REVUE DE STATISTIQUE APPLIQUÉE

M. GIRAULT

Un problème fondamental de la statistique mathématique : l'échantillonnage

Revue de statistique appliquée, tome 12, nº 1 (1964), p. 17-24
<http://www.numdam.org/item?id=RSA_1964__12_1_17_0>

© Société française de statistique, 1964, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

UN PROBLÈME FONDAMENTAL DE LA STATISTIQUE MATHÉMATIQUE : L'ÉCHANTILLONNAGE

M. GIRAULT
Professeur à l'Institut de Statistique, Paris

Dans la plupart des expériences faites actuellement, tant en recherche théorique qu'en recherche appliquée, les modèles déterministes ne conviennent pas : si l'on renouvelle plusieurs fois une expérience dans des conditions jugées analogues, les résultats ne sont pas identiques.

Exemple : Fabrication de tôles en alliage léger :

Une usine fabrique des tôles : des lingots d'alliage léger sont transformés en tôles par une série de laminages : les mêmes suites d'opérations effectuées à partir des mêmes matières premières ne produisent pas des produits identiques. Des mesures d'épaisseurs au centre de la bande obtenue fournissent par exemple les résultats suivants (en microns). 2 483, 2 509, 2 473, 2 457, 2 513...

On reconnaîtra là l'exemple-type de toute fabrication industrielle. Ici on peut parler d'expériences au sens classique du terme : on en connaît les conditions (tout au moins les principales). On sait renouveler une expérience et l'on peut le faire à tout moment.

Il existe par ailleurs tout un ensemble de phénomènes qui intéressent l'ingénieur ou l'administrateur et qui se présentent d'une manière différente : le phénomène étudié se produit parfois ; on sait reconnaître s'il se produit mais on n'est pas libre de le provoquer :

Exemple : Pluies recueillies au mois d'août à l'observatoire du Parc Saint-Maur - Hauteurs en mm :

en 1880	: 60	1925 : 93
1885	: 66	1930 : 81
1890	: 43	1935 : 76
1895	: 42	1940 : 6
1900	: 60	1945 : 89

a cet exemple se rattachent les phénomènes météorologiques, climatologiques, etc..., de nombreux phénomènes économiques, commerciaux et ceux que l'on étudie dans les "Sciences Humaines".

Tous ces phénomènes, qu'ils soient provoqués ou simplement observés, entrent dans un schéma général qu'on désigne sous le nom d'épreuve.

Une épreuve est caractérisée par un ensemble de conditions qui la décrivent :

- fabrication par tel procédé, dans tel atelier, à partir de matières spécifiées.

- Mois d'août au Parc Saint-Maur.

Ces "conditions" sont appelées des hypothèses.

Chaque épreuve produit un résultat auquel on s'intéresse :

- épaisseur de la tôle obtenue.

- hauteur de la pluie recueillie.

Dans les phénomènes étudiés en physique classique, les mêmes conditions d'une épreuve provoquent les mêmes résultats ; on dit que les mêmes "causes" produisent les mêmes "effets" ou que les hypothèses de l'épreuve déterminent le résultat.

Il n'en va pas de même pour la catégorie de phénomènes indiqués plus haut : les mêmes hypothèses n'entraînent pas les mêmes résultats.

Naturellement, ce qu'on désigne sous le nom d'épreuve scientifique est une abstraction (aussi bien dans les modèles déterministes que dans les autres). Jamais les conditions d'une épreuve ne sont complètement observées et, par suite, ne sont complètement décrites. Si l'on voulait être parfaitement rigoureux et lucide, il faudrait bien reconnaître qu'il est impossible de reproduire deux fois la même épreuve réelle. Cette attitude serait aussi stérile qu'elle voudrait être rigoureuse et l'efficacité de la méthode scientifique est si flagrante qu'elle n'a pas besoin d'être défendue : méthode qui admet délibérément un certain flou dans les clichés qu'elle prend et qu'elle retient. Toutefois, le passage d'un ensemble d'épreuves réelles à une épreuve abstraite soulève des difficultés : une épreuve doit être assez clairement définie pour que plusieurs observateurs soient également capables de reconnaître si elle se produit. Certaines conditions, considérées comme essentielles seront complètement spécifiées (facteurs contrôlés) pour le reste, il est tout de même nécessaire de donner quelques indications.

Une épreuve abstraite étant définie ; on désigne par "population" l'ensemble des résultats qu'on peut obtenir en effectuant une telle épreuve : ceux qu'on a obtenus ; mais aussi ceux qu'on obtiendra. Il s'agit donc d'un ensemble abstrait qui n'est jamais complètement connu ; mais dont on peut connaître des échantillons.

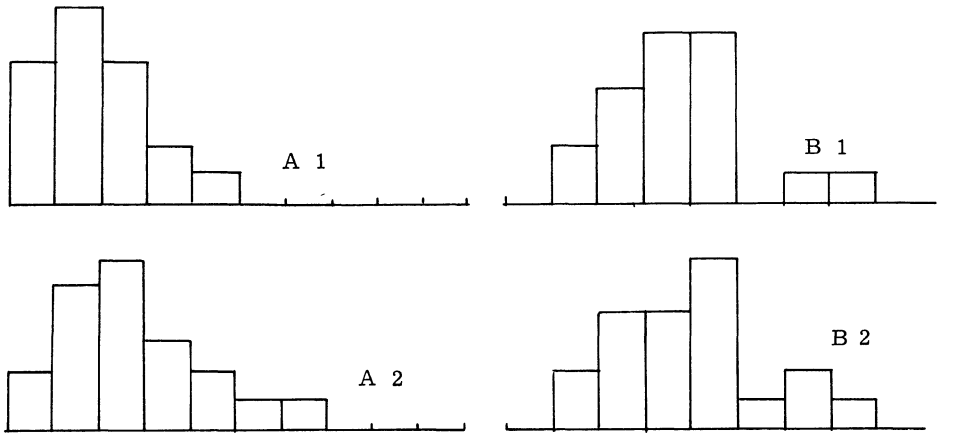
On appelle échantillon de taille n l'ensemble de n résultats d'une même épreuve.

Ainsi nous avons donné ci-dessus un échantillon de taille 10 de la population des hauteurs de pluies observables aux mois d'août au Parc Saint-Maur.

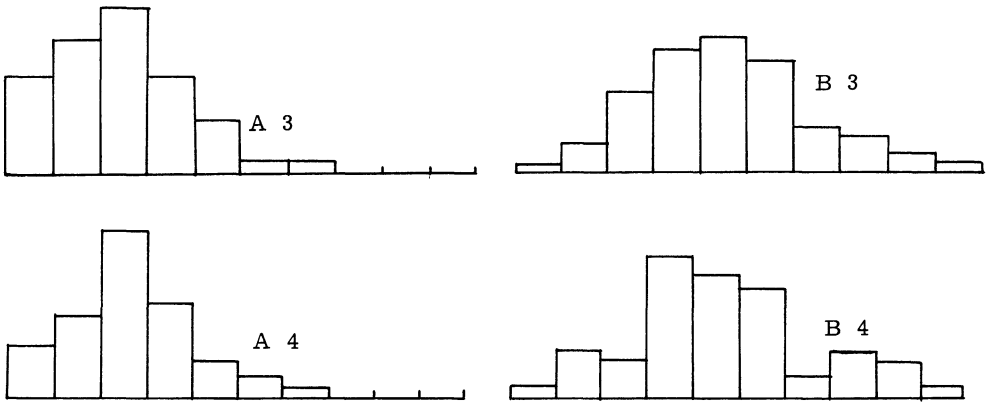
Plusieurs échantillons issus d'une même population ne sont pas identiques ; toutefois on peut observer qu'ils ne sont pas arbitraires :

Exemples : arrivées de bateaux dans un port : on note chaque jour le nombre de bateaux d'un certain type venant au port. Nous allons comparer deux types d'échantillons, notés A et B. Il s'agit d'un même port mais de deux types de bateaux. On admet que tous les échantillons (A) sont parents et qu'il en est de même des échantillons (B).

Echantillons d'effectifs, N = 20



Echantillons d'effectifs N = 100



Les histogrammes précédents mettent en évidence une certaine ressemblance entre les échantillons parents : (échantillons A entre eux et échantillons B entre eux). Corrélativement ils accusent les différences qui distinguent échantillons A d'échantillons B.

Ces analogies et ces différences sont d'autant plus nettes que les effectifs des échantillons sont grands.

Le rôle de la statistique descriptive est de révéler d'une manière aussi nette que possible ces caractères de parenté : soit en présentant des échantillons d'une manière globale (histogramme, fonctions de répartition...) soit en donnant des indices.

Ainsi, dans l'exemple ci-dessus les moyennes des 8 échantillons sont :

effectif N = 20 $\begin{cases} a_1 \text{ (moyenne de } A_1) = 1,35 & b_1 = 3,25 \\ a_2 = 2,25 & b_2 = 3,5 \end{cases}$

effectif N = 100 $\begin{cases} a_3 = 1,91 & b_3 = 4,0 \\ a_4 = 1,87 & b_4 = 4,11 \end{cases}$

Toutefois, quels que soient les effectifs des échantillons observés, les considérations précédentes, de la statistique descriptive, ne permettent pas d'exprimer clairement ni ces analogies ni ces différences ; car les caractères qualitatifs retenus sont trop vagues. Dans cette voie, il n'est pas possible de donner des règles précises qui permettraient de distinguer d'une manière objective des échantillons non parents.

La théorie de l'échantillonnage se propose précisément d'étudier et de décrire des échantillons, donc de caractériser des échantillons parents. Ici comme ailleurs, pour élaborer une théorie il faut construire un modèle abstrait sur lequel le raisonnement mathématique peut s'exercer rigoureusement : c'est le modèle aléatoire ou calcul des probabilités.

Dans cette théorie, chaque évènement[1] pouvant résulter de l'épreuve est affecté d'une mesure : le degré de croyance qu'on lui attribue et qu'on appelle la probabilité de cet évènement. Les probabilités des évènements possibles sont liées les unes aux autres par certaines relations qui permettent un calcul et qui caractérisent la structure de cette théorie.

Pendant longtemps (XVIIIe et XIXe siècles) on a cherché à utiliser le Calcul des Probabilités pour effectuer une véritable induction scientifique au sens classique du terme ; c'est-à-dire de "remonter" des "effets" aux "causes". Naturellement, la connaissance des "causes" ne peut s'exprimer ici qu'en langage probabiliste. La théorie est simple et claire ; elle est basée sur l'application du théorème dit de Bayes, théorème très banal en lui-même ; toutefois l'application de cette théorie se heurte le plus souvent (pas toujours) à une difficulté grave qui en limite l'intérêt : la méconnaissance a priori des lois de probabilité.

Le point de vue actuel sur cette question est très différent et le problème dit de jugement sur échantillon se trouve posé d'une tout autre manière : Si l'on cherche à connaître le comportement d'une famille d'échantillons parents, c'est en définitive pour agir, pour prendre des décisions. C'est à ce problème plus complet que s'attache actuellement la statistique mathématique. On ne recherche plus une connaissance pour elle-même ; ni des "causes" (le terme a bien vieilli !) mais une règle d'action. Nous allons illustrer le procédé sur un exemple emprunté au contrôle de fabrication.

Un problème de décision statistique.

Un atelier fabrique en série des pièces. Celles-ci doivent satisfaire certaines conditions techniques pour être déclarées bonnes. L'étude de

---

(1) On appelle évènement toute partie de l'ensemble des possibles

cette fabrication a conduit à admettre qu'en régime "normal" chaque pièce à la probabilité p = 0,02 d'être mauvaise (et donc la probabilité p = 0,98 d'être bonne) et que ces probabilités sont mutuellement indépendantes pour toutes les pièces fabriquées.

Pour s'assurer que les conditions de fabrication restent satisfaisantes (qu'il n'y a pas de déréglages d'appareils par exemple) on prélève 100 pièces. Celles-ci sont contrôlées et soit K le nombre de pièces mauvaises trouvées dans l'échantillon. A la connaissance de K, on veut fixer une règle d'action : le choix entre les deux termes de l'alternative suivante :

1/ Admettre que la production est normale et donc continuer à produire dans les mêmes conditions, ou bien

2/ admettre que la production est perturbée et donc arrêter la fabrication pour vérifier les appareils.

Les difficultés du choix.

Si certains appareils sont mal réglés, la qualité de la production peut encore être représentée par le schéma précédent mais avec une valeur du paramètre p supérieur à 0,02. Or quelle que soit la valeur réelle de p, le nombre K de pièces défectueuses peut prendre toutes les valeurs entières de 0 à 100 ; de sorte qu'il est impossible de déduire de K la valeur de p avec certitude. Quelle que soit la décision prise, on risque de se tromper.

Soit en choisissant (2) quand les conditions sont normales. (on commet alors une erreur dite de première espèce).

Soit en décidant (1) alors que certains appareils sont déréglés (on commet une erreur dite de seconde espèce).

La statistique mathématique apprend à réduire considérablement ces risques.

Si la fabrication est "normale" ; compte tenu des hypothèses admises, le calcul des probabilités nous apprend que le nombre K peut varier de 0 à 100 par valeurs entières en obéissant très sensiblement à la loi de Poisson de paramètre m = 2. Si tous les entiers de 0 à 100 sont des valeurs possibles de K, celles-ci sont très inégalement probables :

En désignant par P la probabilité d'avoir K = n, on a :

$$P_0 = 0,1353 \quad P_1 = 0,2707 \quad P_2 = 0,2707$$
$$P_3 = 0,1804 \quad P_4 = 0,0902 \quad P_5 = 0,0361$$
$$P_6 = 0,0120 \quad P_7 = 0,0034 \quad P_8 = 0,0009$$

Prob (d'avoir K > 8) = 0,00023 Prob (d'avoir K > 10) = 0,000008

Donc en fait la distribution de probabilité de K est très fortement concentrée sur les petites valeurs (0, 1, 2, 3, ...) et il existe un très petit intervalle ayant une très forte probabilité de contenir K.

Choisissons un seuil de probabilité (0,95 par exemple) et cherchons le plus petit intervalle ayant cette probabilité de contenir K ; c'est l'intervalle [0 ; 4] dit intervalle d'acceptation.

La règle est alors la suivante :

Si l'on obtient K < 4 on décide (1) (pas d'intervention)

Si l'on obtient K > 4 on décide (2) (intervention)

La probabilité de commettre une erreur de 1ere espèce (intervention inutile) est 0,05 (très exactement ici 0,053) c'est la probabilité d'obtenir K > 4 lorsque p = 0,02. On pourrait être tenté de réduire ce risque en choisissant un intervalle d'acceptation plus grand : [0 ; 6] par exemple ; qui a la probabilité 0,995 de contenir K. Ce faisant, on augmenterait l'autre risque ; car cet intervalle plus grand aurait plus de chances de contenir le nombre K, même si ce dernier obéit à une loi de Poisson de paramètre > 2. Supposons par exemple que, par suite de mauvais réglages, P prenne la valeur 0,06. Dans ces conditions K obéirait à la loi de Poisson de paramètre 6. La probabilité de ne pas intervenir (à tort) qui est seulement 0,28 pour l'intervalle [0 ; 4] deviendrait égale à 0,60 pour l'intervalle d'acceptation [0 ; 6]. Il y a donc conflit entre les deux risques.

Test m = 2 contre m > 2

K obéit à la loi de Poisson de paramètre m

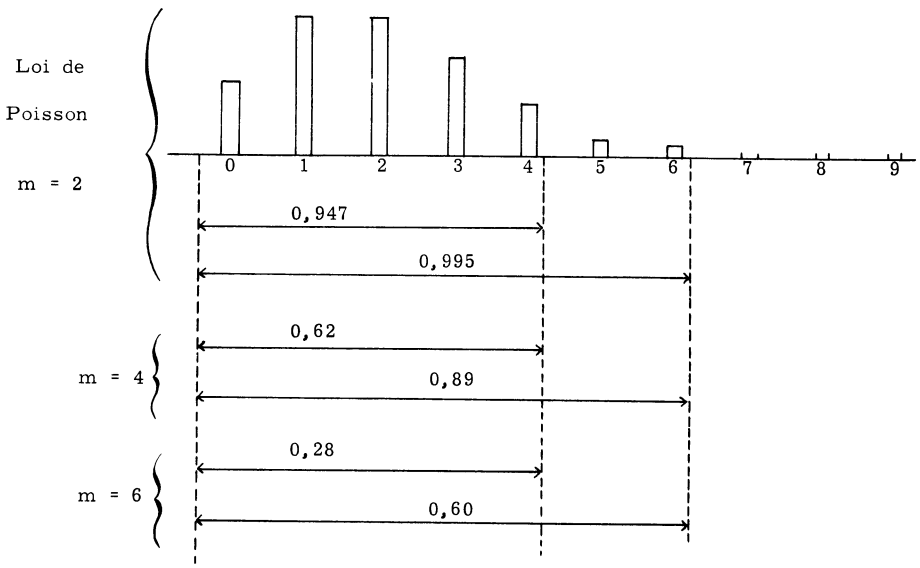


Tableau donnant les probabilités qu'ont les intervalles de contenir K si m = 2 ; m = 4 ou m = 6.

Pour trancher entre ces risques, il faut tenir compte de leurs probabilités et surtout de leurs coûts. Ces techniques assez complexes relèvent de la Science des décisions ; elles seront abordées par Mon-

sieur MORLAT dans son exposé.

Quoi qu'il en soit des problèmes théoriques de décision, il faut remarquer que la marge de choix de l'intervalle d'acceptation est faible, du moins lorsque l'effectif de l'échantillon observé est grand. La probabilité accuse une décroissance très brutale à partir de certaines valeurs (ici k = 5). Si un déréglage se produit, il sera en fait très vite repéré par la méthode précédente.

Terminons ce rapide survol de la statistique en mentionnant quelques problèmes importants qui constituent autant de chapitres de la théorie :

Lois d'échantillonnage : Etudes des lois exactes ou de loi approchées de certaines "statistiques".

Tests dits "non paramétriques". Tests basés sur les rangs des valeurs constituant l'échantillon. Ils ont l'avantage de ne pas faire intervenir les lois de probabilité des quantités étudiées.

Les considérations d'économie conduisent aux plans séquentiels d'échantillonnage (ou échantillonnage progressif) et aux notions de plans d'expérience. Cette dernière théorie, extrêmement séduisante n'a toutefois été développée jusqu'ici que dans le cadre très étroit d'analyse de variance laplacienne. Son intérêt pratique n'en demeure pas moins incontestable.

DISCUSSION

(Président ; M. DELAPORTE)

M. le Président remercie vivement M. GIRAULT de son exposé, dans lequel il a rappelé la plupart des problèmes qui sont à la base même des études de statistique mathématique qu'on est amené à faire dans les applications de cette science, et ceci, en partie, sous la forme du calcul des probabilités.

M. DEYMIE indique qu'en ce qui concerne l'arrivée des navires dans un port cité par le conférencier, cette arrivée est soumise à certaines contraintes qui enlèvent à certaines arrivées leur caractère aléatoire :

- l'heure de la marée pour les ports d'estuaires ;
- les services réguliers ;
- le fait que les navires tendent à éviter d'être à quai le dimanche, jour où les dockers ne travaillent pas.

Peut-on appliquer à ces arrivées ou services une loi de répartition comme la loi de Poisson ?

M. GIRAULT précise qu'il n'a jamais dit que l'arrivée de ces bateaux obéissait à une loi de Poisson, parce qu'il y aurait, en effet, des objections à faire à ce sujet.

Il s'agit, en effet, des bateaux arrivés pendant toute la durée d'un jour donné et sans tenir compte de l'heure ; d'autre part, il ne s'agit pas de la totalité des bateaux arrivés au port, mais d'une catégorie particulière (en fait, des pétroliers).

L'analyse proposée a un caractère global ; il serait, théoriquement possible de faire une analyse fine pour décrire par des lois de probabilité différentes, les arrivées des divers types de bateaux composant la flotte ; mais c'est là une autre question.

M. le Président remercie encore M. GIRAULT qui a esquissé, dans son exposé, tous les problèmes d'une manière extrêmement simple, alors qu'en réalité il existe des difficultés très sérieuses dans ces questions.