G. DARMOIS

Sur le rendement des observations statistiques

Journal de la société statistique de Paris, tome 78 (1937), p. 310-319
<http://www.numdam.org/item?id=JSFS_1937__78__310_0>

IV

SUR LE RENDEMENT DES OBSERVATIONS STATISTIQUES

Le but de cette conférence est d'examiner, sur quelques exemples, la manière dont les observations statistiques peuvent servir à la connaissance, et comment, dans certains cas, on peut en obtenir le rendement maximum. Il ne parait guère contestable que les observations servent à quelque chose. J'entends par là que l'acquisition de connaissances nouvelles ne résulte pas de l'application, à un corps initial de connaissances, du seul raisonnement déductif.

Depuis des siècles, nous tirons des résultats de l'observation et de l'expérimentation, et très probablement nous serons amenés à leur demander plus encore.

Mon désir est que ces remarques semblent bien, comme elles le sont, évidentes. Il ne paraîtra pas inutile alors de réfléchir sur quelques points importants de ce mécanisme d'acquisition.

Nous allons nous placer dans le cas assez favorable où l'on sait bien ce que l'on cherche. Comme vous le verrez, on peut atteindre souvent cette position par une méditation préalable, qui donne une forme précise aux hypothèses à essayer.

Prenons quelques exemples :

Un monsieur se fait fort, par un procédé qu'il n'a pas à indiquer, de retrouver des métaux; cette foi en des capacités analogues se trouve actuellement assez répandue. A-t-il raison? Quelle question pouvons-nous poser à l'expérience pour nous éclairer sur ces aptitudes?

Prenons un autre exemple, tout à fait analogue, peut être représentatif d'une tendance moins répandue, et que j'emprunte à un ouvrage de R. A. Fisher (1). Une dame se fait fort de reconnaître, en goûtant une tasse de thé où l'on a mis un peu de lait, si le thé a été versé d'abord, ou si c'est le lait qu'on a versé le premier.

Encore un exemple emprunté à R. A. Fisher, et se rattachant à des expériences de Darwin, sur la croissance des plantes. Une opinion soutenue par quelques observateurs était que les plantes résultant de croisements étaient supérieures sous le rapport de la taille, aux plantes obtenues par autofécondation. Il faut, dans un tel cas, organiser une expérience et en interpréter les résultats.

On rencontrera souvent des questions analogues à la suivante :

De deux types d'engrais recommandés, quel est celui qui est supérieur à l'autre, et donne des récoltes meilleures. Est-il vrai, comme le prétendent certains observateurs, que l'un donne 3 quintaux de plus à l'hectare.

Tels sont les types de problèmes les plus simples sur lesquels nous allons réfléchir. Pouvons-nous acquérir, par l'observation, par l'expérimentation, des connaissances substantielles sur ces divers points? Quelles relations nos conclusions auront-elles avec la structure de l'expérience, existe-t-il des moyens de perfectionner la précision des résultats..., etc...?

Problème de la tasse de thé. — Étudions en détail, en suivant R. A. Fisher (2), l'exemple de la tasse de thé. Nous nous proposons de voir si cette dame possède quelque don la distinguant de la population commune. Pour cela, nous allons mettre sous une forme nette, capable de servir de base à une théorie, l'hypothèse que nous voulons examiner. Nous ferons celle-ci :

Cette dame choisit une réalité au hasard. — Nous allons voir qu'on peut tirer e là des conséquences précises, qu'il est possible de soumettre au contrôle de l'expérience.

Supposons par exemple qu'on ait préparé huit tasses, quatre de chaque catégorie. Ces huit tasses seront présentées dans un ordre fixé par tirage au sort.

L'épreuve consiste à désigner parmi ces huit tasses les quatre qui appartiennent à une catégorie déterminée. Notre hypothèse entraîne la connaissance des probabilités des différentes combinaisons possibles, qui peuvent être :

4 réussites, 3 réussites, 2 réussites, 1 réussite, 0 réussite. Le calcul est très élémentaire. Il y a C_8^4 combinaisons possibles de quatre tasses, soit 70. On trouve que ces 70 cas possibles se répartissent en :

1 16 36 16 1

un de ces cinq cas se réalise nécessairement. Que faudra-t-il en conclure. Il faut ici faire une convention qui se présente dans tous les problèmes analogues.

Seuil de signification. — Imaginons qu'il y ait quatre réussites,

Ce triomphe doit-il entraîner notre rejet de l'hypothèse faite? On peut nous dire que cette réussite complète est due au hasard. C'est vrai, mais la probabilité d'une telle réussite est $\frac{1}{70} = \frac{1,4}{100}$.

Il faut convenir une fois pour toutes qu'une hypothèse peut être abandonnée quand s'est produit un événement qu'elle disait être rare (de même qu'une hypothèse doit être abandonnée quand s'est produit un événement qu'elle disait impossible).

Bien entendu, il nous arrivera d'abandonner une hypothèse vraie, en suivant cette règle de conduite, mais ce risque sera d'autant plus petit que nous aurons été plus exigeant sur la rareté de l'événement significatif.

Il est commode d'adopter comme seuil de signification la probabilité de $\frac{5}{100}$. Le risque est encore petit, et nos exigences ne nous mènent pas trop loin. Si désirant diminuer notre risque, nous avions fixé $\frac{1}{100}$ dans le problème traité, une réussite complète ne nous aurait pas encore permis de prendre une décision. Si nous adoptons ce point de signification de $\frac{5}{100}$, avons-nous le droit de conclure si l'expérience comporte au moins trois réussites. La probabilité, en choisissant au hasard, d'obtenir au moins trois réussites est $\frac{17}{70} = \frac{24}{100}$. Quand nous disons que ce résultat, qui pourrait paraître frappant à certains observateurs, n'a pas de signification, nous disons seulement que le hasard seul peut le produire, avec une probabilité beaucoup plus élevée que notre seuil.

Avec cette organisation de l'expérience, il faut, pour que le résultat soit significatif, et nous conduise à abandonner notre hypothèse, qu'il n'y ait que des réussites.

Il est à remarquer, comme le montre immédiatement la symétrie de notre loi de probabilité, que s'il y avait une réussite négative complète, c'est-à-dire si les quatre tasses étaient toutes de l'autre catégorie, nous serions amenés à abandonner notre hypothèse également.

En résumé, le but d'une expérience est d'atteindre ce niveau de signification. On pourrait l'obtenir en faisant plusieurs fois de suite l'épreuve des 8 tasses.

Supposons que cette épreuve soit recommencée 10 fois, et que le sujet ait obtenu au moins deux réussites complètes. On sait que la probabilité des nombres 0, 1, 2... de réussites complètes, est donnée par le développement du binôme.

$$\left(\frac{1}{70} + \frac{69}{70}\right)^{1}$$

La somme des 9 premiers termes (qu'on calcule par les deux derniers) est entre $\frac{1}{100}$ et $\frac{2}{100}$. Par conséquent le niveau de signification serait atteint.

Réalisation des conditions que donnerait le hasard. — Il est très important dans ces sortes d'expériences, de prendre les précautions qui assurent une application raisonnable du calcul des probabilités. Nous renverrons pour ce point à l'ouvrage de R. A. Fisher, nous contentant de signaler que le mot anglais de « randomisation » qui traduit cette préoccupation, ne me parait pas avoir d'équivalent dans notre langue.

Le premier exemple ou la recherche des métaux. — Il se trouve qu'une expérience a été faite à Lyon sur cette question (on pourra relire sur ce point l'intéressante chronique de M. Houllevigne dans le *Temps* du 24 décembre 1935.)

Dans un appartement qui comprenait 10 pièces, une masse d'argent de 850 grammes, était placée dans une pièce. On fait faire des recherches à 86 opérateurs. Chacun d'eux donne le résultat qu'il obtient. On recommence l'opération 10 fois de suite et l'on compte pour chaque opérateur, le nombre des succès qu'il a obtenus. Les résultats furent les suivants :

31 n'avaient aucune solution exacte;
33 avaient une solution exacte;
14 avaient deux solutions exactes;
7 en avaient trois;
1 en a obtenu quatre.

Personne n'en avait obtenu plus de quatre.

Quelle hypothèse allons-nous essayer d'éprouver. Ce sera toujours la même.

Le choix se fait au hasard. — Cette hypothèse permet de déterminer complètement la probabilité qu'a un chercheur de réussir 0, 1, 2... 10 fois. Ces probabilités sont encore données par le développement du binôme (1).

$$\left(\frac{1}{10} + \frac{9}{10}\right)^{10}$$

On trouve, en commençant par l'autre bout, et arrondissant les valeurs :

0,35	0,39	0.19	0,06	0,01
0	1	2	3	4 et plus.

(1) La masse d'argent changeait de pièce à chaque fois, ce qui modifie un peu les résultats mais très peu.

Comme il y a 86 chercheurs, la valeur probable du nombre de ceux qui auront trouvé 0, 1, 2... réponses exactes s'obtient en multipliant par 86, ce qui donne, en arrondissant un peu :

$$N_0 = 30 \quad N_1 = 33{,}3 \quad N_2 = 16{,}7 \quad N_3 = 4{,}9 \quad | \quad N = 1{,}1 \text{ pour 4 et plus.}$$

On ne peut qu'être frappé d'abord de la façon dont les résultats de cette expérience dessinent la loi de probabilité.

On peut aller plus loin, et se demander si cet ensemble de résultats manifeste ou non un écart significatif avec les indications de la théorie, faite d'après l'hypothèse.

Pour cela, on peut utiliser une méthode indiquée depuis longtemps par Karl Pearson, et qui consiste à calculer la quantité χ^2, où les N' reprisent les résultats expérimentaux, les N les résultats de la théorie :

$$\chi^2 = \frac{(N'_0 - N_0)^2}{N_0} + \frac{(N'_1 - N_1)^2}{N_1} + \ldots..$$

La loi de probabilité de χ^2 est connue de façon très approchée, et donnée par des tables. On trouve ici $\chi^2 = 1{,}35$.

On peut ainsi calculer la probabilité pour qu'un tel écart soit atteint ou dépassé. On trouve une probabilité de l'ordre de $\frac{3}{4}$. Des écarts de cet ordre, ou plus grands, sont donc tout à fait courants dans notre hypothèse.

Ainsi, les résultats de cette expérience n'ont absolument rien de significatif.

L'hypothèse peut être conservée. On peut considérer le choix comme fait au hasard.

Remarque. — Il faut signaler ici qu'un observateur a obtenu quatre réussites. Or, la probabilité pour un chercheur déterminé, d'obtenir ce nombre de réussites étant de l'ordre du $\frac{1}{100}$, on pourrait être tenté de commettre la grave erreur d'attribuer à ce chercheur des qualités exceptionnelles. Comme il s'agit d'une erreur en quelque sorte classique, il convient de la traiter un peu longuement. En somme, nous avons repéré dans un ensemble nombreux l'exemple le plus saisissant, et nous essayons de dire qu'il indique quelque chose.

Or, sur 86 chercheurs, il y en a toujours un qui réussit mieux que les autres, de même que dans une suite de mesures faites au hasard, il y en a toujours une qui est la plus grande. Mais la plus grande valeur d'une suite aléatoire n'a pas du tout la même loi de probabilité qu'une valeur quelconque, et c'est en regard de sa loi de probabilité qu'il faut juger le plus grand nombre.

C'est ainsi qu'on a pu s'étonner que le chiffre 7 figure si peu souvent dans le développement décimal du nombre π, après avoir choisi précisément le chiffre 7 parce qu'il y figure le moins souvent. Si l'on cherche la loi de probabilité du nombre qui figure le moins souvent, on constate que l'écart entre le nombre réel et sa valeur probable n'est plus du tout significatif.

Résumé des principes. — Nous avons donc le moyen d'acquérir des connaissances sur une hypothèse dès quelle peut être précisée numériquement par une loi de probabilité.

Nous pouvons la rejeter si une observation rare se produit, si un écart exceptionnel est révélé par l'expérience.

Pour la question des engrais, nous pouvons juger l'hypothèse suivante :

Les deux engrais ont même effet, les fluctuations observées étant dûes au hasard, ou si l'on veut, la différence de récolte n'est pas significative.

Si l'on nous a affirmé que la différence est de 3 quintaux à l'hectare, nous pouvons essayer de juger si la différence avec 3 quintaux est significative.

Nous parvenons ainsi à une recherche un peu plus compliquée, où il s'agit d'estimer certaines grandeurs, et où va jouer un grand rôle la question du rendement des observations.

Rendement des observations. Un exemple. — Si nous voulons estimer la densité de certains organismes, compter des globules ou des germes, nous pouvons opérer sur une solution fortement diluée, et estimer la nouvelle densité. Comme « Student » l'a montré (3), ce problème peut être ramené à l'estimation du paramètre d'une loi de Poisson (loi des petits nombres ou des petites probabilités).

Si m est le nombre moyen d'organismes par échantillon prélevé sur la masse diluée, la probabilité pour qu'un échantillon contienne 0, 1, 2... organismes est donné par la suite :

$$e^{-m} \quad me^{-m} \quad \frac{m^2}{2} e^{-m} \quad \frac{m^3}{3!} e^{-m} \ldots\ldots$$

Chaque échantillon prélevé fournit une valeur de la variable aléatoire entière qui suit cette loi de Poisson, et notre but est de déduire des observations une estimation de m.

Comptons n de ces échantillons, nous aurons une suite :

$$n_0 \quad n_1 \quad n_2 \quad n_3 \ldots\ldots$$

de somme n, n_i indiquant combien d'échantillons contenaient un nombre i d'organismes.

On peut par exemple, choisir n_0 et se demander si l'on peut en déduire une estimation de m.

Bien entendu, la valeur probable de $\frac{n_0}{n}$ est e^{-m}. Posons $e^{-m} = M$.

On sait que l'écart type de la variable $\frac{n_0}{n}$ est donnée par la formule classique :

$$\sigma^2_{\frac{n_0}{n}} = \frac{M(1-M)}{n}$$

Si l'on considérait, comme il est naturel, le logarithme de $\frac{n_0}{n}$ comme une estimation de $-m$, on a pour les fluctuations (petites) de $\frac{n_0}{n}$

$$\delta\left(\log \frac{n_0}{n}\right) = \frac{1}{M}\, \delta\left(\frac{n_0}{n}\right)$$

Si donc on pose :

$$\log \frac{n_0}{n} = -m',$$

on aura :

$$\sigma^2_{m'} = \frac{1}{M^2} \frac{M(1-M)}{n} = \frac{e^m - 1}{n}$$

Plus cette valeur est petite, plus l'estimation est précise. R. A. Fisher a proposé (dans 1 et 2) de considérer l'inverse $\frac{1}{\sigma^2}$ et de l'appeler l'information, relative à m, extraite des observations par le calcul de m'

$$I^{m'} = \frac{n}{e^m - 1}$$

On pourrait aussi estimer m à l'aide de l'ensemble des n_i. La méthode la plus naturelle est celle de la moyenne arithmétique :

$$m'' = \frac{n_1 + 2n_2 + 3n_3 + \ldots\ldots}{n}$$

On sait que d'une manière générale, on a pour la moyenne arithmétique :

$$\sigma^2_{m''} = \frac{\sigma^2_x}{n}$$

σ_c étant l'écart type de la variable.

On obtient ici :

$$\sigma^2_{m''} = \frac{m}{n} \qquad\qquad I^{m''} = \frac{n}{m}$$

On peut d'ailleurs démontrer qu'aucune estimation ne peut être plus précise que m'', et que la connaissance de m'' est équivalente, pour l'estimation de m, à l'ensemble des observations (une telle estimation est dite exhaustive).

$\frac{n}{m}$ sera donc dite l'information totale fournie par les n observations.

On voit que m' n'extrait que la fraction

$$\frac{m}{e^m - 1}$$

toujours inférieure à l'unité, de cette information totale.

Avant de poursuivre l'étude de cet exemple particulier, rassemblons les notions générales qu'il représente.

Notion d'information. — On peut considérer qu'à un ensemble de n observations indépendantes, est attachée une grandeur $n\,\iota$, information totale.

Toute méthode d'estimation dont l'écart type est σ fournit une information :

$$\frac{1}{\sigma^2} = nj,$$

inférieure ou égale à $n\ i$.

Signification concrète de l'information. — Si deux méthodes fournissent des informations, dont les valeurs respectives soient nj et knj, la seconde méthode fournira les mêmes renseignements avec un nombre d'observations divisé par k.

Cette information apparaît donc, ainsi que le remarque R. A. Fisher, comme l'élément principal d'un devis d'expérience à faire.

Comme il ne semble guère douteux que nous entrerons dans la voie d'expériences assez vastes, pour tout ce qui regarde ces problèmes de la connaissance acquise par l'observation statistique, il y a une question d'emploi des ressources de toute nature qu'il faut se poser au début. Il est évidemment désirable qu'on écarte les méthodes qui ne donneraient qu'une information trop petite, et qu'on tende à employer les méthodes, s'il en existe, qui donnent le maximum d'information.

Retour à l'exemple précédent. — Nous avons vu que le dénombrement des échantillons vides, ou des ensemencements stériles, donne une fraction

$$\frac{m}{e^{m} - 1}$$

de l'information totale.

On ne pourrait pas, comme il est naturel d'y penser, diminuer m par dilution ou réduction, car ce qu'on cherche, c'est la densité primitive

$$D = \alpha m = A\mu\ .$$

α, A sont les facteurs de dilution. On voit que c'est $\frac{\delta m}{m} = \frac{\delta \mu}{\mu} = \frac{\delta D}{D}$ qui compte.

Il faut donc en réalité calculer l'information relative à Log m, qui a la valeur

$$\frac{n m^2}{e^{m} - 1}$$

le rapport à l'information totale restant le même.

L'optimum a lieu pour le maximum de :

$$\frac{m^2}{e^{m} - 1}$$

Le minimum est obtenu pour $m = 1{,}6$ environ.

La partie d'information extraite est alors environ 0,4. Par conséquent, la méthode évidemment plus commode, qui consiste à ne dénombrer que les échantillons stériles, est, même dans les meilleures conditions, d'un très mauvais rendement relativement à la méthode de dénombrement complet.

Un autre exemple relatif à la loi de Gauss. — Nous allons donner ici quelques résultats, d'ailleurs connus, mais qui illustrent les notions précédentes.

Si l'on observe une population gaussienne, où l'on veut estimer la valeur centrale et la dispersion par l'écart type σ, on peut démontrer que les méthodes classiques qui emploient la moyenne arithmétique et l'écart moyen quadratique, sont les meilleures; elles sont d'ailleurs exhaustives. Ainsi, l'information relative à la valeur centrale est $\frac{n}{\sigma^2}$, ce qui est l'information fournie par la moyenne arithmétique.

Mais prenons, comme on le fait parfois, la valeur médiane. Alors l'information obtenue est seulement de :

$$\frac{2}{\pi}\,\frac{n}{\sigma^2}$$

314 observations traitées par la médiane ne donnent pas mieux que 200 traitées par la moyenne.

Un autre problème moins classique est relatif à la dispersion. Au lieu de l'estimer par la formule habituelle :

$$\sigma'^2 = \frac{\Sigma\,(x_i - x_o)^2}{n - 1}$$

où x_0 est la moyenne des n observations, on propose souvent de calculer la longueur de l'intervalle restant après avoir écarté, aux deux bouts de la distribution, un certain pourcentage des observations, soit p.

Cette longueur est de la forme $k\,\sigma''$, k étant une fonction de p donnée par les tables de la loi de Gauss, et σ'' fournit alors une estimation de σ.

Il est d'abord certain que ces méthodes ont un rendement inférieur à la méthode classique puisque la méthode classique est exhaustive. Il est facile d'étudier la précision, ou ce qui revient au même, l'information attachée à ce procédé.

On peut démontrer (4) que le meilleur des intervalles à choisir est celui qui laisse à chaque bout la fraction $\frac{6{,}917}{100}$. On ne perd pas grand'chose en choisissant l'intervalle qui laisse à chaque bout $\frac{10}{100}$. Il faut alors, d'après les tables, diviser la longueur de cet intervalle par 2,56. On trouve alors :

$$\sigma^2_{\sigma''} = \frac{\sigma^2}{N}\cdot\frac{(2{,}28)^2}{(2{,}56)^2} = \frac{\sigma^2}{2N}\cdot\frac{10}{7}$$

$\frac{10}{7}$ est la valeur approchée du facteur de $\frac{\sigma^2}{2N}$. On perd donc environ $\frac{30}{100}$ de l'information.

J'ai voulu donner sur cet exemple si familier l'ordre de grandeur des pertes que l'on risque en n'employant pas la méthode la plus efficace.

Il y a donc un très grand intérêt à démontrer l'existence d'une méthode plus efficace que toutes les autres. C'est le résultat qu'a obtenu R. A. Fisher, au moins pour les observations nombreuses, et ce résultat a été complété à divers points de vue par les recherches de Dugué (5).

Le théorème sur un cas assez général. — Nous supposons que nos hypothèses nous permettent de préciser la loi de probabilité du caractère à mesurer; il ne subsiste plus dans cette loi que des paramètres. Nous allons donner le résultat dans le cas d'un seul paramètre. Soit :

$$f\,(x,\, m)\; dx.$$

la densité de probabilité; les observations ont donné : x_1, x_2..... x_n.

Il existe alors une limite de précision pour les estimations à comportement gaussien, ou ce qui revient au même, une limite à l'information extraite. Si l'on pose :

$$\int \frac{1}{f} \left(\frac{\partial f}{\partial m}\right)^2 dx = i.$$

on démontre que l'information totale est $n\ i$.

De plus, et c'est un résultat très important :

Il existe toujours une méthode qui extrait une information dont la partie principale est $n\ i$.

Cette estimation est celle qui rend maximum le produit :

$$f\,(x_2 m)\; f\,(x_2 m)\; \ldots\ldots\; f\,(x_n m)$$

D'autre part, on peut, partant d'une méthode dont le rendement n'est pas le meilleur, trouver par des méthodes approchées l'estimation de meilleur rendement.

Ce complément a une valeur pratique assez grande (6).

La méthode préconisée par R. A. Fisher a reçu de lui le nom de « method of maximum likelihood », ou plus brièvement méthode de l'optimum.

G. Darmois.

BIBLIOGRAPHIE

1. R. A. Fisher : *Statistical Methods for Research Workers*, 1re édition en 1925; 6e édition en 1936. Oliver and Boyd, Édinbourg et Londres.
2: R. A. Fisher : *The design of experiments*. Oliver and Boyd, 1935.
3. « Student » : *On the error of counsing with an haemacytometer. Biometrika*, 1907.
4. Trumas L. Kelley : *Statistical Method*. New-York, The Macmillan Company, p. 75-76.
5. Daniel Dugué : *Comptes rendus de l'Académie des Sciences de Paris*, 1936.
6. R. S. Koshal : *Journal of the Royal Statistical Society*, XLVI, 1933, p. 303. (Amélioration de la méthode des moments par approximation vers l'optimum.)
7. J. B. Hutchinson : *The application of the method of Maximum Likelehood to the estimation of linkage*. Genetics, XIV, p. 514, 1929.