REVUE DE STATISTIQUE APPLIQUÉE

A. VESSEREAU

Les méthodes statistiques appliquées au test des caractères organoleptiques

Revue de statistique appliquée, tome 13, n° 3 (1965), p. 7-38

<http://www.numdam.org/item?id=RSA_1965__13_3_7_0>

LES MÉTHODES STATISTIQUES APPLIQUÉES AU TEST DES CARACTÈRES ORGANOLEPTIQUES (1)

A. VESSEREAU

Professeur à l'Institut de Statistique, Université de Paris

SOMMAIRE

---

(1) Conférence publiée dans les "Annales de la Nutrition et de l'Alimentation (1965 XIX N° 2). Reproduite avec l'autorisation du Centre National de la Recherche Scientifique.

INTRODUCTION

L'appréciation directe des caractères organoleptiques au moyen de tests exécutés par une ou plusieurs personnes, soulève des difficultés bien connues. Sauf dans le cas de comparaisons concernant des produits très différents, le jugement porté par un même sujet peut varier d'une épreuve à l'autre ; d'un sujet à l'autre le jugement peut ne pas être le même ; à ces sources de variations "intra" et "inter-sujets", s'ajoute la variabilité propre aux objets mêmes soumis au test.

L'intérêt des méthodes statistiques pour l'interprétation des résultats n'est guère contesté ; leur rôle dans l'organisation des expériences n'est pas moins important. Les statisticiens se refusent, à juste titre, à interpréter les résultats d'expériences dont ils n'ont pas établi, conseillé, ou approuvé le plan, et cela pour deux raisons : une expérience mal conçue peut être impuissante à révéler les différences que l'on cherche à mettre en évidence - ce qui est plus grave, elle peut conduire à des conclusions tout à fait erronées ; l'exemple le plus simple concerne la comparaison de deux termes A et B, où l'ordre de présentation peut influencer la réponse au même titre que la "vraie différence" entre A et B. Il va sans dire que, pour les situations les plus courantes, des plans simples peuvent être exécutés et interprétés sans qu'il soit besoin d'avoir recours à un statisticien qualifié : mais s'écarter des protocoles d'exécution et des règles d'interprétation tels qu'ils ont été exactement définis, présente des risques qu'il faut éviter de courir.

La bibliographie concernant les méthodes statistiques applicables ou susceptibles d'être appliquées au test des caractères organoleptiques, surtout d'origine anglo-américaine, est très importante. On peut y distinguer :

- les articles qui traitent, d'un point de vue théorique, de méthodes générales dont le champ d'application contient le domaine particulier qui nous concerne ici : on les trouvera notamment dans *Biometrika, Annals of Mathematical Statistics, Journal of American Statistical Association, Journal of the Royal Statistical Society* : en fait il n'est guère de méthode statistique dont l'utilisation pour les tests sensoriels ne puisse être envisagée ;

- les articles dans lesquels la partie théorique, plus ou moins développée, sert de base à des applications concrètes ; beaucoup de ceux-ci ont été publiés au cours des dix dernières années, dans *Biometrics* ;

- enfin les publications qui portent sur des sujets particuliers, que l'on trouvera notamment dans *Food Research* et *Food Technology*.

L'interprétation statistique a pour ambition d'étendre les résultats d'une expérience particulière à un ensemble plus général : ce sera l'ensemble des répétitions de la même épreuve, susceptibles d'être effec-

tuées, dans les mêmes conditions, par un sujet ou un groupe de sujets - ou l'ensemble des résultats que l'on obtiendrait sur une "population" de sujets dont le groupe exécutant l'expérience constitue un "échantillon" valable. Mais cette généralisation ne peut pas être faite avec certitude ; l'énoncé des conclusions comporte des risques d'erreur - celles-ci doivent être formulées en termes probabilistes.

On rappellera rapidement, pour éviter d'avoir à y revenir par la suite, les deux moyens principaux utilisés pour l'interprétation statistique :

La notion d'estimation et celle d'intervalle de confiance. Les résultats expérimentaux permettent de calculer la valeur d'un -ou de plusieurs- paramètres tels que : proportion de préférences pour l'un des termes comparés, note moyenne attribuée à un produit, etc. Cette valeur numérique constitue une *estimation* de la *valeur théorique* qui caractérise l'ensemble auquel les observations sont rattachées en tant qu'échantillon. En même temps que l'estimation on calcule un *intervalle de confiance* attaché à celle-ci : intervalle tel qu'on ait une quasi certitude de ne pas se tromper en affirmant qu'il contient la valeur théorique. Cette "quasi certitude" est mesurée par une probabilité $(1 - \alpha)$ voisine de 1, que l'on appelle le "niveau de confiance" (on prend très souvent une probabilité égale à 0,95, ou "95 chances sur 100"). La probabilité complémentaire (α), par exemple 0,05 ("5 chances sur 100") est le risque d'erreur attaché à l'intervalle de confiance.

La notion de test statistique. Assez malencontreusement le même terme de "test" s'applique à l'épreuve des caractères sensoriels et à son interprétation statistique. On évitera les confusions en utilisant, pour ce deuxième sens du même mot, l'expression de "test statistique", ou "test d'hypothèse".

Le test statistique prend d'abord en considération une hypothèse H_o, appelée "hypothèse nulle", qui, sommairement dite, est que "les résultats expérimentaux peuvent s'expliquer par le seul fait du hasard". Par exemple, lorsque deux boissons dont l'une est plus sucrée que l'autre sont présentées à un dégustateur, celui-ci, s'il ne perçoit pas la différence, a une chance sur deux de les classer correctement. L'hypothèse H_o est ici définie par la probabilité $p_o = 1/2$; si une seule épreuve est faite, et quel qu'en soit le résultat, il n'y a pas de raison valable de la rejeter. Mais si dans cinq épreuves indépendantes le dégustateur donne cinq classements corrects, on est naturellement conduit à rejeter l'hypothèse, et à conclure que la différence a été perçue ; en effet, dans le cas de classements "au hasard", l'évènement observé aurait moins de 5 chances sur 100 de se produire $\left(\text{exactement } \left(\frac{1}{2}\right)^5 = \frac{1}{32} = 0,03\right)$.

Au test statistique est donc attaché un premier risque, dit "risque de 1ère espèce", qui est la probabilité de rejeter l'hypothèse H_o bien que celle-ci soit vraie. Dans l'exemple précédent, si l'on choisit pour cette probabilité la valeur $\alpha = 0,05$, on est conduit à ne rejeter l'hypothèse que si cinq classements indépendants ont été corrects. Cependant, en adoptant cette règle, il arrivera 3 fois sur 100 que des classements purement aléatoires feront conclure -à tort- que la différence a été perçue.

On posera donc .

Risque de 1ère espèce :

$$\alpha = \text{Pr. [rejeter } H_0 / H_0 \text{ vraie]}$$

Lorsque l'hypothèse H_0 n'est pas rejetée, on dit que les résultats ne sont pas "significatifs" ; α est le "seuil de signification" pour lequel on adopte souvent, mais pas obligatoirement, la valeur 0,05 (5 chances sur 100) ou, si α ne peut prendre que des valeurs discrètes, la valeur immédiatement inférieure à 0,05.

Le fait que les résultats ne permettent pas de rejeter -au seuil de signification choisi- l'hypothèse H_0 ne prouve pas que celle-ci soit vraie. Reprenant l'exemple des deux boissons sucrées, considérons la règle de décision suivante :

Le dégustateur effectue cinq classements ; si tous sont corrects, on rejette l'hypothèse $H_0\left(p_0 = \frac{1}{2} = \frac{5}{10}\right)$; dans le cas contraire, on accepte cette hypothèse. Le risque de 1ère espèce est :

$$\alpha = \Pr\left[\text{ rejeter } p_0 = \frac{5}{10} \middle/ p_0 = \frac{5}{10}\right] = \left(\frac{5}{10}\right)^5 = 0,03$$

Supposons que le dégustateur ait une "certaine sensibilité" qui lui permette de classer correctement les boissons en moyenne 6 fois sur 10. L'hypothèse vraie n'est pas l'hypothèse $H_0\left(p_0 = \frac{5}{10}\right)$, mais l'hypothèse $H_1\left(p_1 = \frac{6}{10}\right)$. Conformément à la règle fixée, l'hypothèse H_0 sera acceptée si le dégustateur ne répond pas correctement dans les 5 épreuves -évènement dont la probabilité est $1 - \left(\frac{6}{10}\right)^5 = 0,92$.

Ainsi il y a 92 chances sur 100 pour que la règle adoptée ne révèle pas la sensibilité d'un dégustateur caractérisé par

$$p_1 = \frac{6}{10}$$

Le risque de 2e espèce associé à l'hypothèse alternative p_1 est :

$$\beta = \Pr.\left[\text{ accepter } p_0 = \frac{5}{10} \middle/ p_1 = \frac{6}{10}\right] = 0,92$$

Le tableau ci-après donne les valeurs de β associées à différentes valeurs de p.

p	$\beta(p)$
0,5.....	0,97
0,6.....	0,92
0,7.....	0,83
0,8.....	0,67
0,9.....	0,41
1,0.....	0,00

D'une façon générale on posera :

Risque de 2ème espèce.

$$\beta = \text{Pr. [accepter } H_o/H \text{ vraie]}$$

Alors que le risque de 1ère espèce a une valeur fixe choisie à l'avance, le risque de 2e espèce dépend de l'alternative $H \neq H_o$. Il diminue au fur et à mesure que cette alternative "s'éloigne" de l'hypothèse nulle H_o. La relation entre β et H (dans le cas actuel entre β et p) définit la "fonction", ou la "courbe d'efficacité" du test statistique, c'est-à-dire en fait, de la règle de décision adoptée. La quantité $1 - \beta$ est la "puissance" du test pour l'alternative H. Pour une hypothèse H_o donnée, et un seuil de signification α également donné, il y a intérêt à choisir un test statistique aussi puissant que possible à l'égard d'une alternative H_o à laquelle on s'intéresse particulièrement et, si un tel test existe, le test "le plus puissant" à l'égard de toute alternative. La notion de puissance du test intervient dans le choix d'un plan d'expérience ; comme la puissance augmente avec le nombre d'épreuves (ou de répétitions) l'adoption d'un plan permettant un test plus puissant amène une économie dans le nombre d'épreuves. L'"efficacité relative" de deux plans P_1 et P_2 peut se définir comme le rapport n_1/n_2 (généralement exprimé en p. 100) des nombres de répétitions pour lesquels (α, β) se trouvent avoir les mêmes valeurs.

I - GENERALITES SUR LES OBJECTIFS ET LES METHODES

A - Les objectifs.

Dans le domaine des applications pratiques, les principaux problèmes qui se posent sont les suivants :

1/ *Le contrôle de qualité.*

Il s'agit de s'assurer que la qualité d'un produit reste constante. Les épreuves auront donc pour but de rechercher si une *différence* perceptible existe ; elles seront généralement exécutées par un petit nombre "d'experts", auxquels on demandera d'identifier un produit présenté sous forme codée, en même temps que le produit standard ou témoin, également codé.

2/ *Les problèmes de préférence.*

Ceux-ci se posent notamment lors de l'élaboration de produits nouveaux ; les tests de comparaison sont généralement effectués en premier lieu par des experts qualifiés, puis à un stade ultérieur par un panel de consommateurs représentatif du public auquel on destine le produit. Il y a intérêt à ce que les réponses puissent être analysées en fonction de facteurs tels que l'âge, le sexe, la résidence, la catégorie socio-professionnelle, etc.

3/ *L'analyse des facteurs susceptibles de modifier les caractéristiques d'un produit.*

Il peut s'agir du choix des matières premières, des procédés de fabrication et aussi du conditionnement, des conditions de stockage, etc.

Détection de différences et problèmes de préférences seront généralement impliqués dans de telles recherches.

4/ *La sélection et le perfectionnement d'un jury d'experts.*

La sensibilité aux différences, la constance des réponses, seront les points essentiels à examiner.

On trouvera des considérations intéressantes sur l'ensemble de ces problèmes dans R. A. Bradley [6] et D. D. Mason and E. J. Koch [32] ; sur le problème particulier des tests de préférence dans C. I. Bliss [3] et sur celui de la sélection et de l'entraînement des experts dans A. O. Mackey and P. Jones [31], G. Bennett, B. M. Spahr and M. L. Dodds [2], M. G. Tarver and B. H. Ellis [38], Vessereau [43].

Les recherches à caractère plus théorique concernent :

5/ La possibilité de situer sur une échelle quantitative les réponses à des intensités croissantes d'un même stimulus (échelle d'intensité) ou les préférences à l'égard de stimuli de nature quelconque (échelle de préférences).

6/ Le choix d'une échelle unique pour les intensités des quatre goûts fondamentaux : N. T. Gridgeman [18].

7/ La mesure du seuil de perception d'un stimulus ; la distribution des seuils de perception dans un groupe d'individus : E. K. Harris [22].

8/ La comparaison de différentes méthodes objectives ou subjectives : W. G. Kauman, J. W. Gottstein and D. Lantican [29], D. E. W. Schumann and R. A. Bradley [36], [37].

B - Les méthodes.

On peut distinguer :

- celles qui ne font pas appel à la notion de mesure ; dans cette catégorie entrent les tests de classement relatifs à deux ou plusieurs produits ;

- celles qui obligent le sujet à formuler sa réponse sous forme d'une indication numérique, c'est-à-dire d'une "note" mesurant le degré d'intensité ressenti, ou le degré de préférence.

L'interprétation des premières repose essentiellement sur les propriétés de l'analyse combinatoire, de la loi binomiale ou de la loi multinomiale. Les secondes doivent faire appel à un "modèle", ou "distribution théorique", tel que la loi normale ; le modèle est appliqué soit aux réponses elles-mêmes, soit à une variable calculée à partir des réponses et dont on a des raisons de penser qu'elle épouse de plus près le modèle théorique.

Les méthodes faisant appel à la cotation permettent d'utiliser la théorie des plans d'expérience et les techniques d'interprétation de l'"analyse de la variance" ; mais elles prêtent à critique, tout modèle *à priori* pouvant être suspecté d'arbitraire. Les travaux de Thurstone [39], [40], [41] et de R. A. Bradley [7], [9], [10], [11] sur les comparaisons par paires, ont permis de jeter un pont entre les deux méthodes en introduisant dans les problèmes de classement un modèle sous-jacent dont

il est possible de tester la valeur. En fait, l'interprétation des mêmes résultats expérimentaux par diverses méthodes conduit généralement à des conclusions très voisines : C. I. Bliss, M. L. Greenwood and E. S. White [4], J. E. Jackson and M. Fleckenstein [28].

On peut d'autre part classer les expériences suivant que, dans une même épreuve, le nombre de termes différents présentés à un même sujet est égal ou supérieur à deux. Les comparaisons par paires ont l'avantage d'être particulièrement simples -elles permettent généralement au sujet d'exercer dans les meilleures conditions ses facultés sensorielles ; le test "triangulaire" -largement utilisé dans les problèmes d'identification- est une variante dans laquelle l'un des deux produits est représenté deux fois.

Le problème de la "cohérence" des réponses se pose lorsqu'un même sujet est invité à comparer deux à deux plusieurs produits (comparaisons telles que AB, AC, BC). Les réponses ne sont pas "cohérentes" si elles sont telles que $A > B$, $B > C$, $C > A$. D'autre part, l'homogénéité des réponses doit être examinée lorsque plusieurs sujets participent aux épreuves.

Ces problèmes, qui se posent notamment à un "jury" dont la mission est d'établir le classement d'un certain nombre de produits ont fait l'objet d'études particulièrement importantes de M. G. Kendall [30].

Les méthodes séquentielles -méthodes dans lesquelles le nombre d'épreuves n'est pas fixé à l'avance, celles-ci étant poursuivies jusqu'à ce que les résultats permettent de conclure avec des risques fixés à l'avance- ne semblent pas avoir fait l'objet de nombreuses applications dans le domaine des tests organoleptiques. La méthode classique de Wald, ainsi qu'une méthode due à Rao sont toutefois citées par R. A. Bradley [6] comme pouvant être utilisées pour la sélection des experts en dégustation.

Les classifications qui précèdent -tant en ce qui concerne les objectifs que les méthodes- ont, comme beaucoup de classifications, un caractère conventionnel, et elles ne s'appliquent que de façon imparfaite à un champ de recherches et d'applications particulièrement complexe. Elles peuvent cependant être de quelque utilité dans l'étude des cas les plus importants qui fait l'objet des paragraphes suivants.

II - LES TESTS D'IDENTIFICATION (OU DE DIFFERENCE)

A - Définition des principaux tests.

Les tests les plus couramment utilisés -notamment en contrôle de qualité et pour la sélection des experts- sont :

Le *pair-test*. Deux produits A et B sont présentés ; le sujet doit répondre à une question telle que "quel est celui des deux qui est... le plus fort ?... le plus salé ?... le plus aromatique ?...".

Le *test "duo-trio"*. L'un des deux produits -A par exemple- est présenté au sujet qui prend connaissance de ses caractères organoleptiques ; A et B sont ensuite présentés en secret, et le sujet est invité à les reconnaître. Une variante de ce test est le test *dual-standard*, dans lequel les deux produits sont d'abord présentés en clair.

Le *test "triangulaire"*. Il comporte trois produits codés, dont deux sont identiques ; le sujet doit désigner celui qui est différent des deux autres. Pour que l'expérience soit "équilibrée", les 6 séquences AAB, ABA, BAA, BBA, BAB, ABB, doivent être testées, dans un ordre de présentation tiré au sort.

Gridgeman et Hopkins considèrent que la réponse donnée dans ces trois tests peut être correcte pour l'une ou l'autre des raisons suivantes :

- le sujet a réellement perçu la différence ;

- le sujet n'a pas perçu la différence et a donné "par hasard" une réponse correcte.

Soit p_s la probabilité du premier évènement (p_s peut aussi être considéré comme la proportion de sujets ayant, dans un panel important, l'aptitude sensorielle qui permet de répondre correctement à la question posée). Soit d'autre part p_h la probabilité que la réponse correcte soit due au seul hasard. La probabilité p d'une réponse correcte -pratiquement, la proportion de réponses correctes dans un grand nombre de répétitions- est :

$$p = p_s + (1 - p_s)\, p_h$$

Cette relation permet d'estimer p_s à partir de p (résultat expérimental) et p_h (qui est défini par le type d'épreuve choisi) :

$$p_s = \frac{p - p_h}{1 - p_h}$$

Dans le "pair-test" et le test "duo-trio", on a $p_h = \frac{1}{2}$,

$$p = p_D = \frac{1 + p_s}{2} \qquad p_s = 2p_D - 1$$

p_D varie de 1/2 à 1 lorsque p_s varie de 0 à 1.

Dans le test "triangulaire", $p_h = \frac{1}{3}$,

$$p = p_\Delta = \frac{1 + 2\,p_s}{3} \qquad p_s = \frac{3\,p_\Delta - 1}{2}$$

p_Δ varie de 1/3 à 1 lorsque p_s varie de 0 à 1.

Pour que le résultat d'une expérience soit "significatif", il faut, dans le "pair-test", que p_D s'écarte suffisamment (dans un sens ou dans l'autre) de la valeur $p_h = \frac{1}{2}$ (qui correspond à l'identité des deux termes comparés) ; dans le test "duo-trio", que p_D soit suffisamment *plus élevé* que $p_h = \frac{1}{2}$, et dans le test "triangulaire" que p_Δ soit suffisamment *plus élevé* que $p_h = \frac{1}{3}$.

D'une façon plus précise, supposons que l'expérience comporte N répétitions et que l'on ait choisi un seuil de signification $\alpha = 0,05$. Soit n le nombre de réponses en faveur de l'un des deux termes (A ou B, cas du "pair-test"), ou le nombre d'identifications correctes ("duo-trio", ou "triangulaire-test"). L'hypothèse nulle est rejetée (résultats significatifs) dans les circonstances suivantes :

Pair-test : n est extérieur à l'intervalle (n_1, n_2), n_1 et n_2 étant les valeurs de la variable binomiale $\left(p = \frac{1}{2},\ N \text{ répétitions}\right)$ telles que :

$$Pr[n \leqslant n_1] \leqslant 0,025 \qquad Pr[n \geqslant n_2] \leqslant 0,025$$

Test "duo-trio" : n est supérieur à n_2, valeur de la même variable binomiale, telle que :

$$Pr[n \geqslant n_2] \leqslant 0,050$$

Test "triangulaire" : n est supérieur à n_2, valeur de la variable binomiale $\left(p = \frac{1}{3},\ N \text{ répétitions}\right)$ telle que :

$$Pr[n \geqslant n_2] \leqslant 0,050$$

Les tables de la loi binomiale donnent les valeurs de n_1 et n_2 ; pour un nombre de répétitions élevé, on peut utiliser l'approximation normale, qui conduit à rejeter l'hypothèse nulle lorsque :

$$\left|\frac{2n - N}{\sqrt{N}}\right| > 1,96 \quad (\text{"pair-test"})$$

$\left|\frac{2n - N}{\sqrt{N}}\right| > 1,64$ (test "duo-trio") $\left|\frac{3n - N}{\sqrt{2N}}\right| > 1,64$ ("test triangulaire")

B - Comparaison des tests.

Dans le test "triangulaire", la place laissée au hasard est plus petite que dans les deux autres. Mais la véritable comparaison des tests repose sur le "risque de 2e espèce", c'est-à-dire sur la probabilité d'accepter l'hypothèse nulle, alors que l'hypothèse vraie est une alternative $p \neq \frac{1}{2}$ ("pair-test"), $p > \frac{1}{2}$ (test "duo-trio") ou $p > \frac{1}{3}$ (test "triangulaire"). Ce risque décroît au fur et à mesure que p_s (lié à p par les relations indiquées plus haut) augmente de 0 à 1. Il décroît d'autre part lorsque le nombre de répétitions N augmente.

Il est possible, au moyen des tables de la loi binomiale, de tracer les courbes d'efficacité des trois tests. Cette étude mathématique a été faite notamment par J.W. Kopkins [26], N. T. Gridgeman [16], J.W. Hopkins and N.T. Gridgeman [27]. Elle conduit à la conclusion que, dans l'hypothèse où une même valeur de p_s (probabilité de détecter la différence) leur est applicable, les tests se classent dans l'ordre d'efficacité croissante suivant :

"pair" "duo-trio" "triangulaire"

Cependant, pour un même nombre d'épreuves, le premier test est plus économique que les deux autres puisqu'il n'exige que deux produits par épreuve alors que le test "duo-trio" et le test "triangulaire" en demandent trois. Si l'on compare les tests à nombre égal de produits utilisés, on constate [27] que le test triangulaire reste le plus puissant, mais que la puissance relative des deux autres tests peut s'intervertir suivant les valeurs de p_s.

Il n'est pas évident d'autre part qu'une même valeur de p_s soit applicable aux 3 tests. Ce point ne peut être éclairci que par des expérience directes, dans lesquelles les mêmes sujets appliquent aux mêmes

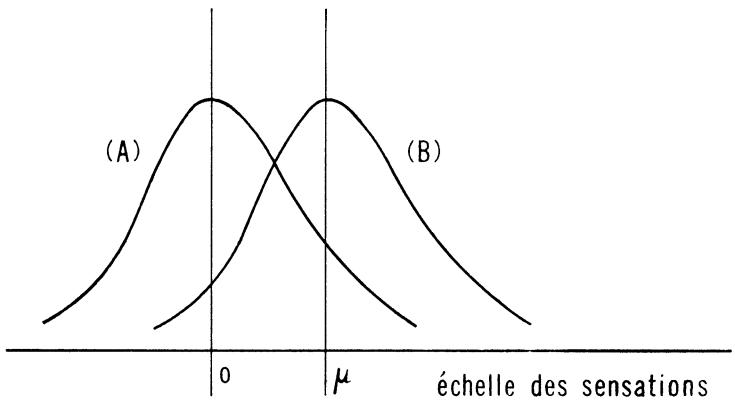
produits chacun des tests, toutes précautions étant prises pour qu'aucun biais ne s'introduise au cours du déroulement des essais. De telles expériences sont décrites notamment dans [16] et [27]. Dans certains cas, on constate que p_s est plus élevé dans le "pair-test" que dans le test "triangulaire" et il peut en résulter une efficacité supérieure du "pair-test".

Accessoirement on trouvera dans [26] des résultats intéressants concernant l'effet de l'entraînement des experts sur la proportion de réponses correctes. Il n'est constaté aucun effet "automatique" résultant de la répétition des épreuves par une même personne, mais une amélioration systématique très nette des réponses est constatée lorsque les experts sont soumis, jour après jour, à un entraînement dirigé.

C - Le modèle mathématique de Bradley et de Ura.

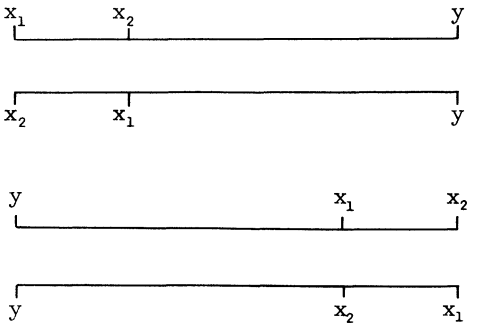
Une contribution importante et originale à l'étude des tests d'identification a été apportée, indépendamment l'un de l'autre, par R. A. Bradley [8] et S. Ura [42].

On admet l'existence d'une échelle quantitative sur laquelle peuvent être situées les impressions sensorielles ressenties sous l'action des stimuli -on admet d'autre part que, sous l'action du stimulus A, considéré comme témoin, les impressions sont distribuées normalement autour d'une moyenne choisie comme origine, avec un écart-type σ, tandis que sous l'action du stimulus B, ces impressions sont distribuées normalement autour d'une moyenne μ, avec le même écart-type σ. Toutes les impressions reçues -et les réponses qui en découlent- sont supposées stochastiquement indépendantes. Moyennant ces hypothèses, il est possible d'exprimer les pourcentages théoriques de réponses exactes -p_Δ pour le test "triangulaire", p_0 pour le test "duo-trio"- en fonction du rapport μ/σ.



Dans le cas du test "*triangulaire*" les sensations ressenties sous l'action de A -qui est le stimulus représenté 2 fois- sont désignées par x_1 et x_2 ; x_1 et x_2 sont 2 valeurs indépendantes de la variable normale $(0,\sigma)$. La sensation ressentie sous l'action de B est désignée par y ; y appartient à la distribution normale (μ,σ).

La réponse formulée par le sujet est correcte dans les quatre situations suivantes :



$x_1 \leqslant x_2 \leqslant y \qquad x_2 - x_1 < y - x_2$

$x_2 \leqslant x_1 \leqslant y \qquad x_1 - x_2 < y - x_1$

$y \leqslant x_1 \leqslant x_2 \qquad x_2 - x_1 < x_1 - y$

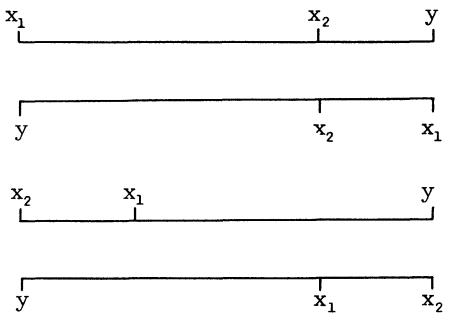
$y \leqslant x_2 \leqslant x_1 \qquad x_1 - x_2 < x_2 - y$

La somme des probabilités correspondant à ces situations, soit p_Δ, a été calculée par Bradley et par Ura qui ont obtenu des expressions différentes (mais naturellement équivalentes) de cette quantité en fonction de μ/σ.

D'autre part, Ura a également calculé l'expression de p_Δ dans l'hypothèse où la distribution des sensations n'est pas normale, mais rectangulaire ; l'expression correspondante est désignée par $p_{\Delta u}$.

Dans le cas du test "*duo-trio*", x_1 se rapporte au témoin présenté en clair avant l'épreuve ; dans celle-ci les sensations sont désignées par x_2 et y.

Les quatre situations amenant des réponses correctes sont les suivantes :



$x_1 \leqslant x_2 \leqslant y$

$y \leqslant x_2 \leqslant x_1$

$x_2 \leqslant x_1 \leqslant y \qquad x_1 - x_2 < y - x_1$

$y \leqslant x_1 \leqslant x_2 \qquad x_2 - x_1 < x_1 - y$

Elles permettent, comme précédemment, de calculer la proportion théorique de réponses exactes en fonction de μ/σ, soit p_0, en supposant les distributions de x et y normales (expressions de Bradley et de Ura) et p_{0u} en faisant, avec Ura, l'hypothèse de distributions rectangulaires.

[8] contient un tableau des valeurs de p_Δ, $p_{\Delta u}$, p_0 et p_{0u} pour toute une gamme de valeurs de μ/σ ; un extrait en est donné ci-après.

μ/σ	p_Δ	$p_{\Delta u}$	p_0	p_{0u}
0,0.....	0,333	0,333	0,500	0,500
0,8.....	0,389	0,381	0,555	0,546
1,6.....	0,526	0,497	0,680	0,656
2,4.....	0,681	0,647	0,806	0,786
3,2.....	0,810	0,793	0,895	0,894
4,0.....	0,898	0,899	0,947	0,950

Il est remarquable de constater que les valeurs obtenues à partir de deux distributions aussi différentes que la distribution normale et la distribution rectangulaire sont très voisines : la forme du modèle rendant compte de la dispersion des sensations semble donc avoir peu d'importance.

Bradley observe encore que, moyennant certaines hypothèses supplémentaires, il est possible d'exprimer p_D en fonction de p_Δ.

a) Suivant la théorie de Hopkins-Grindgeman rappelée plus haut, on a en effet :

$$p_D = \frac{1 + p_s}{2} \qquad p_\Delta = \frac{1 + 2p_s}{3}$$

d'où en éliminant p_s :

$$p_{DH} = (1 + 3p_\Delta)/4 \qquad \text{(l'indice H rappelle Hopkins)}$$

b) Un autre raisonnement dû à David est le suivant :

Dans le test "triangulaire", les probabilités de désigner le premier témoin A, le deuxième témoin A et le produit B (réponse exacte) sont respectivement :

$$(1 - p_\Delta)/2 \qquad (1 - p_\Delta)/2 \qquad p_\Delta$$

Admettant que les probabilités attachées à A et à B restent valables dans le test "duo-trio" où ne figure qu'un témoin, on obtient pour expression de p_D en fonction de p_Δ :

$$p_{DD} = \frac{p_\Delta}{(1 - p_\Delta)/2 + p_\Delta} = \frac{2p_\Delta}{1 + p_\Delta} \qquad \text{(l'indice D rappelle David)}$$

Du tableau détaillé figurant dans [8] on extrait le tableau ci-après qui, pour quelques valeurs de p_Δ (calculées à partir de μ/σ dans l'hypothèse de normalité) donne les valeurs de p_{DH} et p_{DD} obtenues par les formules qui précèdent ; ces valeurs sont rapprochées de la valeur p_D déduite directement de μ/σ.

μ/σ	p_Δ	p_D	p_{DH}	p_{DD}
0,0.....	0,333	0,500	0,500	0,500
0,8.....	0,389	0,555	0,542	0,560
1,6.....	0,526	0,680	0,644	0,689
2,4.....	0,681	0,806	0,761	0,810
3,2.....	0,810	0,895	0,857	0,895
4,0.....	0,898	0,947	0,923	0,946

On constate que les deux formules -surtout celle de David- sont en bon accord avec la valeur p_D calculée directement.

La théorie de Bradley permet de comparer le test "triangulaire" et le test "duo-trio" sans faire intervenir la quantité p_s (probabilité de détecter la différence) puisque les valeurs p_Δ et p_D peuvent se calculer sur la même base objective de la quantité μ/σ caractérisant l'écart des stimuli A et B. Les courbes d'efficacité ont été tracées par Ura [42]

pour N = 20 répétitions et le seuil de signification $\alpha = 0,05$. Elles montrent que, pour ces conditions, le test "triangulaire" est uniformément plus puissant que le test "duo-trio". Bradley [8] a d'autre part calculé qu'au seuil $\alpha = 0,05$, et pour un risque de 2e espèce $\beta = 0,15$ associé à l'alternative $\mu/\sigma = 3$, le test "triangulaire" demande 9 répétitions alors que le test "duo-trio" en exige 12 ; l'efficacité du premier test par rapport au second est, dans ces conditions, de 12/9 = 130 p. 100.

D - Identification au moyen de deux critères.

Ce problème a été étudié par N. T. Gridgeman [21]. On applique le test "duo-trio" à deux produits A et B, une première fois sur un critère h, une deuxième fois, et indépendamment, sur un autre critère k (par exemple h et k sont le goût et l'arôme). On suppose qu'entre A et B il existe de petites différences Δh et Δk.

Dans une population de sujets, on distingue :

- ceux qui détectent à la fois h et k - en proportion p_{hk} ;
- ceux qui détectent seulement h - en proportion p_h ;
- ceux qui détectent seulement k - en proportion p_k ;
- ceux qui ne détectent ni h ni k - en proportion $1 - p_{hk} - p_h - p_k$.

Comme dans les situations précédemment étudiées, une réponse peut être correcte, soit parce que la différence a été réellement perçue, soit du fait d'un hasard heureux. Sur un ensemble de N sujets appartenant à la population considérée, les réponses se classent de la façon suivante :

		Réponses sur h		Total
		Correctes	Incorrectes	
Réponses sur k	Correctes	R_{11}	R_{21}	R_k
	Incorrectes...	R_{12}	R_{22}	$N - R_k$
	Total...	R_h	$N - R_h$	N

On établit facilement les deux tableaux ci-après qui donnent :

- le premier, les valeurs théoriques (espérances mathématiques) des R_{ij} en fonction des quantités p_{hk}, p_h et p_k.

- le deuxième, les estimations de ces quantités en fonction des R_{ij}.

	Coefficients de				
	p_{hk}	p_h	p_k	1	
$E(R_{11})$.	+3	+1	+1	+1	$\times \frac{N}{4}$
$E(R_{12})$.	−1	+1	−1	+1	
$E(R_{21})$.	−1	−1	+1	+1	
$E(R_{22})$.	−1	−1	−1	+1	

	Coefficients de				
	R_{11}	R_{12}	R_{21}	R_{22}	
$\hat{p}_{hk}$...	+1	−1	−1	+1	$\times \frac{1}{N}$
$\hat{p}_h$ ...	0	+2	0	−2	
$\hat{p}_k$ ...	0	0	+2	−2	

L'auteur propose, à partir de ces données, deux séries de tests d'hypothèses :

Tests concernant la différence entre les produits [$\Delta(A, B)$]. L'hypothèse nulle ($p_h = p_k = p_{hk} = 0$) peut être testée contre l'une ou l'autre des deux alternatives suivantes :

- les deux critères h et k révèlent une différence :

$$[p_h + p_{hk} > 0 \quad \text{et} \quad p_k + p_{hk} > 0]$$

- le critère h ou le critère k révèle une différence :

$$[p_h + p_{hk} > 0 \quad \text{ou} \quad p_k + p_{hk} > 0]$$

Tests concernant la différence entre les critères [$\Delta(h, k)$].

On peut prendre pour hypothèse nulle que la différence $\Delta(A, B)$ est décelée également bien au moyen de h et de k -soit $p_h = p_k \geqslant 0$. L'hypothèse alternative est $p_h \neq p_k$. On peut aussi prendre pour hypothèse nulle que quiconque détecte par h détecte par k, soit $p_h = p_k = 0$; l'hypothèse alternative est alors $p_h \neq 0$ ou $p_k \neq 0$.

Un exemple numérique est donné dans [21] ; il concerne la comparaison, par goût et arôme, de 3 marques de rhum et un témoin.

E - Généralisation des tests d'identification.

N. T. Gridgeman [19] a traité le problème plus général de l'identification de N objets ayant un caractère commun parmi M objets présentés. Les identifications correctes, en nombre $R(0 \leqslant R \leqslant N)$ proviennent soit de ce que le caractère a été effectivement perçu (probabilité p_s), soit d'un hasard heureux dans le cas de non perception. Les $(N - R)$ réponses incorrectes proviennent de la non perception et d'un hasard malheureux.

Les expressions de Pr. (R), de l'espérance mathématique E(R), et de la variance de R sont données dans le cas général (M et N quelconques) ainsi que dans les cas particuliers suivants : ($M = 2N$) et ($N = 1$). Dans ce dernier cas on a évidemment (en posant $q_s = 1 - p_s$) :

$$\text{Pr.}[R = 1] = p_s + \frac{1 - p_s}{M} = p_s + \frac{q_s}{M} \qquad \text{Pr.}[R = 0] = \frac{M - 1}{M} q_s$$

$$E(R) = p_s + \frac{q_s}{M} \qquad \text{var}(R) = \frac{M - 1}{M^2} q_s (M p_s + q_s)$$

expressions qui contiennent les cas particuliers du test "triangulaire" ($M = 3$) et du test "duo-trio" ($M = 2$).

Dans le cas d'absence de perception (hypothèse nulle $p_s = 0$) on trouve :

$$\text{Pr.}(R) = \left[\frac{N!\,(M - N)!}{(N - R)!}\right]^2 \times \frac{1}{M!\,R!\,(M - 2N + R)!}$$

et lorsque $M = 2N$:

$$Pr.(R) = \frac{1}{(2N)!}\left[\frac{(N!)^2}{R!\,(N-R)!}\right]^2$$

Par exemple :

R	M = 4 N = 2	M = 6 N = 3	M = 8 N = 4	M = 10 N = 5
0.....	1/6	1/20	1/70	1/252
1.....	2/3	9/20	16/70	25/252
2.....	1/6	9/20	36/70	100/252
3.....	-	1/20	16/70	100/252
4.....	-	-	1/70	25/252
5.....	-	-	-	1/252

Les formules qui précèdent permettent -au moins théoriquement- de tester l'hypothèse nulle $p_s = 0$ sous un seuil de signification donné. En fait, l'emploi de schémas plus compliqués que les schémas classiques ne présenterait un réel intérêt que si leur efficacité était meilleure. L'étude théorique de cette efficacité serait très complexe- et vraisemblablement sans grand intérêt, car il y a tout lieu de penser que p_s décroît à mesure que le schéma devient plus compliqué.

III - LES TESTS DE CLASSEMENT ET DE PREFERENCES PAR PAIRES

Le classement de deux produits suivant l'intensité d'un caractère (dire quel est le plus doux, le plus amer, le plus sucré...) ou suivant la préférence (dire quel est le meilleur) s'effectue par le moyen d'un "pair-test" comme il a été indiqué au paragraphe précédent. Lorsqu'il s'agit de comparer plusieurs produits qui diffèrent par l'intensité d'un même stimulus (concentration croissante d'une même substance) il est raisonnable d'imaginer qu'à l'échelle des intensités du stimulus correspond une échelle d'intensités des sensations (cf. loi de Fechner $S = \alpha + \beta \log \frac{C}{C_o}$). L'existence d'une échelle objective applicable aux préférences est moins évidente : les préférences ont un caractère individuel ; la préférence peut passer par un maximum lorsque l'intensité d'un stimulus déterminé augmente ; enfin, dans le cas de stimuli complexes, on doit s'attendre à ce que la loi des préférences soit également complexe.

Le classement des produits peut s'effectuer en une seule fois et s'accompagner d'une note mesurant le degré de préférence. Mais il peut aussi, plus commodément, être décomposé en comparaisons portant sur toutes les paires de produits pris deux à deux : c'est dans ces conditions que les jugements s'exercent normalement avec le maximum de sensibilité.

On traitera dans ce paragraphe uniquement des comparaisons par paires (avec une extension aux "triplets"), qui ont fait l'objet de nombreuses publications récentes. Les produits à comparer sont désignés par $T_1, T_2, \ldots, T_i, \ldots, T_t$; le nombre de paires est égal à $\frac{t(t-1)}{2}$.

A - La méthode de Thurstone-Mosteller (33), (39), (40), (41).

Elle suppose l'existence d'une échelle continue de sensations (intensités ou préférences). On fait l'hypothèse que dans une population de sujets, la distribution des sensations X_i pour un même stimulus (T_i) est normale, avec la valeur moyenne S_i ; pour le stimulus (T_j) la distribution de X_j est normale avec la moyenne S_j et la même variance. Si, dans la comparaison (T_i, T_j), les sensations sont corrélées, on admet que la corrélation est la même pour toutes les combinaisons (i, j). Dans ces conditions, la distribution de la différence $U = X_i - X_j$ est normale, avec la moyenne $S_i - S_j$ et une variance que l'on peut prendre pour unité. La probabilité que (T_i) soit classé supérieur à (T_j) dans la comparaison (T_i, T_j) est :

$$\pi_{ij} = \Pr.[X_i > X_j] = \Pr.[U > 0] = \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[U-(S_i-S_j)]^2} \, dU$$

$$\pi_{ij} = \int_{-(S_i-S_j)}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \, dy$$

La fréquence p_{ij} des comparaisons (i, j) où T_i est trouvé supérieur à T_j constitue l'estimation de π_{ij} ; les tables de la loi normale donnent ensuite les estimations $[\hat{S}_i - \hat{S}_j]$ des écarts $[S_i - S_j]$.

Le stimulus (i) figure dans (t - I) paires et l'on a :

$$\frac{1}{t} \sum_{j=1}^{t} (\hat{S}_i - \hat{S}_j) = \hat{S}_i - \bar{S}$$

où $\bar{S}$ désigne la moyenne arithmétique des $\hat{S}_i$. On peut donc placer les t valeurs $\hat{S}_i$ en prenant pour origine leur moyenne $\bar{S}$.

Mosteller a donné un test qui permet d'éprouver la validité de ce modèle. Une difficulté d'application se présente lorsque, dans la comparaison (i, j) le produit (T_i) n'est jamais préféré $(p_{ij} = 0)$ ou est toujour préféré $(p_{ij} = 1)$; les valeurs de $\hat{S}_i - \hat{S}_j$ deviennent alors $-\infty$ ou $+\infty$; on a proposé -assez arbitrairement- d'adopter à leur place des valeurs telles que (-4, -3, -2) ou (+2, +3, +4).

Morrissey et Gulliksen ont étendu la méthode de Thurstone Mosteller au cas où certaines comparaisons font défaut (élimination de résultats aberrants par exemple).

Il convient de noter que la méthode de Thurstone transforme les fréquences observées p_{ij} en "écarts normaux". Si l'expérience a été organisée convenablement, il est possible d'utiliser ces écarts dans une analyse de variance permettant de tester l'influence de facteurs tels que l'ordre de présentation des produits.

B - La méthode de Bliss (4).

Les hypothèses sont les mêmes que celles de Thurstone-Mosteller, mais, au lieu de transformer les fréquences p_{ij} en "écarts normaux", on utilise la transformation des *nombres de préférence* $n_{ij} = Np_{ij}$ (N répétitions de la comparaison i, j) en "rankits", ou "scores for ordinal, or ranked data", mis en tables par Fisher and Yates [14]. n résultats issus d'une distribution normale réduite étant rangés par valeurs croissantes, le "rankit" correspondant au rang r est l'espérance mathématique de la variable occupant ce rang.

Le tableau ci-après donne, à titre d'exemple, la correspondance entre la valeur du "rankit" et la valeur de l'écart normal selon Thurstone-

Mosteller, pour N = 25 répétitions, et pour des nombres de préférences allant de 9 à 20 [exemple tiré de l'expérience décrite en détail dans [4]].

Nombre de préférences	Rankit	Ecart normal	Nombre de préférences	Rankit	Ecart normal
9.....	- 0,34	- 0,36	15.....	+0,24	+0,25
10.....	- 0,24	- 0,25	16.....	+0,34	+0,36
11.....	- 0,14	- 0,15	17.....	+0,44	+0,47
12.....	- 0,05	- 0,05	18.....	+0,55	+0,58
13.....	+ 0,05	+ 0,05	19.....	+0,67	+0,71
14.....	+ 0,14	+ 0,15	20.....	+0,69	+0,84

"Rankits" et "écarts normaux" sont voisins, et il n'est pas étonnant que les conclusions qui ont été obtenues à partir de l'une et l'autre de ces variables soient très comparables. Les "rankits" se prêtent, comme les écarts normaux, à l'analyse de variance ; ils présentent sur ces derniers l'avantage de conserver des valeurs finies pour un nombre de préférences égal à 0 ou à N.

C - La méthode de Bradley-Terry (6), (7), (9), (10), (11).

On représente l'effet sensoriel des différents produits par des paramètres :

$$\pi_1 , \pi_2 \ldots\ldots\ldots \quad \pi_i \ldots\ldots\ldots \quad \pi_j \ldots\ldots\ldots \quad \pi_t$$

tels que :

$$\pi_i \geqslant 0 \quad \text{et} \quad \sum_i \pi_i = 1$$

et l'on pose :

$$\Pr.[X_i > X_j] = \pi_{ij} = \frac{\pi_i}{\pi_i + \pi_j}$$

On indiquera plus loin comment on peut estimer les quantités π_{ij} à partir des résultats expérimentaux. Montrons d'abord que le modèle de Bradley-Terry se ramène à celui de Thrurstone moyennant la double modification suivante :

- à la distribution normale des sensations on substitue celle du "carré de la sécante hyperbolique" dont la forme est voisine de celle de la loi normale :

$$f(y)\, dy = \frac{1}{4}\left[\operatorname{sech}^2 \frac{y}{2}\right] dy = [e^{y/2} + e^{-y/2}]^{-2}\, dy \qquad (-\infty < y < +\infty)$$

- à S_i on fait correspondre $\ln\pi_i$ (logarithme népérien), soit :

$$\pi_i = e^{S_i}, \qquad \pi_j = e^{S_j}$$

L'expression de Thurstone :

$$\pi_{ij} = \int_{-(S_i - S_j)}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$$

devient ainsi :

$$\pi_{ij} = \int_{\ln \frac{\pi_j}{\pi_i}}^{\infty} [e^{y/2} + e^{-y/2}]^{-2}\, dy$$

On trouve facilement :

$$\int [e^{y/2} + e^{-y/2}]^{-2}\, dy = -\frac{1}{1 + e^{y}}$$

d'où :

$$\pi_{ij} = \left[-\frac{1}{1 + e^{y}} \right]_{\ln \frac{\pi_j}{\pi_i}}^{\infty} = \frac{1}{1 + \frac{\pi_j}{\pi_i}} = \frac{\pi_i}{\pi_i + \pi_j}$$

Les estimations des π_i s'obtiennent par la méthode du maximum de vraisemblance. Les notations de Bradley sont les suivantes :

Il y a n répétitions de chacune des paires de produits (n "blocs incomplets de deux"), et d'autre part les paires de produits sont en nombre $\frac{t(t-1)}{2}$.

Dans chaque bloc où (T_i) et (T_j) sont comparés, on donne le rang 1 au produit préféré, et le rang 2 à l'autre. r_{ijk} désigne le rang de T_i dans la k^{me} répétition (i, j), (k = 1, 2, ..., n). On a :

$$r_{ijk} + r_{jik} = 3$$

La somme des rangs de T_i est :

$$\sum r_i = \sum_k \sum_{j \neq i} r_{ijk}$$

Posant d'autre part :

$$a_i = 2n(t-1) - \sum_k \sum_{j \neq i} r_{ijk} = 2n(t-1) - \sum r_i$$

on trouve que les estimations p_i s'obtiennent par résolution du système de t équations :

$$\frac{a_i}{p_i} - n \sum_{j \neq i} (p_i + p_j)^{-1} = 0 \qquad (i = 1, 2, \ldots, t)$$

Ce système n'étant pas linéaire, les solutions doivent, dans le cas général, être recherchées par itération à partir de valeurs initiales approchées ; des formules pour celles-ci sont données par O. Dykstra [12]. Des tables donnent directement les valeurs p_i en fonction de $\sum r_i$ dans les cas suivants :

t = 3	n = 1, 2, ..., 10	R. A. Bradley (9)
t = 4	n = 1, 2, ..., 6	
t = 4	n = 7, 8	R. A. Bradley (10)
t = 5	n = 1, 2, ..., 5	

Exemple : Trois produits (t = 3) et trois répétitions (n = 3).

$\sum r_1$	$\sum r_2$	$\sum r_3$	p_1	p_2	p_3
6	9	12	1,00	0,00	0,00
6	10	11	1,00	0,00	0,00
7	8	12	0,67	0,33	0,00
7	9	11	0,70	0,22	0,07
7	10	10	0,71	0,14	0,14
8	8	11	0,45	0,45	0,09
8	9	10	0,50	0,31	0,10
9	9	9	0,33	0,33	0,33

On peut faire ici la remarque que les notations de Bradley sont peu pratiques. En affectant les rangs 1 ou 0 à T_i, suivant qu'il est préféré ou non (au lieu de 1 et 2), la somme des rangs pour T_i donnerait directement le nombre de fois où ce produit a été préféré dans les $n(t-1)$ comparaisons où il figure. Ce nombre, désigné par Bradley par $\sum_{j \neq i} f_{ij}$ [et par d'autres auteurs (Kendall, Dykstra) par a_i] est lié à $\sum r_i$ (qu'il eût été plus logique d'appeler r_i) par la relation mentionnée ci-dessus :

$$a_i = 2n(t-1) - \sum r_i$$

L'hypothèse nulle dans le modèle de Bradley est l'hypothèse (H_o) $\pi_i = \frac{1}{t}$ quel que soit i ; l'hypothèse alternative est l'hypothèse (H_1) $\pi_i \neq \frac{1}{t}$ pour certains i. Le test de l'hypothèse nulle est basé sur le rapport des vraisemblances calculées sous hypothèses H_o et H_1. En désignant la vraisemblance par L (likelihood), on trouve, pour expression du logarithme (népérien) du rapport des vraisemblances :

$$\ln \lambda_1 = \ln \frac{L/H_o}{L/H_1} = B_1 \ln 10 - \frac{nt(t-1)}{2} \ln 2$$

avec :

$$B_1 = n \sum_{i<j} \log (p_i + p_j) - \sum_i a_i \log p_i -$$

(ln désigne le logarithme népérien et log le logarithme dans la base 10).

L'hypothèse nulle est rejetée si la probabilité attachée à λ_1 (ou $\ln \lambda_1$) est inférieure au seuil de signification choisi. Dans la pratique :

- pour n grand, $-2 \ln(\lambda_1)$ peut être considéré comme suivant la loi de χ^2 à $(t-1)$ degrés de liberté ;

- pour les valeurs de n et t figurant dans les tables de Bradley, celles-ci donnent directement les valeurs B_1, et les probabilités correspondantes P (P = probabilité que la valeur correspondante de B_1 ne soit pas dépassée dans le cas de l'hypothèse nulle). Par exemple, pour t = 3 et n = 3.

Σr_1	Σr_2	Σr_3	B_1	P
6	9	12	0,000	0,0117 (*)
6	10	11	0,829	0,0820
7	8	12		
7	9	11	1,840	0,2226
7	10	10	2,077	0,4336
8	8	11		
8	9	10	2,511	0,8906
9	9	9	2,709	1,0000

(*) Dans ce cas seul l'hypothèse nulle peut être rejetée au seuil 0,05.

Bradley a proposé d'autres tests qui s'appliquent au cas où les n répétitions se décomposent en plusieurs groupes homogènes (différents groupes d'experts ; on peut ainsi tester l'homogénéité des résultats entre experts).

D'autre part, un test de "convenance du modèle" ("goodness of fit") peut être effectué. L'hypothèse nulle correspondant au modèle :

$$H_o \qquad \overline{\pi}_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \qquad \text{pour tout } (i, j)$$

est testée contre l'hypothèse alternative :

$$H_1 \qquad \overline{\pi}_{ij} \neq \frac{\pi_i}{\pi_i + \pi_j} \qquad \text{pour certains i et j}$$

$\overline{\pi}_{ij}$ désignant la probabilité que T_i soit classé supérieur à T_j.

Ce test, lui aussi, peut être adapté au cas où les répétitions se séparent en plusieurs groupes.

Des exemples numériques d'application de la méthode de Bradley-Terry figurent notamment dans [4], [7], [17], [25].

La méthode a été adaptée par O. Dykstra [13] au cas où le nombre de répétitions n'est pas le même pour toutes les paires ; l'auteur traite le problème général, et le cas particulier où le nombre de répétitions est égal à n (constant) pour certaines paires et est nul pour les autres. Ce cas est particulièrement important lorsque t est grand, car le nombre total de paires peut devenir trop élevé pour les possibilités expérimentales. L'auteur indique comment il est possible de choisir un certain nombre de paires $\left(\text{parmi les } \frac{t(t-1)}{2}\right)$ de façon à conserver une expérience "équilibrée".

R. M. Abelson et R. A. Bradley [1] ont étudié le cas où les produits T_i résultent de la combinaison factorielle de m facteur A_1, A_2, ... A_s, A_m ayant des nombres de niveaux égaux à a_1, a_2, ..., a_s, ... a_m. On a alors $t = a_1 \times a_2 \times \ldots a_m$.

Le cas de deux facteurs à deux niveaux est spécialement traité ; au-delà, les calculs paraissent trop compliqués pour que des applications pratiques puissent être envisagées.

Enfin Bradley-Pendergrass ont étendu la méthode à la comparaison des produits par "blocs incomplets de 3". Aux produits :

$$T_1,\ T_2 \ \ldots\ T_i \ \ldots\ T_t$$

on attache encore des paramètres :

$$\pi_1,\ \pi_2 \ \ldots\ \pi_i \ \ldots\ \pi_t$$

tels que :

$$\pi_i \geq 0 \qquad \sum_i \pi_i = 1,$$

et l'on écrit :

$$\Pr.\,[T_i > T_j > T_k] = \pi_i^2\ \pi_j\ /\Delta_{ijk}$$

avec :

$$\Delta_{ijk} = \pi_i^2(\pi_j + \pi_k) + \pi_j^2(\pi_i + \pi_k) + \pi_k^2(\pi_i + \pi_j) \qquad \begin{cases} i,j,k = 1,2,\ \ldots,t \\ i \neq j \neq k \end{cases}$$

qui généralise la condition :

$$\Pr.\,[T_i > T_j] = \frac{\pi_i}{\pi_i + \pi_j}$$

des "blocs de deux".

Des exemples d'application sont donnés par G. T. Park [34] ; sur les mêmes matériels expérimentaux, sont comparées la méthode par blocs de 2 et la méthode par blocs de 3 ; l'auteur conclut que cette dernière est applicable à certains tests sensoriels, et que dans certains cas, elle peut être la plus efficace.

En définitive, la méthode de Bradley (blocs de 2) constitue une contribution importante à l'étude des classements par paires ; le modèle imaginé a généralement été confirmé par le "test statistique de convenance" appliqué à différents matériels expérimentaux. L'application pratique est aisée dans le domaine des valeurs de t et n où les tables numériques existent. Au-delà, le calcul des p_i présente des difficultés devant lesquelles on peut hésiter (pour le test d'hypothèse nulle, on peut utiliser l'approximation de χ^2, qui ne demande pas de grands calculs supplémentaires).

Une difficulté doit cependant être signalée. Elle se présente lorsqu'un produit est toujours préféré à tous les autres, ou toujours jugé inférieur à tous les autres.

Dans le premier cas, $T_i > T_j$ pour tout $j \neq i$, le modèle donne $p_i = 1$ et $p_{j \neq i} = 0$; les produits autres que T_i ne sont donc pas classés entre eux. On peut dans ce cas reprendre l'analyse en supprimant ce produit.

Dans le deuxième cas, $T_i < T_j$ pour tout $j \neq i$, on a $p_i = 0$, auquel correspond sur l'échelle des sensations ($e^{\hat{S}i} = p_i$) la valeur $\hat{S}_i = -\infty$.

D - La méthode de Scheffe (35).

Les particularités de cette méthode sont les suivantes :

- chaque paire (T_i, T_j) est testée deux fois : dans l'ordre (T_j, T_i) et dans l'ordre (T_i, T_j) - le sujet ignore naturellement la nature et l'ordre des produits présentés ;

- une note de préférence est donnée au premier produit de la paire. Si celui-ci est T_i (l'autre étant T_j) la note correspond à un barême tel que le suivant :

0	pas de préférence
+3	(i) fortement préféré à (j)
+2	(i) modérément préféré à (j)
+1	(i) légèrement préféré à (j)
-3	(j) fortement préféré à (i)
-2	(j) modérément préféré à (i)
-1	(j) légèrement préféré à (i).

En fait, le barême est au choix de l'expérimentateur (le barême +1, -1 correspond aux schémas étudiés précédemment).

Avec t produits il y a $\frac{t(t-1)}{2}$ paires "non-ordonnées" et $t(t-1)$ paires "ordonnées" ; si n répétitions sont faites, il y a au total $nt(t-1)$ résultats. Ceux-ci sont analysés par la méthode générale d'analyse de la variance, qui permet de tester :

- s'il y a des différences significatives entre les produits ;

- si l'ordre de présentation a un effet sur les résultats ;

- si le modèle adopté est statistiquement acceptable -c'est-à-dire, en prenant par exemple le cas de 4 produits T_1, T_2, T_3, T_4, s'il existe 4 paramètres θ_1, θ_2, θ_3, θ_4, tels que les différences des produits deux à deux soient effectivement mesurées par les écarts $(\theta_1 - \theta_2)$, $(\theta_1 - \theta_3)$, $(\theta_1 - \theta_4)$, $(\theta_2 - \theta_3)$, $(\theta_2 - \theta_4)$, $(\theta_3 - \theta_4)$.

Les estimations $\hat{\theta}_i$ des paramètres -qui correspondent aux écarts normaux moyens de Thurstone, aux ln p_i de Bradley et aux "rankits" moyens de Bliss- s'obtiennent comme il est indiqué dans le tableau ci-après :

Coefficients à appliquer à la somme des notes du premier terme de la comparaison.

	T_1-T_2	T_2-T_1	T_1-T_3	T_3-T_1	T_1-T_4	T_4-T_1	T_2-T_3	T_3-T_2	T_2-T_4	T_4-T_2	T_3-T_4	T_4-T_3	
$\hat{\theta}_1$	+1	-1	+1	-1	+1	-1							$\times \frac{1}{8n}$
$\hat{\theta}_2$	-1	+1					+1	-1	+1	-1			
$\hat{\theta}_3$			-1	+1			-1	+1			+1	-1	
$\hat{\theta}_4$					-1	+1			-1	+1	-1	+1	

On peut faire au modèle de Scheffe les objections suivantes :

L'échelle des cotations est arbitraire ; lorsque les répétitions sont faites par plusieurs sujets, ceux-ci ne l'utilisent généralement pas de la même façon.

On applique la méthode d'analyse de la variance à une variable qui n'est pas normale.

Pour échapper à cette deuxième critique, C. I. Bliss [4] a proposé de transformer les notes brutes en "rankits". En fait, dans l'exemple traité par cet auteur, celui-ci constate que l'analyse des notes brutes et des notes transformées en "rankits" conduit pratiquement aux mêmes conclusions.

E - La méthode de M. G. Kendall.

On l'exposera en suivant l'exemple donné par l'auteur dans [30]. On pourra aussi consulter Youden [44].

Un juge est chargé de classer les produits A, B, C, D, E, F, en effectuant les 15 comparaisons deux à deux. Dans une comparaison, le rang 1 est attribué au produit préféré et le rang 0 à l'autre ; le rang 1/2 correspond au cas où il n'y a pas de préférence : c'est ce qui a lieu dans la comparaison (en fait non effectuée) du produit avec lui-même.

Les résultats sont présentés dans une "matrice des rangs", telle que la suivante :

	A	B	C	D	E	F	Total
A.....	1/2	1	1	0	1	1	4 1/2
B.....	0	1/2	0	1	1	0	2 1/2
C.....	0	1	1/2	1	1	1	4 1/2
D.....	1	0	0	1/2	0	0	1 1/2
E.....	0	0	0	1	1/2	1	2 1/2
F.....	0	1	0	1	0	1/2	2 1/2

Le classement des produits est donné par les sommes de rangs (totaux de lignes ou "scores") :

A, C } 4 1/2 B, E, F } 2 1/2 D : 1 1/2

Kendall observe que ce classement présente certaines anomalies : par exemple, A a été préféré à C et cependant se trouve *ex œquo* avec celui-ci ; D occupe la dernière place et cependant a été préféré à A qui est classé en tête (on reviendra dans un paragraphe suivant sur le problème des "classements incohérents", spécialement étudié par le même auteur).

Kendall propose d'"améliorer" les scores en adoptant la règle suivante : pour chaque produit, le nouveau score s'obtient en ajoutant le

score des produits auxquels il a été jugé supérieur et la moitié du score des produits auxquels il a été jugé équivalent. Par exemple, pour A et B, on obtient :

A. 1/2 (4 1/2) + 2 1/2 + 4 1/2 + 0 + 2 1/2 + 2 1/2 = 14 1/4.

B. 0 + 1/2 (2 1/2) + 0 + 1 1/2 + 2 1/2 + 0 = 5 1/4.

Les scores initiaux et les nouveaux scores sont indiqués dans les 1ère et 2ème lignes du tableau ci-après.

	A	B	C	D	E	F	Classement
1...	4 1/2	2 1/2	4 1/2	1 1/2	2 1/2	2 1/2	AC BEF D
2...	14 1/4	5 1/4	11 1/4	5 1/4	5 1/4	5 1/4	A C BDEF
3...	34 1/8	13 1/8	26 5/8	16 7/8	13 1/8	13 1/8	A C D BEF
4...	83 1/16	36 9/16	69 9/16	42 9/16	36 9/16	36 9/16	A C D BEF

En appliquant une deuxième et une troisième fois le même procédé, on obtient les scores qui figurent aux lignes 3 et 4 du tableau.

On démontre que les scores successifs peuvent s'obtenir en élevant aux puissances 2, 3, 4, ... la matrice des rangs initiaux, et en sommant les résultats par ligne. On démontre aussi que les scores ainsi calculés convergent vers des valeurs limites (dans l'exemple ci-dessus on aboutit à la stabilité du classement dès la 3e puissance).

Kendall donne de cette méthode une interprétation géométrique. Les 6 sommets de l'hexagone A B C D E F sont joints deux à deux, et le sens de la flèche indique le sens de la préférence (A $\longrightarrow$ B signifie que A est préféré à B). Le score initial d'un produit est le nombre de flèches partant du sommet correspondant augmenté de 1/2 (qui correspond à la 1/2 préférence A $\longrightarrow$ A).

Le score modifié de A s'obtient en comptant le nombre de circuits de deux lignes successives, dont la 1ère est issue de A, et dont les flèches sont orientées dans le sens des préférences.

On trouve ainsi les 10 circuits.

A $\longrightarrow$ B $\longrightarrow$ D ; A $\longrightarrow$ B $\longrightarrow$ E ; A $\longrightarrow$ C $\longrightarrow$ B ; A $\longrightarrow$ C $\longrightarrow$ D ;
A $\longrightarrow$ C $\longrightarrow$ E ; A $\longrightarrow$ C $\longrightarrow$ F ; A $\longrightarrow$ E $\longrightarrow$ D ; A $\longrightarrow$ E $\longrightarrow$ F ;
A $\longrightarrow$ F $\longrightarrow$ B ; A $\longrightarrow$ F $\longrightarrow$ D ;

auxquels on ajoute (comptant pour 1/2) les circuits tels que :

A $\longrightarrow$ A $\longrightarrow$ B, A $\longrightarrow$ B $\longrightarrow$ B, etc... (8 circuits)

et comptant pour 1/4, le circuit A $\longrightarrow$ A $\longrightarrow$ A.

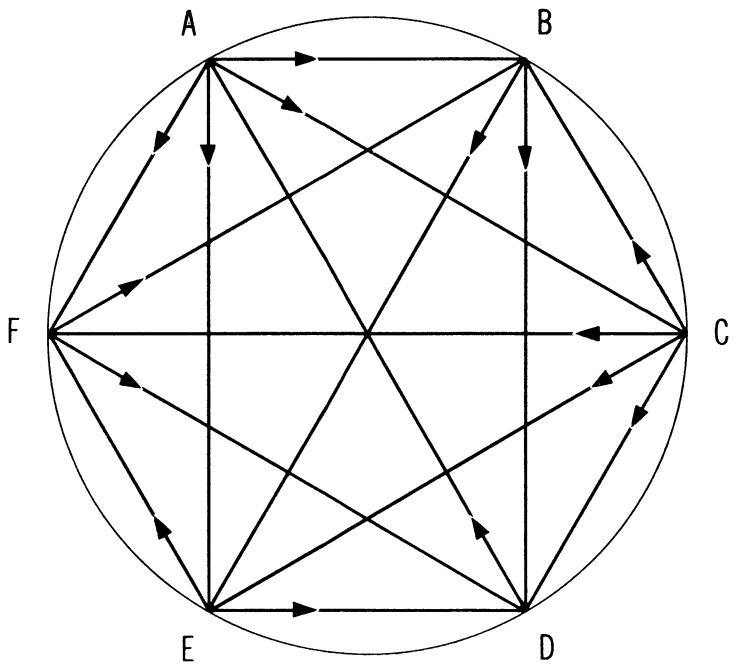
Le nouveau score est ainsi égal à 14 1/4.

Cette façon de procéder élargit, suivant Kendall, la notion de comparaison par paires ; 2 produits tels que A et B sont comparés, non pas directement, mais "à travers" les autres produits (ACB, ADB, AEB, AFB). Elle se généralise en considérant des circuits de plus de deux lignes.

Lorsque plusieurs sujets effectuent des comparaisons, une matrice des rangs peut être établie pour chacun d'eux. La façon la plus simple d'obtenir un classement moyen consiste à additionner ces matrices, puis à prendre pour scores les totaux de lignes. Pour obtenir des scores "améliorés" suivant le procédé qui a été décrit, on a le choix entre deux méthodes :

- élever aux puissances successives chacune des matrices initiales, puis additionner les matrices ainsi recalculées ;

- sommer les matrices initiales et élever cette matrice-somme aux puissances successives.



Kendall conseille la deuxième méthode qui est plus simple, mais sans se déclarer en mesure de justifier cette position.

Une comparaison expérimentale des deux méthodes a été faite par N. T. Gridgeman [20]. L'expérience porte sur 6 arômes et 4 sujets, d'où 4 matrices initiales 6×6, dont les totaux de lignes donnent les scores applicables aux 6 produits ; l'auteur imagine deux autres groupes de 4 matrices 6×6 contenant des valeurs numériques (0, 1, 1/2) différentes, mais telles que les scores applicables aux différents produits conservent les mêmes valeurs que dans le premier cas. Les matrices sont élevées jusqu'à la puissance 6. Sauf dans un cas, les deux méthodes donnent, pour les trois groupes considérés, des scores finaux très voisins.

N. T. Gridgeman est d'avis que la méthode de Kendall modifie peu, ou rarement, les scores initiaux. Il lui oppose une objection d'ordre fondamental : si l'on considère que les produits peuvent être caractérisés

par des paramètres susceptibles d'une interprétation probabiliste, la matrice initiale constitue un échantillon "au hasard" des matrices susceptibles d'apparaître par répétition des épreuves, et il n'y a pas de raison de chercher à l'améliorer par telle ou telle manipulation des données.

F - Comparaison des différentes méthodes.

Plusieurs auteurs ont appliqué deux ou plusieurs des méthodes qui viennent d'être décrites à un même ensemble de données expérimentales. La comparaison la plus complète figure dans Bliss [4]. Elle concerne 4 produits (combinaisons factorielles de 2 insecticides et de 2 fongicides appliqués à des pommiers) ; les 6 comparaisons par paires (classement suivant le goût de compotes fabriquées avec les pommes récoltées) ont été faites 6 fois (6 journées de dégustation) par chacun de 25 dégustateurs différents (900 résultats expérimentaux).

Pour rendre comparables les paramètres attachés aux différents modèles, on a effectué un changement d'origine tel que l'un des produits soit toujours caractérisé par la valeur 1.

Le tableau ci-après donne les valeurs des paramètres.

	T_1	T_2	T_3	T_4
Thurstone - Ecarts normaux ($\hat{S}_i$).....	0,11	1,00	-0,41	-0,70
Bliss - Rankits (*), 1ère méthode.....	0,11	1,00	-0,41	-0,70
2e méthode.....	0,08	1,00	-0,41	-0,67
Scheffe - Degrés de préférence, notes...	0,04	1,00	-0,29	-0,75
" rankits...	0,06	1,00	-0,29	-0,77
Bradley-Terry - ln p_i.....	0,04	1,00	-0,40	-0,64
Kendall - Scores, matrice initiale.....	0,11	1,00	-0,43	-0,68
matrice élevée au carré.....	0,18	1,00	-0,44	-0,74
(*) Bliss propose deux méthodes d'analyse des résultats.				

Dans tous les cas l'ordre des produits est le même ; d'autre part, les valeurs numériques obtenues pour les paramètres sont généralement voisines.

On ne peut donc pas conclure qu'une méthode est préférable aux autres. Le choix de la méthode à appliquer dans un cas particulier doit être fixé en fonction de la plus ou moins grande complexité des calculs à effectuer, des tables dont on dispose (écarts normaux, "rankits", tables de Bradley) et des renseignements que l'on désire tirer de l'analyse : possibilité de tester les résultats dans leur ensemble, de tester des sous-groupes à l'intérieur des répétitions complètes (dégustateurs différents, différentes séances de comparaison, etc.), de tester l'ordre de présentation, la convenance du modèle adopté, etc.

IV - CLASSEMENT OU COTATION SIMULTANES DE PLUS DE DEUX PRODUITS

A - Méthode par cotation.

La cotation d'une série de produits soulève les objections qui ont été faites plus haut à la méthode de Scheffe. C'est sans doute la raison pour laquelle on ne trouve que peu d'études statistiques récentes sur cette question.

Parmi les travaux relativement anciens, on peut citer ceux de J. W. Hopkins [23], [24], de D. D. Mason and E. J. Koch [32] et de C. I. Bliss, M. L. Greenwood and M. H. McKenrick [5].

Il est conseillé de ne pas inclure plus de 6 -à la rigueur 8 ou 9- produits dans la même comparaison. Différentes échelles peuvent être utilisées. Par exemple [24] :

échelle de 0 à 5 pour le degré d'amertume de jus de tomates ;

" −5 à +5 " de sensation salée des mêmes produits.

(les notes négatives correspondent à l'insuffisance, les notes positives à l'excès, le zéro à l'optimum).

Il y a intérêt à organiser l'expérience en choisissant l'un des nombreux plans qui peuvent être interprétés par la méthode d'analyse de la variance. Hopkins observe que l'échelle des notes ayant une étendue limitée, moyenne et variance ne sont pas indépendantes ; la variance a tendance à diminuer aux extrémités de l'échelle, à être maximum vers le milieu ; il suggère de stabiliser la variance par une transformation du type arc sin $\sqrt{x}$.

En définitive, bien que l'application de la méthode d'analyse de la variance à des notes données à partir d'appréciations sensorielles ne soit pas statistiquement orthodoxe, elle ne paraît pas soulever d'objections très graves lorsque le nombre de répétitions est suffisamment élevé (20 à 25 dans les expériences décrites par Hopkins).

B - Méthode par classement.

Les m objets sont rangés dans l'ordre de préférence (ou d'intensité croissante de la sensation éprouvée). Après avoir transformé les rangs (1 à m) en "rankits", on peut appliquer à l'ensemble des mn résultats (n répétitions) la méthode d'analyse de la variance, particulièrement intéressante lorsque le plan d'expérience permet l'analyse simultanée de plusieurs facteurs : M. L. Greenwood and R. Salerno [15].

Le classement de m produits, problème qui se pose notamment aux n membres d'un jury, peut faire l'objet d'un test global. Les rangs de classement sont inscrits dans un tableau à double entrée.

A l'intérieur de chaque ligne, les rangs sont, dans un ordre quelconque, les nombres entiers de 1 à m (leur total est $m(m+1)/2$).

La comparaison porte sur les totaux de colonnes $R_i . S_i$ tous les experts ont donné le même classement, ces totaux sont, dans un certain

Produits / Experts	T_1		T_i		T_m	Total
E_1.....	r_{11}		r_{1i}		r_{1m}	$\frac{m(m+1)}{2}$
E_j.....	r_{j1}		r_{ji}		r_{jm}	$\frac{m(m+1)}{2}$
E_n.....	R_{n1}		r_{ni}		r_{nm}	$\frac{m(m+1)}{2}$
Total.....	R_1		R_i		R_m	$\frac{nm(m+1)}{2}$

ordre, les nombres n, 2n, 3n, ..., mn. A l'opposé, les classements des experts peuvent se compenser de telle sorte que tous les totaux soient voisins de $\frac{n(m+1)}{2}$.

L'hypothèse nulle (H_o) que l'on teste est l'hypothèse que, pour chaque juge, les rangs ont été attribués au hasard (c'est-à-dire en donnant une équi-probabilité aux m! permutations des m rangs). On calcule la quantité :

$$S = \sum_i \left[R_i - \frac{n(m+1)}{2} \right]^2$$

dont les valeurs extrêmes sont 0 (lorsque tous les R_i sont égaux) et $\frac{mn^2(m^2-1)}{12}$ lorsque tous les classements sont identiques, et dont la loi de distribution pour différentes combinaisons (n, m) a été établie dans l'hypothèse nulle. Les tables de S donnent la valeur minimum que cette variable doit atteindre pour que cette hypothèse puisse être rejetée à un seuil de signification donné - c'est-à-dire pour que l'on puisse conclure que les produits sont significativement différents.

V - COHERENCE ET HOMOGENEITE

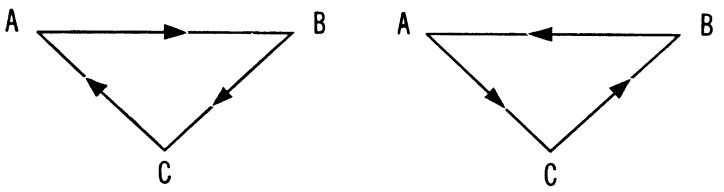
On revient, dans ce paragraphe, sur la notion d'homogénéité, déjà évoquée dans le paragraphe précédent, et l'on introduit la notion de cohérence ; la première s'applique aux écarts constatés entre individus, groupes d'individus, ou groupes de résultats repérés par l'existence d'un facteur contrôlé, tandis que la deuxième concerne chaque individu particulier.

A - Incohérence et ordonnabilité.

Lorsqu'un individu répète une comparaison (A, B) le fait qu'il classe les produits, tantôt dans l'ordre A > B, tantôt dans l'ordre A < B n'est pas considéré comme une "incohérence" ; cela signifie simplement que dans son appréciation intervient une part de "hasard" - si c'est le hasard seul qui la dicte, on constatera à la longue autant de préférences pour A que pour B.

L'incohérence d'un individu peut se manifester dès qu'il compare deux à deux plus de deux produits ; par exemple s'il déclare A > B, A > C, et C > A (non transitivé). En fait, le terme d'incohérence (que nous adoptons, faute de mieux, comme traduction de l'anglais "inconsistence") prête à confusion. En effet, puisque dans la comparaison de chaque paire intervient un élément aléatoire, le phénomène dit d'"incohérence" n'est pas incompatible avec l'existence d'un ordre de préférence se manifestant à l'occasion de la répétition des épreuves (ordonnabilité).

L'ensemble des 3 comparaisons telles que AB, AC, BC, constitue suivant la terminologie de Kendall une "triade". L'incohérence correspond à une "triade circulaire" - c'est-à-dire à des flèches tournant dans le même sens dans la représentation géométrique par un triangle de sommets A, B, C.



Dans le cas de t produits, on peut former $\frac{1}{6}(t-1)(t-2)/6$ triades. Kendall a montré que le nombre de triades circulaires d peut varier de :

$$d = 0 \text{ à } d = t(t^2-1)/24 \quad \text{si t est impair}$$

$$d = 0 \text{ à } d = t(t^2-4)/24 \quad \text{si t est pair.}$$

et il a calculé, sous l'hypothèse nulle, les fréquences attachées aux différentes valeurs de d, pour t = 3, 4, 5, 6. On pourra donc rejeter cette hypothèse (et conclure à l'existence d'un classement non aléatoire) si le nombre de triades circulaires constaté est inférieur à celui qui correspond à une probabilité de dépassement choisie comme seuil significatif.

Si les résultats des comparaisons sont inscrits dans une matrice $t \times t$, avec la valeur 1 pour le produit préféré et la valeur 0 pour l'autre, la diagonale étant elle-même garnie de 0 (cf. matrice donnée à titre d'exemple p. 29, les 1/2 de la diagonale étant remplacés par des 0) on obtient le nombre de triades circulaires à partir des totaux de lignes (scores a_i des différents produits), en appliquant la relation de Kendall :

$$d = \frac{t(t-1)(2t-1)}{12} - \sum \frac{a_i^2}{2}$$

(dans l'exemple cité on trouve d = 5). On voit que d est une fonction linéaire de la somme des carrés des scores (donc pour t donné, de la variance des scores). On peut tester l'hypothèse nulle directement à partir de la quantité $S = \sum a_i^2$ et effectivement Gridgeman [20] a construit une table donnant les probabilités de dépassement attachées aux différentes valeurs de S jusqu'à t = 10.

Par exemple, pour t = 6, on a :

$$\text{Pr.}[S > 55] = \text{Pr.}[d > 0] = 0,000$$

$$\text{Pr.}[S = 55] = \text{Pr.}[d = 0] = 0,022$$

$$\text{Pr.}\,[S \geqslant 53] = \text{Pr.}\,[d \leqslant 1] = 0,051$$

$$\text{Pr.}\,[S \geqslant 51] = \text{Pr.}\,[d \leqslant 2] = 0,120$$

$$\text{Pr.}\,[S \geqslant 49] = \text{Pr.}\,[d \leqslant 3] = 0,208$$

$$\text{Pr.}\,[S \geqslant 47] = \text{Pr.}\,[d \leqslant 4] = 0,398$$

$$\text{Pr.}\,[S \geqslant 45] = \text{Pr.}\,[d \leqslant 5] = 0,509$$

$$\text{Pr.}\,[S \geqslant 43] = \text{Pr.}\,[d \leqslant 6] = 0,773$$

$$\text{Pr.}\,[S \geqslant 41] = \text{Pr.}\,[d \leqslant 7] = 0,919$$

$$\text{Pr.}\,[S \geqslant 39] = \text{Pr.}\,[d \leqslant 8] = 1,000$$

B - Homogénéité.

Chaque fois que cela est possible, il y a intérêt à examiner si les facteurs qui sont susceptibles d'introduire un élément d'hétérogénéité dans les résultats, interviennent de façon significative.

Lorsque les données peuvent être exploitées par une analyse de variance, les facteurs dont on veut connaître l'influence doivent figurer dans le plan d'expérience de telle façon que leur effet puisse être isolé et testé.

Dans les comparaisons par paires, la méthode de Bradley et Terry permet de tester les différences entre plusieurs sous-groupes dont la réunion constitue l'ensemble des répétitions.

Enfin, divers coefficients d'homogénéité ont été proposés, également applicables à la méthode des comparaisons par paires (coefficient de Cochran, coefficient d'agrément de Kendall).

CONCLUSION

De l'exposé qui précède, on doit retenir que la statistique mathématique a pu mettre au point, au cours des dernières années, à l'intention des praticiens et des chercheurs, un nombre assez considérable de méthodes susceptibles de s'adapter à des situations très diverses. Cependant -et cela n'a rien d'étonnant dans un domaine aussi subjectif que celui des caractères organoleptiques- les problèmes de "mesure" sont encore loin d'être parfaitement éclaircis ; leur approfondissement ne pourra se faire que par une étroite collaboration entre les chercheurs et ceux qui sont professionnellement confrontés avec les exigences de l'application pratique.

BIBLIOGRAPHIE

[1] ABELSON R.M. and BRADLEY R.A. - Biometrics, 1954, *10*, 487-502.

[2] BENNET G., SPAHR B. M. and DOODS M.L. - Food Technology, 1956, *10*, 205-208.

[3] BLISS C. I. - Applied Statistics, 1960, *9*, 8-19.

[4] BLISS C. I., GREENWOOD M.L. and WHITE E.S. - Biometrics, 1956, *12*, 381-402.

[5] BLISS C.I., GREENWOOD M.L. and Mc KENRICK M. H. - Food Technology, 1953, *7* , 491-495.

[6] BRADLEY R. A. - Biometrics, 1953, *9*, 22-38.

[7] BRADLEY R.A. - Biometrics, 1954, *10* , 375-391.

[8] BRADLEY R. A. - Biometrics, 1963, *19*, 385-397.

[9] BRADLEY R. A. and TERRY M. F. - Biometrika, 1952, *39*, 324-345.

[10] BRADLEY R. A. - Biometrika, 1954, *41* , 502-537.

[11] BRADLEY R. A. - Biometrika, 1955, *42*, 450-470.

[12] DYKSTRA O. - Biometrics, 1956, *12*, 301-306.

[13] DYKSTRA O. - Biometrics, 1960, *16*, 176-188.

[14] FISHER R. A. and YATES F. - Statistical Tables for Biological, agricultural and medical research., 1953, Oliver and Boyd ed., London.

[15] GREENWOOD M.L. and SALERNO R.- Food Research, 1949, *14* , 314-319.

[16] GRIDGEMAN N. T. - Food Technology, 1955, *9*, 148-150.

[17] GRIDGEMAN N. T. - Biometrics, 1955, *11*, 335-343.

[18] GRIDGEMAN N. T. - Biometrics, 1958, *14*, 458-556.

[19] GRIDGEMAN N. T. - Biometrics, 1959, *15*, 298-306.

[20] GRIDGEMAN N. T. - Biometrics, 1963, *19*, 213-227.

[21] GRIDGEMAN N. T. - Biometrics, 1963, *19*, 398-405.

[22] HARRIS E. K. - Biometrics, 1960, *16*, 245-259.

[23] HOPKINS J. W. - Biometrics, 1950, *6*, 1-16.

[24] HOPKINS J. W. - Biometrics, 1953, *9*, 1-21.

[25] HOPKINS J. W. - Biometrics, 1954, *10*, 391-399.

[26] HOPKINS J. W. - Biometrics, 1954, *10*, 521-530.

[27] HOPKINS J. W. and GRIDGEMAN N. T. - Biometrics, 1955, *11*, 63-68.

[28] JACKSON J. E. and FLECHKENSTEIN M. - Biometrics, 1957, *13*, 51-64.

[29] KAUMAN W. G., GOTTSTEIN J. W. and LANTICAN D. - Biometrics, 1956, *12* , 127-153.

[30] KENDALL M. G. - Biometrics, 1955, *11*, 43-62.

[31] MACKEY A. O. and JONES P. - Food Technology, 1954, *8*, 527-530.

[32] MASON D. D. and KOCH E. J. - Biometrics, 1953, *9*, 39-46.

[33] MOSTELLER F. - Psychometrika, 1951, *16* , 3-9 et 203-218.

[34] PARK G. T. - Biometrics, 1961, *17*, 251-259.

[35] SCHEFFE H. - Journ. Amer. Stat. Assoc., 1952, *47*, 381-400.

[36] SCHUMANN D. E. W. and BRADLEY R. A. - Biometrics, 1957, *13*, 496-510.

[37] SCHUMANN D. E. W. and BRADLEY R. A. - Biometrics, 1959, *15*, 405-416.

[38] TARVER M. G. and ELLIS B. H. - Industrial Quality Control, 1961, *17-12*, 22-26.

[39] THURSTONE L. L. - Psychol. Rev., 1927, *34*, 273-286.

[40] THURSTONE L. L. - Amer. Journ. Psych., 1927, *38*, 368-389.

[41] THURSTONE L. L. - Psychometrika, 1945, *10*, 237-253.

[42] URA S. - Rep. Stat. Appl. Res., J. U. S. E., 1960, *7*, 107-119.

[43] VESSEREAU A. - Revue de Statistique Appliquée, 1953, *1*, 74-81.

[44] YOUDEN W. J. - Revue de Statistique Appliquée, 1962, *10*, 71-75