MATHÉMATIQUES ET SCIENCES HUMAINES

G. TH. GUILBAUD

Les problèmes de la statistique

Mathématiques et sciences humaines, tome 135 (1996), p. 33-50

<http://www.numdam.org/item?id=MSH_1996__135__33_0>

© Centre d'analyse et de mathématiques sociales de l'EHESS, 1996, tous droits réservés.

L'accès aux archives de la revue « Mathématiques et sciences humaines » (http://msh.revues.org/) 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/

Math. Inf. Sci. hum. (34e année, n°135, *1996*, pp.33-50)

LES PROBLÈMES DE LA STATISTIQUE

G. Th. GUILBAUD[1]

RÉSUMÉ — *Le texte qu'on va lire est une réimpression du Chapitre VI du* Traité de Sociologie *(tome premier), publié en 1959 aux Presses Universitaires de France sous la direction de Georges Gurvitch. Celui-ci était en effet assez convaincu de l'importance de la Statistique dans sa discipline pour vouloir qu'une présentation en soit faite dans son* Traité.

On verra que le parti pris par G. Th. Guilbaud fut non de décrire des techniques de la Statistique en vogue chez les sociologues de cette époque (comme par exemple, le test du Khi-deux ou l'Information de Shannon, comme indicateur du degré d'indépendance entre deux variables), mais de s'efforcer de faire comprendre l'esprit des méthodes statistiques, et leur généralité : ce qu'indique bien le titre du Chapitre : "Les problèmes de la Statistique".

Il en est résulté un texte tout à fait remarquable et qui, comme on dit "n'a pas pris une ride" quelque quarante ans après avoir été écrit et dont tout utilisateur ou praticien de méthodes statistiques (peu importe que ce soit en Sociologie ou dans une autre discipline) peut encore faire son profit.

C'est pourquoi le traité de G. Gurvitch étant depuis longtemps introuvable, il a paru opportun à la rédaction de Mathématiques, Informatique et Sciences humaines *de le mettre de nouveau à la disposition des lecteurs.*

SUMMARY — The Problems of Statistics.

This paper is a reprint of the chapter VI of the Traité de Sociologie, *edited in 1959 by G. Gurvitch and published by the Presses Universitaires de France.*

In this paper, G. Th. Guilbaud tries to make clear the spirit of statistical methods and their generality.

This quite remarkable text is not at all out of date 40 years after ; every one who is practising statistical methods will be able to profit by it.

La synthèse, tentée à chaque époque, de l'ensemble des résultats d'une discipline scientifique n'est jamais que provisoire. On entend souvent des usagers de la statistique se plaindre de ce que les instruments d'analyse qu'ils sont amenés à utiliser ont été forgés pour d'autres : pourquoi ce qui vient de la biométrie ou de l'économie serait-il bon pour le sociologue ou le psychologue ? Mais pour la statistique, comme pour la mathématique, comme pour bien d'autres disciplines, il ne s'agit évidemment pas de faire table rase : un vêtement coupé sur mesures, peut-être, mais le tissu n'est-il pas le même pour tous ?

[1] Centre d'Analyse et de Mathématique Sociales (C.A.M.S.) - Paris.

Il serait déraisonnable, que l'on soit sociologue ou biologiste, de vouloir constituer de toutes pièces une statistique particulière qui n'emprunterait rien aux statistiques étrangères. Tout emprunt est périlleux : on est tenté, dit-on, d'appliquer des recettes ; ce qui est vrai. Il est excellent de dénoncer le péril. Quel est le remède ? C'est d'abord de reconnaître l'existence d'une science statistique, largement indépendante de son objet — ou, pour dire mieux, une statistique générale.

Pour donner une idée de ce qu'elle est, la formulation mathématique est souvent indispensable — elle sera ici évitée au maximum, non sans risques — et l'étude historique de l'évolution des idées est d'une grande utilité : on se limitera, dans ce qui va suivre, à un canevas assez grossier.

Pour un excellent statisticien français du siècle dernier, la situation de la science statistique était claire : "La statistique est la science des faits sociaux" (Moreau de Jonès, 1847, *Éléments de statistique*). Que s'est-il donc passé depuis cette date ? La statistique a-t-elle renié ses origines ? L'objet véritable de la statistique n'a jamais été les sociétés humaines en tant que telles, mais seulement certains aspects, certains modes d'appréhension de la réalité sociale. D'abord le nombre — c'est-à-dire d'une part la quantité (ou plutôt la quotité) et d'autre part le nombreux. Ensuite (et, précisément, il faut bien le dire : surtout) la détection de quelques régularités — la recherche des "lois". Mortalité, natalité — dès les origines — montrent un "ordre" global au-delà des apparences contingentes. N'oublions pas que dès que, cette régularité a été constatée, elle inspire une confiance telle qu'on ne tarde pas à en faire la base d'une industrie (les assurances) "tout comme on utilise la force du vent et le courant des rivières" (M. Block, 1878).

De ces origines de la statistique le nom reste (mais qui entend encore État dans Statistique ?) et bien plus encore reste la situation faite à la statistique dans l'enseignement et la recherche : il est fort rare qu'on enseigne la statistique toute seule, la statistique pure. Ce sera statistique et économie, statistique et biologie, statistique et psychologie et ainsi de suite, et les mots d'économétrie, biométrie, psychométrie, souvent, n'ont pas d'autre signification. Dans quelques cas même l'analyse en deux composants est presque impossible : ainsi la démographie.

Et pourtant, il convient de reconnaître que la statistique s'est efforcée à l'autonomie : une méthodologie statistique, née à la conscience claire de ses buts il y a 50 ans — à prétentions universelles quoique sans statut épistémologique reconnu. Elle n'a point réclamé l'épithète de *Novum Organum* : elle y avait cependant quelques titres.

Le choix du mot de "population", pour désigner tout ensemble statistique, signale les études démographiques des premiers statisticiens ; mais, aujourd'hui, on étudiera sous le nom de population, l'ensemble des automobiles circulant dans un pays donné à une date donnée, aussi bien qu'une colonie microbienne, un lot de tubes électroniques, les mots d'un vocabulaire, ou les hommes d'un groupe. Notons-le, dans le jargon statisticien, le mot "population" désigne l'ensemble lui-même que l'on a en vue et non pas, comme dans le langage usuel, le nombre des individus. Pour désigner ce nombre on dira de préférence : "effectif".

Dans bien des cas, cet effectif peut être considérable. Même si l'on a pu réunir les renseignements individuels, la mise en ordre pose déjà de véritables problèmes de méthode.

Le moteur permanent de toute activité *statistique* est en effet le suivant : le réel est toujours trop complexe et la description exhaustive est impraticable.

Réduisons la difficulté par le choix d'un exemple simplifié ; soit une population humaine bien délimitée ; pour chacun des individus qui la composent je m'intéresse à un caractère unique

bien défini — disons, pour fixer les idées, son âge. Imaginons qu'un recensement complet soit possible : voici le registre où sont portés les noms (ou tels indicateurs d'identité qu'on voudra, matricules ou autres) et les âges en face de chaque nom. Pour commencer il faut dire ceci : ce que le statisticien veut appréhender ce n'est pas à proprement parler la population en tant qu'ensemble des individus, mais l'ensemble des âges. L'anonymat est de règle. On dira : *la distribution statistique*, pour désigner cette population de chiffres qui sera notre unique souci. Que reste-t-il de notre registre quand on y efface les noms ? Une suite de nombres. Si la population est nombreuse il me faut trouver moyen de prendre connaissance de la distribution des âges autrement que par lecture de la liste dans l'ordre du registre (lequel n'est peut-être qu'alphabétique, c'est-à-dire probablement tout à fait étranger à mes préoccupations actuelles). Ranger les individus par rang d'âge, les grouper en classes annuelles ou décennales, sera déjà un progrès, je veux dire un moyen de ne pas se perdre dans la diversité empirique. Simplifier le donné toujours complexe, voilà mon premier mouvement très spontané, mais alors le scrupule vient de rester superficiel si je simplifie trop.

L'idéal ne serait-il pas un discours ordonné — non pas ordonné comme une liste quelconque dont il faut attendre la fin pour savoir vraiment — mais ordonné d'un ordre efficace et didactique : tel que dès les premiers mots je vois les grands traits de l'objet à décrire et qui se poursuive en perfectionnant *progressivement* la description, de sorte que je sois libre de m'arrêter où je veux, obtenant à chaque fois la plus grande connaissance possible pour le nombre limité de renseignements que j'ai accepté d'entendre.

De pareilles méthodes de description progressive sont bien connues : le petit jeu des "portraits", les questionnaires destinés à identifier les plantes, les procédures de diagnostic de l'analyse chimique et bien d'autres schémas taxonomiques procèdent d'une logique commune. On peut dire, en une certaine mesure, que le statisticien décide d'opérer selon des principes analogues en construisant une suite de classements emboîtés les uns dans les autres, une série de filtres, ou de cribles, de plus en plus fins.

Conservons notre exemple simple, trop simple, d'une distribution d'âges. Supposé connu l'effectif total de la population étudiée, je donnerai les effectifs de quelques classes, d'abord très larges, par exemple : moins de 20 ans, plus de 60 ans, entre 20 et 60. Ce sera un premier degré de la classification, premier coup d'œil panoramique, que je peux raffiner en subdivisant par degrés successifs. Mais je suis libre de choisir les coupures entre classes : on peut alors songer à définir les classes, non par des limites d'âge, mais par leur contenu, cette façon de faire aura peut-être le mérite d'être mieux adaptée à l'objet de l'étude, le crible n'étant pas forgé d'avance. Dans cette seconde perspective prendront place des renseignements tels que le suivant : la moitié de la population a moins de 38 ans, l'autre moitié 38 ou davantage (âge médian). Après quoi l'on indiquera quatre fractions égales (les quartiles) et ainsi de suite : déciles, centiles (généralement : "quantiles" ou bien "fractiles", si l'on ne veut pas préciser le fractionnement), technique d'un emploi fort commode et très répandu à l'heure présente.

On peut introduire bien des variantes, tout en conservant le style classificatoire. Indiquons seulement en passant quelques directions possibles. Tout renseignement, dans ce style, se compose de trois données numériques : un intervalle de la variable étudiée (l'âge) défini par ses deux frontières et son contenu (en fraction de l'effectif total). Mais la question peut être posée de diverses façons : Que contient l'intervalle de 18 à 45 ans ? Quel est l'âge au-dessous duquel se trouve tant pour cent de la population ? (L'ensemble de réponses constitue la *fonction de répartition*) ou bien encore : Quel est le plus petit intervalle contenant tant pour cent de la population ? (L'ensemble des réponses constitue la *fonction de concentration*). Il reste à organiser une série de questions de ce style, ou plutôt plusieurs séries, entre lesquelles on choisira la plus opportune selon la nature de la population étudiée et du caractère distribué. Ces questionnaires devront être en principe prolongeables jusqu'à l'examen complet de la distribution, de façon à laisser libre le degré de précision de la connaissance requise.

Malgré ses variantes, ce premier style n'épuise pas les possibilités de description progressive. Une autre échelle est bien connue ; celle qui utilise les *moyennes*. Il n'est peut-être pas besoin d'être statisticien pour comprendre ce qu'on veut dire, et que c'est un renseignement sur la distribution, si l'on dit que l'âge moyen est de 34 ans. Mais il n'est pas besoin d'être grand clerc non plus pour s'apercevoir que deux distributions fort différentes peuvent montrer la même moyenne : le renseignement donné par la moyenne est donc insuffisant, tout comme l'est celui que donne la médiane. Or la moyenne ordinaire peut être considérée comme le premier renseignement d'une série qui, comme la série des fractiles, peut procurer une description progressive (et complète si on le veut) de la distribution. Il s'agit de la série des moments. Nous ne pouvons nous attarder aux détails techniques ; il suffira d'esquisser l'idée générale. On sait qu'à la moyenne d'un caractère numérique (comme l'âge), on a l'habitude d'associer la moyenne des carrés, la moyenne des cubes, des puissances quelconques de cette variable. Le procédé semble moins naturel que celui des classes, mais son emploi se justifie, comme on sait, par certaines commodités. Citons en passant l'une d'elles. La description d'une distribution isolée n'est pas la fin ultime : il faudra savoir comparer diverses populations, et en particulier comparer entre elles plusieurs parties d'une même population. Or, il suffit d'un instant de réflexion pour apercevoir, par exemple, que la médiane, tout comme les autres fractiles, qui se recommandaient par la simplicité de leur définition, sont difficiles à utiliser dans les opérations de partage et de réunion des populations. Supposons qu'on ait deux populations dont on connaît les médianes respectives : on les réunit pour constituer une nouvelle population ; la nouvelle médiane n'est pas fonction des deux médianes partielles. Par contre la loi de composition des moyennes est simple ; la moyenne générale est une moyenne des moyennes partielles. C'est un privilège considérable. Il s'agit alors de construire une suite de caractéristiques numériques, possédant le même privilège d'additivité, et pouvant, à partir de la moyenne comme premier terme, constituer une description progressive. Ce rôle, on le montre aisément, peut être joué par les moyennes de x, de x^2, x^3, etc., de x^k qu'on nommera *moment* d'ordre k. L'approfondissement de cette idée conduit à la *fonction génératrice* de Laplace et aux fonctions dites *caractéristiques* chez les modernes.

La description par les moments est assurément moins intuitive que la description par les fractiles. Il peut être utile de chercher à mieux saisir la signification des moments et à se rendre compte des liaisons qui existent entre les deux styles descriptifs. Quelques exemples suffiront à indiquer la voie. Il s'agit toujours d'une variable (ci-dessus, c'était l'âge) que nous noterons (x) — qui est susceptible de prendre diverses valeurs, disons pour fixer les idées : $x = 0, 1, 2, ...$, etc.

Une distribution statistique associe, à chacune de ces valeurs possibles, un effectif, nombre des individus qui réalisent soit $x = 0$, soit $x = 1$, etc. — ou bien une fréquence, rapport de ces effectifs partiels à l'effectif global. Désignons ces fréquences par f_0, f_1, f_2, etc. Ce sont des nombres (essentiellement positifs) qui vérifient d'abord l'égalité :

$$f_0 + f_1 + f_2 + \text{etc.} = 1.$$

Si l'on donne la moyenne, que nous appellerons $M^{(1)}$ on aura :

$$f_1 + 2f_2 + 3f_3 + \text{etc} = M^{(1)}$$

et si l'on donne aussi le second moment $M^{(2)}$:

$$f_1 + 4f_2 + 9f_3 + 16f_4 + \text{etc.} = M^{(2)}.$$

On peut alors se rendre compte que ces conditions lient les fréquences f : elles ne peuvent plus être quelconques. Bien entendu, il y a une infinité de distributions qui ont les mêmes

moments $M^{(1)}$ et $M^{(2)}$ — mais elles ont cependant quelque chose de commun, elles forment une famille, dont on peut tenter une analyse complète.

Supposons, par exemple, que quatre valeurs de x soient possibles : $x = 0, 1, 2$, ou 3, et partons de la distribution "uniforme" qui affecte 25% de la population à chaque valeur. Alors $M^{(1)} = 1{,}5$ et $M^{(2)} = 3{,}5$. Si l'on examine toutes les distributions qui ont les mêmes deux moments, on verra que les valeurs intermédiaires $x = 1$ et $x = 2$, comportent ensemble la moitié de l'effectif, l'autre moitié étant située aux extrêmes ($x = 0$ et $x = 3$). On verra aussi qu'à chacune de ces valeurs extrêmes doit être attribué au moins 1/6 de l'effectif total et au plus 1/3.

De tels exercices permettent de mieux apercevoir la portée des propositions générales qui établissent un lien entre la connaissance des moments et certaines particularités de la répartition. Parmi les propositions, il faut citer celle qu'établit Bienaymé en 1853 et qui fut généralisée par Tchebycheff. Supposons qu'on connaisse les deux premiers moment, c'est-à-dire la moyenne et l'écart-type ; la règle de Bienaymé permet alors de construire des intervalles encadrant la moyenne et contenant certainement telle proportion de l'effectif qu'on désire. Ainsi, pour en revenir aux âges, si je sais que la moyenne est 34, l'écart-type étant 7 — je peux en plaçant deux écarts-types de part et d'autre, obtenir l'intervalle 20-48, et affirmer que les trois quarts *au moins* de la population sont contenus dans cette classe centrale. Ainsi se constitue une sorte de dictionnaire qui permet de traduire en style de classes, divers renseignements donnés en style de moments.

*
* *

Avant d'aller plus loin, il faut revenir à la notion primitive de distribution statistique. L'exemple sommairement présenté plus haut concernait un caractère numérique simple ; à chaque individu était associé un nombre. Or il faut s'attendre, et pas seulement dans les sciences sociales, à trouver des caractères dont la donnée ne peut se réduire à un nombre : soit qu'il s'agisse de caractère qualitatif non mesurable, soit que tout en étant mesurable il faille plusieurs nombres pour le faire.

On commencera dans tous les cas par dessiner la variété des caractères envisagés. Le cas le plus simple sera celui d'un petit nombre de catégories bien définies : l'état civil par exemple. Dans ce cas la distribution statistique consiste à donner simplement les effectifs de chaque catégorie : célibataires, mariés, veufs, divorcés, et il n'y a guère à dire au delà. Rien qui rappelle ici médianes ni moyennes. Mais il peut se faire que les catégories soient plus nombreuses, et, cela va souvent ensemble, moins bien tranchées. On peut penser aux catégories dites socio-professionnelles de la statistique officielle — ou bien à une enquête d'opinion. Positions, statuts, attitudes — sociales ou non — ne se laissent pas toujours facilement énumérer. Il faut cependant commencer par analyser cette "variété". On peut parfois établir une *gamme*, c'est-à-dire donner un sens à la locution "situé entre" et ranger les catégories en une suite bien ordonnée. Dans ce cas la médiane (et les fractiles) reprennent signification — mais non pas la moyenne ni aucun autre moment (du moins tant que la gamme ne s'est pas transformée en échelle de mesure, ce qui peut être impossible)[2].

S'il est radicalement impossible de constituer une gamme, il existe peut-être quand même une configuration plus ou moins visible. On s'en rendra compte en essayant des groupements partiels : s'il paraît impossible de réunir en une seule classe A, B, C, sans y adjoindre D, cela peut être compris comme une indication de *situation* respective sur ce qu'on pourrait nommer

---

[2] La présence ces cotes numériques ne doit pas donner le change : il n'est pas permis, sans précautions, de traiter des numéros d'ordre comme des nombres ordinaires soumis aux règles d'addition et de multiplication.

une "palette" des catégories. En d'autres termes, c'est l'étude des coupures possibles qui devra mettre en lumière la topologie du système. Un cas intéressant est celui où la configuration spatiale souhaitée est données d'emblée : soit comme schéma figuratif (grille, treillis, arbre généalogique, etc.), soit comme réalité géographique. Dans ce dernier cas la description par classes n'est autre que la description par régions, et l'on voit l'intérêt que peut présenter entre autres la recherche des régions d'étendue minimum contenant tel pourcentage de l'effectif total.

Le cas géographique n'est d'ailleurs guère éloigné de celui où l'on a affaire à un caractère mesurable mais complexe, c'est-à-dire qui doit être représenté par plusieurs nombres. Il ne faut pas abandonner trop facilement cette complexité : supposons qu'une enquête sur les budgets domestiques ait permis de caractériser chaque ménage d'une population de ménages par sa dépense et son revenu. Au lieu de traiter séparément les deux distributions, on devra d'abord exploiter leur association. La représentation graphique est ici d'un grand secours : un point du plan, défini par ses deux coordonnées, représentera chaque unité de notre population. Les deux styles qu'on a dit plus haut sont ici permis : soit dessiner des coupures de classes (emboîtées ou disjointes) en évaluant leur contenu ; soit calculer des moyennes et les divers moments (deux moyennes, deux écarts-types, le premier coefficient de corrélation et la suite). Toutefois, il est bien clair que les instruments déjà définis — fonctions de répartition, fonction de concentration, fonction caractéristiques — seront plus difficiles à manier.

Quelles que soient les difficultés spéciales qu'on rencontre, on voit se constituer les modalités d'emploi de la notion extrêmement générale de distribution statistique et de ses descriptions progressives. C'est cette logique — ou cette grammaire si l'on préfère — en quoi consiste la statistique dite descriptive. Il arrive assez souvent qu'on présente cette partie, apparemment la plus simple, de la technique statistique comme un premier chapitre. Ce n'est pas mal faire — mais il convient aussitôt de souligner que ce premier chapitre ne peut, en aucune manière, constituer ses développements de façon autonome.

C'est là un point fort important — parfois méconnu, non sans dommages graves — et qu'il convient d'exposer.

*
* *

Nous avons tenté, dans les pages précédentes, de situer, très sommairement, quelques-uns des systèmes usuels de "filtres" permettant de saisir les distributions statistiques selon une démarche progressive. Cette progression a été organisée, on l'a dit, pour autoriser l'arrêt en cours de route : si la description peut, en droit, aller aussi loin qu'on le désire — en fait, on ne va pour ainsi dire jamais jusqu'au bout. Combien de fois n'arrivera-t-il pas qu'on se contente de dire médianes et déciles — ou même simplement une région à 95 % — ou bien encore la moyenne et le second moment (sous forme de variance ou d'écart-type), rarement le troisième (sous forme d'indice d'asymétrie) — c'est-à-dire, d'une façon générale, les premiers renseignements seulement.

On doit certes revendiquer le droit d'employer ces descriptions incomplètes — en vue desquelles justement les techniques ont été mises au point — on doit pouvoir se contenter de dessiner les "grandes lignes" et négliger les "détails". Mais on aperçoit alors toute l'importance d'une statistique descriptive bien faite, et que ce n'est ni simple ni facile : apprendre à distinguer les détails de l'essentiel, ou plus précisément à hiérarchiser les détails d'inégale importance. Va-t-on devoir se borner aux conseils du sens commun et faire confiance aux divers praticiens, chacun en son domaine ? Il arrivait souvent dans le passé, et il arrive encore aujourd'hui, que le Traité de Statistique se cantonne en cette position prudente, qu'il dise les règles formelles du calcul sans se croire autorisé à parler de leur emploi. On enseigne à correctement calculer des moyennes (ajoutons même des écarts-types et des coefficients de corrélation), en feignant

d'ignorer que la première tentation de celui qui va se trouver confronté aux résultats touffus de l'enquête sera de "réduire" les données empiriques à cette moyenne, à cet écart-type, à ce coefficient et de s'en contenter — et que sa seconde tentation sera de tout rejeter comme trop simpliste.

Que n'a-t-on pas dit sur les insuffisances de la moyenne (pourquoi la médiane a-t-elle été moins malmenée ?). Que n'a-t-on pas trouvé à répliquer sur ses vertus ? C'est une des tâches — et non des moindres — de la statistique d'exposer dans quel cadre d'hypothèses telle ou telle réduction des observations empiriques est légitime ou non.

*

* *

Il est traditionnel de rappeler le cas, assez spécial, de l'astronome ou du physicien qui, ayant fait des mesures — de la déclinaison d'une étoile ou d'une chaleur spécifique ou du magnétisme terrestre — trouve naturel, dit-on, d'effectuer la moyenne de toutes les mesures effectuées et de déclarer que le nombre *unique* ainsi obtenu représente le résultat de l'ensemble des expériences. Mais, en pareille matière, on suppose :
1) l'existence d'une *vraie* valeur de la grandeur à mesurer ;
2) la présence inévitable d'éléments perturbateurs et variables qui déforment cette valeur vraie ;
3) la construction d'un modèle théorique exposant le mécanisme de ces erreurs de mesure ;
4) quelques raisons de préconiser l'*estimation* de la vraie valeur par le ministère de la moyenne.

La théorie (statistique) des erreurs d'observation s'est bien constituée à partir d'intuitions immédiates : valeur centrale, valeur la plus probable, compensation des écarts, etc.. Mais elle a nécessité d'assez longues constructions mathématiques. Citons l'un des fondateurs de la doctrine, Laplace : "Les phénomènes de la nature sont le plus souvent enveloppés de tant de circonstances étrangères, un si grand nombre de causes perturbatrices y mêlent leur influence qu'il est très difficile de les reconnaître. On ne peut y parvenir qu'en multipliant les observations ou les expériences, afin que les effets étrangers venant à se détruire réciproquement, les *résultats moyens* mettent en évidence ces phénomènes et leurs éléments divers" — voilà le niveau du sens commun, mais voici les exigences du calcul : "On détermine par la théorie des probabilités, les *résultats moyens* les plus avantageux ou qui donnent le moins de prise à l'erreur. Mais cela ne suffit pas ; il est, de plus, nécessaire d'apprécier la probabilité que les erreurs de ces résultats soient comprises dans des limites données ; sans cela on n'a qu'une connaissance imparfaite du degré d'exactitude obtenu. Des formules propres à ces objets sont donc un vrai perfectionnement de la méthode des sciences... l'analyse qu'elles exigent est la plus délicate et la plus difficile de la théorie des probabilités..." (Laplace, *Introduction à la théorie analytique des probabilités*, Paris, 1814).

Sans entrer ici dans le détail, il convient de noter que même dans le cas où une simple moyenne de mesures suffit à l'estimation, on ne présentera jamais cette estimation comme vraie valeur. On croit que cette vraie valeur existe — on sait qu'elle est inaccessible. Mais s'il faut donner un chiffre, on présente le moins imprudent, le plus avantageux, "celui qui donne le moins prise à l'erreur" — en l'assortissant, comme le réclame Laplace, d'une marge d'incertitude.

Le résumé de la série des observations comportera non pas un seul, mais deux nombres — dans certains cas simples ce seront la moyenne et la variance — le premier comme la meilleure estimation possible, le second comme indication de la grandeur à craindre pour l'erreur

inévitable. Finalement d'ailleurs — et grâce, principalement, à Laplace lui-même[3] - se constitue une méthode d'analyse des résultats d'expériences, méthode essentiellement *statistique*.

Il convient, cependant — quelles que soient la beauté et la force de cette construction — de ne pas perdre de vue le problème posé : les mesures de notre astronome visaient une réalité bien définie, car il existait une valeur vraie (c'est, du moins, le postulat). On peut bâtir une comparaison : on trouve des traces de projectiles sur un mur, on suppose que quelqu'un s'est exercé en tirant à la cible, où était la cible ? Mais ce serait une tout autre affaire s'il s'agissait d'une cible mobile. Et encore une autre, s'il n'y avait pas eu de cible du tout.

Or, s'il est naturel de prendre la moyenne de plusieurs mesures faites dans des conditions homogènes et donnant des résultats très voisins les uns des autres, pour représenter l'ensemble, il est non moins évident que l'âge moyen, ou le salaire moyen dans une population, ne se justifie pas du tout de la même manière, ne pouvant évidemment pas avoir une signification du même ordre : âge "vrai", salaire "vrai" — cela ne signifierait rien.

Il faut encore revenir à Laplace : on détermine, dit-il, le résultat moyen le plus avantageux, celui qui donne le moins de prise à l'erreur. Il ne s'agit pas de dire la valeur *vraie*, il s'agit d'un pari. Ce n'est pas sans raison qu'on voit se côtoyer la logique du jugement probable et la logique du jeu ? Mais pour éclairer la conduite de notre astronome, il nous faut, comme pour guider un joueur dans ses martingales, un "modèle" bâti selon les principes de la théorie des probabilités. Dans le cas des erreurs d'observation, Laplace nous propose le célèbre, trop célèbre, modèle qui répartit les probabilités d'écart selon une exponentielle du carré de l'écart. Forme assez subtile, qu'avait entrevue Moivre, et à laquelle l'opinion publique, injuste à son ordinaire, devait attacher le nom de Gauss[4].

Sans nous arrêter à la forme particulière de cette loi, retenons seulement qu'il y a une loi, fort précise. Loi des erreurs a-t-on dit quelquefois, mais en fait loi de probabilité des erreurs, cependant susceptible de vérifications expérimentales (tentées pour la première fois en 1838 par l'astronome Bessel). Pour savoir si un dé est honnête, c'est-à-dire pour vérifier que l'as a bien une chance sur six, ni plus ni moins, on jettera le dé plusieurs fois de suite et l'on fera la statistique des points obtenus, de même ici, on doit prévoir une certaine conformité (mais non pas rigoureuse) entre la loi de probabilité des erreurs et le relevé statistique des écarts, c'est-à-dire finalement la forme de la population des mesures. Non pas seulement un histogramme ayant vaguement la forme d'une "cloche", mais une conformité analytique qu'il faudra préciser.

On voit ainsi apparaître une convergence entre la statistique et la théorie des erreurs (convergence entrevue par Laplace, développée par Cournot et Poisson, mais guère ferme avant la fin du XIXème siècle).

Chaque mesure est une unité statistique, un individu — leur ensemble constitue une population. La distribution des mesures n'est pas quelconque : elle n'est pas non plus soumise à une "loi" comme celle de Kepler imposant une forme aux orbites planétaires. La meilleure façon de se représenter les choses est le tirage au sort : la population des observations doit être considérée comme un échantillon d'une population plus vaste, la population des observations *possibles*. Voir le réel comme prélèvement aléatoire au sein du possible, c'est la clef de tout le système : on va confronter les statistiques du réel (issues des relevés, enquêtes, recensements, etc.), avec la statistique du possible (qui s'appelle Calcul des Probabilités).

---

[3] Dès les premiers mémoires de 1774, se trouve posé le problème : déterminer le *milieu* que l'on doit prendre entre plusieurs observations données d'un même phénomène.

[4] Laissons, pour le moment impunie, la maladresse de celui qui imagina de supprimer le nom propre et de dire non plus Loi de Gauss, ni même de Laplace, mais "Loi normale". Canonisation qui devait faire bien des ravages, et qui n'a pas cessé d'en faire, à l'heure où j'écris ces lignes.

Du même coup la notion même d'erreur, et celle, corrélative, de valeur vraie, ne sont plus très nécessaires. Chaque mesure faite est un "tirage" ou du moins est analogue à ce qu'on appelle ainsi quoique, à l'inverse de ce qui a lieu dans les loteries, les divers numéros ne sont pas également probables car la population des boules dans le sac a une certaine structure ; il y a par exemple concentration autour d'une valeur "centrale" ou "typique" (mais finalement pas plus "vraie" qu'une autre).

Voici donc la première idée de modèle probabiliste : les caractères d'une population soumise à l'observation sont rattachés à ceux d'une population idéale ou population-mère, dont la population empirique est "issue". Les premières formes de ce schéma qui viennent à l'esprit rappellent très directement la loterie, ou — selon le vocabulaire traditionnel — l'*urne* et ses boules. Avant le temps même où l'on essayait de se représenter comme un jeu de hasard le mécanisme des erreurs d'observations (en astronomie d'abord), ce type d'explication avait déjà pénétré l'étude scientifique des phénomènes humains. Mais la métaphore de l'urne n'est-elle pas trop simpliste ? Laplace lui-même attire notre attention là-dessus : "J'ai vu des hommes désirant ardemment d'avoir un fils n'apprendre qu'avec peine les naissances des garçons dans le mois où ils allaient devenir pères. S'imaginant que le rapport de ces naissances à celles des filles devait être le même à la fin de chaque mois, ils jugeaient que les garçons déjà nés rendaient plus probables les naissances prochaines des filles. Ainsi l'extraction d'une boule blanche dans une urne qui renferme un nombre limité de boules blanches et noires dans un rapport donné, accroît la probabilité d'extraire une boule noire au tirage suivant. Mais cela cesse d'avoir lieu quand le nombre des boules de l'urne est illimité, comme on doit le supposer, pour assimiler ce cas à celui des naissances". Mais qu'est-ce donc que cette "urne" infinie ? A dire vrai l'imagerie de l'urne et des boules, adoptée dans un louable souci didactique, est vraiment de celles dont on peut se demander si elles n'ont pas fait plus de mal que de bien. Le "modèle" a-t-il donc besoin d'être un objet que l'on puisse toucher et manipuler ? Ce qu'on désire c'est décrire un *processus aléatoire* adéquat — et non pas faire réaliser un projet de machine par un artisan. L'idée même de processus aléatoire a mis longtemps à se dégager, mais elle est aujourd'hui suffisamment claire pour qu'on abandonne, quand il le faut, le musée des accessoires.

Examinons de plus près l'avantage qu'on obtient quand on veut bien se représenter les résultats d'une enquête statistique comme résultats d'un processus aléatoire. Il suffira d'évoquer le schéma le plus simple, qu'on désigne ordinairement sous le nom de Bernoulli, une suite d'épreuves répétées, indépendantes les unes des autres, et à probabilités fixes. A chaque épreuve un certain événement peut se réaliser ou non, la probabilité de réalisation étant p. On effectue N épreuves et on note par n le nombre de réussites. La fréquence (proportion des réussites au nombre des épreuves, soit $n : N = f$) ne peut être prédite de façon déterminée. Il y a mieux : toutes les valeurs de f sont *possibles*. Mais elles sont inégalement probables, et l'on peut montrer par le calcul que celles de ces valeurs qui diffèrent beaucoup de p sont très peu probables. (D'autant moins que N est plus grand — préciser cet énoncé c'est établir avec Jacques Bernoulli la première loi de "grands nombres" connue).

Supposons maintenant qu'on effectue plusieurs séries de N épreuves. A chacune de ces séries est associée une fréquence f : d'où une population, au sens statistique du terme, et une distribution statistique pour les f. Que peut-on dire de cette distribution : il convient ici de ne point abuser d'une "loi des grands nombres" vaguement imaginée. Il faut d'abord affirmer — comme tout à l'heure, pour un f unique — que *toute* distribution des f est *a priori* possible. Mais on ajoutera que les diverses formes sont inégalement probables.

On a donc obtenu de cette manière un modèle qui peut fournir des distributions statistiques ; toutes les distributions ainsi fabriquées ont bien quelque chose de commun — une "loi" — mais cette loi ne s'exprime pas comme conformité à un archétype. L'aléatoire une fois introduit, au départ de la construction, demeurera toujours présent. Et cet aléa même nous

donne la souplesse dont nous avons besoin, avec la manière de s'en servir. Car il nous faut savoir conduire proprement une vérification expérimentale.

Nous avons déjà eu l'occasion de signaler que les premiers essais de la science statistique se sont développés à l'occasion de l'étude des phénomènes démographiques. Buffon relève la proportion des sexes à la naissance dans les paroisses alentour de Montbard, vers 1770. Cette proportion n'est pas vraiment fixe, elle varie d'une paroisse à l'autre. Peut-on trouver un caractère commun malgré la diversité ? A-t-on le droit d'additionner tous les résultats ensemble pour calculer un taux "moyen" de masculinité. C'est évidemment à un schéma de Bernoulli que pensèrent les statisticiens du XVIIIème siècle, non sans quelques tâtonnements. Reprenons les notations précédentes : N sera le nombre des naissances dans un certain lieu et pour une certaine période, n le nombre de garçons, f la proportion n : N. Mais comment choisir p, paramètre fondamental du modèle, probabilité d'une naissance masculine ? C'est le problème de l'estimation des paramètres, dont la position remonte à Bayes (probabilité des causes, 1765). Divisons la difficulté et opérons en deux étapes.

Supposons d'abord que nous ayons une idée *a priori* sur cette probabilité. Pourquoi pas $p = 1/2$, par exemple ? C'est une théorie : il faut la vérifier. On va donc faire le relevé statistique des f. Mais quel qu'il soit il est, par construction même, *possible* ; mais plus ou moins probable. Et l'on sent bien que si ce qu'on a constaté est trop improbable au regard de la théorie, la théorie aura tort. Comme le rapporte Diderot de l'abbé Galiani : il faut dire que les dés sont pipés si le bateleur amène cinq fois de suite le brelan de six. La mise en forme régulière de ce procédé de *décision* conduit à la construction de ce qu'on nomme parfois "tests d'hypothèse" et qui sont de véritables "critères statistiques". Il s'agit de qualifier de vraisemblable ou de (hautement) invraisemblable une hypothèse présentée, qui sera en conséquence acceptée ou rejetée. Si l'hypothèse est rejetée, on devra en changer, mais progressivement. Et l'on peut encore construire des critères qui peuvent y aider. Par exemple on pourra examiner l'hypothèse : la probabilité d'une naissance masculine (dont on postule l'existence) peut-elle être la même dans le pays A et dans le pays B ? Cette hypothèse n'est pas incompatible avec l'observation de deux fréquences différentes f_A et f_B — mais elle devient très peu vraisemblable si la différence est grande. De tels examens, assortis évidemment d'une procédure chiffrée, pourraient s'appeler critère (ou test) d'homogénéité. C'est ainsi qu'en comparant les registres de baptêmes à Londres et à Paris, Laplace conclut (par un calcul un peu différent de nos tests modernes mais analogue en son esprit) : "Il y a trois cent mille à parier contre un qu'à Londres la possibilité des baptêmes de garçons est plus grande qu'à Paris. Cette probabilité approche tellement de la certitude qu'il y a lieu de rechercher la cause de cette supériorité". Cette conclusion doit être soulignée : le jugement de la statistique n'est qu'une incitation à la recherche. Notons, en passant, que Laplace s'oriente vers une explication sociologique, et non biologique.

Enfin, venons-en à l'estimation proprement dite. On a construit un mécanisme aléatoire, réglé selon un paramètre (ici p) et capable de produire une distribution (ici de la variable f). De même que p ne "détermine" pas la distribution observable, de même il est impossible d'obtenir à proprement parler une détermination de la valeur du paramètre p — mais seulement d'aboutir, relativement au choix de p, à des jugements de probabilité. Dans certains problèmes (il n'est pas indispensable de préciser ici) on peut donner un sens précis à l'expression : *maximum de vraisemblance*. Mais il convient de signaler aussi la notion importante de *résumé exhaustif*. L'observation qu'on a faite doit être l'une des possibilités de réalisation offertes par le modèle et cette possibilité est, par le choix de la valeur du paramètre p, dotée d'une probabilité. On peut fort bien concevoir que plusieurs réalisations aient la même probabilité, quel que soit le choix de p. Il en résultera que la connaissance complète de l'observation n'est pas nécessaire, et que l'on sait tout ce qu'on doit savoir quand on a simplement donné quelques caractéristiques de la distribution observée. Ainsi, pour nous limiter à un cas très simple si l'on

a observé plusieurs épreuves réalisant ou non un certain événement, le nombre des succès suffit à guider le choix de la probabilité inconnue.

On voit alors apparaître ici un lien très profond entre la théorie de l'estimation et la statistique descriptive. La moyenne par exemple, deviendra une procédure justifiée si elle résume véritablement et de façon complète ce qu'on doit savoir de l'observation *pour* pouvoir choisir les paramètres du modèle. On peut ainsi donner un sens précis à la technique spontanément adoptée par les praticiens : mais il faut, on le voit, un cadre d'hypothèses, c'est-à-dire un modèle.

En présence d'une distribution statistique on se demande en effet — pour employer un langage usuel et imprécis — si telle ou telle particularité de la distribution "ne serait pas simplement due au hasard". Pour donner un sens non équivoque à une pareille question il faut dire comment on se représente ce "hasard". Une fois cette déclaration faite, on voit bien l'intérêt d'une analyse qui écarte le contingent, le non-significatif (sous forme d'aléas) pour ne retenir que le permanent, l'essentiel. C'est donc finalement une possibilité de réponse à la question : comment décrire sans être submergé par les détails. Et comme on le rappelait en commençant, c'est la question-clé de toute statistique.

Mais il reste à choisir ce "hasard" qui sert à discerner l'important de ce qui ne l'est pas. Nous abordons ici un domaine tout à fait fondamental de la statistique, laquelle ne peut se passer de "hasard", mais quel hasard ? Dans les débuts, les jeux ont fourni le matériel nécessaire : d'abord les dés qui donnaient 6, 36, ..., chances égales, puis la loterie c'est-à-dire la fameuse *urne* peuplée de boules colorées en proportions variables. Bien entendu, la composition de cette urne, qu'il suffit d'imaginer pour faire les calculs, doit rester constante : ou, comme on dit, il faut remettre la boule dans l'urne après chaque tirage. La commodité de ce modèle a causé quelques dommages : certains ont pu croire que c'était le seul modèle possible, d'autres que c'était le modèle idéal, le meilleur de tous les modèles. Bien entendu ces idées fausses, vous ne les trouverez pas chez Pascal, Bernoulli, Laplace ou Cournot — mais qui donc, de nos jours, apprend la statistique en lisant les classiques ?

Arrêtons-nous un instant sur quelques préjugés. L'idée d'un hasard "pur" donc d'un hasard — plus pur qu'un autre — l'idée d'un phénomène "parfaitement aléatoire" — ce furent de idées-forces mais non pas des idées claires. Il est intéressant de trouver ça et là des témoignages de la difficulté. Voici S.F. Lacroix (1816) : "Lorsque les diverses chances d'un jeu sont rigoureusement d'une égale possibilité, tant par la construction des instruments aléatoires que par la manière de s'en servir, les événements passés ne sauraient avoir aucune influences sur les événements futurs." Et voici J. Bertrand (1889) : " Supposons dix millions d'électeurs. Attribuons six millions de votes à un parti et quatre seulement à la minorité. On forme mille collège de dix mille électeur chacun : tout candidat qui réunira plus de cinq mille suffrages sera élu. L'opinion approuvée par les quatre dixièmes des votants serait représentée proportionnellement par quatre cents députés. Les lois du hasard ne lui accordent rien : sur mille représentants pas un seul pour elle. Le calcul réduit à zéro pour ainsi dire, la vraisemblance de tout autre hypothèse. Supposons... qu'un joueur s'engage à payer autant de millions qu'il se trouvera de députés de la minorité... On ne pourrait pas (c'est la réponse rigoureuse, sinon acceptable) lui offrir équitablement plus d'un centime. Ce centime pourrait lui coûter cher. Les minorités, même beaucoup moindres, obtiennent quelque représentant. Les électeurs n'étant pas associés par le sort, les influences locales triomphent des lois du hasard..." On ne parlerait plus ainsi aujourd'hui, donnant un sens si restrictif aux mots "hasard" et "lois du hasard". Mais la conclusion de Bertrand doit être remarquée "c'est avec grande défiance qu'il faut, sur les traces de Condorcet, éclairer les sciences morales et politiques par le flambeau de l'Algèbre".

C'est souvent ainsi que les choses se passent : on commence par élever des barrières ; "voilà le seul vrai hasard" : après quoi il est facile de rejeter tout effort de conquête : "Ce hasard-

là (ainsi défini, c'est-à-dire délimité) ne sert à rien en tel domaine" — par exemple dans les sciences de l'homme !

Mais continuons notre lecture ; c'est toujours J. Bertrand : "Beaucoup de joueurs, préoccupés de la *régularité* nécessaire des *moyennes*, cherchent, dans les coups qui précèdent celui qu'ils vont jouer une indication et un conseil... l'illusion repose sur un sophisme ; on allègue la loi de Bernoulli comme certaine alors qu'elle n'est que probable. Sur vingt mille épreuves à la roulette, la noire ne peut sortir plus de dix mille cinq cents fois — si les dix mille premières parties ont donné six mille noires, les dix mille suivantes ont contracté une dette envers la rouge. *On fait trop d'honneur à la roulette : elle n'a ni conscience ni mémoire*".

Le calcul des probabilités (et la statistique) devaient-ils donc être ainsi contraints : ne jamais étudier que ce qui n'a ni conscience ni mémoire. Mais la phrase, belle et appelée au succès, était déjà démentie avant d'avoir été écrite. Conscience ? Quand Montmort et Bernoulli se demandent comment conseiller le joueur de Hère (ancêtre de notre Baccara) — quand Bertrand lui-même se demande s'il faut tirer ou non à cinq au Baccara — de quoi s'agit-il donc ? Des premières recherches sur les jeux où la réflexion et l'habileté des joueurs se mêlent au hasard "pur". Mémoire ? Les partis de Pascal et Fermat ne sont-ils pas déjà dans le temps irréversible, où ce qui est dépend de ce qui fut ? Et n'est-ce pas l'humble début de l'analyse séquentielle moderne ?

La seconde moitié du XIXème siècle s'engageait dans une voie trop étroite en réservant le calcul des probabilités aux modes les plus aveugles de production du hasard. Ce fut le mérite de H. Poincaré et de A.A. Markoff de reprendre une vieille tradition, un moment interrompue[5].

"Battre" les cartes", dit Littré, c'est "les mêler *afin que* le hasard seul préside à la distribution". C'est là définir une procédure par ses intentions. Mais les gestes traditionnels réalisent-ils correctement leur fin ? Pourquoi, quand le jeu a été "battu" assez longtemps, admettons-nous que les divers rangements sont également probables ? Que le roi de carreau a autant de chances d'être ici que là ? Rien n'est moins évident, nous fait remarquer Poincaré, car le point de départ, la disposition des cartes avant le battage, résultent de péripéties de la partie précédente — et l'on prétend réellement effacer toute trace de ce qui s'est passé avant de jouer la seconde manche. Il s'agit, pour parler comme Bertrand, d'abolir toute mémoire. Or chacun des gestes élémentaires, en quoi consiste le battage usuel, n'est pas dépourvu de mémoire : il consiste à effectuer quelques permutations, c'est-à-dire à passer d'un ordre à un autre. Mais un seul geste ne peut transformer un ordre donné en n'importe quel autre : c'est bien pourquoi la confiance qu'on a dans l'efficacité du battage est proportionnée à la durée des manipulations ; trop courte, elle ne vaudrait rien. Deux états successifs du paquet de cartes ne sont pas indépendants l'un de l'autre, s'ils sont assez rapprochés. Comment se fait-il que cette dépendance diminue quand la distance augmente ?

Pour Poincaré ce problème devait servir à éclairer bien d'autres phénomènes physiques : le mélange des liquides ou des gaz, par exemple. Prenons encore une autre métaphore ; celle de la promenade au hasard dans un réseau de rues. Supposons qu'à chaque carrefour je tire au sort la direction que je vais prendre. Cet aléa n'abolit pas complètement l'enchaînement : quelqu'un qui connaîtrait seulement mon point de départ ne se trouve pas, pour prévoir ma situation actuelle, dans l'ignorance ou l'indifférence totale. Mais cependant on pressent que la connaissance de la position initiale donne une information de plus en plus faible à mesure que ma randonnée se poursuit. C'est cette intuition vague qu'il s'agit de préciser par le calcul. Ce calcul fournit des résultats qui n'étaient pas faciles à prévoir : ainsi, si le réseau de rues est régulier, chaque carrefour présentant quatre directions, on peut parier que le promeneur passera par tel carrefour

---

[5] J. Bertrand lui-même en est conscient dans son chapitre "Les lois de la statistique" ; il écrit : "Toutes les manières de consulter le hasard ne sont pas équivalentes ; sans vouloir le contester, on s'est montré souvent trop peu sévère dans le choix à faire entre elles".

désigné d'avance, quel que soit son point de départ ; il suffit d'être patient. La sagesse des nations disait déjà : tous les chemins mènent à Rome — sans préciser la durée du pèlerinage. Mais aurait-elle été capable de prévoir que le pari cessait d'être recommandable pour d'autres réseaux ? Si l'on prend un réseau régulier mais situé dans l'espace (à chaque carrefour il y a six directions), il n'est plus du tout certain, ni même presque certain, qu'on doive passer à la longue par tel point désigné d'avance.

"Promenade au hasard" peut être une désignation métaphorique. Si l'on veut parler un langage abstrait, on imaginera, avec Markoff (1907), une série d'états possibles (ce sont les divers rangements du paquet de cartes, les divers carrefours du réseau de rues). Le passage d'un état à un autre est aléatoire — et cet aléa sera défini, non pas par la probabilité *d'être* en tel état, mais par la probabilité de *passer* d'un état à un autre. L'une des applications proposées par Markoff mérite un peu d'attention : il s'agit de l'emploi de la statistique dans la description du langage. On a depuis fort longtemps constaté que les fréquences des lettres dans un texte écrit sont assez stables — sans être bien entendu rigoureusement constantes. Ce constat avait souvent été utilisé pour aider au décryptement des messages transmis par le truchement d'une écriture secrète. Peut-on représenter cette stabilité (et ces fluctuations) par un modèle probabiliste ? Une urne contenant les lettres selon des proportions données est un modèle trop grossier. Markoff propose de lier la probabilité d'apparition d'une lettre à la nature de la lettre précédente. Prenant pour exemple un texte de Pouchkine (*Eugen Onegin*) il montre que si le rapport des fréquences entre voyelles et consonnes est à peu près constant, il est cependant impossible de se représenter les fluctuations autour de cette moyenne comme résultant du "simple hasard" que donne l'urne de Bernoulli — par contre, si l'on imagine deux urnes dont l'une servira à tirer le successeur d'une voyelle et l'autre celui d'une consonne — le modèle devient beaucoup plus satisfaisant.

Avec les recherches de Markoff et de Poincaré, le départ était donné et l'on vit se développer une étude systématique des processus aléatoires, dans lesquels le hasard joue son rôle fondamental sans exiger qu'à chaque instant il soit fait table rase de tout le passé du phénomène. Les événements successifs sont bien aléatoires, bien que mutuellement "enchaînés". Le traitement systématique des probabilités "en chaîne" permettait d'enrichir prodigieusement l'arsenal des modèles, jusqu'alors presque exclusivement tributaire du matériel de casino.

Le développement des idées théoriques et des schémas probabilistes abstraits va évidemment de pair avec celui des applications en divers domaines scientifiques. Sous le nom de Mécanique statistique, plusieurs synthèses se construisirent prolongeant les premières recherches sur les gaz et donnant un fondement nouveau à la thermodynamique classique. Les sciences de la vie commencèrent aussi à poser des problèmes du même genre ; il est bien clair que la présence d'un enchaînement est essentiel pour l'étude quantitative des populations humaines ou animales. Divers modèles de processus aléatoires ont été créés en effet directement à la demande de la démographie (mortalité, natalité, migrations, épidémies, etc.).

Comme c'est dans le domaine de l'étude des phénomènes sociaux que l'on a le plus souvent remarqué l'inadéquation des hypothèses d'indépendance — on peut y voir un champ tout préparé à l'emploi des processus aléatoires moins sommaire que celui de l'urne de Bernoulli. C'était d'ailleurs bien la pensée profonde des statisticiens de la fin du siècle précédent. Les tentatives de W. Lexis (vers 1880) par exemple — aujourd'hui un peu oubliées dans les manuels au profit des travaux postérieurs de K. Pearson — pour étudier les formes de la dispersion dans les phénomènes sociaux sont à inscrire en ce sens. Par la suite, la recherche théorique en statistique s'est trouvée plus souvent mise au service du biologiste et les exigences propres de la sociologie n'ont guère été prises en considération ; sous-développement analogue à regretter en ce qui concerne la statistique économique.

La réduction du hasard à certains schémas privilégiés considérés comme seuls donnant le "vrai" hasard, comme seuls représentant l'aléatoire à l'état "pur", a joué, depuis un siècle, un rôle beaucoup plus important qu'on ne croit. Les manières de consulter le sort sont très diverses, et peuvent donner des résultats extrêmement variés ; on a presque honte d'énoncer pareille banalité, mais il s'est passé ceci : on a pu énoncer des "lois du hasard" lesquelles ne font que mettre en forme rigoureuse (en précisant les hypothèses nécessaires) la croyance courante en une régularisation à la longue, une compensation des écarts ou des erreurs. On ne s'étonnait donc plus trop que le hasard soit défini de façon si étroite.

L'histoire des enquêtes statistiques sur les inégalités économiques et sociales fournirait maintes illustrations de ces vues trop courtes. Les hommes et les groupes sont inégaux : on peut se demander, comme faisait Rousseau quelle est la part de l'inégalité "naturelle" et la part de ce qui résulte de l'action humaine. On peut prétendre, avec Lassalle, que l'inégalité des richesses résulte des aléas de la vie sociale. Un problème est posé : quel est le rôle du hasard ? Lorsque Pareto (1896), pour combattre les thèses lassaliennes, cherche si le "hasard" explique ou non les inégalités de revenus — pour lui, et ses contemporains, cela signifie seulement : peut-on assimiler les distributions que donnent les statistiques, fiscales ou autres, à la loi connue des erreurs d'expérience, à savoir la loi de Gauss. Il est vrai que la biométrie naissante avait constaté que cette loi prestigieuse représentait assez convenablement certaines inégalités physiques : les tailles ou les poids d'un groupement humain homogène dessinent en effet, assez bien, la fameuse courbe. Était-ce une raison suffisante pour établir l'équivalence ? Bien entendu, Pareto dut conclure par la négative : les revenus ne sont pas statistiquement distribués selon la loi dite "normale". Mais il fit davantage : il montra que les distributions de revenus ont une forme très remarquable, assez stable à travers les époques et les sociétés les plus diverses, et il en proposa une formulation analytique. Cette formule devait attirer, plus tard, l'attention par son étrange universalité : la distribution des populations des villes, la concentration industrielle, diverses statistiques géographiques, linguistiques, etc., présentent les mêmes caractères que les statistiques originellement compilées par Pareto. Ne pouvait-on alors penser qu'il s'agissait d'une loi fort générale aussi formelle peut-être que celle de Laplace et Gauss.

Mais d'où vient donc le privilège de cette dernière ? La justification la plus souvent invoquée en faveur de la loi normale est celle qui résulte des propriétés établies par Laplace (1812), généralisées par Liapounoff (1901) et quelques autres : il s'agit de la distribution d'une somme dont chaque composante est aléatoire. Moyennant quelques hypothèses sur ces composantes on établit que la loi de probabilité de la somme est voisine d'être gaussienne quand le nombre des composantes est très grand. L'application d'une telle proposition est évidente : ainsi une erreur de mesure peut être considérée comme résultant de l'*addition* d'un très grand nombre d'erreurs partielles, indépendantes les unes des autres, toutes ces erreurs partielles étant à peu près du même ordre de grandeur, entrant chacune pour une faible part dans l'erreur totale. On peut traduire ces hypothèses en un langage mathématique assez précis pour pouvoir alors appliquer le théorème cité. Chaque fois que l'on pourra, de cette manière, se représenter la grandeur d'un phénomène comme résultant de l'addition d'une multitude de variables, on pourra essayer d'utiliser le même schéma : il faudra examiner si les divers éléments composants sont indépendants (en probabilité) et si aucun d'eux n'est prépondérant.

On comprend bien que ces conditions sont assez restrictives et qu'en particulier dans les phénomènes économiques ou sociaux il soit difficile de les admettre. Mais le privilège apparent de la loi normale a été mieux compris le jour où l'on s'est demandé, avec Paul Lévy (1935), quelles sont les lois statistiques qui peuvent être loi-limite d'une somme de termes en grand nombre. Il y a, on le savait, la loi de Gauss — mais aussi quelques autres, dont certaines ont une parenté remarquable avec la loi de Pareto. C'est le groupe très important des distributions "stables", c'est-à-dire dont la forme n'est pas modifiée par l'addition des grandeurs.

On retrouve par cette voie l'exigence fondamentale de toute explication : à savoir un modèle, ici un processus aléatoire — et l'on éclaire du même coup la nature profonde de certaines permanences statistiques, ou si l'on préfère, on se rend mieux compte de la nature des régularités introduites par les "grands nombres".

*

* *

Il faut nous arrêter quelque peu à la "loi des grands nombres". C'est à Jacques Bernoulli (1703) que l'on doit le premier énoncé correct d'une proposition qui n'est pas sans lien avec l'intuition commune et même l'expérience de chacun. Mais à vouloir réduire le théorème à une maxime de bon sens, on risque quelques graves erreurs. Il faut d'abord bien voir que le calcul de Bernoulli ne s'applique qu'à un modèle très particulier (répétition d'une épreuve qui peut fournir l'un ou l'autre de deux résultats contraires avec des probabilités constantes). Poisson avait bien vu qu'il faudrait s'affranchir de ces hypothèses trop restrictives si l'on veut énoncer une proposition qui mérite un titre si ambitieux : "loi des grands nombres" ; à vrai dire ce titre était plutôt un programme de recherches, une invite à généraliser les premiers schémas simples. Depuis Poisson, la réalisation du programme a été fort avancée. Mais pendant que se poursuivait le travail proprement mathématique, on essayait aussi d'établir une communication entre les divers théorèmes asymptotiques et l'idée intuitive de la "régularisation du hasard" par la répétition de l'expérience ou par l'extension de l'enquête. Ce qui fut bien plus difficile que d'établir les théorèmes. Chacun se forge au niveau de son expérience, une loi de grands nombres personnelle, parfois beaucoup plus riche que les constructions du calcul des probabilités — mais en tous les cas difficilement comparable.

Il ne s'agit pas ici de savoir qui est coupable d'abus. Sont-ce les statisticiens, les mathématiciens ou les "autres" ? Ne tentons pas une enquête difficile. Mais observons les conséquences. Dans certaines discussions, ou querelles, tout semble se passer comme si "loi des grands nombres" signifiait que pour comprendre un phénomène il faille toujours s'adresser au plus grand nombre possible d'observations. Mais qui ne voit que cette recommandation apparemment banale risque d'induire, pour enrichir la collection, à "mettre dans le même sac" des choses qui ne le souffrent guère. Danger réel, semble-t-il, en histoire et en sociologie. Et l'on dira que "la loi des grands nombres ne s'applique pas". Mais quelle loi ? Il aurait peut-être mieux valu ne pas suivre Poisson, laisser "loi" et "grands nombres", et parler comme on fait souvent aujourd'hui de formulation asymptotique, de convergence en probabilité, etc.

Reprenons le schéma classique : une fréquence f dépend du nombre d'épreuves N ; que se passe-t-il lorsque N augmente indéfiniment ? On ne peut dire sans précautions de langage que f a une limite — puisque cette fréquence est aléatoire. Pour mesurer la longueur d'un cercle on y inscrit des polygones réguliers et l'on dit que le périmètre du polygone, qui est une fonction du nombre des côtés, a pour limite la longueur du cercle. C'est dire que l'on peut obtenir une approximation aussi précise qu'on voudra à condition d'être libre de choisir des polygones ayant un aussi grand nombre de côtés qu'il le faudra. Dans le cas présent, il n'est pas incorrect de dire que la fréquence f est une approximation de la probabilité p — mais on ne peut pas dire que l'on peut obtenir une approximation *aussi précise qu'on voudra* ; cela signifie en effet que l'écart entre ce qu'on observe et ce qu'on veut atteindre (entre le périmètre polygonal et la circonférence du cercle) peut devenir inférieur à telle exigence de précision qu'on veut. Ce n'est pas possible ici : l'écart entre fréquence et probabilité, il est toujours *possible* qu'il soit grand. Mais on devra dire : une approximation *aussi sûre qu'on le désire.* C'est-à-dire que les chances seront aussi faibles qu'on le souhaite d'avoir un grand écart entre f et p : l'approximation n'est pas plus ou moins précise, il est seulement plus ou moins probable qu'elle soit assez précise. Dans le cas de la circonférence la précision intervient seule, ici interviennent simultanément précision et sécurité : marges d'erreur et probabilité de ne pas dépasser ces marges.

D'un point de vue mathématique très étroit on se satisfait souvent d'une affirmation d'existence : la limite (définie correctement, ici : en probabilité) existe. D'un point de vue plus large, il est nécessaire (et l'usager y tient beaucoup) de préciser les modalités de cette convergence. Quelle valeur de N choisir pour qu'on puisse parier à tant contre un que l'écart entre f (observable) et p (idéal) soit inférieur à telle valeur donnée. C'est dire que les lois de grands nombres doivent être raffinées pour être vraiment utiles.

Leur utilité est d'ailleurs multiple : jusqu'ici nous avons insisté sur les liens logiques qui existent entre une statistique apparemment descriptive et une statistique formelle et probabiliste : pour séparer l'important de ce qui ne l'est pas, on considère le réel comme échantillon prélevé au sein du possible. Mais quand on parle d'échantillon on évoque bien plutôt la technique de l'enquête : c'est l'observation qui n'est qu'une partie du réel observable. Si je veux étudier les caractéristiques du travail (tel ou tel "travail") dans la population active française, je ne puis la plupart du temps songer à l'examen complet. L'enquête exhaustive est impossible. Mais comment passer de la connaissance de l'échantillon à celle de la population totale ? L'idée première, spontanément appliquée avant toute théorie, consiste à souhaiter que l'échantillon soit une image fidèle, une sorte de miniature de la population, du moins en ce qui concerne les distributions statistiques étudiées — que l'échantillon soit *représentatif*. Pour cela il faut évidemment avoir une certaine connaissance de la population totale — et savoir que penser des liaisons qui peuvent exister entre les caractères déjà connus et ceux qu'on veut étudier.

Si je m'intéresse spécialement à la durée du travail — et si, pour ce faire, je choisis un certain nombre d'individus dans chaque région et chaque profession — il est clair que j'aurais dû réfléchir au préalable :
1) A la liaison entre profession, région et durée du travail ;
2) A la façon de prélever l'échantillon dans chaque compartiment (région-profession).

Réfléchir : peut-être n'est-il pas possible, ni même nécessaire de transformer cette réflexion en un calcul proprement dit. Cependant l'expérience montre qu'il est bien plus difficile qu'on ne croit d'échapper aux illusions sans une discipline vraiment sévère.

La sévérité extrême est, en un sens, représentée par le sondage classique avec tirage au sort (et ses modalités : strates, grappes, etc.) — qui nous ramène directement aux calculs que nous avions évoqués dans les pages précédentes.

L'urne — ou ses perfectionnements — réapparaît et l'on sait estimer la confiance à attribuer aux résultats. L'enquête sociologique ne peut ignorer — et n'ignore pas — les raffinements modernes de la théorie des sondages et des modalités de son application. Mais il faut bien constater qu'elle ne peut pas toujours mettre à profit les remarquables édifices théoriques offerts. On peut alors regretter que l'affaire prenne forme de dilemme : sondage par "quota" *ou bien* sondages aléatoire. Il serait urgent d'établir des formes intermédiaires — ou plus exactement de montrer qu'un minimum d'exigences théoriques est indispensable pour bien conduire une enquête. Pour cela, la construction de modèles représentant au mieux la procédure de choix est nécessaire. Le terrain n'est pas complètement inexploré : il convient de profiter de toute la richesse des méthodes de planification des expériences (design of experiment). Longtemps réservée à l'expérimentation biologique, agricole, médicale — cette théorie se développe rapidement, et devrait bientôt constituer une véritable logique (et pratique) de l'expérimentation. Elle doit pouvoir assimiler, sans les détruire, les diverses "règles de la méthode" constituées en plusieurs sciences d'observation. Les orientations récentes de la théorie des probabilités donnent bon espoir en ce sens. Signalons surtout la tendance qui s'affirme de traiter les problèmes de la Décision statistique comme des problèmes d'action plus que de connaissance : il n'y a pas de règle du raisonnement inductif, disait Neyman, mais des règles de conduite inductive. Ce point de vue, praxéologique, a donné déjà d'importants

résultats dans le contrôle statistique industriel — et a permis nombre de synthèses efficaces sur le plan théorique. Mais ce n'est probablement qu'un petit commencement.

NOTE BIBLIOGRAPHIQUE GÉNÉRALE

HISTORIQUE

Il est fort *utile* de connaître les grands classiques : Buffon, Laplace, Cournot, Poisson, Quetelet, Galton, Poincaré, ..., et, quand on le peut, de les lire dans l'original.

Pour avoir une idée de la situation au début du siècle, on pourra jeter un coup d'œil sur les articles : Calcul des probabilités, Théorie des erreurs, Statistique — dans l'*Encyclopédie des sciences mathématiques* (Paris-Leipzig, 1900-1911).

Quelques renseignements utiles dans Westergaard (*Contributions to the history of statistics*, London, 1932) — pour la période antérieure à Laplace il faut voir aussi Todhunter (*A history of the mathematical theory of probability*, 1865).

TECHNIQUES ET MÉTHODES

1 - Pour qui redouterait la mathématique, les textes cités ci-après de : Borel (1, 2), Fréchet, Darmois (2, 4), Vessereau, Yule et Kendall, Yates, pourront éclairer les idées fondamentales et les principales techniques.

2 - Sans exiger une véritable spécialisation mathématique, mais seulement quelques connaissances et de l'habitude : Fisher (1, 2), Neyman, Darmois (1, 3), Morice et Chartier, Finney.

3 - Plus difficile : Cramér, Blackwell et Girshick, Feller, Savage.

4 - Servira d'ouvrage de référence et de bibliographie : Kendall.

BIBLIOGRAPHIE SÉLECTIONNÉE

BLACKWELL, D., GIRSHICK, M.A., *Theory of Games and Statistical Decisions*, New York, Wiley, 1954, 355 p.
BOREL, E., *Les probabilités et la vie*, coll. "Que sais-je ?", n°91, Paris, Presses Universitaires de France, 1943.
BOREL, E., *Probabilité et certitude*, *ibid.*, n°445, Paris, 1950.
CRAMÉR, M., *Mathematical methods of statistics*, Princeton, 1956, 575 p.
DARMOIS, G., *Statistique mathématique*, Paris, Doin, 1928.
DARMOIS, G., *Statistique et applications*, coll. A. Colin, n°174.
DARMOIS, G., *Les mathématiques et la psychologie*, mémorial des sciences mathématiques, fasc. 98, 51 p.
DARMOIS, G., *L'analyse des corrélations*, les conférences du Palais de la Découverte, 17 janvier 1948, 20 p.
FELLER, W., *An introduction to probability theory and its applications*, New York, London, John Wiley & Sons, Inc., Chapman & Hall, Ltd., 2nd ed., 2 vol., 1966.
FISHER, R.A., *Statistical methods for research workers*, Oliver and Boyd, Edingurgh, 1ère éd., 1925, 354 p.
FISHER, R.A., *The design of experiments*, *ibid.*, 1ère éd., 1935, 242 p.
FORTET, R., *Calcul des probabilités*, Paris, C.N.R.S., 1950, 330 p.

FRECHET, M., *Les mathématiques et le concret*, Paris, Presses Universitaires de France, 1955, 438 p.
FRECHET, M., HALBWACHS, *Le calcul des probabilités à la portée de tous*, Paris, Dunod, 1924, 297 p.
HALPHEN, E., *La notion de vraisemblance, essai sur les fondements du calcul des probabilités et de la statistique mathématiques*, publications de l'Institut de Statistique de l'Université de Paris, vol.4, fasc.1, 1955, pp.41-92.
KENDALL, M., STUART A., *The advanced theory of statistics*, 3 vol., London, Griffin, 1977.
MORICE, E., CHARTIER, F., *Méthode statistique*, 2 vol., Paris, INSEE, 1954.
MORLAT, G., "L'usage du calcul des probabilités", *Revue d'Economie politique*, 1956, n°6, pp.889-907.
NEYMAN, J., *First course in probability and statistics* , New York, Holt, 1950, 350 p.
SAVAGE, L.J., *The foundations of statistics*, New York, Wiley, 1954, 294 p.
VESSEREAU, A., *La statistique*, coll. "Que sais-je ?", n°281, Paris, Presses Universitaires de France, 1947.
WESTERGAARD, H., *Contributions to the history of statistics*, London, King, 1932, 280 p.
YATES, F., *Méthodes de sondage pour recensements et enquêtes*, Paris, Dunod, 1951.