mais les résultats des psychologues ne font que confirmer ceux obtenus précédemment par M.-P. Lecoutre (1983, 1991) dans des situations "plus naturelles" pour les psychologues.

groupes (psychologues/statisticiens) sur la seule base des réponses à l'un ou l'autre des deux cas “question 1 × situation 1” et “question 3 × situation 4” permet de classer correctement 32 sujets sur les 45, soit 71%. Le classement n'est amélioré que d'un sujet lorsqu'on l'effectue sur la base du croisement des deux cas, ce qui montre leur redondance[24].

Serait-ce que les psychologues sont trop respectueux des statistiques dont ils ne maîtrisent pas suffisamment les principes, ou qu'ils en attendent trop ?

6.4. L'ÉTUDE DE ROSENTHAL ET GAITO (1963)

Nous allons particulièrement détailler cette étude car elle sert de base à notre expérience 2 présentée dans la section 6.5.

Elle a été la première étude portant sur la manière dont les chercheurs en psychologie interprètent les seuils de signification associés à un test. Les auteurs ont interrogé 19 personnes, réparties en deux groupes : 10 psychologues universitaires, tous docteurs, et 9 étudiants diplômés (*graduate*). Face à une série de 14 seuils de signification (valeurs *p*, de 0.001 à 0.90), supposés associés à des résultats expérimentaux, les sujets étaient invités à exprimer leur “degré de conviction dans les résultats expérimentaux comme une fonction des valeurs *p* correspondantes” (*Expressions of degree of belief in research findings as a function of associated p levels*), au moyen d'une échelle en six points (de 0 pour l'absence totale de confiance à 5 pour l'extrême confiance). Pour chaque valeur *p* les sujets devaient donner deux réponses, selon qu'on supposait les résultats expérimentaux issus d'un échantillon de taille 10 ou 100.

Cette expérience concerne un aspect fondamental du jugement des chercheurs. En effet, le chercheur est constamment confronté à des seuils *p* dans sa pratique : d'une part, dans les expériences qu'il réalise où il est amené à en calculer, ne serait-ce que pour satisfaire les normes de publication; d'autre part, dans son examen de la littérature scientifique où il ne dispose bien souvent que des seuls seuils de signification pour juger des résultats. Le jugement sur le seuil *p* intervient donc comme une étape nécessaire, un préalable à l'interprétation des résultats; ceci ne signifiant d'ailleurs pas que ce jugement ne résulte pas d'une activité complexe. Dans cette étude, le sujet est placé dans une situation relativement abstraite, générique, sensée représenter le noyau commun à tout test statistique. À la différence de l'expérience 1 “psychologues et statisticiens”, on met à l'arrière-plan l'aspect conclusion, décision de l'analyse statistique pour privilégier davantage l'aspect jugement. On peut également estimer que le sujet est ici dégagé des contraintes et des normes inhérentes à la situation de publication.

Résultats

Les auteurs présentent pour résultat quatre courbes (2 groupes × 2 conditions d'effectif) sur un même graphique, portant la confiance moyenne en ordonnée et le seuil *p* en abscisse (*cf.* la Figure 3). Ces courbes sont semblables : la confiance est une fonction monotone décroissante de *p*, non linéaire de type exponentiel. (Pour $p = 0.70$ et $p = 0.90$, la confiance est nulle pour tous les sujets, dans les deux conditions de taille d'échantillon. Aussi ces valeurs sont retirées des analyses par les auteurs.) Dans chacun des deux groupes, et pour toute valeur de *p*, l'effectif $N = 100$ donne lieu à plus de confiance que l'effectif $N = 10$. Par ailleurs les chercheurs sont toujours moins confiants que les étudiants, la courbe la plus élevée des chercheurs ($N = 100$) étant constamment inférieure ou quasi égale à la courbe la plus basse des étudiants ($N = 10$).

[24] Étant donné la modalité de réponse d'un sujet, on affecte celui-ci au groupe pour lequel cette même modalité a la fréquence (absolue) la plus élevée. Par exemple, pour la question 1, situation 1, les sujets répondant “Efficace” ou “?” sont classés comme psychologues, les autres comme statisticiens. Il est même possible d'aller beaucoup plus avant dans la distinction des groupes. Dans un premier temps, nous avons numérisé les résultats en procédant à une analyse des correspondances sur les données recodées sous forme disjonctive complète. Puis nous avons réalisé une analyse discriminante sur les facteurs obtenus. La discrimination est (quasi) parfaite puisque tous les sujets sauf un sont classés correctement. Cependant, ce résultat ne doit pas faire illusion car il est dû, en bonne partie, au simple fait que les sujets sont peu nombreux par rapport aux questions×situations et que presque tous les sujets ont des patrons de réponses différents.

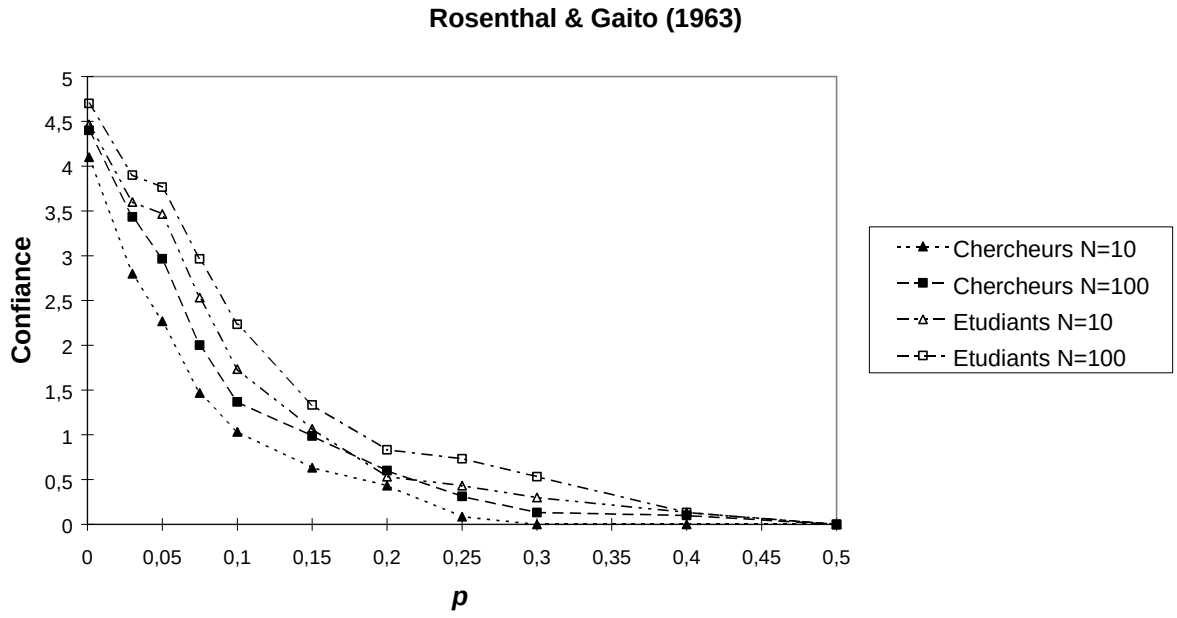


Figure 3

Confiance moyenne exprimée par des chercheurs et des étudiants en psychologie en fonction du seuil observé et de la taille de l'échantillon (N=10 ou N=100). D'après Rosenthal et Gaito (1963)

Cette expérience a été répliquée, apparemment dans des conditions identiques, par Beauchamp et May (1964). Pour l'essentiel ils rapportent les mêmes résultats, recueillis dans une autre université auprès de deux groupes d'une dizaine de sujets chacun. Le seul point de désaccord concerne l'existence ou non de "l'effet chute" dont il est question ci-dessous.

Rosenthal et Gaito font porter leur analyse essentiellement sur deux points.

- L'existence d'un effet de chute brutale de confiance juste après la valeur $p = 0.05$, effet attendu, selon eux, et explicable par le rôle particulier de frontière entre significatif et non significatif joué par cette valeur (en revanche un tel effet, bien qu'également attendu par les auteurs, n'apparaît pas pour 0.01).
- La plus grande confiance attribuée au cas $N = 100$, qu'ils interprètent en considérant que les sujets font appel, non seulement à la notion d'erreur de première espèce (puisque la confiance décroît avec p), mais aussi, et au moins intuitivement, à la notion d'erreur de deuxième espèce (qu'on sait diminuer avec l'effectif, toutes choses égales par ailleurs).

Remarques

Cette étude appelle un certain nombre de remarques.

- En préliminaire, on notera que les auteurs commettent des erreurs d'interprétation quant aux tests; par exemple quand ils énoncent qu'en passant de $p = 0.01$ à $p = 0.03$ on augmente en moyenne de 2 erreurs de type I pour 100 répliques (ce n'est le cas que si H_0 est vraie). C'est encore plus flagrant dans leur réponse à Beauchamp et May (Rosenthal et Gaito, 1964), où, explicitement, ils considèrent que la probabilité que l'effet "chute à 0.05" soit réel est donnée par 1 - p (p étant la valeur obtenue pour le test qu'ils ont appliqué).

- La consigne est des plus ambiguës. Que faut-il entendre ici par "conviction (ou confiance) dans les résultats expérimentaux" ? On peut écarter l'hypothèse que c'est l'honnêteté du chercheur qu'il s'agit de juger. La notion de confiance pourrait aussi renvoyer, pour les sujets, à la qualité du recueil des données, des procédures de mesure, *etc.*; mais en ce cas on ne voit pas bien pourquoi cette confiance serait dépendante de p et de N.

Ce qui nous apparaît comme le plus plausible est : "confiance en la réalité de l'effet", autrement dit "confiance en la vérité de l'hypothèse alternative".

Rien n'est précisé non plus quant au test d'où provient la valeur p; mais là il y a tout lieu de supposer que les auteurs se réfèrent à la situation la plus courante du test usuel d'une hypothèse nulle d'absence d'effet (au moyen d'un t de Student ou d'un F, par exemple) et que c'est bien ainsi que le comprennent les sujets.

Malgré l'ambiguïté la consigne n'a apparemment pas posé de problème aux sujets; tout au moins Rosenthal et Gaito n'en font pas mention, non plus que Beauchamp et May lors de leur réplique.

- L'effet “chute à 0.01 ou 0.05” (*cliff effect*)[25] serait, comme le considèrent les auteurs, très vraisemblable, étant donné que ces valeurs sont souvent prises comme des limites conventionnelles, du moins dans les publications. Mais en réalité, dans les courbes présentées, cet effet ne se manifeste que chez les étudiants (pour $p = 0.05$ uniquement, et dans les deux cas $N = 10$ et $N = 100$). Et d'ailleurs, quand Rosenthal et Gaito (1964) affirment trouver cet effet dans les données de Beauchamp et May (que ces derniers n'ont pas retrouvé, du moins de manière “significative”), ce n'est encore que dans le groupe des étudiants. En conséquence on serait plutôt tenté de conclure, et ce serait déjà un résultat intéressant, que la pratique, dont on supposera pourvus les chercheurs, a éliminé cet effet; encore faudrait-il admettre qu'il existait lorsque ces mêmes chercheurs étaient étudiants.

Oakes (1986) a tiré argument de ces résultats pour conclure, à l’appui de son hypothèse de “significativité”, que la “chute” de confiance observée indiquait que des résultats expérimentaux hypothétiques prennent soudainement statut de réalité dès lors qu’ils sont associés à un seuil observé juste inférieur à la barre des 0.05. Le problème est que, comme nous venons de le voir, dans les données fournies cette “chute” n’est que faible chez les étudiants et n’apparaît pas chez les chercheurs.

- Que la courbe de confiance soit décroissante en fonction de p est un résultat attendu. Mais pour ce qui est de la forme précise de la fonction, les auteurs se bornent à en noter le caractère exponentiel et ne la commentent pas outre mesure; or cela ne va pas de soi. Ainsi, étant donné la pratique, erronée dans le cadre fréquentiste traditionnel mais tenace et fréquente (c'est d'ailleurs celle des auteurs, voir la première remarque ci-dessus), de considérer $1 - p$ (ou $1 - \alpha$) comme la probabilité que l'hypothèse alternative soit vraie, et ce donc indépendamment de N, on pouvait plutôt s'attendre à obtenir la même droite, $confiance = 1 - p$, pour les deux valeurs de N. Que ce ne soit pas le cas reste donc à expliquer.

- L'explication du fait que les courbes $N = 100$ sont supérieures, en confiance, à celles $N = 10$, n'est pas aussi évidente que les auteurs le laissent supposer. En fait il ne s'agit même pas d'une réelle explication car la notion d'erreur de deuxième espèce est invoquée sans aucune justification, comme si c'était la seule possibilité. Cette interprétation a d'ailleurs été critiquée par Bakan (1966, p. 430) et Oakes (1986, pp. 26-29). Ainsi pour Bakan c'est le contraire de ce qu'affirment Rosenthal et Gaito qui est vrai, d'un point de vue normatif : le rejet de l'hypothèse nulle est davantage en faveur d'un effet important dans le cas d'un petit échantillon car il faut alors que l'effet soit important pour être mis en valeur par un test moins puissant.

Royall (1986) s'est intéressé au même type de problèmes, en étudiant le rapport de vraisemblance des hypothèses et en se limitant au test de deux hypothèses simples (ponctuelles). Il conclut que les deux énoncés sont valables, tout dépendant de la façon précise de poser le problème : si le seuil observé ne sert que par comparaison à un seuil prédéfini (selon l'approche de Neyman et Pearson), on conclut à une plus grande évidence contre l'hypothèse nulle dans le cas du plus grand échantillon, alors qu'on conclut à l'inverse quand on prend en compte la valeur précise du seuil observé (selon l'approche de Fisher[26]). Seulement son étude ne vaut que pour le premier cas (Neyman et Pearson); en effet, pour ce qui concerne le second, il ne démontre pas son résultat mais ne fait que l'illustrer au moyen d'un exemple et il est facile de trouver des contre-exemples allant en sens opposé, ce qui signifie que les deux résultats sont possibles, tout dépendant des valeurs de N et de l'écart entre les deux hypothèses. Il est donc impossible de faire une prévision spécifique.

---

[25] Remarquons qu'à la place d'un “effet chute”, caractérisé par la *chute* du degré de confiance après la valeur $p = 0.05$, il est tout aussi possible de parler “d'effet plateau”, ce que nous ferons par la suite. Dans ce cas, ce qui est caractéristique c'est un niveau *plus élevé* de confiance pour la valeur 0.05 (donc un plateau entre 0.03 et 0.05) que ce que donnerait la simple interpolation de la courbe, la chute pour les valeurs suivantes n'en étant qu'une conséquence (on “rattrape” la courbe). Tout dépend en fait de la courbe que l'on prend pour référence (figurée en pointillé dans le dessin suivant).

[26] Alors même qu'il décrit succinctement les deux théories des tests et qu'il fait référence à Neyman et Pearson, il est étonnant que Royall ne cite pas Fisher dans cet article. Peut-être est-ce justement parce qu'il ne traite pas correctement ce cas.

L'interprétation de la consigne et les cadres théoriques

Examinons plus précisément comment on peut prédire la forme et les positions relatives des courbes, selon l'interprétation qu'on donne à la consigne et le cadre théorique de référence qu'on adopte.

Pour cela, comme nous l'avons dit plus haut, nous considérons que la situation de base est celle du test usuel d'une hypothèse nulle ponctuelle d'absence d'effet ($H_0 : \delta = 0$), mais la consigne "conviction dans les résultats expérimentaux" peut donner lieu aux quatre interprétations suivantes :

I1: "*Confiance en la vérité de l'hypothèse alternative (H_1)*" ou, de façon (formellement) équivalente, "*Confiance en la fausseté de l'hypothèse nulle (H_0)*". C'est une interprétation qui nous semble très plausible, au vu des résultats décrits dans la section 6.1.

I2: "*Confiance en la décision prise (en faveur de H_1 pour un résultat significatif, de H_0 sinon)*". Cette interprétation nous apparaît comme la moins probable, mais elle ne peut être totalement exclue. Il est possible que des sujets qui auraient l'habitude d'une pratique décisionnelle du test (accepter H_0 aussi bien que H_1) expriment, comme dans le cas précédent, leur confiance en faveur de H_1 pour les faibles valeurs du seuil p (c'est-à-dire celles qui ne dépassent pas le risque de première espèce qu'ils se donnent), mais aussi leur confiance en faveur de H_0 pour les autres valeurs du seuil.

I3 "*Confiance en ce que l'effet vrai soit important*". Cette éventualité est à considérer en raison de la confusion parfois faite par les chercheurs entre significativité statistique et significativité substantielle (*cf.* la sous-section 2.2.3.).

I4 "*Confiance en ce que l'effet vrai soit de même sens que celui observé*". Oakes tient cette interprétation pour la plus probable (1986, p. 27).

Nous allons considérer les théories statistiques suivantes :
- L'approche de Fisher.
- L'approche de Neyman et Pearson.
- L'approche bayésienne standard, ou fiducio-bayésienne. Dans cette approche on utilise une distribution *a priori* vague, non-informative; se limitant ainsi à "ce que les données ont à dire", sans tenir compte d'informations extérieures.

Nous ne retenons pas comme pertinent le test des rapports de vraisemblance car cela supposerait que l'on soit confronté au test de deux hypothèses ponctuelles ou que l'on dispose d'informations sur la classe des hypothèses alternatives, ce qui n'est pas le cas. Nous ne considérons pas plus une approche bayésienne générale, où la distribution *a priori* choisie pourrait être quelconque, car ce choix, pour ne pas être totalement arbitraire, requiert un minimum d'informations qui n'est précisément pas fourni ici[27].

Posons maintenant que la confiance est bien exprimée par une probabilité sur l'hypothèse, conditionnelle aux données observées puisque les sujets sont ici supposés confrontés à des résultats effectivement obtenus.

Tous les croisements entre les questions et les cadres théoriques énumérés ci-dessus ne sont pas pertinents :

- I1 peut être traitée dans les trois théories (bien que triviale dans le cadre fiducio-bayésien comme nous le verrons ci-dessous).
- I2 correspond à la théorie de Neyman et Pearson uniquement, puisque cette interprétation suppose une approche décisionnelle.

[27] Si l'on veut tout de même illustrer cette approche, on peut, par exemple, retenir le modèle que Lindley (1957) a choisi pour l'énoncé d'un paradoxe bien connu chez les statisticiens. La distribution d'échantillonnage est normale, de variance connue; la probabilité *a priori* de H_0 est fixée à une valeur non nulle, le reste de la probabilité étant réparti uniformément sur un large intervalle. On trouve alors, pour les trois énoncés I1, I3, I4, que la probabilité *a posteriori* de H_1 est de forme exponentielle, la courbe $N = 10$ étant toujours supérieure à la courbe $N = 100$ (la forme exacte des courbes dépendant de la variance supposée connue et de la probabilité *a priori* de H_0).

- I3 et I4 renvoient à l'approche bayésienne uniquement. (I3 pourrait être traitée dans le cadre des théories traditionnelles en posant pour hypothèse nulle une valeur particulière qui serait jugée importante, mais cela n'est pas pertinent ici puisqu'aucune information spécifique n'est fournie aux sujets.)

Les prédictions associées aux différentes théories statistiques

Voyons ce que l'on peut attendre en fonction de ces divers cadres théoriques et interprétations retenus.

(1) Prédictions associées à la théorie de Fisher.

Du point de vue de Fisher, il n'y a pas d'hypothèse alternative particulière à considérer autre que la négation de l'hypothèse nulle et la seule interprétation pertinente en ce cas est I1.

Selon Fisher, le seuil p observé, à lui seul, est indicateur du degré "d'évidence" à l'encontre de l'hypothèse nulle, et par conséquent du degré "d'évidence" en faveur de l'hypothèse alternative. Les courbes doivent donc être identiques quel que soit N et être une simple fonction décroissante de p, sans plus de précision quant à la forme, ce qui est trop peu contraignant pour être vraiment intéressant.

(2) Prédictions associées à la théorie de Neyman et Pearson.

C'est à cette théorie que se réfèrent implicitement Rosenthal et Gaito puisqu'ils ont recours à la notion de risque de deuxième espèce.

Oakes (1986, p. 26) a contesté la pertinence du concept de risque de type I (α), et donc de la théorie de Neyman et Pearson, dans le cas de l'expérience de Rosenthal et Gaito, en arguant que les sujets sont confrontés à des valeurs p, donc calculées *a posteriori*. Cette critique est sans fondement car, s'il est indéniable que p est calculé à partir des résultats, cela ne préjuge en rien de la manière dont fonctionnent les sujets quant au test de signification et à la tâche qu'on leur propose. Ils peuvent très bien avoir en tête un seuil fixé (α) par rapport auquel ils jugent les valeurs p pour considérer le résultat significatif ou non.

Soit H_0 l'hypothèse nulle et H_A son complémentaire c'est-à-dire l'ensemble des H_i (i=1,…), les hypothèses alternatives possibles (nous nous plaçons dans le cas d'un ensemble discret d'hypothèses, mais les résultats sont identiques dans le cas continu). Soit α le risque de première espèce, fixé à l'avance, β_i les risques de deuxième espèce associés aux hypothèses H_i, P_0 et P_i les probabilités *a priori* des hypothèses H_0 et H_i. Notons S l'événement "le résultat est significatif" (c'est-à-dire $p \leq \alpha$), et $\neg S$ l'événement contraire.

Les probabilités *a posteriori* des hypothèses, conditionnellement au résultat obtenu (S ou $\neg S$), deviennent, par simple application du théorème de Bayes :

[1] $Pr(H_0|S) = Pr(H_0 \cap S) / Pr(S)$
$= Pr(S|H_0)Pr(H_0) / [Pr(S|H_0)Pr(H_0) + \Sigma\{Pr(S|H_i)Pr(H_i)\}]$
$= \alpha P_0 / [\alpha P_0 + \Sigma\{(1-\beta_i)P_i\}]$ (La sommation s'entendant pour toutes les hypothèses alternatives.)

[2] $Pr(H_A|S) = \Sigma Pr(H_i|S) = 1 - Pr(H_0|S)$
de même :

[3] $Pr(H_0|\neg S) = (1-\alpha)P_0 / [(1-\alpha)P_0 + \Sigma\beta_i P_i]$

[4] $Pr(H_A|\neg S) = \Sigma Pr(H_i|\neg S) = 1 - Pr(H_0|\neg S)$

Examinons d'abord l'effet de N.

Les probabilités *a priori* des hypothèses, P_0 et les P_i, sont bien sûr inconnues; mais elles sont constantes, de même que α. Par ailleurs, on sait que quand l'effectif augmente, β_i diminue (quelle que soit H_i). Il en résulte que :

[1] $Pr(H_0|S)$ diminue,
[2] $Pr(H_A|S)$ augmente,
[3] $Pr(H_0|\neg S)$ augmente,
[4] $Pr(H_A|\neg S)$ diminue.

Pour I1.

Ce sont les cas [2] et [4] qui sont pertinents. *Quand le résultat est jugé significatif*, la probabilité que H_A soit alors vraie est plus grande dans le cas du plus grand échantillon. Mais c'est l'inverse *quand le résultat*

est jugé non significatif. On devrait donc observer une inversion de l'ordre des courbes de part et d'autre de l'abscisse correspondant à α.

Pour I2.

Ce sont les cas [2] et [3] qui sont pertinents. Alors la probabilité que H_A soit vraie, compte tenu d'un résultat jugé significatif, est toujours plus grande dans le cas du plus grand échantillon. De même pour H_0, dans le cas d'un résultat jugé non significatif. La courbe pour $N = 100$ devrait toujours être supérieure à celle pour $N = 10$.

En ce qui concerne la forme des courbes, quelle que soit l'interprétation de la consigne, on devrait observer que chaque courbe (pour N fixé) est constituée de deux plateaux, l'un correspondant à la zone où $p \leq \alpha$, l'autre à la zone où $p > \alpha$, puisque les expressions [1] à [4] ne dépendent de p qu'à travers S et donc α.

Enfin il faut noter que tous ces résultats précisent la forme des courbes et leurs positions relatives, mais qu'on ne peut pas davantage caractériser les courbes dans la mesure où P_0 et les β_i et P_i demeurent inconnus.

(3) Prédictions associées à la théorie fiducio-bayésienne.

Pour I1.

Puisque nous considérons une distribution *a priori* non informative, continue, la probabilité *a priori* de l'hypothèse nulle ponctuelle est zéro, et donc aussi sa probabilité *a posteriori*. La probabilité *a posteriori* de l'hypothèse alternative est par conséquent un, quels que soient p et N.

Pour I3.

On trouve que pour un même p, la probabilité que l'effet vrai soit important est plus grande dans le cas du plus petit échantillon. À N constant, les courbes en fonction de p sont décroissantes, non linéaires, de forme à peu près exponentielle.

Prenons, par exemple, et sans perte de généralité, le cas d'une comparaison à un degré de liberté, pour deux groupes appariés de taille N. S'il est naturel, comme nous l'avons dit, de considérer que le test évoqué par la consigne est bilatéral relativement à la question de la seule existence de l'effet (I1), ici, dans le cas d'un degré de liberté, il paraît tout aussi naturel de considérer que la question d'un effet important renvoie à un effet de signe déterminé.

Soit d et s l'effet et l'écart-type observés, δ_0 la limite qu'on se fixe d'un effet important. L'effet vrai δ est distribué selon un t généralisé (voir, par exemple, B. Lecoutre, 1984a).

$\delta \in d, s \sim t_q(d, s^2/N)$ $\quad (q = N - 1)$ soit $t = (\delta - d)\sqrt{N}/s \sim t_q$

Soit γ la probabilité que l'effet vrai soit important :

$Pr(\delta > \delta_0 \mid d, s) = \gamma$

$Pr(t > (\delta_0 - d)\sqrt{N}/s) = \gamma$ (où t est distribué suivant le t de Student usuel)

$Pr(t > \delta_0\sqrt{N}/s - d\sqrt{N}/s) = \gamma$

or $d\sqrt{N}/s$ est la statistique de test observée, et est donc égale à la valeur t_q^{1-p} de la fonction de répartition de la distribution du t de Student à q degrés de liberté correspondant à une probabilité $1 - p$. Donc :

$Pr(t > \delta_0\sqrt{N}/s - t_q^{1-p}) = \gamma$

p étant fixé, quand N augmente, $\delta_0\sqrt{N}/s$ augmente (dans la mesure où δ_0 est fixé et où l'on considère s comme constant) et t_q^{1-p} diminue (puisque q augmente); le membre de droite de l'inéquation augmente donc. Il en résulte que la probabilité γ diminue. (On notera que si t_q^{1-p} diminue, c'est que d diminue : conserver un même p avec un N plus grand est plutôt indicateur d'un effet plus faible.)

Pour I4.

Il suffit de poser $\delta_0 = 0$ dans le cas précédent pour trouver le résultat. En fait, on obtient simplement pour probabilité *a posteriori* de l'effet vrai la réinterprétation fiducio-bayésienne du résultat du test

unilatéral (*cf.*, par exemple, Pratt, 1965; Lecoutre, 1984a, 1984b; Casella et Berger, 1987). Soit, par exemple pour un effet observé positif :

$Pr(\delta>0|\text{données}) = 1 - p$ ou $1 - p/2$, selon que le test est unilatéral ou bilatéral.

(Le résultat est bien sûr semblable dans le cas d'un effet négatif.)

Ce résultat est indépendant de l'effectif, par conséquent les courbes $N = 10$ et $N = 100$ devraient être des droites confondues.

• Notons enfin que si l'on considère que l'hypothèse nulle est pratiquement toujours fausse, ainsi qu'il a été dit dans la section consacrée aux critiques des tests (*cf.* 2.1.11.), alors, pour I1, on pourra avoir une confiance maximum en l'hypothèse alternative, quels que soient p et N. On retrouve donc le même résultat que dans le cadre fiducio-bayésien.

Ces résultats sont résumés dans le tableau suivant ("10=100" signifie symboliquement "les courbes $N = 10$ et $N = 100$ sont superposées").

	I1	I2	I3	I4
Fisher	$Pr = f(1-p)$ 10=100			
Neyman-Pearson	$p \le \alpha$: Pr=cste, 10<100 $p > \alpha$: Pr=cste, 10>100	$p \le \alpha$: Pr=cste, 10<100 $p > \alpha$: Pr=cste, 10<100		
Fiducio-bayésien	$Pr = 1$ 10=100		$Pr \cong \exp(ap+b)$ 10>100	$Pr = 1-p$ 10=100
Bayésien (Lindley)	$Pr \cong \exp(ap+b)$ 10>100		$Pr \cong \exp(ap+b)$ 10>100	$Pr \cong \exp(ap+b)$ 10>100

Tableau 23

Type de courbes attendu selon la théorie statistique et l'interprétation de la consigne considérées

Selon le cadre de référence et la formulation choisis, différents résultats sont donc justifiables. Si l'on compare les résultats observés dans les expériences précédentes à ces prédictions, c'est le cadre fiducio-bayésien, associé à l'interprétation d'effet intéressant (I3), qui justifie des courbes non linéaires plus ou moins compatibles avec les courbes moyennes obtenues (le cadre fishérien également, puisqu'il est compatible avec toute fonction monotone). On serait donc tenté de penser que ce cadre est approprié pour la description des jugements naturels des sujets, ce qui recouperait les résultats de M.-P. Lecoutre (1991). Seulement, outre que l'interprétation I3 ne nous semble pas la plus vraisemblable, la prédiction, quant à la position relative des deux courbes, de la supériorité du cas $N = 10$ est contredite par les observations. Sur ce dernier point, seule la théorie de Neyman et Pearson, associée à l'interprétation I2 (peu plausible), prédit l'ordre observé.

Les courbes moyennes observées ne sont donc pas compatibles avec l'hypothèse que les sujets se conformeraient à une seule et même théorie statistique.

La fonction puissance: essai d'un modèle psychophysique

Admettons, comme précédemment, que les sujets ont bien compris la confiance comme celle vis-à-vis de l'hypothèse et regardons cette confiance comme fonction de la probabilité $1 - p$. La courbe est alors une fonction croissante monotone et non linéaire. La fonction exponentielle n'est pas celle qui s'ajuste le mieux aux points observés, la fonction *puissance* est une meilleure approximation, soit : $confiance = a(1 - p)^b$. En estimant les données à partir de la figure présentée par Rosenthal et Gaito on obtient, pour chacun des quatre groupes, les r^2 suivants (qui donnent la proportion de variance "expliquée" par la courbe théorique[28]) :

1) Chercheurs, $N = 10$: $r^2 = 0.996$ ($a = 4.132$, $b = 12.421$)

2) Chercheurs, $N = 100$: $r^2 = 0.992$ ($a = 4.518$, $b = 9.743$)

[28] En réalité il s'agit ici, cas de la régression non linéaire, d'un pseudo r^2, et donc d'une pseudo part de variance expliquée. Il est donné par : $r^2 = 1 - Vr / Vt$ où Vr est la moyenne quadratique des erreurs, $Vr = \Sigma(\text{valeur théorique} - \text{valeur observée})^2 / N$, et Vt est la variance totale de la variable dépendante. Ce pseudo r^2 sera négatif si l'ajustement est plus mauvais que celui donné par la droite moyenne.

3) Étudiants, $N = 10$: $r^2 = 0.986$ ($a = 4.685$, $b = 8.426$)
4) Étudiants, $N = 100$: $r^2 = 0.988$ ($a = 4.893$, $b = 7.041$)

L'ajustement est assez remarquable. Il peut même être légèrement amélioré dans le cas des étudiants en ôtant le point à $p = 0.05$ pour éliminer "l'effet chute". La qualité de cet ajustement peut même paraître surprenante, compte tenu du fait que les points observés sont calculés sur peu de sujets (9 et 10, respectivement pour les "chercheurs" et les "étudiants") et à partir d'une échelle assez peu sensible (en six points). Toutefois, il ne s'agit ici que des courbes moyennes. Une indication de la variabilité interindividuelle est fournie par Rosenthal et Gaito qui précisent que les écart-types décroissent également avec p; ils vont de 0.99 pour $p = 0.001$ à 0.24 pour $p = 0.50$. Compte tenu de l'échelle utilisée, cela semble traduire une relative homogénéité des sujets. Il est regrettable à ce propos que Beauchamp et May ne présentent pas les courbes obtenues lors de leur réplique, et ne disent rien de leur forme.

La fonction puissance fait penser à la loi psychophysique de Stevens (1962). Il serait intéressant de retrouver ici, dans un domaine qui implique une activité de jugement élaborée de la part du sujet, une loi portant sur des "sensations". Pour aller dans ce sens on peut arguer de toutes les idées fausses qui règnent à propos du test de signification et du fait que la logique de celui-ci est loin d'être naturelle, pour envisager que les sujets, en général, font appel en ce domaine à leur intuition au moins autant, sinon plus, qu'à un jugement élaboré. Plus spécifiquement encore, on pourrait penser que la perception des seuils p par les chercheurs est un cas particulier de la perception des nombres, pour laquelle Schneider *et al.* (1974) ont mis en évidence que le "nombre psychologique" est une fonction puissance du nombre.

Mais cette relation demande à être confirmée. D'une part la consigne, et donc ce qui est jugé, doivent être éclaircis, d'autre part il conviendrait de mettre en évidence la loi au niveau individuel.

Expérience complémentaire

Ultérieurement Nelson, Rosenthal et Rosnow (1986) ont cherché à compléter cette première étude, d'une part en améliorant la représentativité de leur échantillon, d'autre part en introduisant de nouveaux facteurs expérimentaux. Ils ont envoyé un questionnaire à 294 psychologues américains, tirés au sort, et ont reçu 85 réponses. La consigne est cette fois un peu plus précise puisqu'il s'agit d'indiquer le degré de confiance dans l'hypothèse de recherche (toujours sur une échelle en six points). Les facteurs expérimentaux pris en compte sont :

- le seuil,
- la taille de l'échantillon ($N = 10$ ou $N = 100$), ou le fait que la recherche soit une réplique,
- la grandeur de l'effet (mesurée par un coefficient de corrélation r, donc relative),
- l'importance attribuée à la recherche (soit il s'agit d'une expérience biomédicale où l'on mesure le taux de survie de sujets humains, soit il s'agit d'un questionnaire de psychologie portant sur des attitudes),
- le degré d'expérience du chercheur.

En plus de l'effet du seuil p, ils constatent :

- qu'un effet "chute" est présent pour 0.05 et 0.10;
- que la confiance est plus grande pour le plus grand effectif (confirmant ainsi les résultats antérieurs), et plus encore dans le cas d'une réplique;
- que la confiance grandit avec la taille de l'effet mais que cela est tempéré par le degré d'expérience des sujets.

Rien n'est dit sur l'influence de l'importance attribuée à la recherche; très certainement son effet est non significatif. Encore une fois, rien n'est dit, non plus, sur la forme particulière de la courbe de confiance, sinon bien sûr qu'elle décroît avec le seuil p.

6.5. EXPÉRIENCE 2 : PERCEPTION DES SEUILS OBSERVÉS

Cette expérience, qui est une réplique de l'expérience de Rosenthal et Gaito de 1963 que nous venons de présenter, a pour objectif principal l'étude individuelle de la perception des seuils de signification. Les courbes présentées par Rosenthal et Gaito ne concernaient que des moyennes, ce qui est insuffisant pour étudier les processus cognitifs sous-jacents. Les courbes moyennes ne renvoient à une réalité psychologique que si l'on peut montrer l'homogénéité des sujets.

Notre objectif est double: (1) identifier des classes de comportements et (2) voir si la fonction puissance ("modèle psychophysique") se retrouve au niveau individuel.

6.5.1. Méthode

Facteurs expérimentaux

La situation présentée au sujet est celle d'une expérience qu'il a réalisée pour tester l'efficacité d'un traitement. Nous avons fait varier trois facteurs. Les deux premiers sont les mêmes que dans l'expérience originale, c'est-à-dire la valeur du seuil *p* du test statistique effectué et la taille *N* de l'échantillon. Un troisième facteur a été ajouté afin de déterminer l'influence possible de la forme de la consigne. Il s'agit de savoir si deux formes différentes, équivalentes d'un point de vue formel, le sont également du point de vue psychologique.

- Le seuil observé *p*.
Douze valeurs ont été retenues : 0.001, 0.01, 0.03, 0.05, 0.07, 0.10, 0.15, 0.20, 0.30, 0.50, 0.70, 0.90 (il y en avait quatorze dans l'expérience de Rosenthal et Gaito).
Ce nombre de valeurs résulte d'un compromis entre disposer de suffisamment de points pour tracer la courbe, d'une part, et éviter que la tâche des sujets ne devienne trop fastidieuse, d'autre part. Compte tenu du rôle des résultats significatifs dans la pratique des tests, les petites valeurs de *p* ont été plus représentées que les grandes.

- La taille *N* de l'échantillon.
Deux modalités ont été retenues : $N = 10$ et $N = 100$; ce sont les mêmes que dans l'expérience originale.

- La formulation de la consigne.
Deux formes complémentaires et équivalentes d'un point de vue formel ont été étudiées : on demandait au sujet d'indiquer, soit son degré de confiance en l'hypothèse alternative (selon laquelle le traitement a réellement un effet), soit sur son degré de confiance en l'hypothèse nulle (selon laquelle le traitement n'a pas d'effet). Ces modalités correspondent aux formulations habituellement données par les chercheurs quand ils sont interrogés sur l'interprétation d'un test statistique (*cf.* 6.1.), la première correspondant davantage à la consigne de l'expérience originale.

Chaque sujet a été affecté à l'une des deux formulations de la consigne et a été soumis aux 24 combinaisons des facteurs *p* et *N*.

Matériel

Afin d'obtenir des courbes individuelles présentant suffisamment de finesse pour pouvoir les comparer à une fonction précise, nous avons choisi d'utiliser une échelle analogique pour les mesures de confiance. Des échelles comportant un nombre limité de points, comme celle de l'expérience originale (échelle en six points) auraient été trop imprécises. Nous avons également rejeté la solution de demander directement au sujet une évaluation chiffrée de son degré de confiance (par exemple en lui demandant d'attribuer une note entre 0 et 10 ou 0 et 100) pour ne pas favoriser *a priori* le choix du complément de la valeur *p* présentée. Par ailleurs, la droite analogique présentée, qui mesure 10 cm, ne comporte pas de marques pour des valeurs particulières (telle une confiance "moyenne") pour éviter d'éventuels effets d'ancrage.

Une échelle se présente ainsi :

N = 10 *p* = .001

Les 24 échelles (12 valeurs p × 2 tailles d'échantillons) sont regroupées en un petit cahier de format 21 cm × 10 cm comprenant une échelle par page. Chaque sujet répond donc pour l'ensemble des conditions. Les 24 échelles sont présentées dans un ordre aléatoire, différent pour chaque sujet. Ce mode de présentation a été choisi afin d'éviter une cohérence artificielle qui aurait pu être induite par une présentation simultanée et/ou ordonnée des échelles.

Consigne

La consigne est plus explicite que celle de Rosenthal et Gaito. En particulier, il est spécifié que la confiance demandée porte bien sur l'*effet du traitement expérimental*. Cela implique que l'interprétation I2 de la consigne de Rosenthal et Gaito que nous avons faite en 6.4. ne peut être retenue ici (confiance en H_1 pour les résultats jugés significatifs, confiance en H_0 pour les autres). Il est également indiqué qu'il s'agit de groupes appariés pour éviter que le sujet se demande si N se réfère à l'effectif par groupe ou à l'effectif total. Cependant cette consigne reste assez proche de l'originale pour que les conditions soient comparables.

(1) Consigne “hypothèse alternative”
Cette première forme est la suivante :

Je vous demande de vous placer dans la situation suivante.

Vous avez effectué une expérience pour tester l'effet d'un traitement.
A partir des résultats recueillis vous avez réalisé un test statistique (disons un *t* de Student pour groupes appariés).
Je vais vous présenter une liste de seuils observés (valeurs de *p*) correspondant à divers résultats possibles, soit sur un échantillon de **N = 10** sujets, soit sur un échantillon de **N = 100** sujets. Pour chacun de ces deux échantillons et chacun de ces *p* possibles vous devrez indiquer le **degré de confiance** que vous avez dans l'hypothèse que **le traitement a réellement un effet**.
Pour cela je vous demande de cocher un point sur un segment dont l'extrémité gauche représente une confiance nulle et l'extrémité droite une confiance totale. Vous cocherez un point d'autant plus vers la gauche que votre confiance est faible, et d'autant plus vers la droite que votre confiance est grande.

Je vous demande de répondre le plus spontanément possible et sans revenir en arrière.

Je vous remercie de votre participation.

(2) Consigne “hypothèse nulle”
Dans cette seconde forme la phrase précisant la tâche à réaliser est changée en :

[...] vous devrez indiquer le **degré de confiance** que vous avez dans l'hypothèse nulle que **le traitement n'a pas d'effet**.

En fait le sujet a la possibilité de revenir en arrière pour éventuellement corriger une réponse dans le cas où il exprime un doute concernant une condition présentée antérieurement (par exemple, s'il pense avoir confondu les conditions $p = .01$ et $p = .10$). Cela s'est produit seulement cinq fois.

Sujets

Le premier groupe de sujets (consigne “hypothèse alternative”) comprend 19 chercheurs (les 19 chercheurs en psychologie de la région parisienne qui ont passé l'expérience 1 “psychologues et statisticiens”), et le second (consigne “hypothèse nulle”) 18 autres chercheurs comparables.

Tous les chercheurs contactés ont accepté de répondre à l'étude. On supposera donc qu'il n'existe pas, à ce niveau, de biais de sélection, tel qu'il y aurait pu en avoir si seuls ceux se sentant suffisamment “armés” en statistiques avaient accepté de répondre.

Dans le but d'estimer, au moins grossièrement, la variabilité intra-sujet des réponses, nous avons procédé à un retest pour trois des sujets du premier groupe[29]. Cette seconde passation s'est effectuée environ six mois après la première, dans des conditions identiques.

Remarque

En ce qui concerne les 19 psychologues du premier groupe qui ont participé aux deux expériences, la présente expérience a été passée systématiquement en premier. En effet, cet ordre de passation était celui qui posait le moins de problèmes car l'expérience sur les seuils se déroulait rapidement et facilement (dans le cas de la première forme de la consigne) et présentait en outre peu de risque d'influencer les réponses à l'expérience "psychologues et statisticiens".

Passation

La passation était individuelle, au cours d'un rendez-vous fixé à l'avance. Après une brève présentation du but de la recherche ("Dans le cadre d'une thèse, je m'intéresse à l'utilisation des statistiques par les chercheurs en psychologie, *etc.*") la consigne est lue au sujet et lui est présentée sur une feuille séparée restant à sa disposition. Après lecture de la consigne le sujet a la possibilité de poser des questions (mais aucun sujet n'a utilisé cette possibilité; tous ont trouvé la consigne claire), puis le cahier de réponses lui est présenté.

Il n'a pas été proposé d'exemples destinés à ce que les sujets s'entraînent à la tâche, "calibrent" leur réponses. Nous avons estimé en effet que ce calibrage était déjà réalisé en raison de la nature même de la tâche et des sujets : ce sont des chercheurs confirmés qui ont donc l'habitude de lire des articles mentionnant des résultats de tests statistiques, et ils savent que les seuils ne peuvent varier qu'entre 0 et 1.

Une pré-expérience, effectuée sur six sujets, a montré que la consigne, dans sa première forme, était claire et la tâche réalisée rapidement et sans difficulté.

Dans le cas de la première forme de la consigne, la passation n'a pas posé de problème et s'est déroulée rapidement; la durée de passation a été de l'ordre de 5 minutes en moyenne, et dans tous les cas elle restée inférieure à 10 minutes. Par contre, la deuxième forme de la consigne s'est révélée "perturbatrice" pour la plupart des sujets.

Cela nous a conduit à effectuer des analyses séparées de chacune des deux formes.

Mesure des jugements de confiance

Les résultats des jugements de confiance, longueurs sur le segment de droite analogique de 10 cm de long, ont été mesurés à 0.5 mm près et ont été ramenés à des mesures variant entre 0 (confiance nulle, extrémité gauche du segment) et 1 (confiance extrême, extrémité droite du segment). Autrement dit, un écart de confiance de 0.10 correspond à un intervalle de 1 cm sur le segment de droite.

le sujet 11 (Figure C19), ayant des résultats qui nous sont apparus inexplicables et aberrants, a été exclu des analyses. En effet, pour $N = 10$ et $N = 100$, et au contraire de ce que l'on obtient chez les autres sujets, les deux courbes ne sont pas monotones, même approximativement, ne présentent pas la décroissance attendue de la confiance en fonction du seuil, et ne se ressemblent pas. Les analyses porteront donc sur 18 sujets pour chacune des deux formes de la consigne.

Nous considérerons d'abord les résultats relatifs à la consigne "hypothèse alternative" (première forme); ceux relatifs à la consigne "hypothèse nulle" seront présentés dans la section 6.5.5.

Dans l'analyse des résultats qui suit, certaines des conclusions descriptives seront généralisées en utilisant des procédures bayésiennes standard. Pour la méthode, on pourra se référer à Bernard (1986, 1991) et Lecoutre *et al.* (1995) pour ce qui concerne les procédures sur des fréquences, et à Lecoutre (1984a, 1996) pour les procédures sur des variables numériques. Les calculs ont été réalisés au moyen des logiciels LesProportions et PAC (Lecoutre et Poitevineau, 1996, 1992).

[29] Le retest n'avait pas été prévu à l'origine, et nous n'avons pu y procéder que pour les sujets dont les cahiers de réponse étaient suffisamment bien identifiés, ce qui explique le faible nombre de retests.

6.5.2. Consigne “hypothèse alternative” : courbes moyennes

La figure suivante (Figure 4) présente les courbes moyennes obtenues sur les 18 sujets retenus. Sont indiqués les r^2 pour la fonction puissance et la fonction exponentielle.

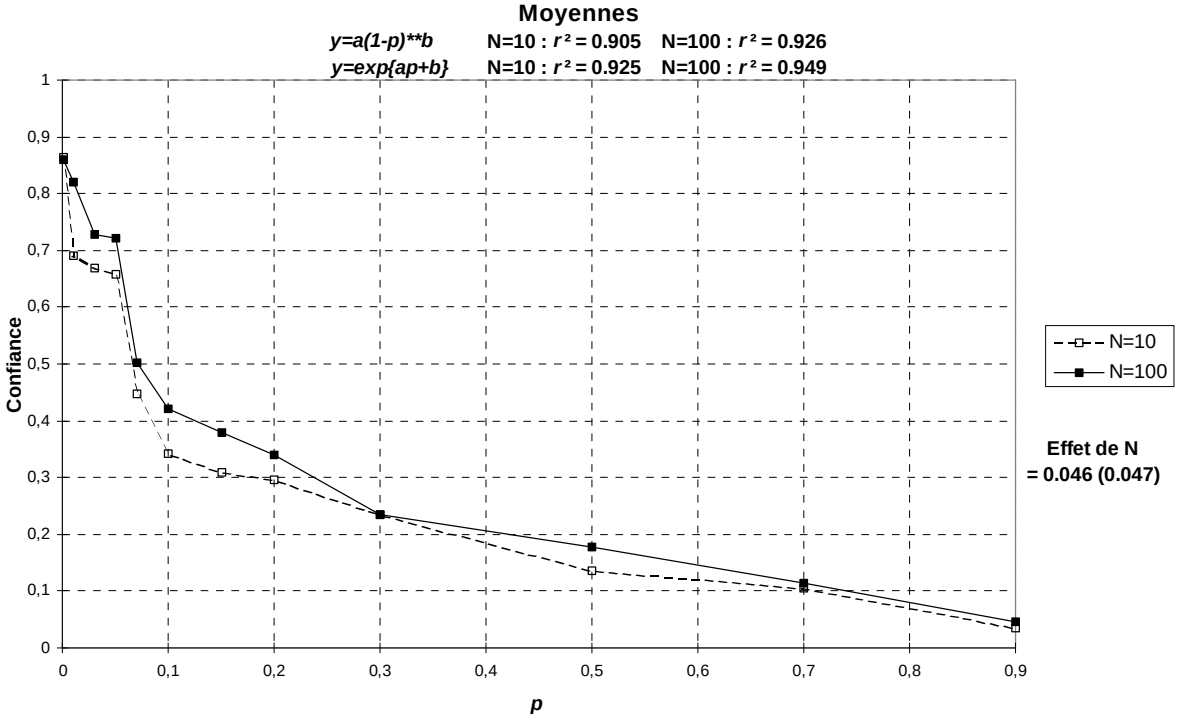


Figure 4

Moyenne des confiances exprimées par des chercheurs en psychologie en fonction du seuil observé (p) et de la taille de l'échantillon (N=10, courbe en pointillé, ou N=100, courbe en trait plein). Les ajustements en termes de r^2 sont rapportés, pour chaque condition de N, pour la fonction puissance (modèle psychophysique) et la fonction exponentielle. L'effet de N est la moyenne, sur les douze seuils, des différences “courbe N=100” - “courbe N=10”; entre parenthèses est indiquée la moyenne des valeurs absolues de ces différences

Dans les grandes lignes notre expérience, après celles de Beauchamp et May (1964) et de Nelson *et al.* (1986), confirme les résultats de Rosenthal et Gaito (*cf.* la Figure 3 présentée en 6.4.) : même type de courbes, de caractère exponentiel, et même effet global de la taille N de l'échantillon, la courbe moyenne pour $N = 100$ étant partout supérieure ou égale à la courbe moyenne pour $N = 10$.

L'effet moyen de l'effectif N est ici de 0.046, moyenne sur les douze seuils des différences “courbe 100 - courbe 10” ($Pr(\ 0.011 < \textit{effet vrai} < 0.081) = 0.90$). Chez Rosenthal et Gaito, sans tenir compte des seuils au delà de .50 pour lesquels la confiance était nulle, cet effet est, dans les mêmes unités, de 0.063 chez les chercheurs et de 0.052 chez les étudiants. Si l'on se limite aux huit seuils communs aux deux expériences, on trouve : 0.044 (ici), 0.066 (chercheurs, 1963), 0.053 (étudiants, 1963). L'effet est donc limité en intensité, correspondant à un écart de l'ordre du demi centimètre sur le segment de 10 cm, et la différence observée entre les deux expériences apparaît négligeable.

L'utilisation d'un type différent d'échelles (droite analogique au lieu d'une échelle en six points), le changement de population et de consigne n'ont donc pas fondamentalement affecté le phénomène[30].

Notons tout de même certaines différences entre les résultats de ces deux expériences.

[30] De manière à essayer de voir si l'introduction d'une échelle analogique pouvait tout de même avoir eu un effet, nous avons recodé les données en six points (0, 1/5, 2/5, …) et recalculé les courbes moyennes. Bien sûr, il est impossible d'affirmer que c'est ce que l'on aurait réellement obtenu avec une échelle en six points. Ces courbes moyennes sont présentées dans la Figure C21 de l'annexe C. Les modifications sont minimes, les courbes “recodées” et “brutes” sont extrêmement proches; on notera simplement que l'effet plateau à $p = 0.05$ pour $N = 10$ est encore plut net avec le recodage en six points (et l'effet de N devient identique à celui des étudiants : 0.052). Nous pensons donc que les différences entre les deux expériences ne tiennent pas à la différence de procédure.

Nos courbes moyennes ressemblent davantage à celles des étudiants de Rosenthal et Gaito qu'à celles des chercheurs, en raison de la présence d'un "effet plateau" pour le seuil .05 (nous reviendrons sur ce point ultérieurement) et d'un degré de confiance comparable pour les points jusqu'à $p = .05$.

Au delà de $p = .10$, nos courbes sont pratiquement linéaires, avec un niveau de confiance nettement plus élevé que dans l'expérience originale. On remarquera tout particulièrement qu'en moyenne la confiance n'est pas nulle pour des seuils de signification relativement élevés ($p > .30$).

L'ajustement par une fonction puissance est satisfaisant, mais il est cependant un peu moins bon que dans l'expérience originale; il est également moins bon que celui donné par un modèle exponentiel.

Mais les courbes moyennes masquent en réalité des comportements bien différents, comme on va le voir maintenant à l'examen des courbes individuelles.

6.5.3. Consigne "hypothèse alternative" : identification de classes de sujets

Les courbes obtenues, individuelles et moyennes, sont présentées en annexe C.

Consigne "hypothèse alternative" : Figures C1 à C32.

- Figures C1 - C19 : courbes individuelles.
- Figures C20 - C24 : courbes moyennes.
- Figures C25 - C32 : courbes individuelles (test/retest).

Si la plupart des sujets ont utilisé pratiquement toute l'étendue de l'échelle (avec une confiance maximum d'au moins 0.85), deux sujets (6 et 10) n'ont pas dépassé une confiance de 0.5, et un sujet (5) n'a pas dépassé 0.70.

Étude de la monotonie des courbes individuelles

En dehors même de toute considération théorique, la monotonie des courbes est une propriété attendue. Bien que les courbes soient globalement monotones, il n'en va pas systématiquement de même dans le détail. Mais, étant donné la nature de la tâche, il y a tout lieu de s'attendre à un écart à la monotonie parfaite du fait de possibles fluctuations dans la réponse du sujet (ne serait-ce qu'au niveau moteur), représentant une sorte d'"erreur de mesure". Nous avons donc considéré que des écarts de confiance inférieurs ou égaux à 0.05 (soit 5 mm), voire à 0.10, entre deux points consécutifs ne devaient pas être pris en compte pour juger du caractère monotone ou non des courbes. Même ainsi il n'est pas exceptionnel d'observer des écarts à la monotonie; cependant, des retests effectués pour trois sujets (qui seront présentés dans la section suivante) montrent que des écarts de confiance de l'ordre de 0.20 entre les deux passations ne sont pas rares, ce qui laisse à penser que le critère de 0.10 est loin d'être trop élevé (à condition de supposer que les deux passations renvoient, pour les jugements concernés, au même état psychologique du sujet). Le tableau ci-dessous indique, pour chaque sujet et chacune des deux courbes, le nombre de points pour lesquels l'écart est supérieur au critère ($\Delta = 0.05$, $\Delta = 0.10$ et $\Delta = 0.20$).

Sujet	Δ = 0.05		Δ = 0.10		Δ = 0.20	
	N=10	*N*=100	*N*=10	*N*=100	*N*=10	*N*=100
1	2	0	1	0	1	0
2	0	0	0	0	0	0
3	3	1	0	0	0	0
4	1	1	1	0	0	0
5	0	4	0	3	0	2
6	1	0	1	0	1	0
7	2	2	1	1	0	0
8	2	1	0	0	0	0
9	0	0	0	0	0	0
10	1	0	0	0	0	0
12	1	1	0	1	0	0
13	2	2	1	2	0	1
14	1	0	0	0	0	0
15	2	2	0	1	0	0
16	2	1	1	0	0	0
17	0	0	0	0	0	0
18	3	1	2	1	1	0
19	0	0	0	0	0	0
courbes monotones	5	8	11	12	15	16

Tableau 24

Nombre de points par lesquels les courbes de confiance s'écartent de la monotonie selon le critère de monotonie adopté (Δ), pour chaque sujet et pour chacune des deux conditions d'effectif.
En grisé les cas de monotonie

La monotonie nous paraît assez bien respectée au niveau individuel. Quatre sujets ont des réponses totalement monotones avec Δ = 0.05, six avec Δ = 0.10 et quatorze avec Δ = 0.20. Il existe cependant un certain nombre d'écarts qui peuvent difficilement être négligés. Ainsi quatre sujets (1, 5, 6 et 13) présentent un écart supérieur à Δ = 0.30 (soit 3 cm) pour l'une ou l'autre des courbes (mais non pour les deux simultanément).

Il faut rejeter l'hypothèse que ces points "aberrants" correspondent au premier jugement du sujet et que celui-ci n'aurait pas encore calibré ses réponses : aucune des échelles correspondantes ne figurait en tête du cahier de réponse.

Une autre explication de ces grands écarts pourrait être une confusion, soit sur le seuil à juger, par exemple lire $p = .01$ au lieu de $p = .10$ ou inversement, soit sur l'effectif. Ainsi, en ce qui concerne le sujet 1, sa confiance "aberrante" en $p = .20$ pour $N = 10$ (0.55) est pratiquement identique à celle donnée pour $N = 100$ (0.535) et pourrait résulter d'une confusion sur l'effectif. Ceci est plus ou moins confirmé par le fait que cette "aberration" a disparu dans la réplique que ce sujet a effectuée (*cf.* Figure C26). La valeur très basse (0.080) donnée par le sujet 5 pour $p = .03$ ($N = 100$) est très voisine de celle donnée pour $p = .30$ (0.075) et pourrait s'expliquer par une confusion sur le seuil. Toujours pour ce sujet et pour $N = 100$, il est possible qu'il y ait également inversion de $p = .001$ et $p = .01$ (ce qui serait cohérent avec les réponses pour $N = 10$).

Pour le sujet 13, la valeur du point aberrant (confiance = 0.040, $p = .10$, $N = 100$), qui correspond à une échelle placée en 9ème position dans le cahier de passation, est proche de celle accordée pour le 8ème jugement (confiance = 0.060, $p = .50$, $N = 10$). Il pourrait s'agir d'une répétition du jugement précédent.

Les retests

Ces retests concernent trois sujets (1, 15, 17) et ont été effectués dans un délai de quatre à six mois après la première passation, dans des conditions semblables (passation individuelle). Ce délai paraît plus que suffisant pour éliminer un éventuel effet de mémorisation de leurs réponses par les sujets (effet d'ailleurs assez peu plausible étant donné le nombre de réponses à fournir). Ces sujets nous ont déclaré que leurs connaissances statistiques n'avaient pas évoluée entre les deux passations. Les résultats sont présentés en annexe C, Figures C25 à C32.

La fidélité, pour ces sujets, est bonne, mais des variations de confiance de l'ordre de 0.15 ou 0.20 ne sont pas rares. Cet ordre de grandeur nous semble corroboré par le fait suivant. Chez les sujets 1 (pour $N = 10$) et

15 (pour $N = 10$ et $N = 100$), on peut observer, par rapport au niveau en $p = .50$, une remontée de la confiance en $p = .70$ qui peut atteindre 0.17. Or, il nous semble peu plausible que la confiance en $p = .70$ soit réellement plus grande qu'en $p = .50$, et nous préférons y voir une indication de la variabilité intra sujet, cohérente avec les écarts constatés entre test et retest.

Effet de la taille N de l'échantillon

Pour juger de l'équivalence de deux niveaux de confiance (à un seuil fixé), nous nous sommes donné le même critère que pour juger de la monotonie des courbes ($\Delta = 0.10$, voire $\Delta = 0.20$). Sur les Figures C1 à C19 de l'annexe C, un effet chiffré de la taille N de l'échantillon est précisé. Il s'agit, d'une part, de la moyenne, sur les douze seuils, des différences "courbe $N = 100$" - "courbe $N = 10$", et d'autre part, indiquée entre parenthèses, d'une distance entre les deux courbes calculée comme la moyenne des valeurs absolues des différences.

Nous avons regroupé les sujets en trois classes :

- Classe "$10 \leq 100$" : la courbe "$N = 100$" est supérieure à la courbe "$N = 10$" pour la majorité des points (soit au moins 7), et les courbes sont égales pour les autres points (dans les limites d'approximation utilisées). On groupe dans cette classe 8 sujets (1, 4, 6, 10, 14, 15, 16, 18).

- Classe "$10 \geq 100$" : nous adoptons la même définition que pour la classe précédente, en inversant le rôle des courbes. On groupe dans cette classe 2 sujets (7 et 8); on pourrait également placer le sujet 7 dans la classe suivante, car il ne présente cet effet que pour les seuils allant de .05 à .50, les autres seuils donnant lieu à l'égalité.

- Classe "$10 = 100$" : les conditions pour les deux classes précédentes ne sont pas remplies et de plus les deux courbes sont égales, dans les limites d'approximation convenues. On groupe dans cette classe 8 sujets (2, 3, 5, 9, 12, 13, 17, 19). En réalité les sujets 5 et 13 présentent des écarts supérieurs au critère maximum d'équivalence ($\Delta = 0.20$), mais nous les avons tout de même inclus dans cette classe parce que les écarts ne sont pas systématiquement dans un sens et aussi pour ne pas multiplier les classes.

Il est donc rare (2 sujets sur 18) d'accorder, pour un même seuil, plus de confiance au résultat provenant de l'échantillon le plus faible ($Pr(\varphi < 0.22) = 0.90$, φ représentant la fréquence parente de tels sujets). Il est vraisemblable que ces sujets considèrent, comme Bakan (1966), qu'un même seuil atteint avec un faible effectif est le signe d'un effet plus fort et que c'est cette intensité de l'effet qui joue sur le niveau de confiance (pour atteindre le même seuil p alors que N diminue, il faut que l'effet augmente, et/ou que l'écart-type diminue).

Les deux autres cas sont aussi fréquents l'un que l'autre (pour chacun d'eux $Pr(0.26 < \varphi < 0.63) = 0.90$). L'effet dans le sens "$10 \leq 100$" peut être simplement expliqué en supposant que les sujets appliquent ici le résultat statistique, qu'ils connaissent, selon lequel la précision des estimations que l'on peut faire est d'autant meilleure que l'effectif est grand. Pour interpréter le cas "$10 = 100$", il est tentant d'invoquer l'hypothèse de représentativité de Kahneman et Tversky, qui a pour conséquence la "loi des petits nombres" (un échantillon de faible effectif est perçu comme apportant une information aussi précise qu'un échantillon plus important), et qui semblerait donc adaptée à une bonne partie des sujets (8/18). Cependant, dans la situation particulière qui nous occupe, où les sujets ne sont pas totalement "naïfs" mais possèdent un minimum de connaissance en statistique, d'autres interprétations sont à considérer. En particulier, nul doute que certains sujets considèrent que "l'effectif ne doit pas intervenir dans le jugement car le p en tient déjà compte" (et c'est d'ailleurs bien ainsi que Fisher présente le seuil de signification, *cf.* 1.1.). Deux sujets l'ont spontanément déclaré, mais il est impossible de savoir si cela en concerne d'autres. Au plus, cela laisse six sujets pour lesquels l'hypothèse de représentativité pourrait s'appliquer.

En fait, pour discuter plus avant de l'effet de N, il est nécessaire de tenir compte simultanément du type de courbes, aspect que nous allons aborder maintenant, en relation avec les modèles théoriques de la section 6.4.

Forme des courbes et courbes théoriques

Comme il était prévisible, les courbes sont globalement décroissantes en fonction du seuil p. Au vu des résultats nous distinguons trois formes principales de courbes :

- Courbes de type linéaire (LIN), correspondant à 4 sujets (4, 7, 18, 19).

- Courbes de type “tout ou rien” (TOR), correspondant à 4 sujets (3, 5, 9, 17). Ces courbes sont caractérisées par la juxtaposition de deux parties horizontales (la coupure se faisant probablement au seuil *p* correspondant au risque de première espèce retenu par le sujet). Pour classer une courbe dans cette catégorie, nous avons retenu comme critères la présence de deux plateaux, au moins approximativement, et d'un saut brutal entre ceux-ci.

- Courbes de type exponentiel (EXP), correspondant à 10 sujets (1, 2, 6, 8, 10, 12 à 16). Nous groupons dans cette classe toute courbe de forme approximativement exponentielle, ce qui inclut notamment la fonction puissance.

Pour chacune des classes et chacune des valeurs de *N*, nous avons calculé la courbe moyenne, qui correspond cette fois à des sujets homogènes. Si l'on admet, hypothèse très forte, que les sujets sont parfaitement homogènes, tant en ce qui concerne la famille de courbes, qu'en ce qui concerne les valeurs des paramètres, la courbe moyenne devient une bonne estimation d'une courbe “vraie”. On peut au moins considérer que la courbe moyenne corrige quelques fluctuations aléatoires sans signification psychologique particulière.

Un premier résultat intéressant réside dans le fait que chez tous les sujets le type de courbes est le même pour les deux conditions d'effectif $N = 10$ et $N = 100$, ce qui indique une cohérence certaine des sujets dans leurs jugements. On peut également en inférer que l'effet de *N*, quand il existe, est limité à une translation de la courbe, c'est-à-dire à une action sur l'intensité et qu'il n'affecte pas plus profondément les processus en jeu.

Nous discutons maintenant les résultats obtenus pour chacune de ces trois classes en référence aux prévisions théoriques décrites dans la section 6.4. Un modèle théorique devra correctement décrire conjointement la forme de la courbe et l'effet de *N* pour être retenu.

Le tableau suivant met en relation le type de courbes et l'effet de *N*.

	10=100	10≤100	10≥100	
LIN	1	2	1	4
TOR	4	0	0	4
EXP	3	6	1	10
	8	8	2	18

Tableau 25

Association entre le type de courbes (en lignes) et l'effet de N (en colonnes)

Classe LIN (*sujets* 4, 7, 18, 19; *Figures* C11 *à* C14)

Sur les figures il est indiqué, pour chacune des courbes, le r^2 relatif à l'ajustement des points par rapport à une droite $y = a(1-p)+b$ ainsi que l'estimation des paramètres.

Seul le cadre fiducio-bayésien associé à l'énoncé I4 (effet vrai allant dans le même sens que l'effet observé) prévoit une droite, avec $a = 1$ et $b = 0$ et un effet “10 = 100” pour *N*. On constate un bon accord des courbes moyennes (Figure C23) avec ce modèle. Au lieu d'une conformité à une théorie fiducio-bayésienne, on peut tout aussi bien voir chez les sujets une illustration de l'usage abusif, et fréquent, des tests traditionnels consistant à penser que $1 - p$ est la probabilité que l'hypothèse alternative soit vraie (*cf.* 2.2.1.). Dans la mesure où le cadre fiducio-bayésien fournit une réinterprétation naturelle de cet usage abusif (*cf.* 3.5.), il n'y a pas de réelle opposition entre ces deux interprétations, et nous qualifierons ces sujets de “fiducio-bayésiens naturels”.

La tendance linéaire est bien marquée chez les quatre sujets, mais les ajustements sont plus ou moins bons (les r^2 varient de 0.795 à 0.984). Pour le sujet 4, l'ajustement est notablement affaibli par les valeurs pour $p = .90$ qui présentent une nette décroissance par rapport à celles pour $p = .70$. L'effet de *N* varie d'un sujet à l'autre; il est surtout notable pour les sujets 7 et 18, mais de sens opposé d'où l'effet moyen faible (effet observé : 0.033, $Pr(-0.075 < \text{effet vrai} < 0.141) = 0.90$). La compatibilité avec le modèle statistique vaut donc surtout pour les résultats du sujet 19 qui sont particulièrement remarquables : l'ajustement des points est excellent[31] et les deux courbes $N = 10$ et $N = 100$ sont très proches (la distance entre les deux courbes est inférieure à 0.05).

[31] Par rapport à la droite $y = 1-p$, on trouve, respectivement pour les deux courbes, $r^2 = .956$ et $r^2 = .973$.

Classe TOR (*sujets* 3, 5, 9, 17; *Figures* C15 *à* C18)

Les quatre sujets ont en commun un très fort effet de "chute" de confiance après le seuil $p = .05$ (.07 pour le sujet 5 et $N = 10$), celle-ci devenant nulle ou très faible au delà, alors qu'elle est très forte (sauf pour le sujet 5) en deçà, et un effet de *N* du type "10 = 100". Les courbes moyennes (Figure C24) les résument bien.

Le sujet 5 est celui qui pose le plus de problèmes. Nous avons discuté plus haut de la possibilité de confusions de lecture. Si nous admettons toutes les hypothèses formulées à cette occasion, le début de sa courbe $N = 100$ ressemblerait à celui de la courbe $N = 10$. À l'inverse, les résultats du sujet 17, qui n'utilise que les valeurs extrêmes, sont presque caricaturaux et n'ont absolument pas varié au cours de la réplique qu'il a effectuée. On ne peut expliquer ces résultats simplement par la théorie de Neyman et Pearson, car si la présence des plateaux correspond bien aux prévisions dans le cadre de cette théorie, en revanche on ne retrouve pas du tout l'effet attendu pour *N*. Nous qualifierons ainsi ces quatre sujets de "pseudo neyman-pearsoniens", dans la mesure où leurs résultats paraissent influencés par cette théorie d'une manière naïve.

De façon voisine, on peut invoquer ici une interprétation "hybride" (au sens de Gigerenzer, 1993), la théorie de Neyman et Pearson expliquant les plateaux, et la théorie de Fisher expliquant l'effet de *N* ("10 = 100"). Les résultats sont également en accord avec l'hypothèse de significativité de Oakes (1986) (*cf.* 6.2.2.), qui exprime le fait que pour ces sujets le jugement de confiance s'apparente à une décision, ou, tout au moins, présente un caractère binaire. Dans la mesure où les interprétations précédentes, et qui sont en réalité différentes formulations d'une même conception, semblent convenir, il n'est pas nécessaire ici de faire appel à l'hypothèse de représentativité de Kahneman et Tversky (compatible avec l'effet "10 = 100").

Notons enfin qu'il pourrait exister, pour $N = 10$, un "double effet de chute" puisque l'on constate, sur la courbe moyenne, une diminution importante entre $p = .001$ et $p = .01$; ce qui ne serait pas explicable par les hypothèses précédentes. Cependant cet effet est imputable quasi exclusivement au seul sujet 5, et, compte tenu de l'amplitude des variations intra individuelles, son existence est loin d'être assurée.

Classe EXP (*sujets* 1, 2, 6, 8, 10, 12 *à* 16; *Figures* C1 *à* C10).

Si cette catégorie est majoritaire, elle est également la plus disparate. Ainsi les courbes du sujet 6 ne sont de forme exponentielle que pour $.05 \leq p \leq .15$. Nous les aurions caractérisées comme TOR (existence d'un plateau pour $p \geq .15$, et d'un autre, plus approximatif, pour $p \leq .05$) si ce n'est que la décroissance après $p = .05$ ne paraît pas suffisamment brutale. Ce sujet apparaît plutôt comme intermédiaire entre les classes TOR et EXP. Quant aux courbes du sujet 2, elles évoquent davantage une sigmoïde qu'une exponentielle ou une fonction puissance.

Les courbes moyennes sont données dans la Figure C22 (inclure ou non le sujet 6, marginal, ne change pratiquement pas les résultats). Ces courbes sont encore plus semblables à celles de Rosenthal et Gaito que les courbes moyennes pour l'ensemble des sujets, ce qui laisse penser que ces auteurs ont eu affaire en majorité à des sujets de type EXP. On remarque que l'effet de *N* est un peu plus manifeste. Sur l'ensemble des seuils, le cas $N = 100$ est supérieur au cas $N = 10$, la différence moyenne observée (0.070) étant très semblable à celle obtenue par Rosenthal et Gaito ($Pr(0.014 < \textit{différence vraie} < 0.126) = 0.90$). Ce dernier résultat amène déjà à rejeter l'hypothèse que les sujets se conformeraient, dans leurs jugements, au modèle fiducio-bayésien associé à un énoncé d'effet vrai important (I3) puisque celui-ci prévoit bien une forme de courbe comparable mais un effet "10 ≥ 100" (seul le sujet 8 correspond à ce cas).

L'effet de *N*, dans le sens "10 ≤ 100", est à interpréter comme nous l'avons fait précédemment lors de la présentation de cet effet, à savoir qu'il est sans doute dû aux connaissances statistiques des sujets ("un plus grand échantillon est préférable").

Remarque

Puisque les 18 sujets considérés ici ont également participé à l'expérience 1, il était intéressant de mettre en rapport les résultats correspondants. Dans ce but nous avons croisé les réponses aux Questions×Situations de l'expérience "psychologues et statisticiens" avec les types de courbes (LIN, TOR, EXP) et les types d'effets de *N* ("10 = 100", "10 ≤ 100", "10 ≥ 100") de l'expérience "seuils *p*", mais nous n'avons pas trouvé d'effet discernable; en raison des faibles effectifs, il n'est pas possible, non plus, d'affirmer qu'il n'existe pas d'effet. Nous n'avons rien distingué de particulier, non plus, en effectuant une analyse des correspondances où les types de courbes et d'effets de *N* étaient portés en variables supplémentaires.

6.5.4. Consigne “hypothèse alternative” : modèle “psychophysique”

Le deuxième objectif de notre expérience était d’examiner la question d'un modèle “psychophysique”, étudiée par l’ajustement de la fonction puissance (ou exponentielle). Cette question ne se pose plus que pour les sujets de la classe EXP.

Sur les figures sont indiqués les r^2 pour la fonction puissance $y = a(1-p)^b$ et la fonction exponentielle $y = \exp\{ap+b\}$[32]. Nous avons ajusté la fonction puissance à $1 - p$ et non directement à p. La raison en est que c'est la quantité $1 - p$ qui se rapporte directement à l'hypothèse alternative (évoquée dans la consigne), et les résultats pour la deuxième forme de la consigne (*cf.* 6.5.5.) semblent confirmer que c'est ce qui paraît naturel aux sujets.

Au niveau moyen les ajustements sont très bons, avec une légère supériorité du modèle exponentiel.

Au niveau individuel les r^2 vont de 0.698 à 0.974 pour la fonction puissance, et de 0.708 à 0.975 pour la fonction exponentielle (si l'on ne tient pas compte des résultats du sujet 6, marginal, pour lesquels les ajustement sont moins bons, les r^2 allant de 0.678 à 0.885). Ces valeurs peuvent paraître élevées mais, à titre de comparaison, on gardera à l'esprit que les r^2 sont de 0.848 ($N = 10$) et 0.822 ($N = 100$) pour le sujet 9 pour lequel ces modèles ne sont pas vraisemblables puisqu'il est classé en TOR. On notera également que la valeur de l'exposant, pour la fonction puissance, varie entre 2.281 et 26.594 pour $N = 10$ et entre 1.317 et 15.838 pour $N = 100$ (pour la courbe moyenne, l'exposant vaut 6.370 pour $N = 10$ et 4.884 pour $N = 100$). Cette grande variabilité inter sujets des exposants est un phénomène bien connu en psychophysique et n'a donc rien d'extraordinaire. Compte tenu de la variabilité intra sujet, il serait vain de vouloir différencier ici les deux types de courbes, puissance et exponentielle, qui apparaissent toutes les deux compatibles avec les résultats des sujets de la classe EXP. On remarquera d’ailleurs que la fonction puissance est loin d'être la seule invoquée dans le domaine des jugements perceptifs (voir, par exemple, Piéron, 1963).

Le résultat remarquable est que le jugement émis donne lieu à une courbe semblable à celles rencontrées en psychophysique; tout se passe donc comme si les sujets de cette classe évaluaient le seuil p comme s'il s'agissait d'une grandeur physique.

On peut encore chercher à rapprocher ces résultats de ceux obtenus par Schneider *et al.* (1974) en ce qui concerne la perception des nombres en général. Les valeurs que nous observons pour les exposants sont très différentes de celles données par ces auteurs, qui sont, pour des données agrégées, de 0.70 ou 0.80 selon la condition expérimentale et correspondent donc à une courbure de sens opposé. Cette différence de courbure ne disqualifie pas forcément l'hypothèse que la perception des seuils de signification s'inscrit dans celle du système des nombres en général. En effet, Schneider *et al.* remarquent que l'exposant varie selon les nombres, mais ils n’ont présenté à leurs sujets que des nombres nettement supérieurs à 1 (entre 17 et 312). On peut imaginer que la présentation de nombres entre 0 et 1 amènerait à un changement plus radical d'exposant, et plus précisément que 1 serait un point d'inflexion de la courbe; mais ceci reste une hypothèse.

On constate également qu'il se maintient, en moyenne, un léger effet plateau pour le seuil $p = .05$, au moins dans le cas $N = 10$, alors que, du fait de la variabilité intra sujet, cet effet n'est pas vraiment distinguable, sauf pour quelques sujets, au niveau des courbes individuelles (ou alors il faudrait également admettre un effet des seuils .03 ou .07, voire .15 ou .20). On peut supposer, pour les sujets de la classe EXP, qu’il n'y a pas d'autre cause à l'effet plateau que l'effet de norme sociale dont l'intensité s'avère dans ce cas très faible : 0.052 pour $N = 10$, 0.024 pour $N = 100$[33] (il serait d’ailleurs pertinent d'inclure ici les résultats du groupe LIN, ce qui diminuerait encore l'intensité de l'effet). Cette faible intensité a pour conséquence que l'hypothèse de “significativité” de Oakes n'est pas adaptée à cette classe de sujets, puisque selon cette hypothèse les sujets devraient essentiellement juger en tout ou rien et donc présenter un fort effet plateau (ou chute).

6.5.5. Consigne “hypothèse nulle”

Les courbes obtenues pour la consigne “hypothèse nulle”, individuelles et moyennes, sont présentées en annexe C dans les Figures C33 à C40.

Il s'agit de savoir si les deux formes de consigne, équivalentes du point de vue formel dans le sens où elles sont complémentaires, le sont également du point de vue psychologique. Bien que le but d'une recherche

[32] Nous avons également calculé l'ajustement par rapport au modèle $y = \exp\{ap+b\}+c$, mais les résultats sont peu améliorés (le gain sur le r^2 est de 0.063 au maximum, et de 0.012 en moyenne); et il est plus difficile de comparer ce modèle à la fonction puissance puisqu'il possède un paramètre de plus.

[33] Le calcul consiste à prendre la différence entre la confiance réelle pour $p = .05$ et celle donné par le modèle exponentiel dont les paramètres ont été estimés sans tenir compte du point $p = .05$.

soit le plus souvent de réfuter l'hypothèse nulle, cette deuxième forme est souvent évoquée par les chercheurs pour exprimer les résultats d'un test (*cf.* 2.2.) et devrait donc leur être assez familière. Nous ne nous attendions pas à de grandes différences, mais plutôt à un simple décalage en intensité des courbes et à davantage de courbes LIN que dans le cas de la première formulation du fait que le seuil *p* a trait directement à l'hypothèse nulle.

Dix-huit sujets ont répondu à cette deuxième forme. Leurs comportements nous ont conduit à les répartir en trois sous-groupes :

- Le premier sous-groupe est constitué de 5 sujets.
Les sujets répondent très facilement et rapidement. Mais on s'aperçoit, au vu des résultats, qu'ils répondent comme si on leur avait présenté la première forme de la consigne, un seuil faible donnant lieu à une grande confiance, un seuil fort à une faible confiance. Ceci s'est confirmé au cours d'un entretien suivant la passation : les sujets ont "automatiquement" (sans s'en rendre compte) transformé la consigne en sa première forme.

- Le deuxième sous-groupe est constitué de 5 sujets.
Les sujets prennent bien conscience de la forme exacte de la consigne, mais, bien que cette consigne soit parfaitement et facilement comprise selon leurs propres dires, la tâche se révèle très difficile à réaliser et même impossible. En effet, les sujets déclarent spontanément, dès le début de la passation, que leur tendance naturelle est de répondre à l'inverse de la question posée (donc de répondre à la première forme), et qu'il leur faut un effort certain pour se plier à la consigne. Cette tendance naturelle est si forte qu'après quelques échelles les réponses sont inversées, les sujets ne s'apercevant de leur modification de comportement qu'au bout de quelque temps, en général. Ils le manifestent de façon explicite, en déclarant typiquement : "Ah non, je me suis trompé ! Je me rends compte que j'ai inversé depuis un moment.".

- Le troisième sous-groupe est constitué de 8 sujets.
Ce dernier cas est semblable au précédent, sauf que les sujets parviennent au bout de la tâche, sans déclarer pendant la passation avoir effectué d'inversions. Mais les mêmes remarques que dans le sous-groupe précédent sont souvent faites à l'issue de la passation : tendance naturelle à répondre à l'inverse de la question posée (mais à laquelle il ne leur semble pas avoir succombé); situation jugée "inverse", ou même "bizarre". La durée moyenne de la passation est bien supérieure à celle correspondant à la première formulation, ce qui confirme les dires des sujets quant à la plus grande difficulté de la tâche. Il n'est pas rare que cette durée atteigne ici 10 à 15 minutes.

Le résultat remarquable est ici que la majorité des sujets (10/18) n'ont pas pu respecter la consigne et accomplir la tâche qui leur était demandée. Si l'on rapproche ce résultat de la facilité des réponses à la première forme de la consigne, cela nous amène à l'énoncé suivant :

Il existe une forme naturelle pour juger des seuils p, qui consiste à exprimer la confiance dans ***l'hypothèse alternative*** *(l'hypothèse de recherche).*

Rien ne permettait de prévoir un effet aussi manifeste. Dans les études mentionnées dans la section 6.1. les énoncés relatifs à H_0 sont davantage choisis comme représentant une forme usuelle de compte-rendu d'un test statistique que ceux relatifs à H_1. De plus, l'énoncé relatif à H_0 pouvait *a priori* sembler plus facile puisque, selon l'idée abusive communément partagée par les chercheurs (*cf.* 2.2.1.), le seuil *p* s'applique directement à cette hypothèse (il est considéré comme étant la probabilité que H_0 soit vraie), alors que pour parler de H_1 (tout aussi abusivement) il est nécessaire de procéder à l'opération supplémentaire, certes élémentaire, du passage au complément à 1 du seuil.

Il faudrait maintenant confirmer cet effet et déterminer ses conditions d'obtention; il conviendrait en particulier de soumettre les mêmes sujets aux deux formes de consigne. Dans la mesure où il est vraisemblable que l'hypothèse de recherche était ici identifiée à l'hypothèse alternative (puisque c'est la pratique la plus courante et que la première phrase de la consigne pouvait l'induire), il serait intéressant de vérifier si la difficulté persiste lorsque l'hypothèse de recherche est identifiée à l'hypothèse nulle.

Nous allons maintenant étudier plus précisément les résultats du troisième sous-groupe de sujets qui sont les seuls à avoir accompli la tâche. Ces résultats sont présentés dans les Figures C33 à C40 de l'annexe C où nous avons porté le complément à 1 des mesures originales de confiance pour faciliter la comparaison avec les résultats relatifs à la formulation "hypothèse alternative" de la consigne. Dans ce qui suit, par confiance il faudra

donc entendre “confiance en l'hypothèse alternative”. Par ailleurs, LIN, TOR et EXP renvoient aux mêmes classes que dans les analyses précédentes.

Très globalement la monotonie est respectée, et il est possible que certains des écarts soient à imputer à la difficulté de la tâche. Seul le sujet 23 se distingue par l'ampleur de ses écarts pour ce qui concerne la courbe $N = 10$ où, à part les points en .05 et .15, la confiance est à peu près constante autour de 0.30 (ce qui pourrait correspondre à des inversions sur les hypothèses jugées).

L'effet de la taille N de l'échantillon est bien moins net que pour la première forme de la consigne. Si quatre sujets peuvent être répartis, plus ou moins bien, dans les classes considérées précédemment (“$10 = 100$” pour le sujet 26, “$10 \leq 100$” pour le sujet 24, “$10 \geq 100$” pour les sujets 25, 27), deux autres sujets (21, 22) forment une nouvelle classe “10×100” où les deux courbes se croisent. Quant aux deux derniers sujets (23, 28), nous hésitons à les classer dans cette dernière catégorie car la cohérence des jugements du sujet 23 est sujette à caution, et le croisement observé pour le sujet 28 tient essentiellement aux deux points en .15 et .20; ceci correspond d’ailleurs à ses commentaires spontanés puisqu’il dit juger en tout ou rien indépendamment de N pour $p < .05$ et $p > .15$, et faire intervenir N uniquement quand $.05 \leq p \leq .15$.

Seuls trois sujets nous semblent assez bien correspondre aux classes de courbes identifiées précédemment : le sujet 21 en LIN, le sujet 27 en TOR et le sujet 25 en EXP. Les cinq autres sujets sont plus difficiles à caractériser, intermédiaires entre TOR et EXP, sans qu'il apparaisse vraiment de nouvelle forme (éventuellement une sigmoïde pour le sujet 28).

Malgré cette relative incertitude quant au classement des sujets, il apparaît tout de même que ceux-ci présentent encore une assez bonne cohérence de leurs jugements entre les deux conditions d'effectifs (en mettant à part le sujet 23) : si les deux courbes ne sont pas nécessairement de la même classe (par exemple le sujet 22 pourrait être classé en TOR pour $N = 100$ et en EXP pour $N = 10$), elles restent toutefois assez semblables.

En résumé, il est beaucoup plus difficile ici de faire une typologie des 8 sujets ayant accompli la tâche, ce qui confirme encore les difficultés liées à la deuxième forme de la consigne (“confiance en l’hypothèse nulle”).

6.5.6. Discussion

Les théories statistiques que nous avons envisagées en 6.4. ne fournissent pas véritablement de modèles adéquats des jugements des sujets (sauf pour les sujets 8, 19, et éventuellement 22) qui ne se comportent donc pas comme des statisticiens intuitifs rigoureux. Pour autant, les comportements des sujets relativement à la formulation “confiance en l’hypothèse alternative” de la consigne nous apparaissent assez nettement typés, et, dans la mesure où nous les supposons représentatifs des comportements des chercheurs en psychologie, nous dirons qu'un chercheur se conduit, dans le jugement qu'il porte sur le résultat d'un test de signification :

- Soit comme un statisticien intuitif naïf (dans le sens où cela ne correspond pas réellement aux théories statistiques), selon l'une deux écoles “fiducio-bayésienne naturelle” (classe LIN) ou “pseudo neyman-pearsonienne” (classe TOR). Dans ce cas, l'effectif intervient peu dans le jugement, les sujets considérant que le seuil p en tient déjà compte.
- Soit comme un sujet placé dans une situation de psychophysique, où le seuil p serait le stimulus et agirait comme une caractéristique physique de l'expérience (classe EXP). À ce jugement de type perceptif, il y a toute chance que se superpose alors un effet (de translation) de l'effectif selon un principe de statistique (“plus N est grand, mieux c'est”).

Il est remarquable de constater qu’à l’exception d’un petit nombre de sujets de la classe TOR, la confiance n’est pas nulle pour des seuils de signification élevés ($p \geq .50$). Tout se passe donc comme si la majorité des chercheurs accordent encore une chance à l'hypothèse alternative, même dans le cas d'un résultat du test très “négatif”. Autrement dit, un tel résultat ne leur paraîtrait pas (à raison) suffisant pour écarter définitivement une hypothèse.

En fait, l'hypothèse de “significativité” de Oakes, proposée pour expliquer le comportement des chercheurs vis-à-vis du test de signification, ne s'applique réellement bien ici qu'aux quatre sujets de la classe TOR et n'a donc qu'une portée restreinte. L'hypothèse de “représentativité” de Kahneman et Tversky, a également une portée limitée dans cette situation de jugement des seuils de signification observés.

Si l'effet “plateau à 0.05”, observé dans les courbes moyennes est en partie imputable aux quelques sujets se comportant en “tout ou rien”, il semble exister aussi comme effet de norme sociale, bien que très faible,

chez les sujets de la classe EXP, dans la condition $N = 10$. Peut-être existe-t-il aussi dans le cas $N = 100$, on ne voit d'ailleurs pas de raison pour que ce ne soit pas le cas, mais alors notre expérience n'est pas assez sensible pour le mettre en évidence.

Il peut être rassurant de constater qu'une minorité seulement des chercheurs jugent en "tout ou rien" (4/18). Même si la théorie de Neyman et Pearson a largement pénétré la psychologie, au moins par la terminologie avec les "risques de 1ère et 2ème espèces", la "puissance", elle ne semble pas avoir fondamentalement influencé les esprits pour ce qui est de l'appréciation "privée" du chercheur, et le jugement de Rozeboom, selon lequel il n'y aurait pas de différence appréciable entre $p = .04$ et $p = .06$, est donc toujours en vigueur :

> "And what scientist in his right mind would ever feel there to be an appreciable difference between the interpretive significance of data, say, for which one-tailed $p = .04$ and that of data for which $p = .06$, even though the point of "significance" has been set at $p = .05$?" (Rozeboom, 1960)

Les critiques de cette théorie (*cf.* 2.1.8.) sont donc, pour beaucoup, sans objet (ou serait-ce que les arguments ont porté ?).

Bien sûr, nos conclusions sont relatives à la situation particulière dans laquelle nous avons placé les chercheurs, c'est-à-dire une situation idéalisée. Pour déterminer si nous avons bien atteint ici un "noyau" de la situation du chercheur, il serait intéressant de vérifier si les résultats se confirment dans une situation écologiquement plus valide, par exemple en plaçant le chercheur face à un article, emprunté soit à la discipline du chercheur soit à une autre discipline pour contrôler l'effet de la connaissance du domaine. On pourrait ainsi fournir comme donnée un court article dans lequel ne figurerait comme résultat que la mention du seuil p, tout en conservant la même tâche de jugement sur une échelle.

* * *

Les différentes études présentées ici confirment l'existence et la prégnance des idées fausses que les psychologues entretiennent à l'égard des tests statistiques traditionnels. Mais l'expérience "psychologues et statisticiens" montre que sur certains points au moins, les statisticiens eux-mêmes commettent des abus d'interprétation. Si cette expérience met en lumière une certaine soumission des chercheurs en psychologie au test statistique, ceci est tempéré par les résultats de l'expérience sur les seuils observés qui révèle que la majorité des chercheurs ont tendance à porter un jugement nuancé en fonction du seuil observé et non à fonctionner en tout ou rien en opposant systématiquement significatif et non significatif.

Nos résultats conduisent à penser que le rôle de l'heuristique de "représentativité" de Kahneman et Tversky, proposée pour expliquer les représentations des sujets dans des situations d'inférence, est variable selon les situations. Si cette heuristique permet de rendre compte des résultats de la situation de prédiction dans l'expérience 1, d'autres hypothèses/explications (le seuil contient déjà l'effet de N; le jugement est apparenté à une décision en tout ou rien) apparaissent plus vraisemblables dans la situation de jugement des seuils observés de l'expérience 2, au moins pour une grande majorité des sujets.

CONCLUSION

CONCLUSION

Les théories des tests statistiques majoritairement (sinon exclusivement) appliquées en psychologie sont les théories fréquentistes de Fisher et de Neyman et Pearson qui divergent à bien des points de vue et qui ont toutes deux été fortement critiquées. Si, du point de vue méthodologique, un certain nombre de ces critiques peuvent apparaître comme “techniques” dans la mesure où un aménagement des procédures pourrait suffire à y répondre, il reste que les critiques fondamentales, différentes pour chacune des deux théories, sont aussi incontournables que l'est, à l'heure actuelle, l'utilisation des tests dans les publications scientifiques. Ces critiques fondamentales concernent l'adéquation de ces théories à la démarche expérimentale. D'une part, la théorie des tests de Fisher met bien l'accent sur l'expérience (unique) effectivement réalisée et tient compte de la richesse des résultats au travers du seuil observé, mais elle dénature le processus de corroboration de l'hypothèse d'intérêt en réduisant celui-ci au rejet d'une autre hypothèse qui ne joue qu'un rôle d'“homme de paille”. D'autre part, la théorie de Neyman et Pearson teste bien, en principe, l'hypothèse d'intérêt, mais elle se disqualifie en se fondant sur une vision purement décisionnelle de l'inférence qui fonctionne en “tout ou rien” et assimile tous les résultats significatifs (resp. tous les résultats non significatifs) entre eux.

Aucune de ce ces deux théories des tests n’apparaissant appropriée à l’usage, il n’est pas surprenant que la pratique actuellement dominante en psychologie consiste à en utiliser un amalgame. Mais un tel amalgame ne peut pas être justifié d’un point de vue formel; en outre, en perdant la cohérence interne spécifique à chacune des deux théories, il conduit inévitablement à des abus d’utilisation. C’est donc à une véritable distorsion de l'outil statistique qu’on assiste dans la pratique. Cette distorsion est encouragée par les manuels de statistique appliquée qui constituent, directement ou indirectement, les références de base pour les chercheurs en psychologie. En effet, dans les manuels que nous avons examinés, les théories des tests statistiques sont rarement rapportées fidèlement. De plus, la plupart de ces manuels comportent des abus d'interprétation explicites, particulièrement dans les exemples d'application, et fournissent donc un mode d’emploi trompeur. Il en résulte bon nombre de mésusages du tests de signification, les chercheurs tendant à lui faire dire plus que ne le permet la théorie.

Les mésusages ressortent différemment selon les situations, comme le montrent notre réanalyse d'articles publiés et nos travaux expérimentaux.

Dans le cadre d'une publication, les chercheurs se comportent essentiellement en “tout ou rien” en ce qu'ils s'en tiennent à la dichotomie test significatif/non significatif et commentent peu les résultats des tests, d'où des conclusions d'ordre décisionnel. Quand le résultat est significatif, il est simplement conclu à l'existence de l'effet et la question de son intensité est éludée. Quand le résultat est non significatif, soit rien n'est dit de plus et on peut admettre que cela équivaut au constat d'ignorance qui convient dans une telle situation, mais n’est évidemment pas satisfaisant, soit il est conclu de façon positive, mais abusive, à la démonstration de l'absence d'effet.

L'expérience “psychologues et statisticiens” montre que les psychologues tendent, plus que des statisticiens, à surestimer le poids d’un résultat significatif et à sous-estimer le rôle des effets observés dans l’interprétation des résultats. Ainsi, quand le résultat est significatif, la confusion entre significativité statistique et significativité substantielle, qui était éludée dans les articles publiés, apparaît. Il se confirme qu’un résultat non significatif est très souvent abusivement interprété comme la démonstration d'une absence d'effet, ceci même par les statisticiens. La situation d'un effet observé faible et non significatif, courante dans les publications, est particulièrement favorable à cet abus. Même dans la situation conflictuelle d’un effet observé fort mais non significatif (relativement rare dans les publications), un tiers des psychologues et des statisticiens concluent à l’absence d’effet. Le fait que les statisticiens ne soient pas à l'abri de ces mésusages implique que le remède aux abus d'interprétation ne peut être recherché simplement par une formation plus poussée des psychologues aux tests statistiques.

Dans le cadre de l'expérience sur les seuils observés, où les contraintes liées à la situation de publication sont en principe absentes, si un petit nombre de chercheurs juge le seuil observé en tout ou rien par rapport au critère de 5%, en revanche, le jugement de la plupart d'entre eux est nuancé, dépendant d'une manière continue de la valeur du seuil sans que la valeur de 5% joue un rôle particulier. Et dans un nombre important de cas, le jugement de confiance s'apparente à un véritable jugement psychophysique.

La théorie de Neyman et Pearson, par son caractère binaire, favorise l'absence de commentaire : un résultat particulier ne sert que d'intermédiaire pour atteindre une simple décision en faveur d'une des deux hypothèses en jeu. Sous cet aspect, et uniquement sous celui-ci, elle fournit une description du comportement des chercheurs dans la situation de publication. D'un autre côté, la théorie de Fisher correspond mieux, par le

rôle nuancé qu'y joue le seuil observé, à ce qui a été observé dans l'expérience sur les seuils. En outre, il se dégage des questionnaires de type “question de cours” sur l'interprétation des tests que les chercheurs, accordant une probabilité à une hypothèse, sont également des “bayésiens sauvages”. Il semblerait donc que l'amalgame des théories des tests statistiques qui se manifeste dans les publications ne renvoie pas à une représentation fixée une fois pour toutes mais que les chercheurs se rapprochent davantage de l'une des théories statistiques des tests ou de l'autre selon le contexte[34].

À l’examen, les abus d'interprétation ne sont pas de simples erreurs de compréhension des théories statistiques, mais ils vont à l'évidence dans le sens d'une correction intuitive des défauts méthodologiques des tests; certains auteurs, comme Bakan (1966), Phillips (1973) ou M.-P. Lecoutre *et al.* (1997) parlent d'ailleurs d'ajustements ou de biais adaptatifs. Ainsi, quand la quantité 1 - *p* est interprétée comme la probabilité que l'hypothèse alternative, généralement identifiée à l'hypothèse de recherche, soit vraie (*cf.* 2.2.1.), c'est une distorsion qui permet de répondre à une question pertinente du point de vue méthodologique (“Dans quelle mesure l'hypothèse de recherche est-elle soutenue par les données ?”). Quand cette même quantité est prise pour la probabilité d'obtenir une réplique significative (*cf.* 2.2.2.), c'est bien sûr parce que la question de la réplicabilité d'un résultat est cruciale dans les sciences expérimentales. La confusion des significativités statistique et substantielle (*cf.* 2.2.3.) est une manière de réintroduire la grandeur des effets. L'acceptation de l'hypothèse nulle (*cf.* 2.2.4.) montre que le but du chercheur est bien d'adopter, au moins provisoirement, une hypothèse. Enfin, l'omission de la condition dans l'énoncé de la probabilité de commettre une erreur de type I ou II (*cf.* 2.2.5.) n'est qu'une autre formulation de la réponse à la question de la vérification de l’hypothèse par les données (donner la probabilité de se tromper, sachant qu'on a décidé en faveur de telle hypothèse, c'est donner la probabilité que cette hypothèse soit fausse).

Quelle que soit l’approche adoptée, l’inadaptation méthodologique des tests de signification est donc manifeste. Néanmoins, leur pratique par les chercheurs en psychologie est socialement adaptée à une norme qui, surmontant toutes les critiques, fait l’objet d’un véritable consensus social impliquant tous les acteurs de la recherche scientifique : prescripteurs, auteurs et censeurs. Ce consensus est révélé par la stéréotypie, mais aussi la pauvreté, de la présentation des résultats statistiques dans les revues expérimentales; il est sans nul doute l’une des raisons essentielles de la longévité de l’usage des tests.

Les conduites des chercheurs peuvent ainsi se décrire en termes d'assimilation et d'accommodation permettant d'atteindre un état d'équilibre représenté par un consensus social :

Assimilation. Les chercheurs transforment les théories statistiques pour les adapter à leurs besoins qui concernent le degré de corroboration des hypothèses, la grandeur des effets; d'où l'apparition des abus précédents.

Accommodation. Les chercheurs transforment leurs questions pour s'adapter à la théorie des tests. De “Jusqu'à quel point l'hypothèse est-elle confortée par les données ?” ou “Quelle est l'intensité de l'effet ?” ils sont passés à “L'effet est-il significatif ?”. La formulation même des hypothèses et la planification des expériences ont été affectées; il n'est pas exceptionnel que l'hypothèse psychologique d'une étude se réduise à la mise en évidence de l'effet significatif de tel ou tel facteur. Cette transformation peut être superficielle (par exemple chez certains des chercheurs interrogés par M.-P. Lecoutre, 1983, qui se déclaraient insatisfaits de la situation actuelle et ne pratiquaient les tests que pour satisfaire aux normes de publication) ou plus profonde (nous avons souvent rencontré des chercheurs qui, en dehors de tout contexte de publication, exposaient spontanément leurs questions uniquement en termes de “Est-ce significatif ?”).

Une voie de recherche serait d'analyser plus finement les attitudes des chercheurs, de chercher à distinguer ce qui relève d'un conformisme social et ce qui est intégré par le chercheur. Par exemple, on peut se demander quand et comment s'opère le passage de l'attitude nuancée observée lors de l'expérience sur les seuils à l'attitude plus tranchée manifeste dans les publications. Bien entendu, un travail préliminaire est de confirmer les résultats obtenus à partir des 20 articles retenus ici en réanalysant les 55 articles restants (il sera particulièrement intéressant de voir s'il ne se manifeste toujours pas d'abus en cas de résultat significatif).

Perspectives

Il ne peut y avoir de psychologie “quantitative” si l'on se limite à un catalogue d'effets “significatifs” ou “non significatifs” et s'il ne se constitue pas un corpus de résultats eux-mêmes quantitatifs. Si, en ce qui concerne l'utilisation des tests statistiques en psychologie, la situation a peu évolué jusqu'à présent, il va

[34] Nous considérons ici que la population dont sont issus les chercheurs ayant participé à nos expériences est comparable à la population correspondant aux chercheurs dont nous avons réanalysé les articles.

certainement en aller tout autrement dans un avenir très proche. En effet, l'*American Psychological Association* (APA) a récemment lancé une “force d'intervention” (*Task Force*) sur l'inférence statistique dont l'objet était de reconsidérer la pratique du test de signification. À la suite d'une première réunion (14 et 15 décembre 1996), cette “force d'intervention” a déjà émis un certain nombre d'avis.

Le test de signification n'est pas irrémédiablement condamné et reste une procédure acceptable, mais son statut se trouve considérablement modifié par les recommandations suivantes :

- l’ouverture à d'autres méthodes d'analyse des résultats, entre autres les méthodes bayésiennes;
- le rapport systématique de la grandeur des effets et des intervalles de confiance correspondants;
- la reconnaissance des études exploratoires bien formulées et bien conduites avec des traitements quantitatifs appropriés des résultats (en réaction contre les abus de la démarche hypothético-déductive);
- l’application du principe de parcimonie au choix des plans d'expérience et des analyses.

Il semble probable que ces premières considérations seront confirmées et entérinées par l'APA[35] et ces recommandations ont donc toutes chances de fournir dans un avenir proche les nouvelles normes de publications. On peut d'ores et déjà en tirer un certain nombre de conséquences prévisibles.

La grandeur de l'effet observé devrait être de plus en plus fréquemment rapportée, notamment parce qu'elle est facile à comprendre et à mettre en œuvre (certains logiciels la fournissent déjà).

Il est également aisé de prédire le succès de l'intervalle de confiance, d'autant que cette procédure est déjà familière aux psychologues. La recommandation de recourir à des plans d'expérience et des analyses simples devrait faciliter son usage, en complément, ou même en remplacement du test de signification. L'utilisation de l'intervalle de confiance marquera un progrès méthodologique indéniable par rapport à la situation actuelle. En particulier, la variabilité des mesures étant exprimée de façon manifeste dans la largeur de l'intervalle, cela devrait réduire l'abus consistant à inférer une absence d'effet sur la seule base d'un résultat non significatif.

En contrepartie, on devrait assister à une prolifération des abus d'interprétation de l'intervalle de confiance en termes probabilistes, c'est-à-dire à une interprétation bayésienne de cet intervalle; à moins que les promoteurs de l’intervalle de confiance parviennent à en faire comprendre et admettre l’interprétation correcte, ce qui nous paraît bien peu vraisemblable. Seul le recours explicite aux méthodes bayésiennes pourrait permettre de faire cesser la duplicité consistant à recourir à la fois à un cadre de justification fréquentiste et à un cadre d’interprétation bayésien.

Enfin, une conjecture plus spéculative est que les méthodes bayésiennes finiront par s’imposer.

Dans la mesure où la demande des psychologues en de nouvelles méthodes statistiques se développera, le rôle des statisticiens devrait s'accroître. Ce rôle concerne à la fois l'enseignement de la statistique, puisqu'il faudra alors enseigner ces méthodes, et la conception de logiciels informatiques qui sont maintenant incontournables. Or l’examen des publications actuelles, tant en statistique mathématique qu’en statistique appliquée, révèle dans tous les domaines une évolution considérable et apparemment inéluctable vers la théorie bayésienne. On peut donc s’attendre à ce que le rôle accru des statisticiens favorise en retour l’utilisation des méthodes bayésiennes en psychologie.

Déjà en 1974, les importants développements de la théorie bayésienne amenaient Winkler à discuter une liste de raisons pour expliquer son peu d'utilisation dans l’analyse statistique en psychologie expérimentale :

> "The points discussed in that section as contributing factors to the theory-practice gap in statistical analysis in experimental psychology include philosophical conviction, tradition, statistical training, lack of 'availability', computational difficulties, reporting difficulties, and perceived resistance by journal editors." (Winkler, 1974, pp. 137-138)

Si nous reprenons ces raisons, il apparaît de nos jours que les conditions pour que les chercheurs en psychologie aient recours à des méthodes bayésiennes pour analyser leurs données sont maintenant objectivement réunies.

(1) Conviction philosophiques. Les fréquents abus d'interprétation des tests de signification montrent que bien souvent les chercheurs émettent des énoncés fiduciaires comme M. Jourdain faisait de la prose : ils

[35] Il est à remarquer que la nécessité de réaliser des répliques des expériences est curieusement absente des recommandations de cette “force d'intervention”. C'est pourtant une suggestion qui a souvent été formulée par les critiques des tests et qui fait plutôt l'objet d'un consensus en sa faveur (en biologie la règle est de réaliser au moins deux ou trois fois l'expérience, pour autant que possible, avant de soumettre un résultat à publication). De même, on notera que l'étude de la puissance n'est pas mentionnée bien que son plus ardent défenseur en psychologie, Cohen, fasse partie des cosignataires de ce rapport (faut-il y voir un désaveu ?).

n'hésitent pas à donner un énoncé probabiliste concernant l'hypothèse (et donc la valeur du paramètre) qui les intéresse. C'est particulièrement évident dans le cas de l'intervalle de confiance où la probabilité est appliquée à l'intervalle particulier calculé à partir des données recueillies. Aussi, il semble bien qu'un grand nombre de psychologues interprètent, au moins implicitement, la probabilité comme une mesure de l'incertitude et n'auraient donc rien à objecter, sur le plan philosophique, aux méthodes bayésiennes. Par ailleurs, Winkler a insisté sur le fait que ce ne sont pas tant les convictions philosophiques des chercheurs à l'égard des probabilités qui sont importantes que l'utilisation cohérente et appropriée de l'approche choisie.

(2) Tradition. La possibilité d'utiliser les méthodes bayésiennes en complément des tests de signification n'implique pas une rupture avec les pratiques traditionnelles.

(3) Enseignement des statistiques. L'enseignement des statistiques à l'attention des psychologues évolue peu (ou pas), et bien que le théorème de Bayes soit souvent mentionné dans le cadre d'une présentation de la théorie des probabilités, la théorie statistique bayésienne n'est en général pas abordée. Mais cela pourrait changer si la pression des psychologues ou des statisticiens s'avérait suffisante.

(4) Manque de disponibilité et difficultés de calcul. Les méthodes bayésiennes impliquent des calculs complexes, nécessitant le recours à des ordinateurs individuels. Des ouvrages et des programmes qui couvrent un grand nombre des situations rencontrées en pratique sont déjà disponibles (*cf.* 3.5.5.), et on peut s'attendre dans un avenir proche à un développement de manuels et de logiciels généraux incluant ces méthodes.

(5) Difficultés de communication. Certes, la nécessité de spécifier la distribution initiale utilisée allonge les comptes-rendus de résultats. Mais, en réalité, Winkler faisait essentiellement allusion à la nécessité de décrire suffisamment dans les articles la méthode utilisée, peu familière aux lecteurs comme aux "referees". Ceci reste bien sûr d'actualité, mais ne peut être un obstacle de fond à l'utilisation de ces méthodes. Évidemment, cet argument perdra d'autant de sa valeur que les méthodes bayésiennes seront plus couramment utilisées.

(6) Résistance (réelle ou supposée) des éditeurs de journaux scientifiques et tradition. Assurément il s'agit là d'un point capital étant donné le rôle actuel de norme que le test de signification joue encore dans les pratiques de publication. Mais, comme nous venons de le voir, avec les nouvelles recommandations de l'APA, la situation pourrait évoluer rapidement et les éditeurs ne devraient pas s'opposer systématiquement à l'usage de méthodes bayésiennes.

La méthode fiducio-bayésienne, qui porte l'attention sur la grandeur de l'effet et qui s'avère déjà un garde-fou utile contre certains abus, est, par son statut privilégié d'objectivité, une méthode de choix pour la constitution d'un corpus de résultats quantitatifs, prenant explicitement en compte la question de l'importance des effets, au niveau moyen comme au niveau individuel. L'approche bayésienne fournit en outre toutes les procédures nécessaires pour les décisions liées à la conduite des expériences, d'une part en ce qui concerne le choix préalable des tailles d'échantillon pour pouvoir conclure avec assez de précision et avec une bonne garantie, et d'autre part en ce qui concerne l'interruption ou au contraire la prolongation d'une expérience en cours.

Il sera intéressant, quelque temps après la publication des recommandations finales de l'APA, de reprendre la réanalyse systématique d'articles publiés (toujours avec des outils fiducio-bayésiens) pour juger de l'évolution des pratiques.

Pour expliquer finalement que les choses n'aient pas évolué jusqu'ici, et qu'il faudra sans doute encore du temps pour que de nouvelles méthodes soient effectivement adoptées, nous pouvons évoquer la loi de Hofstadter (mentionnée, pour un autre propos, par Oakes, 1986, p. viii) :

"Cela prend toujours plus de temps que vous ne pensez, même si vous prenez en compte la loi de Hofstadter."

BIBLIOGRAPHIE

BIBLIOGRAPHIE

ABDI, H. (1987) - *Introduction au traitement statistique des données expérimentales*. Grenoble: Presses Universitaires de Grenoble.

ALBERT, J. (1995) - Teaching inference about proportions using Bayes and discrete models. *Journal of Statistics Education*, **3**. (Par courrier électronique.)

ANDERSON, J. R. (1991) - The adaptative nature of human categorization. *Psychological Review*, **98**, 409-429.

BACHER, F. (1998) - L'utilisation des modèles dans l'analyse des structures de covariance. *L'Année Psychologique*, à paraître.

BAKAN, D. (1966) - The test of significance in psychological research. *Psychological Bulletin*, **66**, 423-437.

BARNARD, G. A. (1947) - The meaning of a significance level. *Biometrika*, **34**, 179-182.

BARTKO, J. J. (1991) - Proving the null hypothesis. *American Psychologist*, **46**, 1089.

BASSOK, M., WU, L.-L. & OLSETH, K. L. (1995) - Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition*, **23**, 354-367.

BEAUCHAMP, K. L. & MAY, R. B. (1964) - Replication report: Interpretation of levels of significance by psychological researchers. *Psychological Reports*, **14**, 272.

BERGER, J. O. (1985) - *Statistical Decision Theory and Bayesian Analysis*. (2ème édition) New York: Springer-Verlag.

BERGER, J. O. & SELLKE, T. (1987) - Testing a point null hypothesis: The irreconciability of *P* values and evidence. (Avec discussion). *Journal of the American Statistical Association*, **82**, 112-122.

BERGER, R. L. & HSU, J. C. (1996) - Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, **11**, 283-319.

BERKSON, J. (1938) - Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association*, **33**, 526-542.

BERKSON, J. (1941) - Comments on Dr. Madow's "Note on tests of departure from normality" with some remarks concerning tests of significance. *Journal of the American Statistical Association*, **46**, 539-541.

BERKSON, J. (1942) - Tests of significance considered as evidence. *Journal of the American Statistical Association*, **37**, 325-335.

BERNARD, J.-M. (1986) - Méthodes d'inférence bayésienne sur des fréquences. *Informatique et Sciences Humaines*, **68-69**, 89-133.

BERNARD, J.-M. (1991) - Inférence bayésienne et prédictive sur les fréquences. *In* H. ROUANET, M.-P. LECOUTRE, M.-C. BERT, B. LECOUTRE, J.-M. BERNARD & B. LEROUX (Eds.), *L'inférence statistique dans la démarche du chercheur*. Berne: Peter Lang.

BERNARD, J.-M. (1996) - Bayesian interpretation of frequentist procedures for a Bernoulli process. *The American Statistician*, **50**, 7-13.

BERNARD, J.-M. (1997) - Bayesian analysis of tree-structured categorized data. *Revue Internationale de Systémique*, sous presse.

BERNARD, J.-M., BLANCHETEAU, M. & ROUANET, H. (1985) - Le comportement prédateur chez un forficule, Eurobellia Moesta (Géné). *Biology of Behaviour*, **10**, 1-22.

BERNARDO, J. M. & SMITH, A. F. M. (1994) - *Bayesian Theory*. Chichester: Wiley.

BERRY, G. (1986) - Statistical significance and confidence intervals (Editorial). *Medical Journal of Australia*, **144**, 618-619.

BINDER, A. (1963) - Further considerations on testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, **70**, 107-115.

BOLLES, R. (1962) - The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports*, **11**, 639-645.

BONDY, W. A. (1969) - A test of an experimental hypothesis of negligible difference between means. *The American Statistician*, **23**, 28-30.

BRAITMAN, L. E. (1988) - Confidence intervals extract clinically useful information from data (Editorial). *Annals of Internal Medicine*, **108**, 296-298.

BRAITMAN, L. E. (1991) - Confidence intervals assess both clinical significance and statistical significance (Editorial). *Annals of Internal Medicine*, **114**, 515-517.

BURKE, C. J. (1953) - A brief note on one-tailed tests. *Psychological Bulletin*, **50**, 384-387.

BURKE, C. J. (1954) - Further remarks on one-tailed tests. *Psychological Bulletin*, **51**, 587-590.

CAMILLERI, S. F. (1962) - Theory, probability, and induction in social research. *American Sociological Review*, **27**, 170-178.

CARVER, R. P. (1978) - The case against statistical significance testing. *Harvard Educational Review*, **48**, 378-399.

CASELLA, G. & BERGER, L. (1987) - Reconciling Bayesian and frequentist evidence in the one-sided testing problem. (Avec discussion). *Journal of the American Statistical Association*, **82**, 106-111.

CHOW, S. L. (1988) - Significance test or effect size? *Psychological Bulletin*, **103**, 105-110.

CHOW, S. L. (1989) - Significance tests and deduction: Reply to Folger (1989). *Psychological Bulletin*, **106**, 161-165.

CHOW, S. L. (1991) - Some reservation about power analysis. *American Psychologist*, **46**, 1088.

CHOW, S. L. (1996) - *Statistical Significance: Rationale, Validity and Utility*. London: Sage.

CIANCIA, F., MAITTE, M., HONORÉ, J., LECOUTRE, B. & COQUERY, J.-M. (1988) - Orientation of attention and sensory gating: An evoked potential and reaction time study in Cat. *Experimental Neurology*, **100**, 274-287.

CLARK-CARTER, D. (1997) - The account taken of statistical power in research published in the *British Journal of Psychology*. *British Journal of Psychology*, **88**, 71-83.

CLÉMENT, E. & RICHARD, J.-F. (1997) - Knowledge of domain effects in problem representation: The case of Tower of Hanoi isomorphs. *Thinking and Reasoning*, **3**, 133-157.

COHEN, J. (1962) - The statistical power of abnormal-social psychological research: A Review. *Journal of Abnormal and Social Psychology*, **65**, 145-153.

COHEN, J. (1969) - *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

COHEN, J. (1990) - Things I have learned (so far). *American Psychologist*, **45**, 1304-1312.

COHEN, J. (1992) - A power primer. *Psychological Bulletin*, **112**, 155-159.

COHEN, J. (1994) - The earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.

CORNFIELD, J. (1966) - Sequential trials, sequential analysis and the likelihood principle. *The American Statistician*, **20**, 18-23.

CORROYER, D. & ROUANET, H. (1994) - Sur l'importance des effets et des indicateurs dans l'analyse statistique des données. *L'Année Psychologique*, **94**, 607-624.

COWLES, M. & DAVIS, C. (1982) - On the origins of the .05 level of statistical significance. *American Psychologist*, **37**, 553-558.

COX, D. R. (1958) - Some problems connected with statistical inference. *Annals of Mathematical Statistics*, **29**, 357-372.

COX, D. R. (1977) - The role of significance tests. *Scandinavian Journal of Statistics*, **4**, 49-70.

CRAIG, J. R., EISON, C. L. & METZE, L. P. (1976) - Significance tests and their interpretation: An example utilizing published research and ω^2. *Bulletin of the Psychonomic Society*, **7**, 280-282.

CROW, E. L. (1991) - Response to Rosenthal's comment "How are we doing in soft psychology?". *American Psychologist*, **46**, 1083.

DAR, R. (1987) - Another look at Meehl, Lakatos, and the scientic pratices of psychologists. *American Psychologist*, **42**, 145-151.

DURAND, J.-L. (1997) - Analyse de l'ouvrage de N. Guéguen, *Manuel de statistique pour psychologues*, Paris: Dunod, 1997. *L'Année Psychologique*, **97**, sous presse.

DWYER, J. H. (1974) - Analysis of variance and the magnitude of effects: A general approach. *Psychological Bulletin*, **81**, 731-737.

EDWARDS, W. (1965) - Tactical note on the relation between scientific and statistical hypotheses. *Psychological Bulletin*, **63**, 400-402.

EDWARDS, W., LINDMAN, H. & SAVAGE, L. J. (1963) - Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

EFRON, B. (1986) - Why isn't everyone a Bayesian. *The American Statistician*, **40**, 1-5.

EVANS, S. J. W., MILLS, P. & DAWSON, J. (1988) - The end of the p value? (Editorial). *British Heart Journal*, **60**, 177-180.

FALISSARD, B. & LANDAIS, P. (1995) - Les statistiques en médecine: et s'il était temps de prendre un peu de distance? *Médecine thérapeutique*, **1**, 775-781.

FALK, R. & GREENBAUM, C. W. (1995) - Significance tests die hard. The amazing persistence of a probabilistic misconception. *Theory & Psychology*, **5**, 75-98.

FAVERGE, J.-M. (1975) - *Méthodes statistiques en psychologie appliquée*. (3 vol.) (7ème édition) Paris: Presses Universitaires de France. (1ère édition: 1950)

FISHER, R. A. (1948) - Conclusions fiduciaires. *Annales de l'Institut Henri Poincaré*, **10**, 191-213.

FISHER, R. A. (1955) - Statistical methods and scientific induction. *Journal of the Royal Statistical Society (B)*, **17**, 69-78.

FISHER, R. A. (1959) - Mathematical probability in the natural sciences. *Technometrics*, **1**, 21-29.

FISHER, R. A. (1962) - Some examples of Bayes's method of the experimental determination of probabilities *a priori*. *Journal of the Royal Statistical Society (B)*, **24**, 118-124.

FISHER, R. A. (1990a) - *Statistical Methods for Research Workers*. (Réimp. 14ème édition de 1970) *In* J. H. BENNETT (Ed.), *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford: Oxford University Press. (1ère édition: 1925, London: Oliver and Boyd.)

FISHER, R. A. (1990b) - *The Design of Experiments*. London: (Réimp. 8ème édition de 1966) *In* J. H. BENNETT (Ed.), *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford: Oxford University Press. (1ère édition: 1935, London: Oliver and Boyd.)

FISHER, R. A. (1990c) - *Statistical Methods and Scientific Inference*. (Réimp. 3ème édition de 1973) *In* J. H. BENNETT (Ed.), *Statistical Methods, Experimental Design, and Scientific Inference*. Oxford: Oxford University Press. (1ère édition: 1956, London: Oliver and Boyd.)

FOLGER, R. (1989) - Significance tests and the duplicity of binary decisions. *Psychological Bulletin*, **106**, 155-160.

FOWLER, R. L. (1985) - Testing for substantive significance in applied research by specifying nonzero effect null hypotheses. *Journal of Applied Psychology*, **70**, 215-218.

FREEMAN, P. R. (1993) - The role of *p*-values in analysing trial results. *Statistics in Medicine*, **12**, 1443-1452.

FREIMAN, J. A., CHALMERS, T. C., SMITH, H. & KUEBLER, R. R. (1978) - The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *New England Journal of Medicine*, **299**, 690-694.

FRICK, R. W. (1995a) - A problem with confidence intervals. *American Psychologist*, **50**, 1002-1003.

FRICK, R. W. (1995b) - Accepting the null hypothesis. *Memory & Cognition*, **23**, 132-138.

FRICK, R. W. (1996) - The appropriate use of null hypothesis testing. *Psychological Methods*, **1**, 379-390.

GIGERENZER, G. (1993) - The superego, the ego, and the id in statistical reasoning. *In* G. KEREN & C. LEWIS (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale, N-J: Lawrence Erlbaum associates.

GLASS, G. V. (1976) - Primary, secondary, and meta-analysis of research. *Educational Researcher*, **5**, 3-8.

GOLD, D. (1969) - Statistical tests and substantive significance. *The American Sociologist*, **4**, 42-46.

GOOD, I. J. (1984) - An error by Neyman noticed by Dickey (C209). *Journal of Statistical Computation and Simulation*, **20**, 159-160.

GOODMAN, S. N. (1992) - A comment on replication, *P*-values and evidence. *Statistics in Medicine*, **11**, 875-879.

GOODMAN, S. N. & BERLIN, J. A. (1994) - The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, **121**, 200-206.

GRANT, D. A. (1962) - Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychological Review*, **69**, 54-61.

GREENWALD, A. G, GONZALEZ, R., HARRIS, R. J. & GUTHRIE, D. (1996) - Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology*, **33**, 175-183.

GROUIN, J.-M. & LECOUTRE, B. (1996) - Probabilités prédictives: un outil pour la planification des expériences. *Revue de Statistique Appliquée*, **XLIV**, 21-35.

GUTTMAN, L. (1979) - Cyril Burt and the careless star worshippers. *L'Echo des Messaches*, **9**, 7-8.

GUTTMAN, L. (1983) - What is not what in statistics? *The Statistician*, **26**, 81-107.

HAASE, R. F., WAECHTER, D. M. & SOLOMON, G. S. (1982) - How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology*, **29**, 58-65.

HAGOOD, M. J. (1941) - The notion of a hypothetical universe. *In* M. J. HAGOOD, *Statistics for Sociologists*. New York: Reynal & Hitchcock, 612-616. (Reproduit dans Morrison & Henkel, 1970).

HARCUM, E. R. (1990) - Methodological versus empirical literature: Two views on casual acceptance of the null hypothesis. *American Psychologist*, **45**, 404-405.

HARRIS, M. J. (1991) - Significance tests are not enough. *Theory & Psychology*, **1**, 375-382.

HAYS, W. L. (1963) - *Statistics for Psychologists*. New York: Holt, Rinehart & Winston.

HAYS, W. L. (1971) - *Statistics for the Social Sciences*. (2ème édition) New York: Holt, Rinehart & Winston.

HICK, W. E. (1952) - A note on one-tailed and two-tailed tests. *Psychological Review*, **59**, 316-318.

HOC, J.-M. (1983) - *L'analyse planifiée des données en psychologie*. Paris: Presses Universitaires de France.

HODGES, J. L., & LEHMANN, E. L. (1954) - Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society (B)*, **16**, 261-268.

HOGBEN, L. (1957) - *Statistical Theory: The Relationship of Probability, Credibility, and Error. An examination of the Contemporary crisis in Statistical Theory from a Behaviourist Viewpoint*. London: Allen and Unwin.

JEFFREYS, H. (1961) - *Theory of Probability*. (3ème édition) Oxford: Clarendon. (1ère édition: 1939.)

JOHNSTONE, D. (1988) - Comments on Oakes on the foundations of statistical inference in the social and behavioral sciences: The market for statistical significance. *Psychological Reports*, **63**, 319-331.

JONES, L. V. (1952) - Tests of hypotheses: One-sided *vs*. two-sided alternatives. *Psychological Bulletin*, **49**, 43-46.

JONES, L. V. (1954) - A rejoinder on one-tailed tests. *Psychological Bulletin*, **51**, 585-586.

KADANE, J. B. (1995) - Prime time for Bayes. *Controlled Clinical Trials*, **16**, 313-318.

KAHNEMAN, D. & TVERSKY, A. (1972) - Subjective probability: A judgement of representativeness. *Cognitive Psychology*, **3**, 430-454.

KIRK, R. E. (1982) - *Experimental Design*. (2ème édition.) Belmont: Brook-Cole.

KISH, L. (1959) - Some statistical problems in research design. *American Sociological Review*, **24**, 328-338.

KOEHLER, J. J. (1996) - The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, **19**, 1-53.

LAFORGE, R. (1967) - Confidence intervals or test of significance in scientific research? *Psychological Bulletin*, **68**, 446-447.

LECOUTRE, B. (1984a) - *L'analyse bayésienne des comparaisons*. Lille: Presses Universitaires de Lille.

LECOUTRE, B. (1984b) - Réinterprétation fiducio-bayésienne du test F de l'analyse de la variance. *L'Année Psychologique*, **84**, 77-83.

LECOUTRE, B. (1985) - How to derive Bayes-fiducial conclusions from usual significance tests. *Cahiers de Psychologie Cognitive*, **5**, 553-563.

LECOUTRE, B. (1991) - Du test de signification à l'inférence fiducio-bayésienne. *In* H. ROUANET, M.-P. LECOUTRE, M.-C. BERT, B. LECOUTRE, J.-M. BERNARD & B. LEROUX (Eds.), *L'inférence statistique dans la démarche du chercheur*. Berne: Peter Lang.

LECOUTRE, B. (1994) - Inférence statistique et raisonnements inductifs. *Psychologie Française*, **39**, 141-151.

LECOUTRE, B. (1996) - *Traitement statistique des données expérimentales: des pratiques traditionnelles aux pratiques bayésiennes*. Saint-Mandé: CISIA. (Programmes Windows fournis avec l'ouvrage.)

LECOUTRE, B. (1997) - Et si vous étiez un bayésien « qui s'ignore » ? *La Revue de Modulad*, **18**, 81-87.

LECOUTRE, B. & CHARRON, C. (1997) -Bayesian analysis of the degrees of implications between binary attributes. Soumis pour publication.

LECOUTRE, B., DERZKO, G. & GROUIN, J.-M. (1995) - Bayesian predictive approach for inference about proportions. *Statistics in Medicine*, **14**, 1057-1063.

LECOUTRE, B., GUIGUES, J.-L. & POITEVINEAU, J. (1992) - Distribution of quadratic forms of multivariate Student variables (AS 278). *Applied Statistics*, **41**, 617-627.

LECOUTRE, B. & LECOUTRE, M.-P. (1979) - A propos d'une expérience d'apprentissage perceptif incident: quelques aspects de la démarche d'analyse des données et méthodes fiduciaires. *Psychologie Française*, **24**, 269-276.

LECOUTRE, B., MABIKA, B. & DERZKO, G. (1997a) - Comparison of two Weibull distributions with unequal parameter shapes. Soumis pour publication.

LECOUTRE, B. & POITEVINEAU, J. (1992) - PAC, Programme d'Analyse des Comparaisons. *Guide d'utilisation et manuel de référence*, Saint-Mandé: CISIA.

LECOUTRE, B., POITEVINEAU, J., DERZKO, G. & GROUIN, J.-M. (1997b) - Désirabilité et faisabilité des méthodes bayésiennes en analyse de variance: application à des plans d'expérience complexes utilisés dans les essais cliniques. *Biométrie et Méthodes bayésiennes*, sous presse.

LECOUTRE, M.-P. (1982) - Comportements des chercheurs dans des situations conflictuelles d'analyse de données expérimentales. *Psychologie Française*, **27**, 1-8.

LECOUTRE, M.-P. (1983) - La démarche du chercheur en psychologie dans des situations d'analyse statistique de données expérimentales. *Journal de Psychologie Normale et Pathologique*, **3**, 275-295.

LECOUTRE, M.-P. (1988) - Attitude des sociologues à l'égard de la notion de généralisation des résultats: questionnaire sur l'inférence statistique. *Bulletin de Méthodologie Sociologique*, **20**, 5-11.

LECOUTRE, M.-P. (1991) - Et … le point de vue des chercheurs? Quelques éléments de réflexion. *In* H. ROUANET, M.-P. LECOUTRE, M.-C. BERT, B. LECOUTRE, J.-M. BERNARD & B. LEROUX (Eds.), *L'inférence statistique dans la démarche du chercheur*. Berne: Peter Lang.

LECOUTRE, M.-P. (1992) - Cognitive models and problem spaces in «purely random» situations. *Educational Studies in Mathematics*, **23**, 557-568.

LECOUTRE, M.-P., DURAND, J.-L. & CORDIER, J. (1990) - A study of two biases in probabilistic judgments: Representativeness and equiprobability. *In* J.-P. Caverni, J.-M. Fabre, M. Gonzalez (Eds.), *Cognitive Biases*, North Holland: Amsterdam.

LECOUTRE, M.-P., POITEVINEAU, J. & LECOUTRE, B. (1997) - Uses, abuses, and misuses of significance tests in the scientific community: Won't the Bayesian choice be unavoidable? *Soumis pour publication*.

LECOUTRE, M.-P. & ROUANET, H. (1993) - Predictive judgements in situations of statistical analysis. *Organizational Behavior and Human Decision Processes*, **54**, 45-56.

LEE, P. M. (1989) - *Bayesian Statistics: An Introduction*. London: Oxford University Press.

LEHMANN, E. L. (1959) - *Testing Statistical Hypotheses*. New York: Wiley.

LEHMANN, E. L. (1993) - The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, **88**, 1242-1249.

LÉPINE, D. & ROUANET, H. (1975) - Introduction aux méthodes fiduciaires: inférence sur un contraste entre moyennes. *Cahiers de Psychologie*, **18**, 193-218.

LEVIN, J. R. (1967) - Misinterpreting the significance of "explained variation". *American Psychologist*, **22**, 675-676.

LINDLEY, D. V. (1957) - A statistical paradox. *Biometrika*, **44**, 187-192.

LINDLEY, D. V. & PHILLIPS, L. D. (1976) - Inference for a Bernoulli process (a Bayesian view). *The American Statistician*, **30**, 112-119.

LOFTUS, G. R. (1993) - Editorial comment. *Memory & Cognition*, **21**, 1-3.

LOFTUS, G. R. & MASSON, M. E. J. (1994) - Using confidence intervals with within-subject designs. *Psychonomic Bulletin & Review*, **1**, 476-490.

LUTZ, W. & NIMMO, I. A. (1977) - The inadequacy of statistical significance (Editorial). *European Journal of Clinical Investigation*, **7**, 77-78.

LYKKEN, D. (1968) - Statistical significance in psychological research. *Psychological Bulletin*, **70**, 151-159.

McGRAW, K. O. (1991) - Problems with the BESD: A comment on Rosenthal's "How are we doing in soft psychology?". *American Psychologist*, **46**, 1084-1086.

McNEMAR, Q. (1960) - At random: Sense and nonsense. *American Psychologist*, **15**, 295-300.

MARKS, M. R. (1951) - Two kinds of experiment distinguished in terms of statistical operations. *Psychological Review*, **58**, 179-184.

MARKS, M. R. (1953) - One and two-tailed tests. *Psychological Review*, **60**, 207-208.

MEEHL, P. E. (1967) - Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, **34**, 103-115.

MEEHL, P. E. (1978) - Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, **46**, 806-834.

MELTON, A. W. (1962) - Editorial. *Journal of Experimental Psychology*, **64**, 553-557.

MIALARET, G. (1996) - *Statistiques*. Paris: Presses Universitaires de France.

MOHER, D., DULBERG, C. S. & WELLS, G. A. (1994) - Statistical power, sample size, and their reporting in randomized controlled trials. *Journal of the American Medical Association*, **272**, 122-124.

MORRISON, D. E. & HENKEL, R. E. (1969) - Significance tests reconsidered. *The American Sociologist*, **4**, 131-140.

MORRISON, D. E. & HENKEL, R. E. (Eds.) (1970) - *The Significance Test Controversy*. London: Butterwoths.

NATRELLA, M. G. (1960) - The Relation between confidence intervals and tests of significance. *The American Statistician*, **14**, 20-22.

NELSON, N., ROSENTHAL, R. & ROSNOW, R. L. (1986) - Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist*, **41**, 1299-1301.

NEYMAN, J. (1935) - Sur la vérification des hypothèses statistiques composées. *Bulletin de la Société Mathématique de France*, **63**, 246-266.

NEYMAN, J. (1950) - *First Course in Probability and Statistics*. New York: Holt.

NEYMAN, J. (1952) - *Lectures and Conferences on Mathematical Statistics and Probability*. (2ème édition) Washington: Graduate School U.S. Department of Agriculture.

NEYMAN, J. & PEARSON, E. S. (1928a) - On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. *Biometrika*, **20A**, 175-240.

NEYMAN, J. & PEARSON, E. S. (1928b) - On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika*, **20A**, 263-294.

NEYMAN, J. & PEARSON, E. S. (1933a) - On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *Series A*, **231**, 289-337.

NEYMAN, J. & PEARSON, E. S. (1933b) - The testing of statistical hypotheses in relation to probabilities *a priori*. *Proceedings of the Cambridge Philosophical Society*, **29**, 492-510.

NISBETT, R. & ROSS, L. (1981) - *Human Inference: Strategies and Shortcomings of Social Judgments*. Englewood Cliffs (N. J.): Prentice Hall.

NUNNALLY, J. C. (1960) - The place of statistics in psychology. *Educational and Psychological Measurement*, **20**, 641-650.

NUNNALLY, J. C. (1975) - *Introduction to Statistics for Psychology and Education*. New York: McGraw-Hill.

OAKES, M. (1986) - *Statistical Inference: A Commentary for the Social and Behavioural Sciences*. New York: Wiley.

O'BRIEN, T. C. & SHAPIRO, B. J. (1968) - Statistical significance - What? *Mathematics Teacher*, **61**, 673-676.

O'GRADY, K. E. (1982) - Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, **92**, 766-777.

PEARSON, E. S. (1955) - Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society (B)*, **17**, 204-207.

PHILLIPS, L. D. (1973) - *Bayesian Statistics for Social Scientists*. London: Nelson.

PIAGET, J. (1975) - *Les mécanismes perceptifs*. (2ème édition) Paris: Presses Universitaires de France. (1ère édition: 1961.)

PIÉRON, H. (1963) - La psychophysique. *In* P. FRAISSE & J. PIAGET (Eds.), *Traité de psychologie expérimentale. II Sensation et motricité*. Paris: Presses Universitaires de France.

PITZ, G. F. (1972) - Quick and dirty data analysis with a Bayesian flavor. *Manuscrit non publié*. Southern Illinois Univ.

POITEVINEAU, J. & BERNARD, J.-M. (1986) - La série des programmes IBF. *Informatique et Sciences Humaines*, **68-69**, 135-137.

POITEVINEAU, J. & LECOUTRE, B. (1997) - Some statistical misconceptions in Chow's Statistical significance. *Behavioral and Brain Sciences*, sous presse.

POPPER, K. R. (1973) - *La logique de la découverte scientifique*. Paris: Payot. (Réimp. 1984; 1ère édition originale, en Allemand: 1939; 1ère traduction en Anglais: 1959.)

PRATT, J. W. (1965) - Bayesian interpretation of standard inference statements (avec discussion). *Journal of the Royal Statistical Society (B)*, **27**, 169-203.

PRENTICE, D. A. & MILLER, D. T. (1992) - When small effects are impressive. *Psychological Bulletin*, **112**, 160-164.

REUCHLIN, M. (1962) - *Les méthodes quantitatives en psychologie*. Paris: Presses Universitaires de France.

REUCHLIN, M. (1976) - *Précis de statistique*. Paris: Presses Universitaires de France.

REUCHLIN, M. (1992) - *Introduction à la recherche en psychologie*. Paris: Nathan.

RICHARDSON, J. T. E. (1996) - Measures of effect size. *Behavior Research Methods, Instruments, & Computers*, **28**, 12-22.

ROBERT, Ch. (1992) - *L'analyse statistique bayésienne*. Paris: Économica.

ROBERT, Cl. (1995) - *L'empereur et la girafe*. Paris: Diderot éditeur, Arts et Sciences.

ROBERT, M. (1994) - Stratégies méthodologiques. *In* M. RICHELLE, J. REQUIN & M. ROBERT (Eds.), *Traité de psychologie expérimentale*. Paris: Presses Universitaires de France.

ROGERS, J. L., HOWARD, K. I. & VESSEY, J. T. (1993) - Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, **113**, 553-565.

ROSENTHAL, R. (1979) - The “file-drawer” and tolerance for null results. *Psychological Bulletin*, **86**, 638-641.

ROSENTHAL, R. & GAITO, J. (1963) - The interpretation of levels of significance by psychological researchers. *Journal of Psychology*, **55**, 33-38.

ROSENTHAL, R. & GAITO, J. (1964) - Further evidence for the cliff effect in the interpretation of levels of significance. *Psychological Reports*, **15**, 570.

ROSENTHAL, R. & RUBIN, D. B. (1979) - Comparing significance levels of independent studies. *Psychological Bulletin*, **86**, 1165-1168.

ROSENTHAL, R. & RUBIN, D. B. (1982) - A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, **74**, 166-169.

ROTHMAN, K. J. (1978) - A show of confidence (Editorial). *New England Journal of Medicine*, **299**, 1362-1363.

ROUANET, H. (1967) - *Les modèles stochastiques d'apprentissage*. Paris: Gauthier-Villars.

ROUANET, H. (1986) - Modèles en tout genre et pratiques statisticiennes. *Comportements*, **4**, 113-124.

ROUANET, H. (1991a) - Les pratiques statisticiennes en question. *In* H. ROUANET, M.-P. LECOUTRE, M.-C. BERT, B. LECOUTRE, J.-M. BERNARD & B. LEROUX (Eds.), *L'inférence statistique dans la démarche du chercheur*. Berne: Peter Lang.

ROUANET, H. (1991b) - Les tests statistiques revisités. *In* H. ROUANET, M.-P. LECOUTRE, M.-C. BERT, B. LECOUTRE, J.-M. BERNARD & B. LEROUX (Eds.), *L'inférence statistique dans la démarche du chercheur*. Berne: Peter Lang.

ROUANET, H. (1996) - Bayesian methods for assessing importance of effects. *Psychological Bulletin*, **119**, 149-158.

ROUANET, H. (1997) - Statistical practice at stake. *In* H. ROUANET, M.-P. LECOUTRE, M.-C. BERT, B. LECOUTRE, J.-M. BERNARD & B. LEROUX (Eds.), *Statistical Inference in the Strategy of the Researcher*. Berne: Peter Lang. Sous presse.

ROUANET, H., LEROUX, B. & BERT, M.-C. (1987) - *Statistique en sciences humaines: procédures naturelles*. Paris: Dunod.

ROUANET, H. & LECOUTRE, B. (1983) - Specific inference in ANOVA: From significance tests to Bayesian procedures. *British Journal of Mathematical and Statistical Psychology*, **36**, 252-268.

ROUANET, H., LÉPINE, D. & HOLENDER, D. (1978) - Model acceptability and the use of Bayes-fiducial methods for validating models. *In* J. REQUIN (Ed.), *Attention and Performance VII*. Hillsdale: Lawrence Erlbaum Associates.

ROUANET, H., LÉPINE, D. & PELNARD-CONSIDÈRE, J. (1976) - Bayes-fiducial procedures as practical substitutes for misplaced significance testing: An application to educational data. *In* D. N. M. DE GRUIJTER & L. J. T. VAN DER KAMP (Eds.), *Advances in Psychological and Educational Measurement*. New York: Wiley.

ROYALL, R. M. (1986) - The effect of sample size on the meaning of significance tests. *The American Statistician*, **40**, 313-315.

ROZEBOOM, W. W. (1960) - The fallacy of the null hypothesis significance test. *Psychological Bulletin*, **57**, 416-428.

SALSBURG, D. (1994) - Intent to treat: The *reductio ad absurdum* that became gospel. *Pharmacoepidemiology and Drug Safety*, **3**, 329-335.

SALSBURG, D. S. (1985) - The religion of statistics as practiced in medical journals. *The American Statistician*, **39**, 220-223.

SAVAGE, R. J. (1957) - Nonparametric statistics. *Journal of the American Statistical Association*, **52**, 332-333.

SCHEFFÉ, H. (1959) - *The Analysis of Variance*. New York: Wiley.

SCHERVISH, M. J. (1995) - *Theory of Statistics*. New York: Springer Verlag.

SCHMIDT, F. L. (1996) - Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, **1**, 115-129.

SCHNEIDER, B., PARKER, S., OSTROSKY, D., STEIN, D. & KANOW, G. (1974) - A scale for the psychological magnitude of number. *Perception & Psychophysics*, **16**, 43-46.

SCHUIRMANN, D. J. (1987) - A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, **15**, 657-680.

SCHWARTZ, D. (1984) - Statistique et vérité. *Journal de la Société Statistique de Paris*, **125**, n°2, 74-83.

SELDMEIER, P. & GIGERENZER, G. (1989) - Do studies of statistical power have an effect on the power of studies. *Psychological Bulletin*, **105**, 309-316.

SELVIN, H. C. (1957) - A critique of tests of significance in survey research. *American Sociological Review*, **22**, 519-527.

SERLIN, R. C. & LAPSLEY, D. K. (1993) - Rational appraisal of psychological research and the good-enough principle. *In* G. KEREN & C. LEWIS (Eds.), *A Handbook for Data Analysis in the Behavioral Sciences: Methodological Issues*. Hillsdale: Lawrence Erlbaum Associates.

SHAFER, G. (1976) - *A Mathematical Theory of Evidence*. Princeton: Princeton University Press.

SIEGEL, S. (1956) - *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

SKIPPER, J. K., GUENTHER, A. L. & NASS, G. (1967) - The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, **1**, 16-18.

SMITH, M. L. & GLASS, G. V. (1977) - Meta-analysis of psychotherapy outcome studies. *American Psychologist*, **32**, 752-760.

SPIEGELHALTER, D.J., FREEDMAN, L.S. & PARMAR, M.K.B. (1994) - Bayesian approaches to randomized trials (with discussion). *Journal of the Royal Statistical Society (A)*, **157**, 357-416.

STERLING, T. D. (1959) - Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association*, **54**, 30-34.

STERLING, T. D., ROSENBAUM, W. L. & WEINKAM, J. J. (1995) - Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, **95**, 108-112.

STEVENS, S. S. (1962) - The surprising simplicity of sensory metrics. *American Psychologist*, **17**, 29-39.

STEVENS, S. S. (1968) - Measurement, statistics, and the schemapiric view. *Science*, **161**, 849-856.

STRAHAN, R. F. (1991) - Remarks on the Binomial Effect Size Display. *American Psychologist*, **46**, 1083-1084.

STUDENT (1908) - The probable error of a mean. *Biometrika*, **6**, 1-25.

TULLOCK, G. (1959) - Publication decisions and tests of significance: A comment. *Journal of the American Statistical Association*, **54**, 593.

TYLER, R. (1931) - What is statistical significance? *Educational Research Bulletin*, **10**, 118-142.

TVERSKY, A. & KAHNEMAN, D. (1971) - Belief in the law of small numbers. *Psychological Bulletin*, **76**, 105-110.

VAUGHAN, G. M. & CORBALLIS, M. C. (1969) - Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. *Psychological Bulletin*, **72**, 204-213.

WELLEK, S. & MICHAELIS, J. (1991) - Elements of significance testing with equivalence problems. *Methods of Information in Medicine*, **30**, 194-198.

WILSON, K. V. (1961) - Subjectivist statistics for the current crisis. *Contemporary Psychology*, **6**, 229-231.

WILSON, W., MILLER, H. L. & LOWER, J. S. (1967) - Much ado about the null hypothesis. *Psychological Bulletin*, **67**, 188-196.

WINCH, R. F. & CAMPBELL, D. T. (1969) - Proof? No. Evidence? Yes. The significance of tests of significance. *The American Sociologist*, **4**, 140-143.

WINER, B. J. (1971) - *Statistical Principles in Experimental Designs*. (2ème édition) New York: McGraw-Hill. (1ère édition: 1962.)

WINKLER, R. L. (1974) - Statistical analysis: Theory versus practice. *In* C.-A. S. STAEL VON HOLSTEIN (Ed.), *The Concept of Probability in Psychological Experiments*. Dordrecht: Reidel.

YATES, F. (1964) - Sir Ronald Fisher and the design of experiments. *Biometrics*, **20**, 307-321.

ANNEXES

QUELQUES RAPPELS

La distribution d'échantillonnage et la vraisemblance

Étant donné une taille n d'échantillon et une statistique S_n à laquelle on s'intéresse (la moyenne, par exemple), la distribution des valeurs que peut prendre cette statistique pour un échantillon de taille n, étant donné également les paramètres dont cette distribution dépend, est appelée la distribution d'échantillonnage de la statistique S_n .

Une hypothèse statistique H détermine la distribution d'échantillonnage de la statistique. En particulier $Pr(D|H)$ désignera la probabilité d'observer un événement D étant donné les valeurs des paramètres spécifiées par H.

Pour fixer les idées, considérons une variable dichotomique comme la latéralité (droitier/gaucher, en excluant la possibilité d'être ambidextre). L'hypothèse H_1 fixe que la proportion p de droitiers dans une certaine population est de 2/3. Admettons que dans un échantillon au hasard de taille 10 on obtienne 8 droitiers. Sous l'hypothèse d'indépendance des observations, la variable aléatoire D, proportion observée de droitiers, est distribuée selon une loi binomiale de paramètres 10 et 2/3 et il est facile de calculer $Pr(D = 8/10 \mid H_1: p = 2/3)$. Ici, D est la variable et p est une constante et, en utilisant la convention d'écrire des majuscules pour les variables et des minuscules pour les constantes, on écrira $Pr(D|p)$.

Mais, partant du fait qu'on a observé 8/10, on peut très bien calculer $Pr(D = 8/10 \mid H_2: p = 3/4)$ ou $Pr(D = 8/10 \mid H_3: p = 1/2)$, *etc.*, c'est-à-dire faire varier les valeurs de p en maintenant constante la valeur d, soit $Pr(d|P)$ qui est alors considéré comme une fonction du paramètre p. Cette fonction est appelée la fonction de vraisemblance (*likelihood function*) du paramètre et $Pr(D = 8/10 \mid H_1: p = 2/3)$ est ici la vraisemblance de l'hypothèse $p = 2/3$. Cette notion de vraisemblance a été introduite par Fisher en 1925. Il ne s'agit pas d'une probabilité, la somme des $Pr(d|P)$ n'étant pas égale à un, mais dans les cas simples où cette somme est finie, il est toutefois facile de renormaliser ces quantités.

Ainsi le même élément peut être envisagé selon deux perspectives différentes. Pour bien marquer qu'on s'intéresse à la vraisemblance, on utilise parfois la notation v ou l (pour *likelihood*) au lieu de Pr, et pour insister sur le changement de point de vue on inverse les termes par rapport au signe de conditionnement; une même valeur pourra donc être désignée par $Pr(D|H)$ ou $v(H|D)$.

La vraisemblance est utilisée par les statisticiens bayésiens, et aussi par les statisticiens traditionnels; par exemple, dans le cadre de la théorie de Neyman et Pearson la région critique est déterminée par l'étude du rapport des vraisemblances de l'hypothèse testée et de l'hypothèse alternative.

Le théorème de Bayes

Le théorème de Bayes, ou plus simplement formule de Bayes, permet le passage de la probabilité *a priori* d'un événement (pour ce qui nous occupe il s'agit d'une hypothèse sur la valeur d'un paramètre) à sa probabilité *a posteriori*, compte tenu de la réalisation d'un autre événement (le recueil de données dans notre cas). À propos du théorème de Bayes on parle parfois de *probabilités inverses*, puisqu'il y a renversement des termes par rapport au signe de conditionnement; on évoque aussi la question de la *probabilité des causes*.

- Dans le cas d'un ensemble discret d'événements H_i (i = 1,…), exhaustifs et exclusifs, la formule de Bayes donne la probabilité que l'événement H_i se réalise sachant qu'on a observé l'événement D, appartenant à un autre ensemble d'événements, et s'énonce :

  $$Pr(H_i|D) = Pr(D|H_i)\, Pr(H_i) / \sum\{Pr(D|H_i)\, Pr(H_i)\}$$

  $Pr(D|H_i)$ est la probabilité d'échantillonnage de D sous H_i; c'est encore la vraisemblance de H_i.
  $Pr(H_i)$ est appelée la probabilité *a priori* de H_i, ou encore sa probabilité initiale.
  $Pr(H_i|D)$ est appelée la probabilité *a posteriori* de H_i, ou sa probabilité finale.
  Le dénominateur, qui n'est autre que $Pr(D)$, est appelé la probabilité prédictive de D.

- Dans le cas où l'on s'intéresse non plus à des événements discrets mais à des variables continues (disons θ et v) qui admettent une densité de probabilité (notée f), la formule donne alors la densité de probabilité de θ conditionnellement à v et s'écrit :

$$f(\theta|v) = f(v|\theta)\, f(\theta) \,/\, \int_{\Theta} f(v|\theta)\, f(\theta)\, d\theta$$

Du point de vue mathématique, le théorème de Bayes est la conséquence directe de la définition des probabilités conditionnelles et ne pose donc aucun problème. C'est l'application de ce théorème qui est parfois contestée, le problème se ramenant en fait à celui de l'interprétation de la probabilité. Ainsi, Neyman (1950, 1952) considère que son application à des jugements sur des hypothèses n'est pertinente que lorsqu'on peut donner une interprétation purement fréquentiste de la probabilité d'une hypothèse.

Puisque le théorème de Bayes ne fait pas problème, il est évident qu'une analyse statistique dite bayésienne, par opposition à une analyse ne mettant en jeu que la conception fréquentiste (traditionnelle), ne peut se définir par le simple fait qu'elle ferait appel au théorème de Bayes. Pour reprendre les termes de Kadane (1995), on dira qu'une analyse est bayésienne quand les paramètres, qui n'ont pas été observés, ou l'idée qu'on s'en fait sont traités comme aléatoires, alors que les données, qui ont été observées, sont traitées comme fixées.

RÉANALYSE FIDUCIO-BAYÉSIENNE D'ARTICLES PUBLIÉS

Tableaux et Figures des résultats

Tableaux et Figures B1 - B4

Tableaux/Figures B1 et B2 : distribution des effets calibrés observés *E* pour les analyses principales (B1) et secondaires (B2).
Tableaux/Figures B3 et B4 : distribution des coefficients de corrélation observés *r* pour les analyses principales (B3) et secondaires (B4).

Tableaux B5 - B10

Tableaux B5 et B6 : Croisement du résultat du test de signification et des conclusions descriptives, selon le type d'analyses (principale/secondaire), pour les effets calibrés (tableau B5) et pour les corrélations (tableau B6).
Tableau B7 et B8 : Croisement du résultat du test de signification et des inférences, selon le type d'analyses (principale/secondaire), pour les effets calibrés (tableau B7) et pour les corrélations (tableau B8).
Tableau B9 et B10 : Croisement des conclusions descriptives et des inférences, selon le type d'analyses (principale/secondaire), pour les effets calibrés (tableau B9) et pour les corrélations (tableau B10).

Effet calibré E

Analyses principales (N = 224)

Effet calibré	*n*	%
[0.0-0.2]	17	7.6
]0.2-0.4]	48	21.4
]0.4-0.6]	41	18.3
]0.6-0.8]	35	15.6
]0.8-1.0]	28	12.5
]1.0-1.2]	30	13.4
]1.2-1.4]	10	4.5
]1.4-1.6]	4	1.8
]1.6-1.8]	5	2.2
]1.8-2.0]	2	0.9
]2.0-5.2]	4	1.8

Tableau B1

Distribution des effets calibrés E

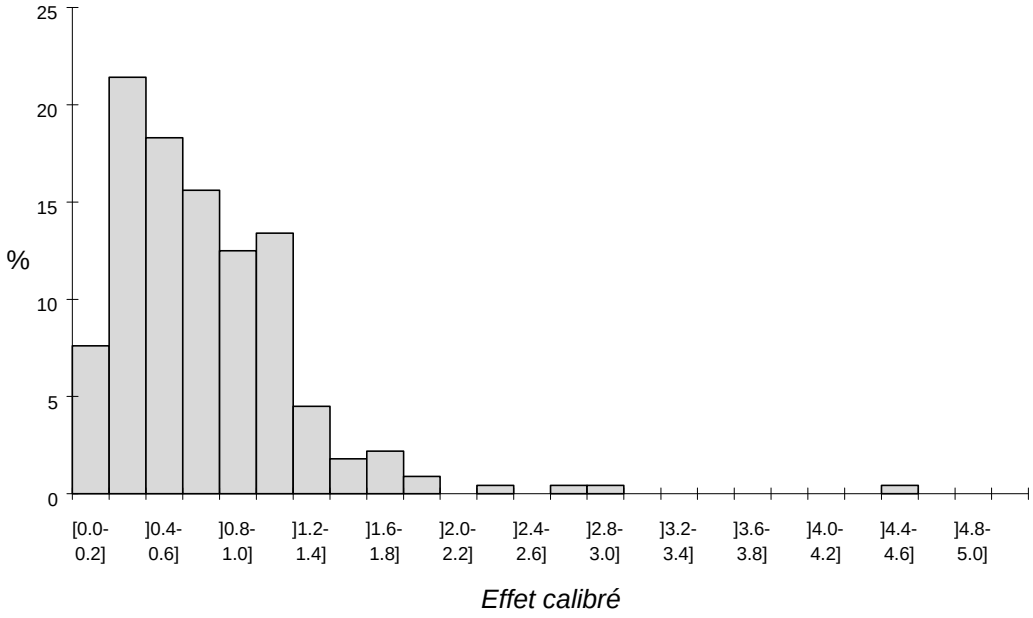


Figure B1

Histogramme des effets calibrés E

Moyenne pondérée	=	0.715
Médiane "pondérée"	=	0.619
Moyenne équipondérée	=	0.834
Médiane "équipondérée"	=	0.749

Effet calibré E

Analyses secondaires (N = 160)

Effet calibré	*n*	%
[0.0-0.2]	30	18.8
]0.2-0.4]	43	26.9
]0.4-0.6]	31	19.4
]0.6-0.8]	15	9.4
]0.8-1.0]	11	6.9
]1.0-1.2]	12	7.5
]1.2-1.4]	6	3.8
]1.4-1.6]	5	3.1
]1.6-1.8]	0	0.0
]1.8-2.0]	1	0.6
]2.0-5.2]	6	3.8

Tableau B2

Distribution des effets calibrés E

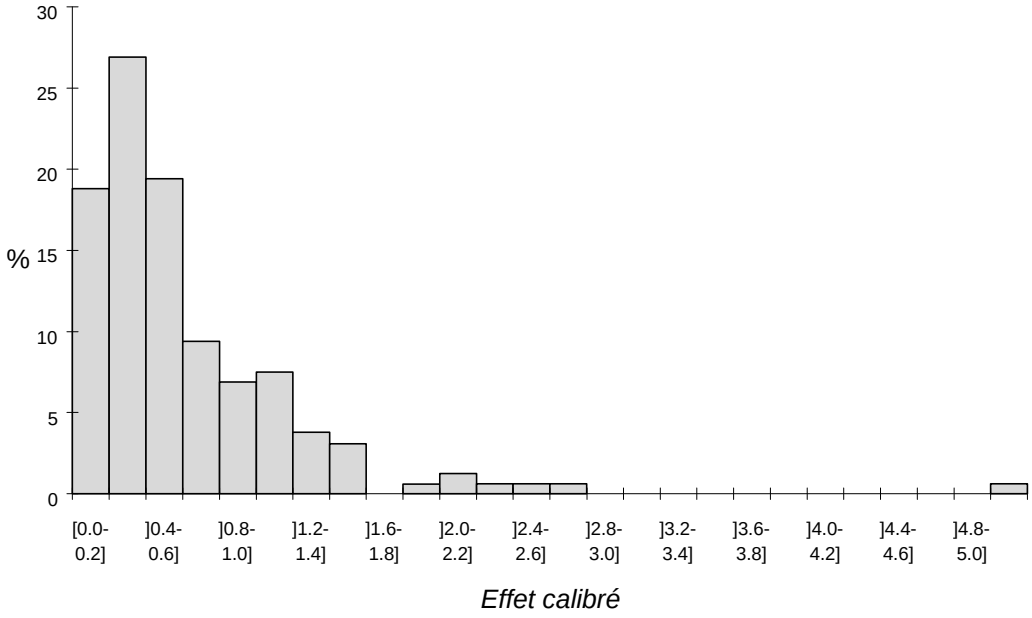


Figure B2

Histogramme des effets calibrés E

Moyenne pondérée	=	0.618
Médiane “pondérée”	=	0.437
Moyenne équipondérée	=	0.750
Médiane “équipondérée”	=	0.718

Coefficient de corrélation

Analyses principales (N = 175)

r	*n*	%
[0.0-0.1]	42	24.0
]0.1-0.2]	38	21.7
]0.2-0.3]	30	17.1
]0.3-0.4]	28	16.0
]0.4-0.5]	15	8.6
]0.5-0.6]	3	1.7
]0.6-0.7]	10	5.7
]0.7-0.8]	8	4.6
]0.8-0.9]	1	0.6
]0.9-1.0]	0	0.0

Tableau B3

Distribution des coefficients de corrélation r

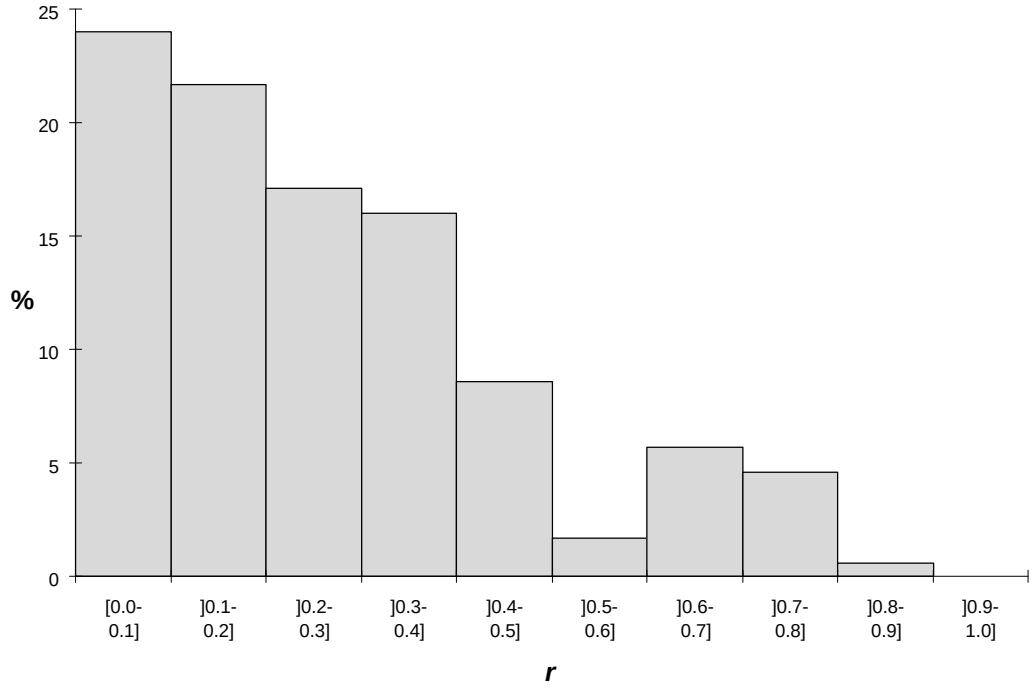


Figure B3

Histogramme des coefficients de corrélation r

Moyenne pondérée	=	0.268
Médiane “pondérée”	=	0.230
Moyenne équipondérée	=	0.240
Médiane “équipondérée”	=	0.196

Coefficient de corrélation

Analyses secondaires (N = 176)

r	*n*	%
[0.0-0.1]	42	23.9
]0.1-0.2]	38	21.6
]0.2-0.3]	31	17.6
]0.3-0.4]	27	15.3
]0.4-0.5]	10	5.7
]0.5-0.6]	13	7.4
]0.6-0.7]	6	3.4
]0.7-0.8]	2	1.1
]0.8-0.9]	6	3.4
]0.9-1.0]	1	0.6

Tableau B4

Distribution des coefficients de corrélation r

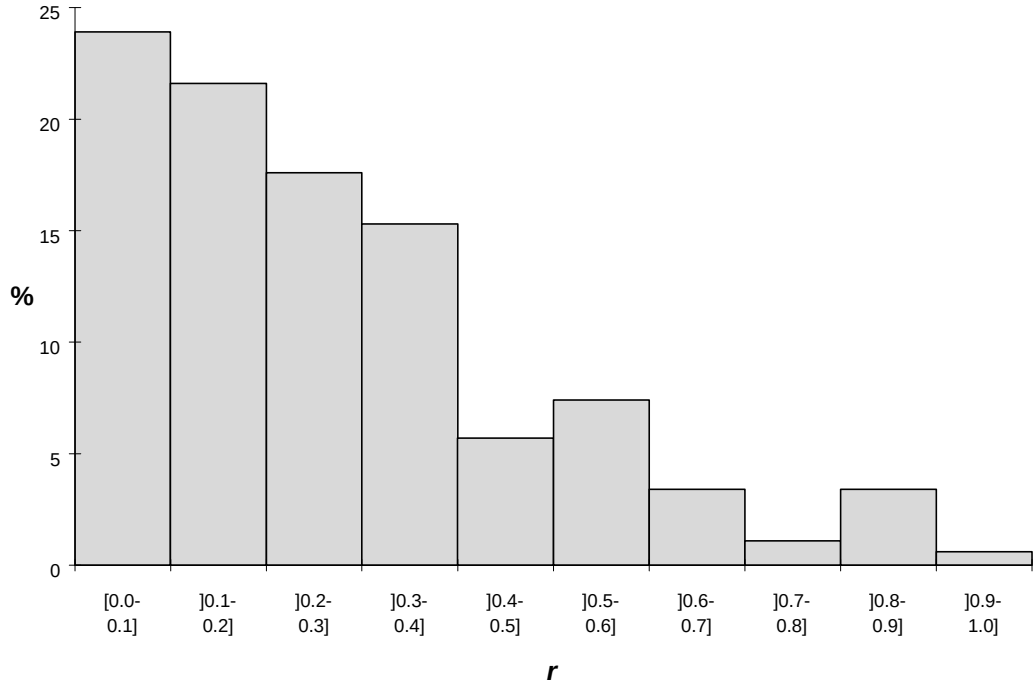


Figure B4

Histogramme des coefficients de corrélation r

Moyenne pondérée	=	0.278
Médiane “pondérée”	=	0.220
Moyenne équipondérée	=	0.325
Médiane “équipondérée”	=	0.297

Effet calibré

Analyses principales

	Conclusion sur l'effet observé E			
test	négligeable	moyen	notable	
Significatif	1 *(0.6%)*	87 *(50.9%)*	83 *(48.5%)*	*N*=171 *(*76.3%*)*
N.S.	16 *(30.2%)*	37 *(69.8%)*	0	*N*=53 *(23.7%)*
Total	17 *(7.6%)*	124 *(55.4%)*	83 *(37.1%)*	*N*=224

Analyses secondaires

	Conclusion sur l'effet observé E			
test	négligeable	moyen	notable	
Significatif	3 *(3.2%)*	48 *(51.6%)*	42 *(45.2%)*	*N*=93 *(58.1%)*
N.S.	27 *(40.3%)*	40 *(59.7%)*	0	*N*=67 *(41.9%)*
Total	30 *(18.8%)*	88 *(55.0%)*	42 *(26.3%)*	*N*=160

Tableau B5

Croisement du résultat du test de signification et des conclusions descriptives, pour les analyses principales (sous-tableau supérieur) et pour les analyses secondaires (sous-tableau inférieur)

Coefficient de corrélation

Analyses principales

	Conclusion sur le coefficient observé r			
test	négligeable	moyen	notable	
Significatif	0	58 *(72.5%)*	22 *(27.5%)*	*N*=80 *(45.7%)*
N.S.	42 *(44.2%)*	53 *(55.8%)*	0	*N*=95 *(54.3%)*
Total	42 *(24.0%)*	111 *(63.4%)*	22 *(12.6%)*	*N*=175

Analyses secondaires

	Conclusion sur le coefficient observé r			
test	négligeable	moyen	notable	
Significatif	4 *(4.0%)*	66 *(66.0%)*	30 *(30.0%)*	*N*=100 *(56.8%)*
N.S.	38 *(50.0%)*	38 *(50.0%)*	0	*N*=76 *(43.2%)*
Total	42 *(23.9%)*	104 *(59.1%)*	30 *(17.0%)*	*N*=176

Tableau B6

Croisement du résultat du test de signification et des conclusions descriptives, pour les analyses principales (sous-tableau supérieur) et pour les analyses secondaires (sous-tableau inférieur)

Effet calibré

Analyses principales

	Conclusion sur l'effet vrai ε					
test	négligeable	moyen	notable	non négl.	non notable	?
Significatif (*N*=171)	0	27 *(15.8%)*	30 *(17.5%)*	88 *(51.5%)*	23 *(13.5%)*	3 *(1.8%)*
N.S. (*N*=53)	3 *(5.7%)*				39 *(73.6%)*	11 *(20.8%)*
Total (*N*=224)	3 *(1.3%)*	27 *(12.1%)*	30 *(13.4%)*	88 *(39.3%)*	62 *(27.7%)*	14 *(6.2%)*

Analyses secondaires

	Conclusion sur l'effet vrai ε					
test	négligeable	moyen	notable	non négl.	non notable	?
Significatif (*N*=93)	1 *(1.1%)*	10 *(10.8%)*	19 *(20.4%)*	46 *(49.5%)*	17 *(18.3%)*	0
N.S. (*N*=67)	3 *(4.5%)*				54 *(80.6%)*	10 *(14.9%)*
Total (*N*=160)	4 *(2.5%)*	10 *(6.2%)*	19 *(11.9%)*	46 *(28.8%)*	71 *(44.4%)*	10 *(6.2%)*

Tableau B7

Croisement du résultat du test de signification et des inférences, pour les analyses principales (sous-tableau supérieur) et pour les analyses secondaires (sous-tableau inférieur).
En grisé, les impossibilités

Coefficient de corrélation

Analyses principales

	Conclusion sur le coefficient vrai ρ					
test	négligeable	moyen	notable	non négl.	non notable	?
Significatif (*N*=80)	0	33 *(41.2%)*	19 *(23.8%)*	23 *(28.8%)*	5 *(6.2%)*	0
N.S. (*N*=95)	0				93 *(97.9%)*	2 *(2.1%)*
Total (*N*=175)	0	33 *(18.9%)*	19 *(10.9%)*	23 *(13.1%)*	98 *(56.0%)*	2 *(1.1%)*

Analyses secondaires

	Conclusion sur le coefficient vrai ρ					
test	négligeable	moyen	notable	non négl.	non notable	?
Significatif (*N*=100)	0	39 *(39.0%)*	16 *(16.0%)*	26 *(26.0%)*	16 *(16.0%)*	3 *(3.0%)*
N.S. (*N*=76)	0				74 *(97.4%)*	2 *(2.6%)*
Total (*N*=176)	0	39 *(22.2%)*	16 *(9.1%)*	26 *(14.8%)*	90 *(51.1%)*	5 *(2.8%)*

Tableau B8

Croisement du résultat du test de signification et des inférences, pour les analyses principales (sous-tableau supérieur) et pour les analyses secondaires (sous-tableau inférieur).
En grisé, les impossibilités

Effet calibré

Analyses principales

Conclusion sur l'effet observé E \ *Conclusion sur l'effet vrai ε*	négligeable	moyen	notable	non négl.	non notable	?
négligeable (*N*=17)	3 *(17.6%)*				14 *(82.4%)*	0
moyen (*N*=124)		27 *(21.8%)*		35 *(28.2%)*	48 *(38.7%)*	14 *(11.3%)*
notable (*N*=83)			30 *(36.1%)*	53 *(63.9%)*		0
Total (*N*=224)	3 *(1.3%)*	27 *(12.1%)*	30 *(13.4%)*	88 *(39.3%)*	62 *(27.7%)*	14 *(6.2%)*

Analyses secondaires

Conclusion sur l'effet observé E \ *Conclusion sur l'effet vrai ε*	négligeable	moyen	notable	non négl.	non notable	?
négligeable (*N*=30)	4 *(13.3%)*				25 *(83.3%)*	1 *(3.3%)*
moyen (*N*=88)		10 *(11.4%)*		23 *(26.1%)*	46 *(52.3%)*	9 *(10.2%)*
notable (*N*=42)			19 *(45.2%)*	23 *(54.8%)*		0
Total (*N*=160)	4 *(2.5%)*	10 *(6.2%)*	19 *(11.9%)*	46 *(28.8%)*	71 *(44.4%)*	10 *(6.2%)*

Tableau B9

Croisement des conclusions descriptives et des inférences, pour les analyses principales (sous-tableau supérieur) et pour les analyses secondaires (sous-tableau inférieur). En grisé, les impossibilités

Coefficient de corrélation

Analyses principales

Conclusion sur le coefficient observé r \ *Conclusion sur le coefficient vrai ρ*	négligeable	moyen	notable	non négl.	non notable	?
négligeable (*N*=42)	0				42 *(100.0%)*	0
moyen (*N*=111)		33 *(29.7%)*		20 *(18.0%)*	56 *(50.5%)*	2 *(1.8%)*
notable (*N*=22)			19 *(86.4%)*	3 *(13.6%)*		0
Total (*N*=175)	0	33 *(18.9%)*	19 *(10.9%)*	23 *(13.1%)*	98 *(56.0%)*	2 *(1.1%)*

Analyses secondaires

Conclusion sur le coefficient observé r \ *Conclusion sur le coefficient vrai ρ*	négligeable	moyen	notable	non négl.	non notable	?
négligeable (*N*=42)	0				42 *(100.0%)*	0
moyen (*N*=104)		39 *(37.5%)*		12 *(11.5%)*	48 *(46.2%)*	5 *(4.8%)*
notable (*N*=30)			16 *(53.3%)*	14 *(46.7%)*		0
Total (*N*=176)	0	39 *(22.2%)*	16 *(9.1%)*	26 *(14.8%)*	90 *(51.1%)*	5 *(2.8%)*

Tableau B10

Croisement des conclusions descriptives et des inférences, pour les analyses principales (sous-tableau supérieur) et pour les analyses secondaires (sous-tableau inférieur). En grisé, les impossibilités

RÉSULTATS DE L'EXPÉRIENCE SUR LES SEUILS OBSERVÉS

Consigne "hypothèse alternative" : Figures C1 à C32

Figures C1 - C19 : courbes individuelles

Les sujets sont classés par type de courbe, dans l'ordre EXP, LIN, TOR, le sujet 11, atypique, étant reporté à la fin.
Pour tous les sujets, l'effet de la taille N de l'échantillon est indiqué. Il s'agit, d'une part, de la moyenne, calculée sur les douze seuils, des différences "courbe $N = 100$" - "courbe $N = 10$", et d'autre part et indiquée entre parenthèses, de la moyenne des valeurs absolues de ces différences (qui est une distance entre les deux courbes).
Pour les sujets de la classe EXP (Figures C1 à C10), sont indiqués, pour chacune des courbes, les r^2 relatifs à l'ajustement des points par rapport à la fonction puissance $y = a(1-p)^b$ et par rapport à la fonction exponentielle $y = \exp\{ap+b\}$.
Pour les sujets de la classe LIN (Figures C11 à C14), sont indiqués, pour chacune des courbes, les r^2 relatifs à l'ajustement des points par rapport à la droite $y = a(1-p)+b$, ainsi que les paramètres de la droite ajustée.

Figures C20 - C24 : courbes moyennes

Figure C20 : courbes moyennes pour l'ensemble des 18 sujets.
Figure C21 : courbes moyennes pour l'ensemble des 18 sujets des échelles recodées en six points.
Figure C22 : courbes moyennes pour les 10 sujets de la classe EXP.
Figure C23 : courbes moyennes pour les 4 sujets de la classe LIN.
Figure C24 : courbes moyennes pour les 4 sujets de la classe TOR.

L'effet de N est rapporté pour toutes ces courbes.

Les retests des sujets 1, 15 et 17 : Figures C25 à C32

Figures C25 - C32 : courbes individuelles des sujets 1 et 15 (test /retest)

À chaque sujet correspond quatre figures : d'abord les résultats de la première passation, puis ceux de la seconde passation, et enfin les comparaisons (test/retest) pour chaque condition d'effectif.
Pour chacune des conditions d'effectif, l'effet test/retest est indiqué. Il s'agit, d'une part, de la moyenne, sur les douze seuils, des différences "test - retest", et d'autre part et indiquée entre parenthèses, de la moyenne des valeurs absolues de ces différences (qui est une distance entre les deux courbes).

Commentaires

- Sujet 17.
Les courbes du retest de ce sujet ne sont pas présentées car elles sont absolument identiques à celles de sa première passation : quel que soit N, la confiance est totale pour $p \leq .05$, et nulle au delà.

- Sujet 1 (Figures C25 à C28).
Alors que l'on constate, pour la première passation, que la courbe $N = 100$ donne toujours lieu à plus de confiance que la courbe $N = 10$ (aux approximations près, pour deux points), il en va autrement pour la seconde passation. Pour cette dernière, aucune tendance à la supériorité de l'une ou l'autre des courbes ne se dégage, et nous pourrions décrire ce cas comme un cas d'égalité entre les deux courbes (bien que pour

$p = .10$ on observe un écart de confiance supérieur à 0.30). Cette différence entre les deux passations peut être relativisée si l'on remarque que l'effet "N=10 ≤ N=100" de la première passation est dû aux seuls trois points d'abscisse .07, .10 et .15 (voire aussi .20 si l'on admet l'hypothèse d'une confusion et que la "vraie" confiance pour ce seuil est d'environ 0.20), et qu'en dehors de ceux-ci les jugements sont comparables.
Globalement, la forme des courbes est semblable d'une passation à l'autre (de type exponentiel).
Pour $N = 10$, les courbes sont très voisines, sauf en $p = .20$. Mais, comme nous l'avons remarqué plus haut, ce point peut être expliqué par une confusion lors de la première passation. D'ailleurs la réplique va bien dans ce sens, la confiance passant alors à 0.165 pour ce point. L'écart maximum enregistré entre les deux passations est alors de 0.16 (pour $p = .50$). Nous considérons que la fidélité est très bonne, compte tenu de la nature de la tâche. Du fait des variations constatées entre les passations, nous aurions tendance à augmenter notre critère d'équivalence jusqu'à $\Delta = 0.15$ ou $\Delta = 0.20$. La courbe de la réplique nous apparaît alors monotone, malgré les quelques fluctuations au niveau des premiers points.
Pour $N = 100$, les courbes sont encore très voisines pour $p \leq .01$ et $p \geq .30$, mais divergent nettement dans l'intervalle, les valeurs pour la réplique étant abaissées. L'écart maximum est de 0.425 (en $p = .20$). Là aussi, nous admettrons que la courbe de la réplique est relativement monotone.

- Sujet 15 (Figures C29 à C32).
Les commentaires à formuler sur les résultats de ce sujet sont très proches de ceux formulés pour le sujet 1 : même forme des courbes, et même effet de N, mais encore plus marqué. Ici, l'abaissement de la confiance au cours de la réplique, pour $N = 100$, se constate sur pratiquement l'ensemble des seuils p (le maximum de baisse, 0.315, est atteint pour $p = .10$). Il est à noter que le sujet a spontanément déclaré, au cours de la réplique, que la valeur de N n'intervenait pas dans ses jugements, ce que l'on constate effectivement (si l'on excepte $p = .01$, pour lequel l'écart dépasse 0.30), alors que la courbe $N = 100$ est systématiquement supérieure à celle $N = 10$ dans la première passation. Interrogé sur ce point, le sujet n'a pu nous fournir d'explication de son changement d'attitude.

Consigne "hypothèse nulle" : Figures C33 à C40 (courbes individuelles)

Figures C1 - C19 : courbes individuelles

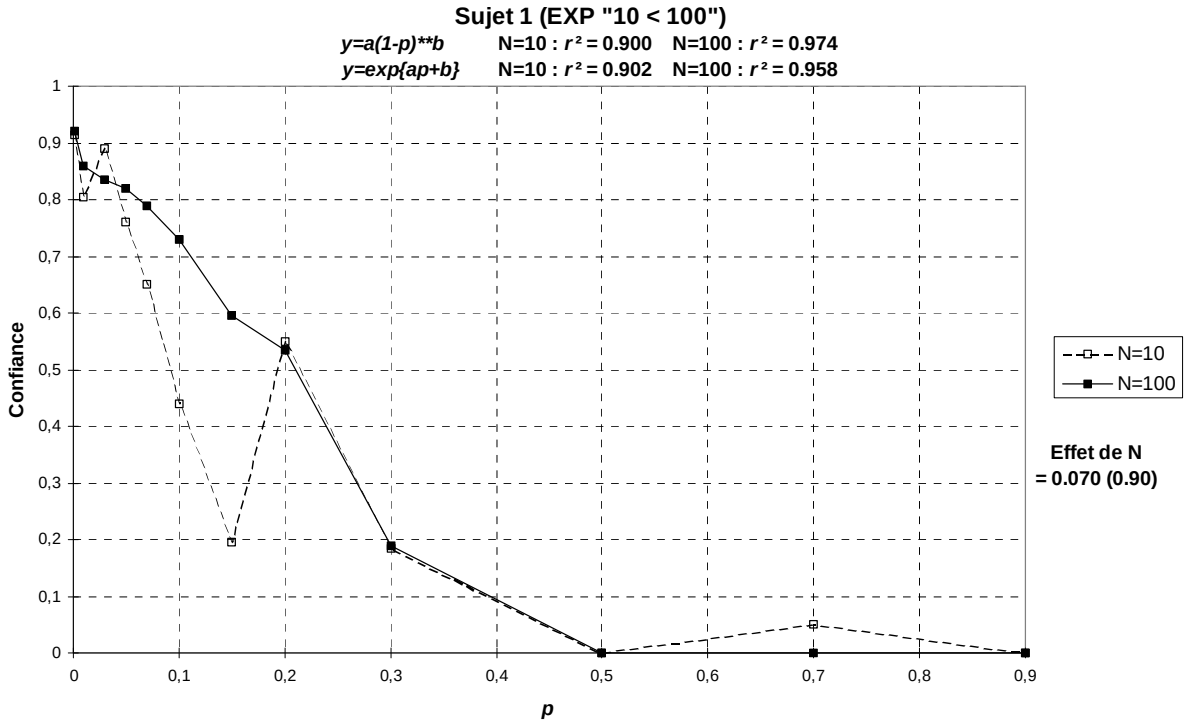


Figure C1

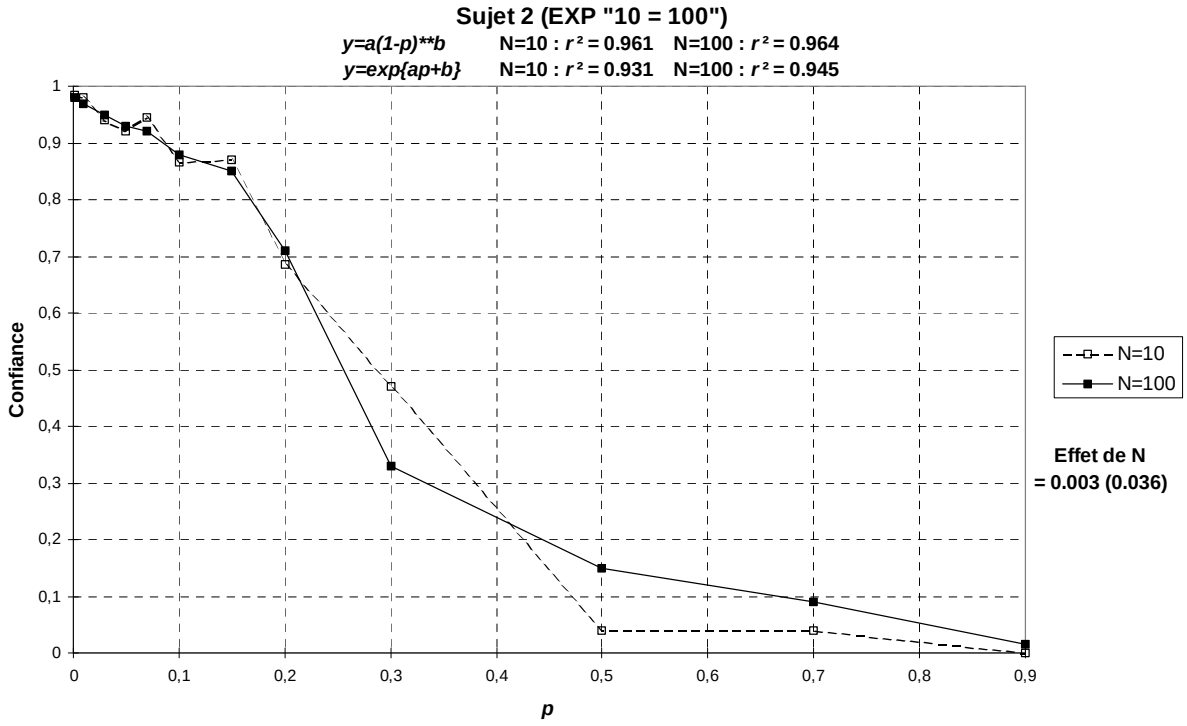


Figure C2

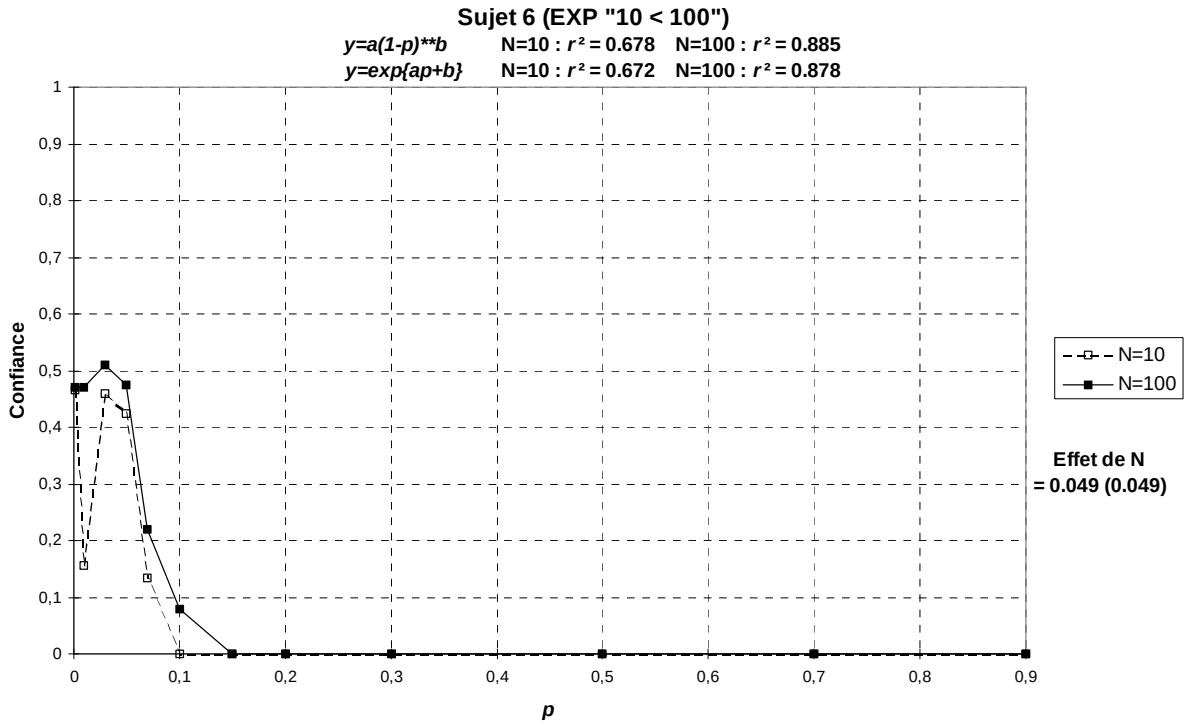


Figure C3

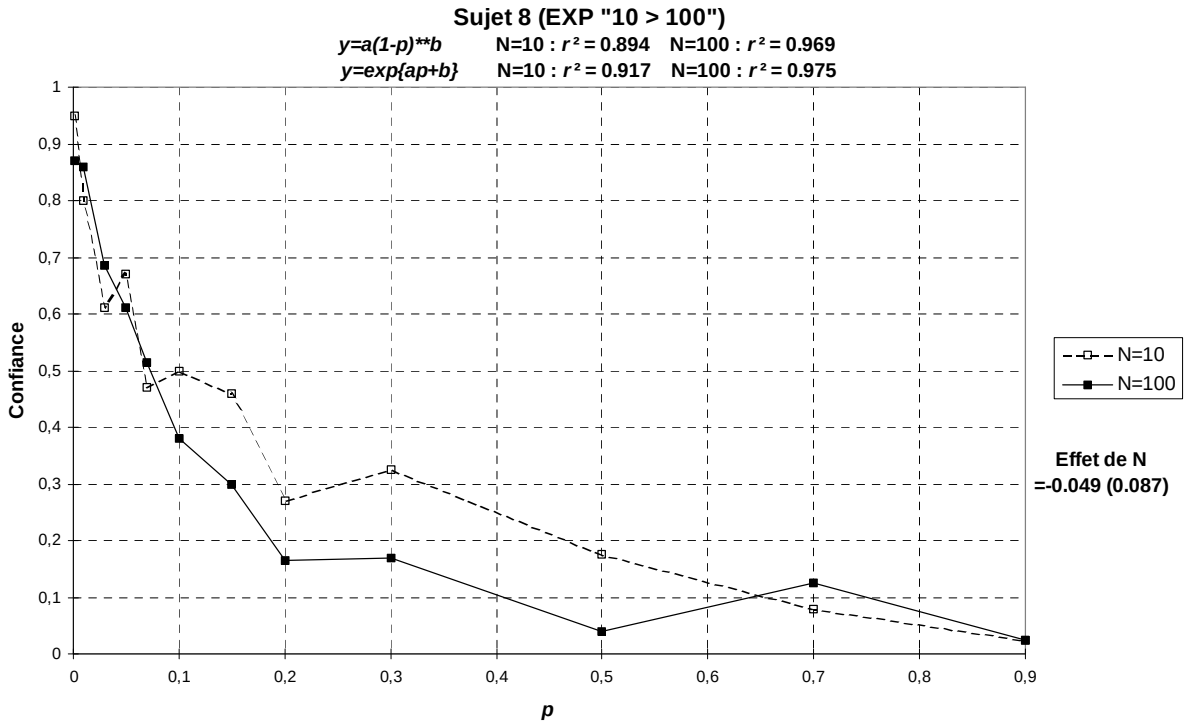


Figure C4

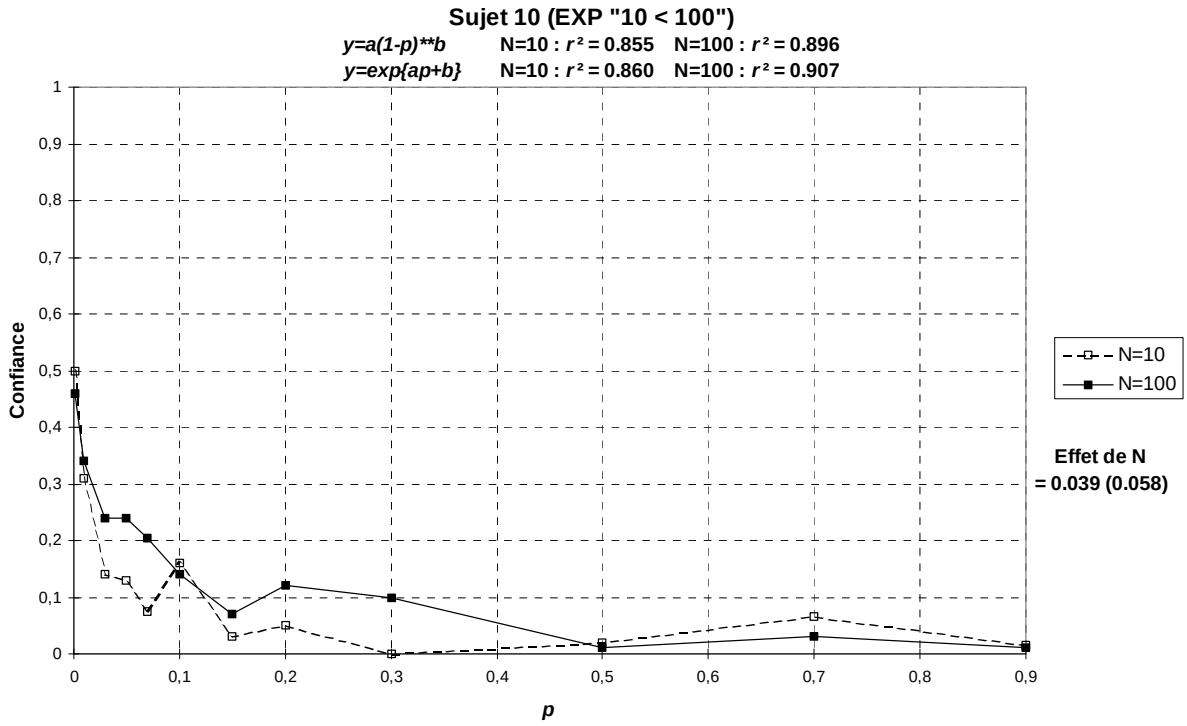


Figure C5

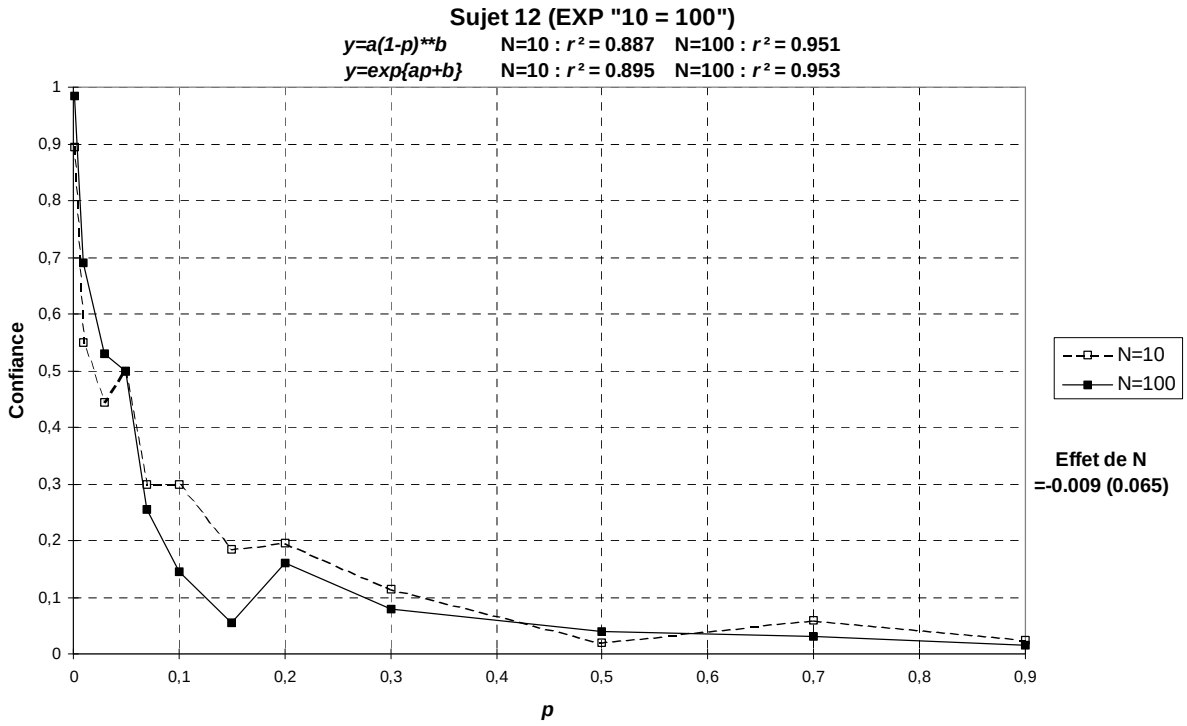


Figure C6

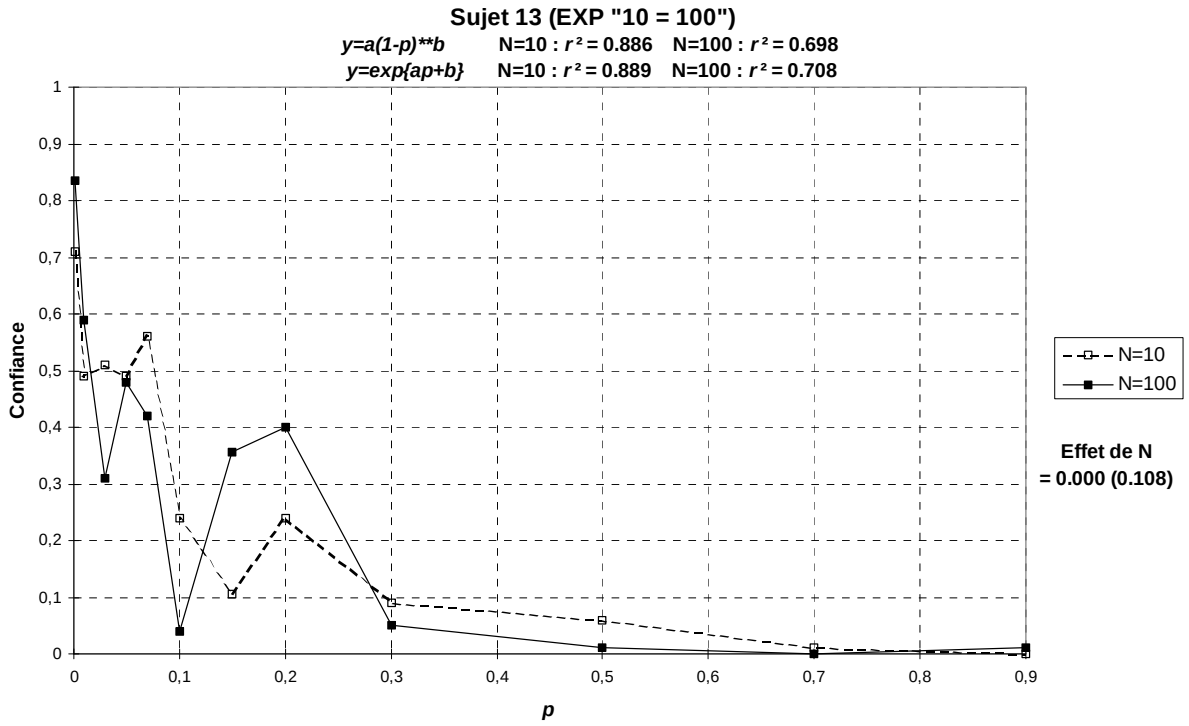


Figure C7

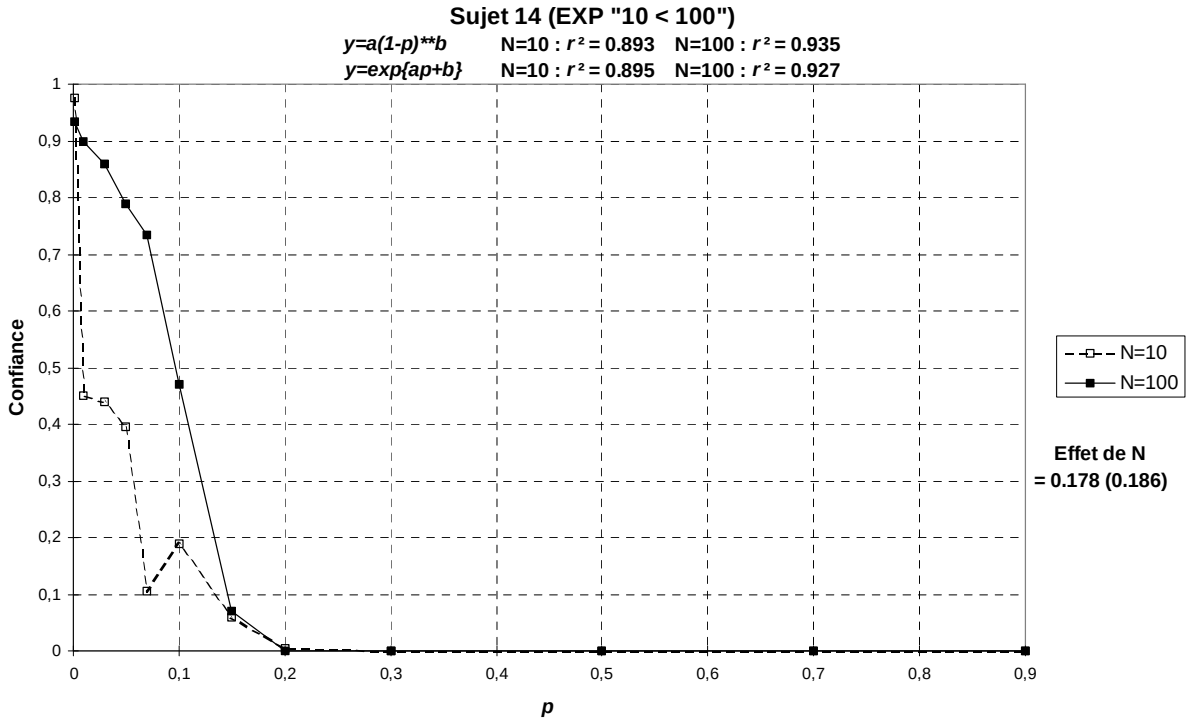


Figure C8

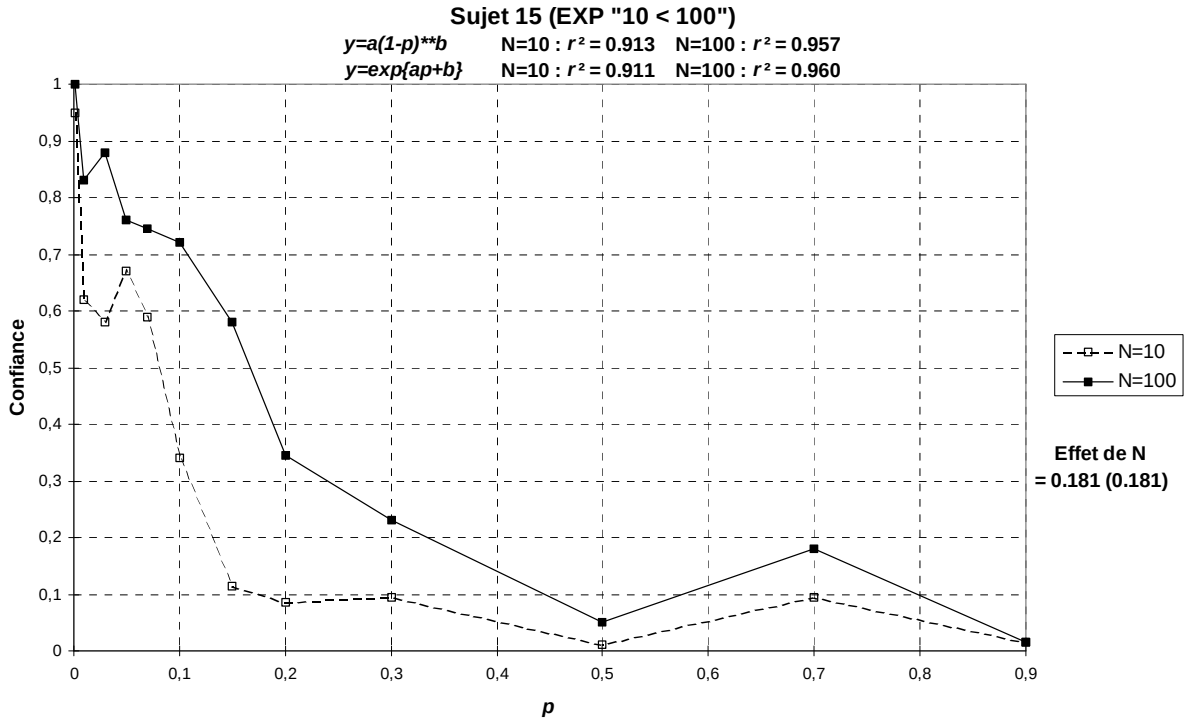


Figure C9

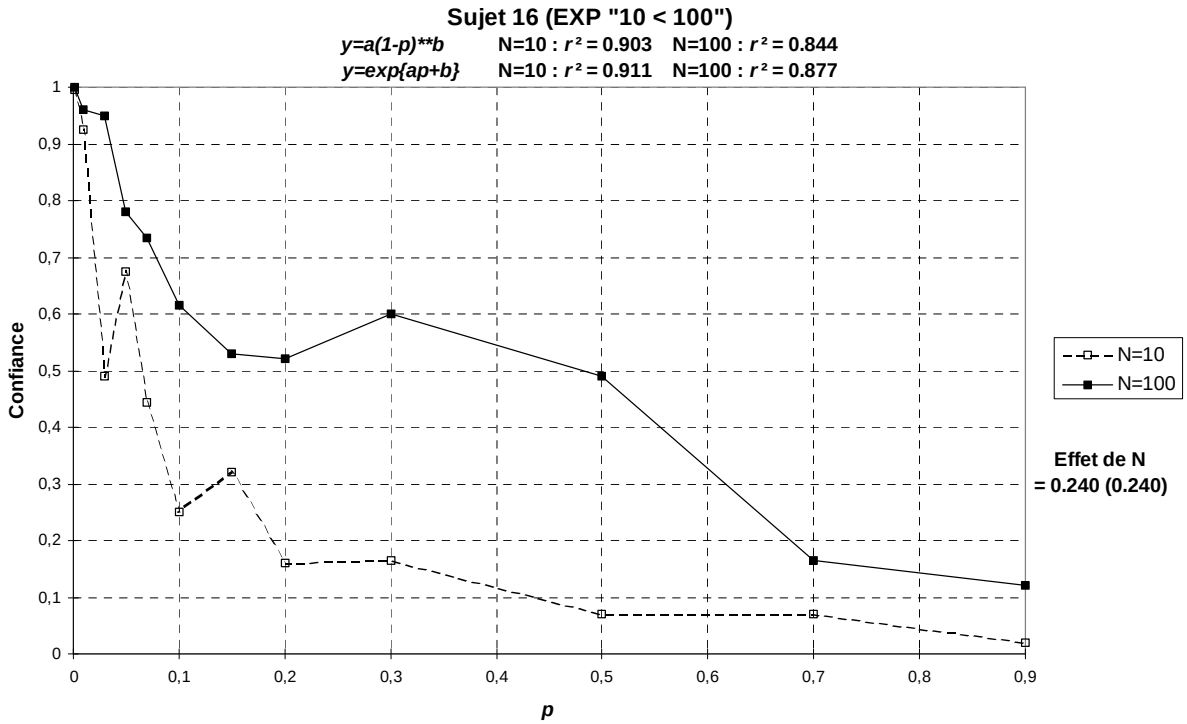


Figure C10

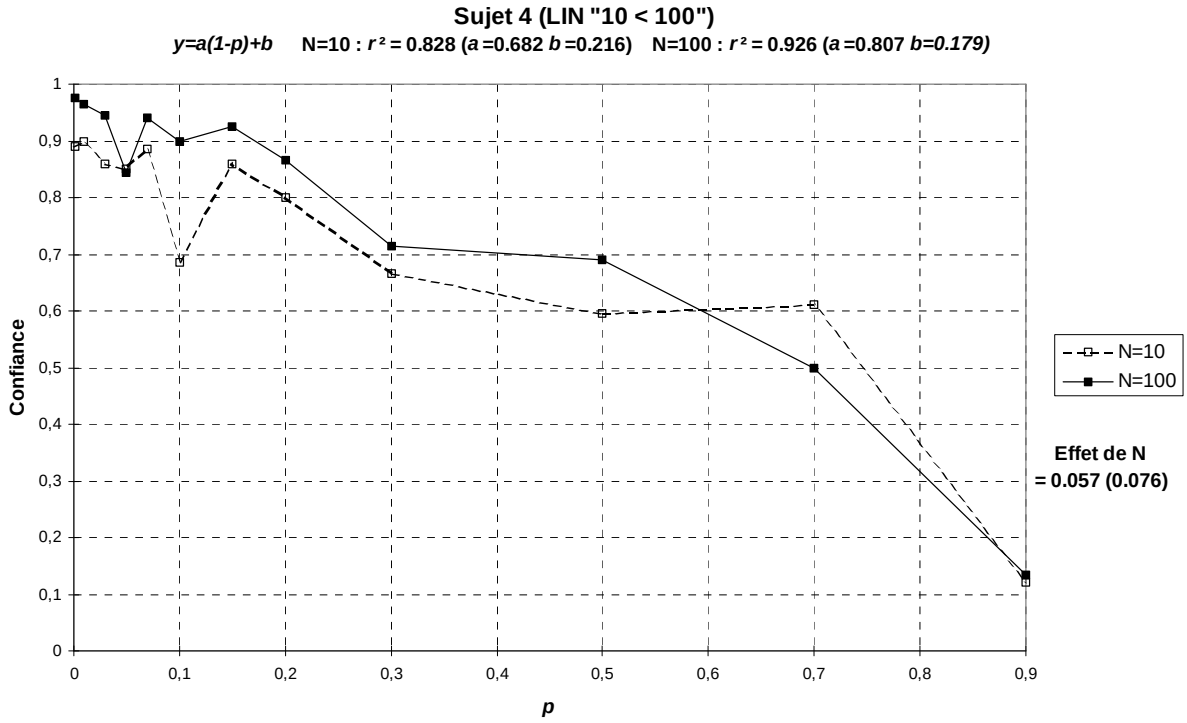


Figure C11

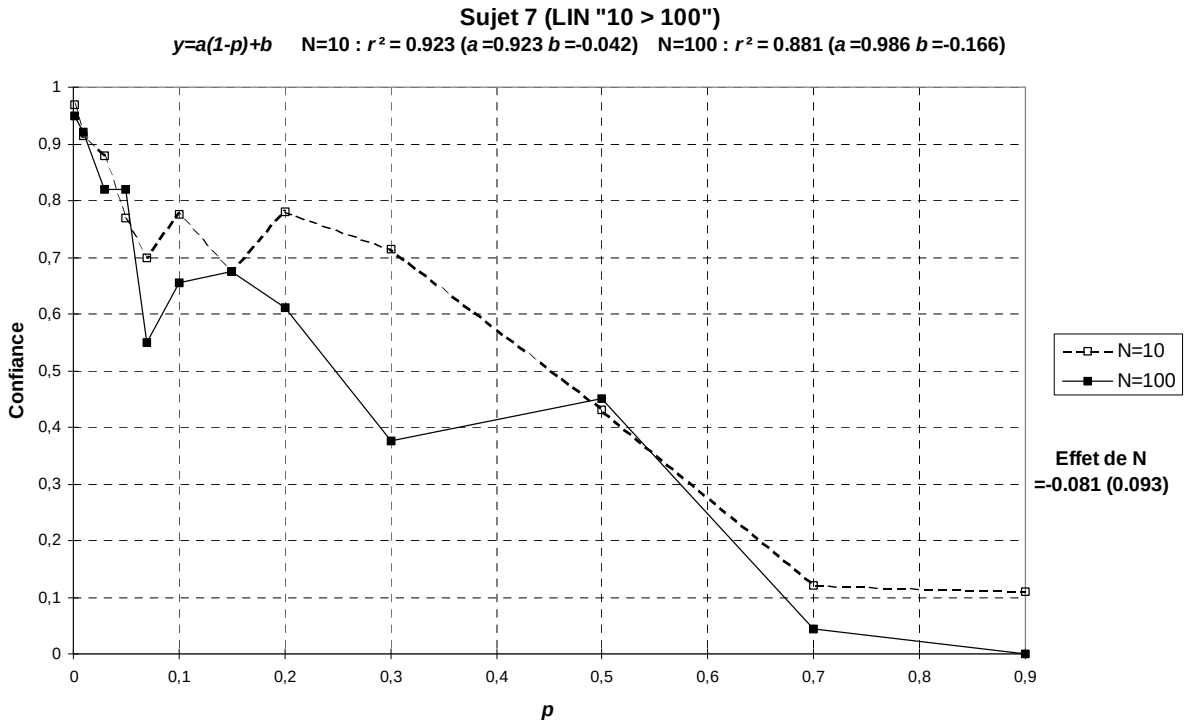


Figure C12

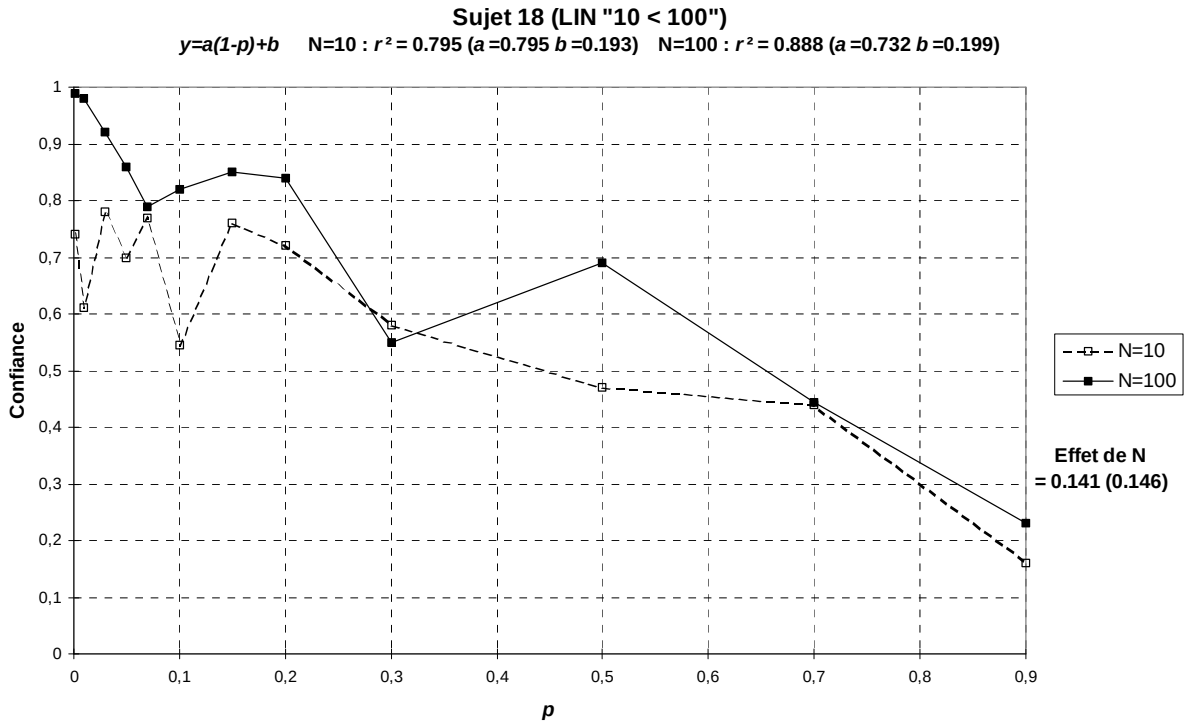


Figure C13

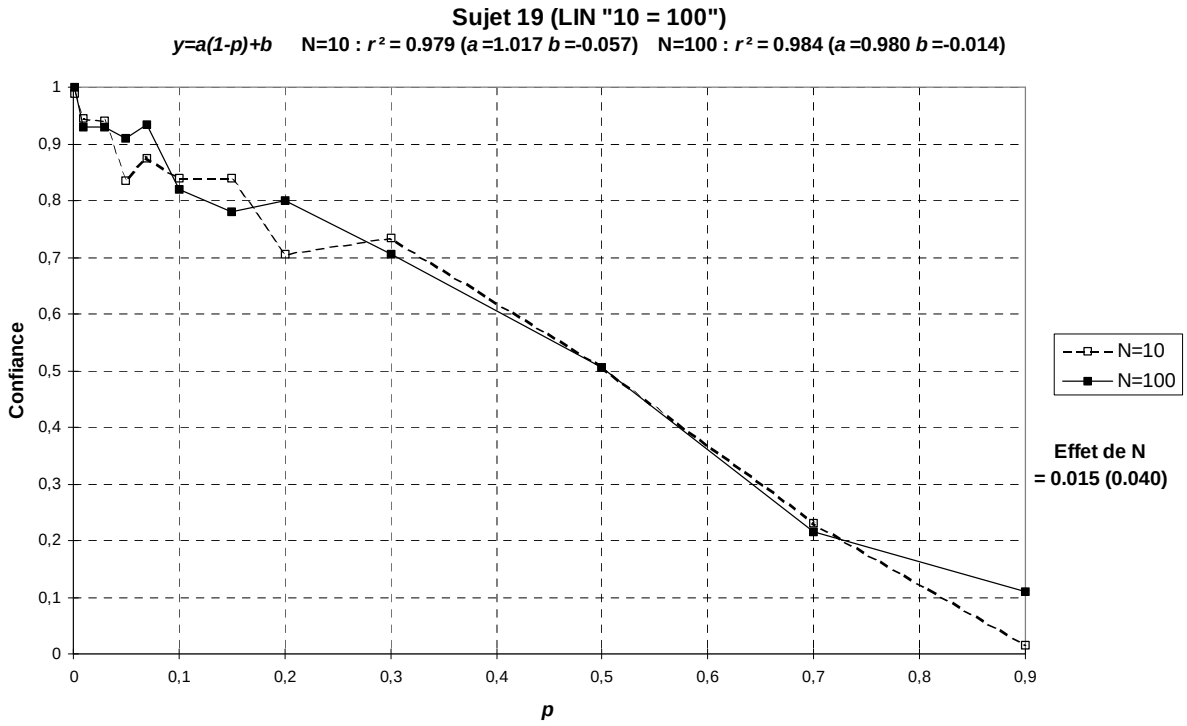


Figure C14

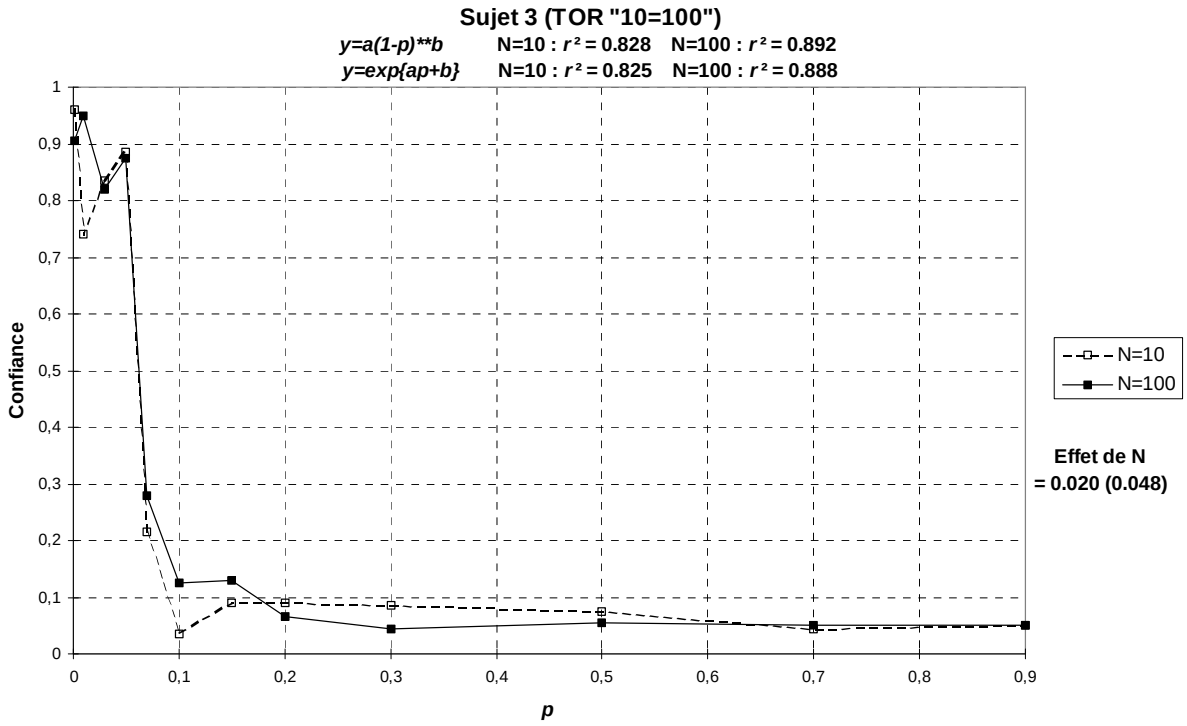


Figure C15

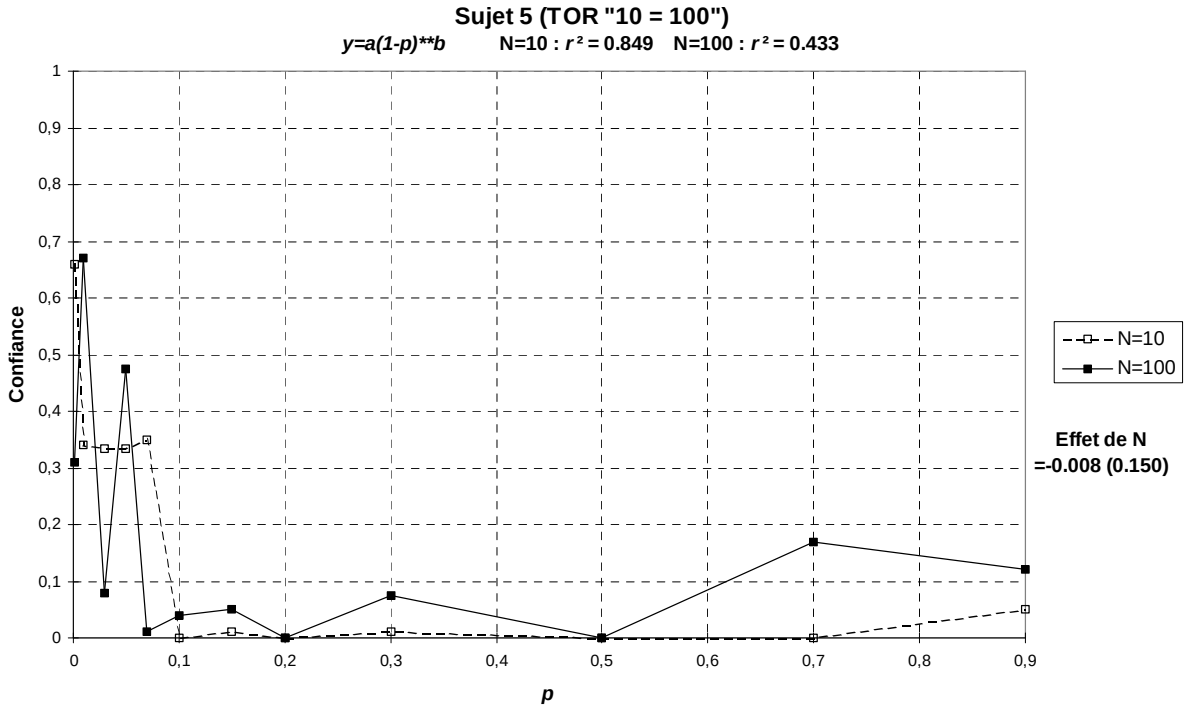


Figure C16

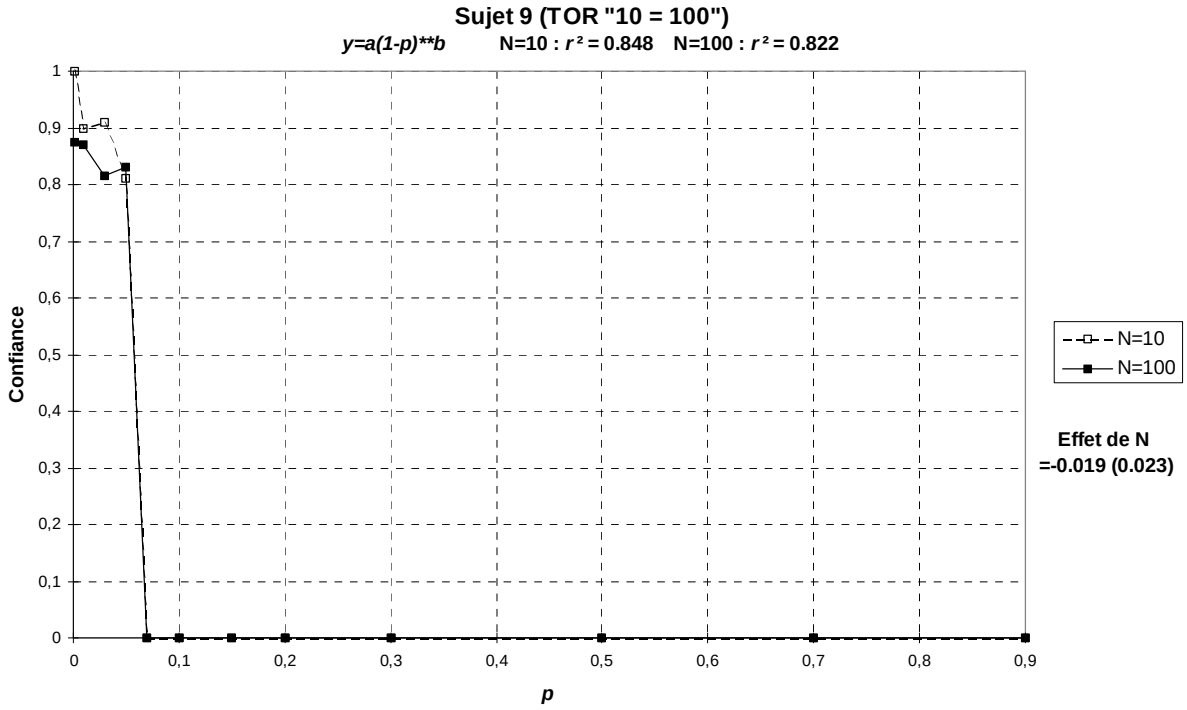


Figure C17

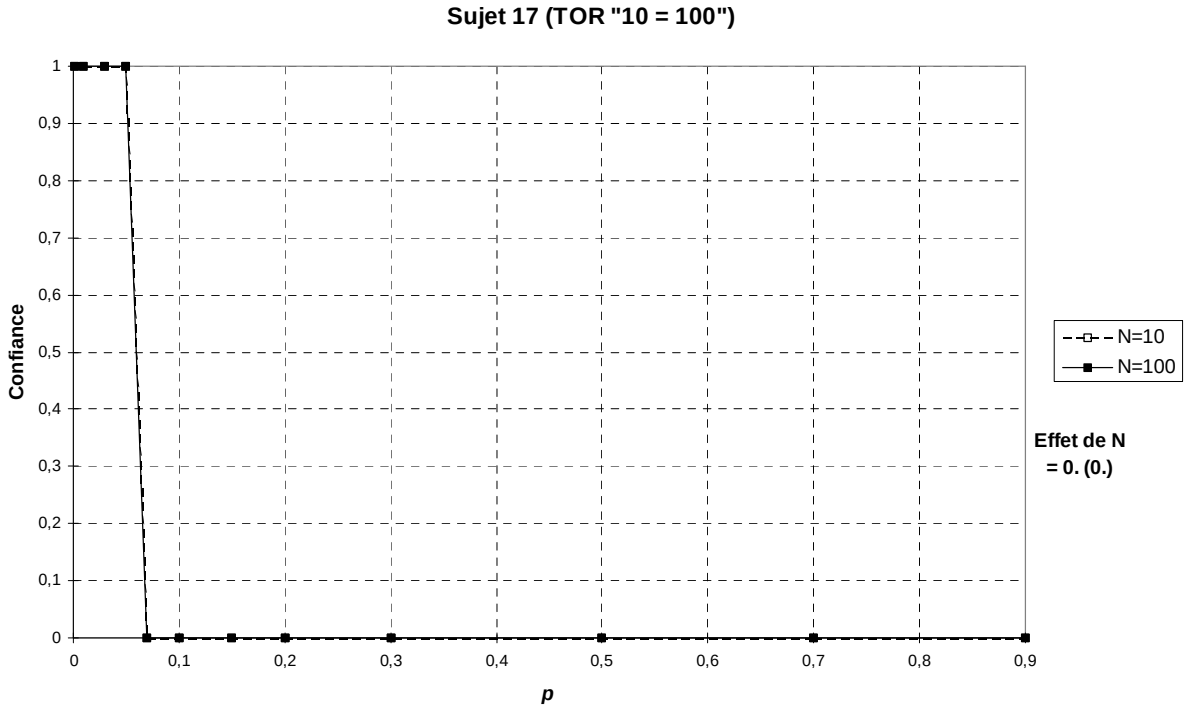


Figure C18

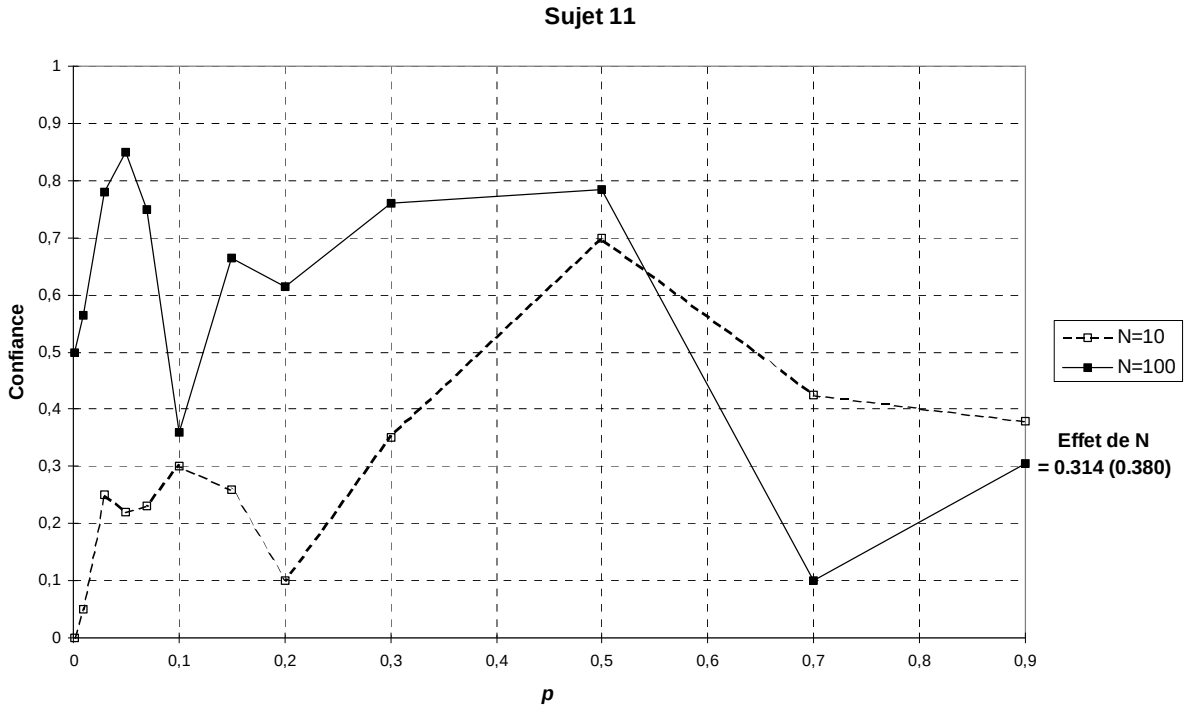


Figure C19

Figures C20 - C24 : courbes moyennes

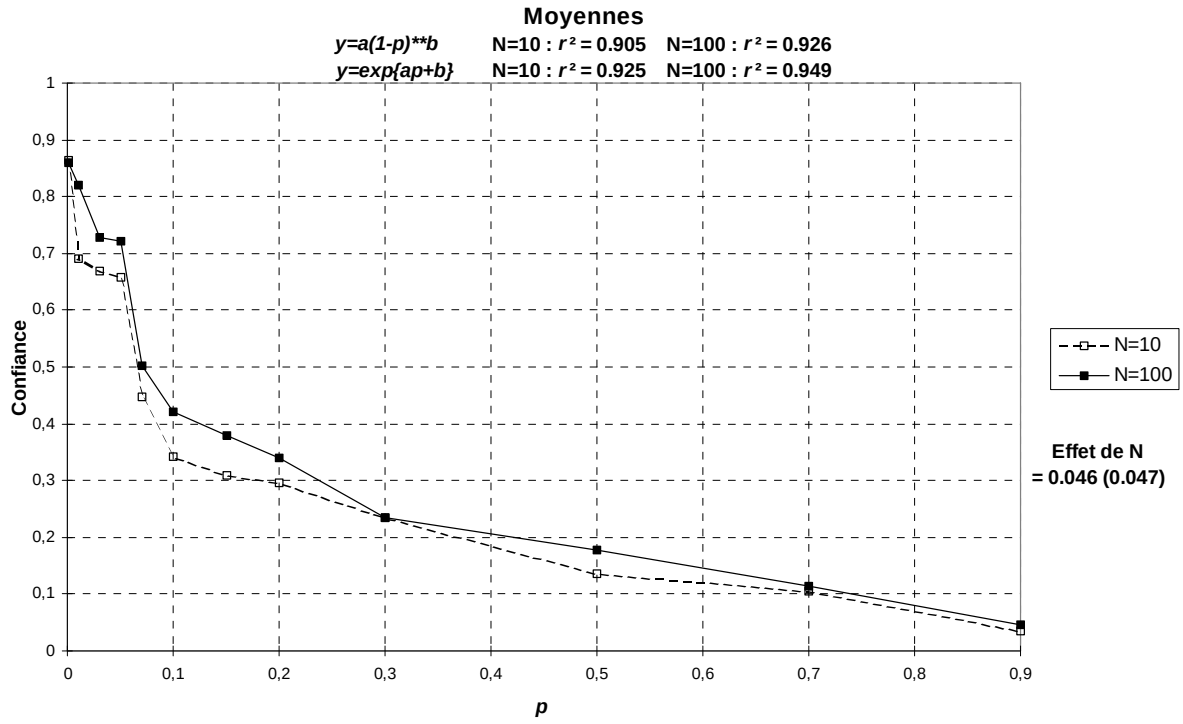


Figure C20

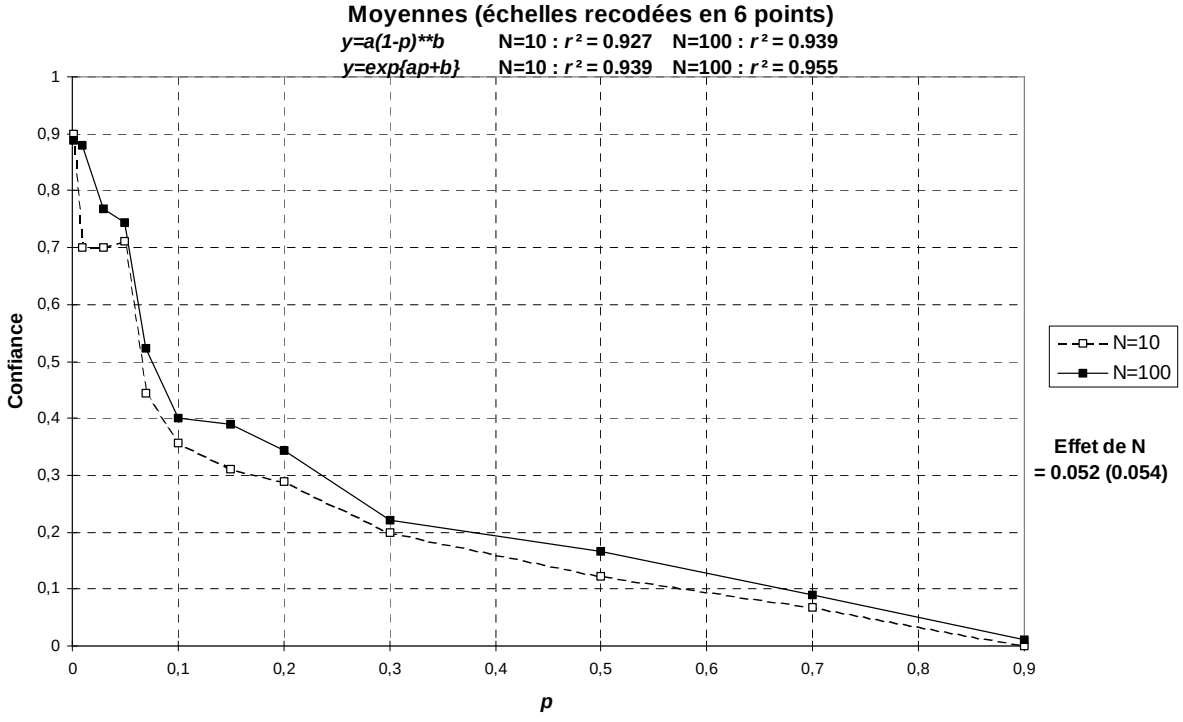


Figure C21

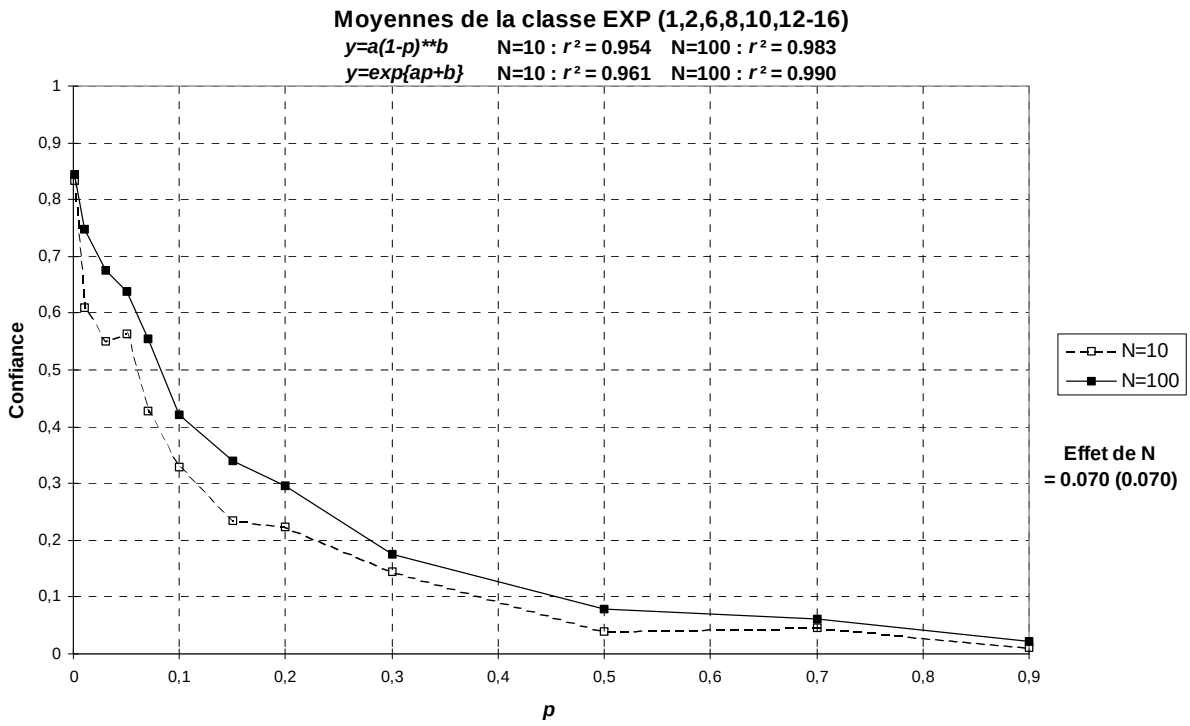


Figure C22

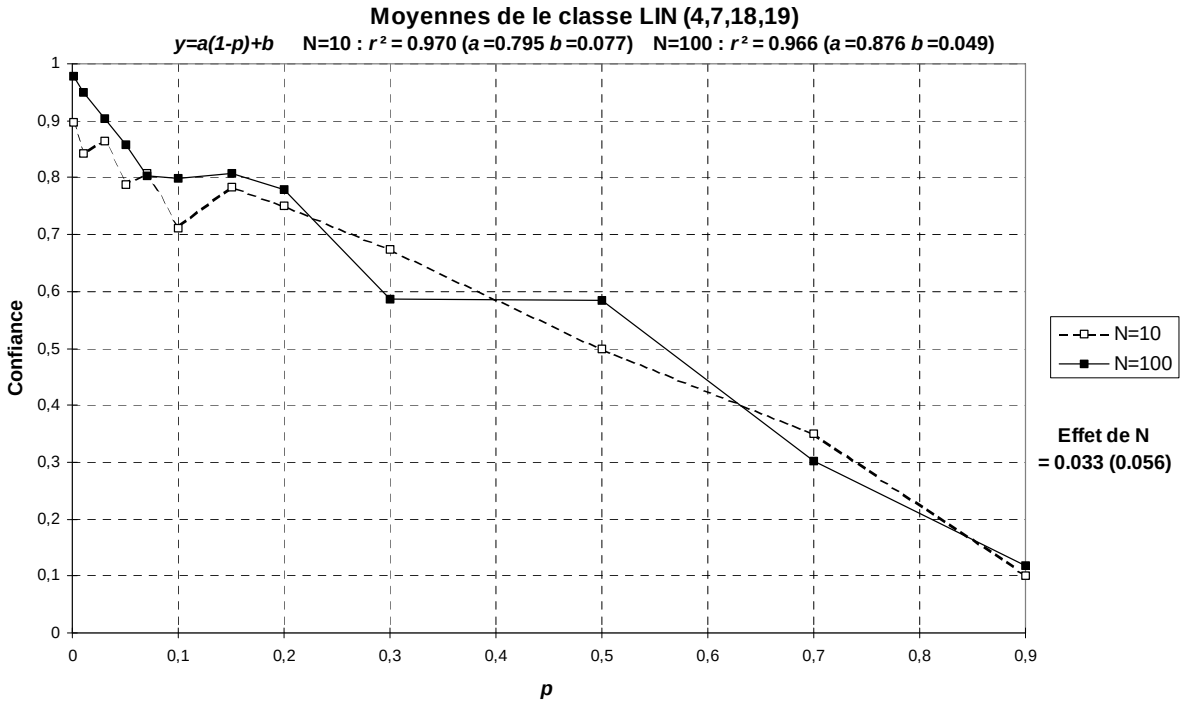


Figure C23

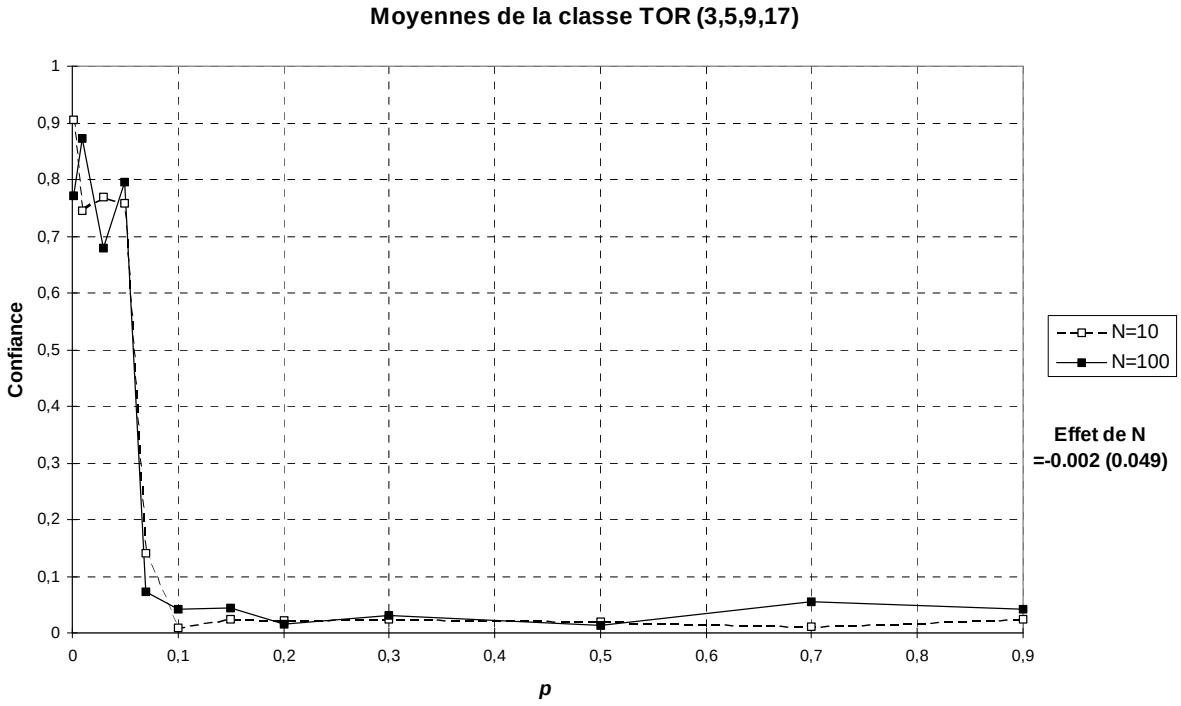

Moyennes de la classe TOR (3,5,9,17)
Confiance
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
0,1
0,2
0,3
0,4
0,5
0,6
0,7
0,8
0,9
p
N=10
N=100
Effet de N
=-0.002 (0.049)

Figure C24

Figures C25 - C32 : courbes individuelles des sujets 1 et 15 (test / retest)

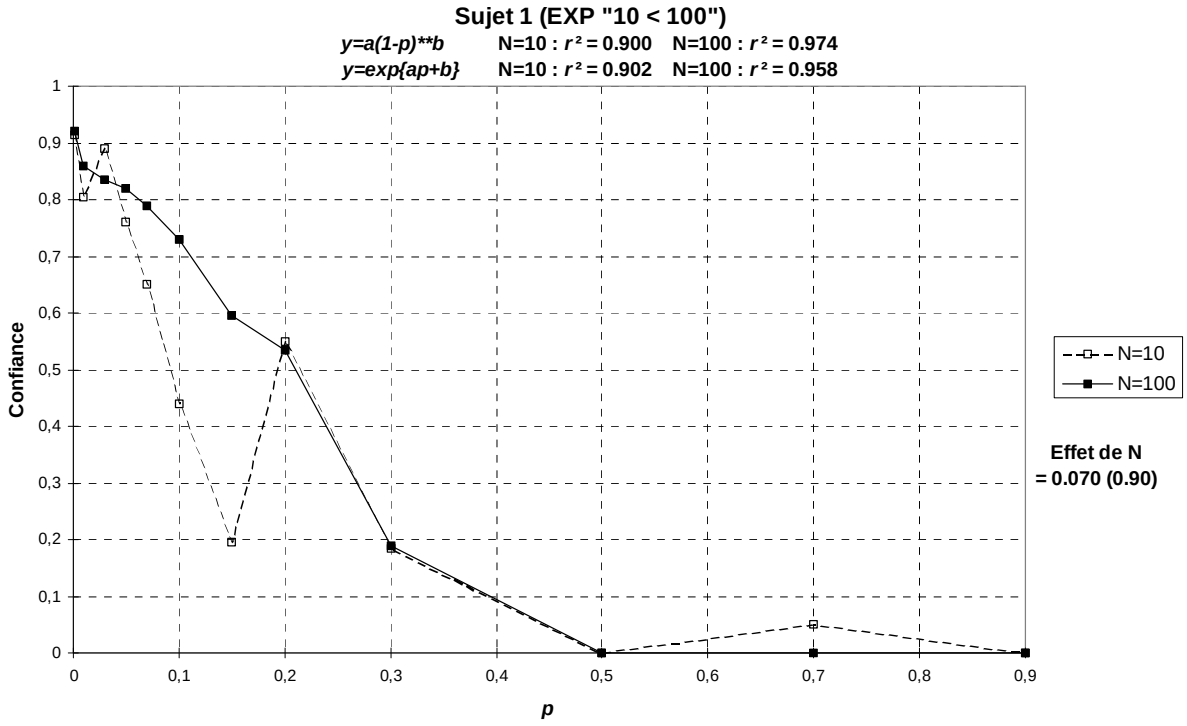


Figure C25

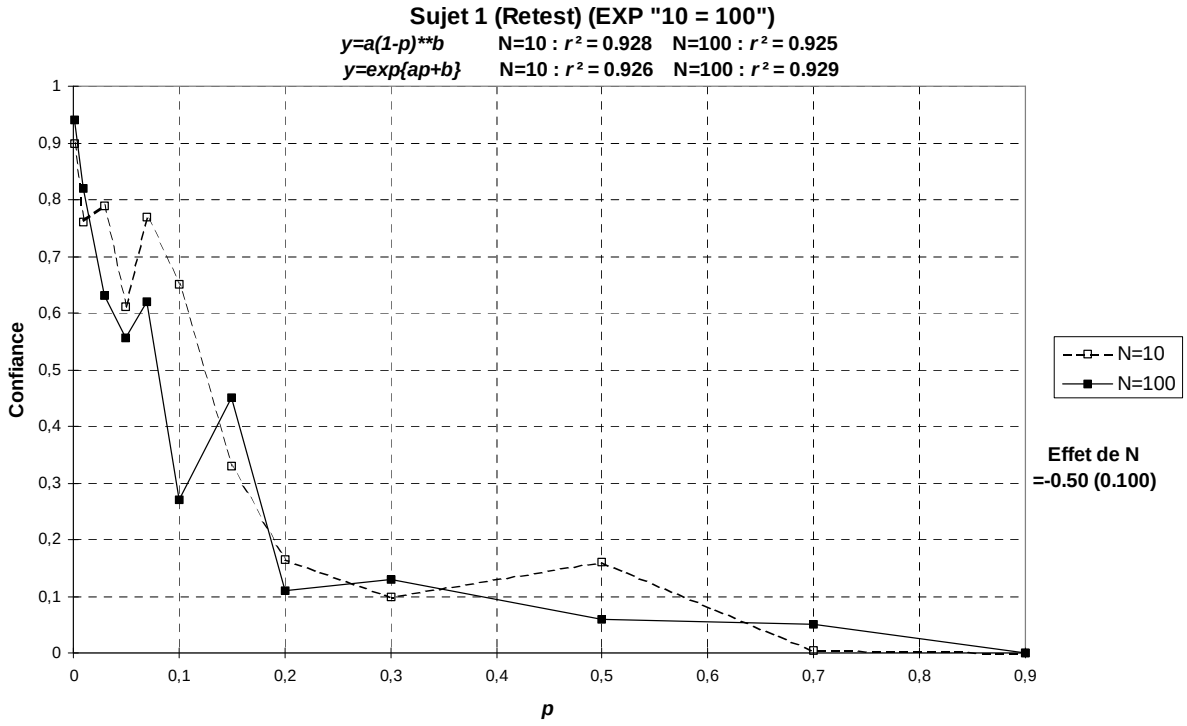


Figure C26

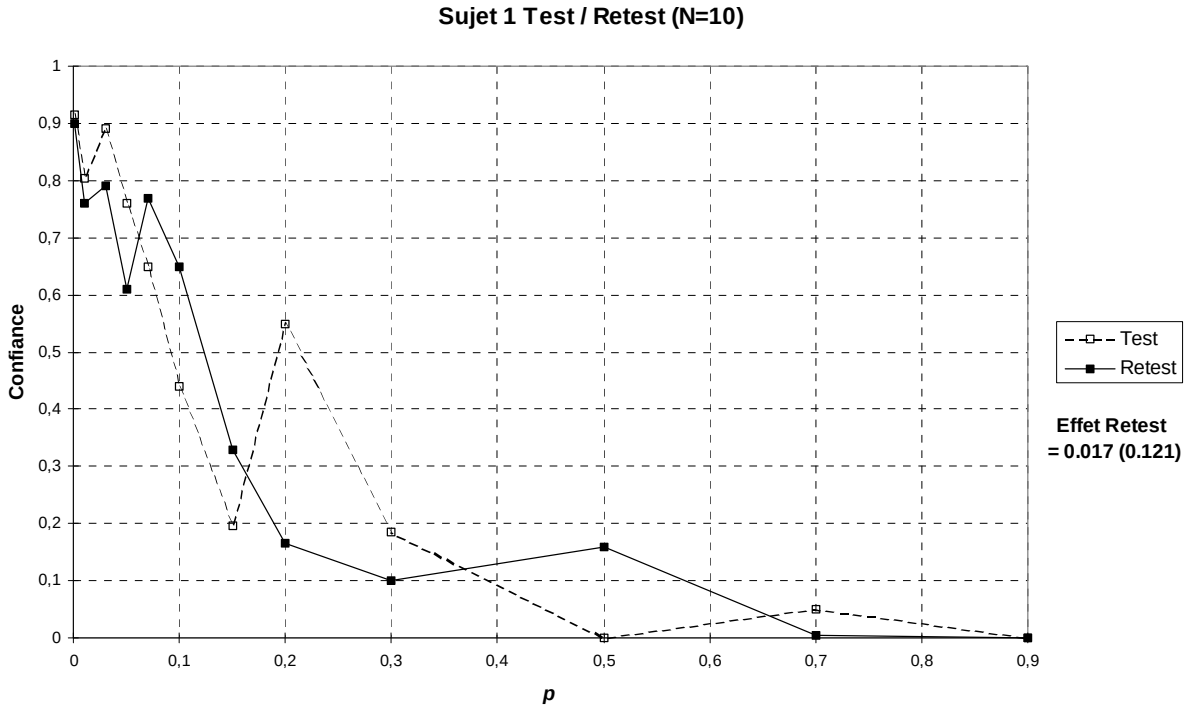


Figure C27

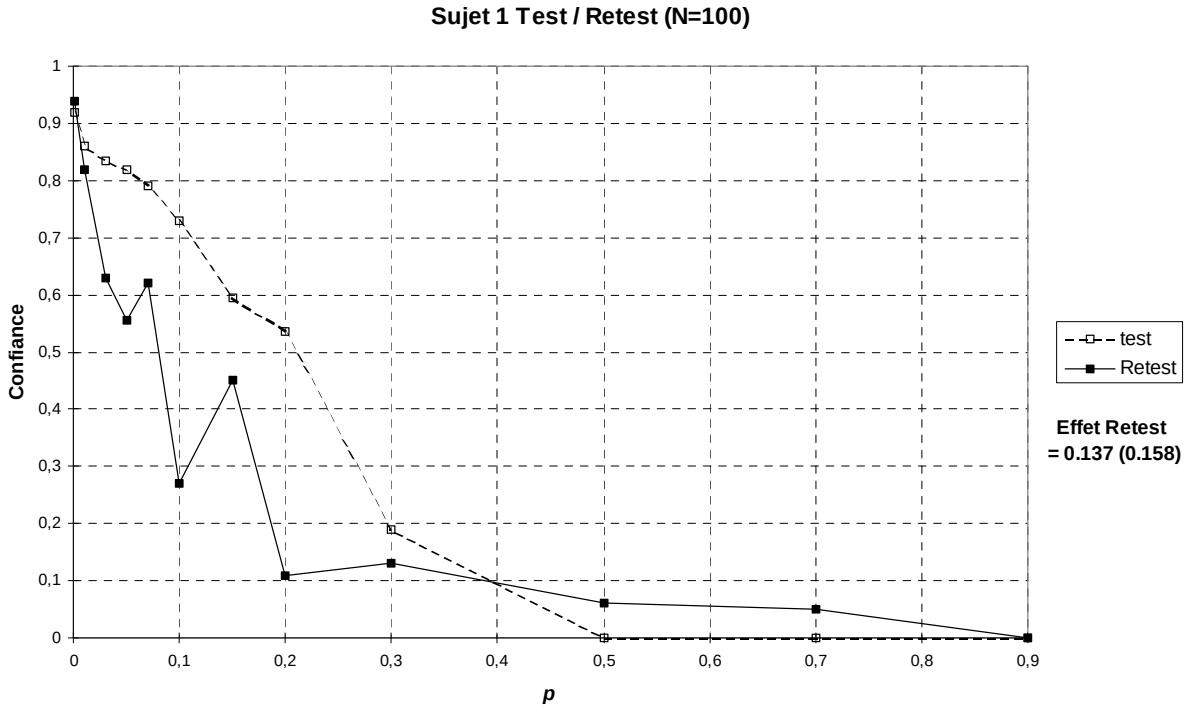


Figure C28

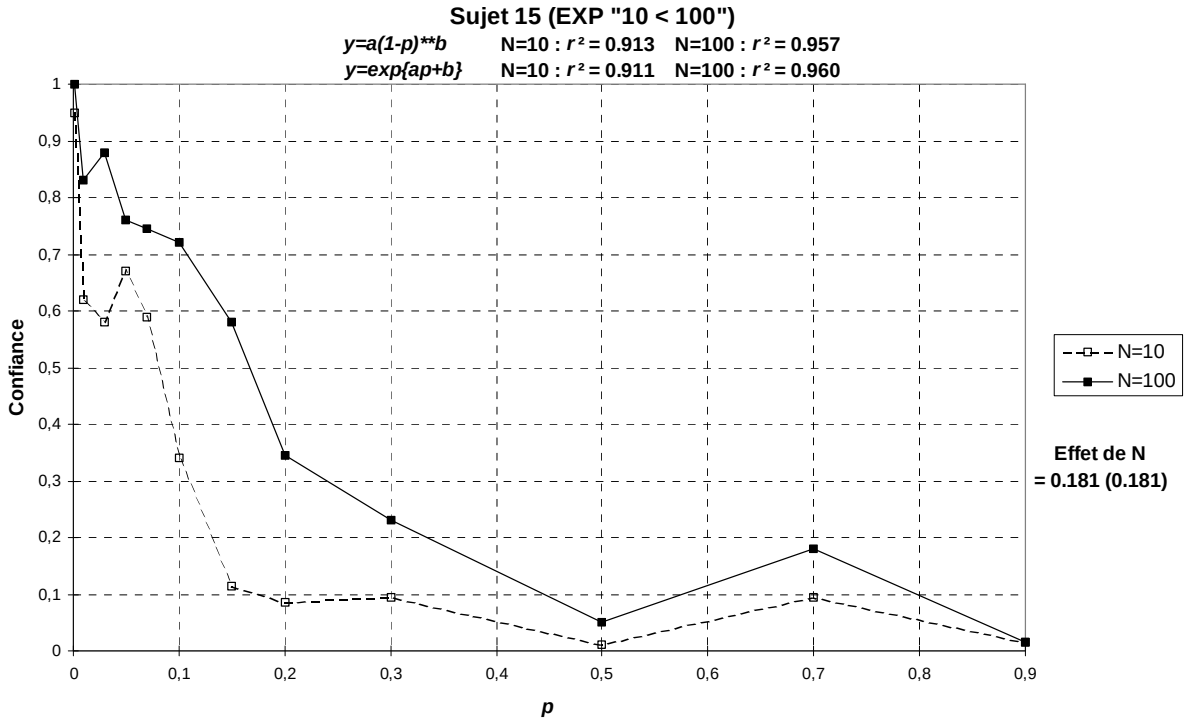


Figure C29

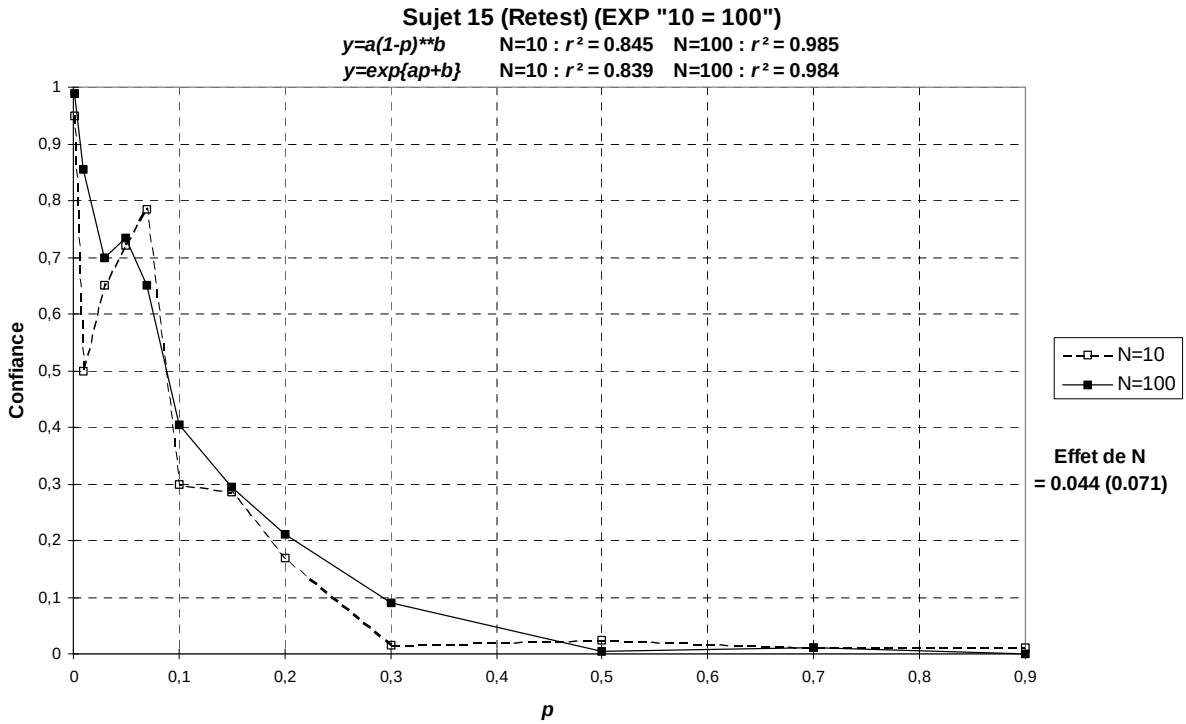


Figure C30

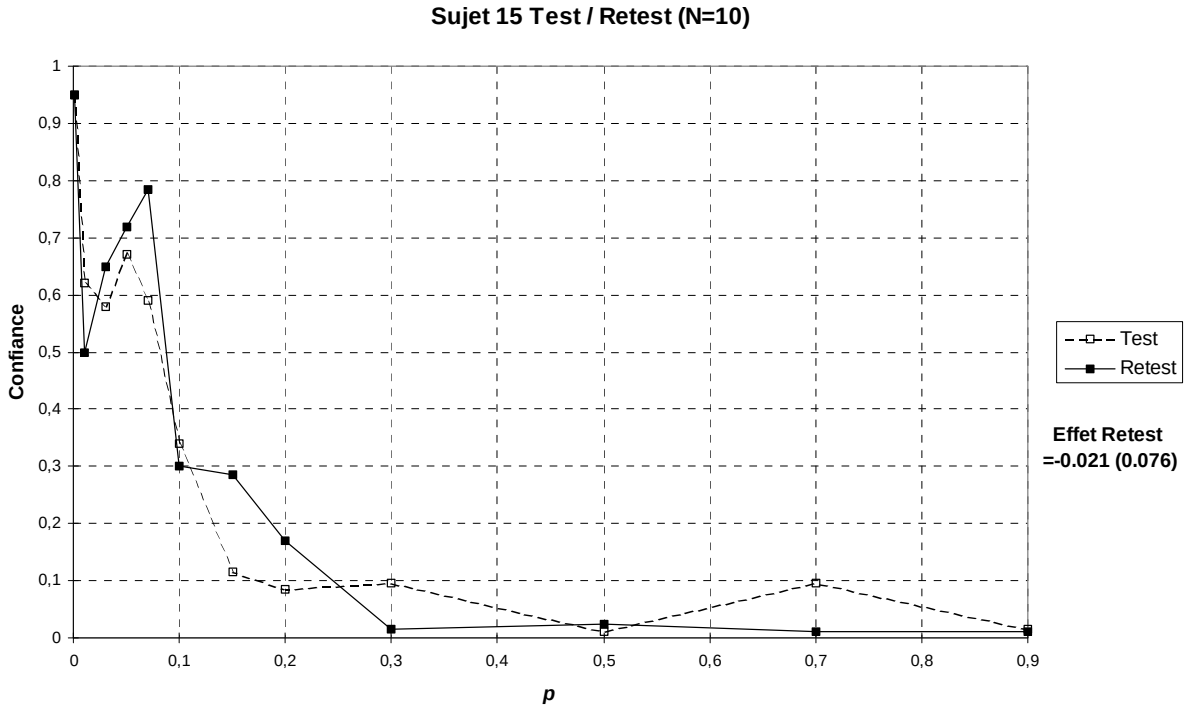


Figure C31

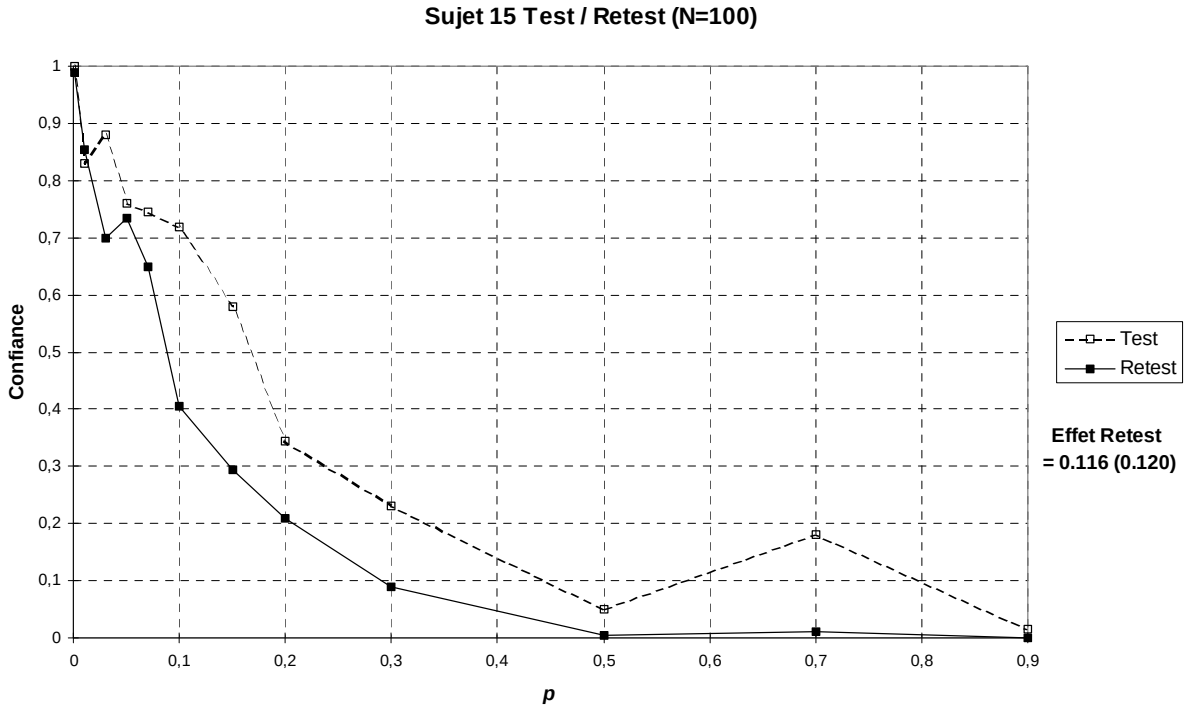


Figure C32

Figures C33 - C40 : courbes individuelles pour la consigne "hypothèse nulle"

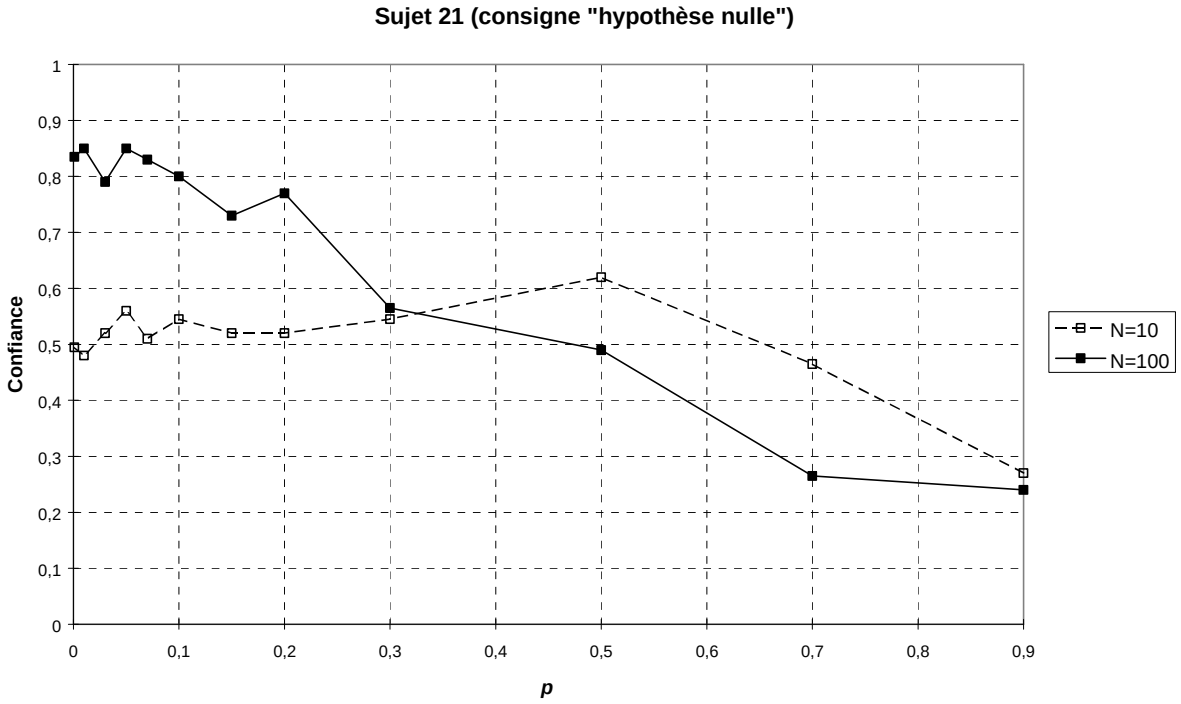


Figure C33

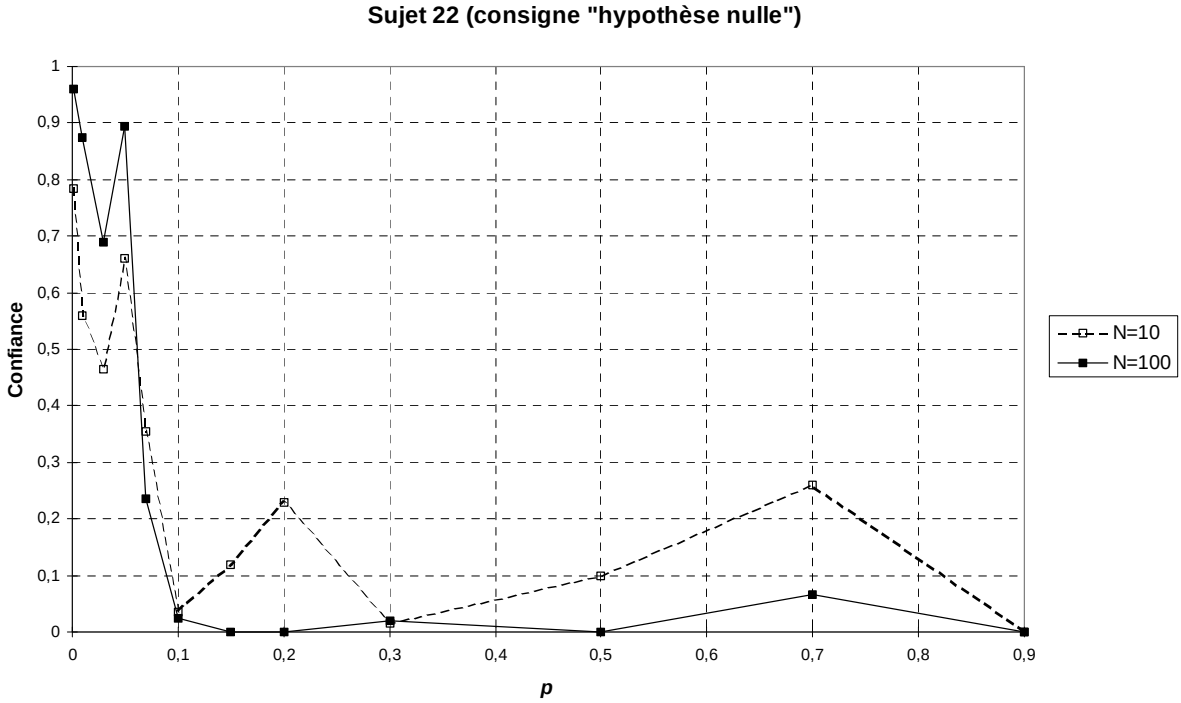


Figure C34

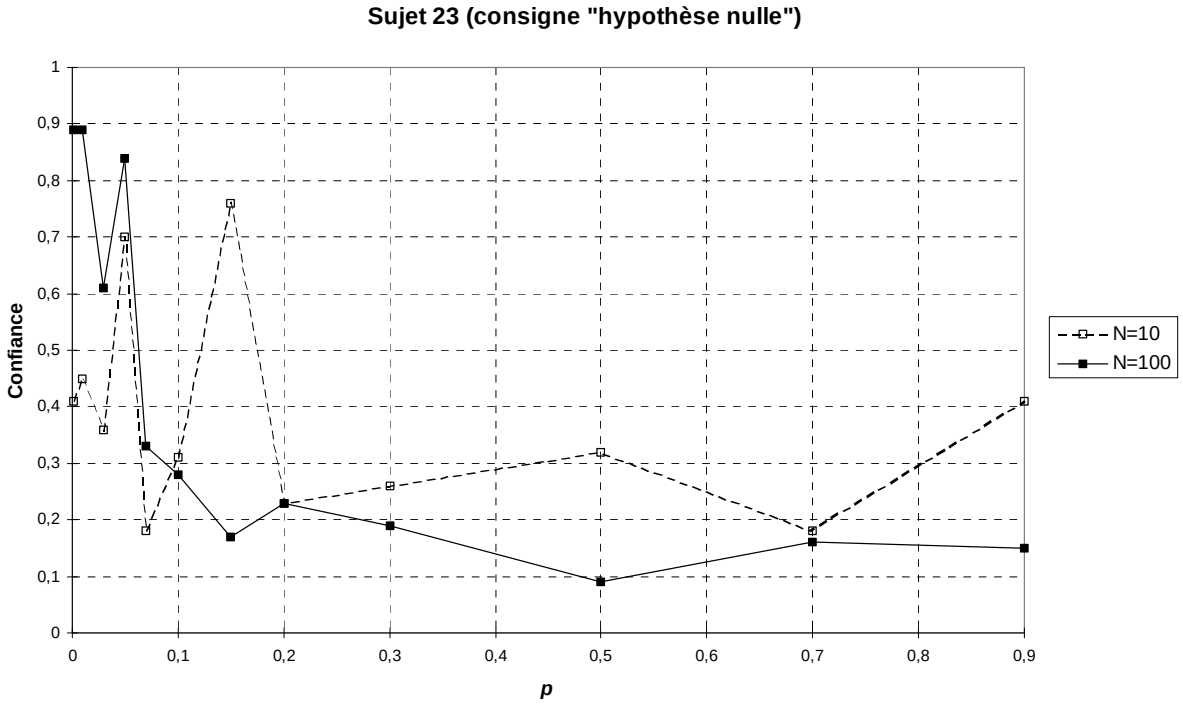


Figure C35

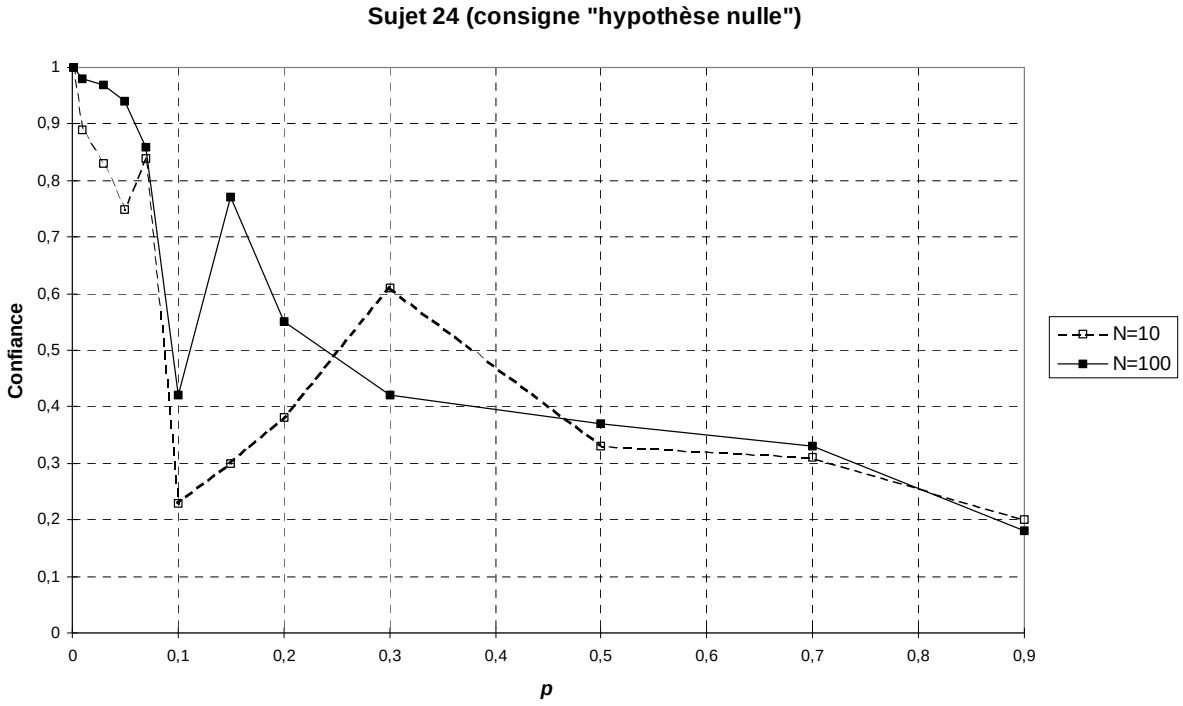


Figure C36

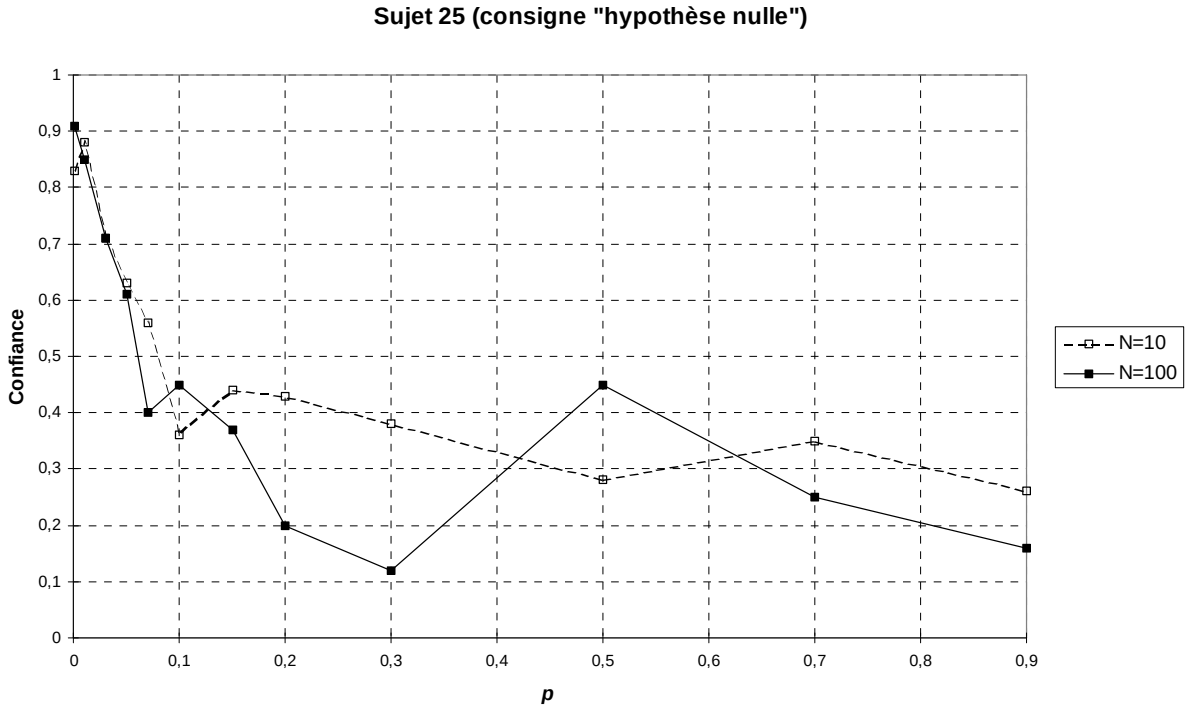


Figure C37

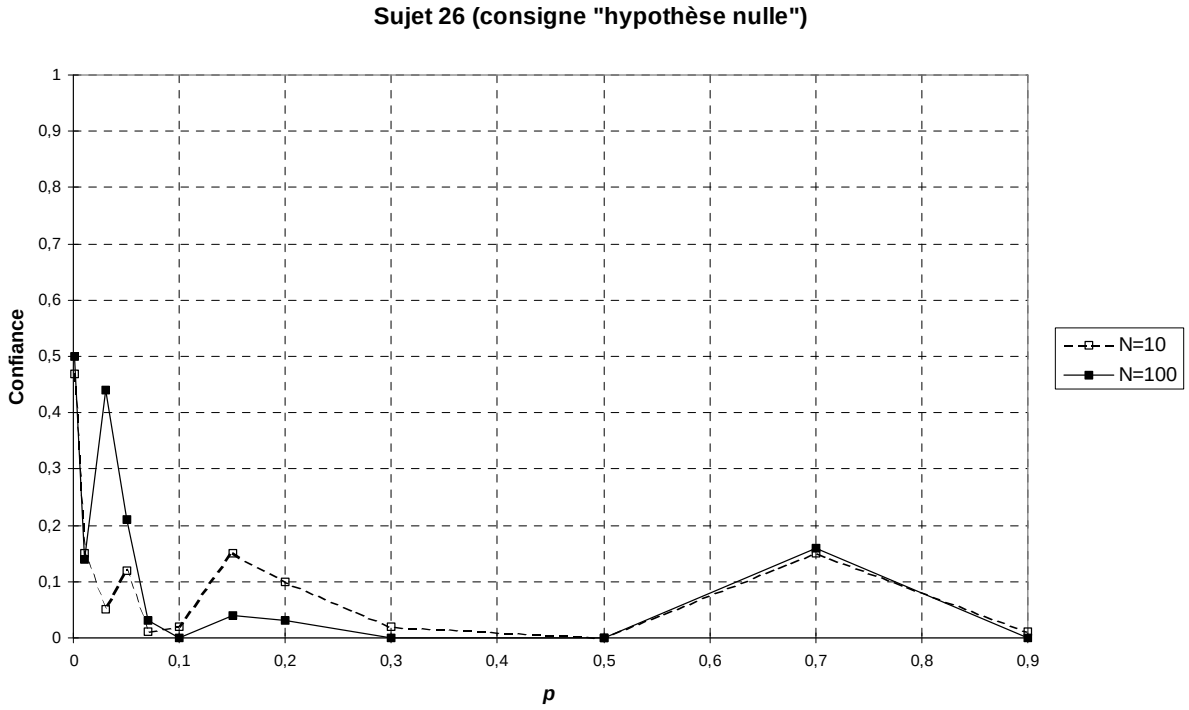


Figure C38

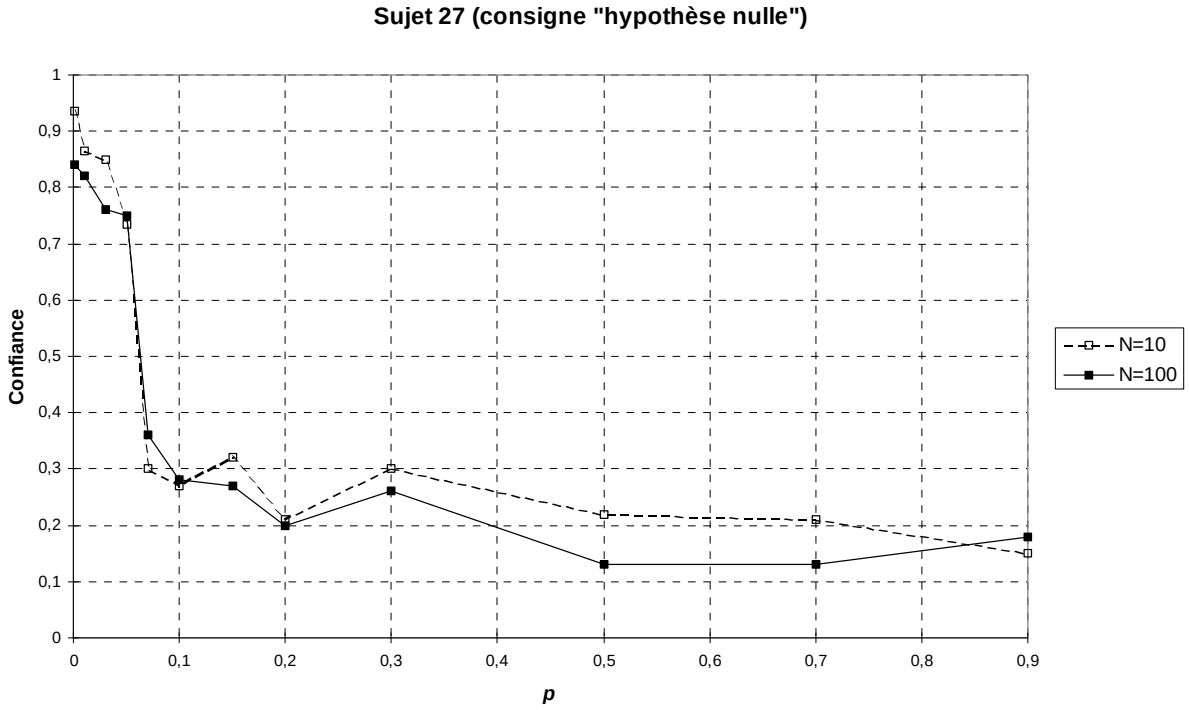


Figure C39

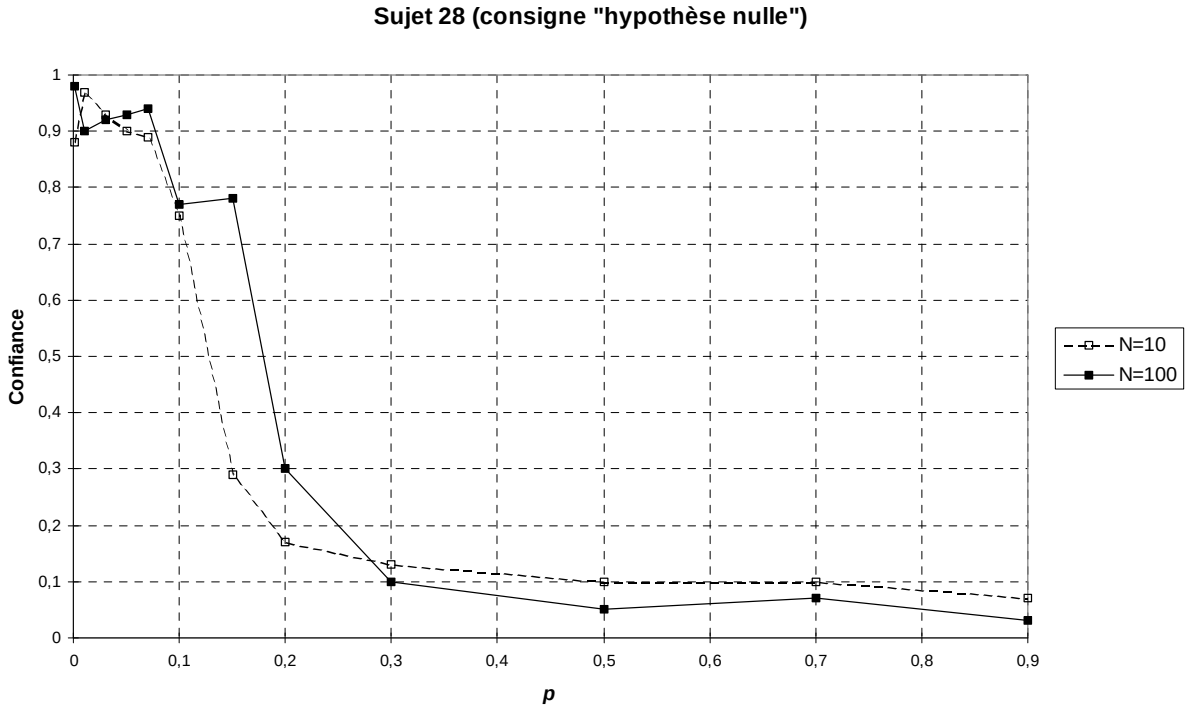


Figure C40