JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

MAURICE DUMAS

Le groupage des observations et les corrections qu'il nécessite dans le calcul des moments

Journal de la société statistique de Paris, tome 88 (1947), p. 175-189

<http://www.numdam.org/item?id=JSFS_1947__88__175_0>

IV

LE GROUPAGE DES OBSERVATIONS ET LES CORRECTIONS QU'IL NÉCESSITE DANS LE CALCUL DES MOMENTS

1. Certaines grandeurs sont par nature continues, mais les mesures que l'on en fait ne peuvent être que discrètes : même si des mesures de longueur sont obtenues de façon relativement précise et même si cette précision fait qu'il est raisonnable de les exprimer par un nombre comportant des centièmes de millimètre, la série statistique constituée par ces mesures a exactement la même allure que celle qui correspondrait à une grandeur de nature discrète, comme le nombre de boules d'une urne. Ainsi, toute série statistique sur laquelle on peut être amené à raisonner comporte ou peut comporter des répétitions de certaines mesures; toute série peut donc être considérée comme régulièrement (1) répartie en classes; si le caractère considéré est par nature continu, l'étendue commune à toute les classes est au moins égale à l'unité du dernier ordre conservé pour exprimer les mesures.

Soit une série statistique, dite série initiale, constituée directement par un ensemble de mesures. D'après ce qui précède cette série est répartie en classes. L'étendue de ces classes est souvent jugée trop faible, notamment dans l'un et l'autre des deux cas suivants :

a) On désire faire différents calculs à partir des mesures de la série : manifestement, il est plus simple de calculer les moments de divers ordres d'une série d'un grand nombre de mesures lorsque cette série comprend de nombreuses répétitions que lorsqu'elle n'en comprend que peu ou pas du tout; le moyen de faire apparaître des répétitions est de choisir une étendue de classe supérieure à l'étendue des classes initiales, de définir grâce à elle de nouvelles classes et enfin de répartir les mesures de la série initiale dans ces classes; bien entendu,

(1) Nous disons qu'une série est régulièrement répartie de ω en ω lorsqu'elle est repartie dans des classes dont les etendues respectives sont toutes egales à ω et que toutes les mesures de l'une quelconque des classes, sont remplacées par la moyenne arithmétique des limites de cette classe.

sauf exception, un résultat de calcul obtenu en agissant comme si les mesures faisant partie d'une même classe avaient toutes la même valeur n'est pas exactement égal à celui que donnerait le calcul direct avec les données de la série initiale : dans le cas d'un moment, on simplifie les calculs mais on n'obtient qu'une approximation du moment correspondant, relatif à la série initiale.

b) On désire avoir de la série une vue d'ensemble relativement expressive grâce à une construction graphique telle que celle d'un histogramme : pour qu'un tel graphique soit effectivement expressif, il faut que d'une classe à la suivante les effectifs ne varient pas d'une façon désordonnée et un moyen pour y parvenir est précisément de choisir de façon convenable à la fois une étendue de classe relativement grande et une origine de classes.

Nous nous proposons d'évoquer ici quelques questions en rapport avec ce qui précède. Pour cela nous considérons deux valeurs d'étendue, à savoir ω et $\Omega = k\,\omega$ (k étant un nombre entier positif). La série régulièrement répartie de ω en ω est dite série initiale; pour les calculs ou pour la représentation par un histogramme expressif, on a fait choix d'une série régulièrement répartie de Ω en Ω et cette série est dite série retenue. Les origines des classes de la série retenue coincident toutes avec des origines de classes de la série initiale.

2. Soit à établir des relations entre les moments d'ordre 1 et 2 de la série initiale et les moments correspondants de la série retenue, tous ces moments étant comptés à partir d'une même origine, d'ailleurs quelconque.

Ce problème admet une solution classique dans le cas particulier où ω est nul, c'est-à-dire dans un cas où la série initiale est représentable non par les extrémités supérieures des tuyaux d'orgues d'un histogramme mais par la courbe des densités de la loi de probabilité d'une variable aléatoire continue.

Sheppard a établi que si l'on calcule les moments (1) à partir d'une origine d'ailleurs quelconque on a

$$(m'_1)_o \backsimeq (m'_1)_\Omega$$
$$(m'_2)_o \backsimeq (m'_2)_\Omega - \frac{\Omega^2}{12}$$

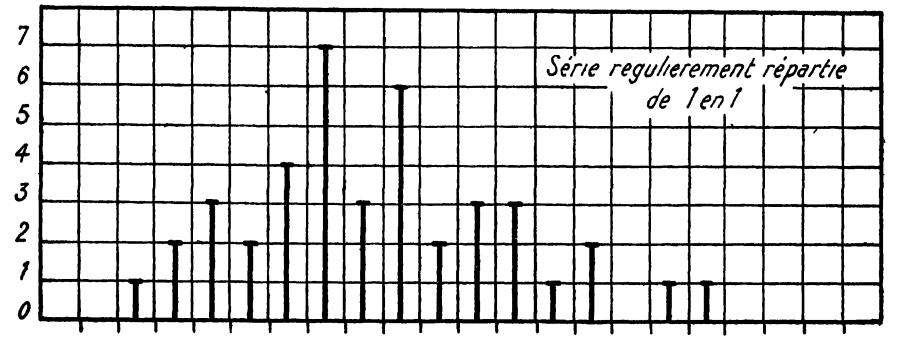
De la démonstration, je ne rappelle rien ici (Voir annexe I), si ce n'est que, basée sur des développements en série d'Euler—Maclaurin, elle suppose d'une part que la courbe des densités satisfait à certaines conditions de continuité; d'autre part que cette courbe n'a aucun extremum trop pointu eu égard à Ω; enfin qu'à ses deux extrémités cette courbe se raccorde convenablement à l'axe des abscisses. En particulier, la démonstration n'est pas valable lorsque la courbe est en U.

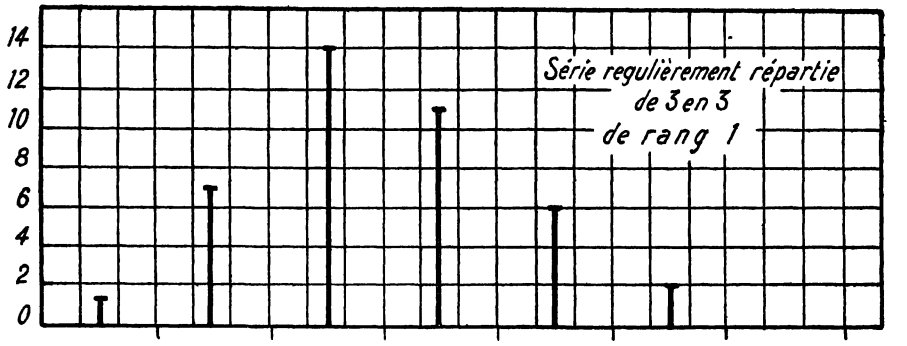
Peut-être les conditions moyennant lesquelles les formules de Sheppard sont valables ne sont-elles pas toujours présentes à l'esprit de celui qui utilise ces

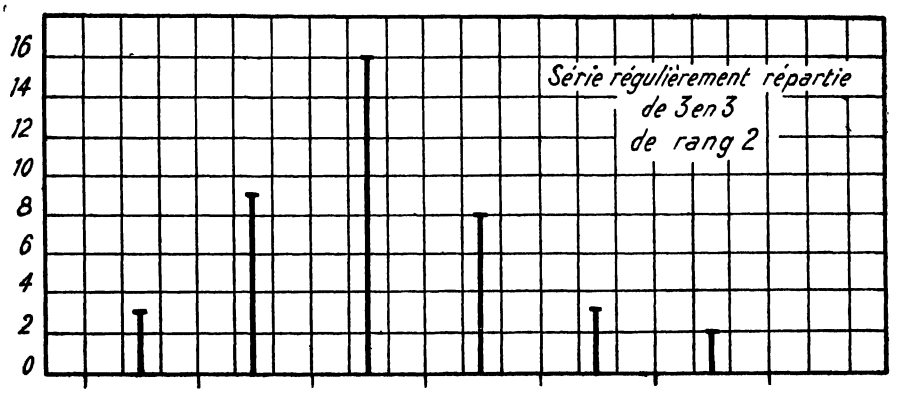
---

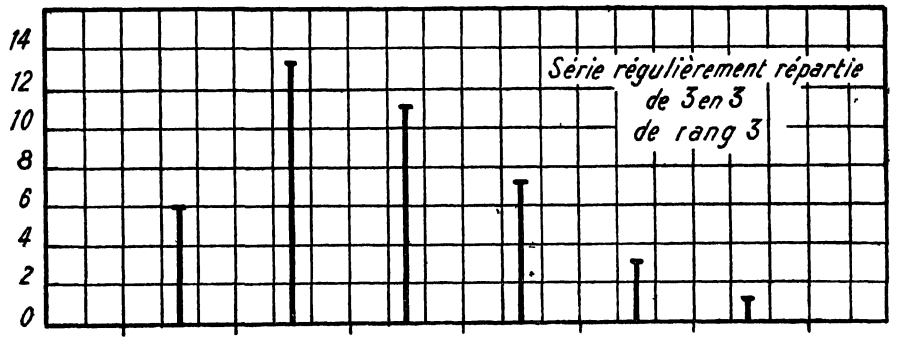
(1) m désigne un moment d'une variable aléatoire; toutefois, on écrit μ quand ce moment est compté par rapport à l'espérance mathématique de la variable.
m' désigne un moment d'une série statistique; toutefois on écrit μ' quand ce moment est compté par rapport à la moyenne arithmétique de la série.
De plus, le cas échéant, nous rappelons en indice en dehors d'une parenthèse la valeur de l'étendue des classes de la série.

formules, car il est rare me semble-t-il qu'il en soit fait mention. Je les donne en annexe telles que je les ai établies après quelques recherches faites à l'occa-









sion d'un travail sur l'application des méthodes statistiques au domaine des

techniques industrielles, sujet qui, même limité au cas de la statistique à un caractère, est vaste et dont j'ai essayé de faire le tour avec un autre ingénieur, M. Maheu.

3. C'est précisément M. Maheu à qui je dois le raisonnement qui est détaillé à l'annexe II et dont je vais essayer de donner ici une idée.

Le haut de la figure représente la série initiale, qui est une série répartie de ω en ω. On constitue une série régulièrement répartie de Ω en $\Omega = k\,\omega$ en groupant k intervalles consécutifs de la première série (sur la figure $k = 3$) et on la représente en élevant au milieu de cet intervalle une droite de longueur convenable. La série initiale donne naissance non pas à une seule série régulièrement répartie de Ω en Ω mais à k telles séries. Pour chacune de ces séries il est possible de faire les calculs correspondant à $(m'_1)_\Omega$ et à $(m'_2)_\Omega$. On trouve que la moyenne arithmétique des k valeurs différentes de $(m'_1)_\Omega$ est précisément égale à (m'_1), et que la moyenne arithmétique des k valeurs différentes de $(m'_2)_\Omega$ est égale à $(m'_2) + \frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right)$.

Il s'agit là non pas d'égalités approchées, mais d'égalités rigoureuses obtenues par des considérations en quelque sorte de simple arithmétique. D'après ces résultats, à la condition, mais à la condition seulement, que l'on consente à assimiler une grandeur faisant partie d'un certain ensemble à la moyenne arithmétique des grandeurs de cet ensemble, on peut écrire relativement à la série retenue

$$(m'_1)_\Omega^{(r)} \simeq (m'_1)_\omega \qquad (m'_2)_\Omega^{(r)} \simeq (m'_2)_\omega + \frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right).$$

A la limite pour $\omega = o$ et par suite pour $k = \infty$ nous retrouvons les expressions de Sheppard, mais nous les retrouvons après nous être instruits. En effet :

— d'une part, nous connaissons maintenant les expressions analogues à celles de Sheppard mais relatives au cas où ω est différent de 0;

— d'autre part, nous avons appris que les formules de Sheppard se justifient d'une certaine manière sans que l'on ait à faire aucune hypothèse ni sur la continuité ni sur la forme aux limites de la série initiale; en particulier ces formules sont vraies pour des séries initiales évoquant des courbes en U (un exemple correspondant à ce cas fait l'objet de l'annexe III);

— enfin, nous connaissons la vraie signification du signe de l'égalité approchée : c'est l'assimilation d'une grandeur à la moyenne arithmétique de la population dont fait partie cette grandeur.

4. Pour donner une idée des populations dont nous venons de parler, il faudrait faire état pour chacune d'elles non seulement comme ci-dessus de sa moyenne arithmétique, mais aussi de quelques indices de dispersion. Sans doute (Voir l'annexe V) est-il possible d'exprimer quelques-uns de ces indices au moyen de quantités relativement simples à former, mais il suffit en tout cas de prendre quelques exemples concrets de séries statistiques pour constater la grande dispersion (1) des $(m'_1)_\Omega$ et des $(m'_2)_\Omega$ autour de leurs moyennes arithmétiques respectives. Notamment on constate sur l'exemple faisant

(1) Les $(\mu'_2)_\Omega$ ont une dispersion inférieure à celles de $(m'_2)_\Omega$.

l'objet de l'annexe IV que la quantité $(m'_2)_\Omega^{(r)} - (m'_2)_\infty$ varie de part et d'autre de sa moyenne arithmétique $\frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right)$ lorsque r varie, mais que, lorsque k est petit, elle varie tellement qu'elle est fréquemment négative, de même qu'inversement elle est fréquemment égale à plus du double de cette quantité. C'est là un fait général; il faut reconnaître que ce fait laisse quelque peu sceptique sur l'intérêt qu'il y a à faire subir des corrections égales à $\frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right)$.

De ce point de vue deux cas sont à considérer :

— ou bien k est petit : l'intérêt dont il vient d'être parlé se traduit lorsque k est petit par ceci qu'il n'y a guère plus d'une chance sur deux que le résultat corrigé soit plus près de la réalité que le résultat non corrigé; en l'occurrence, faire ou non la correction calculée devient une question de bon vouloir ou de flair, mais ce n'est pas quelque chose qui s'impose comme correspondant à un fait presque certain;

— ou bien k est relativement grand : lorsque k est grand il n'en va pas de même que plus haut; la correction apparaît comme ayant alors bien des chances d'être dans le bon sens, ce qui conduit à la faire; en la faisant, on s'aperçoit d'ailleurs qu'elle est d'un ordre de grandeur voisin de celui de la quantité que l'on corrige. ce qui justifie quelques doutes sur la précision du résultat corrigé.

5. Que l'on ait 99 chances sur 100 ou seulement 51 sur 100, cela n'empêche que dans l'ensemble il est préférable de faire la correction. Encore ne faut-il la faire que dans les cas correspondant à la démonstration. Par exemple si, connaissant la série initiale, on veut estimer le moment d'ordre 2 d'une certaine série régulièrement répartie de Ω en Ω correspondant à cette série initiale, on peut prendre ce moment d'ordre 2 à peu près égal à $(m'_2)_\infty + \frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right)$ et l'on a plus d'une chance sur 2 d'être de la sorte plus près de la réalité que si l'on s'était borné à le prendre égal à $(m'_2)_\infty$.

D'auçuns pourront trouver cet exemple très intéressant; ce n'est pas mon cas, car le plus généralement, sinon exclusivement, ce que l'on désire avoir c'est une valeur approchée des moments de la série initiale (de celle qui comprend de nombreuses classes) à partir des moments d'une certaine série régulièrement répartie de Ω en Ω, moments qui sont faciles à calculer car un petit nombre seulement de classes est en jeu. La question se pose donc de savoir si pour ce problème il est logique de se servir des expressions trouvées.

Je ne le crois pas et je ne le croirais que si cela venait à m'être démontré, car les deux problèmes mettent en cause des populations entièrement différentes : d'une part la population des k séries régulièrement réparties de Ω en Ω auxquelles la série initiale donne naissance; d'autre part, la population beaucoup plus nombreuse que la précédente de toutes les séries imaginables régulièrement réparties de $\overline{\omega}$ en ω auxquelles on impose seulement comme conditions d'avoir, dans des groupes donnés de k classes consécutives, des effectifs également donnés.

A supposer même que dans un cas concret l'on soit parvenu à déterminer tous les individus de cette dernière population, sans doute ne va-t-on pas

attribuer des probabilités égales à chacun de ces individus. Le problème est donc de l'ordre de ceux des probabilités des causes. Sa solution classique doit être cherchée dans une voie faisant intervenir des probabilités *a priori* et la solution est fonction de l'hypothèse de probabilités *a priori* retenue.

6. La remarque qui précède m'amène naturellement à revenir sur ceci, qu'en groupant en classes d'étendue Ω des observations constituant une série statistique, on avait parfois en vue d'aboutir à une série dont les effectifs de classes ne variaient pas de façon désordonnée d'une classe à la suivante. On dispose pour cela de deux paramètres, à savoir la valeur k du rapport Ω/ω et l'origine de l'une des classes d'étendue Ω. Finalement, ayant agi par tâtonnements sur ces deux paramètres, on retient une série plutôt que toute autre, et si on la retient, c'est parce que l'on est prêt à faire l'hypothèse que son histogramme ou, mieux, son polygone des fréquences, donne une meilleure idée de la courbe des densités de la loi de probabilité inconnue qui est à l'origine de la série initiale que ne le fait l'histogramme de cette série. Cette hypothèse qui est tout à fait de la même nature qu'une hypothèse de probabilités *a priori*, permet notamment de résoudre le problème suivant qui est du plus haut intérêt pratique : partant d'une série d'observations, qui est une série régulièrement répartie de ω en ω, à quelle valeur doit-on estimer les moments des divers ordres de la loi de probabilité non discrète qui est à son origine ?

Une solution consiste :

— à considérer pour différentes valeurs de $\Omega = k\,\omega$, toutes les séries régulièrement réparties de Ω en Ω auxquelles la série initiale peut donner lieu;

— à faire un choix parmi toutes ces séries, c'est-à-dire à retenir l'une d'elles parce que l'on estime d'après les effectifs de ses classes successives, qu'elle doit être une image à peu près fidèle de la loi de probabilité qui est à l'origine de la série initiale;

— à admettre que l'histogramme, ou mieux que le polygone des fréquences, de la série obtenue sont très voisins de la courbe des densités de la loi en cause ci-dessus, et à calculer en conséquence la valeur cherchée, relative à un moment de cette loi.

On trouve par des considérations simples de centres de gravité et de moments d'inertie :

Cas de l'histogramme :

$$m_1 = (m'_1)_\Omega \qquad m_2 = (m'_2)_\Omega + \frac{\Omega^2}{12}.$$

Cas du polygone des fréquences :

$$m_1 = (m'_1)_\Omega \qquad m_2 = (m'_2)_\Omega + \frac{\Omega^2}{6}.$$

Ainsi en particulier lorsque l'on part d'une série régulièrement répartie de Ω en Ω que l'on juge être une image à peu près fidèle d'une loi, il faut, pour avoir le moment d'ordre 2 de cette loi, partir du moment d'ordre 2 de la série et ajouter à ce dernier une quantité de l'ordre de grandeur de $\Omega^2/12$, ou mieux de $\Omega^2/6$.

7. Cette règle s'oppose nettement à celle que l'on suit chaque fois que l'on pense résoudre le problème considéré en appliquant la correction de Sheppard,

puisque cette correction loin de conduire à ajouter quelque chose, conduit à retrancher $\Omega^2/12$.

Quel parti faut-il donc prendre dans le cas du problème d'estimation en cause ici? M. Risser, qui a vu l'alternative, s'est prononcé, sans longue explication, en faveur du-$\Omega^2/12$. Moi-même, ne serait-ce que pour le plaisir de soutenir une opinion peu répandue, je préconise d'ajouter $\Omega^2/6$ plutôt que de retrancher $\Omega^2/12$ et cela pour deux raisons.

La première est que le fait de passer par l'intermédiaire du polygone des fréquences, offre le grand avantage que l'on voit ce que l'on fait : on a sous les yeux la ligne à laquelle correspondent les valeurs trouvées, on sait à quoi s'en tenir, on peut chiffrer les répercussions de quelques retouches, etc...

La seconde est que le retranchement de $\Omega^2/12$ n'est justifié que moyennant un très grand nombre d'hypothèses : hypothèse qu'il existe une fonction continue, ayant une expression mathématique unique dans tout l'intervalle de variations à considérer, et dont la courbe représentative limite exactement dans chacun des intervalles de classes des aires proportionnelles aux effectifs de ces classes, et surtout hypothèse que cette fonction, dont on a bien rarement une idée quelque peu nette, satisfait aux conditions aux limites, aux conditions d'absence de maximum « pointu », etc..., nécessaires pour que la formule de Sheppard fournisse une bonne approximation des moments lui correspondant.

Et je termine cet exposé en remarquant que si réellement, comme je le préconise, les formules de Sheppard n'ont pas à être appliquées d'une façon générale et en quelque sorte automatique, dans le domaine de l'estimation des moments relatifs à la loi à laquelle une série appartient, cela leur enlève la plus grande partie de l'intérêt qu'on leur attache ordinairement.

Annexe I. — Démonstration, conditions de validité et généralisation de la formule de Sheppard.

Partons d'une loi de probabilité dont la fonction des densités est $\delta(x)$. Découpons l'aire bordée par la courbe des densités en trapèzes curvilignes dont les côtés parallèles entre eux ont pour abscisses $(j \mp 0{,}5)\ \Omega$.

Posons :

$$f_j = \int_{(j-0{,}5)\frac{\Omega}{2}}^{(j+0{,}5)\frac{\Omega}{2}} \delta(x)\,dx \quad ; \text{ d'où (Voir } N.\ B.\ 1\text{)} :$$

$$f_j = \Omega\,\delta(j\,\Omega) + \frac{\Omega^3}{24}\frac{d^2\delta}{dx^2}(j\,\Omega) + \ldots\ldots \tag{1}$$

Les f_j sont des fréquences qui définissent une série régulièrement répartie de Ω en Ω, pour laquelle on a :

$$(m'_2)_\Omega = \sum_j f_j \Omega^2 j^2 = [\text{d'après (1)}]\ \Omega \sum_j \Omega^2 j^2 \delta(j\,\Omega) + \frac{\Omega^3}{24}\sum_j \Omega^2 j^2 \frac{d^2\delta}{dn^2}(j\,\Omega) + \ldots \tag{2}$$

d'où en remplaçant les sommes Σ par des intégrales (Voir $N.\ B.\ 2$) :

$$(m'_2)_\Omega \backsimeq \int_{-\infty}^{+\infty} x^2 \delta(x)\,dx + \frac{\Omega^2}{24}\int_{-\infty}^{+\infty} x^2 \frac{d^2\delta}{dx^2}(x)\,dx. \tag{3}$$

Comme l'intégration par parties de la dernière intégrale montre que cette intégrale est égale à 2, on a finalement :

$$(4) \qquad (m'_2)_\Omega \backsimeq m_2 + \frac{\Omega^2}{12},$$

ce qui est la formule de Sheppard.

N. B. 1. — La formule (1) suppose que l'on est en droit d'appliquer dans l'intervalle d'indice j la formule d'Euler-Maclaurin; elle suppose donc que la courbe de la fonction $\delta(x)$ n'a dans cet intervalle ni point de discontinuité, ni point anguleux.

Dans la suite, on se limite aux deux termes écrits explicitement, ce qui n'est acceptable qu'à la condition que la dérivée d'ordre 3 soit à peu près constante dans l'intervalle considéré. De ce fait, notamment, ces termes ne suffisent pas, sauf exception, pour bien représenter une fonction présentant un maximum pouvant être qualifié de « pointu », compte de la valeur de Ω.

N. B. 2. — Soit $g(x)$ une fonction à considérer lorsque l'on envisage de remplacer une somme Σ par une intégrale : par exemple, pour passer de (2) à (3) on considère $g(x) = x^2\,\delta(x)$ et aussi $g(x) = x^2 \frac{d^2\delta}{dx^2}(x)$.

A propos de tels remplacements, notons que la formule d'Euler Maclaurin (Voir *N. B. 1*) permet d'écrire :

$$(5) \qquad \int_{(j-0,5)\Omega}^{(j+0,5)\Omega} g(x)\,dx \backsimeq \Omega\, g(j\Omega) + \frac{\Omega^3}{24}\frac{d^2 g}{dx^2}(j\Omega) + \frac{\Omega^5}{1920}\frac{d^4 g}{dx^4}(j\Omega).$$

$$(6) \qquad \frac{dg}{dx}[(j+0,5)\Omega] - \frac{dg}{dx}[(j-0,5)\Omega] \backsimeq \Omega\frac{d^2 g}{dx^2}(j\Omega) + \frac{\Omega^3}{24}\frac{d^4 g}{dx^4}(j\Omega).$$

$$(7) \qquad \frac{d^3 g}{dx^3}[(j+0,5)\Omega] - \frac{d^3 g}{dx^3}[(j-0,5)\Omega] \backsimeq \Omega\frac{d^4 g}{dx^4}(j\Omega).$$

L'élimination des quantités $\frac{d^2 g}{dx^2}(j\Omega)$ et $\frac{d^4 g}{dx^4}(j\Omega)$ entre ces équations conduit à :

$$\Omega\, g(j\Omega) \backsimeq \int_{(j-0,5)\Omega}^{(j+0,5)\Omega} g(x)\,dx - \frac{\Omega^2}{24}\left[\frac{dg}{dx}[(j+0,5)\Omega] - \frac{dg}{dx}[(j-0,5)\Omega]\right] + \frac{7\,\Omega^4}{5760}\left[\frac{d^3 g}{dx^3}[(j+0,5)\Omega] - \frac{d^3 g}{dx^3}[(j-0,5)\Omega]\right],$$

d'où, à supposer que de part et d'autre de chacune des limites de classes, les dérivées d'ordres 1, 3, etc... soient toutes les mêmes, ce qui impose pratiquement que $g(x)$ ait une expression analytique unique dans tout l'intervalle $(-\infty; +\infty)$:

$$(8) \qquad \sum^{+\infty} \Omega\, g(j\Omega) \backsimeq \int_{-\infty}^{+\infty} g(x)\,dx - \frac{\Omega^2}{24}\left[\frac{dg}{dx}(+\infty) - \frac{dg}{dx}(-\infty)\right] + \frac{7\,\Omega^4}{5760}\left[\frac{d^3 g}{dx^3}(+\infty) - \frac{d^3 g}{dx^3}(-\infty)\right].$$

Pour que le remplacement d'une somme par une intégrale soit acceptable suivant la formule

$$(9) \qquad \sum_{=-\infty}^{+\infty} \Omega\, g(j\Omega) \backsimeq \int_{-\infty}^{+\infty} g(x)\,dx,$$

il faut, d'après le calcul ci-dessus et sa généralisation évidente, non seulement qu'aux limites les fonctions $\frac{dg}{dx}, \frac{d^3 g}{dx^3}$ et autres dérivées d'ordre impair soient nulles, mais encore que les formules (5) à (7) puissent être écrites, ce qui suppose que les fonctions $\int g(x)\,dx, \frac{dg}{dx}, \frac{d^3 g}{dx^3}$ et autres dérivées d'ordre impair satisfont aux conditions de continuité, d'absence de maximum « pointu » etc..., évoquées au *N. B. 1.*

On a une idée des erreurs susceptibles d'être faites lorsque ces conditions ne sont pas toutes respectées en considérant à titre d'exemple les deux égalités suivantes :

$$\Sigma_{u=0}^{5} u^2 (5-u)^{50} \simeq 1^2 \times (5-1)^{50} = 1{,}27.10^{30};$$

$$\int_0^5 u^2 (5-u)\,dx = 5^{53} \frac{2!\,50!}{53!} = 1{,}58.10^{32}.$$

L'intégrale est dans ce cas 125 fois plus grande que la somme Σ.

N. B. 3. — Les conditions moyennant lesquelles la formule de Sheppard est applicable sont celles qui viennent d'être indiquées; il suffit de remplacer $g(x)$ d'une part par $x^2 \delta(x)$, de l'autre, par $x^2 \frac{d^2 \delta}{dx^2}$.

Remarquons toutefois que les conditions aux limites sont en fait un peu différentes de celles qui résultent de ce qui précède. Examinons ce point en prenant pour plus de généralité a et b comme limites au lieu de $\mp \infty$. Si on fait intervenir dans (1) le terme $+ \frac{\Omega^5}{1920} \frac{d^4 \delta}{dx^4}$ ($/\ \Omega$) et si dans chacune des intégrations par parties telles que celle faite pour passer de (3) à (4) on écrit explicitement la quantité qui intervient par ses valeurs aux limites, des simplifications se produisent dans les calculs et l'on trouve :

$$(10) \qquad (m'_2)_\Omega = m_2 + \frac{\Omega^2}{12} - \frac{\Omega^2}{6} \left| x\,\delta(x) \right|_a^b + \frac{\Omega^4}{360} \left| 3\frac{d\delta}{dx} + x\frac{d^2\delta}{dx^2} \right|_a^b.$$

Les conditions aux limites se réduisent donc à celles qui apparaissent sur cette formule.

D'un autre point de vue, (10) doit remplacer (4) lorsque les termes entre barres ne peuvent pas être considérés comme nuls et lorsque a et b sont les limites des classes extrêmes d'étendue Ω (hors de cette dernière condition, la formule (1) ne serait pas légitimement employée dans le raisonnement qui conduit à (10)). Par exemple, soit la loi ayant pour intervalle de variations $(0; +\infty)$ et pour fonction des densités $\delta(x) = \sqrt{\frac{2}{\pi}}\, e^{-\frac{x^2}{2}}$; d'où :

$$E[x] = 0{,}798,\ \delta(o) = 0{,}798 \quad \text{et} \quad \frac{d^2\delta}{dx^2}(o) = -0{,}798;$$

considérant les moments par rapport au point d'abscisse 1, on a :

$$m_1 = -0{,}202 \text{ et } m_2 = 0{,}404$$

et d'après (10) pour $\Omega = 1$:

$$(m'_2)_\Omega = 0{,}404 + 0{,}083 - 0{,}133 - 0{,}002 = 0{,}352.$$

A remarquer l'importance relative du terme en $\Omega^2/6$.

Annexe II. — Démonstration des formules :

Moyenne arithmétique de $(m'_1)_\Omega = (m'_1)_\omega$

Moyenne arithmétique de $(m'_2)_\Omega = (m'_2)_\omega + \frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right)$.

Les moments en cause ci-dessous sont tous comptés par rapport à un point qui est à une distance $d_0 + \omega/2$ du début de la classe de rang 1 de la série régulièrement répartie de ω en ω qui est la série initiale, si bien que la classe de rang i de cette série est caractérisée par la valeur $d_i = d_0 + i\,\omega$; la fréquence des mesures dans cette classe est f_i.

Les k séries régulièrement réparties de Ω en $\Omega = k\,\omega$ auxquelles la série initiale donne naissance ont respectivement pour moments d'ordre 2 : $(m'_2)^{(1)}_\Omega$, $(m'_2)^{(2)}_\Omega$, ..., $(m'_2)^{(k)}_\Omega$. Chacune de ces quantités est une fonction linéaire des f_i; par exemple :

$$(m'_2)^{(1)}_\Omega = (f_1 + f_2 + \ldots + f_k)\left(d_0 + \frac{k+1}{2}\,\omega\right)^2$$
$$+ (f_{k+1} + f_{k+2} \ldots + f_{2k})\left(d_0 + \Omega + \frac{k+1}{2}\,\omega\right)^2 + \ldots$$

Par suite on peut poser :

$$\frac{1}{k}\sum_{j=1}^{k}(m'_2)^{(j)}_\Omega = f_1 C_1 + f_2 C_2 \ldots + f_i C_i + \ldots$$

Pour déterminer C_i, on remarque que f_i intervient dans k classes d'étendue Ω, à savoir :

— dans la classe se terminant en $d_0 + (i + 0{,}5)\,\omega$, caractérisée par $d_0 + (i + 0{,}5 - 0{,}5\,k)\,\omega$;

— dans la classe se terminant en $d_0 + (i + 1{,}5)\,\omega$, caractérisée par $d_0 (i + 1{,}5 - 0{,}5\,k)\,\omega$; etc...

— enfin, dans la classe se terminant en $d_0 + (i + k - 0{,}5)\,\omega$, caractérisée par $d_0 + (i + 0{,}5\,k - 0{,}5)\,\omega$.

Par suite :

$$k\,C_i = \sum_{j=1}^{k}[d_0 + (i - 0{,}5\,k - 0{,}5)\,\omega + j\,\omega]^2$$
$$= k\,[d_0 + (i - 0{,}5\,k - 0{,}5)\,\omega]^2 + 2\,\omega\,[d + (i - 0{,}5\,k - 0{,}5)\,\omega]\,0{,}5\,k\,(k+1)$$
$$+ \omega^2\,k\,(k+1)\,(2k+1)/6$$
$$= k\,[d_0 + i\,\omega]^2 + \text{etc...}$$

et finalement :

$$C_i = d_i^2 + \frac{\omega^2}{12}(k^2 - 1),$$

d'où l'expression annoncée relative aux moments d'ordre 2.

L'autre expression s'établit encore plus simplement que la précédente.

Annexe III. — Exemple d'une répartition en U.

Moments calculés par rapport au milieu de la classe de rang 0, de la série donnée.

a) Cas de la série donnée, qui est une série régulièrement répartie de $\omega = 1$ en ω :

$$(m'_1)_\omega = 16{,}74;\ (m'_2)_\omega = 430{,}24.$$

b) Cas des séries régulièrement réparties de $\Omega = 10$ en Ω correspondant à la série donnée.

Il y a 10 séries de l'espèce; leurs $(m'_2)_\Omega$ respectifs sont les suivants, mis en regard de la valeur j, laquelle indique que le début d'une classe de la série en cause coïncide avec le début de la classe de rang j de la série donnée :

$j = 1 : 437{,}25$; $j = 2 : 437{,}65$; $j = 3 : 395{,}75$; $j = 4 : 412{,}05$; $j = 5 : 431{,}45$;
$j = 6 : 435{,}35$; $j = 7 : 466{,}95$; $j = 8 : 451{,}75$; $j = 9 : 455{,}95$; $j = 10 : 460{,}75$.

D'où

$$\frac{1}{10}\,\Sigma_{j=1}^{10}\,(m'_2)_\Omega^{(j)} \quad 438{,}49.$$

c) Comme $\frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right)$ égale dans le cas ci-dessus 8,25 et que

$$438{,}49 - 8{,}25 = 430{,}24 = (m'_2)_\omega$$

on a bien :

$$\frac{1}{k}\,\Sigma_j\,(m'_2)_\Omega^{(j)} = (m'_2)_\omega + \frac{\Omega^2}{12}\left(1 - \frac{1}{k^2}\right).$$

Série donnée :

RANGS	EFFECTIFS	RANGS	EFFECTIFS
0	13	19	0
1	5	20	3
2	7	21	4
3	2	22	1
4	4	23	3
5	1	24	2
6	2	25	5
7	0	26	6
8	3	27	8
9	1	28	4
10	2	29	5
11	1	30	7
12	1	31	2
13	0	32	16
14	0	TOTAL .	100
15	1		
16	0		
17	2		
18	1		

Annexe IV. — Exemple d'une répartition en cloche.

Série donnée, qui est une série de poids, régulièrement répartie de $\omega = 0{,}1\ g$ en ω :

POIDS	EFFECTIFS	POIDS	EFFECTIFS	POIDS	EFFECTIFS	POIDS	EFFECTIFS	POIDS	EFFECTIFS
3,5	1	4,8	1	5,6	6	6,4	2	7,2	2
3,6	0	4,9	2	5,7	3	6,5	6	7,3	0
à		5,0	3	5,8	3	6,6	6	à	...
4,2	0	5,1	2	5,9	6	6,7	5	8,1	0
4,3	1	5,2	1	6,0	8	6,8	4	8,2	1
4,4	1	5,3	5	6,1	4	6,9	3		100
4,5	0	5,4	1	6,2	6	7,0	1	TOTAL	
4,6	1	5,5	6	6,3	4	7,1	3	GÉNÉRAL.	
4,7	2								

Moments calculés par rapport à la valeur 3,5; les origines des classes sont, suivant les cas : 3,45; 3,55; 3,65, etc... :

a) Cas de la série donnée $(m'_1)_\omega = 2{,}481$; $(m'_2)_\omega = 6{,}7179$; $(\mu'_2)_\Omega = 0{,}5625$;

b) Cas de séries régulièrement réparties de Ω en Ω; les résultats sont donnés dans l'ordre correspondant aux origines des classes situées en 3,55; en 3,65; en 3,75, etc...

Ω = 0,3		Ω = 1,0			Ω = 2,1			
$(m'_1)_\Omega$	$(m'_2)_\Omega$	$(m_1)_\Omega$	$(m'_2)_\Omega$	$(\mu'_2)_\Omega$	$(m_1)_\Omega$	$(m'_2)_\Omega$	(Suite)	
2,468	6,6820	2,500	6,9375	0,6875	2,507	7,4365	2 436	6,7914
2,478	6,6942	2,470	6,7885	0,6876	2,544	7,6644	2,410	6,5578
2,497	6,7975	2,470	6,7825	0,6816	2,581	7,8871	2,384	6 3820
		2,490	6,8505	0,6504	2,555	7,7959	2,400	6,3774
		2,480	6,7995	0,6491	2,487	7 4601	2,437	6,5965
		2,460	6,7335	0,6819	2,482	7,3228	2,411	6,4387
		2,470	6,7385	0,6376	2,456	7,1362	2,427	6 4773
		2,470	6,6325	0,5316	2,472	7,1586	2,506	6,8992
		2,470	6,7065	0,6056	2,509	7,6995	2,501	7,0219
		2,530	7,0345	0,6336	2,462	7,0828	2,580	7,4502
							2,554	7,4596
(1) *2,481*	6,7246	*2,481*	6,8004	0,6446			*2,481*	7,0846 (1)
(2)	—0,0067		—0,0825	0,0825				— 0,3667 (2)
(3)	*6,7179*		*—6,7179*	0,5621				*6,7179* (3)
(4) 0,018	0,0518	0,020	0,1104	0,0456			0,058	0,4824 (4)

(1) Moyenne arithmetique (2) $-\frac{\Omega^2}{12}\left(1-\frac{1}{k^2}\right)$ (3) Total (1) + (2) (4) Écart-moyen-apparent quadratique

Annexe V. — Indices de dispersion divers.

Cas des $(m'_1)_\Omega$. Les $(m'_1)_\Omega$ des k séries régulièrement réparties de $\Omega = k\,\omega$ en Ω issues d'une même série régulièrement répartie de ω en ω admettent un écart-moyen-apparent quadratique σ_1 qui s'exprime au moyen des quantités suivantes :

S_ι : somme des effectifs des classes de la série initiale ayant pour rangs ι; $\iota + k$; $\iota + 2k$; ...

σ_s : écart-moyen-apparent quadratique de la série régulièrement répartie de ω en ω ayant comme effectifs de classes respectivement S_1; S_2;... S_k.

On établit :

$$\frac{\sigma_1^2}{\omega^2} = \sigma_s^2 + \frac{k^2-1}{12} - \frac{k}{N^2}\sum_{\iota=1}^{k-1}\sum_{j=\iota+1}^{k}(j-\iota)\,S_\iota\,S_j.$$

Exemple : Avec les données de l'annexe IV et Ω = 1,0, on a :

$$S_1 = 13;\ S_2 = 13;\ S_3 = 10;\ S_4 = 8;\ S_5 = 11;$$

$$S_6 = 12;\ S_7 = 9;\ S_8 = 10;\ S_9 = 10;\ S_{10} = 4;$$

$$\sigma_s^2 = 7{,}7099 \text{ et } \sum_{\iota=1}^{k-1}\sum_{j=\iota+1}^{k}(j-\iota)\,S_\iota\,S_j = 15921.$$

Donc :

$$\sigma_1^2 = 0{,}01\ [7{,}7099 + 8{,}2500 - 15{,}9210] = 0{,}000389 \backsimeq 0{,}020^2,$$

qui est la valeur trouvée à l'Annexe IV.

Cas des $(m'_2)_\Omega$. Les $(m'_2)_\Omega$ des k séries régulièrement réparties de Ω en Ω issues de la série initiale particularisée en ceci qu'elle comprend des classes en nombre multiple de k et que les effectifs de chacune de ces classes sont tous égaux entre eux, admettent σ_2 pour écart-moyen-apparent quadratique avec

$$\sigma_2^2 = \frac{\Omega^4}{180}\left(1 - \frac{1}{k^2}\right)\left(1 + \frac{11}{k^2}\right).$$

N. B. — On remarque que dans le cas particulier en cause ici, le σ_2 est, pour k grand, voisin de $\Omega^2/\sqrt{180}$ et par suite voisin de $\Omega^2/12$, terme correctif de Sheppard.

M. Dumas.

DISCUSSION

M. Risser. — M. Dumas a bien voulu me communiquer ses bonnes feuilles; il m'a ainsi permis de rectifier les observations que j'avais formulées immédiatement à la suite de sa très intéressante communication, alors que je n'avais pas présentes à l'esprit diverses considérations touchant à l'emploi de la méthode de Sheppard, et à l'abus que l'on pouvait en faire lorsque diverses circonstances relatives à son emploi n'étaient pas remplies.

Je tiens tout d'abord à signaler l'intérêt de l'étude de notre collègue, des remarques qu'elle présente à propos de la justification du classement des observations, de la détermination des moments du premier et du second ordre conduisant à caractériser un tableau de mesures après l'avoir remplacé par un polygone de trapèzes ou de rectangles, non seulement dans le cas où l'on adopte pour l'échelle des abscisses la grandeur ω, mais encore dans celui où l'on fait intervenir $k\omega$, avec $k = 1, 2, 3...$). Qu'il me soit permis à ce propos de rappeler un exemple classique dû à Czuber, relatif à la liste des hauteurs en centimètres de 125 pins âgés de 9 ans; en effet de ce tableau qui donne des renseignements sur la collection, on ne peut tirer une vue d'ensemble. En l'absence de toute indication, on a supposé que les longueurs sont mesurées au demi-centimètre près, c'est-à-dire que la mesure 180 comprend les individus ayant de 179 cm. 5 à 180 cm. 5, On a été ainsi conduit à faire plusieurs classements de la collection; Czuber a considéré comme normaux ceux qui correspondaient respectivement à des amplitudes de 15 et 10 centimètres.

Toutes les méthodes élémentaires donnant le moyen de calculer la moyenne, la médiane, la dispersion, les quartiles supposent que le contenu d'une classe est concentré sur son abscisse moyenne, ce qui ne s'applique rigoureusement qu'au cas des variables discontinues.

Si l'on cherche à tenir compte de la distribution des individus dans toute l'étendue de la classe, on peut recourir à la méthode de Sheppard, *mais l'on ne doit point perdre de vue les multiples hypothèses que l'on doit introduire en l'occurrence.*

En effet, si l'on désigne par $y = f(x)$, l'équation de la courbe de fréquence, et N son aire totale, on sait que $y = \int_a^b f(x)dx$; quant à l'aire comprise entre les abscisses $(x_i \pm 1/2)$, elle a pour valeur :

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

Alors que le moment d'ordre p a pour valeur :

$$\mu_p = \frac{1}{N} \sum_{a'}^{b'} x_i^p y_i, \text{ avec } a' = a + \tfrac{1}{2}, \quad b' = b - \tfrac{1}{2},$$

le moment calculé d'après l'équation de la courbe s'exprime au moyen de la relation :

$$\mu'_p = \frac{1}{N} \int_a^b x^p f(x)\, dx.$$

Sheppard cherche quelle correction doit être appliquée à μ_p pour trouver μ'_p, et fait état de l'équation :

$$N \mu_p = \sum_{a'}^{b'} x_i^p f(x_i) + \frac{1}{24} \sum_{a'}^{b'} x_i^p f''(x_i) + \frac{1}{1920} \sum_{a'}^{b'} x_i^p f^{(IV)}(x_i) + \ldots,$$

et utilise ensuite la formule sommatoire de Max Laurin.

Or dans le calcul de $\int_a^b x^p f''(x)\, dx$, et de $\int_a^b x f^{(IV)}(x)\, dx$, il ne laisse subsister respectivement que :

$$p(p-1) \int_a^b x^{p-2} f(x)\, dx, \text{ et } p(p-1)(p-2)(p-3) \int_a^b x^{p-4} f(x)\, dx.$$

Or cette hypothèse est vérifiée par la courbe de Gauss, les limites étant infinies; elle l'est également pour le groupe des courbes de Pearson et pour les limites finies ou infinies que chacune d'elles comporte; toutefois, je n'ai pas pu ces jours derniers me procurer le mémoire original de Sheppard, et vérifier s'il en est de même pour le type particulier de la courbe en U de Pearson (dont on trouve des exemples dans l'examen de la nébulosité de l'atmosphère).

Les nombreuses hypothèses qu'introduit ainsi Sheppard sont loin d'être remplies; aussi est-il sage de ne point faire état de la formule suivante résultant de l'emploi de ces hypothèses :

$$\mu_2 = \mu'_2 + {}^1/_{12}.$$

Je signale que MM. Traynard et moi avons tenu à appeler l'attention du lecteur de l'ouvrage *Les principes de la statistique mathématique* sur la différence profonde existant entre la formule de Sheppard, et celles qui résultent d'un calcul direct des moments de l'ensemble des fréquences représenté dans un cas par des trapèzes, et dans l'autre par des rectangles.

En effet, M. Traynard — dans les Annales de l'École normale (1909) — a exposé d'après Pearson ce calcul des moments, et trouvé les formules suivantes :

Ensemble de trapèzes $\mu'_1 = \mu_1 = 0, \quad \mu'_2 = \mu_2 + {}^1/_6.$
Ensemble de rectangles $\mu'_1 = \mu_1 = 0, \quad \mu'_2 = \mu_2 + {}^1/_{12}$ (ou $\mu_2 = \mu'_2 - {}^1/_{12}$).

M. Dumas nous a apporté une contribution importante en distinguant le classement de la distribution par bandes de largeur ω, de celui caractérisé par des bandes de largeur $\Omega = k\omega$ (k pouvant prendre des valeurs entières) et en comparant $(m'_1)_\Omega$ et $(m'_2)_\Omega$ respectivement à $(m'_1)_\omega$ et $(m'_2)_\omega$.

La communication de M. Dumas sera lue — j'en suis convaincu — avec le plus vif intérêt, car elle nous incite à procéder à un examen attentif des distributions statistiques, et nous incite à utiliser dorénavant les formules approchées qui ont été présentées par notre distingué collègue.

M. DUMAS est heureux que son respecté collègue M. Risser, se rencontre maintenant avec lui pour inciter à la prudence dans l'emploi de la formule de Sheppard et s'excuse de ne pas avoir, dans son exposé, cité M. Traynard, dont il connaissait bien, cependant, le beau travail de 1909. Il saisit cette occasion pour attirer l'attention sur ce que le raisonnement de Sheppard reproduit par M. Risser suppose que la différence $b - a$ est entière ou infinie : s'il n'en est pas ainsi et si $(a\,;\,b)$ est le plus grand intervalle à l'intérieur duquel $f(x)$ n'est pas nul, il convient de déterminer a'' et b'' tels que l'on ait $a \leqslant a'' < b'' \leqslant b$ et que $b'' - a''$ soit le plus grand entier contenu dans $b - a$, puis d'appliquer à l'intervalle $(a''\,;\,b'')$ la formule (10) de l'annexe 1, avec $\Omega = 1$.