

ANNALES DE L'I. H. P.

E.J. GUMBEL

Les valeurs extrêmes des distributions statistiques

Annales de l'I. H. P., tome 5, n° 2 (1935), p. 115-158

http://www.numdam.org/item?id=AIHP_1935__5_2_115_0

© Gauthier-Villars, 1935, tous droits réservés.

L'accès aux archives de la revue « Annales de l'I. H. P. » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

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

Les valeurs extrêmes des distributions statistiques

PAR

E. J. GUMBEL

Introduction

Considérons les résultats d'un nombre fini N d'observations portant sur une distribution statistique illimitée. Parmi ces résultats, il y en aura toujours une valeur (ou plusieurs) qui dépassera toutes les autres, tout en restant finie puisque évidemment nous ne pouvons observer que des valeurs finies. Il s'agit de préciser cette valeur maximum et d'analyser la manière dont elle croît avec le nombre des observations.

Des parties importantes de ce problème ont déjà été résolues. En effet, on a reconnu que cette valeur possède les propriétés d'une variable statistique, c'est-à-dire qu'il existe une distribution de « la plus grande valeur », dépendant du nombre des observations et de la distribution, dite initiale, envisagée. M. von BORTKIEWICZ (3 ; 4) a calculé « la plus grande valeur » dans le cas d'une loi initiale de Gauss et pour un petit nombre d'observations. M. L. C. H. TIPPETT (32) a calculé les premiers moments, pour les mêmes conditions. M. von MISES (29) a donné une valeur approximative de l'espérance mathématique, valable pour un grand nombre d'observations et pour une distribution initiale soumise à certaines conditions. M. TRICOMI (33) a calculé directement l'espérance mathématique dans le cas de la loi de Gauss et pour un grand nombre d'observations. C'est pour les mêmes conditions que M. DE FINETTI (7) a établi des tables numériques de la valeur médiane.

Un travail important est dû à M. E. L. DODD (5) qui a calculé l'espérance mathématique de la dernière valeur pour les distributions initiales de Gauss et sa transformation logarithmique, pour les distributions de Makeham et de Pearson, et pour la loi des événements rares. M. M. FRÉCHET (10) a introduit la notion très féconde de distribution finale stable, qui reproduit la distribution initiale. En partant des valeurs de la variable divisée par une valeur moyenne, il arrive à une distribution finale pour laquelle les moments d'ordre élevé divergent. Enfin, M. R. A. FISHER (8) a calculé les premiers moments pour une autre distribution finale stable. C'est elle que nous allons retrouver.

Dans ce qui suit nous allons systématiser et élargir ces résultats en considérant des conditions moins rigides que les précédentes et qui permettront d'appliquer la théorie générale à d'autres distributions que celles envisagées jusqu'à présent. En outre nous élargirons le problème fondamental en examinant la distribution de la $m^{\text{ième}}$ valeur au lieu de nous borner à la valeur maximum, m étant un nombre assez petit par rapport au nombre d'observations. Nous envisagerons les premières et les dernières valeurs et nous calculerons l'espérance mathématique et les moments de ces valeurs dans le cas d'une distribution initiale satisfaisant à certaines conditions et pour un grand nombre d'observations.

Soit $w(x)$ une distribution continue d'une variable statistique x , à variation illimitée dans les deux sens et dont l'espérance mathématique \bar{x} est

$$(1) \quad \bar{x} = \int_{-\infty}^{+\infty} zw(z)dz.$$

La probabilité d'une valeur inférieure à x sera

$$(2) \quad W(x) = \int_{-\infty}^x w(z)dz$$

grandeur qui satisfait aux conditions limites $W(-\infty) = 0$; $W(\infty) = 1$. La seconde de ces équations définira ce que nous appellerons *la condition de l'aire*. La valeur médiane \tilde{x} est définie par

$$(1a) \quad W(x) = \frac{1}{2}$$

et la, ou les valeurs dominantes par

$$(1b) \quad w'(\tilde{x}) = 0; \quad w''(\tilde{x}) < 0.$$

On nomme intervalle probable la différence entre deux valeurs x'' et x' telles que

$$(2a) \quad W(x'') = \frac{3}{4}; \quad W(x') = \frac{1}{2}.$$

La probabilité pour que la variable soit contenue dans l'intervalle $x' < x < x''$, ou qu'elle la surpasse est égale à $1/2$. La moitié de cet intervalle sera appelée l'écart probable ρ .

Une distribution sera appelée symétrique par rapport à l'origine si $w(-x) = w(x)$. Dans ce cas la probabilité d'une valeur inférieure à $-x$ sera

$$W(-x) = 1 - W(x)$$

Si une distribution symétrique a une dominante égale à zéro, les trois moyennes coïncident. En effet, dans ce cas

$$W(0) = \frac{1}{2}$$

et l'espérance mathématique

$$\bar{x} = \int_{-\infty}^{+\infty} z dW(z)$$

est nulle. Cette propriété remarquable de la coïncidence des trois moyennes se rencontre particulièrement dans une distribution de Gauss. La *dispersion* se définit par

$$(3) \quad \sigma^2 = \int_{-\infty}^{+\infty} (z - \bar{x})^2 w(z) dz$$

et on appelle *écart type*, sa racine carrée positive. La différence entre l'espérance mathématique et la dominante divisée par l'écart type

$$(4) \quad \delta = \frac{\bar{x} - \tilde{x}}{\sigma}$$

mesure l'asymétrie. La différence entre la variable et une moyenne divisée par une mesure des écarts constitue la variable réduite.

Les moments autour de l'espérance mathématique sont définis par

$$(5) \quad \mu_n = \int_{-\infty}^{+\infty} (z - \bar{x})^n w(z) dz.$$

La condition de l'aire donne

$$(5a) \quad \mu_0 = 1$$

et la définition de l'espérance mathématique

$$(5b) \quad \mu_1 = 0$$

ce qui montre que seuls les moments d'ordre $n > 1$ peuvent définir de nouvelles caractéristiques. Pour caractériser une distribution on introduit les quotients de PEARSON

$$(6a) \quad \rho_1 = \frac{\mu_3}{\sigma^3}$$

$$(6b) \quad \rho_2 - 3 = \frac{\mu_4 - 3\mu_2^2}{\mu_2^2}$$

dont le premier mesure pareillement l'asymétrie de la distribution considérée et le second sa différence avec une distribution de GAUSS.

Enfin on définit la fonction caractéristique $\varphi(t)$ d'une distribution $w(x)$, par l'intégrale

$$(7) \quad \varphi(t) = \int_{-\infty}^{+\infty} w(z)e^{(z-\bar{x})t} dz.$$

Quand on connaît cette fonction on peut en déduire tous les moments puisque :

$$(8) \quad \mu_n = \left(\frac{d^n \varphi(t)}{dt^n} \right)_{t=0}.$$

Le problème des valeurs extrêmes dont nous allons nous occuper exige la connaissance du comportement asymptotique des distributions. On doit avoir en vertu de la condition de l'aire

$$w(x) \rightarrow 0$$

pour les valeurs absolues croissantes de la variable. En outre, la dérivée tendra vers zéro et sera négative ou positive suivant que la variable elle-même sera respectivement positive ou négative. Pour des valeurs suffisamment élevées de la variable, la probabilité $W(x)$ diffère peu de l'unité si ces valeurs sont positives et de zéro si elles sont négatives. C'est pour cette raison que, souvent, on fait usage de la règle de L'HOPITAL, sous les deux formes suivantes :

$$(9') \quad \frac{w(x)}{1 - W(x)} = - \frac{w'(x)}{w(x)}$$

$$(9) \quad \frac{w(x)}{W(x)} = \frac{w'(x)}{w(x)}$$

dont la première (seconde) vaut pour les grandes valeurs positives (négatives) et dont on tire les valeurs asymptotiques de $W(x)$.

Pour des distributions symétriques ces deux relations sont identiques. Dans ce qui suit nous nous bornerons à envisager uniquement des distributions pour lesquelles les équations précédentes sont vérifiées soit presque toujours, soit asymptotiquement.

Le signe de la dérivée logarithmique qui s'y trouve sera négatif ou positif suivant que la variable sera elle-même positive ou négative. Sa valeur absolue peut augmenter, rester constante ou décroître, lorsque la variable croît en valeur absolue.

Pour les dérivées d'ordre supérieur on peut se demander si les relations analogues

$$(10') \quad \frac{w^{(v)}(x)}{w(x)} = (-1)^v \frac{w^v(x)}{(1 - W(x))^v},$$

$$('10) \quad \frac{w^{(v)}(x)}{w(x)} = \frac{w^v(x)}{W^v(x)},$$

sont vérifiées asymptotiquement pour des grandes valeurs absolues de la variable. Les équations (9') ('9) en sont des cas particuliers correspondant à $v = 1$. Les deuxièmes membres de ces équations sont des puissances d'ordre v des dérivées logarithmiques. Les équations (10) ne sont que les conséquences de (9). Posons pour le premier membre

$$\frac{w(x)}{1 - W(x)} = x$$

et négligeons les différences avec le second. Alors on obtient avec la même précision

$$\frac{w''(x)}{w(x)} = x^2.$$

En posant (10') pour v arbitraire on prouve facilement que cette relation vaut aussi pour la dérivée suivante. Par la même méthode on arrive à ('10) en partant de ('9).

Si la distribution initiale est exponentielle

$$w(x) = \frac{1}{2} e^{-|x|}$$

les conditions (10) sont remplies pour tout $x \neq 0$. Pour cette raison nous appellerons les distributions $w(x)$ qui vérifient ces équations, des *distributions du type exponentiel*.

Nous nous occuperons tout particulièrement de ce type de distribution, en imposant en outre aux distributions de ce type la condition que la valeur absolue de la dérivée logarithmique ne diminue jamais sans avoir une limite lorsque la variable croît.

Autrement dit : Pour une distribution du type exponentiel tous les moments existent, propriété qu'il est raisonnable de demander pour une distribution initiale.

I. — Les distributions des $m^{\text{ièmes}}$ valeurs et leurs dominantes

Quand on fait N observations ou quand on prend N échantillons sur une distribution initiale illimitée dans les deux sens, on peut classer les valeurs numériques qui constituent ces N observations par ordre de grandeur. Elles seront toutes finies. Il y en aura une qui sera *la plus grande*, la dernière valeur, x_1 , une autre immédiatement inférieure x_2 , que l'on peut considérer comme la plus petite parmi les deux plus grandes, en général une $m^{\text{ième}}$ qui sera la plus petite parmi les m plus grandes. Considérons ces valeurs en partant de l'autre extrémité de la suite. De ce point de vue, il en existe une valeur, x_N , la première, qui sera la plus petite (mais dont la valeur absolue peut être assez grande). En général il existe une $m^{\text{ième}}$ valeur d'*en bas* qui sera la plus grande parmi les m plus petites et qui sera désignée par

$$x_{N-m+1} = {}_m x.$$

La plus grande, la dernière valeur de x est une variable statistique, c'est-à-dire qu'il existe une probabilité $w_1(x_1, N)dx_1$ pour que cette valeur soit comprise entre x_1 et $x_1 + dx_1$.

Donc w_1 représente la distribution de la dernière valeur. La valeur \tilde{u}_1 de x_1 pour laquelle w_1 est maximum sera appelée *la dominante* de la dernière valeur. En outre, il existe une espérance mathématique \bar{u}_1 , un écart type σ_1 , et, généralement, un $n^{\text{ième}}$ moment $\mu_{n,1}$ de cette distribution.

De même, les $m^{\text{ièmes}}$ valeurs d'*en haut* et d'*en bas* sont des variables statistiques ayant des distributions w_m et ${}_m w$, des valeurs dominantes \tilde{u}_m et ${}_m \tilde{u}$, des espérances mathématiques \bar{u}_m et ${}_m \bar{u}$ et des $n^{\text{ièmes}}$ moments $\mu_{n,m}$ et ${}_m \mu_n$. Le champ de variation de toutes les variables x_m et ${}_m x$ est identique au champ de variation de la variable x de la

distribution initiale, puisque à toute valeur de la variable peut être attribuée une densité de probabilité pour qu'elle soit la $m^{\text{ième}}$. Voilà pourquoi il est légitime d'écrire les distributions des $m^{\text{ièmes}}$ valeurs dans la forme :

$${}_m\mathbf{w}(x, N) \quad \text{et} \quad \mathbf{w}_m(x, N).$$

On aura évidemment

$$\dots {}_m\tilde{u} < {}_{m-1}\tilde{u} < \dots < \tilde{u}_m < \tilde{u}_{m-1} \quad \text{pour} \quad m = 1, 2, 3 \dots$$

et la même relation vaut pour les espérances mathématiques.

Il s'agit d'abord de calculer ces dominantes et leurs probabilités et de trouver les relations qui existent entre ces distributions.

La distribution de la $m^{\text{ième}}$ valeur d'en haut est donnée par

$$(II') \quad \mathbf{w}_m(x, N) = \binom{N}{m} m W(x)^{N-m} (1 - W(x))^{m-1} w(x),$$

et la distribution de la $m^{\text{ième}}$ valeur d'en bas, par

$$(II) \quad \mathbf{w}_m(x, N) = \binom{N}{m} m W(x)^{m-1} (1 - W(x))^{N-m} w(x).$$

En effet, nous imposons la condition que parmi les résultats des N observations il y en ait un qui soit supérieur (ou inférieur) à $N - m$ autres résultats et inférieur (ou supérieur) à $m - 1$ autres.

Le nombre m peut varier de $m = 1$ jusqu'à $m = \frac{N}{2}$ ou jusqu'à $m = \frac{N+1}{2}$ suivant que le nombre des observations N est pair ou impair. Les distributions (II) sont valables pour n'importe quelle valeur de N et de m , mais les cas $N \gg 1$ et $m \ll \frac{N}{2}$ nous intéressent tout particulièrement du point de vue des valeurs extrêmes.

Pour les valeurs de N et de m qui nous intéressent, les distributions (II) constituent des systèmes de courbes illimitées dans les deux directions, mais placées à la fin et au commencement de la distribution initiale. Pour simplifier l'écriture nous laissons de côté la variable x , qui a des signes inverses dans les équations (II) lorsque l'espérance mathématique de la distribution initiale est nulle. Il est également inutile de répéter toujours la variable N . Nous écrivons donc tout simplement

$$\begin{aligned} \mathbf{w}_m(x, N) &= \mathbf{w}_m, \\ {}_m\mathbf{w}(x, N) &= {}_m\mathbf{w} \end{aligned}$$

Pour $m = 1$ on obtient la distribution connue de la dernière et de la première valeur

$$(I2') \quad \mathbf{w}_1 = N W^{N-1} w \quad \mathbf{w}_1 = N(1 - W)^{N-1} w \quad (I2)$$

$$(I3') \quad = \frac{dW^N}{dx} \quad = - \frac{d(1 - W)^N}{dx} \quad (I3)$$

Ce sont seulement ces deux distributions qui ont été traitées jusqu'à présent.

Les courbes représentant les distributions des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas se coupent toutes en un point qui correspond à la valeur médiane de la distribution initiale. Pour des distributions initiales symétriques par rapport à l'origine on aura la dominante

$${}_m \tilde{u} = - \tilde{u}_m,$$

les espérances mathématiques

$${}_m \bar{u} = - \bar{u}_m,$$

et enfin les distributions

$${}_m \mathbf{w}(-x) = \mathbf{w}_m(x).$$

Dans ce cas les valeurs absolues des dominantes et des espérances mathématiques des $m^{\text{ièmes}}$ valeurs, comptées dans un sens ou dans l'autre, sont égales, ainsi que leurs densités de probabilité. Il en résulte que les moments d'ordre pair seront égaux et les moments d'ordre impair différent seulement du signe.

Dans le cas général, nous allons nous limiter d'abord au calcul des dominantes. On les déduit des équations suivantes :

$$\mathbf{w}_m \cdot \left(\frac{N - m}{W} w - \frac{m - 1}{1 - W} w + \frac{w'}{w} \right) = 0,$$

$${}_m \mathbf{w} \cdot \left(\frac{m - 1}{W} w - \frac{N - m}{1 - W} w + \frac{w'}{w} \right) = 0,$$

lesquelles ont la forme d'équations différentielles, mais qui ne sont valables que pour une ou plusieurs valeurs de la variable. Écartons les solutions banales correspondant à des valeurs infinies ; il s'agit de trouver les valeurs finies, fonctions de m et de N , qui sont les racines de ces équations. Au lieu de chercher des solutions exactes nous nous limitons encore une fois aux grandes valeurs de N et aux valeurs de m telles que $m \ll N$. Si l'on peut employer vers la fin et

vers le commencement de la distribution initiale la règle de L'Hopital sous les formes (9), ces équations mènent à

$$(14') \quad W(\tilde{u}_m) = 1 - \frac{m}{N},$$

$$(14) \quad W_{(m)}(\tilde{u}) = \frac{m}{N}.$$

Les équations (9) sont des conditions dont on tire le nombre d'observations nécessaires pour que le calcul des dominantes à l'aide de (14) soit justifié. Ce nombre sera d'autant moins grand que la distribution initiale tendra plus vite vers zéro.

On peut expliquer ces résultats par un raisonnement trop simple pour être rigoureux. Soit \tilde{u}_m une grandeur observée, telle qu'il y en ait m autres supérieures ou égales. La probabilité d'une valeur supérieure ou égale à \tilde{u}_m , sera

$$1 - W(\tilde{u}_m).$$

Parmi N observations on doit s'attendre à ce que $N(1 - W(\tilde{u}_m))$ soient supérieures ou égales à \tilde{u}_m , nombre qui doit être lui-même égal à m , ce qui vérifie bien notre formule.

La dominante de la $m^{\text{ième}}$ valeur pour N observations ne dépendra d'après (14) que du quotient $\frac{m}{N}$. Donc elle sera égale à la dominante de la $k.m^{\text{ième}}$ valeur pour $k.N$ observations ($k = 1, 2, \dots$). On aura pour des valeurs de N suffisamment grandes

$$\tilde{u}_k(kN) = \tilde{u}_1(N); \quad {}_k\tilde{u}(kN) = {}_1\tilde{u}(N).$$

La dominante de la dixième valeur pour 100.000 observations sera égale à la dominante de la dernière valeur pour 10.000 observations. D'ailleurs, la manière dont la dominante dépend du quotient $\frac{m}{N}$ change suivant les différentes distributions initiales.

Il est évident que la dominante de la $m^{\text{ième}}$ valeur d'en haut (ou d'en bas) augmente (ou diminue) avec le nombre d'observations. En effet, on obtient en dérivant (14) par rapport à N

$$(15') \quad \frac{d\tilde{u}_m}{dN} = \frac{m}{N^2 w(\tilde{u}_m)},$$

$$(15) \quad \frac{d_m \tilde{u}}{dN} = - \frac{m}{N^2 w_{(m)}(\tilde{u})}.$$

Donc, la valeur absolue de la dominante de la $m^{\text{ième}}$ valeur augmente avec le nombre d'observations.

Il sera facile de compléter ce résultat par une règle générale qui servira pour évaluer l'ordre de grandeur des dominantes. Posons, pour les valeurs absolues des dérivées logarithmiques de la distribution initiale aux $m^{\text{ièmes}}$ valeurs extrêmes dominantes :

$$\frac{Nw(\tilde{u}_m)}{m} = \alpha_m; \quad \frac{Nw({}_m\tilde{u})}{m} = {}_m\alpha,$$

valeurs qui ont la dimension réciproque de la variable. Si les grandeurs α sont indépendantes de N ou tendent vers une constante, soit en augmentant, soit en diminuant, pour des N croissants, c'est-à-dire pour des valeurs croissantes de la variable, les $m^{\text{ièmes}}$ valeurs dominantes seront fonctions linéaires du logarithme du nombre d'observations, que l'on peut écrire à l'aide de la variable réduite.

$$(I4 a') \quad \tilde{u}_m = \bar{x} + ks \lg \frac{N}{m}$$

$$({}'I4 a) \quad {}_m\tilde{u} = \bar{x} - ks \lg \frac{N}{m},$$

où k est une constante et s l'écart type de la distribution initiale. En effet, pour une seule observation la plus grande valeur est identique à l'espérance mathématique et pour N observations il faut ajouter un multiple d'une mesure des écarts.

Si au contraire les valeurs α augmentent sans limite avec N , les $m^{\text{ièmes}}$ valeurs dominantes augmenteront plus lentement que le logarithme du nombre d'observations

$$(I4 b') \quad \tilde{u}_m = \bar{x} + ksf \left(\lg \frac{N}{m} \right)$$

$$({}'I4 b) \quad {}_m\tilde{u} = \bar{x} - ksf \left(\lg \frac{N}{m} \right).$$

Le traitement des $m^{\text{ièmes}}$ valeurs n'est donc légitime que pour des distributions initiales, telles que les deux premiers moments ne divergent pas, condition que nous avons posée dès le commencement et qui exclut le cas où les α diminueraient sans limite.

Quant à la manière dont les dominantes dépendent de m , on obtient

en dérivant (I4) par rapport à m (procédé légitime pour de grandes valeurs de m , qui pourtant doivent être petites par rapport à N)

$$(I6') \quad \frac{d\tilde{u}_m}{dm} = - \frac{1}{Nw(\tilde{u}_m)}$$

$$('I6) \quad \frac{d_m \tilde{u}}{dm} = \frac{1}{Nw(m\tilde{u})}.$$

Les signes s'expliquent aisément ; en effet plus m augmente, plus la dominante de la $m^{\text{ième}}$ valeur d'en haut diminue et plus celle de la $m^{\text{ième}}$ valeur d'en bas augmente. Compte tenu de (II) et (I4), les densités de probabilité des dominantes deviennent, pour les valeurs élevées de N et des petites valeurs de m :

$$(I7') \quad \mathbf{w}_m(\tilde{u}_m) = N\psi(m)w(\tilde{u}_m)$$

$$('I7) \quad {}_m\mathbf{w}(m\tilde{u}) = N\psi(m)w({}_m\tilde{u})$$

où

$$(I7 a) \quad \psi(m) = \frac{m^m e^{-m}}{m!}$$

est la probabilité de l'espérance mathématique dans la loi des événements rares. Pour des grandes valeurs de m , les densités de probabilité des dominantes tendront vers

$$(I7 b') \quad \mathbf{w}_m(\tilde{u}_m) = \frac{Nw(\tilde{u}_m)}{\sqrt{2\pi m}}$$

$$('I7 b) \quad {}_m\mathbf{w}(m\tilde{u}) = \frac{Nw({}_m\tilde{u})}{\sqrt{2\pi m}}.$$

La densité de probabilité des dominantes augmentera avec m . En effet d'après (I6) et (I7) on aura

$$\frac{d\mathbf{w}_m(\tilde{u}_m)}{dm} = - \mathbf{w}_m(\tilde{u}_m) \left(\frac{1}{2m} + \frac{w'(\tilde{u}_m)}{Nw^2(\tilde{u}_m)} \right)$$

$$\frac{d_m \mathbf{w}(m\tilde{u})}{dm} = {}_m\mathbf{w}(m\tilde{u}) \left(- \frac{1}{2m} + \frac{w'({}_m\tilde{u})}{Nw^2({}_m\tilde{u})} \right).$$

En introduisant la dominante par (I4) dans (9), la règle de L'Hopital devient pour la valeur dominante

$$(I8') \quad x_m \equiv \frac{Nw(\tilde{u}_m)}{m} = - \frac{w'(\tilde{u}_m)}{w(\tilde{u}_m)}$$

$$('I8) \quad m^x \equiv \frac{Nw({}_m\tilde{u})}{m} = \frac{w'({}_m\tilde{u})}{w({}_m\tilde{u})}.$$

Donc les formules précédentes deviennent

$$(19) \quad \frac{d\mathbf{w}_m(\tilde{u}_m)}{dm} = \frac{\mathbf{w}_m(\tilde{u}_m)}{2m}$$

$$('19) \quad \frac{d_m \mathbf{w}(m\tilde{u})}{dm} = \frac{{}_m \mathbf{w}(m\tilde{u})}{2m}.$$

Pour des valeurs croissantes de m , les distributions des $m^{\text{ième}}$ valeurs se resserrent autour des dominantes. Le champ de variation de la $m^{\text{ième}}$ valeur sera plus petit que celui de la première ou dernière valeur. L'espérance mathématique de la $m^{\text{ième}}$ valeur tendra vers la dominante et l'écart type diminuera pour des valeurs croissantes de m .

On ne peut pas établir sans imposer aucune restriction la même relation entre la densité de probabilité de la dominante et le nombre d'observations. En effet, en introduisant (18) dans (17) on obtient

$$(20') \quad \mathbf{w}_m(\tilde{u}_m) = - \frac{w'(\tilde{u}_m)}{w(\tilde{u}_m)} m\psi(m)$$

$$('20) \quad {}_m \mathbf{w}(m\tilde{u}) = \frac{w'(m\tilde{u})}{w(m\tilde{u})} m\psi(m).$$

ou d'après les valeurs α définies par (18)

$$(20 a') \quad \mathbf{w}_m(\tilde{u}_m) = \alpha_m m\psi(m)$$

$$('20 a) \quad {}_m \mathbf{w}(m\tilde{u}) = {}_m \alpha m\psi(m).$$

Dans ces relations seulement le premier facteur du second membre, la dérivée logarithmique de la valeur dominante, dépendra de N . La densité de probabilité des dominantes augmentera avec le nombre d'observations si la valeur absolue de cette dérivée logarithmique augmente avec la variable. Dans ce cas la distribution de la $m^{\text{ième}}$ valeur se reserre pour des nombres croissants. Si la dérivée logarithmique est constante, c'est-à-dire si l'on est dans le cas d'une distribution initiale exponentielle, les densités de probabilité des dominantes seront indépendantes du nombre d'observations. Enfin, si la grandeur absolue de la dérivée logarithmique diminue pour des valeurs croissantes de la variable, les densités de probabilité de la dominante diminueront et la distribution s'élargira. Ce sont seulement les deux premiers cas qui nous intéressent du point de vue pratique puisque pour le troisième la précision de la $m^{\text{ième}}$ valeur diminuera avec un nombre d'observations croissant.

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

Du point de vue théorique, il est intéressant de noter que M. FRÉCHET (10) a construit une distribution initiale d'une variable positive pour laquelle la valeur absolue de la dérivée logarithmique diminue sans limite. Les moments d'ordre supérieur de cette distribution divergent. La dominante de la dernière valeur augmente plus vite que le logarithme du nombre d'observations. Enfin, la distribution de la dernière valeur s'étale sans cesse pour un nombre croissant d'observations. Plus on fait d'observations, moins précis sera le résultat. C'est pour éviter de telles catastrophes que nous nous bornons au traitement des distributions initiales du type exponentiel.

Il existe une relation intéressante entre les densités de probabilité correspondant aux dominantes des $m^{\text{ième}}$ s valeurs d'en haut et d'en bas, à savoir :

$$(21) \quad \frac{w_m(\tilde{u}_m)}{{}_m w({}_m \tilde{u})} = \frac{w(\tilde{u}_m)}{w({}_m \tilde{u})}.$$

Le quotient de la densité de probabilité correspondant à la dominante de la $m^{\text{ième}}$ valeur d'en haut, par la densité de probabilité de la dominante de la $m^{\text{ième}}$ valeur d'en bas est égal au quotient des densités de probabilité de ces valeurs elles-mêmes dans la distribution initiale. On peut établir une relation analogue pour les valeurs médianes de la dernière et de la première valeur. D'après la définition (1a) celles-ci sont données par

$$\begin{aligned} W(\tilde{u}_1) &= 2^{-\frac{1}{N}} & W({}_1 \tilde{u}) &= 1 - 2^{-\frac{1}{N}} \\ &= 1 - \frac{1}{N} \lg 2 & &= \frac{1}{N} \lg 2 \end{aligned}$$

pourvu que N soit suffisamment grand. La médiane de la dernière (ou de la première) valeur sera un peu supérieure (ou inférieure) à la dominante, ces différences disparaissant d'ailleurs lorsque le nombre d'observations croît. En introduisant ces valeurs dans les distributions (12) on aura, par analogie avec (21)

$$(21a) \quad \frac{w_1(\tilde{u}_1)}{{}_1 w({}_1 \tilde{u})} = \frac{w(\tilde{u}_1)}{w({}_1 \tilde{u})}.$$

Au lieu de traiter les médianes des $m^{\text{ièmes}}$ valeurs nous allons envisager plus tard les propriétés générales des moyennes des distributions (11) pour de grands nombres N.

Jusqu'à présent nous n'avons envisagé que la dominante et sa densité de probabilité, ainsi que la question de savoir comment ces grandeurs dépendent des nombres N et m . Il nous reste à voir comment la distribution de la $m^{\text{ième}}$ valeur varie en fonction de x .

Comparons entre elles, les densités de probabilité w_{m+1} et w_m d'un côté, ${}_{m+1}w$ et ${}_m w$ de l'autre pour toutes les valeurs de la variable. En appliquant (14) aux équations (11), on a

$$\frac{w_{m+1}}{w_m} = \left(\frac{1}{W} - 1 \right) : \left(\frac{1}{W(\tilde{u}_m)} - 1 \right)$$

$$\frac{{}_m w}{{}_{m+1} w} = \left(\frac{1}{W} - 1 \right) : \left(\frac{1}{W({}_m \tilde{u})} - 1 \right).$$

Mais W est une fonction toujours croissante de x . Donc pour $x \begin{matrix} \leq \\ \geq \end{matrix} \tilde{u}_m$ on aura

(22')
$$w_{m+1} \begin{matrix} \leq \\ \geq \end{matrix} w_m$$

et pour

$$x \begin{matrix} \leq \\ \geq \end{matrix} {}_m \tilde{u},$$

(22)
$${}_{m+1} w \begin{matrix} \leq \\ \geq \end{matrix} {}_m w.$$

Pour les valeurs a) *antérieures* et b) *postérieures* de la dominante, la densité de probabilité de la $m^{\text{ième}}$ valeur d'en haut est respectivement a) *moindre* ou b) *plus grande* que la densité de probabilité de la $m+1^{\text{ième}}$ valeur. Pour les valeurs a) *antérieures* et b) *postérieures* de la dominante de la $m^{\text{ième}}$ valeur d'en bas, la densité de probabilité pour que la valeur x soit la $m^{\text{ième}}$ est respectivement a) *plus grande* et b) *moindre* que la densité de probabilité pour que cette valeur soit la $m+1^{\text{ième}}$. Le sens des égalités (22) est très clair. La courbe de la distribution w_{m+1} coupe celle de la distribution postérieure w_m en un point qui correspond précisément à sa dominante. La courbe de la distribution ${}_{m+1} w$ coupe la courbe de la distribution antérieure ${}_m w$ au point représentant sa dominante. Plus particulièrement : la densité de probabilité de la dominante de la dernière valeur (la plus grande) est égale à la densité de probabilité pour que cette valeur soit aussi la seconde des plus grandes valeurs. Et la densité de probabilité de la dominante de la première valeur (qui est la plus petite) est égale à la densité de probabilité pour que cette valeur soit également la seconde.

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

Ces relations sont valables pour des valeurs de N constantes et des valeurs variables de m ; on peut cependant les établir également dans le cas de m constant et N variable. On a en effet en vertu de (11) :

$$w_m(x, N + 1) = \frac{N + 1}{N + 1 - m} W w_m(x, N)$$

d'où l'on tire

$$(23') \quad w_m(x, N + 1) \underset{\leq}{\geq} w_m(x, N)$$

si

$$W(x) \underset{\leq}{\geq} 1 - \frac{m}{N + 1},$$

ce qui pour des valeurs suffisamment grandes de N équivaut d'après (14) à la condition

$$x \underset{\leq}{\geq} \tilde{u}_m.$$

De même, on trouve

$$(23) \quad {}_m w(x, N + 1) \underset{\leq}{\geq} {}_m w(x, N)$$

si

$$x \underset{\leq}{\geq} {}_m \tilde{u}.$$

Pour des valeurs a) *antérieures* et b) *postérieures* de la dominante de la $m^{\text{ième}}$ valeur d'en haut, la densité de probabilité pour qu'une valeur x soit la $m^{\text{ième}}$ pour N observations est a) *plus grande* et b) *moindre* que la densité de probabilité pour qu'elle le soit pour $N + 1$ observations. Pour des valeurs a) *antérieures* et b) *postérieures* de la dominante de la $m^{\text{ième}}$ valeur d'en bas, la densité de probabilité pour qu'une valeur x soit la $m^{\text{ième}}$ pour N observations, est a) *moindre* et b) *plus grande* que la densité de probabilité pour qu'elle le soit pour $N + 1$ observations. La signification des égalités (23) est simple : La courbe de la distribution de la $m^{\text{ième}}$ valeur d'en haut ou d'en bas pour $N + 1$ observations coupe la courbe de la distribution de la $m^{\text{ième}}$ valeur d'en haut (ou d'en bas) pour N observations en un point qui correspond à sa valeur dominante.

Notre raisonnement nous a montré jusqu'ici, quelles sont les distributions des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas, leurs dominantes et leurs relations pour des valeurs consécutives de m et de N . Les

médianes s'approchent des dominantes pour les distributions de la première et de la dernière valeur. Le calcul de l'espérance mathématique d'après les formules (11) est trop compliqué, on peut cependant affirmer en tout cas que si les distributions des $m^{\text{ièmes}}$ valeurs se resserrent pour un nombre d'observations croissant, l'espérance mathématique tendra vers la dominante. En première approximation on peut admettre que ces deux valeurs sont égales, si la grandeur absolue de la dérivée logarithmique augmente avec la variable. Ce théorème assez important qui dérive de (20) a été démontré pour le cas particulier $m = 1$ par M. von MISES (30).

Celui-ci a montré que l'espérance mathématique de la dernière valeur tend vers la dominante, pourvu qu'on ait

$$\frac{1 - W(x+k)}{1 - W(x)} \rightarrow 0.$$

Lorsque k est constant et positif, cette condition peut s'écrire :

$$(24) \quad \frac{w(x+k)}{w(x)} \rightarrow 0.$$

Elle exige que la distribution initiale tende vers zéro plus vite qu'une exponentielle ; pour un nombre croissant d'observations la distribution se resserre alors autour de la dominante. Ces deux propriétés peuvent être séparées : si une distribution se resserre, l'espérance mathématique tend vers la dominante, mais ces deux valeurs peuvent se rapprocher même si la distribution ne se resserre pas.

Si la distribution se resserre, l'écart type tendra vers zéro. Il s'agit maintenant de préciser la manière dont cet écart tend vers sa limite pour savoir ce que devient la différence entre l'espérance mathématique et la dominante divisée par l'écart type. Pour évaluer approximativement l'écart type correspondant à la distribution de la première et de la dernière valeur nous avons fait usage (23) de cette égalité entre les trois moyennes pour admettre que dans le cas d'une distribution initiale de GAUSS les distributions de la dernière et de la première valeur peuvent être approchées également par une distribution de GAUSS. Nous avons employé (15) la même méthode pour obtenir l'ordre de grandeur de l'écart type de la distribution de la $m^{\text{ième}}$ valeur. Une telle approximation cependant ne suffit pas.

II. — Distributions finales

Il s'agit de calculer l'espérance mathématique, l'écart type et les moments plus élevés des distributions de la $m^{\text{ième}}$ valeur. Dans ce but nous allons d'abord donner à ces distributions une forme plus simple qu'elles ont dans le cas d'un grand nombre d'observations, et qui se prête mieux au calcul.

Le problème consiste à voir si effectivement les distributions

$$w_m(x, N) \quad \text{et} \quad w_m(x, N)$$

des $m^{\text{ièmes}}$ valeurs tendent vers une forme bien définie lorsque le nombre d'observations croît. Dans l'affirmative cette forme sera appelée la *distribution finale*.

Les conditions analytiques du problème peuvent être écrites

$$\frac{N}{m} \rightarrow \infty; \quad w(x) \rightarrow 0,$$

d'où l'on tire

$$W(x) \rightarrow 1.$$

pour les $m^{\text{ièmes}}$ valeurs d'en haut et

$$W(x) \rightarrow 0,$$

pour les $m^{\text{ièmes}}$ valeurs d'en bas.

Jusqu'à un certain point ce problème est analogue à celui traité par POISSON et BORTKIEWICZ, lesquels se proposaient de chercher vers quelle forme tend la distribution de BERNOULLI, lorsque le nombre d'observations augmente indéfiniment, la probabilité initiale tendant vers zéro. Les deux problèmes présentent des analogies mais aussi des différences. Dans le problème de POISSON la variable reste discontinue puisqu'elle représente un nombre de cas. Ici, elle reste continue, puisque c'est une valeur de la variable initiale. L'analogie est évidente pour $m = 1$; en effet dans ces cas les deux problèmes traitent des alternatives, tandis que l'introduction de $m > 1$ crée trois cas distincts. Aussi notre distribution initiale dépend-elle de la variable, tandis que la probabilité initiale dans le problème de POISSON en est indépendante. Voilà pourquoi il faut s'attendre à ce que l'existence d'une distribution finale soit liée à des conditions analytiques concer-

nant la nature de la distribution initiale, circonstance qui n'existe pas dans le problème de POISSON. En tout cas il existe une $m^{\text{ième}}$ valeur remarquable, à savoir la dominante. En effet, nous avons trouvé que sa densité de probabilité est proportionnelle à la probabilité de la dominante dans la loi des événements rares.

Envisageons d'abord la distribution $w_m(x, N)$ de la $m^{\text{ième}}$ valeur d'en haut. D'après (11') elle peut s'écrire :

$$(25') \quad w_m = \binom{N-1}{m-1} \left(\frac{1}{W} - 1 \right)^{m-1} \frac{dW^N}{dx},$$

formule dont (12') est un cas particulier correspondant à $m = 1$.

Développons d'abord $W(x)$ au voisinage de la dominante \tilde{u}_m . On aura

$$(26') \quad W(x) = W(\tilde{u}_m) + \frac{x - \tilde{u}_m}{1!} w(\tilde{u}_m) + \frac{(x - \tilde{u}_m)^2}{2!} w'(\tilde{u}_m) + \dots \\ = 1 - \frac{m}{N} + \frac{m(x - \tilde{u}_m)N}{N \cdot 1! \cdot m} w(\tilde{u}_m) - \frac{m(x - \tilde{u}_m)^2 N^2}{N \cdot 2! \cdot m^2} w^2(\tilde{u}_m) + \dots$$

d'après (14') et (9'). Les termes écrits sont les trois premiers termes du développement de

$$(27') \quad V_1(x) = 1 - \frac{m}{N} e^{-(x - \tilde{u}_m) \frac{N}{m} w(\tilde{u}_m)}.$$

Il est légitime de poser $W(x) = V_1(x)$ pourvu qu'on puisse écrire approximativement

$$(28') \quad w^{(v)}(\tilde{u}_m) = (-1)^v \frac{N^v}{m^v} w^{v+1}(\tilde{u}_m) \\ = (-1)^v \frac{w^{v+1}(\tilde{u}_m)}{(1 - W(\tilde{u}_m))^v}$$

d'après (14'). Cette condition signifie que la distribution initiale est du type exponentiel défini dans le premier chapitre. En égalant les deux expressions (26') et (27') et en introduisant une nouvelle variable y_m par la *transformation dite finale* des $m^{\text{ièmes}}$ valeurs d'en haut

$$(29') \quad y_m = \frac{N}{m} w(\tilde{u}_m) (x - \tilde{u}_m)$$

on aura approximativement

$$W^N(x) = e^{-me^{-y_m}}$$

et

$$(30') \quad \frac{dW^N(x)}{dx} = Nw(\hat{u}_m)e^{-y_m - e^{-y_m}}.$$

Pour $m = 1$, on obtient d'après (12') la distribution finale de la plus grande valeur

$$(31') \quad w_1 = Nw(\hat{u}_1)e^{-y_1 - e^{-y_1}}.$$

La probabilité pour que la dernière valeur soit inférieure à x sera

$$(32') \quad \mathfrak{W}_1 = e^{-e^{-y_1}},$$

formule déduite par R. A. FISHER (8).

Revenons au cas général $m > 1$. Nous connaissons la forme finale du dernier facteur de (25'). En appliquant le même développement au deuxième facteur on arrive, pour des valeurs suffisamment grandes de N , à la relation

$$\begin{aligned} \left(\frac{1}{W} - 1\right)^{m-1} &= \left(\frac{1}{1 - \frac{m}{N}e^{-y_m}} - 1\right)^{m-1}. \\ &= \frac{m^{m-1}}{N^{m-1}} e^{-(m-1)y_m}. \end{aligned}$$

Mais pour ces valeurs, on a, d'après la formule de Stirling

$$\frac{m^{m-1}}{N^{m-1}} \binom{N-1}{m-1} = \frac{m^m}{m!}$$

Donc, la forme finale de la distribution (11') de la $m^{\text{ième}}$ valeur d'en haut sera

$$(33') \quad w_m = \frac{m^m}{(m-1)!} \frac{N}{m} w(\hat{u}_m) e^{-my_m - me^{-y_m}}.$$

Pour un nombre suffisamment grand d'observations la distribution de la $m^{\text{ième}}$ valeur d'en haut peut être représentée par la formule doublement exponentielle (33'), pourvu que la distribution initiale soit elle-même du type exponentiel. La formule devient rigoureuse si la distribution initiale est exponentielle.

Cette forme finale peut être employée pour des distributions qui ne sont pas soumises à la condition (24) et qui par conséquent ne se resserrent pas pour un nombre croissant d'observations.

Pour $m = 1$ la somme (32') a déjà été envisagée par M. R. A. FISHER (8) ; cet auteur part cependant d'un point de vue opposé à celui que nous avons adopté. Nous construisons la distribution finale de la plus grande valeur pour une distribution initiale donnée et pour cela nous employons le procédé classique qui consiste à partir d'une distribution où figure le nombre N d'observations et à chercher la forme à laquelle on arrive si N augmente indéfiniment. M. FISHER au contraire cherche une distribution initiale. Il veut que la probabilité de la plus grande valeur, d'être inférieure à x pour N observations, soit égale à la probabilité initiale d'une valeur inférieure à une fonction linéaire de x . En d'autres termes, il cherche les distributions initiales qui satisfont à l'équation

$$(34) \quad W^N(x) = W(a_N x + b_N).$$

Les solutions de cette équation sont considérées comme des distributions finales. Pour $a_N = 1$, c'est-à-dire pour une distribution illimitée dans les deux sens, on trouve la somme (32'). Mais pour $a_N \neq 1$, c'est-à-dire pour des distributions limitées soit à gauche, soit à droite la distribution finale, appelée « penultimate form » diffère de notre solution, ce qui n'est pas étonnant puisque notre construction de la distribution finale est tout à fait différente du postulat de M. R. A. FISHER. D'ailleurs cette deuxième distribution finale est identique à celle trouvée par M. FRÉCHET (10).

Nous allons déduire maintenant la forme finale de la distribution de la $m^{\text{ième}}$ valeur d'en bas par les procédés employés dans le cas de la $m^{\text{ième}}$ valeur d'en haut. En effet, cette distribution peut être écrite

$$('25) \quad {}_m w = - \binom{N-1}{m-1} \left(\frac{W}{1-W} \right)^{m-1} \frac{d}{dx} (1-W)^N.$$

formule pour laquelle ('12) est le cas spécial $m = 1$.

Développons W au voisinage de la dominante qui cette fois est $m\tilde{u}$; on aura

$$('26) \quad W(x) = \frac{m}{N} + \frac{m}{N} \frac{(x - m\tilde{u})}{1} \frac{N}{m} w_{(m\tilde{u})} + \frac{m}{N} \frac{(x - m\tilde{u})^2}{2!} \frac{N^2}{m^2} w_{(m\tilde{u})}^2 + \dots$$

d'après ('14) et ('9). Les termes écrits sont les trois premiers termes du développement de

$$('27) \quad {}_1 V(x) = \frac{m}{N} e^{(x - m\tilde{u}) \frac{N}{m} w_{(m\tilde{u})}}.$$

Il est légitime d'égaliser ces deux expressions pourvu qu'on puisse poser approximativement

$$('28) \quad w^{(v)}(m\tilde{u}) = \frac{w^{v+1}(m\tilde{u})}{W^v(m\tilde{u})}.$$

Donc nous imposons de nouveau à la distribution initiale un caractère du type exponentiel.

En introduisant une nouvelle variable my par la *transformation finale* des $m^{\text{ièmes}}$ valeurs d'en bas

$$('29) \quad my = \frac{N}{m} w(m\tilde{u})(x - m\tilde{u}),$$

on aura approximativement

$$(1 - W(x))^N = e^{-me^{my}},$$

et

$$('30) \quad -\frac{d}{dx}(1 - W(x))^N = Nw(m\tilde{u})e^{my} - me^{my}.$$

Pour $m = 1$ on obtient, d'après ('12), la distribution finale de la première valeur

$$('31) \quad {}_1w = Nw({}_1\tilde{u})e^{1y} - e^{1y}.$$

La probabilité pour que la première valeur soit inférieure à x sera

$$('32) \quad {}_1\mathfrak{B} = 1 - e^{-e^{1y}}.$$

En appliquant la même approximation aux premiers membres de ('25) on obtient

$$\binom{N-1}{m-1} \left(\frac{W}{1-W}\right)^{m-1} = \frac{m^{m-1}}{(m-1)!} e^{(m-1)my},$$

puisque W est très petit vis-à-vis de 1.

Donc, la forme finale de la distribution ('11) de la $m^{\text{ième}}$ valeur d'en bas sera

$$('33) \quad {}_m w = \frac{m^m}{(m-1)!} \frac{N}{m} w(m\tilde{u}) e^{mmy} - me^{my}.$$

Toutes les formules employées pour la construction de la distribution finale de la $m^{\text{ième}}$ valeur d'en bas sont strictement analogues à celles employées pour la $m^{\text{ième}}$ valeur d'en haut. L'analogie entre les distributions des $m^{\text{ièmes}}$ valeurs d'en haut ('11') et d'en bas ('11) est donc conservée par la transformation finale.

Les distributions finales (33) des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas sont asymétriques. Les unes descendent à gauche plus vite qu'à droite, les autres à droite plus vite qu'à gauche. Quand on remplace les distributions (11), valables pour toutes valeurs de N et m , par les formes finales valables pour des grands N et des petits m , les dominantes et leurs probabilités (17) et (20), de même que les relations (21) sont conservées ; en effet, on les calcule en utilisant les mêmes grandeurs. Cependant les points d'intersection des courbes successives

$$w_{m+1}(x, N); w_m(x, N) \quad \text{et} \quad w_m(x, N); w_m(x, N + 1),$$

n'auront plus la forme simple et intéressante qui existe pour les distributions exactes.

Dans ces distributions finales des $m^{\text{ièmes}}$ valeurs, la distribution initiale ne subsiste plus de façon explicite. Nous arriverons donc à la même distribution finale en partant de différentes distributions initiales, pourvu que ces distributions soient toutes du type exponentiel. Le nombre d'observations N et la distribution initiale $w(x)$ n'entrent que dans la combinaison

$$(35') \quad \frac{N}{m} w(\tilde{u}_m) = x_m$$

et

$$(35) \quad \frac{N}{m} w(\tilde{u}_m) = m\alpha.$$

Les grandeurs x_m et $m\alpha$ sont identiques aux dérivées logarithmiques de la distribution initiale pour les valeurs dominantes. Ces dérivées introduites dans les équations (18), auront des valeurs différentes pour différentes distributions initiales et dépendront en général de N et de m . Elles décident si la distribution finale de $m^{\text{ième}}$ valeur se resserre ou non pour un nombre croissant d'observations.

Les tableaux I et II contiennent les distributions finales des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas pour $m = 1, 2, 3, 4, 5$ en fonction des variables finales y_m et $m\gamma$, ainsi que les probabilités pour la première et la dernière valeur.

Le calcul de ces valeurs est des plus simples, puisque les courbes sont contenues dans des intervalles assez étroits qui se resserrent très vite pour des valeurs croissantes de m . La figure montre l'allure de ces distributions ; pour plus de simplicité nous les avons tracées à partir d'une origine commune.

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

TABLEAU I. — *La distribution finale de la dernière et de la première valeur.*

y	$\mathfrak{B}_1(y_1) = 1 - \mathfrak{B}(-y_1)$	$\frac{w(y_1)}{x_1} = \frac{w_1(-y_1)}{1^2}$	Δ
— 2,5	—	0,00006	0,00451
— 2,0	0,00062	0,00457	0,04614
— 1,5	0,01132	0,05071	0,12865
— 1,0	0,06599	0,17936	0,13769
— 0,5	0,19230	0,31705	0,05083
0,0	0,36788	0,36788	— 0,03717
0,5	0,54525	0,33071	— 0,07606
1,0	0,69220	0,25465	— 0,07614
1,5	0,80003	0,17851	— 0,06030
2,0	0,87345	0,11821	— 0,04259
2,5	0,92120	0,07562	— 0,02731
3,0	0,95143	0,04831	— 0,01901
3,5	0,97025	0,02930	— 0,01132
4,0	0,98185	0,01798	— 0,00699
4,5	0,98895	0,01099	— 0,00429
5,0	0,99328	0,00670	— 0,00263
5,5	0,99592	0,00407	— 0,00160
6,0	0,99752	0,00247	— 0,00097
6,5	0,99850	0,00150	— 0,00059
7,0	0,99909	0,00091	— 0,00036
7,5	0,99945	0,00055	— 0,00022
8,0	0,99967	0,00033	— 0,00013
8,5	0,99980	0,00020	— 0,00007
9,0	0,99987	0,00013	— 0,00006
9,5	0,99993	0,00007	— 0,00003
10,0	0,99996	0,00004	

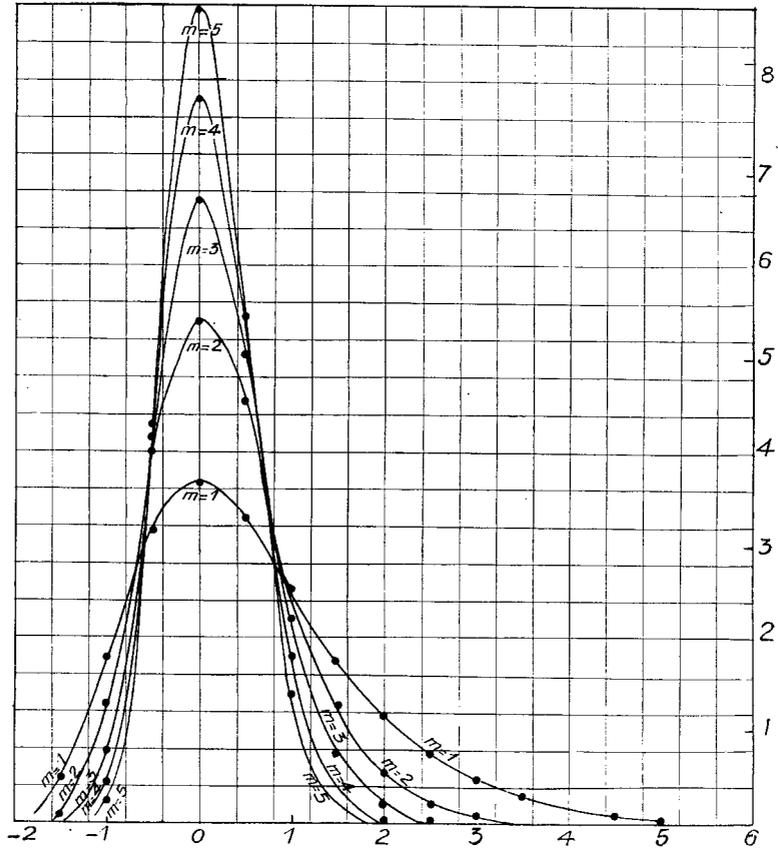
Le tableau I permet le calcul de l'écart probable ρ , défini par (2a).
La transformation finale conduit à

$$\begin{aligned} 2x_0 &= y'' - y' \\ &= 1,57253. \end{aligned}$$

Il y a donc la même probabilité (égale à 1/2) pour que la plus grande et la plus petite valeur soient contenues dans l'intervalle

$$(36) \quad \tilde{u} - \frac{0,32663}{x} < x < \tilde{u} + \frac{1,24590}{x},$$

ou qu'elle le surpasse. Si la valeur absolue de la dérivée logarithmique de la distribution initiale augmente avec la variable, cet écart probable diminuera avec le nombre croissant d'observations. Il faut ajouter l'indice τ aux valeurs \tilde{u} et α , à droite ou à gauche suivant qu'il s'agit de la distribution de la dernière ou de la première valeur.



Les distributions finales des $m^{\text{ièmes}}$ valeurs extrêmes.

La distribution finale de la $m^{\text{ième}}$ valeur d'en bas peut être déduite de la distribution finale de la $m^{\text{ième}}$ valeur d'en haut en remplaçant y_m par $-my$ et \tilde{u}_m par $m\tilde{u}$, circonstance qu'on ne rencontre pour les distributions exactes (11) que pour des distributions initiales symétriques par rapport à l'origine. Alors, les distributions finales des

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

$m^{\text{ièmes}}$ valeurs d'en haut et d'en bas sont symétriques entre eux par rapport à l'axe des ordonnées. Dans le cas général, les distributions finales des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas autour de leurs valeurs dominantes \tilde{u}_m et $m\tilde{u}$, peuvent être condensées d'après les formules (33), en une formule unique

$$(37) \quad w = \frac{m^m}{(m-1)!} x e^{\mp my - m e^{\mp y}}.$$

TABLEAU II. — Les distributions finales.

$\frac{w_m(y_m)}{x_m} = \frac{m w(-my)}{m^x}$				
y	$m = 2$	$m = 3$	$m = 4$	$m = 5$
— 2	0,00008	—	—	—
— 1,5	0,01028	0,00166	0,00028	0,00004
— 1	0,12871	0,07791	0,04418	0,02418
— 0,5	0,40208	0,43024	0,43112	0,41713
0	0,54136	0,67212	0,78147	0,87755
0,5	0,45430	0,50707	0,54803	0,55310
1	0,25939	0,22293	0,17942	0,13943
1,5	0,12746	0,07679	0,04332	0,02360
2	0,05589	0,02230	0,00833	0,00301
2,5	0,02287	0,00584	0,00139	0,00032
3	0,00934	0,00152	0,00023	0,00003
3,5	0,00343	0,00034	0,00003	—
4	0,00128	0,00008	—	—
4,5	0,00048	0,00002	—	—
5	0,00018	—	—	—
5,5	0,00007	—	—	—
6	0,00002	—	—	—

On emploie pour cela le tableau suivant qui donne la correspondance des notations.

Cette écriture uniforme a l'avantage de nous permettre de traiter par la même formule les distributions des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas, ce qui facilite grandement les calculs.

La question de savoir vers quelle forme tend la distribution de la

$m^{\text{ième}}$ valeur pour un nombre croissant d'observations par conséquent, est résolue, pourvu que la distribution initiale soit du type exponentiel.

TABLEAU III. — Distributions finales des $m^{\text{ièmes}}$ valeurs.

	Notation uniforme	Correspondant à la $m^{\text{ième}}$ valeur	
		d'en haut	d'en bas
1	Distribution $w \dots$	w_m	${}_m w$
2	Dominante $\tilde{u} \dots$	\tilde{u}_m	${}_m \tilde{u}$
3	Constante $\alpha \dots$	$\alpha_m = \frac{N}{m} w(\tilde{u}_m)$	${}_m \alpha = \frac{N}{m} w({}_m \tilde{u})$
4	Variable finale $y \dots$	$y_m = x_m(x - \tilde{u}_m)$	${}_m y = {}_m \alpha(x - {}_m \tilde{u})$
5	Signe.....	—	+

III. — Espérance mathématique et moments

Les formes finales des distributions de la $m^{\text{ième}}$ valeur d'en haut et d'en bas permettent le calcul de l'espérance mathématique et des moments, et cela d'une façon très simple, tandis que ce calcul pour les distributions elles-mêmes serait des plus laborieux.

La probabilité pour que la $m^{\text{ième}}$ valeur soit comprise entre x et $x + dx$ est

$$(38) \quad w_m dx = \frac{m^m}{(m-1)!} e^{\mp my - me^{\mp y}} dy,$$

d'après la distribution finale (33) et la transformation finale (29). Nous allons en déduire l'espérance mathématique de la $m^{\text{ième}}$ valeur d'en haut, \bar{u}_m , et d'en bas, ${}_m \bar{u}$, à l'aide de la formule symbolique (37). En exprimant la variable x au moyen de la variable finale y , on aura à l'aide de (1) et de (29)

$$\bar{u} = \frac{m^m}{(m-1)!} \int_{-\infty}^{+\infty} \left(\frac{y}{x} + \tilde{u} \right) e^{\mp my - me^{\mp y}} dy,$$

où nous avons laissé de côté l'indice m .

Donc la différence entre l'espérance mathématique et la valeur dominante de la $m^{\text{ième}}$ valeur sera, d'après les transformations finales (29)

$$(39) \quad \alpha(\bar{u} - \tilde{u}) = \bar{y};$$

l'espérance mathématique de la variable finale sera

$$\bar{y} = \frac{m^m}{(m-1)!} \int_{-\infty}^{+\infty} y e^{\mp(m-1)y - m e^{\mp y}} e^{\mp y} dy.$$

En introduisant une nouvelle variable d'intégration z par

$$(40) \quad e^{\mp y} = \frac{z}{m},$$

on obtient

$$\bar{y} = \frac{\mp 1}{(m-1)!} \int_0^{\infty} (\lg z - \lg m) z^{m-1} e^{-z} dz.$$

Cette intégrale qui paraît compliquée est facile à résoudre par l'introduction des fonctions gamma. On obtient tout de suite

$$\bar{y} = \pm \lg m \mp \frac{d \lg \Gamma(m)}{dm},$$

ou, d'après une propriété bien connue de la fonction gamma,

$$(41) \quad \bar{y} = \pm \lg m \pm \gamma \mp \sum_{r=1}^{m-1} \frac{1}{r},$$

γ étant la constante d'EULER. En introduisant l'indice m , on aura la relation

$$(42) \quad {}_m \bar{y} = -\bar{y}_m,$$

qui est une généralisation des relations

$${}_m \tilde{u} = -\tilde{u}_m; \quad {}_m \bar{u} = -\bar{u}_m,$$

valables seulement pour les distributions initiales symétriques.

L'espérance mathématique de la variable finale dans la distribution de la $m^{\text{ième}}$ valeur d'en haut est égale, mais de signe contraire, à la même quantité pour la $m^{\text{ième}}$ valeur d'en bas.

Pour $m = 1$ on obtient l'espérance mathématique de la dernière et de la première valeur

$$\bar{y}_1 = \gamma; \quad {}_1 \bar{y} = -\gamma,$$

ou en vertu de (39)

$$(43') \quad \bar{u}_1 = \tilde{u}_1 + \frac{\gamma}{\alpha_1}$$

$$(43) \quad {}_1 \bar{u} = {}_1 \tilde{u} - \frac{\gamma}{\alpha_1^2}.$$

E. J. GUMBEL

Pour de très grandes valeurs de m , on peut poser approximativement

$$\gamma = \sum_{r=1}^m \frac{1}{r} - \lg m.$$

En vertu de (41), l'espérance mathématique de la $m^{\text{ième}}$ valeur se rapproche de la valeur dominante si m augmente suffisamment. Cela résulte naturellement du fait que la distribution finale de la $m^{\text{ième}}$ valeur se resserre pour des valeurs croissantes de m .

En posant

$$(44) \quad \lg m + \gamma - \sum_{v=1}^{m-1} \frac{1}{v} = S_{1,m},$$

on obtient, d'après (39), la formule commune à l'espérance mathématique de la $m^{\text{ième}}$ valeur d'en haut et d'en bas

$$\bar{u} = \hat{u} \pm \frac{S_{1,m}}{\alpha},$$

En introduisant les indices, on écrira les espérances mathématiques

$$(45') \quad \bar{u}_m = \hat{u}_m + \frac{S_{1,m}}{\alpha_m},$$

$$(45) \quad {}_m\bar{u} = {}_m\hat{u} - \frac{S_{1,m}}{m\alpha}.$$

Donc la différence entre l'espérance mathématique et la dominante de la $m^{\text{ième}}$ valeur se compose de deux facteurs ayant des propriétés très différentes. L'un, $S_{1,m}$, dépend seulement de m et diminue assez vite lorsque m croît, l'autre, en vertu des formules (3) du tableau III et de (18) est égal à l'inverse de la valeur absolue de la dérivée logarithmique de la distribution initiale, pour la valeur dominante.

Pour le calcul de l'espérance mathématique de la $m^{\text{ième}}$ valeur, on n'a qu'à calculer la dominante à l'aide de la formule (14). L'espérance mathématique de la $m^{\text{ième}}$ valeur se rapproche de la dominante pour des nombres d'observations croissants, pourvu que les termes $1 : |\lg' w(\hat{u})|$ augmentent plus lentement que les valeurs dominantes elles-mêmes. Cependant pour des distributions dans lesquelles la dérivée logarithmique diminue en même temps que croissent les valeurs de la variable, l'espérance mathématique s'éloigne de la dominante, puisque la distribution s'élargit. Nous l'avons évité, en imposant à la distri-

bution initiale un caractère exponentiel. La distribution étant du type exponentiel, les valeurs x augmentent avec \tilde{u} ou diminuent vers une limite, ce qui veut dire que l'espérance mathématique s'approche de la dominante pour un nombre croissant d'observations. Car leur différence relative diminue. De plus, nous connaissons maintenant la manière dont ces deux valeurs se rapprochent l'une de l'autre ; il nous reste à savoir si leur différence diminue plus vite que ne le fait l'écart type de la distribution de la $m^{\text{ième}}$ valeur. Il nous faut donc calculer cet écart.

Traisons le problème général du $m^{\text{ième}}$ moment autour de l'espérance mathématique. Les deux premiers moments des distributions finales des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas sont données par (5a) et (5b).

En partant des distributions finales de la $m^{\text{ième}}$ valeur d'en haut on obtient les distributions finales de la $m^{\text{ième}}$ valeur d'en bas, en remplaçant y par $-y$ et en changeant les positions des indices m . Donc, les moments d'ordre pair de la distribution des $m^{\text{ièmes}}$ valeurs d'en bas et d'en haut seront égaux, et les moments d'ordre impair seront égaux mais de signes contraires ; l'indice m sera à droite pour les uns, et à gauche pour les autres. Il suffira donc de calculer les $m^{\text{ièmes}}$ moments des $m^{\text{ièmes}}$ valeurs d'en haut.

Le $m^{\text{ième}}$ moment de la $m^{\text{ième}}$ valeur d'en haut donné au moyen de y_m , par (5) (38) et (29'), s'écrit

$$(46) \quad x_{m^{\text{ième}}, m}^n = \frac{m^m}{(m-1)!} \int_{-\infty}^{+\infty} (y - \bar{y})^n e^{-(m-1)y - me^{-y}} e^{-y} dy,$$

formule dans laquelle nous avons écrit y au lieu de y_m . En introduisant encore une fois la nouvelle variable d'intégration z par

$$(40') \quad e^{-y} = \frac{z}{m},$$

on obtient

$$(47) \quad x_{m^{\text{ième}}, m}^n = \frac{(-1)^n}{(m-1)!} \int_0^{\infty} (\lg z - \lg m + \bar{y})^n z^{m-1} e^{-z} dz.$$

Pour calculer cette intégrale la première idée qui vient à l'esprit est l'emploi de la formule de binôme ; cela ne nous permettrait cependant de calculer que les moments autour de la dominante, dont on pourrait calculer les moments autour de l'espérance mathématique à l'aide d'une formule de récurrence. On peut éviter ce détour. Par le

procédé employé plus haut on peut réduire le $n^{\text{ième}}$ moment à une $n^{\text{ième}}$ dérivée d'un produit de la fonction gamma par une fonction exponentielle ; on obtient en effet

$$\alpha_{m^{\text{ième}}, n, m}^n = \frac{(-1)^n}{\Gamma(m)} \left(\frac{d^n}{d\phi^n} e^{(\phi-1)(\bar{y}-lgm)} \int_0^\infty z^{m-1+\phi-1} e^{-z} dz \right)_{\phi=1}.$$

Il s'agit maintenant de réduire ces $n^{\text{ièmes}}$ dérivées aux $n^{\text{ièmes}}$ dérivées logarithmiques de la fonction gamma, dont les valeurs numériques sont bien connues. En posant pour la variable dont dépend la fonction gamma

$$\phi = 1 + t$$

le $n^{\text{ième}}$ moment de la distribution finale de la $m^{\text{ième}}$ valeur d'en haut devient

$$\alpha_{m^{\text{ième}}, n, m}^n = \frac{(-1)^n}{\Gamma(m)} \left[\frac{d^n}{dt^n} e^{t(\bar{y}-lgm)} \Gamma(m+t) \right]_{t=0}.$$

En introduisant enfin

$$(48) \quad e^{t(\bar{y}-lgm)} \frac{\Gamma(m+t)}{\Gamma(m)} = \varphi(t),$$

on obtient

$$(49) \quad \alpha_{m^{\text{ième}}, n, m}^n = (-1)^n \varphi^{(n)}(0),$$

où l'exposant entre parenthèses désigne l'ordre de la dérivée correspondante. Notre méthode, qui pourrait sembler un peu compliquée, mène donc au procédé classique de la construction d'une fonction caractéristique. Abstraction faite du signe et du facteur multiplicatif α_m^n la fonction $\varphi(t)$ est bien une fonction caractéristique pour la distribution finale de la $m^{\text{ième}}$ valeur d'en haut. L'expression (49) permet en outre la réduction des $n^{\text{ièmes}}$ dérivées aux $n^{\text{ièmes}}$ dérivées logarithmiques de la fonction gamma. En effet par un théorème connu d'analyse on a

$$(50) \quad \varphi^{(n)} = \sum_{v=0}^{n-1} \binom{n-1}{v} \varphi^{(v)} \lg^{(n-v)} \varphi.$$

Il suffira de traiter les facteurs logarithmiques, car les $\varphi^{(v)}$ sont les moments antérieurs au $n^{\text{ième}}$, dont les deux premières valeurs nous sont connues, par (5a) et (5b).

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

Le dernier membre de la somme disparaît ; en effet

$$(51) \quad \lg \varphi'(0) = \bar{y} - \lg m + \left(\frac{d \lg \Gamma(m+t)}{dt} \right)_{t=0},$$

s'annule en vertu de la formule (48). Les autres valeurs des dérivées logarithmiques de la fonction $\varphi(t)$ sont données par la théorie des fonctions gamma. Posons pour le moment $n - v = \lambda + 1$ ou

$$\lambda = 1, 2, \dots, n - 1.$$

Puisque le facteur $\lambda = 0$ ne nous intéresse plus, on pourra écrire

$$\lg^{(\lambda+1)} \varphi(t) = \lg^{(\lambda+1)} \Gamma(m+t) = \frac{d^{(\lambda+1)}}{dt^{\lambda+1}} \sum_{k=1}^{m-1} \lg(k+t) + \lg^{(\lambda+1)} \Gamma(1+t).$$

En introduisant les dérivées logarithmiques de la fonction gamma, on obtient

$$\lg^{(\lambda+1)} \varphi(0) = (-1)^{\lambda} \lambda! \sum_{k=1}^{m-1} \frac{1}{k^{\lambda+1}} - (-1)^{\lambda} \lambda! \sum_{k=1}^{\infty} \frac{1}{k^{\lambda+1}},$$

ce que nous écrirons

$$(52) \quad \lg^{(\lambda+1)} \varphi(0) = (-1)^{\lambda+1} \lambda! S_{\lambda+1, m},$$

en posant

$$(44a) \quad \sum_{k=m}^{\infty} \frac{1}{k^{\lambda+1}} = S_{\lambda+1, m}.$$

En introduisant ces valeurs (51) et (52) dans la somme (50), on obtient à l'aide de (49)

$$\alpha_{m!}^n \mu_{r, m} = (-1)^n (n-1)! \sum_{v=0}^{n-2} (-1)^{r+n-v} \frac{\alpha_{m!}^v \mu_{r, m}}{v!} S_{n-v, m}$$

En séparant les deux premiers membres, connus par les formules (5a) et (5b), on obtient finalement

$$(53) \quad \alpha_{m!}^n \mu_{n, m} = (n-1)! S_{n, m} + (n-1)! \sum_{v=2}^{n-2} \frac{\alpha_{m!}^v \mu_{r, m} S_{n-r, m}}{v!}.$$

En partant des valeurs connues des deux premiers moments on peut calculer tous les moments, l'un après l'autre, à l'aide de cette formule, pour toutes les valeurs de m . La valeur de N et de la distribution initiale n'y entre que par leurs combinaisons α_m . Les facteurs $S_{\lambda,m}$ sont étroitement reliés aux valeurs bien connues S_λ de la théorie des fonctions gamma. En effet, d'après (44a)

$$S_{\lambda,m} = S_\lambda - \sum_{v=1}^{m-1} \frac{1}{v^\lambda}.$$

Quant aux moments de la distribution finale de la $m^{\text{ième}}$ valeur d'en bas, on les obtient en remplaçant α_m par $m\alpha$. A cela près, le second membre de (53) reste le même pour les moments d'ordre pair et change de signe pour les moments d'ordre impair. En particulier, l'écart type des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas sera

$$(54) \quad \alpha_m \sigma_m = \sqrt{S_{2,m}} = m \alpha_m \sigma.$$

Les moments suivants seront

$$(55) \quad \left\{ \begin{array}{l} \alpha_m^3 \mu_{3,m} = 2S_{3,m} = -m \alpha_m^3 \mu_3, \\ \alpha_m^4 \mu_{4,m} = 6S_{4,m} + 3S_{2,m}^2 = m \alpha_m^4 \mu_4, \\ \alpha_m^5 \mu_{5,m} = 24S_{5,m} + 20S_{2,m}S_{3,m} = -m \alpha_m^5 \mu_5. \end{array} \right.$$

Donc, tous les moments des distributions finales des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas peuvent être calculés par des formules très simples. Les valeurs $S_{\lambda,m}$ ne dépendent plus du nombre d'observations. Par contre, les facteurs α_m et $m\alpha$ à gauche et à droite augmentent avec le nombre d'observations, si la valeur absolue de la dérivée logarithmique de la distribution initiale augmente. Dans ce cas, les moments diminuent pour un nombre croissant d'observations.

Le facteur α n'intervient plus dans l'élément qui mesure l'asymétrie

$$(56) \quad \delta_m = \frac{S_{1,m}}{\sqrt{S_{2,m}}} = -m \delta$$

et dans les quotients des moments

$$(57) \quad \rho_{1,m} = \frac{2S_{3,m}}{S_{2,m}^2} = -m \rho_1$$

$$(58) \quad \beta_{2,m} - 3 = \frac{6S_{4,m}}{S_{2,m}^2} = m \beta_2 - 3.$$

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

Par conséquent ces valeurs sont indépendantes de la distribution initiale et du nombre d'observations. Elles restent les mêmes pour n'importe quelle distribution initiale du type exponentiel dès que le nombre d'observations est suffisamment grand. Elles dépendent seulement du nombre ordinal m et auront la même valeur pour les distributions finales qui se resserrent ou qui sont invariantes si le nombre d'observations augmente. Le calcul des médianes des $m^{\text{ièmes}}$ valeurs conduit d'après (37) aux fonctions Gamma incomplètes. On peut donc déterminer les valeurs numériques, fonctions des α et des dominantes analogues à (45) à l'aide des tables de PEARSON (31).

Pour $m = 1$, c'est-à-dire pour la distribution de la dernière et de la première valeur on aura

$$S_{1,1} = \gamma$$

et

$$S_{\lambda,1} = S_{\lambda} \quad \text{pour} \quad \lambda \geq 2.$$

L'écart type pour ces distributions sera

$$(54a) \quad \sigma_{1,1} = \frac{\pi}{\sqrt{6}} = 1,28\sigma,$$

donc plus grand que l'écart probable calculé par (36).

Le coefficient d'asymétrie devient

$$(56a) \quad \rho_1 = \frac{\gamma\sqrt{6}}{\pi} = -1,2.$$

Donc la différence entre l'espérance mathématique et la dominante de la dernière et de la première valeur est encore 0,45005 fois l'écart type de cette distribution.

Les quotients des moments auront les valeurs numériques suivantes

$$(57a) \quad \rho_{1,1} = 1,13906 = -1,2\rho_1$$

$$(58a) \quad \rho_{2,1} - 3 = 2,4 = 1,2\rho_2 - 3.$$

Cette valeur qui est positive, comme on doit attendre, prouve que les valeurs moyennes de la distribution finale y jouent un rôle prépondérant. Mais, en même temps la distribution de la dernière (ou de la première) valeur est plus étalée du côté droit (ou gauche), qu'une distribution de GAUSS. Les formules (54) et (55) donnent les moments de la distribution finale de la dernière et de la première valeur. Dans ce cas particulier, on obtient pour les quatre premiers moments les valeurs calculées par M. R. A. FISHER (8).

Dans le cas général, on déduit les valeurs $S_{\lambda,m}$ nécessaires pour le calcul des moments à l'aide de la formule de récurrence

$$(59) \quad S_{\lambda,m} = S_{\lambda,m-1} - \frac{1}{(m-1)^\lambda}$$

puisque les valeurs S_λ sont connues. Pour des valeurs croissantes de λ les S_λ tendent assez rapidement vers 1. Il en résulte que les $S_{\lambda,m}$ tendent vers zéro et cela d'autant plus vite, que le nombre ordinal m est plus grand. Le tableau IV contient quelques valeurs de $S_{\lambda,m}$ pour $\lambda, m = 2, 3, \dots, 10$.

TABLEAU IV. — Valeurs de $S_{\lambda,m}$.

λ	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
2	1,64493	0,64493	0,39493	0,28382	0,22132
3	1,20206	0,20206	0,07706	0,04002	0,02439
4	1,08232	0,08232	0,01982	0,00748	0,00357
5	1,03693	0,03693	0,00568	0,00157	0,00059
6	1,01734	0,01734	0,00172	0,00035	0,00010
7	1,00835	0,00835	0,00054	0,00008	0,00002
8	1,00408	0,00408	0,00017	0,00002	
9	1,00201	0,00201	0,00006		
10	1,00099	0,00099	0,00002		
λ	$m = 6$	$m = 7$	$m = 8$	$m = 9$	$m = 10$
2	0,18132	0,15355	0,13314	0,11751	0,10517
3	0,01639	0,01176	0,00885	0,00690	0,00553
4	0,00197	0,00120	0,00078	0,00054	0,00039
5	0,00027	0,00014	0,00008	0,00005	0,00003
6	0,00004	0,00002	0,00001		
7	0,00001				
8					
9					
10					

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

On tire facilement de ces valeurs les caractéristiques des distributions des $m^{\text{ièmes}}$ valeurs en fonction du nombre m . La deuxième colonne du tableau V contient l'espérance mathématique de la variable finale définie par (1) et (39) et calculée d'après (41), la troisième donne la mesure d'asymétrie définie par (4) et calculée d'après (56). Les dernières colonnes contiennent les quotients des moments définis par (6a) et (6b) et calculés d'après (57) et (58). L'espérance mathématique de la variable finale, la mesure d'asymétrie et le premier quotient des moments sont positifs (ou négatifs) pour les $m^{\text{ièmes}}$ valeurs d'en haut (ou d'en bas).

TABLEAU V. — *Caractéristiques des distributions des $m^{\text{ièmes}}$ valeurs.*

m	$\pm \bar{y}$	$\pm \bar{z}$	$\pm \beta_1$	$\beta_2 - 3$
1	0,577	0,450	1,139	2,400
2	0,270	0,337	0,780	1,188
3	0,176	0,280	0,621	0,763
4	0,130	0,244	0,529	0,557
5	0,103	0,220	0,468	0,437
6	0,086	0,201	0,425	0,360
7	0,073	0,187	0,391	0,305
8	0,064	0,175	0,364	0,264
9	0,057	0,165	0,343	0,235
10	0,051	0,157	0,324	0,212

Le tableau V montre que les caractéristiques diminuent pour des valeurs croissantes de m .

Reste à savoir comment les distributions des $m^{\text{ièmes}}$ valeurs se resserrent si m augmente. Pour cela il suffit de traiter le signe des changements des quotients β_1 , β_2 et \bar{z} comme fonction de m .

Puisque

$$\frac{\Delta S_{\lambda,m}}{\Delta m} = -\frac{1}{m^{\lambda}}$$

on aura d'après (57) et (59) en traitant $\beta_{1,m}^2$

$$\text{signe } \frac{\Delta \beta_{1,m}^2}{\Delta m} = 3S_{2,m} - \frac{2S_{2,m}}{m} = -\text{signe } \frac{\Delta m^{\frac{1}{2}}}{\Delta m}$$

et d'après (58)

$$\text{signe } \frac{\Delta \zeta_{2,m}}{\Delta m} = -\frac{S_{2,m}}{m^2} + 2S_{4,m} = \text{signe } \frac{\Delta_m \zeta_2}{\Delta m}.$$

Pour connaître le signe de ces différences, il nous faut une expression analytique pour les sommes $S_{\lambda,m}$. Choisissons la formule d'intégration de EULER, qui donne pour $\lambda \geq 2$.

$$(60) \quad S_{\lambda,m} = \frac{1}{m^{\lambda-1}} \left(\frac{1}{\lambda-1} + \frac{1}{2m} + \frac{\lambda}{12m^2} - \frac{\lambda(\lambda+1)(\lambda+2)}{720m^4} \dots \right)$$

dont on tire

$$(60a) \quad S_{2,m} = \frac{1}{m} \left(1 + \frac{1}{2m} + \frac{1}{6m^2} - \frac{1}{30m^4} \dots \right)$$

$$(60b) \quad S_{3,m} = \frac{1}{2m^2} \left(1 + \frac{1}{m} + \frac{1}{2m^2} - \frac{1}{6m^4} \dots \right)$$

$$(60c) \quad S_{4,m} = \frac{1}{m^3} \left(\frac{1}{3} + \frac{1}{2m} + \frac{1}{3m^2} - \frac{1}{6m^4} \dots \right),$$

formules qui fournissent des approximations suffisantes à partir de $m = 3$. En introduisant ces valeurs on aura

$$(61) \quad \text{signe } \frac{\Delta \beta_{1,m}}{\Delta m} = -\frac{1}{2m} \left(1 - \frac{1}{m} - \frac{5}{6m^2} + \frac{11}{30m^4} \dots \right) = -\text{signe } \frac{\Delta_m \beta_1}{\Delta m}.$$

L'expression précédente sera négative à partir de $m = 2$. Donc, pour un nombre m croissant les quotients $\beta_{1,m}$ et $m\beta_1$ tendront vers zéro. Pour les distributions d'en haut (ou d'en bas) les côtés droits (ou gauches) sont plus importants que les côtés gauches (ou droits), mais cette différence tend à disparaître lorsque m augmente.

Pour le deuxième quotient on aura

$$(62) \quad \text{signe } \frac{\Delta \zeta_{2,m}}{\Delta m} = -\frac{1}{m^3} \left(\frac{1}{3} - \frac{1}{2m} - \frac{1}{2m^2} + \frac{3}{10m^4} \dots \right) = \text{signe } \frac{\Delta_m \zeta_2}{\Delta m}$$

qui sera toujours négatif à partir de $m = 3$. Donc, la différence entre les distributions des $m^{\text{ièmes}}$ valeurs et celle de GAUSS diminue pour des valeurs croissantes de m .

La différence relative entre l'espérance mathématique et la dominante divisée par l'écart type tendra vers zéro. Puisque

$$S_{1,m} \rightarrow 0$$

on aura d'après (44) et (56)

$$m\hat{\sigma}^2 = \hat{\sigma}_m^2 \rightarrow \frac{1}{m^2 S_{2,m}}.$$

Mais d'après (60a) $S_{2,m}$ tend vers zéro comme $\frac{1}{m}$. Pour un nombre croissant d'observations, la différence entre l'espérance mathématique et la dominante diminuera plus vite que l'écart type. L'espérance mathématique s'approchera de la dominante, de telle manière que leurs valeurs respectives seront d'autant plus voisines que m sera plus grand. De plus, l'asymétrie, mesurée par β_1 , tendra à disparaître et la distribution finale de la $m^{\text{ième}}$ valeur tendra, pour des valeurs croissantes de m , vers une distribution pour laquelle $\beta_2 = 3$, c'est-à-dire vers une distribution de GAUSS. Celle-ci peut être regardée comme une limite de la distribution finale (37) de la $m^{\text{ième}}$ valeur. Pour de grandes valeurs de m , qui sont pourtant petites par rapport à $\frac{N}{2}$, on peut poser

$$S_{2,m} = \frac{1}{m}$$

et l'on aura

$$(63) \quad w(x, N) = x \sqrt{\frac{m}{2\pi}} e^{-\frac{my^2}{2}}.$$

Pour avoir la distribution doublement finale de la $m^{\text{ième}}$ valeur d'en haut ou d'en bas pour N observations, il faut ajouter les indices m aux valeurs x et y , d'après la règle établie dans le tableau III. Cette expression vaut pour des grandes valeurs de N et de m , telles que $m \ll N$. Elle permet de retrouver la limite (17b) vers laquelle tend la probabilité de la dominante. On peut aussi arriver, naturellement, à l'expression limite (63) en développant le second exposant dans (37).

La formule (63) confirme l'analogie établie plus haut entre les distributions des $m^{\text{ièmes}}$ valeurs et la distribution de BERNOULLI. La limite de POISSON, analogue à la distribution finale des $m^{\text{ièmes}}$ valeurs extrêmes conduit également, en effet, pour de grandes valeurs de m à la distribution de GAUSS.

En même temps, cette formule pose la question de savoir quelles sont les valeurs caractéristiques et les limites des distributions des $m^{\text{ièmes}}$ valeurs pour de grandes valeurs de m , telles que

$$(64) \quad m \approx \frac{N}{2},$$

questions que nous avons laissées à part jusqu'à présent, puisqu'elles

ne se rapportent plus aux valeurs extrêmes. Nous allons traiter ce cas limite puisqu'il mène à des analogies intéressantes.

Des problèmes analogues ont été traités par J. BERTRAND (1) qui envisage d'abord la plus petite erreur absolue, — c'est-à-dire la plus petite déviation absolue de l'espérance mathématique — d'une distribution initiale de GAUSS, ensuite la seconde, la troisième, etc. Ces valeurs sont donc classées sans tenir compte du signe ; elles ne varient que dans le domaine positif et leurs espérances mathématiques sont assez petites.

Les distributions (11) des $m^{\text{ièmes}}$ valeurs étant établies indépendamment de toute condition sur N et m valent aussi pour (64). Introduisons suivant la méthode de M. EYRAUD (6)

$$(65') \quad m = Np$$

$$(65) \quad N - m = Nq.$$

Les valeurs de p et q qui nous intéressaient pour les $m^{\text{ièmes}}$ valeurs extrêmes étaient $p \rightarrow 0$; $q \rightarrow 1$. Les valeurs que nous allons traiter maintenant seront d'après (64)

$$p \sim \frac{1}{2} \sim q.$$

Suivant la définition (1a) de la médiane, on appelle les $m^{\text{ièmes}}$ valeurs satisfaisant à (64) les *voisines de la médiane*. Les dominantes sont les racines de

$$(66') \quad \frac{Nq}{W} w - \frac{Np - 1}{1 - W} w + \frac{w'}{w} = 0,$$

$$(66) \quad \frac{Nq}{1 - W} w - \frac{Np - 1}{W} w - \frac{w'}{w} = 0.$$

Pour une distribution initiale telle que la médiane ne soit pas trop éloignée de la dominante, les troisièmes termes peuvent être négligés par rapport aux deux autres. Cela est encore plus légitime si le nombre d'observations est suffisamment grand. Donc les dominantes des $m^{\text{ièmes}}$ valeurs voisines de la médiane, données par

$$W(\tilde{u}_m) = q \frac{N}{N - 1} = 1 - W({}_m \tilde{w}),$$

dérivent, pour des valeurs suffisamment grandes de N , des équations (14) valables pour les $m^{\text{ièmes}}$ valeurs extrêmes.

La densité de probabilité $\mathbf{w}(\tilde{u})$ des dominantes des $m^{\text{ièmes}}$ valeurs voisines de la médiane, sera déduite, pour N suffisamment grand, d'après (II) et (14), par

$$(67) \quad \frac{\mathbf{w}_m(\tilde{u}_m)}{w(\tilde{u}_m)} = \sqrt{\frac{N}{2\pi pq}} = \frac{{}_m\mathbf{w}(\tilde{u})}{w(\tilde{u})},$$

relation analogue à (21) et (17b). En effet, on peut écrire ces formules pour les $m^{\text{ièmes}}$ valeurs extrêmes

$$\frac{\mathbf{w}_m(\tilde{u}_m)}{w(\tilde{u}_m)} = \sqrt{\frac{N}{2\pi p}} = \frac{{}_m\mathbf{w}(\tilde{u})}{w(\tilde{u})}.$$

Pour la distribution initiale de GAUSS, M. HAAG (2 ; 28) a démontré que la distribution finale des $m^{\text{ièmes}}$ valeurs voisines de la médiane est du type de GAUSS. On peut étendre aisément ce théorème. Posons pour les $m^{\text{ièmes}}$ valeurs d'en haut, voisines de la médiane

$$W(x) = q + y_m,$$

et pour les valeurs d'en bas correspondantes

$$W(x) = p - my.$$

D'après (II) et (67), on obtient les distributions finales $\mathbf{w}(x)$ par

$$\mathbf{w}_m(x)dx = \sqrt{\frac{N}{2\pi pq}} \left(1 \pm \frac{y}{q}\right)^{Nq} \left(1 \mp \frac{y}{p}\right)^{Np-1} = {}_m\mathbf{w}(x)dx.$$

Les signes supérieurs (ou inférieurs) valent pour l'équation gauche (ou droite) avec l'indice m à la variable y à droite (ou à gauche). Le développement usuel du produit en y conduit aux distributions finales correspondantes à (33) et (63)

$$(68) \quad \mathbf{w}_m(x)dx = \sqrt{\frac{N}{2\pi pq}} e^{-\frac{Ny^2}{2pq}} dy = {}_m\mathbf{w}(x)dx.$$

La condition énoncée en haut permet de poser en première approximation

$$(69) \quad W(x) = w(\tilde{u}) \left(x - \tilde{u} \mp \frac{q - p}{2w(\tilde{u})}\right).$$

où nous avons laissé de côté l'indice m de \tilde{u} . Les variables finales dans les distributions des $m^{\text{ièmes}}$ valeurs voisines de la médiane seront

$$y_m = w(\tilde{u}) \left(x - \tilde{u} \mp \frac{q - p}{2w(\tilde{u})}\right) = {}_m y.$$

Il en résulte que : les distributions finales des $m^{\text{ièmes}}$ valeurs voisines de la médiane d'une distribution initiale satisfaisant à (69) sont du type de GAUSS, avec une espérance mathématique

$$(70) \quad \bar{u}_m = \check{u} \pm \frac{q - \check{p}}{2w(\check{u})} = {}_m\bar{u},$$

et un écart type égal à

$$(71) \quad \sigma_m = \frac{1}{w(\check{u})} \sqrt{\frac{\check{p}q}{N}} = {}_m\sigma.$$

La première formule est analogue à (45). Quant à l'écart type de la $m^{\text{ième}}$ valeur extrême, il sera donné, pour de grandes valeurs de m , en vertu de (54) et (60b), par la formule analogue

$$\sigma_m = \frac{1}{w(\check{u})} \sqrt{\frac{\check{p}}{N}} = {}_m\sigma.$$

Pour la médiane, laquelle au sens ordinaire n'est définie que pour N impair, il faut choisir une valeur \check{m} telle que $N = 2\check{m} - 1$, c'est-à-dire $\check{p} = \frac{1}{2} + \frac{1}{N}$. Pour des valeurs suffisamment grandes on obtient une distribution de GAUSS ayant l'écart type

$$(71 a) \quad \sigma_{\check{m}} = \frac{1}{2w(\check{x})\sqrt{N}}.$$

Il s'ensuit que les distributions des $m^{\text{ièmes}}$ valeurs voisines de la médiane sont analogues aux distributions des $m^{\text{ièmes}}$ valeurs extrêmes pour de grandes valeurs de m .

Conclusions

Pour une distribution initiale $w(x)$ ayant la probabilité $W(x)$ et pour un nombre croissant d'observations, les distributions (II) des $m^{\text{ièmes}}$ valeurs d'en haut et d'en bas conduisent à deux limites correspondantes aux deux limites de la distribution de BERNOULLI. Introduisons \check{p} et q par (65). Pour les $m^{\text{ièmes}}$ valeurs voisines de la médiane, où $\check{p} \simeq q$, et $W \rightarrow q$ ou $W \rightarrow \check{p}$, on revient à la distribution de GAUSS pourvu que la médiane de la distribution initiale ne soit pas trop éloignée de la dominante. Pour les valeurs extrêmes, où $\check{p} \rightarrow 0$ et $W \rightarrow 1$ ou $W \rightarrow 0$, on arrive à la distribution doublement

exponentielle (37) pourvu que la distribution initiale appartienne au type exponentiel. Le calcul des dominantes est le même pour les deux limites. Pour de grandes valeurs de m il existe une analogie complète entre les $m^{\text{ièmes}}$ valeurs extrêmes et les $m^{\text{ièmes}}$ valeurs voisines de la médiane, car les deux distributions sont du type de GAUSS.

Pour appliquer nos résultats portant sur les $m^{\text{ièmes}}$ valeurs extrêmes à une distribution donnée du type exponentiel et à un nombre N donné, on calculera d'abord la dérivée logarithmique. Si sa valeur absolue augmente sans limite pour des valeurs croissantes de la variable, la distribution de la $m^{\text{ième}}$ valeur se resserrera pour un nombre croissant d'observations. Si la dérivée logarithmique tend, soit en augmentant soit en diminuant, vers une constante, la distribution finale ne changera plus de forme.

Puis on calculera la probabilité totale, dont on tirera les dominantes des $m^{\text{ièmes}}$ valeurs à l'aide de (14). Les équations (9) de L'HOPITAL nous permettront de voir si le nombre d'observations est suffisant pour que cette approximation soit justifiée.

Dans ce cas, la variable finale sera donnée par (29), la distribution finale par (33), l'espérance mathématique par (45), l'écart type par (54), les autres moments par (55), le coefficient d'asymétrie par (56) et les quotients des moments par (57) et (58). Si en dehors du nombre des observations, le nombre m est très grand tout en restant petit vis-à-vis de N , la distribution finale de la $m^{\text{ième}}$ valeur pour N observations tendra vers la distribution (63) de Gauss.

Cette théorie est susceptible de nombreuses applications, puisque en particulier, la distribution de GAUSS et la distribution exponentielle appartiennent au type exponentiel. Pour l'une, les distributions des $m^{\text{ièmes}}$ valeurs se resserrent et la précision augmente pour un nombre croissant d'observations ; pour l'autre, elles restent indépendantes de N . Dans ce cas, on ne peut pas augmenter la précision en augmentant le nombre d'observations. Mais pour toute distribution initiale du type exponentiel l'espérance mathématique de la $m^{\text{ième}}$ valeur tendra vers la dominante, en ce sens que leur différence relative disparaît pour un nombre croissant d'observations.

Pour appliquer ces formules à des observations portant sur la $m^{\text{ième}}$ valeur, on calculera la moyenne arithmétique et l'écart type. On obtient alors les deux constantes α et $\tilde{\alpha}$ de la distribution théorique par (45) et (54). Le calcul des moments supérieurs fournit le contrôle

usuel à l'aide de (55). Les valeurs théoriques comparables aux observations peuvent être tirées des tableaux I et II, procédé extrêmement simple.

Parmi les applications connues, nous nous bornerons à citer les problèmes d'artillerie qui ont donné lieu à de nombreux essais (en partie vains) de déterminer la plus grande valeur d'une déviation (2), — les intervalles entre les émissions radio-actives classés d'après leur longueur, enfin la durée extrême de la vie humaine (22 ; 23), c'est-à-dire la distribution des âges des derniers survivants observés pour une série d'années (26). Enfin, M. de MISÈS (30) a traité la théorie des plus grandes valeurs pour deux distributions. Il en a tiré une conséquence particulièrement intéressante pour mesurer l'intelligence d'une minorité dite raciale existant dans un pays de l'Europe Centrale.

Étant données ces applications multiples ainsi que la simplicité du procédé auquel nous aboutissons, nous espérons que cette théorie des *m*^{èmes} valeurs pourra être employée avec succès pour divers calculs pratiques concernant un ensemble d'observations. C'est dans ce but que le présent travail a été entrepris.

(Conférences faites à l'Institut Henri Poincaré en l'hiver 1933-34.)

Université de Lyon, le 30 juin 1934.

BIBLIOGRAPHIE

- 1) BERTRAND I. — *Calcul des Probabilités*. Paris, Gauthiers-Villars, 1907.
- 2) BOREL, E. — *Traité du Calcul des probabilités et de ses applications*, t. IV, fasc. I, J. Haag, Applications au tir. Paris, 1926.
- 3) Von BORTKIEWICZ L. — *Variationsbreite und mittlerer Fehler*. Sitzungsberichte der Berliner mathematischen Gesellschaft, t. 21, p. 3. Berlin, 1922.
- 4) — *Die Variationsbreite beim Gauss'schen Fehlergesetz*. Nordisk Statistisk Tidskrift Bd. I, 1. p. 11 ; 2. p. 12. Stockholm, 1922.
- 5) DODD E. L. — *The greatest and the least variate under general laws of error*. *Transactions of the American Mathematical Society*, vol. 25, n° 4, p. 525. New-York, 1923.

LES VALEURS EXTRÊMES DES DISTRIBUTIONS STATISTIQUES

- 6) EYRAUD H. — *Sur la valeur la plus précise d'une distribution*. Comptes Rendus de l'Académie des Sciences, 199, n° 18, p. 817. Paris, 1934.
- 7) De FINETTI B. — *Sulla legge di probabilità degli estremi*. Metron, vol. IX, n° 3/4, p. 125. Roma, 1932.
- 8) FISHER R. A. et L. H. C. TIPPETT. — *Limiting forms of the frequency distributions of the largest or smallest member of a sample*. Proc. Cambridge, Phil. Soc. 24, Part 2, p. 180. Cambridge 1928.
- 9) FOGELSON S. — *Sur la détermination de la médiane*. Revue trimestrielle de Statistique, t. VII, fasc. 2. Varsovie, 1930.
- 10) FRÉCHET M. — *Sur la loi de probabilité de l'écart maximum*. Annales de la Société polonaise de Mathématique, t. VI, p. 93. Cracovie, 1927.
- 11) GUMBEL E. J. — *La distribuzione dei decessi secondo la legge di Gauss*. Giornale dell'Istituto italiano degli Attuari, Anno III, No. 3, p. 311. Roma 1932.
- 12) — *L'âge limite*. (Résumé) Bulletin de la Société mathématique de France, tome IX, fasc. 3/4, p. 34. Paris 1932.
- 13) — *Die Gauss'sche Verteilung der Gestorbenen*. Jahrbücher für Nationalökonomie und Statistik, 138. Band, III. Folge, Bd. 83, p. 365. Jena 1933.
- 14) — *Die Verteilung der Gestorbenen um das Normalalter*. Aktuarske Vedy, Tome IV, No. 2, p. 65. Prague 1933.
- 15) — *La plus petite valeur parmi les plus grandes*. Comptes Rendus de l'Académie des Sciences, 196, No. 25, p. 1857. Paris 1933.
- 16) — *La plus petite valeur parmi les plus grandes et la plus grande valeur parmi les plus petites*. Comptes Rendus, 197, N° 18, p. 965. Paris 1933.
- 17) — *La distribution limite de la plus petite valeur parmi les plus grandes*. Comptes Rendus, 197, n° 20, p. 1082. Paris 1933.
- 18) — *La distribution limite de la plus grande valeur parmi les plus petites*. Comptes rendus, 197, N° 23, p. 1381. Paris 1933.
- 19) — *L'espérance mathématique de la n ième valeur*. Comptes rendus, 198, n° 1, p. 33. Paris 1934.
- 20) — *Les moments des distributions finales de la première et de la dernière valeur*. Comptes rendus, 198, n° 2, p. 141. Paris 1934.
- 21) — *Les moments des distributions finales de la n ième valeur*. Comptes rendus, 198, n° 4, p. 313. Paris 1934.
- 22) — *Das Alter des Methusalem*. Zeitschrift für schweizerische Statistik und Volkswirtschaft, 69. Jahrg. H. 4. Basel 1933.
- 23) — *L'età limite*. Giornale dell'Istituto italiano degli Attuari. Anno V, No. 1. Roma 1934.
- 24) — *Le paradoxe de l'âge limite*. Comptes Rendus, 199, n° 19, p. 918. Paris, 1934.

E. J. GUMBEL

- 25) — *La distribution finale des valeurs voisines de la médiane*. Comptes Rendus, 199, n° 22, p. 1174. Paris, 1934.
- 26) — *Les plus grands âges en Suisse*. Zeitschrift für schweizerische Statistik und Volkswirtschaft, 70. Jahrg., N. 4, p. 601. Basel, 1934.
- 27) — *Les mêmes valeurs extrêmes et le logarithme du nombre d'observations*. Comptes Rendus, 200, n° 7, p. 509. Paris, 1935.
- 28) HAAG J. — *Sur la combinaison des résultats d'observations*. Comptes Rendus, 179, n° 24, p. 1388. Paris, 1924.
- 29) von MISES R. — *Ueber die Variationsbreite einer Beobachtungsreihe*. Sitzungsberichte der Berliner Mathematischen Gesellschaft, t. 22, p. 3. Berlin, 1923.
- 30) de MISÈS R. — *Problème de deux races*. Recueil Mathématique, t. 41, n° 3, p. 359. Moscou, 1934.
- 31) PEARSON K. — *Tables of the incomplete Γ function*. University Press. Cambridge, 1934.
- 32) TIPPETT L. H. C. — *On the extreme individuals and the range of samples taken from a normal population*. Biometrika, vol. XVII, part 3/4, p. 364. Cambridge, 1925.
- 33) TRICOMI F. — *Determinazione del valore asintotico di un certo integrale*. Rendiconti Accademia dei Lincei, vol. XVII, 1. sem. fasc. 2, p. 116. Roma, 1933.

TABLE DES MATIÈRES

Introduction	115
Les distributions des <i>mêmes</i> valeurs et leurs dominantes.....	120
Les distributions finales	131
Espérance mathématique et moments	140
Conclusions	154