Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(3), p. 127-136.

Le nombre de permutations dans les tests permutationnels
The number of permutations in permutation tests

Louis Laurencelle
Université du Québec à Trois-Rivières

In a first part, the concepts and theory of exact randomization tests are reviewed, together with their implementation for each of a number of customary test situations including simple anova designs. Approximate (or incomplete) randomization tests are considered in the second part, as manageable alternatives to exact tests. We propose a model to calculate the relative power of approximate randomization tests and sketch out some guidelines for the user.

Article first published in Lettres Statistiques, *1987, vol. 8, p. 51-77.*

Fisher, en 1936, proposait une procédure idéale pour étudier la différence entre deux groupes d'observations. Il s'agissait de considérer les deux groupes de données tels qu'observés comme étant une répartition parmi toutes les répartitions possibles des données entre les groupes : sous cette perspective, le rang de la différence obtenue, dans l'ensemble des différences résultant des répartitions possibles, en indiquerait le caractère significatif. « En réalité, le statisticien n'effectue pas cette procédure très simple mais laborieuse; ses conclusions n'ont toutefois d'autres fondements que le fait de concorder avec celles obtenues par cette méthode élémentaire» (Fisher, 1936, p. 58-59, notre traduction). Fisher introduisait là ce qu'on nomme aujourd'hui les « randomization tests », ou tests permutationnels, dérivés de son principe de « randomisation » (Fisher, 1971, p. 17 seq.) pour la construction de plans d'expérience.

Les tests permutationnels à la Fisher occupent maintenant une place modeste dans l'enseignement des statistiques (voir p. ex. Siegel et Castellan, 1988 ; Sokal et Rohlf, 1981), et quantité de tests basés sur ce principe et calculés sur des observations transformées en rangs ont été proposés (Siegel et Castellan, 1988 ; Bradley, 1968 ; Lehmann, 1975). Ces tests ont en commun de calculer une statistique basée sur une configuration observée de résultats, soit $s_o = s(C_o)$, afin de la comparer à l'univers des statistiques découlant de toutes les configurations différentes possibles, soit $s_i = s(C_i)$ pour $i = 1..T$, le nombre T de configurations différentes pouvant être très élevé. Hormis quelques cas très simples, la solution pratique de ces tests doit passer par l'ordinateur. La taille du problème, en termes de temps d'exécution du programme informatique, est alors proportionnelle au produit de trois facteurs : la complexité de la statistique à calculer, la production de chaque configuration distincte, le nombre de configurations. Ce dernier facteur, T, est ordinairement si élevé que la solution est inaccessible en pratique : c'est pour cette raison que Fisher (1936) ne proposait cette procédure qu'idéalement, ou didactiquement. Pour cette raison aussi, on est amené à considérer des approximations des tests permutationnels, les tests paramétriques usuels étant de bons candidats à cette fin (Bradley, 1968, p. 15-44, 88-91 ; Kendall et Stuart, 1979, chap. 31).

Une autre façon de contourner la difficulté issue du nombre élevé de configurations possibles consiste à n'en retenir qu'un échantillon aléatoire. Déjà en 1908, W. S. Gosset, alias « Student », utilisait une méthode d'échantillonnage aléatoire pour construire une distribution empirique de quotients *t* et vérifier sa conformité approximative avec la loi distributionnelle qui depuis porte son nom. Cet emploi de l'échantillonnage aléatoire afin d'étudier le comportement d'un processus quantitatif tombe aujourd'hui sous l'appellation collective de « méthodes Monte Carlo » (Metropolis et Ulam, 1949 ; Laurencelle,

2001); grâce à la vitesse et l'automaticité des calculs par l'ordinateur, ces méthodes se sont répandues avec succès en chimie physique, en acoustique, en hydrodynamique, etc. Le recours à l'échantillonnage aléatoire pour l'estimation et la décision statistiques semble plus récent : voir par exemple Efron (1982) pour la technique du *bootstrap*. Quant à l'application de ces méthodes aux tests de signification statistique, elle apparaît déjà dans Hoeffding (1952), puis Barnard (1963).

Ces tests à permutations aléatoires, ou « approximate randomization tests » (Edgington, 1969), consistent simplement à échantillonner au hasard un nombre de configurations raisonnable parmi toutes celles possibles, puis à déterminer comme tantôt le rang de la statistique initialement observée. En restreignant le nombre de configurations à un maximum raisonnable, la solution par ordinateur de ces tests permutationnels devient généralement possible. Cela est un avantage extraordinaire pour ces tests car ils sont applicables pour toutes sortes de statistiques et d'hypothèses nulles, même alors que des tests appropriés, paramétriques ou non, n'existent pas ou restent mathématiquement insolubles.

Malgré la grande souplesse et la puissance des tests à permutations aléatoires, la tradition d'enseignement et de pratique en statistique ne les a pas vraiment intégrés. Les manuels n'en parlent pas (sauf, p. ex., quelques pages dans Biometry, de Sokal et Rohlf, 1981, p. 787 seq.), et peu d'études sur la structure et la validité de ces tests ont été publiées. Même si on inclut les tests à permutations complètes à la Fisher, une documentation spécifique sur l'application de ces tests pour des données quelconques ne semble pas disponible. Fait exception l'ouvrage de E. S. Edgington, *Randomization tests*, paru en 1980. Cet ouvrage est en fait un traité pratique sur les tests permutationnels « systématiques » ou « aléatoires », dans le vocabulaire de l'auteur. Ce dernier présente des rationnels, des programmes d'ordinateur en Fortran, des exemples numériques. C'est, à notre connaissance, le seul livre dans lequel l'idée combinatoire de Fisher, complétée par l'idée d'un échantillonnage selon l'approche Monte Carlo, a reçu un développement sérieux.

Dans les paragraphes suivants, nous donnerons d'abord des exemples de tests permutationnels, complets ou aléatoires, puis nous tenterons d'établir un barème pour indiquer la faisabilité des tests à permutations complètes selon le nombre de permutations engendrées. Enfin, nous appuyant sur quelques auteurs publiés et des raisonnements statistiques, nous vérifierons la validité des tests à permutations aléatoires et proposerons des critères pour fixer le nombre suffisant de permutations à produire.

Trois exemples de tests permutationnels

Nous verrons dans cette section trois exemples de tests permutationnels, – sur la corrélation linéaire, sur la différence de deux groupes, sur la structure d'une matrice de similitudes à deux dimensions. Ces exemples seront l'occasion de concrétiser notre idée sur la réalisation d'un test permutationnel et nous permettront d'en situer la théorie et la pratique dans leur contexte naturel. Noter, à toutes fins utiles, l'ouvrage de Nijenhuis et Wilf (1975), *Combinatorial algorithms*, qui présente un catalogue important d'algorithmes et de programmes d'ordinateur applicables à nos besoins. Ces algorithmes et programmes produisent des permutations, partitions ou combinaisons d'objets soit dans un ordre séquentiel défini, soit au hasard ; Edgington (1980) fournit aussi des programmes, un pour chaque test qu'il étudie.

Test sur la corrélation.

Nous disposons d'une série de n mesures, $\{ x_i, y_i \}$, dont la corrélation observée est r_o. Sous l'hypothèse nulle, la variable Y n'est pas associée à la variable X, et la permutation de la série des Y_i par rapport aux X_i est indifférente. Le nombre de configurations distinctes correspond, en fixant la série des $\{ x_i \}$, au nombre de permutations des y_i, soit $n!$. Si par exemple la série observée des $\{ y_i \}$ est strictement croissante en fonction des $\{ x_i \}$, d'où $r_o \to 1$,[1] le rang de r_o parmi les $n!$ valeurs produites de r sera 1, et la probabilité d'un score r_o aussi extrême sous un test bilatéral sera $2/n!$.

Une méthode efficace pour engendrer les $n!$ permutations requises est décrite dans Laurencelle (1978). Pour produire des permutations aléatoires, la méthode usuelle consiste à permuter au hasard chaque élément Y_i dans le sous-ensemble supérieur $(Y_i, Y_{i+1}, \ldots, Y_n)$, pour $i = 1$ à $n - 1$, le temps de production étant proportionnel à n (Nijenhuis et Wilf, 1975, p. 62).

Test sur la différence de deux groupes.

Nous avons mesuré la variable X pour n_1 éléments du groupe 1, soit $(x_1\ x_2\ \ldots\ x_{n1})$, et n_2 éléments du groupe 2, soit $(x_{n1+1}\ x_{n1+2}\ \ldots\ x_{n1+n2})$, et $n = n_1 + n_2$. Le nombre de répartitions distinctes des n éléments dans les deux groupes est

[1] La corrélation r de Pearson entre ces deux séries coordonnées de statistiques d'ordre s'approchera de 1, sans forcément l'atteindre, alors que la corrélation de rangs (ou « rho ») de Spearman sera 1.

${}_nC_{n1} = n! / (n_1!n_2!)$. Pour statuer sur la différence observée entre les moyennes $(\bar{X}_1 - \bar{X}_2)_O$, il suffit de considérer la somme des données dans le groupe 1, où $n_1 \leq n_2$.[2] Par l'énumération des ${}_nC_{n1}$ permutations des $\{x_i\}$ entre les groupes, on calcule à chaque fois la somme des n_1 premières valeurs. Il faut noter que, si $n_1 = n_2$, chaque différence $(\bar{X}_1 - \bar{X}_2)$ se répétera en signe opposé, et qu'il n'y a en fait que ${}_nC_{n1}/2$ différences possibles ; cette même remarque et cette même procédure s'appliqueraient aussi à la différence des variances $(s_1^2 - s_2^2)$, à la différence des écarts-types $(s_1 - s_2)$, voire au quotient des variances (s_1^2/s_2^2), lequel se répète en valeur réciproque (s_2^2/s_1^2). S'il s'avère que, pour un test unilatéral, les n_1 plus fortes observations se retrouvent dans le groupe 1, la probabilité extrême de ce résultat est simplement $1 / {}_nC_{n1}$, et la différence sera dite significative si cette quantité est inférieure au seuil de probabilité convenu.

Pour réaliser ce test, il faut donc énumérer les combinaisons relatives au groupe 1. Soit α, un seuil de signification convenu : Ferland (1981, voir aussi Ferland et Laurencelle, 2012) décrit un algorithme qui engendre au plus $\alpha \times {}_nC_{n1}$ combinaisons seulement, réduisant d'un facteur $(1–\alpha)$ le temps d'exécution par rapport à un programme qui examinerait les combinaisons complètes (Nijenhuis et Wilf, 1985, p. 22). On trouvera dans Edgington (1980, p. 84) un algorithme de combinaisons aléatoires dans le groupe 1.

Test de l'existence d'un groupe taxonomique.

L'exemple suivant est tiré du manuel *Biometry* (Sokal et Rohlf, 1981, p. 790-791). Les auteurs rapportent un tableau de « coefficients de similitude » calculés entre 10 espèces comparées deux à deux. L'hypothèse posée concerne l'existence possible d'un groupe taxonomique, incluant trois espèces, disons A, B, C. La statistique envisagée est la suivante : la similitude interne est établie comme la moyenne des similitudes parmi les éléments du groupe pris deux à deux, et la similitude externe est la moyenne des coefficients de chaque élément du groupe avec ceux hors du groupe, la statistique étant la différence (Similitude interne) – (Similitude externe). Malgré son peu de sophistication mathématique, cet exemple décrit une situation où la mesure de base, le coefficient de similitude, n'a pas un comportement statistique connu, et où la statistique testée apparaît relativement complexe. La situation cependant se présente semblable à celle de l'exemple précédent : un groupe-cible, de 3 éléments, mesuré par rapport au groupe complémentaire de 7. Il y a ${}_{10}C_3 = 120$ regroupements différents possibles, et le regroupement (A, B, C) sera jugé significatif à 5%, c.-à-d. présentant un niveau de similitude exceptionnellement élevé, si la statistique observée obtient un rang de 6 (= 0,05 × 120) ou moins parmi les 120 statistiques engendrées.

Le nombre total de permutations

Le nombre de configurations différentes possibles varie selon le problème considéré et la taille des échantillons. Nous examinerons quelques catégories importantes de tests, puis nous présenterons un barème permettant par exemple de déterminer quels problèmes peuvent être réalistement traités par permutations complètes et quels doivent l'être par une méthode d'échantillonnage aléatoire.

Permutations ordinales.

Plusieurs tests statistiques peuvent être ramenés à un problème de permutation ordinale. Nous avons, plus haut, donné l'exemple de la corrélation des séries jumelées $\{X_i, Y_i\}$. Nous pouvons considérer aussi le test des séquences dichotomiques ou polytomiques (*runs test*) et celui des séquences monotones (*runs up and down test* : voir Bradley, 1968); la variance permutative (Laurencelle, 1983; von Neumann et coll., 1941); le spectre de puissance dans l'analyse de Fourier (Barnard, 1963), etc. Comme il y a $n!$ configurations découlant d'un échantillon de n éléments, le nombre d'éléments maximal pour ne pas excéder, disons, 10 000 configurations est $n(10\,000) = 7$, puisque $7! = 5\,040 < 10\,000$ et $8! = 40\,320 > 10\,000$. Pour ne pas dépasser 1 000 000, on aura $n(1\,000\,000) = 9$.

Combinaisons en k groupes.

Les tests les plus connus en statistique se retrouvent dans la catégorie des combinaisons, qu'il s'agisse de la comparaison de deux ou plusieurs moyennes indépendantes, de deux variances indépendantes, ou d'une quelconque statistique basée sur chacun de k groupes. Si l'on considère deux groupes égaux seulement, c.-à-d. $n_1 = n_2 = n$, le nombre de configurations est ${}_{2n}C_n$, et $n(10\,000) = 7$, $n(1\,000\,000) = 11$. La généralisation à k groupes égaux peut se noter ${}_{kn}C_{n(k)} = (k \cdot n)! / (n!)^k$. Par exemple, avec $k = 4$ groupes, $n(10\,000) = 2^3$, $n(1\,000\,000) = 3$. Plusieurs statistiques applicables avec 2 groupes ou plus sont symétriques, en ce

[2] Cette somme, de calcul plus simple, a même distribution que la différence de moyennes et même distribution ordinale que le test t applicable. Le fait de traiter le groupe le plus petit entraîne de plus une réduction des calculs, chaque somme étant établie par l'addition de seulement n_1 valeurs.

[3] Noter que, de toute façon, l'analyse de variance est problématique si l'on a moins que 2 observations par groupe. Avec une seule observation, on ne peut pas estimer la variance intragroupe, ou variance d'erreur, ce qui empêche la construction du quotient F.

sens que l'ordre des groupes n'y joue pas : pensons en particulier au Carré moyen des groupes, qui apparaît au numérateur du quotient *F* en analyse de variance. Pour ces cas, les configurations ne différant que par l'ordre des *k* groupes peuvent être ramenées à un seul exemplaire, d'où le nombre de configurations engendrées devient ${}_{kn}C_{n(k)} / k! = (k{\cdot}n)! / \{(n!)^k \cdot k!\}$, soit considérablement moindre. Edgington (1980, p. 67 seq.) propose un programme Fortran qui engendre seulement les configurations voulues.

Le cas des échantillons inégaux donne lieu, pour 2 groupes, à la formule ${}_{n1+n2}C_{n1}$, et pour *k* groupes, ${}_{n1+n2+\ldots+nk}C_{n1,n2,\ldots,nk} = (\Sigma_j n_j)! / \Pi_j n_j!$. Dans le cas des statistiques symétriques mentionnées précédemment, il faut considérer que les configurations engendrées dans des groupes mutuellement inégaux sont forcément distinctes. Cependant, dans *k* groupes, il peut s'en trouver 2 ou plusieurs qui soient de tailles égales, d'où la formule générale du nombre de configurations distinctes est :

$$\frac{\left(\sum_j n_j\right)!}{\prod_j n_j!} \times \frac{1}{\prod_r f_r!},$$

où f_r indique le nombre de groupes de taille *r*. Par exemple, avec *k* = 6 groupes de tailles respectives 1, 1, 2, 2, 2, 3, le nombre global de configurations est 11! / (1!1!2!2!2!3!) = 831 600, et le nombre de configurations distinctes est un nombre réduit, puisqu'il y a deux ensembles de groupes égaux, $f_1 = 2$, $f_2 = 3$ et $f_3 = 1$, d'où 831 600 / (2!3!1!) = 69 300. Reste le problème de formuler une méthode d'énumération efficace, soit une méthode qui produise seulement les configurations voulues.

Permutations binaires.

La version permutationnelle du test de différence entre deux moyennes jumelées (ou « pairées ») consiste à permuter le signe des *n* différences observées, le nombre de configurations étant 2^n. On a donc ici $n(10\,000) = 13$ et $n(1\,000\,000) = 19$. Laurencelle (1979) propose une méthode d'énumération efficace de ces permutations, nécessitant seulement un total de $2^n - 1$ (plutôt que $n{\cdot}2^n$) permutations de signes.

Permutations ordinales croisées.

Correspondant aux plans d'analyse de variance à *k* groupes indépendants, on peut définir des plans à *k* mesures répétées pour un groupe de *n* sujets, ou sources échantillonnales : le cas *k* = 2 correspond aux permutations binaires. Le nombre de permutations étant *k*! pour chacune des *n* sources, on a $(k!)^n$ configurations différentes. Si l'ordre des conditions n'apparaît pas dans la statistique, comme ce serait le cas pour une analyse en polynômes linéaires orthogonaux (*trend analysis*), on peut considérer que l'arrangement de la première source, ou premier sujet, est fixé, et le nombre de configurations distinctes devient alors $(k!)^{n-1}$. Pour *k* = 4 encore, dans le premier cas on obtient $n(10\,000) = 2$, $n(1\,000\,000) = 4$; dans le second cas, $n(10\,000) = 3$, $n(1\,000\,000) = 5$, soit un de plus que dans le premier cas.

Plans mixtes d'analyse de variance.

Le calcul du nombre de configurations se complique dans le cas de statistiques plus élaborées. Nous prendrons l'exemple d'un plan d'analyse à deux dimensions, de design A × B$_R$, c.-à-d. avec *p* groupes de *n* éléments chacun et *k* mesures répétées sur chaque élément. Ce plan d'analyse donne lieu, typiquement, à 3 tests *F* basés sur 5 Carrés moyens[4], soit *F*(A/S), *F*(B/BS) et *F*(A×B/BS), A désignant le Carré moyen des *p* groupes, S celui intragroupe, B celui des *q* occasions de mesure, A×B celui de l'interaction « groupes × occasions de mesure ». Soit, pour illustration, *p* = 3, *q* = 2 et *n* = 5. Les nombres de configurations différentes pour les Carrés moyens de A et S sont $T(A) = T(S) = (p{\cdot}n)! / \{n!^p{\cdot}p!\} = 126\,126$. Pour B, $T(B) = q!^{p\cdot n-1}$, comme on a vu précédemment, ici T(B) = 16 384. Pour le Carré moyen d'interaction A×B, sa valeur dépend à la fois de la composition des groupes et de la variation des moyennes dans chaque groupe. Retenant le premier sujet dans le premier groupe, on considère d'abord les compositions différentes des *p* groupes (avec *n*–1 sujets dans le premier et *n* dans les autres groupes), l'ordre entre les *p*–1 derniers groupes étant indifférent. À cela, il faut combiner les *q*! permutations pour les $p{\cdot}n-1$ derniers sujets. Cette analyse produit $T(A{\times}B) = (p{\cdot}n-1)! / \{n!^{p-1}(n-1)!(p-1)!\} \times q!^{p\cdot n-1}$, ici 2 066 448 384.[5] Enfin, pour le Carré moyen BS, il faut, pour chaque configuration des groupes, permuter *n*–1 lignes de données dans chaque groupe, d'où $T(BS) = T(A){\times}(q!^{n-1})^p$, ici 516 612 096.

Le tableau 1 donne des indications sur la taille échantillonnale possible selon le nombre maximal consenti de configurations à analyser, ce pour quelques catégories de problèmes. À cause de ses applications nombreuses et pour illustrer les calculs, nous avons considéré aussi la comparaison de deux groupes inégaux. Selon divers paliers du nombre de configurations à analyser, le Tableau 2

---

[4] Les tests permutationnels, à proprement parler, doivent être appliqués aux numérateurs des quotients *F*, ici les Carrés moyens A, B et A×B, plutôt qu'aux quotients *F* eux-mêmes. Noter qu'en général les distributions ordinales d'un quotient et de son numérateur ne sont pas correspondantes. Les quotients *F* spécifiés ici correspondent à un modèle d'analyse à effets déterminés (Winer, 1971).

[5] Cette formule dénombrant les configurations qui produisent des valeurs distinctes des Carrés moyens A×B et BS est donnée sous toutes réserves. Noter que T(A×B) = T(A) × T(B).

Tableau 1. Taille maximale des échantillons selon le nombre maximal de configurations permises et le type de problème

		Nombre maximal de configurations permises				
Problème*		10^3	10^4	10^5	10^6	10^7
A $n!$		6	7	8	9	10
B $(\Sigma_j n_j)! / \Pi_j n_j!$	$k = 2$	6	7	9	11	12
	3	2	3	4	5	5
	4	-	2	2	3	3
	5	-	-	-	2	2
	6	-	-	-	-	2
C $(\Sigma_j n_j)!/\{\Pi_j n_j! \times k!\}$	2	6	8	10	11	12
	3	3	4	4	5	6
	4	2	2	3	3	4
	5	2	2	2	2	3
	6	-	-	2	2	2
	7	-	-	-	2	2
	8	-	-	-	2	2
D $(k!)^n$	2	9	13	16	19	23
	3	3	5	6	7	7
	4	2	2	3	4	5
	5	-	-	2	2	3
	6	-	-	-	2	2

*A. Permutations ordinales B. Combinaisons en k groupes égaux ; le tableau donne n, le nombre d'éléments par groupe C. Combinaisons réduites en k groupes égaux, l'ordre des groupes n'étant pas considéré D. Permutations ordinales de n vecteurs chacun ; la ligne $k = 2$ représente aussi les permutations binaires. Voir texte.

rapporte la taille maximale du second groupe pour toutes les tailles possibles du premier groupe. Dans les deux tableaux, les inscriptions faites en regard de 10 000 000 configurations sont présentées surtout pour illustrer l'impossibilité pratique des tests à permutations complètes, sauf dans les cas où les échantillons sont passablement petits.

Le nombre suffisant de permutations

Les considérations précédentes montrent que les tests à permutations complètes sont impraticables dans la plupart des cas, même en utilisant l'ordinateur. On peut recourir alors aux tests à permutations aléatoires, en explorant seulement un nombre réaliste de configurations. Soit $S_{n,\alpha}$ une statistique basée sur n permutations d'éléments échantillonnaux et telle que $\alpha \cdot n$ statistiques permutationnelles ou moins lui soient supérieures.[6] La statistique $S_{n,\alpha}$ approche la valeur paramétrique s_α quand n croît, de même que la fonction de répartition $P(S_{n,\alpha})$ approche $P(s_\alpha) = 1 - \alpha$. Cet argument de Hoeffding (1952) indique que la distribution empirique issue de configurations aléatoires tend vers la distribution échantillonnale (mathématiquement) exacte de la statistique étudiée. Nous sommes alors placé devant une alternative : d'une part, la production d'un grand nombre de configurations permet d'accroître la précision et, vraisemblablement, la puissance du test, mais nous devons d'autre part garder le coût des calculs sous un seuil

[6] Pour un test unilatéral du côté positif de la distribution. Un raisonnement similaire vaut pour un test unilatéral négatif ou pour un test bilatéral.

Tableau 2. Taille maximale de second groupe (n_2) selon le nombre de configurations permises et la taille du premier groupe (n_1), pour la comparaison de deux groupes indépendants

	Nombre maximal de configurations permises				
n_1	10^3	10^4	10^5	10^6	10^7
2	43	139	445	1412	4471
3	16	37	82	179	389
4	9	19	36	67	121
5	7	13	23	38	62
6	6	10	16	26	40
7	-	8	13	20	29
8	-	-	11	16	23
9	-	-	10	14	19
10	-	-	-	12	17
11	-	-	-	11	15
12	-	-	-	-	14

* On suppose $n_1 \leq n_2$. Noter que, pour $n_1 = 1$, n_2 = (Nombre de configurations permises) – 1.

raisonnable, plafonnant ainsi le nombre de configurations aléatoires à produire.[7]

Une question se pose immédiatement : doit-on échantillonner avec ou sans remise, dans l'ensemble fini des configurations possibles? Raj et Khamis (1958) et, plus récemment, Gabler (1985) montrent que, sous certaines conditions assez larges, l'échantillonnage sans remise est plus efficace que celui avec remise. La comparaison s'applique par exemple à la moyenne, $\bar{X}$, et donne lieu à l'inégalité suivante :

$$\mathrm{Var}(\bar{X}; n') \leq \mathrm{Var}(\bar{X}; n),$$

où $\mathrm{Var}(\bar{X}; n')$ désigne la variance d'une moyenne calculée à partir des seuls n' éléments échantillonnaux distincts, alors que $\mathrm{Var}(\bar{X}; n)$ concerne tous les n éléments de la série. Cette inégalité s'applique lorsqu'on échantillonne au hasard n éléments : la moyenne établie sur tous les éléments est moins précise que celle dont on a exclu les répétitions. L'inégalité s'applique aussi lorsqu'on échantillonne jusqu'à obtenir n' éléments distincts : la moyenne obtenue à partir de ces n' éléments est plus précise que si on lui ajoute les répétitions. Nonobstant cet avantage de l'échantillonnage sans remise, sa mise en œuvre est plus lourde car elle suppose qu'on vérifie à chaque nouvelle configuration formée au hasard si elle fait ou non partie de l'ensemble déjà formé.

En outre, le taux de répétitions d'éléments dépend des tailles respectives de l'échantillon et de la « population ». Le nombre n' d'éléments distincts dans un échantillon de taille n, pigé avec remise dans une population (rectangulaire discrète) de taille N, est (Johnson et Kotz, 1969, p. 251-253) :

$$E(n') = N\cdot[1 - (1 - 1/N)^n] .$$

Le nombre de configurations envisagées, n, est généralement minime par rapport au nombre possible N : dans ce cas, les quotients $(n/N)^p$, $p \geq 2$, tendent vers zéro, et :

$$\begin{aligned} E(n') &= N[1 - (1 - n(1/N) + {}_nC_2(1/N)^2 - {}_nC_3(1/N)^3 + \ldots)] \\ &\approx N[1 - (1 - n/N)] \\ &\approx n , \end{aligned}$$

de sorte que le taux de répétitions escompté est presque nul. Par exemple, pour comparer deux groupes de 10 éléments chacun, le nombre total de combinaisons est ${}_{20}C_{10} = 184\,576$. Si l'on pige $n = 1\,000$ configurations au hasard, $E(n') \approx 997{,}3$ et le taux de répétitions moyen est ici de 0,3%, un résultat pareil apparaissant pour $n = 10\,000$. Autant d'arguments pour se satisfaire d'un échantillonnage avec remise, c.-à-d. sans vérification.

Nous allons considérer ici les tests basés sur une taille échantillonnale calculée comme un multiple de la taille minimale requise, soit :

$$n = k / \alpha - 1 , \quad k \text{ entier} .$$

[7] Kahn et Marshall (1953) et Laurencelle (2001) proposent des stratégies afin d'accroître la précision ou de réduire la taille des échantillons dans les études de type Monte Carlo. Toutefois, ces stratégies ne s'appliquent pas de façon générale au présent problème des tests de signification statistique.

Les tests de cette taille sont exacts sous l'hypothèse nulle H_0 en ce sens que, si H_0 est vraie, la statistique sera déclarée significative dans une proportion α (Hope, 1968 ; Edgington, 1969). Ainsi, pour $k = 1$, la statistique observée sera dite significative si elle est de rang 1, c.-à-d. la plus grande (ou la plus petite) parmi les $n + 1$ statistiques considérées ; pour $k = 2$, elle devra être de rang 1 ou 2, ou généralement avoir l'un des rangs 1, 2, …, k. Quant à la précision ou la puissance de tels tests, ces caractéristiques dépendent à la fois de n ou k, et du modèle asymptotique auquel on les intègre, soit un modèle paramétrique ou un modèle permutationnel.

Modèle paramétrique.

La référence à un modèle paramétrique comporte certains avantages. Prenons l'exemple du test de différence entre deux moyennes indépendantes, pour des populations à variances inconnues et supposément égales. La référence à un modèle paramétrique consiste ici à poser que les éléments sont distribués normalement, avec variances égales, et que sous H_0 la différence standardisée entre les moyennes, $(\bar{X}_1 - \bar{X}_2)/s$, suit la loi t ayant $(n_1 + n_2 - 2)$ degrés de liberté. Le critère de rejet de l'hypothèse nulle repose sur une borne fixe, soit $t_{n1+n2-2[1-\alpha]}$, le centile $100(1 - \alpha)$ de la loi t appropriée pour un test unilatéral. C'est dans un contexte pareil que Hoeffding (1952), puis Hope (1968) et Marriott (1979) tentent de documenter la puissance des tests à permutations aléatoires. Prenant le modèle de la loi normale, Hope (1968) établit que les valeurs $k = 1$ ou $k = 2$ fournissent une efficacité raisonnable. Marriott (1979) conteste ces petites valeurs de k. D'abord, le recours à un test permutationnel est souvent motivé par le caractère problématique de la variable analysée : la puissance associée à une valeur k, $P(k)$, sera moindre sur un modèle t à 3 degrés de liberté que sur un modèle à 10 degrés de liberté ou un modèle normal. C'est dire que, dans la pratique des tests permutationnels, les conclusions de Hope peuvent être inapplicables. L'autre argument de Marriott concerne la probabilité de rejet sous un test permutationnel. Dans un modèle paramétrique, la valeur critique, ou borne, est fixe, disons $t_{[1-\alpha]}$: toute statistique dont la probabilité extrême sous H_0 est inférieure à α sera déclarée significative avec probabilité 1, et toute autre dont la probabilité extrême est supérieure à α sera dite significative avec probabilité 0. Dans un test permutationnel de taille $n = k / \alpha - 1$, la probabilité de rejeter H_0 sera plutôt :

$$P_t = \sum_{i=0}^{k-1} \binom{n-1}{i} (1 - p_t)^{n-i-1} p_t^i ,$$

où p_t est la probabilité extrême de la statistique t. Le calcul montre que, pour $p_t < \alpha$, P_t tend vers 1 lorsque k croît, comme P_t tend vers 0 pour $p_t > \alpha$. La borne de rejet est donc floue, ce qui entraîne une perte de puissance d'ailleurs difficile à établir.[8] Marriott, à l'instar de Besag et Diggle (1977), recommande la valeur $k = 5$ à toutes fins pratiques.

La référence à un modèle paramétrique permet de mieux cerner la notion de borne floue évoquée par Marriott (1979), en nous donnant le moyen de déterminer la précision de la borne permutationnelle. Supposons en effet qu'au lieu d'être prises dans l'ensemble fini des permutations d'un échantillon, les configurations sont formées à neuf à partir d'une population infinie d'éléments. La statistique évaluée observe alors une loi de distribution échantillonnale spécifique, et on peut en évaluer certaines caractéristiques. Prenons la statistique t_{10}, un écart-réduit t à 10 degrés de liberté. Si l'on tire au hasard n valeurs de cette statistique, on peut déterminer la borne empirique $t'_{[1-\alpha]}$ par la statistique d'ordre $t_{(r)}$, soit :

$$t'_{[1-\alpha]} = t_{(r)} \text{, telle que } \text{Prob}\{ t' \geq t_{(r)} \} \leq \alpha \text{ ,} \qquad \text{Déf. 1}$$
$$r = \lceil n \cdot (1-\alpha) + \tfrac{1}{2} \rceil ,$$

où la fonction $\lceil x \rceil$ dénote l'entier égal ou immédiatement supérieur à x. Une autre méthode consiste à interpoler la borne $t'_{[1-\alpha]}$ entre deux statistiques d'ordre, de façon à ce que la surface extrême soit approximativement α.[9] Ainsi :

$$t'_{[1-\alpha]} = (1 - u + s) \cdot t_{(s)} + (u - s) \cdot t_{(s+1)} , \qquad \text{Déf. 2}$$
$$\text{telle que } \text{Prob}\{ t' \geq (1 - u + s) \cdot t_{(s)} + (u - s) \cdot t_{(s+1)} \} \approx \alpha ,$$
$$u = n \cdot (1-\alpha) + \tfrac{1}{2} \text{ , } s = \lfloor u \rfloor ,$$

où la fonction $\lfloor x \rfloor$ dénote la partie entière de x. Cette borne est telle que la proportion de valeurs qui lui sont supérieures est environ α.

La précision de la borne empirique par rapport à la borne paramétrique peut alors être évaluée par le REQM, la racine carrée de l'écart quadratique moyen, soit :

$$\text{REQM}(t'_{[1-\alpha]}) = \{ \text{E}(t'_{[1-\alpha]} - t_{[1-\alpha]})^2 \}^{1/2}$$
$$= \{ [\text{E}(t'_{[1-\alpha]}) - t_{[1-\alpha]}]^2 + \text{var}(t'_{[1-\alpha]}) \}^{1/2} ,$$

expression qui est fonction du biais et de la variance de

---

[8] Cette approche se heurte aussi à une difficulté conceptuelle. Dans un modèle paramétrique, en effet, la puissance s'établit seulement lorsque H_0 est fausse, et la probabilité extrême p_t se calcule alors sur la distribution réelle de la statistique, une distribution autre que celle spécifiée par H_0. À strictement parler, les calculs de Marriott renseignent sur la capacité du test à commettre l'erreur de type I dans une proportion correcte α.

[9] Le principe d'une borne interpolée repose ici sur la distribution rectangulaire des statistiques d'ordre empiriques. Chacune a un poids estimé de $1/n$, et la borne exacte doit définir une somme des poids à gauche égale à $1–\alpha$. Ne pas confondre cette « borne exacte » avec la procédure de test au seuil α exact, réalisée en général à partir d'un nombre aléatoire auxiliaire (voir p. ex. Hogg et Craig, 1978, p. 254-256.).

$t'_{\lceil 1-\alpha \rceil}$. Pour des valeurs asymptotiques de n,[10] le biais tend vers 0 ; quant à la variance de la statistique d'ordre relative à la Déf. 1 ci-dessus, elle s'exprime (Mood, Graybill et Boes, 1974, p. 257) comme :

$$\text{var}(t'_{\lceil 1-\alpha \rceil}) \approx \frac{\alpha(1-\alpha)}{n \cdot f^2_{t_{1-\alpha}}}, \qquad \text{selon Déf. 1}$$

$f_{t_{1-\alpha}}$ étant la densité de la loi distributionnelle en référence à l'abscisse $t_{1-\alpha}$. Le calcul de précision de la borne permutationnelle, pour n forts, se ramène donc à :

$$\text{REQM}(t'_{\lceil 1-\alpha \rceil}) = \sqrt{\frac{\alpha(1-\alpha)}{n}} \cdot f^{-1}_{t_{1-\alpha}}.$$

Par exemple, la borne à 5% de t_{10} est $t_{10[0,95]} \approx 1{,}812$. La densité de t_{10} à cette abscisse (Mood et coll., 1974, p. 250) est 0,081636, et la précision est donnée par :

$$\text{REQM}(t'_{[0,95]}) = \frac{2,66972}{\sqrt{n}}.$$

D'après cette formule, une précision de 0,1 s'obtient au prix de n = 713 échantillons, une de 0,01 au prix de n = 71 275. Si nous connaissons la variance paramétrique et utilisons l'écart-réduit normal plutôt que t_{10}, la borne $z_{[0,95]}$ est 1,645, la densité à cette borne f = 0,103111, et la précision :

$$\text{REQM}(z_{[0,95]}) = \frac{2,11371}{\sqrt{n}},$$

ce qui exige n = 447 pour une précision de 0,1 et n = 44 678 pour 0,01. La variance de la borne interpolée, selon notre Déf. 2 ci-dessus, est approchée par :

$$\begin{aligned} \text{var}(t'_{\lceil 0,95 \rceil}) &= a^2 \cdot \text{var}(t'_{(s)}) + b^2 \cdot \text{var}(t'_{(s+1)}) + \\ &\qquad 2a \cdot b \cdot \text{cov}(t'_{(s)}, t'_{(s+1)}) \\ &\approx \frac{\alpha(1-\alpha)}{n \cdot f^2_{t_{1-\alpha}}}, \qquad \text{selon Déf. 2} \end{aligned}$$

$$a = 1 - b = 1 - u - s,$$

puisque, dans un contexte asymptotique, les statistiques d'ordre voisines $t'_{(s)}$ et $t'_{(s+1)}$ ont même variance et une corrélation parfaite. Pour des valeurs sous-asymptotiques de n, l'erreur dans ces formules devient appréciable, et mieux vaut alors exploiter les approximations plus élaborées de David et Johnson (1954).

Bref, dans le modèle paramétrique, la borne utilisée pour tester la signification de notre statistique est fixe, avec une variance nulle et une précision parfaite. Par contraste, la borne permutationnelle occupe sur l'abscisse une zone floue, d'une largeur inversement proportionnelle à la racine carrée du nombre d'échantillons. Enfin, l'applicabilité d'un modèle paramétrique est douteuse dans le présent contexte : l'évaluation de précision n'est juste que si la variable observée se conforme au modèle stipulé, auquel cas le chercheur, s'il le savait, renoncerait au test permutationnel! La plupart du temps, la variable observée n'est pas catégorisable sûrement ou elle apparaît atypique, et alors la référence à un modèle paramétrique est fautive puisque le test paramétrique a, dans ces circonstances, une puissance réelle inférieure à sa puissance nominale.

Modèle permutationnel.

Dans un modèle strictement permutationnel, nous considérons que les configurations étudiées (n) sont un échantillon représentatif de la population des configurations possibles (N). Une statistique s_o sera dite significative au seuil α si elle occupe les 100α rangs centiles supérieurs de la distribution des N statistiques, c.-à-d. si $s_o \geq s_{(r)}$, $r = N - \lfloor \alpha\, N \rfloor$, où $s_{(i)}$ désigne la i^e statistique d'ordre parmi les N possibles : le cas échéant, l'hypothèse nulle est réputée fausse. Aussi, le domaine de la variable étudiée se restreint aux valeurs ponctuelles constituant l'échantillon original, chacune ayant une « probabilité égale ». Le test permutationnel, basé sur n configurations aléatoires, consiste à établir une borne approximative $s'_{(r)}$, $r = n - \lfloor \alpha \cdot n \rfloor$, et vérifier si la statistique observée dépasse ou non cette borne, c.-à-d. si $s_o \geq s'_{(r)}$.

Le test à permutations complètes définit ici les 100% de puissance[11] et la précision maximale : il s'agit alors d'évaluer la puissance et la précision du test à permutations aléatoires en les comparant aux premières. Ce type d'évaluation concorde avec la motivation et la logique des tests permutationnels, puisqu'on ne s'aventure pas hors du territoire des données observées et qu'on invoque seulement le principe de randomisation, ou répartition aléatoire, déjà familier aux chercheurs. C'est en tout cas l'avis de Bradley (1968) et de Edgington (1973), et Fisher (1936, 1971), mise à part sa réserve sur l'incommodité des calculs, appuie la validité des tests paramétriques sur celle des tests permutationnels, plutôt que l'inverse.

La précision du test à permutations aléatoires pourrait s'évaluer en étudiant le comportement de la borne aléatoire $s'_{(r)}$ mentionnée plus haut. On peut par exemple établir un REQM, selon :

$$\text{REQM}(s'_{(r)}) = \{ E_C(s'_{(r)} - s_{(r)})^2 \}^{1/2},$$

l'espérance E_C étant basée sur les différentes collections contenant chacune n configurations, soit N^n pour un échantillonnage avec remise, et ${}_N C_n$ pour un échantillonnage sans remise. Dans ce dernier cas, le REQM tend régulièrement vers 0 et le rejoint lorsque n croît vers N. Pour des valeurs inférieures de n ou dans les cas plus pratiques d'échantillonnage avec remise, nous n'avons pas réussi à obtenir une expression générale pour estimer le REQM.

[10] C'est-à-dire p. ex. $1/\sqrt{n} \to 0$.

[11] Puisque la condition de départ est que la statistique observée (s_o) occupe la zone critique, dite significative, de la distribution des N configurations de cette statistique.

Tableau 3. Données de puissance d'un test permutationnel selon le nombre de configurations aléatoires produites ($k/\alpha - 1$), en fonction de k et du seuil α

	Puissance						k
α	1	4	8	16	64	128	$P \approx 0{,}99$
.05	0,64151	0,80959	0,86395	0,90330	0,95146	0,96565	1511
0,025	0,63677	0,80709	0,86217	0,90203	0,95082	0,96520	1551
0,01	0,63397	0,80561	0,86111	0,90128	0,95045	0,96494	1574
0,005	0,63304	0,80512	0,86076	0,90103	0,95032	0,96485	1582
0,001	0,63623	0,80473	0,86048	0,90083	0,95022	0,96478	1589

Quant à la puissance du test aléatoire, posons d'emblée que l'hypothèse nulle est fausse, c.-à-d. que la statistique observée s_o occupe les 100α rangs centiles supérieurs,[12] ou $s_o \geq s_{(r)}$. Dans ce cas, la puissance du test aléatoire correspond à la probabilité de rejeter H_0 en utilisant n configurations. Prenons une taille échantillonnale $n = k/\alpha$, $k = 1, 2, \ldots$, selon laquelle on compare s_o parmi les $n-1$ statistiques aléatoires, pigées avec remise. La puissance associée à une valeur de k, $P(k)$, est la somme des probabilités de rejet de H_0 selon la place de s_o parmi les $n-1$ autres statistiques. Pour $k = 1$ par exemple, s_o sera la plus forte si les $n-1$ autres statistiques font partie du sous-ensemble inférieur, avec probabilité $(1-\alpha)^{n-1}$; ou si $n-2$ proviennent du sous-ensemble inférieur, et 1 du sous-ensemble critique à condition que s_o soit plus forte que cette dernière, avec probabilité $(n-1)(1-\alpha)^{n-2}\cdot\alpha\cdot\frac{1}{2}$; ou si, en général, $(n-1-r)$ proviennent du sous-ensemble inférieur et r du sous-ensemble critique, à condition que s_o soit maximale parmi ces r statistiques, selon une probabilité :

$$\binom{n-1}{r}(1-\alpha)^{n-r-1}\,\alpha^r\cdot\frac{1}{r+1}.$$

Pour k général, si $r \geq k-1$ statistiques sont échantillonnées dans le sous-ensemble critique, s_o doit faire partie des k plus fortes valeurs, événement dont la probabilité est $k/(r+1)$. Ainsi, la puissance s'évalue par:[13]

$$P(k) = \sum_{r=0}^{n-1}\binom{n-1}{r}(1-\alpha)^{n-r-1}\,\alpha^r\cdot\left(\frac{k}{r+1}\right)_{k\leq r+1},$$

où $n = k/\alpha$.

Le tableau 3 présente le résultat des calculs par la formule ci-dessus, selon quelques valeurs du seuil α et certaines valeurs de k, de même que la valeur de k requise pour atteindre une puissance de 0,99. Le tableau indique que l'on peut obtenir une puissance satisfaisante, de l'ordre de 0,90, en fixant k à 16, ce qui correspond à seulement $n = 319$ configurations au seuil unilatéral de 0,05, ou à 1 599 au seuil de 0,01. La puissance quasi parfaite de 0,99 s'obtient au prix de tailles plus élevées, soit 30 219 (avec $k = 1\,511$) pour $\alpha = 0{,}05$ et 157 399 (avec $k = 1574$) pour $\alpha = 0{,}01$. Notons ici que la génération de quelques centaines de milliers de configurations aléatoires reste un objectif bien réaliste, compte tenu de l'efficacité temporelle des ordinateurs modernes.

Les calculs affichés au tableau 3, basés sur le modèle permutationnel, diffèrent de ceux de Hope (1968) ou même Marriott (1979), lesquels s'appuient sur d'autres hypothèses. Noter que l'évaluation de puissance pour un échantillonnage sans remise,[14] lequel serait beaucoup plus laborieux, donne à peu près les mêmes résultats pour N fort, ou n/N proche de 0, ce qui est le cas usuel. Sommairement, nous recommanderions d'utiliser à peu près $k = 64$, c.-à-d. $64/\alpha - 1$ configurations aléatoires afin d'assurer une efficacité relative d'au moins 95%.

Même si on se base sur les calculs inspirés du modèle permutationnel, les tests à permutations aléatoires

[12] Noter que le seuil réel α' est déterminé par $\alpha' = \lfloor N\,\alpha \rfloor / N \leq \alpha$. Pour N fort, la différence entre α' et α est négligeable, et nous l'avons ignorée.

[13] En fait, le cumul se fait en deux phases. La phase 1 consiste à considérer les cas où $0 \leq r \leq k-1$ statistiques s_i occupent la zone critique (de grandeur k, i.e. occupée par les k statistiques d'ordre extrêmes) , sans compromettre la significativité de s_o. Dans la phase 2, il y a $r \geq k$ valeurs s_i qui empiètent sur la zone critique, et alors la significativité sera obtenue seulement si s_o fait partie des k plus grandes valeurs dans cette zone, ce qui se produit avec probabilité $k/(r+1)$.

[14] La formule d'évaluation considère $N = N_1 + N_2$, où $N_1 = \lfloor \alpha N \rfloor$. Le nombre total d'échantillons de taille $n-1$ est $N^{(n-1)}$, où $x^{(k)}$ désigne une factorielle descendante, $x(x-1)(x-1)\ldots(x-k+1)$, et $x^{(0)} = 1$; p. ex. $x^{(x)} = x!$. Chaque échantillon est divisé selon qu'il a $(n-1-r)$ statistiques provenant du sous-ensemble inférieur, et r provenant du sous-ensemble critique, ce qui se produit de ${}_{n-1}C_r N_2^{(n-1-r)} N_1^{(r)}$ façons. D'où la formule :

$$P^*(k) = \binom{n-1}{2}\frac{N_1^{(r)}N_2^{(n-1-r)}}{N^{(n-1)}}\cdot\left(\frac{k}{r+1}\right)_{k\leq r+1}$$

demeurent une solution réaliste et efficace pour la plupart des situations. Ces tests sont aussi un choix prudent lorsque le chercheur ne peut garantir l'adéquation d'un modèle paramétrique dans sa situation : dans ce cas, le recours à d'autres tests dits non-paramétriques, consistant par exemple à convertir les données en rangs, à les dichotomiser ou à les transformer d'autres façons, a le double inconvénient de nuire à la limpidité de l'interprétation et de sacrifier peu ou prou la puissance disponible. Les tests permutationnels ont, quant à eux, une logique qui serre de près la logique du design expérimental, soit le principe de randomisation, et l'argument de probabilité qu'ils exploitent est simple et intuitif. Quant à leur puissance, elle approche assez facilement celle des tests paramétriques quand ceux-ci sont applicables, le domaine d'applications des tests permutationnels étant beaucoup plus grand (Bradley, 1968).

Références

Barnard, G. A. (1963). (Commentaire) *Journal of the Royal Statistical Society B*, *25*, 294.

Besag, J., Diggle, P. J. (1977). Simple Monte Carlo tests for spatial patterns. *Applied Statistics*, *26*, 327-333.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs (NJ) : Prentice-Hall.

David, F. N., Johnson, N. L. (1954). Statistical treatment of censored datat, Part I: Fundamental formulae. *Biometrika*, *41*, 228-240.

Edgington, E. S. (1969). Approximative randomization tests. *Journal of Psychology*, *72*, 143-149.

Edgington, E. S. (1973). The random-sampling assumption in "Comment on component-randomization tests". *Psychological Bulletin*, *80*, 84-85.

Edgington, E. S. (1980). *Randomization tests*. New York : Marcel Dekker.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *SIAM Monographs*, no. 38, 92 p.

Ferland, P. (1981). Épreuve distributionnelle sur deux moyennes indépendantes. Rapport inédit, 44 p. Trois-Rivières : Université du Québec à Trois-Rivières.

Ferland, P., Laurencelle, L. (2012). Un algorithme efficace pour la comparaison de deux moyennes indépendantes. *Tutorials in Quantitative Methods for Psychology*, *8*, 11-24.

Fisher, R. A. (1936). « The coefficient of racial likeness » and the future of craniometry. *Journal of the Anthropological Institute*, *66*, 57-63.

Fisher, R. A. (1971). *The design of experiments* (9e édition). New York : Hafner.

Gabler, S. (1985). On unequal probability sampling : sufficient conditions for the superiority of sampling without replacement. *Biometrika*, *71*, 171-175.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *Annals of Mathematical Statistics*, *23*, 169-192.

Hogg, R. V., Craig, A. T. (1978). *Introduction to mathematical statistics* (4e édition). New York: Macmillan.

Hope, A. C. A. (1968). A simplified Monte Carlo significance test procedure. *Journal of the Royal Statistical Society B*, *30*, 582-598.

Johnson, N. L., Kotz, S. (1969). *Distributions in statistics. Discrete distributions*. New York : Houghton Mifflin.

Kahn, H., Marshall, A. W. (1953). Methods of reducing sample size in Monte Carlo computations. *Journal of Operations Research of the Society of America*, *1*, 263-278.

Kendall, M. G., Stuart, A. (1979). *The advanced theory of statistics. Vol. 2, Inference and relationship* (4e edition). New York : Macmillan.

Laurencelle, L. (1983). La variance permutative. *Lettres Statistiques*, *7*, 22 p.

Laurencelle, L. (1978). Permutations complètes par rotations récursives. *Lettres Statistiques*, *4*, 11 p.

Laurencelle, L. (1979). Permutations binaires et combinaisons hypercomplètes. *Lettres Statistiques*, *5*, 16 p.

Laurencelle, L. (2001). *Hasard, nombres aléatoires et méthode Monte Carlo*. Québec : Presses de l'Université du Québec.

Lehmann, E. L. (1975). *Nonparametrics : statistical methods based on ranks*. San Francisco : Holden-Day.

Marriott, F. H. C. (1979). Barnard's Monte Carlo tests : how many simulations? *Applied Statistics*, *28*, 75-77.

Metropolis, N., Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *14*, 335-341.

Mood, A. M. Graybill, F. A., Boes, D. C. (1974). *Introduction to the Theory of Statistics* (3e édition) New York : McGraw-Hill.

Nijenhuis, A., Wilf, H. S. (1975). *Combinatorial algorithms*. New York : Academic Press.

Raj, D., Khamis, S. H. (1958). Some remarks on sampling with replacement. *Annals of Mathematical Statistics*, *29*, 550-557.

Siegel, S., Castellan, N. (1988). Nonparametric statistics for the behavioural sciences (2e édition). New York : McGraw-Hill.

Sokal, R. R., Rohlf, F. J. (1981). *Biometry* (2e édition). San Francisco : Freeman.

"Student », ou W. S. Gosset (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.

Winer, B. J. (1971). *Statistical principles in experimental design*. New York : McGraw-Hill.

Von Neumann, J., Kent, R. H., Bellinson, H. R., Hart, B. I. (1941). The mean square successive difference. *Annals of Mathematical Staitstics*, *12*, 153-162.