JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

R. RISSER

Exposé des principes de la statistique mathématique. Considérations générales

Journal de la société statistique de Paris, tome 76 (1935), p. 281-318

<http://www.numdam.org/item?id=JSFS_1935__76__281_0>

JOURNAL

DE LA

SOCIÉTÉ DE STATISTIQUE DE PARIS

N° 10. — OCTOBRE 1935

I

EXPOSÉ

DES

PRINCIPES DE LA STATISTIQUE MATHÉMATIQUE

CONSIDÉRATIONS GÉNÉRALES

Ire PARTIE

De quelques définitions de la statistique. — Son but scientifique.

Alors que la méthode expérimentale fait appel à des ensembles homogènes ou s'applique à des faits considérés comme homogènes, la statistique déduit des masses hétérogènes sur lesquelles elle étend ses investigations des éléments objectifs, et permet non point de fixer avec certitude tel point de doctrine, mais simplement de limiter une zone probable d'erreur.

Si d'une part les caractères de la méthode statistique sont aussi ceux de la méthode scientifique, et si d'autre part les champs d'investigation de la statistique touchent à des branches de plus en plus nombreuses du savoir humain, peut-on dire avec Cournot que la statistique est une « science qui a pour objet de recueillir et de coordonner des faits nombreux dans chaque espèce, de manière à obtenir des rapports numériques sensiblement indépendants des anomalies du hasard, et qui dénotent l'existence des causes régulières dont l'action s'est combinée avec celle des causes fortuites » (1).

(1) Exposé de la théorie des chances et des probabilités, COURNOT, p. 182.

D'autres auteurs, à la suite d'Achenwal et d'un des maîtres de la statistique et de la démographie en Allemagne, G. von Mayr, en font presque une sorte de sociologie universelle ayant pour objet de décrire et de grouper méthodiquement les phénomènes de toute espèce que la statistique peut parvenir à compter dans la vie des sociétés.

Levasseur, un des plus grands démographes du XIXe siècle, déclare que la statistique est un procédé d'investigation s'appliquant non à une matière unique, mais à des matières extrêmement diverses embrassant les sciences de la nature qui n'ont pas l'unité nécessaire pour constituer une science.

Pour d'autres auteurs, « la statistique est un instrument et non une fin (1) ».

Si elle n'est point la science sociale, elle fournit à l'économie politique, à la démographie des matériaux des plus intéressants quand elle est préparée puis établie par des statisticiens de profession.

En s'attachant spécialement à l'objet même de la statistique, comme l'a fait Léon Say qui la considérait comme la science des dénombrements et lui assignait comme but de mettre en lumière des régularités, on peut plus exactement la définir, ainsi que l'a fait March, comme « la science de l'hétérogène », et la dénommer pléthométrie (2).

Pour le statisticien comme pour tout autre observateur, le but revient à dégager les éléments immédiats de l'observation des modifications qui les affectent, à procéder à une investigation des causes régissant le phénomène à l'étude, et comme le dit si bien Cournot, « épurer en quelque sorte les conditions du sort ».

Dépouillement. — Groupement. — Recherche des valeurs signalétiques ou caractéristiques.

Lorsque l'on veut étudier un phénomène au point de vue statistique, il faut tout d'abord rechercher le moyen le plus approprié à sa mesure et à son classement, puis effectuer les mesures ou les classifications — soit à proprement parler le relevé statistique qui n'est en définitive que la représentation des manifestations unité par unité du phénomène envisagé. — compte tenu des circonstances qui l'accompagnent en un lieu donné et à un instant donné.

Grâce à l'emploi de fiches, de registres..., on note les dites manifestations et l'on en fait le comptage; on procède ainsi à ce qu'on appelle le dépouillement, puis à la formation des groupements et enfin à l'exposé statistique.

Avant de passer à la manière de rechercher et de calculer les valeurs caractéristiques des phénomènes collectifs à l'étude, nous devons indiquer les modes de classement des observations recueillies.

On peut classer des individus ou des manifestations d'un phénomène, soit *d'après les attributs ou qualités*, qu'ils présentent, et faire ainsi apparaître des *séries* de fréquences, soit *d'après leurs caractères quantitatifs*, c'est-à-dire d'après leurs mesures ou leurs grandeurs, — et introduire des *sériations* de fréquences.

C'est ainsi que nous aurons des séries de fréquence pour des fleurs de la

---

(1) *La statistique, ses difficultés, ses procédés, ses résultats*, de M. André Liesse.
(2) *De la méthode dans les sciences*, 2e série. Étude de March.

même espèce ou de la même variété, groupées d'après la couleur de leur corolle, et des sériations pour des fleurs de la même espèce ou de la même variété ordonnées d'après le nombre de leurs pétales.

On peut dans certains cas transformer une série de fréquence en sériation, et aussi passer d'une sériation à une série.

Les séries sont susceptibles d'être classées selon la nature de leurs attributs ou séries de temps, d'espace et de fait; elles peuvent être statiques ou dynamiques à travers le temps.

Ce travail accompli, il faut passer au calcul de moyennes, d'indices, de coefficients et de rapports, à la distribution des éléments autour de leur moyenne, bref à toute une série d'opérations qui rentrent dans le cadre de ce que l'on appelle aujourd'hui la statistique mathématique, dont l'étude se rattache intimement à celle du calcul des probabilités, calcul dont les principes peuvent servir de guide dans l'examen des phénomènes où interviennent des fréquences, à la condition expresse d'apporter dans l'emploi desdits principes un esprit critique sans cesse en éveil.

Fréquences et probabilités. — On peut imaginer des jeux divers de hasard et rechercher les chances d'arrivée de tel ou tel événement; l'analyse combinatoire et certaines méthodes tirées de l'analyse mathématique permettent le calcul de ces chances. On conçoit facilement une pièce de métal ayant la forme d'un jeton cylindrique dont les deux faces identiques sont recouvertes chacune d'un verni de couleur différente mais d'égale densité et de même épaisseur; le lancement de cette pièce permet d'effectuer des expériences dont les résultats peuvent être comparés avec ceux prévus par l'analyse pure.

Faire intervenir alors la notion de probabilité semble chose des plus naturelles au point de vue mathématique et philosophique; peut-on sans inconvénient adopter cette manière de faire dans des domaines différents de celui ressortissant des jeux ? Pas tout à fait.

Lorsque l'on se trouve en présence de phénomènes démographiques, naissances, mariages, décès..., on ne peut que calculer des rapports qu'il serait dangereux d'assimiler à des probabilités, en raison même des circonstances multiples qui peuvent perturber le phénomène à l'étude; on introduit alors des fréquences et l'on peut — si ces fréquences présentent des fluctuations faibles autour de la moyenne d'ensemble — tenter une explication des résultats afférents au phénomène envisagé, en ayant recours à un schéma de tirage de boules de couleur donnée dans une urne de composition définie.

On conçoit fort bien le taux de mortalité à l'âge x de la population d'un pays, mais ce taux est afférent à un ensemble hétérogène, car la population comprend des personnes valides et invalides, figurant parmi les membres ou anciens membres des divers groupes professionnels de la nation, et des personnes sans profession. Or même si l'on arrivait à constituer avec les individus d'âge x des groupes partiels homogènes, on n'aurait encore la possibilité de ne calculer pour eux qu'une fréquence, car les personnes constituant ces groupes partiels n'auraient pas toutes les mêmes conditions de vie, d'habitat, posséderaient des antécédents héréditaires différents. On arriverait finalement si l'on voulait saisir vraiment la probabilité à introduire tête par tête, ou dans des conditions exceptionnelles des groupes tellement peu denses que les chiffres caractérisant

leur mortalité à l'âge x n'auraient plus de sens au point de vue statistique, comme l'indique le théorème de Bernoulli.

Ce que nous venons de dire pour la mortalité, nous pouvons le répéter pour tous les autres phénomènes démographiques qui se trouvent influencés, non seulement par les facteurs physiologiques, mais encore par les circonstances économiques.

Des causes diverses en plus ou moins grand nombre, dont certaines se neutralisent plus ou moins complètement, viennent perturber les phénomènes statistiques et modifier leurs fréquences; ce sont ces causes perturbatrices que le statisticien doit déceler à travers les résultats de l'observation.

Du prolongement des ensembles statistiques et la conception du calcul des probabilités.

L'étude des problèmes posés par la statistique et par certains jeux ainsi que leur généralisation a amené les mathématiciens et les philosophes à se poser maintes questions se rattachant au prolongement des ensembles statistiques, dont l'examen a pu être conduit à bien, grâce à la théorie mathématique des ensembles, et a eu pour conséquence une révision et une mise au point des concepts du calcul des probabilités. Cette révision a eu une influence heureuse sur certains points de doctrine de la statistique mathématique; aussi avons-nous jugé utile de rappeler rapidement comment les répartitions statistiques font intervenir l'intégrale de Stieljes et les ensembles.

Revenons tout d'abord à la mortalité d'un groupe de population, et remarquons que la probabilité pour une tête de mourir entre l'âge o et l'âge $5n$ (n variant de 1 à 20, si l'on suppose que les têtes du groupe soient toutes disparues à l'âge 100) est une quantité qui croît de 0 à 1; nous avons ainsi l'image d'une courbe de probabilité.

Dans une telle répartition, nous n'observons point de discontinuité, alors que celle afférente à la représentation du gain dans le jeu de pile ou face en fait apparaître une relativement simple; en effet, on a la probabilité $\frac{1}{2}$ de gagner ou de perdre l'enjeu fixé à 1 (par exemple), et l'on constate que la loi de répartition $z = f(x)$ qui se réduit à la portion $(-\infty, -1)$ de l'axe des x pour $x < -1$, est formée ensuite d'un élément de verticale $(0, \frac{1}{2})$ correspondant à l'abscisse (-1), puis d'une droite horizontale $z = \frac{1}{2}$ s'étendant de $x = -1$ (à droite) à $x = +1$ (à gauche), puis d'un nouvel élément de verticale (de $z = \frac{1}{2}$ à $z = 1$) pour $x = +1$, et enfin de la portion de droite horizontale indéfinie vers la droite $z = 1$, à partir du point $x = 1$, $z = 1$.

A la notation gaussienne, on substitue pour l'étude de telles représentations discontinues la notation de Stieljes qui définit l'accroissement $df(x)$ de $f(x)$ dans un intervalle infiniment petit, et la somme $\int_a^b df(x)$ représente $[f(b) - f(a)]$, alors même que l'intégrale intermédiaire $\int_a^b f'(x)\,dx$ n'aura aucun sens.

Or dans le cas des courbes discontinues, il peut exister une probabilité pour

que la grandeur z prenne une valeur déterminée, probabilité qui est alors mesurée par l'accroissement brusque de la fonction $f(x)$ lorsque x atteint une valeur particulière.

Dans la figuration d'une probabilité, on peut en définitive se trouver en présence : 1° soit d'une courbe continue; 2° soit d'une courbe continue dans certains intervalles et avec des verticales en certains points; 3° soit enfin d'ensembles parfaits (1).

On est ainsi amené à examiner des fonctions continues non décroissantes, du type des fonctions en escalier; et dont la dérivée n'existe pas en tous les points de l'intervalle où la fonction est définie, type dont un exemple aujourd'hui classique a été suggéré par M. Lebesgue, puis présenté et généralisé par M. E. Borel d'une manière fort ingénieuse (1).

Comme une loi de probabilité à une variable peut être figurée par une répartition sur l'axe des x de masses positives et de somme égale à l'unité, on peut distinguer autant de sortes ou de classes de lois de probabilité que de sortes de masses. La première classe est constituée par des masses situées en certains points A_n ,ces points pouvant être : 1° en nombre fini; 2° constituer un ensemble dénombrable avec un ou plusieurs points d'accumulation; 3° former un ensemble partout dense. La deuxième classe est relative à la répartition de masses sur l'axe des x avec une densité $\varphi(x)$, et la troisième consiste en la répartition de masses dans un ensemble de mesure nulle, sans qu'aucun point contienne de masse finie, répartition appartenant au type de M. Lebesgue envisagé ci-dessus.

Le cas le plus général qui réside dans la répartition de masses des trois classes a été étendu au plan et à l'espace.

En résumé, au collectif empirique correspondant à des ensembles statistiques finis de m éléments, il est substitué un collectif mathématique avec m tendant vers l'infini, qui fait entrer dans la définition de la probabilité mathématique les propriétés dont le collectif empirique est dépourvu; et il est ainsi tenu compte des besoins de la stochastique et de la biométrie, et l'on constate que la contribution de l'expérience adaptée à la théorie des ensembles a fourni une base solide à cette branche des sciences née de la théorie des jeux avec Pascal et Fermat (2).

Après ces considérations d'ordre général, nous abordons maintenant notre sujet qui comprend deux parties consacrées, la première aux séries statistiques et la seconde à la covariation et à la corrélation.

---

(1) Voir *Principes et formules classiques du calcul des probabilités*, de M. E. Borel.

(2) Voir l'article de Bohlmann sur la théorie mathématique des assurances sur la vie dans l'encyclopédie des sciences mathématiques pures et appliquées (traduction et mise au point de Poterin du Motel), la thèse de M. Broggi, le mémoire de 1909 de M. Borel sur le calcul des probabilités, les « Grundlagen der Wahrscheinlichkeitsrechnung », de M. von Mises, les études de MM. Paul Lévy, Fréchet, Cantelli.

SÉRIES STATISTIQUES

CHAPITRE I

A. — Classements. — Tableaux de nombres. — Graphiques.

1. *Données.* — Le résultat des mesures effectuées sur des ensembles d'individus ou d'objets fournit des collections de nombres; alors que les objets et individus étudiés sont les variables, les nombres provenant des mesures sont les données.

Nous sommes conduit à distinguer les variables continues et les variables discontinues, et nous dirons que la taille des conscrits est une variable continue, le nombre des jours de pluie par mois une variable discontinue, mais nous tenons à faire remarquer que cette distinction est assez artificielle.

On peut dire que ce n'est que dans le cas où l'on opère sur des nombres abstraits, comme le rapport des naissances masculines ou féminines au nombre total des naissances, que l'on peut parler de variables continues et que l'on forme alors des *séries homogrades* suivant la terminologie de l'astronome Charlier. On conçoit ainsi la liste des taux de natalité par département ou par région de territoire pour une année donnée, celle des taux de natalité masculine dans un département donné au cours d'une ou plusieurs périodes décennales; on peut de même considérer le rapport des mariages, des décès, à la population totale pour la formation de séries homogrades.

Lorsque le caractère sur lequel porte l'observation est mesurable et susceptible de prendre un certain nombre de valeurs différentes, on dit que les séries statistiques afférentes à ces observations sont des *séries hétérogrades*; tel est le cas de la distribution des tailles des conscrits dans un certain pays à une époque donnée.

2. *Classements. — Graphiques.*

Les données étant fournies par l'observation, le statisticien doit procéder à leur classement, en se guidant sur l'enseignement de l'expérience, sans oublier un instant que la meilleure manière de procéder pour une collection donnée varie avec le résultat à atteindre; il partage la collection en classes contiguës s'étendant entre deux limites d'intervalle constant, et n'introduit qu'un nombre de classes rarement supérieur à vingt.

Le classement effectué, on pourra supposer, soit que toutes les variables d'une même classe ont la valeur moyenne, soit qu'elles sont réparties uniformément dans la classe; de telles hypothèses exigent un certain discernement pour le choix des limites de la classe, choix qui doit être tel que la distribution réelle et la distribution supposée ne soient pas trop différentes.

Au tableau de nombres ainsi constitué figuratif de la répartition, on peut substituer un graphique que l'on peut représenter de deux façons. Après avoir marqué sur l'axe des abscisses des points équidistants correspondant aux valeurs médianes de chaque classe, on élève en chacun de ces points une

ordonnée proportionnelle au nombre des variables correspondantes. Si l'on joint les extrémités de ces ordonnées, on construit le polygone des trapèzes (première présentation), et si l'on prend chaque ordonnée comme axe d'un rectangle dont la hauteur et la base sont respectivement égales à l'ordonnée et à la grandeur de la classe, l'ensemble de ces rectangles constitue la seconde présentation.

3. *Courbe de fréquence. — Courbe des sommes.*

Aux polygones de fréquences, on associe immédiatement les courbes de fréquences toutes les fois que l'on se trouve en présence de variables continues; dans le cas où l'on considère des séries statistiques afférentes à des variables discontinues, on est conduit, si l'on veut construire des courbes de fréquence à admettre, ce qui est d'ailleurs très souvent contraire à la réalité, une répartition uniforme des individus dans toute l'étendue de chacune des classes.

Les statisticiens emploient encore un autre mode de représentation numérique ou graphique qui réside dans l'utilisation des sommes de fréquences, et par suite des polygones de sommes et enfin des courbes de sommes.

B. — Moyenne et moments.

1. *La moyenne et la médiane.* — Nous avons procédé jusqu'ici à l'établissement d'un tableau, à la construction d'un polygone de fréquences; il nous faut maintenant synthétiser les nombres du tableau ou les graphiques par des valeurs typiques qui sont la moyenne et la médiane.

Si pour les grandeurs x_i de la variable (avec $i = 1, 2, 3, \ldots\ldots n$), l'on a obtenu des fréquences y_i, la *moyenne* m est donnée par la formule $m = \frac{\Sigma x_i y_i}{\Sigma y_i}$, qui représente au point de vue de Quételet la valeur centrale à partir de laquelle la nature a réparti l'élément étudié suivant une certaine loi.

La moyenne est l'abscisse du centre de gravité du polygone des rectangles et du polygone des trapèzes.

La *médiane* μ est l'abscisse de la parallèle à *oy* qui partage le polygone des rectangles en deux parties de même aire; pour le polygone des sommes, c'est l'abscisse du point dont l'ordonnée est égale à la moitié du contenu de la collection.

2. *Dispersion.* — Avant de poursuivre plus loin notre exposé, rappelons que la courbe

$$= \frac{N}{\sqrt{2\pi}\,\sigma} e^{\frac{-x^2}{2\sigma^2}}$$

représente la distribution théorique de N objets symétriques autour d'une moyenne prise pour origine, et qu'elle est d'ailleurs appelée courbe de Laplace-Gauss.

Dans une semblable courbe ou distribution, la médiane est confondue avec la moyenne, et les moments

$$\mu_p = \frac{1}{N} \int_{-\infty}^{+\infty} x^p \, y dx$$

sont tels que l'on a :

$$\mu_{2p+1} = o, \text{ et } \mu_{2p} = 1.3. \ldots. (2p-1)\,\sigma^{2p} = \frac{(2p)!\,\sigma^{2p}}{2.4.6.\ldots.2_p} = \frac{(2p)!\,\sigma^{2p}}{2^p \,.\, p!};$$

quant au moment :

$$\mu_2 = \frac{\int_{-\infty}^{+\infty} x^2\, ydx}{\int_{-\infty}^{+\infty} ydx}$$

que l'on désigne sous le nom d'*écart quadratique moyen*, ou encore par σ, il est d'autant plus petit que les valeurs de x sont plus agglomérées autour de la moyenne, ce qui lui a valu d'être appelé *standard-déviation* par les statisticiens anglais, et de caractériser la mesure de la dispersion.

Rappelons à ce propos que M. E. Borel appelle unité la grandeur $\sigma\sqrt{2}$, et désigne par *écart médian*, l'écart qui ayant probabilités égales d'être dépassé ou non, est déterminé par la relation :

$$\frac{1}{N}\int_{-\infty}^{+\infty} ydx = \frac{1}{2}, \quad \text{d'où} \quad x = \pm\, 0{,}67449\,\sigma = 0{,}4769 \text{ } \textit{écart unité.}$$

Ceci étant, on caractérise une distribution quelconque au moyen de la grandeur σ, ou de sa *dispersion*, en faisant état de la formule $\sigma^2 = \frac{\Sigma\,(x-m)^2\,y}{\Sigma\,y}$, qui se ramène à $\sigma^2 = m_2 - m^2$ (m_2 étant le moment d'ordre 2 de la distribution envisagée).

Est-il possible de rattacher à une interprétation analytique les valeurs typiques (moyenne et médiane); c'est ce que nous allons voir :

Soient $x_1, x_2, \ldots. x_N$ les valeurs de la variable rangées dans l'ordre croissant, A la valeur typique, et $(x_i - A)$ l'écart de x_i avec A. On peut déterminer A par la condition que les écarts les plus grands en valeur absolue soient aussi petits que possible; il suffit à cet effet de prendre $A_1 = \frac{x_1 + x_N}{2}$.

Une telle valeur typique A_1 ne présente qu'un intérêt secondaire en statistique, pour cette raison simple que x_1 et x_N sont des valeurs exceptionnelles anormales; elle est toutefois employée par les météorologistes pour la détermination de la température moyenne journalitère.

On peut caractériser une valeur typique A_2 par la condition de rendre minimum la *moyenne des valeurs absolues des écarts*, moyenne que l'on désignera ordinairement par écart moyen.

Posons :

$$A_p = x_p, \quad A_{p+1} = x_{p+1},$$

et formons

$$\Sigma\,|\,x_i - A_p\,| - \Sigma\,|\,x_i - A_{p+1}\,| = (N - 2p)\,(x_{p+1} - x_p);$$

il résulte de là que :

$$\Sigma\,|\,x_i - x_p\,|$$

décroît lorsque p augmente si $(N - 2p) > 0$.

Il s'en suit que si N est *impair*, il y a un individu de rang médian $A_2 = x_{p+1}$, tel que l'on a :

$$\Sigma\,|\,x_i - x_q\,| - \Sigma\,|\,x_i - A_2\,| > o, \quad \text{avec} \quad q \leqslant p$$
$$\Sigma\,|\,x_i - A_2\,| - \Sigma\,|\,x_i - x_r\,| < o, \quad \text{avec} \quad r \geqslant p+2,$$

mais il est essentiel de remarquer qu'il peut y avoir plusieurs individus de même grandeur, et par suite la valeur médiane est indéterminée entre eux.

Si N est pair, il y a deux individus de rang médian $p = \frac{N}{2}$; l'écart est minimum pour les deux abscisses et pour toute abscisse intermédiaire.

On peut enfin définir la valeur typique par la condition de rendre minimum la somme des carrés des écarts, et l'on est dans cette hypothèse conduit à la valeur $A_3 = m$.

En désignant les écarts $\varepsilon_i = | x_i - A |$, à partir d'une valeur quelconque A, on est amené à considérer l'écart moyen à partir de A, soit $\frac{\sum_1^n \varepsilon_i}{N} = e_A$, et l'écart quadratique à partir de A, soit $\sigma_A = \sqrt{\frac{\sum \varepsilon_i^2}{N}}$, et l'on remarque que $\sigma_A > e_A$.

La considération de l'écart quadratique moyen conduit immédiatement au théorème de Bienaymé qui donne la fréquence relative des individus dont l'écart avec A est supérieur à $k\,\sigma_A$.

En effet, dans la somme $\Sigma (x_i - A)^2 = N\sigma_A^2$, il y a n termes pour lesquels

$$(x_i - A)^2 > k^2\,\sigma_A^2,$$

et par suite :

$$\Sigma (x_i - A)^2 \geqslant nk^2\,\sigma_A^2, \text{ ou } N\,\sigma_A^2 \geqslant nk^2\,\sigma_A^2, \text{ soit } \frac{n}{N} \leqslant \frac{1}{k^2};$$

$\frac{n}{N}$ qui est la fréquence cherchée limitée par la fraction $\frac{1}{k^2}$, dépend donc de l'écart donné et de l'écart quadratique moyen σ_A relatif à A.

3. *Valeurs typiques secondaires. — Quartils. — Mode.*

En même temps que la médiane, on fait intervenir les deux quartils qui avec cette même médiane, sont les abscisses des trois parallèles à Oy, partageant l'aire du polygone des rectangles en quatre parties égales; on met ainsi en évidence deux des trois valeurs typiques secondaires.

Quant à l'abscisse afférente à la plus grande fréquence, représentée par M et désignée sous le vocable de *mode* par les statisticiens anglais, elle est relativement difficile à évaluer d'une manière précise.

Grâce à une interpolation au voisinage du maximum visible basée sur l'emploi de la parabole $y = ax^2 + bx + c$ passant par les points (x_i, y_i) avec $i = (1, 2, 3)$ et $y_1 < y_2$, $y_2 > y_3$, on trouve pour l'abscisse du sommet :

$$M = \frac{y_1 - y_3}{2\,(y_1 + y_3 - 2\,y_2)}$$

On ne peut calculer d'une manière un peu précise cette grandeur qu'en secourant à une opération d'ajustement, mais on doit faire remarquer que l'introduction des courbes ajustées nécessite des hypothèses dont la justification est loin d'être absolue; rappelons enfin que l'étude de l'ajustement conduit à la relation $3\,(m - \mu) = m - M$, ou encore $2\,m + M = 3\,\mu$ (1).

(1) Nous indiquerons plus loin un cas d'ajustement signalé par M. Gibrat, où la formule précédente peut être considérée comme fournissant une excellente approximation.

3'. *Médiale. — Interquartal. — Différence moyenne. — Pente moyenne générale.*

A côté des caractéristiques classiques : moyennes de divers ordres, médiane, mode, quartils que nous venons de signaler, est apparue depuis fort peu de temps une caractéristique nouvelle dite *médiale* à laquelle se rattachent les quartals, et qui dépend comme la médiane d'ailleurs de l'ordre des élémentr afférents aux grandeurs à l'étude. Il suffit pour s'en rendre compte de considérer le classement des ouvriers d'une usine d'après les salaires horaires; le salaire de l'ouvrier qui se trouve au milieu de la série donne la valeur *médiane m*, alors que le salaire de l'ouvrier qui est tel que le montant des salaires inférieurs au sien est égal au montant des salaires supérieurs au sien, répond à la *médiale m'*.

March reprenant en 1928 (1) une étude faite par lui en 1898 (2) — où il comparait la distribution des salaires de 13.000 ouvriers métallurgistes américains dont on avait pu apprécier les salaires individuels, — a été conduit à faire des remarques fort judicieuses sur la comparaison de ces diverses caractéristiques, et à utiliser — en vue de mettre en lumière l'inégalité des salaires, l'intervalle des quartils ou interquartil, puis l'interquartil relatif ou rapport de cette dernière quantité à la médiane, et enfin l'interquartal et l'interquartal relatif (3).

Ces divers indices, moyenne, médiale..., ne donnant point encore une idée vraiment satisfaisante de l'inégale répartition des salaires, il était donc nécessaire de faire apparaître des caractéristiques de variabilité qui fussent indépendantes de l'ordre des éléments. C'est à M. Gini que l'on doit les caractéristiques dites *d* « *différence moyenne* » (4).

$$d = \frac{4 \Sigma x_i y_i}{n^2},$$

et de « *pente moyenne générale* » :

$$p = \frac{d n^2}{4 \Sigma y^2}$$

$$\left[\text{qui est une moyenne, puisqu'elle revient à } \frac{\Sigma (x_i \cdot y_i)}{\Sigma (y_i \cdot y_i)}\right],$$

et où figurent le nombre n des éléments de la série, x_i l'écart entre la grandeur d'un élément quelconque et la grandeur de la médiane, et y_i le nombre des éléments compris entre cet élément et la médiane.

Il est utile de rappeler que dans l'expression de d, l'on a fait intervenir non seulement les différences distinctes, mais aussi les différences nulles.

4. *Moments.* — L'ajustement de la courbe des fréquences fait apparaître les sommes $\Sigma x_i^p y_i$, où $y_1, y_2 \ldots\ldots y_n$ sont les contenus des classes d'abscisses respectives $x_1, x_2, \ldots\ldots x_n$ (nombres entiers successifs), et les grandeurs dites

(1) « Différences et corrélation en statistique », *Journal de la Société de Statistique de Paris*, février 1928.

(2) « Quelques exemples de distribution de salaires », *Journal de la Société de Statistique de Paris*, 1898.

(3) Voir, du même auteur, « Les mesures de la variabilité ». Métron, vol. VI, n° 2, p. 56.

(4) Voir l'intéressant article de M. Gini, dans le volume VI de MÉTRON.

moments d'ordre p, soit $m_p = \frac{1}{N} \Sigma x_i^p y_i$ (avec $N = \Sigma y_i$), auxquelles on rattache les moments μ_p par rapport à la moyenne $m_1 = m = \frac{\Sigma x_i y_i}{N}$.

$$\mu_p = \frac{1}{N} \Sigma (x_i - m)^p \cdot y_i = m_p - m \cdot p \cdot m_{p-1} + m^2 \frac{p(p-1)}{2!} m_{p-2} + \ldots\ldots$$

Le calcul des moments peut être effectué en ayant recours seulement à des additions; il suffit à cet effet de calculer les éléments suivants :

$$S_1^0 = y_n,\ S_2^0 = y_n + y_{n-1},\ S_3^0 = S_2^0 + y_{n-2}, \ldots\ldots$$
$$S_1^1 = S_1^0,\ S_2^1 = S_1^1 + S_2^0, \ldots\ldots\ldots\ldots\ S_n^1 = S_{n-1}^1 + S_n^0,$$
$$S_1^2 = S_1^1,\ S_2^2 = S_1^2 + S_1^1, \ldots\ldots\ldots\ldots\ S_n^2 = S_{n-1}^2 + S_n^1,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$

grâce auxquels on trouve $S_n^p = \frac{\Sigma\, r\,(r+1)\ldots(r+p-1)}{p!} = \frac{m_p . N}{p!} +$

$+ m_{p-1} \cdot \frac{N}{2(p-2)!} + \ldots\ldots$; des relations entre m_p, m_{p-1}, ... et S_n^p, on déduit les m_p en fonction des S_n^p et des m_{p-j} avec $j = (1, 2, \ldots\ldots)$:

$$m_0 = N = S_n^0;\ m_1 = m = \frac{S_n^1}{S_n^0};\ m_2 = \frac{2 S_n^2}{S_n^0} - m_1; \ldots\ldots$$

et aussi

$$\mu_2 = \frac{2 S_n^2}{N} - m(m+1);\ \mu_3 = \frac{6 S_n^3}{N} - 3(m+1)\mu_2 - m(m+1)(m+2); \ldots\ldots$$

La précédente méthode a subi une modification qui a pour but d'éviter l'introduction des valeurs trop grandes de S_n^j et qui est fondée sur le principe suivant :

Soient :

$$-(k-1), -(k-2), \ldots\ldots -1, 0, 1, 2, \ldots\ldots (n-k)$$

et

$$y_1, \quad y_2, \quad \ldots\ldots\ldots\ldots\ y_{k-1}, y_k, y_{k+1}, \ldots\ldots\ldots\ y_n,$$

les abscisses et les ordonnées correspondantes.

Formons maintenant les sommes :

$$\begin{cases} S_1^0 = y_n, & S_2^0 = S_1^0 + y_{n-1}, \ldots\ldots & , & S_{n-k}^0 = S_{n-k-1}^0 + y_{k+1}; \\ \sigma_1^0 = y_1, & \sigma_2^0 = \sigma_1^0 + y_2 \quad , \ldots\ldots & , & \sigma_{k-1}^0 = \sigma_{k-2}^0 + y_{k-1}; \\ S_1^1 = S_1^0, & S_2^1 = S_1^1 + S_2^0 \quad , \ldots\ldots & , & S_{n-k}^1 = S_{n-k-1}^1 + S_{n-k}^0; \\ \sigma_1^1 = \sigma_1^0, & \sigma_2^1 = \sigma_1^1 + \sigma_2^0 \quad , \ldots\ldots & , & \sigma_{k-1}^1 = \sigma_{k-2}^1 + \sigma_{k-1}^0; \\ S_1^2 = S_1^1, & S_2^2 = S_1^2 + S_2^1 \quad , \ldots\ldots & , & S_{n-k}^2 = S_{n-k-1}^2 + S_{n-k}^1; \\ \sigma_1^2 = \sigma_1^1, & \sigma_2^2 = \sigma_1^2 + \sigma_2^1 \quad , \ldots\ldots & , & \sigma_{k-1}^2 = \sigma_{k-2}^2 + \sigma_{k-1}^1; \end{cases}$$

d'où l'on déduit :

$$1^0)\ N = S_{n-k}^0 + \sigma_{k-1}^0 + y_k,$$
$$2^0)\ S_{n-k}^1 - \sigma_{k-1}^1 = m N,$$
$$3^0)\ S_{n-k}^2 + \sigma_{k-1}^0 = \frac{m_2 N}{2} + \frac{1}{2}(S_{n-k}^1 + \sigma_{k-1}^1),$$
$$4^0\ S_{n-k}^3 - \sigma_{k-1}^3 = \frac{m_3 N}{6} + \frac{1}{2}(S_{n-k}^2 - \sigma_{k-1}^2) - \frac{1}{6}(S_{n-k}^1 - \sigma_{k-1}^1),$$

et après avoir posé :

$$S^p_{n-k} + \sigma^p_{k-1} = S_p,\ S^p_n - S^p_{k-1} = D_p,$$

on aboutit au système des relations :

$$N = S_o + y_k,\ m\,N = D_1,\ m_2\,N = 2\,S_2 - S_1$$
$$m_3\,N = 6\,D_3 - 3\,D_2 + D_1$$
$$m_4\,N = 24\,S_4 - 36\,S_3 + 14\,S_2 - S_1$$
$$m_5\,N = 120\,D_5 - 240\,D_4 + 150\,D_3 - 30\,D_2 + 76\,D_1$$

et aux valeurs des moments pris par rapport à la moyenne :

$$\mu_2 = \frac{2\,S_2 - S_1}{N} - m^2;\ \mu_3 = \frac{6\,D_3 - 3\,D_2 + D_1}{N} - 3\,m\,\mu_2 - m^3;$$

$$\mu_4 = \frac{24\,S_4 - 36\,S_3 + 14\,S_2 - S_1}{N} - 4\,m\,\mu_3 - 6\,m^2\,\mu_2 - m^4; \ldots\ldots$$

4'. *Corrections.* — Tous les calculs faits jusqu'ici supposent que le contenu d'une classe est concentré sur son abscisse moyenne, et ne correspondent par suite qu'au cas des variables discontinues.

On est donc conduit à corriger les éléments caractéristiques qui en ont été déduits, et on a recours en l'occurrence à la méthode de Sheppard.

Soit $y = f(x)$ l'équation de la courbe de fréquence, N son aire totale ($N = \int_a^b f(x)\,dx$), et y_i l'aire comprise entre les abscisses $x_i \pm \frac{1}{2}$,

Soit :

$$y_i = \int_{x_i - \frac{1}{2}}^{x_i + \frac{1}{2}} f(x)\,dx.$$

Or on vient de calculer $\mu_p = \frac{1}{N} \sum_{a'}^{b'} x_i^p\,y_i$ $\left(\text{avec } a' = a + \frac{1}{2},\ b' = b - \frac{1}{2}\right)$ alors que le moment calculé d'après l'équation de la courbe est $\mu_p = \frac{1}{N}\int_a^b x^p f(x)\,dx$; il faut donc chercher la correction qu'il y a lieu d'appliquer à μ_p.

Comme $y_i = \int_{x_i - \frac{1}{2}}^{x_i + \frac{1}{2}} f(x)\,dx = f(x_i) + \frac{1}{24} f'(x_i) + \frac{1}{1920} f^{(IV)}(x_i) + \ldots\ldots$

on trouve — en faisant usage d'une formule analogue à la formule sommatoire d'Euler-Mac Laurin, et eu égard à une hypothèse qui se trouve vérifiée pour la courbe de dispersion de Gauss dans le cas de limites infinies, et pour le groupe des courbes de Pearson avec les limites finies ou infinies que chacune d'elles comporte, — que $(\alpha)\ \mu_p = \mu'_p + \frac{p(p-1)}{24}\mu'_{p-2} + \frac{p(p-1)(p-2)(p-3)}{1920}\mu'_{p-4}$, après avoir négligé les termes qui contiennent la 5e puissance de l'intervalle unité.

On peut aussi, comme l'a fait M. Traynard (*Annales de l'École Normale*, 1909) en suivant le processus de Pearson, procéder au calcul des moments de la courbe des fréquences représentée soit par le polygone des trapèzes, soit par le polygone des rectangles qui fournissent respectivement pour valeurs

de μ'_p des expressions (β) et (γ) qui sont loin d'être identiques entre elles et à (α).

Il n'est pas sans intérêt de rappeler que si dans le cas envisagé par l'expérience les formules de Sheppard sont applicables, ce sont celles-là qu'il y aura lieu d'utiliser, mais elles devront être soumises au critère de l'erreur probable.

5. *Retour sur l'inégalité de Tchebicheff-Bienaymé. — Rappel de la loi des grands nombres.* — Nous avons fait remarquer que la loi de répartition d'un phénomène est définie d'une manière complète par les masses p_i placées aux points d'abscisses x_i, puisqu'il est possible de calculer immédiatement les moments des divers ordres et en particulier $m_1 = \Sigma\, p_i\, x_i$ ou l'espérance mathématique, et $m_2 = \Sigma\, p_i\, x_i^2$.

Si l'on rapporte les moments d'ordre (2, 3, ...) au point d'abscisse m_1, on introduit ainsi la grondeur σ_x afférente au moment d'ordre 2 dont nous avons fait état pour établir une inégalité due à la fois à Bienaymé et à Tchebicheff, d'où l'on déduit que la probabilité totale P_2 que $|(X)| \leqslant t\sigma_x$ est supérieure ou au moins égale à $\left(1 - \frac{1}{t^2}\right)$.

Tchebicheff a d'ailleurs généralisé ce résultat et montré que pour l'écart moyen d'ordre p, il y avait lieu de substituer à l'inégalité précédente l'inégalité suivante :

$$P_p \geqslant 1 - \frac{1}{t^p}.$$

Ceci étant, considérons une urne renfermant des boules blanches et des boules noires, en proportions respectives p et q, et effectuons n tirages en remettant chaque fois la boule extraite de l'urne, tirages auxquels nous faisons correspondre les grandeurs aléatoires $X_1, X_2, \ldots\ldots X_n$, qui sont indépendantes. On démontre facilement : 1° qu'à la variable aléatoire $Z = \Sigma\, X_i$, on rattache son espérance mathématique, et 2° que l'espérance mathématique de $(Z - np)^2$ est égale à (npq); de là, il résulte que la variable aléatoire $F = \frac{Z}{n}$, ou fréquence des succès, a pour espérance mathématique p, et que l'écart quadratique afférent à la fréquence F ou σ_F a pour valeur $\sqrt{\frac{pq}{n}}$. De ces résultats et de l'inégalité de Tchebicheff-Bienaymé, on déduit une démonstration simple et fort élégante du fameux théorème de Jacques Bernoulli (1), qui jette un pont entre le calcul des probabilités et la statistique et qui s'énonce ainsi qu'il suit :

La probabilité que le rapport du nombre des boules blanches extraites au nombre total des boules sorties, ne s'écarte pas au delà d'un intervalle donné, du rapport du nombre des boules blanches au nombre total des boules contenues dans l'urne, s'approche indéfiniment de la certitude, par la multiplication indéfinie des événements quelque petit que l'on suppose cet intervalle (2).

C. — De la fonction caractéristique.

On sait que la connaissance des moments d'une fonction détermine cette fonction, mais dans le domaine de la statistique, il ne faut point perdre de vue

(1) *Ars conjectandi opus posthumum.* Bâle, 1713.
(2) Laplace, *Introduction à la théorie philosophique des probabilités.*

que les moments du fait de leur caractère expérimental, ne peuvent pas toujours être calculés d'une manière précise.

Aux moments, on peut associer facilement une certaine fonction, dite *fonction caractéristique* de la loi de probabilité donnée; en effet, considérons une série de valeurs x_m d'une variable discontinue de probabilités respectives y_m, et calculons la valeur probable de e^{tx}, c'est-à-dire

$$\sum_1^N y_m e^{tx_m} = \varphi(t) = 1 + tm_1 + \frac{t^2}{2!} m_2 + \ldots\ldots + \frac{t^p}{p!} m_p + \ldots\ldots,$$

les moments m étant évalués à partir de $x = 0$.

On constate de suite que la grandeur m_p n'est autre que $\left(\frac{d^p \varphi}{dt^p}\right)_{t=o}$, et l'on remarque que $\varphi(t)$ satisfait à une équation différentielle d'ordre N à coefficients constants.

Si l'on a deux séries statistiques (x_m, y_m), (x'_n, y'_n), auxquelles correspondent les fonctions caractéristiques φ et φ_1, la fonction caractéristique Φ afférente à $(x_m + x'_n)$, dans l'hypothèse où x_m et x'_n sont indépendantes, n'est autre que $(\varphi \times \varphi_1)$; en effet, la probabilité de $(x_m + x'_n)$ est $(y_m \,.\, y'_n)$ et par suite :

$$\Phi(t) = \sum_1^N \sum_1^{N'} y_m y'_n \,.\, e^{t(x_m + x'_n)} = \varphi(t) \,.\, \varphi_1(t).$$

A titre d'application, considérons une urne contenant des boules blanches et des boules noires (p étant la probabilité de sortie d'une blanche et q celle d'une noire), et faisons n *tirages*, en remettant chaque fois dans l'urne la boule tirée.

La fonction caractéristique pour un tirage est $(pe^t + q)$, et pour n tirages cette fonction n'est autre que :

$$(pe^t + q)^n = \left[p\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots\right) + q\right]^n = \left(1 + pt + \frac{pt^2}{2!} + \frac{pt^3}{3!} + \ldots\right)^n$$
$$= 1 + npt + \frac{t^2}{2!}\left(n^2 . p^2 + npq\right) + \ldots,$$

si l'on prend les moments par rapport à $x = 0$.

On en déduit $m_1 = np$, $m_2 = npq + n^2p^2$, $m_3 = npq\,(q - p) + 3\,n^2p^2 + n^3p^3, \ldots\ldots$

Si l'on compte les abscisses à partir de la valeur probable, on remarque immédiatement que la fonction caractéristique relative aux n tirages n'est autre que $(pe^{qt} + qe^{-pt})^n$, et dans ces conditions l'on a la relation :

$$\varphi(t) = 1 + \frac{t^2}{2!} u_2 + \frac{t^3}{3!} \mu_3 \ldots\ldots = U^n = [(1 - q)\, e^{qt} + qe^{(q-1)t}]^n,$$

d'où l'on déduit la relation de récurrence.

$$\mu_{s+1} = pq\left(n s \mu_{s-1} - \frac{\partial \mu_s}{\partial q}\right),$$

et par suite $\mu_1 = 0$, $\mu_2 = npq$, $\mu_3 = npq\,(q - p), \ldots\ldots$; nous retrouvons ainsi à propos de μ_2 un résultat qui a été obtenu par un procédé tout à fait différent, indiqué d'ailleurs ci-dessus, et nous remarquons que la quantité désignée par certains auteurs, sous le nom d'*unité d'écart*, n'est quatre que $\sqrt{2\,\mu_2}$.

A la loi de fréquence de Poisson $y_k = e^{-A} \cdot \frac{A^k}{k!}$ (dite loi des petites probabilités), on fait correspondre la fonction caractéristique :

$$\varphi(t) = \sum_{o}^{\infty} e^{-A} \frac{A^k}{k!} e^{kt} = e^{A(e^t - 1)},$$

et par suite les valeures de m_s :

$$m_1 = A,\ m_2 = A + A^2,\ m_3 = A + 3A^2 + A^3,\ \ldots\ldots,\ \text{et celles de } \mu_s :$$
$$\mu_2 = A,\ \mu_3 = A,\ \mu_4 = A + 3A^2,\ \ldots\ldots$$

Rappelons enfin qu'à la distribution continue de probabilité $f(x)$, on rattache le moment m_p

$$m_p = \int_a^b f(x)\, x^p\, dx,\ \text{et la fonction caractéristique } \varphi(t) = \int_a^b e^{tx} f(x)\, dx.$$

Si la loi de probabilité est la loi *normale* $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$ (x variant de $-\infty$ à $+\infty$), la fonction caractéristique correspondante est :

$$\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{tx}\, e^{-\frac{x^2}{2\sigma^2}}\, dx = e^{\frac{t^2\sigma^2}{2}},$$

et les moments μ_{2p} et μ_{2p+1} ont respectivement pour valeur :

$$\mu_{2p} = \frac{(2p)!}{2^p\, p!}\, \sigma^{2p},\ \mu_{2p+1} = 0.$$

M. P. Lévy substitue à la fonction caractéristique $\varphi(t)$ la fonction définie par la relation :

$$\Phi(t) = \int_{-\infty}^{+\infty} f(x)\, e^{itx}\, dx,\ \text{d'où l'on déduit } f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Phi(t)\, e^{-itx}\, dt.$$

La formule de la fonction caractéristique de la loi normale donne de suite la propriété classique due à Poisson et à Cauchy, et retrouvée par M. d'Ocagne.

Soient en effet les variables indépendantes x_i ($i = 1, 2, 3, \ldots\ldots n$) suivant la loi normale de probabilité et de dispersions respectives σ_i, la fonction caractéristique de $X = \Sigma x_i$ étant définie par $e^{\frac{-t^2 \Sigma \sigma i^2}{2}}$, il en résulte que X suit également la loi normale, et que sa dispersion est définie par la relation $S^2 = \Sigma\, \sigma_i^2$.

Rappelons enfin qu'à la loi de probabilité définie par un développement de la forme :

$$(1)\ f(x) = A_o f_o(x) + A_1 f_1(x) + \ldots\ldots\ldots + A_p f_p(x) + \ldots\ldots$$

dans lequel $f_0, f_1, f_2, \ldots\ldots f_p \ldots\ldots$ ne sont autres que la fonction classique :

$$\frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

et ses dérivées successives, correspond la fonction caractéristique

$$\varphi(t) = \varphi_o(t)\,[A_o - A_1 t + \ldots\ldots\ldots + (-1)^p A_p t^p + \ldots\ldots\ldots],$$

si l'on fait usage du facteur e^{tx} de Cauchy, et la fonction

$$\Phi(t) = \Phi_o \left[A_o + \sum_{p=1}^{\infty} (-it)^p A_p \right]$$

lorsque l'on a recours au facteur e^{itx}.

Si maintenant, l'on prend pour origine des abscisses la valeur probable de x, et si l'on adopte comme unité de mesure l'écart quadratique, l'on se trouve en présence de lois réduites, auxquelles correspondent les fonctions caractéristiques :

$$\Phi(t) = \Phi_o [1 + A_3 (-it)^3 + \ldots\ldots], \text{ et } \varphi(t) = \varphi_o [1 - \alpha_3 t^3 + \ldots\ldots,$$

suivant que le facteur adopté pour le calcul de la fonction caractéristique est e^{itx}, ou e^{tx}. *Si au lieu de faire sn tirages dans une urne de composition donnée, on procède, comme l'a indiqué Poisson, à une première opération consistant en s extractions successives dans les urnes 1, 2, 3,.....s, ayant les compositions respectives* $(p_1, q_1) \ldots\ldots (p_s, q_s)$ *en boules blanches et boules noires, et si l'on recommence n fois cette même opération*, on peut, — grâce à l'emploi de la fonction caractéristique — définir tout d'abord la fonction caractéristique afférente à la première opération, soit $\lambda(t) = (q_1 + p_1 e^t)(q_2 + p_2 e^t) \ldots\ldots (q_s + p_s e^t)$, puis la fonction caractéristique relative à l'ensemble des n opérations partielles, soit $\varphi(t) = [\lambda(t)]^n$.

Après transport de l'origine des variables aléatoires au point d'abscisse $\dfrac{p_1 + p_2 + \ldots\ldots + p_s}{s}$, on remarque que le logarithme de la fonction caractéristique d'ensemble (c'est-à-dire afférente aux n opérations ou encore aux (ns) tirages) s'écrit :

$$\varphi(\theta) = \frac{\theta^2}{2} + \frac{\theta^3}{\sqrt{n}} a_3 - \frac{\theta^4}{n\sqrt{n}} a_4 + \ldots\ldots, \text{ après avoir posé } t = \frac{\theta}{\sqrt{n \Sigma p_i q_i}}$$

On aurait pu tout aussi bien — comme l'a d'ailleurs indiqué Lexis — faire une première épreuve de n tirages dans l'urne (p_1, q_1), puis une deuxième épreuve de n tirages dans l'urne (p_2, q_2), ... et enfin une s^{eme} épreuve de n tirages dans l'urne (p_s, q_s).

On peut encore généraliser ce qui vient d'être dit en procédant à μ groupes d'épreuves, les nombres d'épreuves de ces groupes étant respectivement $(n_1, n_2, \ldots\ldots n_\mu)$, les probabilités des événements favorables étant $(p_1, p_2, \ldots\ldots p_\mu)$, et les probabilités contraires $(q_1, q_2, \ldots\ldots q_\mu)$, et envisager cet ensemble d'épreuves dans lequel l'unité d'écart u sera défini par :

$$(1)\ u^2 = u_1^2 + u_2^2 + \ldots\ldots\ldots + u_\mu^2,$$

où u_i est l'unité d'écart correspondant au groupe i.

Si maintenant l'on remplace les μ groupes d'épreuves par une série de N épreuves équivalentes, dans lesquelles les probabilités de l'événement favorable et de l'événement contraire résultent des équations $N = \sum_{i=1}^{\mu} n_i$, $Np = \sum_{i=1}^{\mu} n_i p_i$, $Nq = \sum_{i=1}^{\mu} n_i q_i$, l'on démontre que l'unité d'écart U rattachée à cet

ensemble fictif de N épreuves est telle que (2) $U^2 = u^2 + \frac{2}{n} \Sigma \Sigma' n_i n_k (p_i p_k)^2$, le signe Σ' indiquant que la dernière sommation ne s'étend qu'une seule fois à chaque produit $n_i n_k$, c'est-à-dire que l'on ne fait pas intervenir $n_k n_i$, ou encore que l'on suppose $i > k$.

L'opération globale, consistant en s tirages faits dans une même urne de composition p, q, lorsque l'on a eu soin de remettre chaque fois dans l'urne la boule extraite, correspond au cas de Bernoulli, et fournit ce que les statisticiens appellent une *série normale*, c'est-à-dire une série où la dispersion est définie par $\sqrt{spq}$; l'opération préconisée par Poisson des s tirages successifs dans les urnes (p_1, q_1), ... (p_s, q_s) fait apparaître une série dite *hyponormale*, c'est-à-dire telle que sa dispersion soit inférieure à $\sqrt{s p_o q_o}$, avec $p_o = \frac{\Sigma p_i}{s}$, $q_o = \frac{\Sigma q_i}{s}$.

Toujours dans l'hypothèse de remise des boules dans les urnes après chaque tirage, on remarque que le type de tirage ou de schéma préconisé par Lexis, fait apparaître une dispersion globale supérieure à $\sqrt{s p_o q_o}$, et conduit à une série dite *hypernormale*.

CHAPITRE II

POLYGONE BINOMIAL. — COURBES DE PEARSON.— SÉRIES DE BRUNS-CHARLIER.

L'étude d'une série statistique se rapportant à un événement qui dépend de deux probabilités complémentaires peut être effectuée grâce au théorème de Bernoulli associé à la théorie des épreuves répétées.

Si d'une urne renfermant Np boules blanches et Nq boules noires, l'on tire *n boules en remettant à chaque fois la boule tirée*, on sait que la probabilité de la sortie de k boules blanches est égale à :

$$P_k = \frac{n!}{k!\,(n-k)!}\, p^k q^{n-k}, \text{ que l'on écrit encore } \binom{n}{k} p^k q^{n-k}.$$

De plus, l'application du théorème de Bernoulli nous montre que le comptage des boules blanches tirées dans un certain nombre H de séries de n tirages fait apparaître une courbe de fréquence expérimentale dont les ordonnées sont sensiblement proportionnelles à P_k, et que les écarts entre les données expérimentales et les chiffres théoriques sont d'autant plus faibles que le nombre H est plus élevé.

Le terme P_k est *maximum* si l'on a à la fois $(n+1)\,p - 1 < k < (n+1)p$; quant à la valeur approchée P du maximum, elle est définie par l'expression : $\log P = \frac{1}{2} \log (2\,\pi\, npq) + \frac{1}{12n} - \frac{1}{12\,npq}$, à laquelle on rattache les inégalités

$$\frac{1}{\sqrt{2\,\pi\,(n+1)\,pq}} < P < \frac{1}{\sqrt{2\,\pi\, npq}}.$$

Rappelons enfin que l'on peut avec une approximation suffisante remplacer P_k par :

$$\frac{1}{\sqrt{2\pi(n+1)pq}}\, e^{\frac{-t^2}{2pq}}\left[1+\frac{t^3}{6\sqrt{n-1}}\,\frac{q-p}{p^2q^2}\right],$$

où t n'est autre

$$\frac{k+\frac{1}{2}-(n+1)p}{\sqrt{n+1}}.$$

Dans le cas des petites probabilités, Poisson a montré que l'on pouvait — avec une approximation assez grande, alors même que n était assez petit — remplacer P_k par $e^{-A}\cdot\frac{A^k}{k!}$, avec $(n+1)p = A$, et mis en lumière ce fait intéressant que dans l'application de la formule des épreuves répétées, il suffisait de *connaître la valeur la plus probable de k.*

A) *Courbes du binôme.* — La pente p_k du côté du polygone binomial joignant les points (x_{k-1}, y_{k-1}), (x_k, y_k) est égale à $(y_k - y_{k-1})$, puisque $\Delta x_{k-1} = 1$ et que l'ordonnée centrale de ce côté n'est autre que $\frac{y_{k-1}+y_k}{2} = y'_k$; il s'ensuit que l'examen du rapport $\frac{y_k - y_{k-1}}{\frac{y_k+y_{k-1}}{2}}$ conduit à la relation :

$$(1)\quad \frac{p_k}{y'_k} = \frac{\frac{-2}{q-p}\,x'_k}{\frac{2pq(n+1)}{q-p}+x'^2_k},\ \text{où}\ x'_k = k-\frac{1}{2}-\left[(n+1)p-\frac{1}{2}\right],$$

et l'on peut dire que l'équation différentielle de la courbe tangente à chaque côté du polygone binomial en son milieu est du type (2) $\frac{dy}{dx} = -\frac{\gamma xy}{x+a}$.

On constate que dans le cas particulier de $p = q$, qui correspond au jeu de pile ou face, l'équation (1) qui devient

$$\frac{p_k}{y'_k} = \frac{-x'_k}{\frac{n+1}{4}},\ \text{conduit à}\ \frac{dy}{dx} = \frac{-xy}{\sigma^2},$$

et par suite à la courbe classique de Laplace-Gauss.

$$y = y_0\, e^{\frac{-x^2}{2\sigma^2}}.$$

Dans le cas général ($p < q$), la solution de l'équation différentielle (2) est la suivante :

$$(3)\quad y = y_0\left(1+\frac{x}{a}\right)^{\gamma a} e^{-\gamma x},\ \text{avec}\ \gamma = \frac{2}{q-p},\ a \quad \frac{2pq(n+1)}{q-p}.$$

Grâce à un développement en série, on peut mettre tout d'abord y sous la forme :

$$(3)'\ y = y_0\, e^{\frac{-\gamma x^2}{2a}}\left[1+\frac{a\gamma}{3}\frac{x^3}{a^3}-\frac{a\gamma}{4}\frac{x^4}{a^4}+\frac{a\gamma}{5}\frac{x^5}{a^5}-\left(\frac{a\gamma}{6}-\frac{a^2\gamma^2}{18}\right)\frac{x^6}{a^6}\cdots\right],$$

puis limitant la parenthèse au deuxième terme, écrire y ainsi qu'il suit :

(4) $y = y_0 e^{\frac{-z^2}{2}} \left[1 + \frac{z^3}{3\sqrt{a\gamma}}\right]$, après avoir posé $\frac{\gamma x^2}{a} = z^2$, ou encore

(4)' $y = y_0 e^{\frac{-z^2}{2}} \left[1 + \frac{H z^3}{6}\right]$.

C'est d'ailleurs en partant de la formule (4)' que l'on démontre la *relation simple entre l'obscisse M du maximum (mode) et les abscisses m et μ de la moyenne et de la médiane.*

Il suffit en effet de faire tout d'abord état de la relation

$$\int y dz = y_0 F(z) - \frac{H}{6} y_0 (z^2 + 2) e^{\frac{-z^2}{2}}, \text{ avec } F(z) = \int e^{\frac{-z^2}{2}} dz,$$

puis de la notion de la médiane qui fournit l'équation $\int_{-\infty}^{\mu} y dz = \int_{\mu}^{+\infty} y dz$, et enfin tenir compte ce ce que μ est petit pour aboutir à l'équation $H\mu^2 - 6\mu + 2H = o$, dont on choisit la plus petite racine, qui en première approximation se réduit à $\frac{H}{3}$.

Or la moyenne ayant pour valeur $\frac{q-p}{2\sigma} = \frac{H}{2}$, on voit en définitive que l'on a :

$$m - M = 3(m - \mu).$$

B. — Courbes de Pearson.

1. *Équation différentielle de Pearson.*

L'étude du problème relatif à l'extraction de n boules d'une urne renfermant Np boules blanches et Nq boules noires dans l'hypothèse de remise dans l'urne après chaque tirage de la boule extraite, nous a conduit à l'équation différentielle :

$$(2) \quad \frac{dy}{dx} = \frac{-\gamma xy}{x + a}$$

Si maintenant, l'on fait les séries de n tirages, *sans remettre à chaque fois la boule tirée*, on peut se proposer de trouver l'équation différentielle afférente à la sortie de k boules blanches dans une série. La probabilité en question se calcule facilement en concevant que l'on a pris une poignée de n boules; elle est égale au produit du nombre des combinaisons de Np boules blanches k à k par le nombre des combinaisons de Nq boules noires $(n - k)$ à $(n - k)$, divisé par le nombre des combinaisons de N objets n à n. Grâce à un calcul analogue à celui qui a servi pour l'établissement de l'équation (2), on trouve pour définir l'équation différentielle de la courbe tangente au milieu de chaque côté du polygone l'équation :

(5) $\frac{y'}{y} = \frac{a + X}{B_0 + B_1 X + B_2 X^2}$, qui se ramène au type (6) $\frac{y'}{y} = \frac{x}{b_0 + b_1 x + b_2 x^2}$

Si l'intégration de cette équation (6) met en évidence douze types de courbes

la classification de Pearson laisse de côté un certain nombre des courbes précédentes et ne fait apparaître que les types suivants :

Type I $y = y_o\left(1 - \frac{x}{a_1}\right)^{-va_1}\left(1 - \frac{x}{a_2}\right)^{-va_2}$; $a_1 < 0 < a_2$, $v > 0$, $a_1 < x < a_2$; courbe limitée des deux cotés.

Type II $y = y_o\left(1 - \frac{x^2}{a^2}\right)^{-k}$; courbe limitée des deux côtés et symétrique.

Type III $y = y_o\left(1 + \frac{x}{a}\right)^{va} e^{-vx}$; $a > 0$, $v > 0$, $x > -a$; courbe limitée d'un côté.

Type IV $y = y_o\left(1 + \frac{x^2}{a^2}\right)^{-k} c^{v \operatorname{arc\,tg} \frac{x}{a}}$; $k > 0$; courbe dissymétrique illimitée.

Type V $y = y_o\left(1 - \frac{x}{a}\right)^{-va} e^{\frac{-va}{1 - \frac{x}{a}}}$; $v > 0$, $va > 1$ $x < a$; courbe limitée d'un côté.

Type VI $y = y_o\left(1 - \frac{x}{a_1}\right)^{va_1}\left(\frac{x}{a_2} - 1\right)^{va_2}$ $\begin{cases} a_1 < 0 < a_2, v < 0, a_2 < x, \text{ le produit } va_2 \\ \text{doit être plus grand que } -1, \text{ et le pro-} \\ \text{duit } v(a_2 - a_1) \text{ doit être plus petit que } -1 ; \\ \text{Courbe limitée d'un côté.} \end{cases}$

Type VII $y = y_o e^{\frac{-x^2}{2\sigma^2}}$; courbe illimitée symétrique.

2. *Ajustement des courbes de Pearson.*

Faisant état de l'équation différentielle caractéristique des courbes de dispersion, on établit une série de relations entre les moments m_i ($i = 1, 2, 3, 4, \ldots$) et les coefficients de l'équation différentielle ; ceci étant, on peut à l'aide des moments empiriques (m_i' ou μ_i') déterminer les coefficients de la courbe de dispersion, *après avoir eu soin au préalable de fixer le type de courbe, auquel appartient la distribution.*

Pour résoudre la première partie du problème, il suffit de partir de l'équation différentielle :

$$xy = -y'(b_o + b_1 x + b_2 x^2), \text{ et de former } \int yx^n dx$$

qui se ramène à

$$\int yx^n dx = \int y\left[(n-1)\, b_o x^{n-2} + nb_1 x^{n-1} + (n+1)\, b_2 x^n\right] dx,$$

en vertu de ce que le terme tout intégré est nul, lorsque l'intégration est étendue à l'intervalle total de variation de x.

On a donc en définitive la relation de récurrence :

$$m_n = (n-1)\, b_o m_{n-2} + nb_1 m_{n-1} + (n+1)\, b_2 m_n \text{ (avec } m_o = 1,\ m_1 = m)$$

qui nous donne :

$$m_1 = b_1 + 2\, b_2 m_1,\ m_2 = b_o + 2\, b_1 m_1 + 3\, b_2 m_2,\ m_3 = 2\, b_o m_1 + 3\, b_1 m_2 + 4\, b_2 m_3,$$
$$m_4 = 3\, b_o m_2 + 4\, b_1 m_3 + 5\, b_2 m_4,$$

d'où l'on déduit m_1, m_2, m_3, m_4, en fonction de b_i (avec $i = 0, 1, 2$) et par suite les μ_i.

Ceci étant, pour les courbes des types I et VI, on pose :

$$p_1 = 1 - va_1,\ p_2 = 1 + va_2,\ p_1 + p_2 = s = 2 - \frac{1}{b_2},\ p_1 p_2 = p,$$

et l'on obtient alors — en tenant compte de l'équation de la courbe — les relations suivantes :

$$m = m_1 = \frac{a_1 + a_2}{s}; \mu_2 = \frac{b_o}{1 - 3\,b_2} + \frac{2\,b_1^2}{(1 - 2\,b_2)(1 - 3\,b_2)} - \frac{b_1^2}{(1 - 2\,b_2)^2}$$

$$\mu_2 = -\frac{a_1 a_2}{s+1} + \frac{2\,(a_1 + a_2)^2}{s\,(s+1)} - \frac{(a_1 + a_2)^2}{s^2}, \text{ où encore}$$

$$\mu_2 = \frac{p\,(a_2 - a_1)^2}{s^2\,(s+1)}; \mu_3 = \frac{2\,(a_2 - a_1)^3\,p\,(p_2 - p_1)}{s^3\,(s+1)\,(s+2)}; \mu_4 = \frac{3\,(a_2 - a_1)^4\,p\,[p\,(s-6) + 2\,s^2]}{s^4\,(s+1)\,(s+2)\,(s+3)}.$$

En définitive, on est conduit avec Pearson à considérer les quantités β_1 et β_2 :

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{4\,(s^2 - 4\,p)\,(s+1)}{p\,(s+2)^2}, \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\,(s+1)\,[p\,(s-6) + 2\,s^2]}{p\,(s+2)\,(s+3)},$$

qui sont essentiellement positives, et que l'on détermine en partant des éléments expérimentaux; on tire de ces équations les grandeurs s et p :

$$s = \frac{6\,(\beta_2 - \beta_1 - 1)}{3\,\beta_1 + 2\,\beta_2 + 6}, p = \frac{4\,s^2\,(s+1)}{\beta_1\,(s+2)^2 + 16\,(s+1)},$$

grâce auxquelles on calcule p_1 et p_2, puis a_1 et a_2 et ρ, avec $(a_2 - a_1)^2 = \frac{\mu_2\,s^2\,(s+1)}{p}$, $\rho = \frac{s-2}{a_2 - a_1}$.

On doit tenir compte de ce que $b_2 = o$, pour les courbes du type III, et de ce que $a_1 = a_2$ pour celles du type II.

Pour les courbes du type IV, il faut après un changement d'origine, faire état des équations d'identification :

$$b_o = \frac{\rho^2\,a^2}{k^3} + \frac{a^2}{2\,k}, b_1 = \frac{\rho a}{2\,k^2}, b_2 = \frac{1}{2\,k}.$$

et utiliser le processus de calcul exposé ci-dessus à propos des courbes du type I; pour les courbes du type V, il faut faire appel aux équations d'identification

$$\rho = -\frac{2}{b_1}, a = -\frac{b_1}{2\,b_2} = -\frac{2\,b_o}{b_1},$$

et calculer m_1, m_2, m_3, m_4, μ_2, μ_3, μ_4, en partant des valeurs ci-dessus de ρ et de a.

A la courbe du type VII, il faut rattacher les résultats classiques :

$$b_1 = b_2 = 0,\ m = 0,\ m_2 = b_o = \sigma^2,\ \mu_2 = \sigma^2,\ m_3 = \mu_3 = 0,\ m_4 = \mu_4 = 3\,\sigma^4;$$
$$\beta_1 = 0,\ \beta_2 = 3,\ 3\,\beta_1 - 2\,\beta_2 + 6 = 0.$$

3. *Interprétation graphique.*

Les inégalités ou égalités caractéristiques afférentes aux divers types de courbes (solutions de l'équation (6) peuvent être interprêtées graphiquement d'une manière relativement simple.

Il suffit à cet effet d'utiliser le quadrant de droite d'un système d'axes $\beta_1\,\beta_2$ et de considérer les droites correspondant à $s = \infty$, $s = 2$, $s = 0$, $s = -1$, la cubique et la quartique se rattachant respectivement aux relations

$\beta_1 (s+2)^2 + 16 (s+1) = 0$, $1 - s + p = 0$, (s et p ayant les valeurs ci-dessus indiquées).

On prend comme mesure de la dissymétrie $k = \sqrt{\beta_1}$, en choisissant le signe de façon que k soit positif si la courbe s'allonge vers la droite; pour avoir une idée de la forme de la courbe, on utilise le coefficient $k' = \beta_2$, car si $k' > 3$, la distribution est plus étalée que celle de la courbe normale, et si $k' < 3$, cette distribution est plus tassée.

C. — Explication des suites a répartition normale et a répartition asymétrique d'après Kapetyn (1).

1. *Cas de la répartition symétrique.*

Supposons avec Kapetyn que l'on note le 1er mai dans un jardin le nombre de grains d'un certain diamètre. S'il pleut ce jour-là, les grains observés grandiront tous un peu, mais pas dans la même proportion, du fait que certains buissons sont plus exposés à la pluie que d'autres, et que le sol n'étant pas homogène, certains buissons seront de nouveau favorisés, et enfin que certains grains d'un même buisson profiteront plus de l'action de l'eau que d'autres en raison même de leur structure intérieure.

Kapetyn envisage alors les conditions les plus simples du développement des grains. Soit μ_o le diamètre des grains à leur origine; grâce à la pluie, le diamètre de la moitié des grains se trouve augmenté de $(\mu_1 + \varepsilon)$, grains que nous classerons sous la rubrique A. Le diamètre de l'autre moitié des grains qui appartiendront à la rubrique α n'aura augmenté que de $(\mu_1 - \varepsilon)$.

La moyenne des diamètres des grains est maintenant égale à $\mu_o + \mu_1$, les grains A et α présentant respectivement par rapport à la moyenne les écarts $+ \varepsilon$ et $- \varepsilon$.

S'il fait du soleil le lendemain, le diamètre augmente de nouveau inégalement, car il est des grains plus exposés au soleil que d'autres.

Kapetyn *admet* que l'effet du soleil est le même sur les grains A que sur les grains α, c'est-à-dire qu'il est *indépendant du diamètre des grains;* il suppose en outre que le diamètre de la moitié des grains (constituant le groupe B) s'est accru de $(\mu_2 + \varepsilon)$ et que le diamètre de l'autre moitié des grains (formant le groupe β) n'a augmenté que de $(\mu_2 - \varepsilon)$.

Alors que le système des deux groupes (A, α) se rattache à une division homograde du premier ordre, le système [(A, α) (B, β)] se rattache à une division homograde de second ordre.

Ce dernier système comprend les systèmes partiels :

(I) Grains (A B), sur lesquels les deux causes ont été favorables, et dont le diamètre s'écarte de la moyenne de 2ε;

(II) (III) Grains (A β, B α), sur lesquels l'une des causes a été favorable et l'autre défavorable, et dont le diamètre est égal à la moyenne;

(IV) Grains (α β), sur lesquels les deux causes ont été défavorables, et dont le diamètre s'écarte de $- 2\varepsilon$ sur celui de la moyenne.

---

(1) Voir « Skew frequency curves in biology and statistics ». J. C. Kapetyn, 1903. Groningen; voir aussi les travaux de Wicksell dans les *Annales de l'Observatoire de Lund.*

La probabilité qu'un grain appartienne au système (I) est 1/4, aux systèmes (II) et (III) 1/2, et enfin au système (IV) 1/4.

Les circonstances restant semblables, n causes agissant sur les grains considérés, on obtient une répartition homograde de n^e ordre, le nombre des groupes étant de 2^n.

Si de nouveau, la probabilité est $\frac{1}{2}$ pour qu'une cause quelconque produise un écart $+\varepsilon$ correspondant à un effet favorable, et 1/2 pour qu'elle produise un écart $-\varepsilon$ relatif à un effet défavorable, et si de plus l'effet de chaque cause est *indépendant* de celui des autres, le nombre des groupes sur lesquels i causes ont été favorables et $(n-i)$ défavorables est égal à $\frac{n!}{i!\,(n-i)!}$; l'écart afférent à l'effet global de toutes ces causes a pour valeur $[i-(n-i)]\,\varepsilon = (2i-n)\,\varepsilon$.

La probabilité de l'écart $(2i-n)\,\varepsilon$ a pour valeur :

$$\frac{n!}{i!\,(n-i)!}\cdot\left(\frac{1}{2}\right)^n,$$

expression qui pour n assez grand, correspond à une répartition normale des grains.

Il est évident qu'un tel schéma est rare, car les causes élémentaires n'agiront pas de manière que le diamètre de la moitié des grains augmente d'une même quantité $\mu_1+\varepsilon$, et celui de l'autre moitié d'une quantité $(\mu_1-\varepsilon)$, mais elles auront pour effet de faire augmenter les diamètres différemment entre certaines limites.

Bessel a montré (Voir *Astronomische Nachrichten*, t. 15, p. 369-405) que quelque soit l'effet produit par ces causes, pourvu : 1° qu'il y ait beaucoup de causes agissantes; 2° que l'effet d'une cause soit indépendant de celui des autres; 3° que l'effet de chaque cause soit petit relativement à la somme des effets, la répartition des grains sera normale, alors même que la probabilité p de l'effet favorable ne serait pas égale à la probabilité q de l'effet défavorable.

La justification de la loi de Gauss a été donnée d'une part par M. J. W. Lindeberg (1) et d'autre part par M. Paul Lévy (2).

Rappelons à ce propos que l'on peut donner un sens suffisamment précis au terme « erreur d'observation », en considérant que la grandeur soumise à l'observation a une vraie valeur. L'erreur est alors la différence entre la valeur observée et cette vraie valeur, et l'on admet qu'il existe pour cette erreur une loi de probabilité en faisant l'hypothèse fondamentale suivante revenant à regarder l'erreur considérée ε comme formée d'erreurs élémentaires en nombre très grand, (1) $\varepsilon = \varepsilon_1 + \varepsilon_2 + \ldots\ldots + \varepsilon_n$, les ε_i étant des variables aléatoires ayant chacune une loi de probabilité.

Si après avoir admis pour la relation liant ε aux ε_i la forme linéaire la plus simple, et supposé que les ε_i sont indépendantes et dépourvues d'erreur systématique, on arrive — alors même que les lois des erreurs composantes ε_i sont quelconques — à justifier la loi de Gauss, *à la condition que n soit très grand, et aussi qu'aucune des erreurs ε_i ne joue un rôle prépondérant.*

---

(1) Voir « Eine neue erleitung des exponentialgesetzes in der Wahrscheinlichkeitsrechnung » (*Mathematische Zeitschrift*, Band XV, 1922).

(2) Voir son traité des probabilités et son cours d'Analyse de l'École polytechnique.

La justification subsiste encore en présence d'une erreur ε_i prépondérante, si celle-ci suit la loi de Gauss; la proposition ne subsisterait plus malgré la grande valeur de n, si une erreur prépondérante ε_k s'écartait de la loi de Gauss.

2. *Cas de la répartition asymétrique.* — Les expérimentateurs se sont bien vite aperçus qu'il existe bon nombre de phénomènes naturels dont la répartition est nettement asymétrique; dans de telles répartitions, la probabilité de l'écart $+ \varepsilon$ n'est pas égale à celle de l'écart $- \varepsilon$. Aussi Kapetyn s'est-il demandé quelles pouvaient être les causes amenant une répartition asymétrique d'une certaine régularité; il a considéré à cet effet des grains sphériques de diamètre d dont la fréquence observée est $f(d)$ et il a désigné par F (V) la fréquence du volume V.

Si $f(d)$ représente la répartition normale correspondant aux coefficients du développement $(p + q)^4$, la répartition des volumes ou de $\frac{6V}{\pi}$ est fortement asymétrique.

d	$f(d)$	$\frac{6V}{\pi}$
2	1	8
3	4	27
4	6	64
5	4	125
6	1	216

Il résulte de l'examen de ce tableau que la répartition de surface des grains ainsi que celle de toute fonction non linéaire du diamètre sera asymétrique. Kapetyn ajoute que si les trois conditions de Bessel sont satisfaites relativement aux diamètres, elles ne peuvent l'être relativement aux surfaces; il fait de plus observer que si par exemple, l'effet du soleil est indépendant du diamètre des grains, la probabilité pour que ce diamètre augmente de δ est la même pour tous. Il s'en suit que l'effet du soleil est indépendant de l'effet des causes antérieures. Or, si les grains sont sphériques, l'effet du soleil est sensiblement proportionnel à la racine carrée de leur surface S; il découle de là que l'effet de cette cause dépend donc de celui des causes qui ont agi précédemment.

3. *Remarques.* — D'après ce qui vient d'être dit, on pourra prévoir si une répartition sera asymétrique ou non; tel est le cas de la répartition des revenus dans un pays, puisque les causes qui augmentent les revenus ne sont pas indépendantes des revenus acquis, mais leur sont plutôt proportionnelles. C'est sous l'influence de cette remarque signalée par M. Jordan dans son traité de Statistique mathématique (p. 217) que M. Gibrat a entrepris sa très intéressante étude sur les « Inégalités économiques ». Kapetyn conclut en disant que si l'effet Δx d'une cause est indépendant de la grandeur x, la fréquence des écarts des grandeurs x suivra la loi de Laplace; au cas où cet effet est défini par une expression de la forme $\varepsilon f(x)$, [ε dépendant de la cause seule], la répartition sera asymétrique.

En l'occurrence la grandeur $\frac{\Delta x}{f(x)} = \varepsilon$ est indépendante de x; si donc on la considère comme la variation Δz d'une quantité z, alors l'effet des causes

sur z sera indépendant de x, et la quantité z présentera une répartition normale pourvu que la médiane des grandeurs z soit nulle.

La répartition des grandeurs x étant asymétrique, on peut, grâce à la méthode de Kapetyn et de Van Uven, chercher à déterminer une fonction $z = f(x)$, telle que la répartition des z soit normale. Si $[\omega(x)\,dx]$ désigne la probabilité pour que x_i soit compris entre x et $x + dx$, la relation de correspondance entre x et z sera définie par l'équation $\frac{1}{\sqrt{\pi}} e^{-z^2}\,dz = \omega(x)\,dx$, à laquelle on joint la relation classique $P_v = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{z_v} e^{-t^2}\,dt = \psi(z_v)$, qui est la valeur de la probabilité pour que t soit compris entre $-\infty$ et z_v.

Il résulte de là que si x_v correspond à z_v et si $f(x_0) = \infty$, l'on a :

$$P_v = \psi(z_v) = \int_{x_0}^{x_v} \omega(x)\,dx,$$

$$\text{et } \frac{d\,P_v}{dx^v} = \omega(x_v) = \frac{1}{\sqrt{\pi}} e^{-[f(x_v)]^2} f'(x_v).$$

De cette dernière équation on déduit que la fonction $f(x)$ doit être croissante, puisque $\omega(x)$ est positive; de plus, la médiane afférente aux grandeurs z répondant à $z = o$, il s'en suit que la médiane x_m des grandeurs x est la solution de l'équation $f(x_m) = o$, qui n'a qu'une seule racine réelle entre $-\infty$ et $+\infty$, et la valeur de $\omega(x_m)$ est définie par la relation $\omega(x_m) = \frac{f'(x_m)}{\sqrt{\pi}}$.

La moyenne des grandeurs x est :

$A = \int_{x_0}^{x_M} x\,\omega(x)\,dx$, où x_0 et x_M sont telles que $f(x_0) = -\infty$, et $f(x_M) = +\infty$.

M. Gibrat, dans le travail précité, a analysé des séries statistiques rentrant dans les cas suivants :

$$(7)\ \Delta x = \frac{\Delta z}{k} \text{ et } (8)\ \Delta x = x\frac{\Delta z}{k} \text{ ou encore } (8)'\ \Delta x = \frac{(x - x_0)\,\Delta z}{k}.$$

Or, à l'équation (8)' se rattache la condition de correspondance

$$z = a \log (x - x_0) + b,$$

et la relation

$$(9)\ A'^2 M' = X_m^3, \text{ ou } A' = A - x_0,\ X_m = x_m - x_0,\ M' = M - x_0,$$

et les grandeurs A, x_m, M sont respectivement la moyenne, la médiane et l'abscisse (mode) du maximum.

Lorsque a est relativement grand, on constate en développant en série $\frac{A'}{X_m}$ et $\frac{M'}{X_m}$ que la relation (9) se réduit au quatrième ordre près à la condition approchée,

$2A + M = 3x_m$, qui n'est autre que celle donnée plus avant avec des notations quelque peu différentes.

Dans un travail soumis en 1934 à l'Institut des hautes études de Belgique,

M. A. Della Riccia a fait remarquer que dans les recherches statistiques se rattachant aux domaines respectifs de la biologie et de l'économique, ce n'est point à la courbe (G) de fréquence de Gauss (G) $y_1 = Y_1\, e^{-k'^2}$ qu'il faut recourir, mais à la courbe (K_1).

$$(K_1),\ y_3 = Y_3\, e^{-k'^2 \log \rho},\ \left[\text{avec } \rho = \frac{(x - x_0)(x_2 - x_1)}{(x_2 - x)(x_1 - x_0)}\right]$$

qui généralise la courbe de Kapetyn ayant servi de thème aux études statistiques de M. Gibrat.

Au lieu de considérer l'écart ε entre la valeur x dont on cherche la fréquence et la valeur x_1 la plus fréquente [voir formule (G)], l'auteur fait intervenir [formule (K_1)] le logarithme du rapport harmonique entre la valeur x et les trois valeurs caractéristiques suivantes : le minimum x_0, la valeur la plus fréquente x_1, et la valeur maximum x_2; avec tous les probabilistes, il introduit le paramètre Y_3 ou fréquence de la valeur la plus fréquente et le paramètre d'homogénéité.

Il a été ainsi conduit à examiner non seulement la courbe de fréquence, mais aussi la courbe de sommation (K_2) $z = \int y_3\, dx$, et à utiliser les courbes (K_1) et (K_2) pour l'étude des revenus familiaux; il a été amené à ce propos à construire des tables numériques correspondant aux valeurs :

$$\left(\frac{1}{11}, \frac{2}{10}, \frac{3}{9}, \frac{4}{8}, \frac{5}{7}, \frac{6}{6}\right) \text{de } \frac{p}{q} = \frac{x_1 - x_0}{x_2 - x_1}, \text{ et de } k' = \left(\frac{1}{2}, 1, 2\right)$$

D. — Les séries de Bruns-Charlier. — Généralisation.

1. *Le développement de Bruns.* — La deuxième école de statisticiens part de considérations différentes basées sur l'emploi de deux fonctions dérivant de la fonction binomiale, qui sont respectivement la fonction de Laplace-Gauss :

$$f_1(x) = \frac{1}{\sqrt{2\pi kp(1-p)}}\, e^{\frac{-x^2}{2kpq}},$$

lorsque k tend vers l'infini, et la fonction de Poisson $f_2(x) = \frac{\lambda^x\, e^{-\lambda}}{x!}$, lorsque k tend vers l'infini, p tend vers zéro et kp vers λ; elle emploie f_1 et f_2 comme fonctions génératrices dans la représentation des ensembles statistiques.

On peut avec Bruns chercher à représenter la série statistique par un développement de la forme :

$$(10)\ F(z) = \alpha_0\, \varphi(z) + \alpha_1\, \varphi'(z) + \ldots\ldots\ldots + \alpha_n\, \varphi^{(n)}(z) + \ldots\ldots\ldots,$$

où φ représente la fonction $\frac{1}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}}$. et $\varphi^{n)}(z)$ sa dérivée d'ordre n.

Rappelons d'une part que l'on a $\varphi^{(n)}(z) = \varphi(z)\, P_n(z)$, $P_n(z)$ étant un polynôme de degré n (dit polynôme d'Hermite), qui est lié à $P_{n-1}(z)$ par la relation de récurrence

$$P_n = P'_{n-1} - z\, P_{n-1},$$

et d'autre part que les polynômes $P_0, P_1, \ldots P_n, \ldots\ldots$ forment un système orthogonal, c'est-à-dire un système tel que :

$$I_{m,n} = \int_{-\infty}^{+\infty} \varphi(z).\, P_m(z).\, P_n(z)\, dz = o, \text{ avec } m > n$$

$$\text{et } I_{m,m} = \int_{-\infty}^{+\infty} \varphi(z)\, [P_m(z)]^2\, dz = m!$$

Ceci étant, on détermine les coefficients α_0, α_1, ... α_n, ... du développement (10) par le procédé employé pour la série de Fourier, et l'on trouve, après avoir posé $\frac{x}{\sigma} = z$:

$$\alpha_0 = \frac{N}{\sigma},\ \alpha_1 = + \int_{-\infty}^{+\infty} - \frac{x}{\sigma^2} f(x)\, dx = - \frac{m_1}{\sigma^2} N$$

$$\alpha_2 = \int_{-\infty}^{+\infty} f(x) \left(\frac{x^2}{\sigma^2} - 1\right) \frac{dx}{\sigma} = \frac{1}{2!} \left(\frac{m_2}{\sigma^2} - 1\right) \frac{N}{\sigma},\ \alpha_3 = \frac{1}{3!} \left(\frac{- m_3}{\sigma^3} + \frac{3\, m_1}{\sigma}\right) \frac{N}{\sigma}, ...$$

Si l'origine est fixée à la moyenne arithmétique, et si l'on prend $\sigma^2 = \mu_2$, on aboutit aux résultats suivants :

$$\alpha_0 = \frac{N}{\sigma},\ \alpha_1 = 0,\ \alpha_2 = 0,\ \alpha_3 = - \frac{\mu_3}{3!} \frac{N}{\sigma^4} = - \frac{K\, N}{3!\, \sigma}.$$

$$\alpha_4 = \frac{(\mu_4 - 3\, \mu_2^2)\, N}{4!\, \sigma^5} = \frac{(K' - 3)}{4!\, \sigma} N, ...$$

Il est souvent inutile d'aller plus loin que le quatrième terme, et il est bon de rappeler que les coefficients suivants exigent le calcul de moments d'ordre élevé dont la précision diminue rapidement.

On démontre que sous des conditions très larges imposées à $f(x)$ le développement en séries de polynômes P_n est convergent (1).

2. *Développement de Charlier.* — L'astronome Charlier préconise pour la représentation de séries statistiques un développement de la forme :

$$(11)\ f(x) = \psi(x) + c_1 \psi_1 + c_2 \psi_2 + ..., \text{ avec}$$

$$(12)\ \psi(x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

$$(12')\ \psi_r(x) = - \psi_{r-1}(x) + \psi_{r-1}(x - 1), \text{ et } r = (1, 2, 3, ...)$$

Des relations (12) et (12') on déduit :

$$\psi_r(x) = q_r(x)\, \psi(x), \text{ avec } q_0(x) = 1,$$

et l'on montre que les $q_r(x)$ qui sont des polynômes de degré r jouent un rôle analogue à celui des polynômes d'Hermite dans les séries statistiques basées sur l'emploi de la fonction de Gauss.

En effet, si l'on considère la fonction $F(x) = \psi(x)\, q_h(x)\, q_k(x)$, où apparaissent les multiplicateurs q_h et q_k, on démontre que la somme $\sum_{x=0}^{\infty} F(x)$ est nulle pour $h \neq k$, et qu'elle a pour valeur $\frac{k!}{\lambda^k}$ pour $h = k$.

Les critères de convergence de la série de Charlier ont été établis par M^lle^ Pollaczek-Geiringer en 1928 (voir *Skandinavisk-Aktuarietidskrift*, 1928, p. 98).

Si à la fonction de Poisson, on substitue comme fonction génératrice la fonction binomiale :

$$f(x) = \binom{k}{x} p^x (1 - p)^{k-x} = \frac{k!}{x!\,(k - x)!}\, p^x (1 - p)^{k-x},$$

(1) L'étude de la convergence a été effectuée par M. Kneser (Math. Annalen, t. 58) et par M^me^ Myller-Lebedeff (Math. Annalen, t. 64); elle a été reprise par MM. Galbrun et Berger (voir Galbrun, C. R. Acad. des Sciences, 1909, et *Bulletin de la Société mathématique de France*, 1912).

on peut donner à la série statistique la forme :

$$F(x) = f(x) + c_1 \Delta f(x) + c_2 \Delta^2 f(x) + \ldots\ldots + c_k \Delta^k f(x),$$

avec

$$\Delta f(x) = f(x-1) - f(x) = f_1(x),\ \Delta^2 f(x) = \Delta_1 f_1(x) = f_2(x),$$

on trouve que

$$\Delta^r f(x) = q_r(x) f(x),$$

et l'on remarque que q_r qui est un polynôme de degré r est lié à q_{r-1} par la loi de récurrence.

$$q_r(x) = -q_{r-1}(x) + \frac{1-p}{p} \cdot \frac{x}{k+1-x} q_{r-1}(x-1).$$

3. *Généralisation.* — Romanovsky a appliqué le processus analytique exposé précédemment, en introduisant des courbes de fréquence autres que la courbe normale; c'est ainsi qu'il a utilisé la fonction :

(13) $\varphi(x) = \frac{1}{\lambda}\left(1 + \frac{x}{a}\right)^{\nu a} e^{-\nu x}$, ou encore la fonction $\psi(z)$ qui en résulte par le changement de variable $1 + \frac{x}{a} = \frac{z}{\nu a} = \frac{z}{p}$, et par la détermination de λ au moyen de la relation $\int_{-a}^{+\infty} \varphi(x)\, dx = \frac{e^p}{p^p}$.

Ceci étant, il fait intervenir les polynômes :

$$P_k(z) = \frac{1}{z^p e^{-z}} \frac{d^k}{dz^k}\left[z^{p+k} e^{-z}\right] = \frac{1}{\psi(z)} \frac{d^k}{dz^k}\left[z^k \psi(z)\right]$$

qui jouissent des propriétés suivantes :

$$\int_0^\infty \psi(z)\, P_m(z)\, P_n(z)\, dz = 0, \text{ avec } m \neq n,$$

$$\int_0^\infty \psi(z)\, (P_n)^2\, dz = \frac{n!\,\Gamma(p+n+1)}{\frac{a}{p}\,\Gamma(p+1} \text{ avec } m = n;$$

il suppose ensuite que l'on peut figurer la série statistique $F(z)$ au moyen du développement en série ci-dessous :

$$F(z) = \psi(z)\left[\alpha_0 P_0(z) + \alpha_1 P_1(z) + \ldots\ldots + \alpha_n P_n(z) + \ldots\ldots\right],$$

dont on calcule les coefficients α_n à l'aide de la méthode classique :

$$\int_0^\infty F(z)\, P_0(z)\, dz = \frac{p}{a}\alpha_0,\ \int_0^\infty F(z)\, P_n(z)\, dz = n!\,(p+1)(p+2)\ldots(p+n) \cdot \frac{p}{a} \cdot \alpha_n.$$

Si l'on détermine l'origine et les coefficients de la courbe (13) au moyen des valeurs de m_1, m_2, m_3, on remarque alors que :

$\alpha_1 = \alpha_2 = \alpha_3 = 0$, et l'on trouve pour la valeur de $\alpha_4 = \frac{N}{12} \frac{[2\beta_2 - 3\beta_1 - 6]\beta_1^2}{(4+\beta_1)(4+2\beta_1)(4+3\beta_1)}$.

Romanovsky a également étudié des développements en série basés sur l'emploi de la fonction $u_0 = (a_1 + x)^{m_1}\, 1\, (a_2 - x)^{m_2}$.

CHAPITRE III

Choix des grandeurs a priori. — Schémas de Bernoulli, Poisson, Lexis, Borel, Polya.

Supposons que l'on ait effectué N extractions d'une urne renfermant en proportion déterminée p mais inconnue, des boules blanches et noires, — soit en remettant les boules extraites, soit en ne les remettant pas; on a constaté la sortie de n boules blanches. *Peut-on de ce résultat déduire une indication précieuse sur la valeur de la probabilité inconnue.*

Le problème ainsi posé n'est autre que celui que s'est posé Bernoulli, et qu'il a résolu en faisant appel à la loi des grands nombres qu'il a énoncée et démontrée.

Nous pouvons dire que si N est suffisamment grand, $\frac{n}{N}$ est une valeur approchée de p avec une probabilité voisine de 1, et avec Tschuprow nous la désignerons *valeur présumée.*

Quel que soit le mode d'extraction, l'espérance mathématique est la même $E\left(\frac{n}{N}\right) = p$; en ce qui concerne l'écart type il est égal à $\frac{p\,(1-p)}{N}$, si l'on remet les boules extraites dans l'urne.

$$\text{et à } \frac{A-N}{A-1}\ \frac{p\,(1-p)}{N},$$

dans le cas contraire (A désignant le nombre total des boules de l'urne).

A. — Séries homogrades.

Considérons par exemple les naissances masculines m_i rapportées à un nombre fixe S de naissances des deux sexes, dans N régions d'un territoire à une époque donnée; nous pouvons alors former une série de quotients $\frac{m_i}{S}$ ($i = 1, 2, 3, \ldots\ldots$ N), qui peuvent être assimilés aux résultats de N *épreuves* comportant chacune S tirages d'une urne renfermant des boules blanches et noires en proportion p *déterminée mais inconnue.*

Posons $\frac{m_i}{S} = p'_i$, et remarquons que la moyenne :

$$\Sigma\,\frac{p'_i}{N} = \frac{\Sigma\, m_i}{SN}$$

peut être prise comme valeur présumée de la probabilité; p'_i est une valeur approchée à laquelle on rattache la dispersion $\sqrt{\frac{pq}{S}}$, et la moyenne p'_0 est une valeur plus approchée avec la dispersion $\sqrt{\frac{pq}{S}}\,\frac{1}{\sqrt{N}}$.

Si nous admettons que les *épreuves* sont de probabilités indépendantes, on peut calculer $E\,(p'_i - p'_0)^2$, et l'on trouve que l'espérance mathématique de $(p'_i - p'_0)^2$ a pour valeur $\frac{pq}{S}\left(1 - \frac{1}{N}\right)$.

Or l'espérance mathématique de σ'^2 ou $E(\sigma'^2) = E\frac{\Sigma(p'_i - p'_o)^2}{N-1}$ est exactement égale à $\frac{pq}{S}$, ou encore à la valeur probable de la fluctuation autour de p, que l'on désigne par σ_B^2.

$$E(\sigma'^2) = \frac{pq}{S} = \sigma_B^2, \text{ (avec } q = 1 - p)$$

La valeur présumée de $\frac{pq}{S}$, pouvant être calculée en formant, soit : la quantité $\frac{p'_o q'_o}{S}\left(1 - \frac{1}{N}\right)$ dite *valeur calculée*, ou approximativement $\frac{p'_o q'_o}{S}$, soit la quantité σ'^2 dite VALEUR MESURÉE, on est conduit à former la quantité Q définie par l'expression :

$$Q = \sqrt{\frac{\frac{\Sigma(p'_i - p'_o)^2}{N-1}}{\frac{p'_o q'_o}{s}\left(1 - \frac{1}{N}\right)}}$$

Si Q = 1, l'on dit que la série envisagée est à dispersion normale; si de plus le groupement des fréquences expérimentales autour de leur moyenne peut être reproduit au moyen du graphique binomial, on dit alors que les observations sont synthétisées par le *schéma de Bernoulli* qui correspond aux tirages de boules blanches d'une urne de composition constante.

1. *Coefficients de dispersion et de divergence.* — A côté du coefficient Q de dispersion, introduit en statistique en 1877 par Lexis (1), il y a lieu de signaler un coefficient analogue dit coefficient de divergence basé sur l'emploi de l'écart moyen et utilisé dès 1875 par Dormoy.

Alors que primitivement nous faisions les s tirages dans une seule urne, et que nous répétions N fois notre *épreuve*, nous ferons ces tirages dans des urnes U_1, U_2, ... U_s; les N épreuves nous fourniront les résultats m_1, m_2, ... m_N, et feront apparaître le schéma de Poisson.

Les probabilités moyennes p_o et q_o étant prises égales à

$$p_o = \frac{p_1 + p_2 \cdots\cdots + p_s}{s}, \quad q_o = \frac{q_1 + q_2 \cdots\cdots + q_s}{s},$$

on démontre que la dispersion σ_P de la théorie de Poisson est liée à celle σ_B de la série de Bernoulli par la relation.

$$\sigma_P^2 = \sigma_B^2 - \sum_{i=1}^{s} (p_i - p_o)^2 \cdot \frac{1}{s^2}, \text{ et l'on voit ainsi que } Q < 1.$$

Or l'examen des documents statistiques a montré que ce n'est que dans des cas peu nombreux que l'on trouvait pour Q une valeur inférieure à 1, et que l'on pouvait recourir au schéma de Poisson; il fallait donc trouver un schéma nouveau pour expliquer les valeurs très supérieures à 1 et atteignant même 50 du coefficient de dispersion.

(1) LEXIS, « Zur Théorie der Massenerscheinungen in der menschlichen Gesellschaft, 1877. Dormoy, *Journal des Actuaires français*, 1875, *Théorie mathématique des assurances sur la vie*, 1878.

2. *Schéma de Lexis.* — Ce problème a été résolu par Lexis, en supposant que *la première épreuve, ou série de s tirages* était faite dans l'urne U_1 de composition (p_1, q_1), la deuxième épreuve dans l'urne U_2 (p_2, q_2), et ainsi de suite en admettant que les compositions $(p_i\ q_i)$ n'étaient pas très différentes les unes des autres.

Dans l'ignorance où nous sommes des opérations, nous substituons aux urnes U_i $(i = 1, 2, \ldots\ldots N)$ une urne unique de composition $p'_o = \frac{\Sigma\, p'_i}{N}$ où p'_o est une valeur approchée de la moyenne :

$$p_o = \frac{\Sigma\, p_i}{N};$$

Ceci étant, considérons d'une part l'écart σ_P défini par $\sigma_P^2 = \sum_{i=1}^{N} \frac{(p_i - p_o)^2}{N}$ qui nous indiquera si les N urnes ont des compositions voisines ou très dispersées, et d'autre part l'écart type σ_B correspondant à l'hypothèse d'un schéma de Bernoulli avec les probabilités (p_o, q_o), puis faisons intervenir l'écart caractéristique σ_L de la série de Lexis.

$$\sigma_L^2 = E\, \frac{\Sigma\, (p'_i - p_o)^2}{N};$$

Après avoir remarqué que :

$$\sigma_L^2 = \sigma_B^2 + \left(1 - \frac{1}{s}\right) \sigma_P^2,$$

on voit que :

$$Q^2 = \frac{\sigma_L^2}{\sigma_B^2} = 1 + \left(1 - \frac{1}{s}\right) \frac{\sigma_P^2}{\sigma_B^2}.$$

Le schéma de Lexis fait donc apparaître des *séries hypernormales*, et met en lumière les valeurs de Q d'autant plus élevées que la composition des N urnes les unes par rapport aux autres est plus dispersée.

3. *Schéma de Borel.* — M. Borel a réalisé un schéma fort simple afférent aux valeurs de $Q > 1$ en introduisant le tirage *par grappes*, qui correspond au cas où les boules blanches sont — comme les boules noires — attachées par ensemble de $\mathcal{N}$; ce schéma permet de réaliser un coefficient de divergence quelconque.

4. *Schéma des tirages contagieux.* — On peut concevoir des tirages où la boule tirée est remise immédiatement dans l'urne d'où elle est extraite, et des tirages où la boule au contraire n'est point remise; le cas le plus simple que nous allons d'ailleurs considérer ici correspond à une urne unique.

Plaçons-nous dans l'hypothèse de non *remise* dans l'urne de la boule extraite, et soit A le nombre total des boules blanches et noires de l'urne dont la composition avant tout tirage est définie par les éléments p, q; on fait N tirages successifs en ne remettant pas les boules.

La probabilité de tirer x boules blanches est définie par l'expression :

$$P_x = \frac{N\,(N-1)\ldots(N-x+1)\,B\,(B-1)\ldots(B-x+1)\,C\,(C-1)\ldots(C-N+x+1)}{1, 2, 3, \ldots\ldots x\,A\,(A-1)\ldots\ldots(A-N+1)},$$

avec $B = Ap$, $C = Aq$;

quant à la valeur probable de x ou E (x), elle est égale à (Np), c'est-à-dire identique à celle que l'on trouve dans le cas où l'on remet dans l'urne la boule tirée après chaque tirage.

La loi de répartition est voisine d'une loi normale dont l'écart moyen quadratique serait $\sqrt{\frac{N(A-N)pq}{A}}$; l'écart quadratique est en réalité égal à $\sqrt{\frac{N(A-N)pq}{A-1}}$. On remarque donc que ce genre de tirage conduit à une diminution de la dispersion type, et l'on est alors en présence d'un schéma contagieux *sous-dispersé.*

Si l'on considère maintenant (1) une urne de composition initiale (p, q) renfermant Ap boules blanches et Aq boules noires, et supposons qu'après tirage d'une boule d'une couleur l'on remette $(1 + A\gamma)$ boules de la même couleur, on trouve que la probabilité de tirer b boules blanches et n boules noires en N épreuves a pour valeur :

$$P_b = \frac{n!\,b!}{N!}\,\frac{p(p+\gamma)\ldots\ldots[p+(b-1)\gamma]\,q(q+\gamma)\ldots\ldots[q+(n-1)\gamma]}{(1+\gamma)(1+2\gamma)\ldots[1+(b-1)\gamma](1+b\gamma)\ldots[1+(n+b-1)\gamma]},$$

avec $b + n = N$.

Faisant tendre N vers l'infini, et $N\gamma$ vers une constante $\mathcal{C}$, Polya et Eggenberger ont trouvé que μ_2 restait de l'ordre de $\sqrt{N}$, que la loi limite de répartition qui était encore une loi de Laplace, faisait apparaître une *surdispersion* du fait de la présence du coefficient $(1 + \mathcal{C})$.

Suivant les hypothèses que l'on peut faire sur les valeurs Np, Np^2, $N\gamma$, $N\gamma^2$, l'on est amené à la considération de six formes de distribution, et l'on peut expliquer non seulement la loi des grands nombres et la loi des petits nombres, mais encore avec une faible contagion soit les événements non rares, soit les événements rares, avec une forte contagion les événements non rares, et enfin avec une contagion « presque faible » les événements désignés par les auteurs sous la rubrique « presque rares ».

B. — Les séries hétérogrades.

Peut-on caractériser la distribution des tailles d'un million d'hommes de vingt et un ans qui rentre dans les séries hétérogrades, en procédant à la mesure des tailles de 100.000 d'entre eux, ou encore peut-on donner une limite des erreurs dans l'évaluation des constantes caractéristiques de la série.

Si l'on désigne par :

$$m_o, m_1, m_2 \ldots\ldots \mu_o, \mu_1, \mu_2, \ldots\ldots$$

les moments de la population totale,

$$m'_o, m'_1, m'_2 \ldots\ldots \mu'_o, \mu'_1, \mu'_2 \ldots\ldots$$

les moments de la population soumise à l'épreuve statistique, nous voyons, en supposant que les N mesures sont indépendantes et en recourant à des calculs

---

(1) Polya et Eggenberger, *C. R. Académie des Sciences*, Paris, 12 novembre 1928, Zeitschrift fur angewandte Mathematik und Mechanik 3/1923, p. 279-289, Eggenberger. Thèse, Zurich, 1924.

relativement simples, que l'espérance mathématique de $(m'_1 - m_1)^2$ a pour valeur $E(m'_1 - m_1)^2 = \frac{m_2 - m_1^2}{N}$, si l'épreuve comporte la mesure de N individus, valeur que l'on peut écrire encore :

$$\sigma^2_{m'_1} = \frac{\mu_2}{N}.$$

On trouve de même la relation $\sigma^2_{m'_p} = \frac{m_{2p} - m_p^2}{N}$ qui nous montre que les fluctuations d'un moment d'ordre p sont fonction du moment d'ordre double.

En réalité les moments μ'_p de la distribution étant rapportés à la moyenne arithmétique, il est nécessaire de procéder à la détermination des écarts types de $\mu'_2, \mu'_3, \ldots \mu'_k$; les calculs sont un peu plus longs que ceux qui permettent d'établir les résultats précédemment donnés, et montrent par exemple qu'en première approximation la valeur de :

$$E[\mu'_2 - E(\mu'_2)]^2 \text{ n'est autre que } \frac{\mu_4 - \mu_2^2}{N}.$$

CHAPITRE IV

Retour sur les moments. — Critères attachés a certaines fonctions de fréquence.

M. Guldberg, dans des travaux récents, a montré que l'on pouvait définir les critères de divers types de fonction de fréquence, en faisant appel aux relations de récurrence des dites fonctions, alors que Pearson et son école interprétaient une certaine équation différentielle; M. Guldberg revenait à l'étude des propriétés des fonctions de fréquence et procédait à la recherche d'invariants caractéristiques de ces fonctions.

Après avoir défini les moments autour de la moyenne arithmétique :

$$m_r = \sum_1^n \frac{(x_i - m_1)^r}{n},$$

et les moments par rapport à l'origine :

$$\sigma_2 = \frac{\Sigma x'_i}{n},$$

il introduit les semi-invariants de Thiele en recourant à l'identité :

$$e^{\mu_1 t + \frac{\mu_2 t^2}{2!} + \frac{\mu_3 t^3}{3!} + \ldots} = 1 + \frac{\sigma_1 t}{1} + \frac{\sigma_2 t^2}{2!} + \ldots$$

qui fournit la relation de récurrence.

$$\sigma_{r-1} = \mu_1 \sigma_r + \mu_2 \sigma_{r-1} \binom{r}{1} + \mu_3 \sigma_{r-2} \binom{r}{2} + \ldots + \mu_{r-1},$$

expression où le symbole $\binom{r}{p}$ représente le nombre des combinaisons de r objets p à p, (soit C_r^p).

Aux moments σ_r, on peut substituer les moments factoriels $\sigma_{(r)}$ introduits par Steffensen et Sheppard :

$$\sigma_{(r)} = \Sigma\, x_i\,(x_i - 1)\,\ldots\,(x_i - r + 1),$$

qui peuvent d'ailleurs comme les semi-invariants de Thiele, être représentés en recourant à l'identité :

$$\sigma_0 + \frac{\sigma_{(1)}t}{1} + \frac{\sigma_{(2)}t^2}{2!} + \ldots\ldots = \frac{1}{n}\,\Sigma\,(1+t)^{X_i}, \text{ avec } |t| < 1.$$

On remarque enfin que l'on peut utiliser aussi bien les moments, les semi-invariants que les moments factoriels.

A) *La fonction caractéristique associée aux fonctions de fréquence binomiale et de Poisson.*

La fonction caractéristique s'écrit, si l'on emploie l'exponentielle e^{itx}, comme le fait M. P. Lévy :

$$\varphi\,(t) = \Sigma\, e^{itx}\, f\,(x);$$

il s'en suit que si l'on remplace f (X) par la fonction de fréquence binomiale, on obtient :

$$\varphi\,(t) = \sum_{o}^{k} \binom{K}{X} p^X\,(1-p)^{K-X}\, e^{itx} = [pe^{it} + (1-p)]^k$$

et l'on peut déterminer les σ_i grâce à cette dernière équation et à la relation :

$$\varphi^p\,(o) = i^p \sigma_p$$

On peut aussi revenir aux semi-invariants, en remplaçant t par it et en introduisant $\psi\,(t) = \log \varphi\,(t) = \mu_1\, it + \mu_2 \dfrac{(it)^2}{2!} + \ldots\ldots$; on en déduit :

$$\mu_r = \frac{\psi^{(r)}\,(o)}{i^r}$$

Dans le cas de la fonction binomiale, l'on trouve :

$$\mu_1 = kp;\ \mu_2 = kp\,(1-p),\ \mu_3 = kp\,(1-p)\,(1-2p),\ \ldots\ldots$$

et l'on déduit par l'élimination de p et de k entre les $(k+1)$ premiers invariants toute une série de relations dont les deux premières sont :

$$\frac{\mu_3\,\mu_1}{2\,\mu_2^2 - \mu_1 \cdot \mu_2} = 1,\ \frac{\mu_4\,\mu_1}{\mu_1^2\,\mu_2 - 6\mu_1\,\mu_2^2 + 6\mu_2^3} = 1$$

Il résulte de là qu'une série statistique dont les semi-invariants satisfont sensiblement aux relations précédentes, peut être représentée d'une manière approchée par la fonction de fréquence binomiale.

La fonction de fréquence de Poisson $\dfrac{\lambda^X e^{-\lambda}}{X!}$.

Un calcul analogue à celui qui a été esquissé ci-dessus montre que les semi-invariants de la fonction de Poisson sont tous égaux à λ. Cette propriété capi-

tale permet de s'assurer si une série statistique est susceptible d'être représentée par la fonction de Poisson.

Moments factoriels. — Suivant que l'on fait état de la fonction binomiale ou de la fonction de Poisson, les moments factoriels $\sigma_{(r)}$ ont respectivement pour valeur :

$$[k\,(k-1)\,.\,.\,.\,.\,k-(r-1)]\,p^r, \text{ ou } \lambda^r$$

B) *D'un nouvel aspect des fonctions de fréquence.*

M. Guldberg s'est demandé si l'on peut caractériser la stabilité par un procédé différent de celui signalé par Dormoy et Lexis; pour lui, la théorie des fonctions de fréquence comprend quatre problèmes distincts :

1° Le calcul numérique de la fonction;

2° Le processus analytique à employer en vue de substituer à la fonction discontinue une fonction continue, qui pour des valeurs entières de la variable prend les mêmes valeurs que la fonction discontinue;

3° La détermination des moments d'une fonction de fréquence;

4° Une série statistique étant donnée, chercher une fonction de fréquence donnant une représentation approchée de la dite série, et établir — si possible — les critères nécessaires et suffisants pour qu'une fonction déterminée de fréquence remplisse les conditions requises.

C'est Charlier qui le premier a résolu le problème (2°) pour la fonction de Poisson, en introduisant la fonction $\psi(x) = \frac{e^{-\lambda}}{\pi}\int_0^\infty e^{\lambda\cos t}\cos[\lambda\sin t - xt]\,dt$, et a donné le développement :

$$\psi(x) = \frac{e^{-\lambda}\sin\pi x}{\pi}\left[\frac{1}{x} - \frac{\lambda}{1!\,(x-1)} + \frac{\lambda^2}{2!\,(x-2)} + \ldots\ldots + (-1)^n\frac{\lambda^n}{n!\,(x-n)} + \ldots\right];$$

On sait de plus d'après Jorgensen que $\psi(x)$ satisfait à l'équation aux différences finies :

$$\psi(x+1) = \frac{\lambda}{x+1}\psi(x) - e^{-\lambda}\frac{\sin\pi x}{\pi}.$$

Grâce à l'équation aux différences finies afférente à la fonction de Poisson $f(x+1) = \frac{\lambda}{x+1}f(x)$· il est facile de calculer les moments :

$\sigma_r = \sum_0^\infty x^r f(x)$, et l'on trouve la relation de récurrence fort intéressante.

$\sigma_{n+1} = \lambda(\sigma_n + C_n^1\sigma_{n-1} + C_n^2\sigma_{n-2}\ldots\ldots + 1)$, qui peut s'écrire symboliquement $\sigma_{n+1} = \lambda(\sigma+1)^{(n)}$, en ayant soit de remplacer σ^i par σ_i.

1. *Les moments incomplets de M. Frisch.*

On désigne par moments incomplets ${}_t\mu_r$ d'ordre r l'expression :

${}_t\mu_r = \sum_{x=t}^\infty (x-\lambda)^r\,f(x)$, et par moment incomplet autour de l'origine ${}_t\sigma_r$

${}_t\sigma_r = \sum_{x=t}^\infty x^r\,f(x)$.

A la fonction de fréquence de Poisson et à l'équation aux différences finies correspondante, il y a lieu de rattacher l'équation liant ${}_t\sigma_{n+1}$ aux ${}_t\sigma_j$ pour $j = [n, (n-1) \ldots\ldots o]$.

$$ {}_t\sigma_{n+1} = \lambda\,[{}_t\sigma_n + C_n^1 \cdot {}_t\sigma_{n-1} + C_n^2 \cdot {}_t\sigma_{n-2} + \ldots\ldots + C_n^n \cdot {}_t\sigma_o] + t^{n+1} f(t) $$

2. *Caractéristique de la fonction de Poisson.*

L'équation aux différences finies s'écrivant :

$$ (10) \qquad \frac{f(x+1)}{f(x)} \cdot (x+1) = \lambda, $$

il s'en suit que le premier membre de (10) est constamment égal au semi-invariant.

Il résulte de là que si pour une série statistique $(x, \mathcal{H}(x))$, *on calcule la moyenne:*

$$ \mu_1 = \frac{\sum_o^\infty x\ \mathcal{H}(x)}{\sum_o^\infty \mathcal{H}(x)} $$

que l'on désigne par λ, *et si d'autre part, pour toutes les valeurs de* x $(0, 1, 2, \ldots\ldots)$ *on constate que l'expression* $\frac{\mathcal{H}(x+1)}{\mathcal{H}(x)} \cdot (x+1)$ *est sensiblement égale à* λ, *on doit en conclure que la série à l'étude peut être représentée par la formule de Poisson.*

3. *Fonction binomiale. — Sa caractéristique.*

Pour la fonction binomiale $f(x) = C_k^x\, p^x\, (1-p)^{k-x}$, (avec $o \leqslant x \leqslant k$), on peut calculer facilement les $\sigma_r = \sum_o^k x^r f(x)$ au moyen de la formule de récurrence :

$$ (1-p)\,\sigma_{n+1} = pk\,(\sigma+1)^{(n)} + p \sum_{j=1}^{n} C_n^j\, \sigma_{n-j+1}, $$

et l'on trouve :

$$ \sigma_1 = pk, \quad \sigma_2 = pk\,(pk+1-p); $$

quant aux moments incomplets ${}_t\sigma_n$ autour de l'origine et aux moments incomplets :

$$ {}_t\mu_n = \sum_{x=t}^{k} (x - kp)^n f(x) $$

comptés à partir de la valeur la plus probable kp, ils peuvent être évalués en partant de l'équation aux différences finies :

$$ (14) \qquad f(x+1) = \frac{p}{1-p} \cdot \frac{k-x}{x+1} \cdot f(x), $$

comme l'a montré M. Frisch.

La comparaison des semi-invariants μ_1 et μ_2 qui ont respectivement pour valeur :

$$ \mu_1 = kp, \quad \mu_2 = kp\,(1-p), $$

montre que $\mu_2 < \mu_1$; de plus l'examen de l'équation(14)montre que la fonction

$$ (15)\ \psi(x) = \frac{f(x+1)\,(x+1)}{f(x)} + \frac{px}{1-p}, $$

ou encore $\left\{ \frac{f(x+1)}{f(x)}(x+1) + \frac{\mu_1 - \mu_2}{\mu_2} x \right\}$ a la valeur constante $\frac{pk}{1-p} = \frac{\mu_1^2}{\mu_2}$ si la loi de fréquence est du type binomial.

Ceci étant, il est possible de déterminer μ_1 et μ_2 au moyen de l'équation :

$$(16) \qquad \frac{f(x+1)(x+1)}{f(x)} + \frac{\mu_1 - \mu_2}{\mu_2} x = \frac{\mu_1^2}{\mu_2},$$

en recourant aux valeurs de $f(x)$, $f(x+1)$, $f(x+2)$, et l'on peut affirmer que $\psi(x)$ se réduit à une constante, si la loi de fréquence est la loi binomiale.

En définitive, on est conduit lorsque l'on se trouve en présence d'une série statistique $[x, \mathcal{H}(x)]$, à *calculer :*

$$\mu_1 = \frac{\sum_o^k x\, \mathcal{H}(x)}{\sum \mathcal{H}(x)}, \quad \mu_2 = \sum_o^k \frac{(x-\mu_1)^2\, \mathcal{H}(x)}{\sum \mathcal{H}(x)},$$

et à former :

$$\frac{\mathcal{H}(x+1)}{\mathcal{H}(x)}(x+1) + \frac{\mu_1 - \mu_2}{\mu_2} x = \psi(x);$$

Si pour les valeurs $(x = 0, 1, 2, \ldots\ldots, k-1)$, ψ *reste sensiblement constant, on peut dire que la série peut être représentée d'une manière approchée par la fonction* $\binom{k}{x} p^x (1-p)^{k-x}$.

4. *Fonction de fréquence de Pascal. — Sa caractéristique.*

A cette fonction qui est définie par la relation :

$f(x) = \binom{(k+x)}{k} p^k q^x$, on rattache les moments $\sigma_r = \sum_{x=o}^{\infty} x^r f(x)$, qui obéissent à la loi de récurrence.

$$(1-q)\,\sigma_{n+1} = q(k+1)(\sigma+1)^{(n)} + q \sum_{j=1}^{n} C_n^j\, \sigma_{n-j+1}, \text{ et les semi-invariants}$$

$$\mu_1 = \frac{q}{1-q}(k+1), \quad \mu_2 = \frac{q(k+1)}{(1-q)^2} > \mu_1.$$

Si de ces deux dernières relations, l'on tire les valeurs de $(k+1)$ et de q on remarque que l'équation aux différences finies caractéristiques de la fonction de Pascal se met sous la forme :

$$\frac{f(x+1)}{f(x)}(x+1) + \frac{\mu_1 - \mu_2}{\mu_2} x = \frac{\mu_1^2}{\mu_2}$$

Aux trois fonctions de fréquence qui viennent d'être examinées, correspond, une même forme d'équation aux différences finies avec les critères :

$\mu_1 > \mu_2$ pour la première (binomiale); $\mu_1 = \mu_2$ pour la seconde (Poisson); $\mu_1 < \mu_2$ pour la troisième (Pascal.)

L'expression $\psi(x) = \frac{f(x+1)}{f(x)}(x+1) + \frac{\mu_1 - \mu_2}{\mu_2} x$ est constante et égale à $\frac{\mu_1^2}{\mu_2}$.

5. *Fonction de fréquence hypergéométrique.*

M. Guldberg dans ses conférences à l'Institut Henri-Poincaré en juin 1932, a également étudié la fonction :

$$f(x) = \frac{\binom{k}{x}\binom{h}{m-x}}{\binom{k+h}{m}} =$$

$$\frac{k!}{x!\,(k-x)!}\,\frac{h!}{(m-x)!\,(h-m+x)!}\,\frac{m!\,(k+h-m)!}{(k+h)!},$$

que l'on appelle la fonction de fréquence hypergéométrique, et qui n'est autre que la valeur de la probabilité d'extraire dans un tirage de m boules x boules blanches d'une urne renfermant k boules blanches et h boules noires, dans l'hypothèse où l'on ne remet pas les boules extraites; cette fonction satisfait à l'équation aux différences finies :

$$f(x+1) = \frac{(k-x)(m-x)}{(x+1)(h-m+1+x)} \cdot f(x), \text{ où } m \geqslant x \geqslant 0,$$

que l'on peut encore écrire :

$$(17)\quad \frac{f(x+1)}{f(x)} \cdot (x+1)(h-m) + \frac{f(x+1)}{f(x)} \cdot (x+1)^2 + (k+m)\,x - x^2 = km$$

Ceci étant, on détermine les grandeurs $(h-m)$, $(k+m)$, km au moyen des moments factoriels :

$$\sigma_{(1)} = \sigma_1,\ \sigma_{(2)} = \sigma_2 - \sigma_1,\ \sigma_{(3)} = \sigma_3 - 3\,\sigma_2 + 2\,\sigma_1$$

$$\sigma_{(1)} = \frac{mk}{h+k},\ \sigma_{(2)} = \frac{mk\,(m-1)\,(k-1)}{(h+k)\,(h+k-1)},\ \sigma_{(3)} = \frac{mk\,(m-1)\,(k-1)\,(m-2)\,(k-2)}{(h+k)\,(h+k-1)\,(h+k-2)};$$

on peut aussi calculer les trois mêmes grandeurs, en recourant soit aux moments ordinaires σ_i, soit aux semi-invariants μ_i de Thiele.

Dans le cas particulier $h = k$ correspondant à la fonction hypergéométrique symétrique, on a :

$$\mu_1 = \frac{m}{2},\ \mu_2 = \frac{m}{4} \cdot \frac{2k-m}{2k-1},$$

et l'équation aux différences finies devient :

$$(18)\quad \frac{f(x+1)}{f(x)}\,(x+1)\left[\frac{\mu_1 - \mu_1^2 + 4\,\mu_1\,\mu_2 - 3\,\mu_2}{\mu_1 - 2\,\mu_2} + x\right]$$

$$+ \frac{(3\,\mu_1^2 - 4\,\mu_1\,\mu_2 - \mu_2)}{\mu_1 - 2\,\mu_2}\,x - x^2 = \frac{2\,\mu_1^3 - 2\,\mu_1\,\mu_2}{\mu_1 - 2\,\mu_2},$$

$$\text{ou } \psi(x) = \frac{2\,\mu_1^3 - 2\,\mu_1\,\mu_2}{\mu_1 - 2\,\mu_2};$$

si le quotient de $\psi(x)$ par le second membre de (18) pour des valeurs entières et successives de x est voisin de 1, on peut affirmer que la loi de distribution peut être représentée avec approximation suffisante par la fonction hypergéométrique symétrique.

R. Risser.