JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

FRED MILHAUD

Essai d'application de la théorie de l'information à la délimitation des valeurs dites normales ou pathologiques en biologie

Journal de la société statistique de Paris, tome 105 (1964), p. 55-60

<http://www.numdam.org/item?id=JSFS_1964__105__55_0>

VIII

VARIÉTÉ

Essai d'application de la théorie de l'information à la délimitation des valeurs dites normales ou pathologiques en biologie

On a trop souvent l'habitude, lorsqu'on veut déterminer si une mesure biologique est significative de quelque chose que l'on appelle anormal, d'appliquer le tableau des $\theta(x)$. On dit alors que 95 % des données recueillies dans une vaste population se situent dans un intervalle de $+ 2\tau$ à $- 2\tau$, 99 % dans un intervalle de $+ 3\tau$ à $- 3\tau$, etc. Volontiers, on qualifie de pathologique une valeur quand son écart à la moyenne dépasse 2τ.

On oublie que le tableau des θ (x) se fonde sur la loi de Laplace-Gauss. Celle-ci se démontre dans le cadre d'un schéma bien précis, celui de Bernouilli, parfaitement réalisé lorsqu'on détermine les fréquences de deux événements contradictoires. On ne peut l'appliquer aux fréquences des valeurs de mesure d'ordre anatomique ou biochimique que si l'on fait un certain nombre d'hypothèses. Il faut admettre que dans l'espèce animale considérée, à l'âge considéré, ce caractère présente, sauf cas pathologiques bien définis, une valeur en quelque sorte essentielle. Il faut admettre qu'un grand nombre de phénomènes, les uns d'ordre physiologique, les autres d'ordre technique et tenant aux procédés de recherches, peuvent introduire de petites variations. Il faut considérer que ces causes de variations sont également probables dans les deux sens. Les variations étant très nombreuses et très petites, on peut raisonner comme si elles n'étaient susceptibles que de deux valeurs opposées, leur nombre étant aléatoire, et on se trouve amené au schéma de Bernouilli. Mais, il y a là tout un ensemble d'hypothèses simplificatrices dont aucune ne présente de garanties de vérités.

Même si la courbe de fréquence des valeurs obtenues semble unimodale, il n'en résulte pas que l'on ait

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \text{ où } x \text{ est exprimé en } \tau$$

ce qui serait nécessaire pour avoir une répartition conforme au tableau des θ (x). Il y a d'autres formules qui donnent des courbes unimodales. Il faut se contenter de parler de courbes fréquence d'allure gaussienne.

Par ailleurs, il apparaît à la réflexion, qu'en biologie, la moyenne a peu d'intérêt quand elle diffère de la médiane. Supposons des dosages d'urée qui, chez les sujets sains, se répartissent entre 0,20 et 0,50, nous voyons qu'un résultat de 0,60 est plus probable qu'un résultat de 0,10. Le phénomène serait encore plus net si on prenait l'exemple des glycémies. On peut facilement imaginer des populations où les valeurs d'une constante biologique nous donnent une moyenne arithmétique qui soit à la limite du pathologique ou même pathologique.

On pourrait essayer de sauver le schéma de Bernouilli en supposant que les nombreuses petites causes de variation sont multiplicatives, de sorte que ce serait des logarithmes qui se répartiraient sur une courbe de Gauss. On resterait dans les hypothèses indémontrées sans grand avantage pratique.

Ce qui importe, c'est la valeur qu'on a autant de chances de dépasser que de ne pas atteindre, donc la médiane.

Nous avons un moyen de connaître la probabilité d'un écart à la médiane, c'est de considérer le rang auquel il correspond.

Classons les sujets selon la valeur du caractère considéré. Donnons le rang 1 au premier et au dernier, 2 au second et à l'avant-dernier, etc. Soit n le nombre de sujets, le rang le plus élevé correspond à la médiane, et il est égal à $\frac{n}{2}$

La détermination du rang pose un problème pratique qui résulte des ex æquo.

Dans les classements habituels, on détermine le rang seulement d'après le nombre de sujets antérieurs. Par exemple, on peut avoir trois troisièmes ex æquo, le suivant sera sixième, et il n'y aura pas de quatrième ni de cinquième. Nous considérons, nous, que les ex æquo le sont par suite de la précision insuffisante de la mesure, et que la moindre erreur est de leur donner à tous le rang moyen parmi ceux qui leur seraient donnés si on les différenciait mieux. Dans l'exemple ci-dessus, il y aurait un 3e, un 4e, un 5e, il semble sage de donner à chacun des 3 sujets le rang 4.

Le rang nous donne une probabilité.

La probabilité d'un rang $\leqslant 1$ est $\frac{1}{n/2} = \frac{2}{n}$

La probabilité d'un rang $\leqslant 2$ est $\frac{2}{n/2} = \frac{4}{n}$

La probabilité d'un rang $\leqslant 3$ est $\frac{3}{n/2} = \frac{6}{n}$ etc.

Bref, la probabilité d'un rang x est $\leqslant \frac{2\,x.}{n}$ Elle est 1 si on considère le rang $\frac{n}{2}$ qui est celui de la médiane.

Une mesure de constante biologique donne une *information* si elle correspond à une faible probabilité.

Mais l'information n'a pas d'intérêt si elle porte sur une seule mesure.

Si sur 100 sujets je prends les deux plus grands et les deux plus petits et que je ne peux rien en conclure, l'information est inutile. Ce qu'il faut pouvoir mettre en évidence, ce n'est pas seulement une valeur peu probable, c'est une conjoncture peu probable. Si, par exemple, un taux de cholestérol de 4 g par litre intéresse le médecin ce n'est pas seulement parce qu'il dépasse largement la médiane, mais c'est surtout parce que lorsqu'on l'observe on est en droit de supposer comme beaucoup plus probable que si on ne l'avait pas observé, que d'autres mesures s'écartent beaucoup de la médiane, par exemple celles du taux des lipides, du taux des lipoprotéides, du taux d'urée ou de la tension artérielle. On disposera d'un nombre aussi élevé que possible de sujets, aussi représentatifs que possible du tout-venant de la population. On déterminera sur chacun d'eux les valeurs simultanées des constantes biologiques que l'on suppose informantes les unes des autres, sans toutefois qu'il y ait entre elles une corrélation appréciable, au sens strict du mot « corrélation ». On classera les sujets selon le procédé indiqué plus haut pour chacune des constantes étudiées.

Chaque sujet y aura donc plusieurs rangs x_i, chacun correspondant à une probabilité $= \frac{2\,x_i}{n} = \varphi_i$

Il réalise une conjoncture de probabilité $\Pi \frac{2\,x_i}{n} = \Phi$,

Il apporte une information $H = -\Sigma \log \varphi_i$.

Soit t le nombre des constantes étudiées, nous avons

$$H = -\left(\Sigma \log x - t \log \frac{n}{2}\right) = t \log \frac{n}{2} - \Sigma \log x$$

Nous constatons que dans telle des constantes étudiées, nous avons s sujets qui nous donnent $\Pi \frac{2\,x_i}{n} \leqslant 0{,}02$

La probabilité de rencontrer un tel ensemble parmi les sujets tirés au sort est

$$\omega = (0{,}02)^s\,(0{,}98)^{\,m-s}\,\frac{m\,!}{s\,!\,(m-s)\,!}$$

Nous choisirons m de façon à avoir $\omega < 0{,}01$.

Les valeurs qui correspondent à des rangs $\leqslant m$ peuvent être considérées comme informantes.

Il est possible que la valeur de m soit différente dans le classement par ordre de valeurs croissantes ou par ordre de valeurs décroissantes.

A plus forte raison, si ensuite nous considérons les mesures des caractères et que nous évaluons en écarts-types la différence à la moyenne, nous pouvons trouver des valeurs très diverses.

Il est impossible de dire à combien de σ au-dessus de la moyenne et à combien de σ au-dessous d'elle doivent commencer les zones informantes. On le saura dans chaque cas après avoir effectué les recherches ci-dessus.

*
* *

Les raisonnements que nous avons adoptés comportent l'abandon de vieilles habitudes intellectuelles. On a tendance à croire, ainsi que nous l'avons rappelé au début, que tout caractère biologique mesurable présente une valeur en quelque sorte essentielle qui correspond à ce qu'on appelle ordinairement la « normale », ce mot ayant alors un sens différent de celui qu'il a en mathématique. Pour nous, il y a seulement une zone dépourvue d'information, du moins provisoirement. Rien n'exclut l'hypothèse que les limites de cette zone viennent à changer si nous demandons un jour à la mesure considérée des informations différentes de celles que nous pensons actuellement à lui demander. Rien n'exclut non plus l'hypothèse qu'il n'y ait aucune limite à la zone non informante des valeurs possibles d'un caractère. En ce cas, les notions d'anomalies ou de valeurs pathologiques n'auront aucun sens en ce qui le concerne.

*
* *

Il est important de bien préciser la nature du problème abordé.

On pouvait se demander :

1° quelle est la probabilité pour qu'il existe m sujets donnant

$$\Pi \frac{2\,x_i}{n} < 0{,}02$$

2° quelle est la probabilité pour que les m sujets qui donnent les valeurs les plus extrêmes dans un sens donné pour la mesure A réalisent

$$\Pi \frac{2\,x_i}{n} < 0{,}02$$

3° quelle est la probabilité pour qu'une population quelconque de m sujets tirés au sort nous donnent tous

$$\Pi \frac{2\,x_i}{n} < 0{,}02$$

C'est la troisième question que nous nous sommes posée pour dire si une population était remarquable. Nous avons fixé m *a posteriori* à une valeur telle que l'inégalité soit réalisée pour les m premiers sujets.

*
* *

Une objection peut venir à l'esprit, à savoir que la valeur de s dans le sous-ensemble de m sujets est aléatoire d'une expérience à l'autre.

On notera que dans l'hypothèse la plus pessimiste, son écart-type est

$$\sigma_s = \sqrt{s \frac{s}{m} \frac{m-s}{m}} = s \sqrt{\frac{m-s}{m^2}} = s \sqrt{\frac{1}{m} - \frac{s}{m^2}}$$

Nos conclusions sont donc valables si quand on remplace s par $s + 2\,\tau_s$, c'est-à-dire lorsqu'on le multiplie par

$1 + 2\sqrt{\frac{m-s}{m^2}} = 1 + 2\sqrt{\frac{m-s}{m}}$, les conditions exigées sont encore réalisées.

C'est une condition supplémentaire très facilement réalisée elle-même car $2\sqrt{\frac{m-s}{m}}$ est habituellement très petit.

* * *

Il reste qu'une information obtenue avec m trop petit serait sans intérêt.

En effet, les sujets d'un rang $< \alpha$ dans une mesure donnent sûrement $\omega \leqslant 0{,}02$ quels que soient leurs rangs dans une autre mesure. Si, par exemple, nous avons 100 sujets avec 2 séries de rangs de 1 à 50, nous avons $\alpha = 1$. Avec 200 sujets, on aurait $\alpha = 2$, etc. Sur m sujets qui se suivent du 1er au m^e, on a r tranches de α sujets, avec $r = \frac{m}{\alpha}$ Dans la 1re tranche, la probabilité d'avoir $\omega = 0{,}02$ est 1. Dans la 2e, elle est de $\frac{1}{2}$. Dans la 3e elle est de $\frac{1}{3}$, etc. Sur l'ensemble, elle est de $\frac{1}{r!}$

Si les m sujets ne se suivent pas, elle est $< \frac{1}{r!}$

Nous exigeons

$$\frac{1}{r!} < 0{,}01 \qquad \text{ou} \qquad r! > 100$$

Nous exigeons donc, pour affirmer qu'une zone des sujets dont m donnent $H > 1{,}699$ est informante, que deux conditions soient réunies

$$(0{,}02)^{s}\,(0{,}98)^{m-s}\,\frac{m!}{s!\,(m-s)!} < 0{,}01$$

$$r! > 100$$

Nos calculs sont fondés sur la notion que les variables confrontées ne sont pas corrélées. Un syndrome, en effet, est constitué par la rencontre simultanée de valeurs extrêmes de variables habituellement non corrélées. Une petite taille et un poids peu élevé ne constituent pas un syndrome, tandis qu'une petite taille, un métabolisme basal bas, un quotient intellectuel faible, en constituent un. Nous nous sommes placés dans l'hypothèse de variables non corrélées en posant la formule

$$\Pi \frac{2\,x_i}{h} = \Phi_j$$

Nous devons donc confronter exclusivement des variables non corrélées.

Cette règle est en pratique une incommodité, on peut essayer de lui apporter des exceptions.

Soient A et B corrélés entre eux et non avec C.

On remplacera A et B par une variable idéale D, en donnant à chaque sujet un rang en D, X*d*, qui soit la moyenne géométrique de son rang X*a* en A et de son rang X*b* en B. On confrontera D et C

$$H = \log x_d + \log x_c - 2 \log \frac{n}{2} = \frac{\log x_a + \log x_b}{2} + \log x_c - 2 \log \frac{n}{2}$$

avec $H \geqslant 1{,}699$ exigible

ou $H = \log x_a + \log x_b + 2 \log x_c - 4 \log \frac{n}{2}$

avec l'exigence

$$H \geqslant 3{,}398.$$

Les sujets qui ont des valeurs informantes en D ont *ipso facto* des valeurs informantes en A et B, et on reconnaît ainsi des zones informantes en A et B.

C'est un artifice qui peut quelquefois rendre service.

Dr Fred Milhaud