156

QUELQUES REMARQUES SUR L'ESTIMATION EN STATISTIQUE (1)

par R. A. Fisher.

Le procédé classique, datant au moins de l'époque de Gauss, pour éprouver la signification de la différence entre la moyenne observée d'un échantillon normal, et zéro ou quelque autre valeur choisie pour la comparaison, est de diviser la différence par son écart-type, estimé d'après l'échantillon. Si $\bar{x}$ est la moyenne observée de n observations et μ la vraie moyenne de la population de laquelle l'échantillon était tiré, alors on sait depuis longtemps que $\bar{x}$ est distribué dans les différents échantillons suivant une distribution normale dont le centre est μ et ayant une variance égale au $\frac{1}{n}$ ième de celle de la population échantillonnée. Si donc, nous connaissions le véritable écart-type σ de cette population, nous saurions que

$$\frac{(x - \mu)\sqrt{n}}{\sigma}$$

est distribuée normalement avec une variance égale à l'unité et ainsi nous pourrions dire avec exactitude la probabilité avec laquelle une valeur donnée serait dépassée.

(1) Communication présentée le 16 mai, à la Société de Biotypologie.

En fait, la vraie valeur σ n'est pas connue, mais nous avons à sa place une estimation tout à fait satisfaisante, s, définie par

$$s^2 = \frac{1}{n-1} S (x - \bar{x})^2$$

où S représente la sommation sur l'échantillon. Cette estimation est en fait suffisante ; mais c'est, néanmoins, un fait que la valeur de s que l'on aura trouvée diffère en général plus ou moins de la vraie valeur σ. En conséquence, si nous substituons s à σ et si nous calculons :

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

nous n'avons pas le droit de dire que t sera normalement distribué. « Student » s'est le premier demandé comment, en fait, le quotient t est distribué quand il est calculé à partir d'un échantillon de n observations. La solution exacte est donnée par la loi de fréquence :

$$df = \frac{\left(\frac{n-2}{2}\right)!}{\left(\frac{n-3}{2}\right)! \sqrt{\pi (n-1)}} \cdot \frac{dt}{\left(1 + \frac{t^2}{n-1}\right)^{n/2}}$$

C'est une distribution très différente au point de vue mathématique de la loi de Gauss, bien qu'elle s'approche progressivement de cette forme quand n croît indéfiniment. La distribution est cependant exacte et il est possible d'en dresser des tables pour chaque valeur de n. En fait, elle a été plusieurs fois complètement mise en Table. En conséquence, au lieu de dire qu'il y a une probabilité d'une chance sur quarante que

$$\frac{(x - \mu)\sqrt{n}}{\sigma} > 1{,}960$$

affirmation qui ne serait utile que si l'on connaissait σ avec exactitude ; on peut également dire si, par exemple, notre moyenne est fondée sur 15 observations que :

$$t = \frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

a une probabilité de un sur quarante de dépasser la valeur 2,145.

Ce résultat est directement utile, car s n'est pas inconnu, mais est calculable avec exactitude d'après les informations.

Armé de ce nouvel instrument, il est naturel, pour les besoins pratiques, de faire un nouveau pas logique, d'une grande importance théorique, c'est-à-dire de se servir du taux par exemple 2,145 approprié au niveau de signification choisi, de multiplier cette quantité par l'écart-type de la moyenne ainsi qu'il a été estimé, d'ajouter ou de soustraire le produit à la moyenne observée et ainsi d'obtenir les limites de travail pour les valeurs de la moyenne inconnue de la population.

En fait, puisque la distribution de t est connue avec exactitude et puisque t est donné par la formule :

$$t = \frac{(\overline{x} - \mu)\sqrt{n}}{s}$$

qui ne comprend, à part μ, que des quantités directement calculables, à savoir $\overline{x}$ et s, toutes les deux estimations statistiques exhaustives, nous pouvons affirmer, sans nous servir des probabilités à priori, une loi de probabilité pour μ qui correspondra à l'ensemble des résultats trouvés plus haut ; c'est-à-dire que la probabilité pour que μ soit moindre que

$$\overline{x} - 2{,}145\,\frac{s}{\sqrt{n}}$$

est exactement un quarantième.

* * *

Nous pouvons maintenant apprécier la nécessité de la condition que j'ai mentionnée, en rapport avec le test de signification de Student, pour la moyenne d'un échantillon normal. A savoir que les quantités x et s qui, avec le paramètre inconnu μ apparaissent dans l'expression de t, devraient être des estimations suffisantes de la moyenne et de l'écart-type de la population échantillonnée.

Car c'est une garantie de ce qu'elles ont fourni toute l'information que l'échantillon doit donner quant à la nature de la population échantillonnée. Si nous nous étions servi d'estimations

alternatives, si, par exemple, nous avions trouvé la médiane au lieu de la moyenne arithmétique, $\bar{x}$, ou si nous nous étions servi de la formule de Peter, fondée sur la déviation moyenne au lieu de la formule de Bessel, fondée sur le carré moyen, nous pourrions déduire un test de signification entièrement valide, c'est-à-dire nous pourrions avoir trouvé une quantité t, avec une distribution connue exactement pour des échantillons de dimensions données et exprimables, comme t' en fonction du paramètre inconnu en même temps que de quantités directement calculables. Mais, si nous étions allé plus loin et si nous avions substitué t' en fonction de μ et déduit une distribution du paramètre inconnu, la distribution que nous aurions obtenue serait basée seulement sur cette part de l'information utilisable, que nos estimations spéciales de la moyenne et de l'écart-type auraient conservées. La distribution obtenue différerait de celle trouvée en se servant d'estimations suffisantes et les lois de probabilité qu'elle incorpore seraient différentes. Si l'on n'exige pas que toute l'information utilisable soit épuisée, une armée de déductions différentes semblent également admissibles, chacune dépendant du choix personnel du statisticien, et à travers son choix, de la méthode d'estimation que l'on doit employer.

Quand l'estimation suffisante est possible, il n'y a pas de problème ; mais on voit que le traitement exhaustif de cas dans lesquels il n'existe pas d'estimation suffisante nécessite une prompte solution. Ce traitement est à présent parfois possible, mais pas toujours autant que je sache. J'ai parlé d'estimations suffisantes comme contenant en elles-mêmes toute l'information fournie par les données. Ce n'est pas strictement exact. Il y a toujours un peu d'information supplémentaire ou auxiliaire dont nous avons besoin, même avec une estimation suffisante avant qu'elle puisse être utilisée.

Cette petite quantité d'information, dont nous avons besoin, est la dimension de l'échantillonage ou, en général, l'étendue des résultats observés. Nous avons toujours besoin de le savoir pour savoir jusqu'à quel point nous pouvons avoir confiance en notre information. Au lieu de prendre pour accordée la dimension de l'expérience, et de dire que la particularité des cas où l'estimation suffisante est possible, repose sur le fait que l'estimation contient

alors toute l'information recherchée, nous aurions bien pu également inverser notre exposé ; et prenant pour accordée l'estimation par la méthode du maximum de vraisemblance, dire que la particularité était que l'on n'avait besoin de rien de plus que de la dimension de l'expérience pour compléter cette information.

Cet aspect de la question retournée est le plus fructueux des deux, une fois que nous nous sommes convaincus que quand il y a de l'information perdue, cette perte est minimisée en se servant du maximum de vraisemblance. Les cas dans lesquels l'estimation suffisante est impossible sont ceux dans lesquels, en utilisant cette estimation, on a besoin d'extraire de l'échantillon une autre information auxiliaire en plus du simple nombre d'observations qui le compose. La fonction que doit remplir cette information auxiliaire est de distinguer parmi les échantillons de mêmes dimensions ceux desquels des estimations plus ou moins exactes peuvent être extraites ; où, en général, de distinguer entre les échantillons ayant différentes fonctions de vraisemblance, même si elles peuvent être rendues maximum par la même valeur. L'information auxiliaire ne modifie jamais la valeur de notre estimation ; elle détermine sa précision.

Le procédé de cette sorte le plus général possible serait à partir d'un échantillon de n observations de spécifier :

a) l'estimation ou l'ensemble d'estimation des paramètres inconnus ayant la plus grande vraisemblance ;

b) un ensemble de statistiques auxiliaires fonctionnellement indépendantes suffisantes en conjonction avec (*a*) pour permettre aux observations d'être reconstruites en entier et ayant la propriété additionnelle que ces quantités auxiliaires soient toutes distribuées dans les échantillons en distributions indépendantes de paramètres inconnus. Il est aisé de voir que cela peut être fait dans certains cas simples. Par exemple, si μ est le seul paramètre inconnu dans une loi de distribution dont l'élément différentiel est :

$$df = \Phi\,(x - \mu)\,dx$$

les différences entre des observations successives, quand celles-ci sont rangées par ordre de grandeur, fournissent $x - 1$ quantités fonctionnellement indépendantes, calculables d'après les résultats

expérimentaux ; la loi de distribution de chacune d'elles étant évidemment indépendante de μ. Nous pouvons donc regarder un tel ensemble de différences comme spécifiant la configuration de l'ensemble des résultats et en interprétant notre estimation prendre pour sa loi de distribution celle appropriée aux seuls résultats ayant la configuration observée.

Il y a alors un second groupe de solutions par lequel l'estimation peut être faite exhaustive, dépendant comme les statistiques suffisantes, d'une relation fonctionnelle spéciale et, comme elles, résolvant une classe très étendue de problèmes qui se posent dans la pratique.

Et mon mot final sur ce point sera une question dont la réponse, autant que je sache, est inconnue et qui est donc, pour le moment, un défi à notre intuition mathématique :

Puis-je la poser sous cette forme :

Le terrain cultivable d'un village égyptien pré-dynastique est d'inégale fertilité. Étant donnée la hauteur à laquelle le Nil s'élèvera, la fertilité de chaque portion est connue avec exactitude, mais la hauteur de l'inondation affecte les différentes parts du territoire d'une façon inégale. On demande de diviser l'aire entre les différentes familles de ce village de telle sorte que les récoltes des lots assignés à chacun soit dans une proportion déterminée, quel que soit le niveau auquel la rivière s'élève.

Si ce problème est susceptible d'une solution générale, alors il est possible en général de reconnaître quelque chose qui correspond à la configuration de l'échantillon dans le cas simple discuté ci-dessus et un des premiers problèmes de la discussion incertaine aura trouvé sa solution complète.

Sinon, il doit rester encore d'autres problèmes à élucider.