MATHÉMATIQUES ET SCIENCES HUMAINES

M. BARBUT

Problèmes d'enseignement

Mathématiques et sciences humaines, tome 14 (1966), p. 23-30

<http://www.numdam.org/item?id=MSH_1966__14__23_0>

PROBLEMES D'ENSEIGNEMENT

M. BARBUT

UNE INTRODUCTION AU TEST DU χ^2*

On ne trouvera pas ici une théorie du test statistique du χ^2; ni a fortiori de ses différentes variantes, selon qu'on l'utilise à tester une hypothèse simple ou composite, une forme de distribution, l'indépendance des aléas, etc....

Le but poursuivi est simplement de mettre en évidence le contexte algébrique dans lequel se situe ce test, et de montrer comment son introduction peut se faire aisément à partir des structures fondamentales:
Simplexes,
Espaces vectoriels.

I. - LE PROBLEME A RESOUDRE

Dans une population donnée chaque individu peut appartenir à l'une des n catégories d'un caractère donné. Prenons par exemple le caractère: position de l'individu vis-à-vis du syndicalisme et supposons qu'il y ait trois catégories:

I - être pour le syndicalisme
II - être indifférent au syndicalisme
III- être contre le syndicalisme.

Pour faire un test d'hypothèse nous devons avoir une idée a priori de la proportion d'individus dans chaque catégorie; il s'agit, à l'aide d'un sondage, de savoir si notre hypothèse est plausible. C'est à la résolution de ce type de problème que sert le test du χ^2.

Prenons des données numériques pour plus de clarté :
- opinion a priori sur la répartition du caractère :

$$P_1 = \frac{1}{6} \qquad P_2 = \frac{4}{6} \qquad P_3 = \frac{1}{6}$$

* Cette introduction doit être complétée par la lecture de l'article de G. Th. GUILBAUD, "Exercice préparatoire à l'étude du Khi-deux", qui fait suite dans le présent bulletin.

soit la répartition suivante pour 6 individus :

I	II	III
1	4	5

que nous noterons (1, 4, 1).

- sondage effectué sur 6 individus

- résultat obtenu :

I	II	III
2	3	1

noté (2, 3, 1)

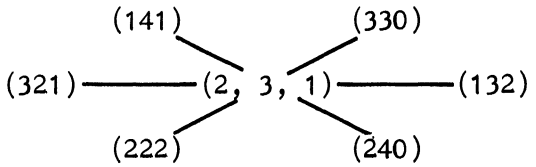
II. - STRUCTURE SUR L'ENSEMBLE DES RESULTATS EVENTUELS.

Avant de répondre à la question posée: "notre opinion (1, 4, 1) sur la répartition du caractère est-elle plausible étant donné le résultat (2, 3, 1) du sondage ou bien, devons-nous modifier cette opinion?", avant même d'effectuer le sondage, nous devons:

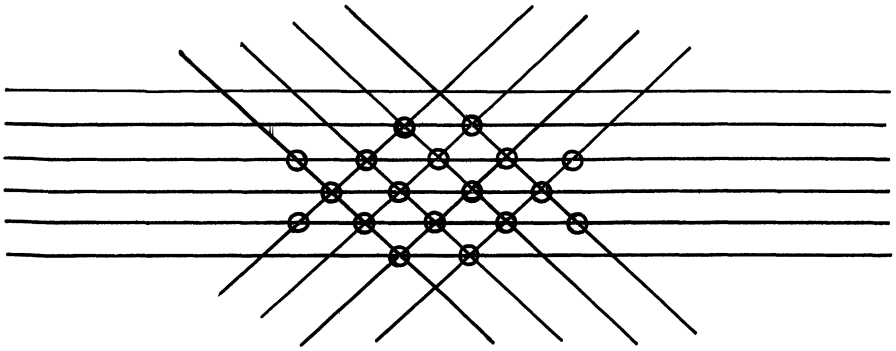
1) nous représenter l'ensemble des éventualités qui pourraient résulter du sondage (ici c'est l'ensemble des triplets (x, y, z) de nombres entiers positifs ou nuls avec $x + y + z = 6$);

2) "ordonner" cet ensemble d'une manière intéressante pour notre problème. Ceci nous permettra en effet de situer de façon précise, dans l'ensemble des résultats possibles du sondage, le constat qui sera fait; et plus précisément de situer ce constat par rapport à notre opinion a priori (1, 4, 1).

Pour "ordonner" cet ensemble d'éventualités il est naturel de dire que les éventualités les plus proches de (2, 3, 1) sont celles qui en résultent par transfert d'un individu d'une catégorie dans l'autre. Les éventualités les plus proches de (2, 3, 1) sont ainsi :



Cela définit une "relation de proximité" entre les triplets (x, y, z): Les voisins de (x, y, z) sont, si x, y et z sont strictement positifs, les 6 triplets de la forme (x-1, y+1, z), (x-1, y, z+1), etc...; nous voyons apparaître un "réseau" où chaque (x, y, z) aura en général 6 voisins. Ce réseau peut être figuré sur un papier triangulaire.



Chaque point du réseau figure un triplet (x, y, z) entouré de ses "voisins". En fait nous n'avons qu'un nombre fini d'éventualités, le réseau est limité et certains points ont moins de 6 voisins. Quatre voisins seulement si l'une des coordonnées est nulle; deux voisins si deux coordonnées sont nulles. Par exemple:

(2, 3, 1)

(3, 2, 1) — (3, 3, 0) — (2, 4, 0) (5, 0, 1) — (6, 0, 0) — (5, 1, 0)

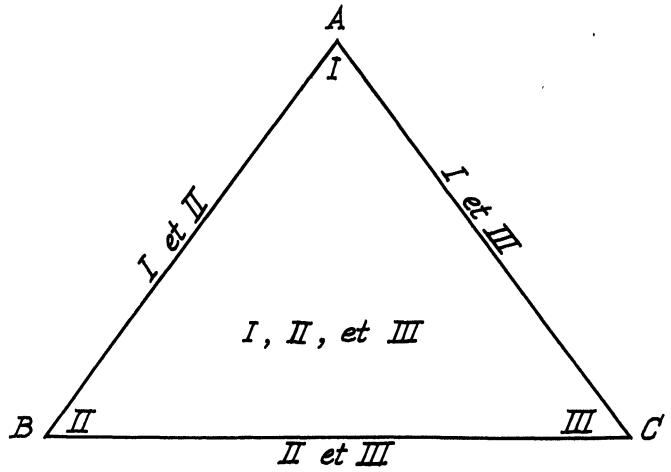
(4, 2, 0)

Traçons le réseau complet. (6, 0, 0), (0, 6, 0), (0, 0, 6) sont les sommets A, B, C d'un triangle (ils n'ont que deux voisins).

en A tous les individus sont dans la catégorie I
en B tous les individus sont dans la catégorie II
en C tous les individus sont dans la catégorie III

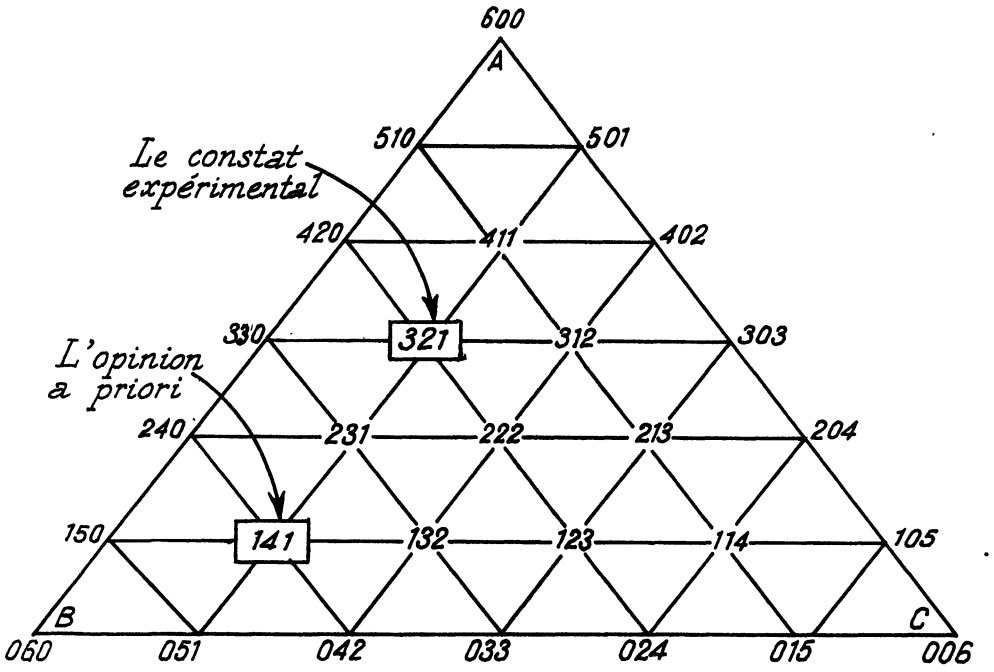
sur la droite AB les individus seront répartis entre les catégories I et II
" AC " " " " " I et III
" BC " " " " " II et III

A l'intérieur du triangle les individus seront répartis entre les 3 catégories.



On reconnait là la représentation en simplexe géométrique de l'ensemble des parties d'un ensemble à 3 éléments, ici l'ensemble {I, II, III}.

parallèlement à la direction BC nous avons x = constante
" " AC " y = constante
" " AB " z = constante



REMARQUE I

Dans la théorie du test χ^2 on dirait ici que c'est un χ^2 à 2 degrés de liberté: notre ensemble d'éventualités a deux degrés de liberté puisque les trois variables x, y, z sont liées par la relation x + y + z = 6. Ce nombre de degré de libertés est celui de la dimension de l'espace auquel appartiennent les vecteurs (x, y, z).

REMARQUE II

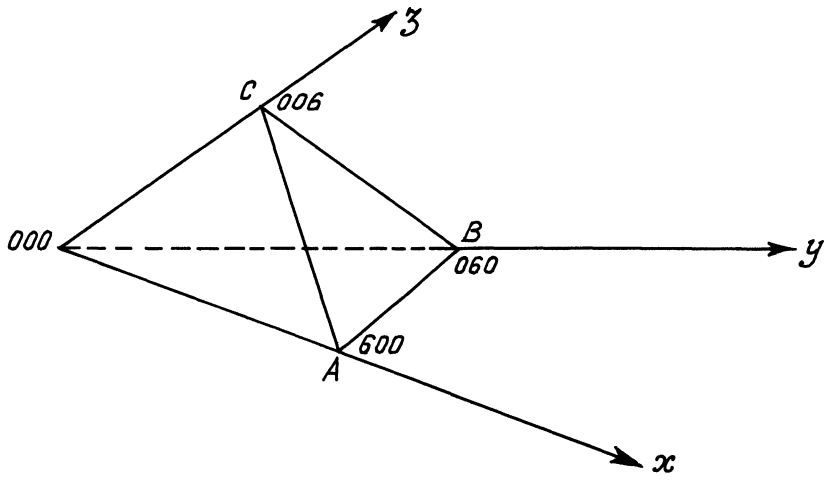
Aux sommets du triangle (simplexe-3) nous avons les trois points (6, 0, 0); (0, 6, 0); (0, 0, 6) ou, si l'on veut, (1, 0, 0); (0, 1, 0); (0, 0, 1) (ce qui revient à considérer les proportions

$$\frac{x}{n}, \frac{y}{n}, \frac{z}{n},$$

avec x + y + z = n).

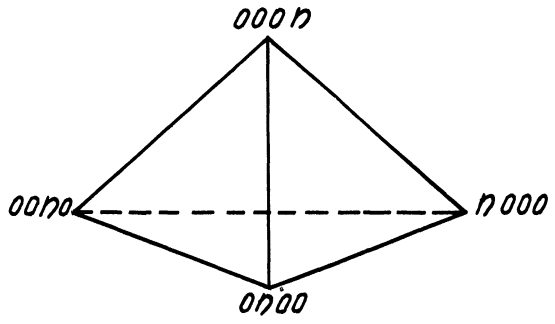
Ces trois vecteurs constituent une base de l'espace vectoriel R^3 et il apparait bien ici qu'algèbre linéaire et simplexe sont étroitement liés. Notre triangle est la section du quadrant $x \geqslant 0$, $y \geqslant 0$, $z \geqslant 0$ de R^3 par le plan

$x + y + z = 6$. (ou $x + y + z = 1$ si l'on se ramène à des pourcentages). Dans le triangle, seuls les points à coordonnées entières nous intéressent ici.



REMARQUE III

Avec 4 catégories pour le caractère étudié, la topologie sur l'ensemble des quadruplets (x, y, z, t) (avec $x + y + z + t = n$) serait construite de la même manière, mais c'est un tétraèdre qui serait "triangulé". Les sommets du tétraèdre correspondraient à (1, 0, 0, 0); (0, 1, 0, 0); (0, 0, 1, 0); (0, 0, 0, 1).



Au delà de 4 catégories même méthode; les $(x_1, x_2, \ldots, x_k)$ sont les "points" à coordonnées entières du simplexe k :

$$x_1 \geqslant 0, \ldots, x_k \geqslant 0, x_1 + x_2 + \ldots + x_k = n$$

l'espace est de dimension k-1 (section de l'espace de dimension k par l'hyperplan, $\Sigma x_i = n$) : le nombre de degrés de liberté est k-1.

III. - MODELE PROBABILISTE

Revenons à notre objectif. Il s'agit de savoir si le résultat (3, 2, 1) est compatible avec notre hypothèse: la composition de la population pour le caractère étudié est $(\frac{1}{6}, \frac{4}{6}, \frac{1}{6})$. Si les 6 individus ont été tirés "au hasard" et si nous pouvons assimiler leur tirage à un tirage (avec remise) de boules dans une urne à 3 catégories, nous avons un certain modèle probabiliste.

Nous prenons pour hypothèse H_0: les proportions d'individus dans chacune des catégories I, II, III sont respectivement $\frac{1}{6}$, $\frac{4}{6}$, et $\frac{1}{6}$, nous connaissons alors la probabilité de chacune des éventualités. Elle est donnée par la loi multinomiale: La probabilité d'avoir le résultat (x, y, z) sous l'hypothèse H_0 est

$$\frac{6!}{x!\ y!\ z!},\ (\frac{1}{6})^x\ (\frac{4}{6})^y\ (\frac{1}{6})^z$$

D'une façon générale, lorsque l'opinion a priori H_0 est donnée par le triplet (p, q, r), elle induit sur l'ensemble des constats possibles (x, y, z) une distribution de probabilité multinomiale :

$$P\ (x, y, z \mid p, q, r) = \frac{n!}{x!\ y!\ z!}\ p^x\ q^y\ r^z$$

A chaque opinion a priori (p, q, r) correspond ainsi une distribution de probabilité sur l'ensemble des éventualités.

IV. - REGLE DE DECISION

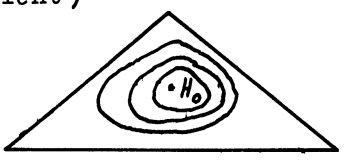
Nous pouvons délimiter autour de (1, 4, 1) qui représente notre hypothèse à tester H_0 une "zone" Z contenant une certaine probabilité. La règle de décision sera la suivante: si le résultat de l'expérience est à l'intérieur de la "zone" nous ne rejettons pas l'hypothèse, si le résultat est en dehors nous la rejettons.

La détermination de la zone où l'on ne rejette pas l'hypothèse peut se faire ainsi*:

1) la zone entoure (1, 4, 1);

2) on se donne une "distance" entre points du triangle, La distance entre le point représentatif de l'hypothèse et le point représentatif du constat est alors un aléa numérique, dont la loi se calcule à partir de la distribution de probabilité multinomiale induite par l'hypothèse H_0.

3) Les valeurs croissantes de la distance définissent autour du point représentatif de H_0 des zones "concentriques"; chacune de ces zones renferme une probabilité (total des probabilités des points qu'elle contient) égale à la probabilité pour que la distance entre hypothèse et constat soit inférieure à la distance correspondant à la frontière de la zone.



* Voir le calcul pratique pour un exercice analogue, dans l'article cité de G. Th. GUILBAUD.

4) On choisit un seuil α ; il lui correspond une zone telle que la probabilité contenue dans Z soit $(1-\alpha)$, la probabilité extérieure à Z soit α.

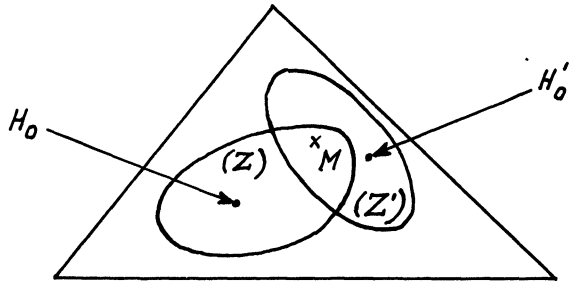
Il n'y a pas de règle générale pour fixer α, sinon dans le cadre de la théorie des décisions statistiques: dans celle-ci il faut tenir compte de ce que coûterait l'erreur commise soit en rejetant l'hypothèse à tort, soit en l'acceptant à tort. Si un coût d'erreur n'a pas de sens dans le problème considéré, le choix de α est assez arbitraire*.

REMARQUE -

Si le constat est à l'intérieur de la "zone", on ne rejette pas l'hypothèse testée; il faut bien remarquer que ce n'est pas équivalent à confirmer cette hypothèse. Nous disons simplement que l'hypothèse faite est compatible avec le résultat de l'expérience, sans plus.

ILLUSTRATION -

Partant d'une hypothèse H_0 nous trouvons la "zone" Z et un résultat d'expérience figuré par le point M.



Nous ne rejettons pas H_0.

Mais si nous étions partis de l'hypothèse H'_0, le résultat de l'expérience aurait également pu nous conduire à ne pas rejeter H'_0, Et pourtant $H_0 \neq H'_0$: il ne faut pas affirmer que l'hypothèse à tester est vraie. Seulement, le test ne nous fournit aucune raison de changer d'opinion.

V. - LE TEST "CLASSIQUE" DU χ^2

Notre étude portant sur un petit nombre d'observations nous avons pu décrire totalement le procédé et expliciter les calculs. Si le sondage porte sur un grand nombre d'individus il se passe des phénomènes nouveaux: On démontre que si la taille n de l'échantillon est grande la loi multinomiale peut être approchée par une loi de Laplace-Gauss à k-1 dimensions lorsque le nombre de catégories est k.

* Sur cette question du choix du seuil, voir: G. Morlat, Statistique et Théorie de la Décision. M.S.H. N° 8, 1964.

La distance entre le résultat observé $(x_1, x_2, \ldots, x_k)$ et l'hypothèse à tester $(x'_1, x'_2, \ldots, x'_k)$ est alors repérée pour des raisons de commodité, par:

$$d\,[(x_1, x_2, \ldots, x_k), (x'_1, x'_2, \ldots, x'_k)] = \sum_{i=1}^{k} \frac{(x_i - x'_i)^2}{x'_i} = Q^2$$

Dans le cas où k = 3: $d[(x, y, z), (p, q, r)] = \frac{(x-p)^2}{p} + \frac{(y-q)^2}{q} + \frac{(z-r)^2}{r}$

La loi de Q^2, qui a priori devrait dépendre de l'hypothèse (p, q, r) de base est en fait indépendante de celle-ci. C'est là la principale des raisons qui font adopter Q^2 comme indicateur de la distance.

Q^2 suit approximativement la loi dite du χ^2 à k-1 degrés de liberté. On ne rejettera pas l'hypothèse à tester si l'écart n'est pas trop grand: plus précisément s'il correspond à un point de la zone Z précédemment définie.

Le test classique du χ^2 est donc basé sur une loi approximative (celle de Laplace-Gauss) et sur un indicateur d'écart commode. Il est applicable si l'effectif de l'échantillon n'est pas trop petit; plus précisément c'est lorsque cet effectif est assez grand que l'usage des tables de χ^2 est licite dans l'application du test: mais on peut effectuer le test avec un sondage très restreint (n=6) pourvu que l'on utilise la loi de probabilité exacte (loi multinomiale).

RESUME