PROBABILITÉS.

SUR LA CONSTANCE DES CAUSES,

CONCLUE DES EFFETS OBSERVÉS,

PAR J. BIENAYMÉ,

INSPECTEUR GÉNÉRAL DES FINANCES.

Extrait des procès-verbaux de la séance du 25 avril 1840
de la Société Philomatique de Paris.

M. Jules Bienaymé expose un principe de probabilités qu'il croit entièrement nouveau, et qui lui paraît susceptible de recevoir des applications continuelles dans les sciences d'observation. Voici en quoi consiste ce principe.

Lorsqu'on a fait un grand nombre d'expériences, ou qu'on a recueilli une masse de renseignements statistiques pour en déduire un certain résultat moyen, on peut les partager en plusieurs groupes, soit d'après l'ordre dans lequel les expériences ont été effectuées, soit d'après toute autre considération particulière à ces expériences. Si l'on détermine ensuite les résultats moyens de chacun de ces groupes naturels, on conçoit qu'ils différeront plus ou moins entre eux, et qu'ils s'écarteront plus ou moins du résultat général. D'ordinaire il se trouvera des écarts d'autant plus grands, que les groupes seront plus multipliés. Il est facile de voir que l'étendue de ces écarts doit dépendre du résultat observé; mais il semble au premier coup-d'œil qu'elle devrait également dépendre de la possibilité que donne aux phénomènes en question la cause ou le système de causes qui les régit. Cependant il n'en est rien, quand ce système de causes reste constant pendant toute la durée des expériences. On démontre sans peine que, dans ce cas, les relations de probabilité qui doivent exister entre le résultat général et les résultats partiels sont absolument indépendantes de la possibilité des phénomènes; il n'entre dans les expressions qui les caractérisent que les résultats seuls des observations faites, même alors que la loi de possibilité des phénomènes est connue à l'avance.

Ainsi, par exemple, si l'on tient note des résultats de 120,000 coups d'un dé ordinaire, on saura d'avance que la possibilité d'amener l'as est $\frac{1}{6}$, de sorte que ce point devra se présenter à peu près $\frac{1}{6}$ de120,000, ou 20,000 fois. Cependant il pourra arriver que l'as, au lieu de paraître à peu près une fois sur 6, ne se montre en 120,000 coups qu'une fois sur 20, c'est-à-dire environ 6,000 coups. Eh! bien, malgré cette différence entre le rapport observé et la possibilité réelle, quand on partagera les 120,000 jets de dé en plusieurs séries, et qu'on examinera le nombre des as dans chacune de ces séries, comme la cause sera restée constante, on trouvera que les séries partielles donnent à peu près des nombres d'as proportionnels au nombre des as contenu dans la série totale; et cela d'autant plus exactement que les séries partielles contiendront chacune plus de coups de dé. Si le rapport total des as s'est trouvé de $\frac{1}{6}$ à peu près, sur les 120,000 épreuves, ce sera encore cette fraction $\frac{1}{6}$ qui régira les séries partielles, et chacune d'elles offrira, à certains écarts près, environ $\frac{1}{6}$ d'as. Si, au contraire, le rapport total s'est trouvé de $\frac{1}{20}$, nonobstant la possibilité constante $\frac{1}{6}$ de la face du dé qui porte l'as, ce sera la fraction $\frac{1}{20}$ qui régira les séries partielles, et il se rencontrera dans chacune d'elles $\frac{1}{20}$ d'as, toujours à certains écarts près, dont la probabilité est assignable.

Cependant, comme il vient d'être dit, chaque série partielle, dans ce dernier cas comme dans le premier, aura été formée sous l'influence de la possibilité $\frac{1}{6}$; et il semblerait *à priori* que cette possibilité dût donner un grand nombre de séries où il se trouverait $\frac{1}{6}$ d'as : de sorte qu'elle n'aurait fourni le rapport extraordinaire $\frac{1}{20}$ sur l'ensemble des 120,000 épreuves que par quelques séries fort rares qui s'écarteraient excessivement du rapport de possibilité $\frac{1}{6}$. Mais cette présupposition serait inexacte; et au contraire, quand l'ensemble des coups de dés fournit le rapport $\frac{1}{20}$, si éloigné du rapport réel de possibilité $\frac{1}{6}$, il devient extrêmement probable, moralement certain (pour employer l'expression de Jacques Bernoulli) que les séries partielles, même très multipliées, s'écarteront peu du rapport $\frac{1}{20}$.

La démonstration du principe qui assure ce résultat est très simple. Supposons qu'on ait exécuté un grand nombre c de tirages dans une urne contenant des boules blanches et des boules noires dans un rapport connu, tel que p soit la possibilité d'amener une boule blanche. Supposons encore qu'il soit sorti a boules blanches et b noires, et qu'on partage le nombre total c des tirages en deux séries, la première de m et la deuxième de n tirages. On sait que la probabilité d'amener dans la première série d'épreuves r boules blanches, dont la possibilité est p, s'exprime par le terme du développement de la puissance m du binôme $p + (1-p)$, dans lequel p a l'exposant r, soit :

$$\frac{m \,.\, m-1 \,.\, m-2 \,\ldots\ldots\, m-r+1}{1 \,.\, 2 \,.\, 3 \,\ldots\ldots\, r}\, p^r (1-p)^{m-r}.$$

Semblablement la probabilité d'amener q boules blanches dans la deuxième série d'épreuves, sera

$$\frac{n \,.\, n-1 \,\ldots\ldots\, n-q+1}{1 \,.\, 2 \,\ldots\ldots\, q}\, p^q (1-p)^{n-q}.$$

Partant le concours des deux événements (r boules blanches dans la première série, et q boules blanches dans la deuxième) aura pour probabilité le produit des deux précédentes, soit :

$$\frac{m \,.\, m-1 \,..\, m-r+1}{1 \,.\, 2 \,..\, r} \times \frac{n \,.\, n-1 \,..\, n-q+1}{1 \,.\, 2 \,..\, q}\, p^{r+q} (1-p)^{m+n-r-q}.$$

Maintenant il convient d'observer que les épreuves sont faites, et qu'il est sorti a boules blanches sur le total $c = m + n$ des épreuves : de sorte que les deux nombres de blanches r et q, dans les deux séries partielles, sont assujétis à la condition $r + q = a$. Chacun de ces nombres ne peut donc varier que depuis o jusqu'à a : ce qui rend impossibles un grand nombre de cas qui pouvaient arriver dans deux séries d'épreuves. Il n'y a dès lors lieu de considérer parmi les valeurs de la probabilité ci-dessus que celles qui sont données par la condition $r + q = a$: et puisque ces valeurs deviennent seules possibles, il faut en faire la somme, et diviser l'expression précédente par cette somme. Or, on reconnaît sans difficulté que la somme dont il s'agit est

$$\frac{m+n\,.\,m+n-1\,.....\,m+n-a+1}{1\quad.\quad2\quad.....\quad a}\,p^a\,(1-p)^{c-a}.$$

$$\text{ou}\quad\frac{c\,.\,c-1\,.....\,c-a+1}{1\,.\,2\,.....\,a}\,p^a\,(1-p)_b.$$

Le quotient de la probabilité ci-dessus par cette somme est

$$\frac{\dfrac{m\,.\,m-1\,..\,m-r+1}{1\,.\,2\,..\,r}\times\dfrac{n\,.\,n-1\,..\,n-q+1}{1\,.\,2\,..\,q}\,p^{r+q}(1-p)^{m+n-r-q}}{\dfrac{c\,.\,c-1\,....\,c-a+1}{1\,.\,2\,.....\,a}\,p^a\,(1-p)^{c-a}}.$$

et l'on voit qu'à cause de $c=m+n$ et $a=r+q$, la possibilité p disparaît complètement de ce quotient.

Ainsi la probabilité de trouver r boules blanches dans la première série, et $q=a-r$ blanches dans la deuxième, quand on partage en deux séries un nombre total c de tirages qui a donné a boules blanches, est simplement

$$\frac{\dfrac{m\,.\,m-1\,....\,m-r+1}{1\,.\,2\,....\,r}\times\dfrac{n\,.\,n-1\,....\,n-q+1}{1\,.\,2\,....\,q}}{\dfrac{c\,.\,c-1\,.....\,c-a+1}{1\,.\,2\,.....\,a}}$$

expression dans laquelle il ne reste plus que les résultats des tirages, c'est-à-dire des faits observés.

Avec un peu d'attention, on reconnaît dans cette expression la suivante :

$$\frac{a\,.\,a-1\,.\,a-2\,....\,a-r+1\times b\,.\,b-1\,.\,b-2\,....\,b-m+r+1}{c\,.\,c-1\,.\,c-2\,.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;.\;c-m+1}$$

$$\times\frac{m\,.\,m-1\,.....\,m-r+1}{1\,.\,2\,.....\,r}$$

qui est la possibilité de tirer r boules blanches et $(m-r)$ noires d'une urne contenant c boules, dont a blanches et b noires, quand on y prend m boules au hazard, sans en remettre aucune.

Les relations de probabilité entre les séries partielles et la série totale des épreuves, ou des expériences, sont donc non seulement indépendantes de la possibilité réelle des événements, mais de plus elles sont les mêmes que si les faits dont se compose une série partielle avaient été tirés au hasard de la série totale des faits observés.

L'application de ce principe (qui s'étend d'ailleurs à tous les cas de probabilités constantes, quel que soit le nombre des espèces d'événements dont le résultat se compose) sera très aisée à faire.

Lorsqu'il importera de connaître si la cause, ou le système de causes, qui a régi une série d'expériences, n'a point subi de variation pendant la durée de ces expériences, il suffira de les diviser en séries partielles, et de calculer si les écarts des résultats moyens de ces subdivisions sont renfermés dans les limites que leur assigne le résultat moyen général, en vertu du nouveau principe.

Il est fort digne de remarque, ajoute M. Bienaymé, qu'on pourra conclure par ce procédé, que la cause a été constante ou variable, sans rien préjuger sur la possibilité réelle qu'elle peut donner aux phénomènes. Cette conclusion subsistera, quand même le résultat moyen s'écarterait complètement de la valeur de cette possibilité, et que par conséquent il donnerait à l'observateur une idée tout-à-fait inexacte de cette valeur. C'est là une conséquence importante, car il en ressort que la statistique, et en général les sciences d'observation, peuvent toujours fournir des données positives sur la constance des lois naturelles, indépendamment de la valeur de ces lois.

Voici la formule à employer quand on divise seulement en deux parties la série des observations, et qu'il ne s'agit que de deux phénomènes exclusifs l'un de l'autre, comme le sont la sortie d'une boule noire et la sortie d'une blanche dans une suite de tirages.

En conservant les lettres déjà employées, on supposera qu'il a été observé a phénomènes d'un certain genre sur un grand nombre c d'expériences, et que le phénomène contraire a par suite eu lieu $c - a = b$ fois. Si l'on prend une série partielle de m de ces observations, on doit trouver, dans l'hypothèse d'une cause constante, que le nombre des phénomènes dont il s'est présenté a sur la masse, est, pour la série partielle, compris entre les limites

$$r = \mathrm{N} \pm u \sqrt{2\, m \frac{a\, b}{c^2} \cdot \frac{c - m}{c}}$$

(N étant le plus grand nombre entier renfermé dans $(m+1)\,\dfrac{a+1}{c+2}$)

avec une probabilité, exprimée par $$\frac{1}{\sqrt{\pi}} \int_{-u}^{u} dt\, e^{-t^2} + \frac{e^{-u^2}}{\sqrt{2\pi \frac{ab(c-m)}{c^3} m}}$$

Le plus souvent la division en deux séries suffira pour manifester la constance ou l'inconstance de la cause. Car on peut remarquer que les limites précédentes sont très étroites.

Fourier, dans les *Recherches statistiques sur Paris*, avait conseillé de séparer les observations en groupes, afin de reconnaître par les écarts des résultats partiels, si l'on pouvait accorder quelque confiance au résultat moyen général. Mais il n'a donné aucune règle à ce sujet. L'incertitude subsistait donc. On avait même appliqué à l'examen des résultats partiels une formule de Laplace, qui se rapporte à un problême très différent du problême actuel : c'est celle qui exprime les écarts probables d'un nombre m de nouvelles épreuves, quand déjà on a fait c expériences qui ont donné a fois le phénomène attendu. Les limites du nombre r des répétitions de ce phénomène dans ces m épreuves nouvelles (et non dans m des c épreuves déjà faites), sont :

$$r = \mathrm{N} \pm u \sqrt{2 \frac{a\, b}{c^2}\, \frac{c+m}{c} m}$$

avec une probabilité $$\frac{1}{\sqrt{\pi}} \int_{-u}^{u} dt\, e^{-t^2} + \frac{e^{-u^2}}{\sqrt{2 \pi m \frac{a\, b\, (c+m)}{c^3}}}$$

Ici N est le plus grand nombre entier contenu dans $(m+1)\,\dfrac{a}{c}$.

On voit que ces limites sont plus grandes que celles qui résultent du principe énoncé. Elles les surpassent dans le rapport de $\sqrt{c+m}$

à $\sqrt{c-m}$ (par exemple de $\sqrt{3}$ à 1, si $m = \frac{1}{2}c$). On était par suite exposé, en les employant, à regarder comme résultats d'une cause constante des écarts beaucoup trop considérables, et qui indiquaient positivement l'existence d'une cause variable.

On comprend sur-le-champ que les limites des écarts des nombres qui ont concouru à former un résultat moyen, doivent être bien moindres que ne le sont celles de nombres qui n'y ont pas contribué, bien que les uns et les autres soient régis par la même possibilité constante. Cette prévision s'accorde avec les formules qui dérivent du principe nouveau. Elles ne dépendent plus que des termes du développement du *binôme des factorielles :* on peut s'en assurer. Et le plus grand terme de cette suite, ainsi que ceux qui l'avoisinent, sont relativement plus grands que les termes correspondants du développement du binôme des puissances. De là des écarts moindres pour une même probabilité.

Extrait de L'Institut, *Journal général des Sociétés et travaux scientifiques de la France et de l'Étranger;* Ire section : Sciences mathématiques, physiques et naturelles; n° 333, du jeudi 14 mai 1840.

PARIS. — Imprimerie d'A. RENÉ et Cie, rue de Seine, 32.