SUR L'INTERPRÉTATION DES RÉSULTATS STATISTIQUES

PAR

H. STEINHAUS (WROCŁAW)

En lisant l'Introduction aux *Erreurs statistiques moyennes* de M. Ritala [1]), mon attention a été attirée par une discussion qui m'a parue intéressante du point de vue des principes; comme elle touche aux questions qui se posent dans la pratique, le fait que la réponse ne comporte pas de formules compliquées ni de raisonnements laborieux me sert de motif pour la publier au risque d'enfoncer une porte ouverte à tout mathématicien; la discussion citée m'a convaincu qu'elles ne l'est qu'à demi pour les personnes dont le métier est d'appliquer les résultats statistiques et non pas d'en analyser les fondements logiques.

Les tables dressées par M. Ritala, médecin finnois, servent pour l'évaluation rapide de l'erreur moyenne („standard error") qu'il faut attacher à un résultat statistique. Je ne parlerai ici que de la Table I [2]); elle donne $\varepsilon=\sqrt{\frac{pq}{n}}$ en fonction de p et de n. La question pratique est d'évaluer l'erreur moyenne ε en connaissant le pourcentage $P=100p$ calculé d'après une statistique embrassant n individus dont $\nu=np$ ont réussi dans l'épreuve.

Il y a ici trois problèmes différents.

Le premier est celui de la détermination du pourcentage observé

$$P=100\frac{\nu}{n}:$$

Cette détermination ne fait qu'exprimer numériquement le résultat d'observation. Ce résultat est hors de doute et la question d'une erreur quelconque est privée de sens; pensons à 30 personnes malades traitées de la même manière, dont la moitié a survécu; le traitement a réussi dans 50% des cas observés: c'est tout ce qu'on peut dire.

[1]) A. M. Ritala, *Zur Berechnung des statistischen mittleren Fehler (Standard error)*, Acta Societatis Medicorum Fennicae „Duodecim", Ser. B. 19, Fasc. 2, Helsinki 1933 (cf. aussi ce volume p. 55).

[2]) Op. cit., p. 31-46.

Le second problème, *problème direct* — comme nous l'appellerons — est de déterminer le pourcentage attendu des réussites quand on connait la probabilité de la réussite; si l'on sait qu'elle est égale à p, ce pourcentage est égal à

$$P=100p=100E\left(\frac{\nu}{n}\right) \tag{1}$$

et il est soumis à l'erreur moyenne ε donnée par la formule classique

$$\varepsilon=\sqrt{\frac{PQ}{n}} \qquad (Q=100-P), \tag{2}$$

ces formules étant à l'abri de toute objection. Supposons que $P=100\%$; cela signifie que la probabilité de la réussite est 1, donc la réussite est sûre. On obtient

$$100E\left(\frac{\nu}{n}\right)=100, \qquad \varepsilon=\sqrt{\frac{100\times 0}{n}}=0;$$

cela veut dire que le pourcentage attendu sur n épreuves est 100% avec l'erreur moyenne zéro, ce qui est bien d'accord avec le bon sens — puisqu'il est certain que l'on aura n réussites sur n épreuves et qu'aucun écart n'est possible. Si l'on avait $p=1/2$, $P=50\%$ et $n=30$, les mêmes formules donneraient

$$100E\left(\frac{\nu}{30}\right)=50, \qquad \varepsilon=\sqrt{\frac{50\cdot 50}{30}}=9{\cdot}15;$$

on écrirait alors $(50\pm 9{\cdot}15)\%$ pour le pourcentage de réussites à prévoir dans une série de 30 essais.

Comment se fait-il que l'on trouve aux pages 22-24 de l'Introduction citée un reproche à la formule (2)?

„In den Fällen, wo $P=0\%$ (oder $P=100\%$) erhalten wir nach der Formel $\varepsilon=\sqrt{\frac{PQ}{n}}$ fur ε den Wert $\pm\sqrt{\frac{0\times 100}{n}}=\pm 0$. Ist also $P=0\%$, so ist sein mittlerer Fehler jedenfalls 0, ganz unabhängig davon, ob die Statistik 100 oder 1000 Fälle enthält. Das stimmt aber nicht mit der statistischen Erfahrung überein. Man ist keinesfalls berechtigt, wenn man nur wenige Fälle (z. B. 30 Operationen, alle mit Erfolg ausgeführt) behandelt hat, die Regel aufzustellen, dass bei dieser Operation die Mortalität gleich Null ist. Das %-Resultat kann ja immerhin (sei es auch $=0$)

bloss einen auf einem Zufall beruhenden, vielleicht exzeptionellen Wert darstellen. Formeln, die solche vernunftwidrige Resultate geben, müssen zurücktreten, und wir sind gezwungen auf andere Weise zu versuchen, die Zuverlässigkeit der Resultate in diesen Fällen ($p=0^0/_0$ bzw. $p=100^0/_0$) zu prüfen" [3]).

En effet, on ne saurait nier que 30 opérations réussies ne suffisent pas pour affirmer avec certitude que la mortalité dans ce genre d'opérations est nulle. Quelle est donc la raison du paradoxe?

La réponse est immédiate: il s'agit ici du troisième problème — *problème inverse* — qui consiste à déterminer quelle est la probabilité d'un événement quand on sait qu'il s'est produit ν fois sur n. On n'est pas autorisé à substituer à ce problème le problème direct, ce qui explique pourquoi les tentatives de le résoudre par les formules (1) et (2) conduisent à des résultats incongrus.

Pour exprimer plus facilement le problème inverse, je parlerai de l'efficacité d'une drogue, en entendant par là la probabilité de son action favorable; on suppose que cette action se manifeste par un symptôme qui peut se produire ou non, le tiers étant exclu. Théoriquement, cette probabilité est égale à la limite de la fréquence relative de réussites dans une suite infinie d'essais; en pratique, on ne connaît l'efficacité qu'approximativement. Pour la déterminer, on peut appliquer la règle de Bayes en oubliant, pour le moment, les objections de toute sorte qui peuvent s'y opposer. Supposons que le chimiste puisse donner à sa drogue une efficacité quelconque entre 0 et 1, et qu'il choisisse cette efficacité pour chaque charge séparément en faisant tourner une sorte de roulette dont le disque porte une graduation uniforme de $0^0/_0$ à $100^0/_0$; une boule indique par son arrêt l'efficacité pour aujourd'hui.

La règle de Bayes répond à trois questions importantes:

1° Quelle est l'efficacité à attendre $E(p)$, si la drogue a agi ν fois sur n? Le sens de cette question est le suivant: quand on ne retient que les échantillons qui ont agi ν fois sur n et quand on calcule leur efficacité moyenne (en prenant la moyenne arithmétique des efficacités), cette moyenne tend vers une limite qui est bien l'efficacité expectée, $E(p)$.

[3]) loco cit., p. 22-23.

2° Quelle est l'erreur moyenne de p? Cette grandeur, $\varepsilon(p)$, est définie par la formule

$$\varepsilon^2(p)=E(p^2)-(E(p))^2.$$

3° Quelle est la probabilité β pour que p dépasse une borne donnée α? On peut définir cette probabilité d'une manière analogue à 1°.

Les formules à employer sont maintenant:

$$1^0\quad E(p)=\int_0^1 x^{\nu+1}(1-x)^{n-\nu}\,dx : \int_0^1 x^{\nu}(1-x)^{n-\nu}\,dx=\frac{\nu+1}{n+2};$$

on aura donc, en désignant $100\,E(p)$ par P,

$$P=\frac{\nu+1}{n+2}\cdot 100; \tag{3}$$

$$2^0\quad \varepsilon^2(p)=\int_0^1 x^{\nu+2}(1-x)^{n-\nu}\,dx : \int_0^1 x^{\nu}(1-x)^{n-\nu}\,dx-(Ep)^2=$$

$$=\frac{\nu+1}{n+2}\,\frac{n-\nu+1}{n+2}\cdot:(n+3);$$

on aura donc, en désignant $100\,\varepsilon(p)$ par $\varepsilon(P)$ et $100-P$ par Q,

$$\varepsilon(P)=\sqrt{\frac{PQ}{n+3}}; \tag{4}$$

$$3^0\quad \beta=\int_\alpha^1 x^{\nu}(1-x)^{n-\nu}\,dx : \int_0^1 x^{\nu}(1-x)^{n-\nu}\,dx;$$

cette formule se simplifie dans le cas $\nu=n$ et devient

$$\beta=1-\alpha^{n+1}. \tag{5}$$

Examinons les exemples suivants:

28 29 30 réussites

respectivement sur

30 30 30 essais.

Les formules (1) et (2) donnent pour efficacités respectives

$$87{\cdot}11\pm 6{\cdot}12^0/_0,\qquad 96{\cdot}68\pm 3{\cdot}28^0/_0,\qquad 100\pm 0^0/_0, \tag{6}$$

tandis que les formules (3) et (4) fournissent les valeurs

$$90{\cdot}63\pm 5{\cdot}07^0/_0,\qquad 93{\cdot}75\pm 4{\cdot}21^0/_0,\qquad 96{\cdot}89\pm 3{\cdot}02^0/_0. \tag{7}$$

C'est le dernier nombre du résultat (6) qui a soulevé la méfiance. On voit le paradoxe disparaître dans (7); en effet, l'efficacité de $96{\cdot}89 \pm 3{\cdot}02\%$ qui correspond aux 30/30 réussites n'a plus rien de choquant pour le bon sens.

Calculons encore la probabilité pour que l'efficacité dans le cas 30/30 dépasse la borne $96{\cdot}89+3{\cdot}02=99{\cdot}91\%$. Cette probabilité β est donnée par (5) avec $\alpha=0{\cdot}9991$ et $n=30$; elle est donc égale à

$$1-0{\cdot}9991^{31}=1-0{\cdot}9731=0{\cdot}0269,$$

donc moindre que 3%. Pour une variable distribuée selon la loi de Gauss, la probabilité qu'elle dépasse sa valeur moyenne plus l'erreur moyenne positive est environ 6 fois plus grande que celle que nous avons trouvée. Il faut donc se garder d'attribuer à l'erreur moyenne calculée d'après les formules exactes les propriétés de la distribution gaussienne; ces propriétés font défaut quand le nombre ν est proche de 0 ou de n.

M. Ritala mentionne une formule donnée par un médecin de mérite, M. H. Poll, et qui doit remplacer (2) quand $\nu=0$ ou $\nu=n$, à savoir $\nu=\dfrac{900}{n+9}$; je ne peux pas juger la valeur de cette formule, son explication étant trop succincte pour que je puisse la comprendre; l'auteur des Tables ne l'accepte pas car elle conduit à un $\varepsilon(p)$ égal à 3 quel que soit le nombre des essais. On peut supposer que si M. Ritala avait consulté à ce propos les mathématiciens de Helsinki, comme J. W. Lindeberg ou R. Nevanlinna, dont il cite les récommandations dans l'Introduction, il aurait évité cette difficulté.

M. Ritala propose la formule

$$\varepsilon(P)=\pm\frac{100}{n}\sqrt{\frac{n-1}{n}}$$

qu'il simplifie ensuite:

$$\varepsilon(P)=\pm\frac{100}{n}. \tag{8}$$

On voit que cette formule diffère peu de ce que donne la nôtre (4) pour $\nu=n$; elle devient dans ce cas

$$100\,\frac{1}{n+2}\sqrt{\frac{n+1}{n+3}}.$$

Or, il y a une différence essentielle: M. Ritala réserve sa formule (8) pour les cas $\nu=0$ et $\nu=n$; nous appliquons (4) dans tous les cas. Les nombres (6) deviennent moyennant (8)

$$87{\cdot}11 \pm 6{\cdot}12\%, \qquad 96{\cdot}68 \pm 3{\cdot}28\%, \qquad 100 \pm 3{\cdot}33\%. \tag{9}$$

Une autre différence apparaît en comparant (9) à (7): nos formules ne donnent jamais l'efficacité de 100%.

Il est temps de résumer nos conclusions:

1. Le problème de déterminer l'efficacité d'un remède, celui de déterminer la mortalité à craindre dans un genre donné d'opérations, et beaucoup d'autres, sont du type *inverse*; ils peuvent être résolus à l'aide des formules (3) et (4).

2. Ces formules donnent un résultat qui diffère de celui fourni par les formules (1) et (2); la différence cesse d'être négligeable quand n, nombre des essais, est petit ou quand n étant quelconque, ν, nombre des réussites observées, est proche de 0 ou de n.

3. La Table 1 de M. Ritala peut être employée pour y lire l'erreur moyenne (4), mais il faut ajouter seulement, pour calculer P, un essai réussi et un essai manqué aux observations (voir (3)), et augmenter le nombre total n de 3 (voir (4)).

4. Il est faux de croire que les épreuves à nombre restreint d'essais ($n<30$) ne suffisent pas pour obtenir des conclusions valables. Toute conclusion tirée d'une épreuve implique une probabilité qui mesure la sûreté de la conclusion. Exemple: la formule (5) donne une probabilité de 27% pour que l'efficacité d'un remède qui a agi 2 fois sur 2 dépasse 90%; la même formule donne une probabilité de $99{\cdot}9\%$ pour que le même remède ait une efficacité dépassant 10%. Il y a donc des conclusions que l'on peut tirer avec un grand „coefficient de sûreté" de 2 essais. Nous avons vu que la probabilité pour qu'un remède que l'on a vu réussir 30 fois sur 30 ait une efficacité dépassant $99{\cdot}91\%$ est moindre que 3%. Voilà une conclusion peu sûre, tirée de 30 essais. Il n'y a donc aucune frontière qui sépare les grands nombres n des petits; quand on donne le degré de sûreté que l'on demande et l'efficacité (qualité) à garantir, la formule (5) nous enseigne quel est le nombre n des essais qu'il faut faire. Cette remarque est importante pour l'examen des produits industriels,

où n est souvent borné par des considérations d'ordre économique ($n \leqslant 40$, par exemple).

5. L'objection à soulever contre l'emploi de la règle de Bayes est la suivante. Comment sait-on que la roulette du producteur est graduée uniformément? Or, il n'y a pas d'hypothèse plus simple et, en la rejetant, nous serions contraints de recourir à une autre ou bien d'introduire des concepts et des méthodes qui, bien que correctes, sont trop subtiles pour l'usage que l'auteur des *Tables* avait en vue. Le calcul des probabilités nous enseigne comment on mesure certains ensembles d'objets; dans la géométrie pratique on mesure les aires en se servant des hypothèses euclidiennes car il n'y a aucun argument en faveur des hypothèses plus compliquées.

L'hypothèse de Bayes dans ce genre de problèmes est une convention qui ne peut être démentie par l'expérience, les tables statistiques ne présupposant rien sur l'univers dont on tire la drogue on un traitement.

Pour être à l'abri de toute objection, il suffirait d'avertir le lecteur que le calcul repose sur l'hypothèse d'une probabilité uniforme à priori. Si l'on voulait, en suivant des exemples célèbres, recourir à la méthode du „maximum likelihood", on obtiendrait justement la „vraisemblance" $100^0/_0$ dans les cas où tous les essais réussissent, et les médecins n'en seraient pas moins choqués par la vraisemblance $100^0/_0$ qu'ils ne l'ont été (avec raison) par la probabilité $100^0/_0$.