

JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

MAURICE FRÉCHET

Sur la loi de répartition de certaines grandeurs géographiques

Journal de la société statistique de Paris, tome 82 (1941), p. 114-122

http://www.numdam.org/item?id=JSFS_1941__82__114_0

© Société de statistique de Paris, 1941, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (<http://publications-sfds.math.cnrs.fr/index.php/J-SFdS>) 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/>

II

SUR LA LOI DE RÉPARTITION DE CERTAINES GRANDEURS GÉOGRAPHIQUES

Introduction. — On trouve dans les *Comptes Rendus de la XXIV^e Session de l'Institut international de Statistique (en 1938)* une intéressante étude intitulée : « Deux types fondamentaux de distribution statistique », où l'auteur, a attiré l'attention sur des statistiques indiquant la répartition de certaines grandeurs géographiques. (Par exemple, chaque colonne donne le nombre des îles, des lacs, ayant une surface comprise entre deux aires données, de même pour le nombre des cours d'eau ayant une longueur comprise entre deux nombres donnés). Il a fait observer que la répartition de ces grandeurs n'obéit pas à la loi statistique si fondamentale qu'on l'a appelée loi normale (et que nous appelons, avec E. B. WILSON, la seconde loi de LAPLACE). Il observe de plus que ces tableaux statistiques paraissent se conformer à des lois de répartition très analogues entre elles qui ne seraient que des exemples d'un second type fondamental de statistique.

Une troisième remarque. — Ce sont là deux observations importantes que nous nous proposons de compléter par une troisième remarque. Ce sera, d'ailleurs, une observation qui se présenterait d'elle même à l'esprit des statisticiens économistes (1) ou psychologues ayant lu l'article cité. Avant même de consulter ces tableaux, ils auraient eu l'idée que les répartitions envisagées présentent dans leur nature une certaine analogie avec la répartition des

richesses ou la répartition des capacités intellectuelles. Or, ces deux dernières sortes de répartition paraissent aussi obéir à des lois très voisines, à la valeur des paramètres près. On a proposé des formules très variées pour les représenter. La première et la plus simple d'entre elles est celle de PARETO. On peut l'exprimer simplement ainsi : Appelons N le nombre des observations où la variable aléatoire envisagée (revenu, valeur d'une succession, note à un examen, etc...) est supérieure à un nombre x . Ceci étant, si l'on représente graphiquement la relation qui existe non pas entre N et x , mais entre leurs logarithmes, on obtient une ligne droite. C'est là une façon d'exprimer la loi de PARETO (2) qui a, en particulier, l'avantage de se prêter à une vérification visuelle.

En fait, on constate que les tableaux statistiques utilisés dans l'article mentionné se représentent assez bien sous cette forme comme le montrent les graphiques ci-joints (3). On voit, en effet, que dans chacun les points obtenus sont assez bien alignés.

Liste des graphiques.

Tableau et graphique	I. — Distribution de la surface des terres émergées en fonction de l'altitude (pour la terre entière).
—	II. — Distribution de la surface des terres émergées en fonction de l'altitude.
—	III. — Distribution des districts en fonction de leurs altitudes.
—	IV. — Distribution des lacs en fonction de leurs surfaces (pour la terre entière).
—	V. — Première distribution des îles marines en fonction de leurs surfaces.
—	VI. — Seconde distribution des îles marines en fonction de leurs surfaces.
—	VII. — Distribution des cours d'eau en fonction de leurs longueurs.

Les graphiques II et III ci-dessous ne sont reproduits que pour mémoire et ont été séparés des précédents parce que ces tableaux n'avaient été présentés dans le mémoire cité que pour montrer l'analogie de la distribution des altitudes suivant leurs nombres (tableau II) ou suivant les surfaces qu'elles occupent (tableau III). On constate, en effet, que les graphiques correspondants ont des formes paraboliques analogues. Mais étant relatifs tous deux à un petit territoire, ils ne pouvaient constituer de bons exemples du second type fondamental et, pour la distribution des altitudes, le tableau I relatif à la terre entière doit leur être préféré.

Expression analytique. — Par des calculs rapides et approximatifs où nous n'avons pas cherché à réduire le plus possible les écarts entre nombres observés

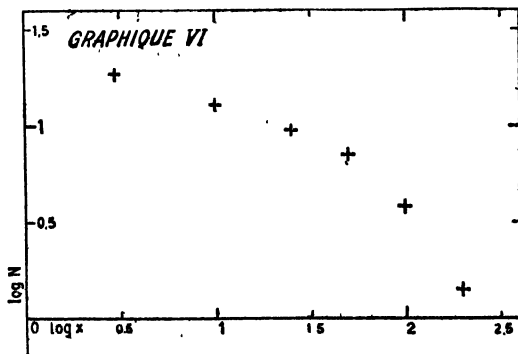
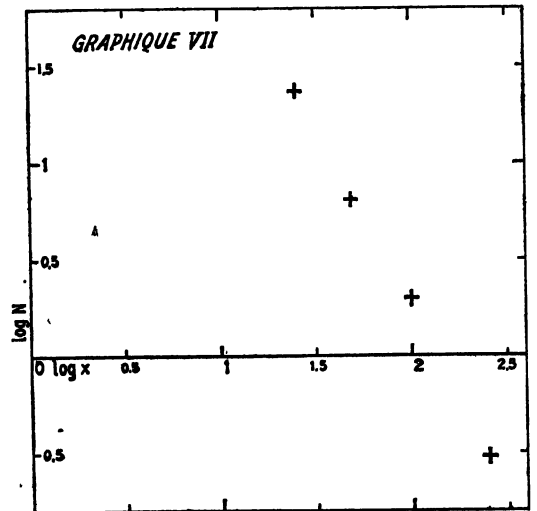
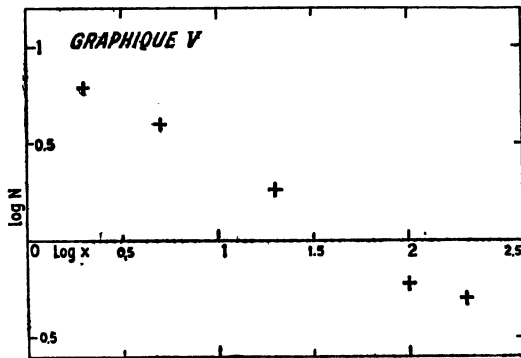
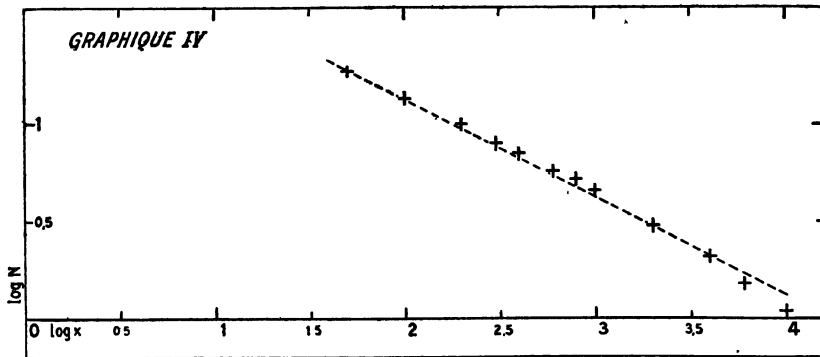
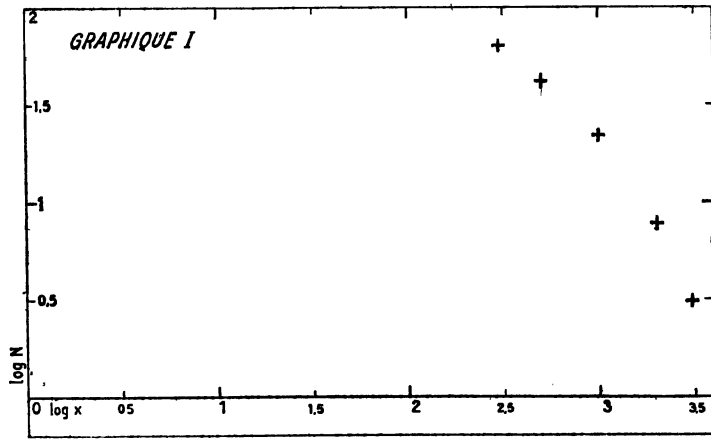
(1) On avait déjà constaté que la répartition des villes, suivant leur population, obéit assez bien à la loi de PARETO définie quelques lignes plus loin.

(2) Son expression analytique est de la forme :

$$N = \frac{A}{x^\alpha}$$

où A et α sont deux constantes positives.

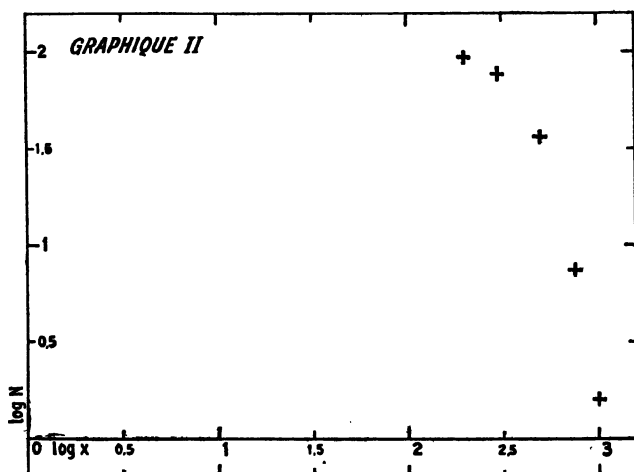
(3) Ces graphiques ont été dressés par le *Laboratoire de calcul de la Faculté des Sciences de Paris (Institut Henri Poincaré, 11, rue Pierre Curie, V^e)*. Ce laboratoire est mis à la disposition des savants, des techniciens et des industriels pour effectuer tous les calculs numériques exigeant quelques connaissances mathématiques. Il pourra rendre de grands services, en particulier aux statisticiens.



et nombres calculés, nous avons pu nous assurer que la distribution des surfaces des lacs (tableau IV) se conforme particulièrement bien à la loi de PARETO. On obtient une remarquable approximation en écrivant :

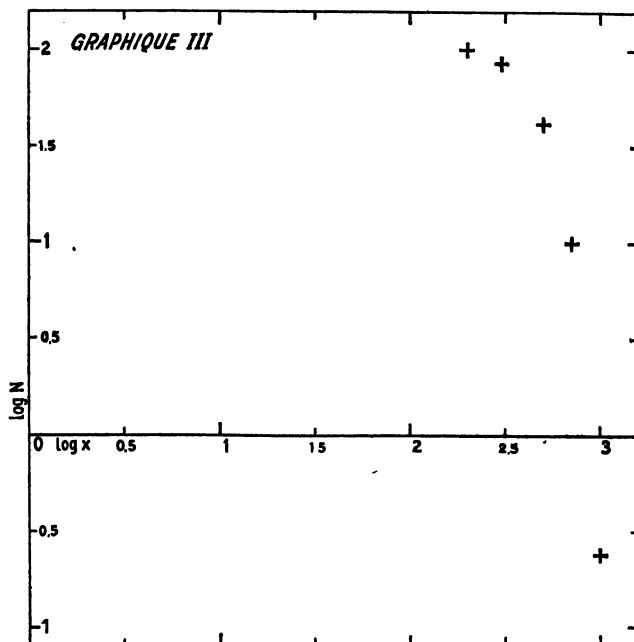
$$\log N = 2,09 - 0,48 \log x,$$

formule qu'on pourrait d'ailleurs améliorer. En écrivant au-dessous des pour-



centages observés qui figurent dans l'article cité, les pourcentages calculés par cette formule, on a le tableau suivant :

Pourcentage observé .	81,5	5,0	3,6	2,0	1,0	1,3	0,5	0,6	1,6	0,9	0,6	0,4	1,4
Pourcentage calculé .	81,4	5,1	3,7	1,7	1,2	1,2	0,7	0,5	1,3	0,9	0,4	0,7	1,2



En groupant les intervalles pour compenser les irrégularités dues au hasard, on obtient une approximation encore meilleure sous la forme suivante :

Surface en km ²	moins de 50	de 50 à 100	de 100 à 200	de 200 à 600	de 600 à 2.000	plus de 2.000
Nombre des lacs en % { observé .	81,5	5,0	3,6	4,3	2,7	3
calculé .	81,4	5,1	3,7	4,1	2,5	3,2

Avec une approximation moins bonne, on obtient les formules :

$$\log N = 1,74 - 0,6 \log x \quad (\text{tableau VI}),$$

$$\log N = 0,90 - 0,4 \log x \quad (\text{tableau V}),$$

$$\log N = 3,6 - 1,67 \log x \quad (\text{tableau VII}),$$

$$\text{et, enfin, } \log N = 4,55 - 1,11 \log x \quad (\text{tableau I}),$$

pour la répartition des altitudes dans le monde, les coefficients de ces formules, obtenus par une méthode rapide, pouvant d'ailleurs être améliorés.

Conclusion. — Les valeurs particulières des coefficients de ces formules, le fait même que la relation entre $\log N$ et $\log x$ est linéaire, ne sont peut-être pas d'un grand intérêt pour les géographes (1). Nous croyons donc utile de faire ressortir celles des conclusions de cette étude qui sont susceptibles de retenir leur attention. D'une part, tout ce à quoi on pourrait s'attendre d'après la définition générale de N , c'est que $\log N$ soit une fonction décroissante de $\log x$. *A priori*, les points représentatifs pourraient donc être dispersés très irrégulièrement pourvu qu'ils satisfassent à cette condition, c'est-à-dire qu'ils descendent en s'éloignant vers la droite. Or, il résulte déjà du mémoire cité : 1° que ces points se répartissent pour chaque tableau sur une courbe assez régulière; 2° que ces courbes régulières qui concernent respectivement les répartitions de grandeurs géographiques très différentes (aire d'un lac, altitude d'une montagne, longueur d'un cours d'eau) ont cependant des formes analogues; 3° que cette forme commune ne ressemble pas à celle de la répartition des erreurs d'observation. Le présent article fait, en outre, ressortir : 4° que la forme commune de ces répartitions géographiques est analogue à celle de la répartition des richesses ou des capacités intellectuelles; 5° qu'en s'inspirant de la signification générale de N , il a été possible de prévoir *a priori* cette analogie (comme cela a été, en fait, notre cas) et ensuite de la vérifier. Nous insistons donc sur le fait qu'au lieu, — ce qui serait d'ailleurs une méthode légitime, — de traduire chaque tableau numérique par une formule dont le choix serait dicté par la considération des *valeurs* des nombres du tableau, nous avons, à l'inverse, pu, en nous basant sur la *signification* des variables envisagées, *prévoir* la forme de la formule représentative, et ensuite vérifier cette prévision.

Remarque. — Il ne faut d'ailleurs pas oublier que, même pour la répartition

(1) Au moment de la correction des épreuves, nous constatons que notre texte se trouvait ainsi d'accord, par avance, avec l'une des intéressantes remarques du baron Murre lors de sa dernière conférence (sur la loi de Pareto). Il observait que ce qui intéressait l'économiste, dans une formule, n'était pas tant sa rigoureuse exactitude que sa facilité d'interprétation.

des fortunes dont l'étude lui a donné naissance, la formule de PARETO ne doit pas être considérée comme finale. Son approximation n'est pas très bonne aux deux extrémités de l'échelle et de nombreuses formules ont été proposées — entre autres, par PARETO lui même, — pour obtenir une meilleure approximation. Mais l'étude qui vient d'être faite au moyen de la première formule de PARETO suffit pour établir le fait essentiel que les répartitions de nombreuses grandeurs géographiques sont semblables entre elles et analogues à celle de la richesse. Si l'on trouve une formule remplaçant avantageusement celle de PARETO pour la répartition des fortunes, il n'est pas certain que cet avantage se conserve pour la répartition des grandeurs géographiques, mais on aura intérêt à chercher au moins à voir s'il en est ainsi.

Maurice FRÉCHET.

DISCUSSION

M. RISSER dit que la communication si intéressante et si vivante de M. FRÉCHET lui suggère quelques remarques suivantes :

On sait que l'emploi conjugué des valeurs de t apparaissant dans la formule représentative de $\Theta(t)$ et des fréquences, nous donne le moyen de vérifier si la loi de dispersion d'une série statistique est normale; si la dispersion est bien du type Laplace GAUSS, les points du diagramme utilisé sont en ligne droite. Le graphique fournit alors, avec une approximation suffisante, la valeur moyenne de la mesure, son écart probable et son écart moyen, ainsi que le module de précision caractérisant les observations.

Lorsqu'on se trouve en présence d'une répartition asymétrique, définie par la fonction $\omega(x)$ par rapport à la grandeur x , on cherche à déterminer une fonction $z = f(x)$ telle que l'on ait $z = \frac{1}{\sqrt{\pi}} e^{-z^2} dz = \omega(x) dx$.

Si l'effet Δx d'une cause est indépendante de la grandeur x , la fréquence des écarts des grandeurs z suivra la loi de Laplace. Si, au contraire, cet effet dépend de x , c'est à dire s'il est défini par une relation de la forme $\Delta x = \lambda f(x)$, λ ne dépendant pas de x , mais de la cause seule, la répartition cessera d'être symétrique. En ce cas, le rapport $\frac{\Delta x}{f x}$ étant indépendant de x , on peut le considérer comme la variation Δz d'une quantité z ; l'effet des causes sur z sera indépendant de x , et z aura tous les caractères d'une répartition normale, à la condition toutefois que la médiane des grandeurs z soit nulle.

C'est cette idée apparaissant dans un mémoire de EDGEWORTH, et reprise ultérieurement par KAPTEYN, puis par VAN UVEN, qui a permis de ramener la répartition asymétrique à la répartition normale.

KAPTEYN a étudié tout spécialement les cas où :

1° L'effet des causes élémentaires est proportionnel à la grandeur x (avec $\Delta x = x \frac{\Delta z}{h}$);

2° L'effet des causes élémentaires est proportionnel à $(X - X_0^2)$.

La méthode de cet astronome est très générale et permet d'interpréter les répartitions de fréquences les plus diverses; seules les difficultés du calcul pourraient en limiter l'application.

Que l'ajustement soit effectué analytiquement ou graphiquement, il importe d'évaluer les écarts entre les valeurs observées et les valeurs calculées des fréquences; si ces écarts sont faibles, on a intérêt à utiliser la première transformation de Kapteyn, attendu qu'elle n'introduit qu'un paramètre.

M. RISSER a d'ailleurs eu l'occasion de l'appliquer à des problèmes relatifs respectivement à l'assurance accidents du travail et à l'assurance automobile.

M. FRÉCHET répond qu'à la fin de sa communication, il s'était précisément associé d'avance, mais sous une forme plus générale, au vœu exprimé par M. RISSER.

Il avait dit qu'il serait désirable de s'assurer si l'on obtient, pour la représentation de la répartition des données géographiques, un meilleur résultat que celui fourni par la formule de PARETO, non seulement au moyen de la formule de MAC ALISTER-GIBRAT recommandée par M. RISSER mais, aussi, par les différentes formules proposées pour remplacer celles de Pareto (1).

Il signale, à cette occasion, qu'on paraît parfois attribuer une certaine vertu (M. Risser n'a d'ailleurs rien dit de semblable) aux lois de probabilités obtenues par une transformation de la variable dans la seconde loi de Laplace (appelée à tort loi de GAUSS). Il doit être bien entendu que ces lois pourraient être aussi obtenues (par une transformation nécessairement différente de la variable) à partir d'une loi simple autre que la seconde loi de LAPLACE. M. FRÉCHET rappelle à ce sujet qu'il a indiqué, dans l'article cité en note ci-dessus, le curieux résultat suivant : en remplaçant x par $a \log (x - x_0) + b$, non seulement on obtient la loi de MAC ALISTER-GIBRAT à partir de la deuxième loi d'erreur de LAPLACE (probabilité élémentaire $\frac{1}{\sqrt{\pi}} e^{-x^2} dx$), mais encore on obtient la loi de Pareto elle-même, à partir de la première loi d'erreur qui avait été proposée par Laplace (probabilité élémentaire $\frac{1}{2} e^{-x} dx$ du moins pour $a \log (x - x_0) + b > 0$).

M. FRÉCHET pense, d'ailleurs, que les meilleures lois représentatives sont celles qui résultent d'un essai d'explication théorique, même si celle-ci n'est pas très convaincante, plutôt que celles qui sont choisies uniquement pour représenter, avec une bonne précision, un tableau numérique de données.

M. René ROY fait observer que c'est peut être l'avantage majeur que présente la loi de l'effet proportionnel, utilisée par GIBRAT pour la description des inégalités économiques; il s'agit d'une transformation logarithmique de la loi normale, qui trouve ainsi sa justification dans le recours à une théorie, en dehors de la bonne adaptation des ajustements aux données de l'observation.

M. Huber signale, dans un ordre d'idées voisin de celui qu'a évoqué M. Fréchet, une loi de concentration de la population présentée par M. Auerbach dans les *Petermanns Mitteilungen* de février 1913. Les villes d'un pays étant rangées par ordre de population décroissante, le produit du rang par la population serait constant. Ainsi, pour les villes allemandes, après quelques oscilla-

(1) On trouvera rappelées quelques unes d'entre elles dans l'article de M. Fréchet « sur les formules de répartition des revenus », paru dans la *Revue de l'Institut International de Statistique* de 1939.

tions pour les grandes villes, ce produit (population en centaines de mille par le rang) se fixerait aux environs de 48. Dans une note publiée dans le *Journal de la Société* d'août septembre 1913, il avait constaté que, pour la France, après quelques oscillations pour les grandes villes, le produit croissait au contraire de 15 pour les villes de 100.000 habitants à 33 pour les villes de 5.000 à 6.000 habitants. La loi ne paraît pas générale. Il serait intéressant toutefois de vérifier si la loi de répartition des villes et communes, suivant leur population, jouit des propriétés constatées par M. FRÉCHET pour d'autres distributions, relatives il est vrai, à la géographie physique et non à la géographie humaine.

M. Fréchet fait remarquer que les écarts des valeurs fournies par les formules d'ajustement par rapport aux données de l'observation présentent peu d'importance pour l'économiste, le géographe, etc..., et il insiste sur la nécessité, pour eux, de traduire avant tout l'allure générale de la distribution.

M. René ROY acquiesce entièrement à cette thèse qu'il a d'ailleurs mise en œuvre pour l'application à la demande de la formule simple de Pareto concernant la distribution des revenus.

* * *

M. Huber, ayant envoyé, après la séance, divers renseignements bibliographiques à M. Fréchet, celui ci nous a adressé la note ci-après, qui complète sa réponse à M. Huber.

M. Fréchet avait indiqué pendant la séance qu'il y aurait discordance entre les résultats de M. Auerbach et ceux de M. Huber, s'il s'agissait de l'application d'une même forme analytique de loi de répartition avec possibilité de choix différents des paramètres, mais que cette discordance n'en serait plus véritablement une si l'on avait imposé d'avance la même valeur du paramètre à deux pays différents.

La forme proposée par M. Auerbach était $Nx = \text{constante}$, N étant le nombre des villes dont la population est supérieure ou égale à x (c'est-à-dire le rang de la ville de population x quand les villes sont rangées par ordre de grandeur). Elle convenait à peu près à l'Allemagne et non à la France. Observons cependant d'abord, comme un fait assez remarquable, que le produit de deux nombres de natures si différentes ne varie, même pour la France, que de 15 à 33. Mais il y a mieux; en effet, la loi $Nx = \text{constante}$ n'est qu'un cas particulier de la loi de Pareto $Nx^\alpha = \text{constante}$, celui où $\alpha = 1$.

De sorte que nous nous trouvons précisément dans le cas prévu par M. FRÉCHET : il résulte des chiffres produits que la loi de Pareto s'applique assez bien à l'Allemagne en prenant $\alpha = 1$; mais, quant à la France, on devrait seulement conclure : ou bien qu'elle ne s'applique pas, ou bien qu'elle ne peut s'appliquer avec une valeur de α voisine de 1.

M. FRÉCHET a eu la curiosité de consulter, après la séance, une statistique donnant la répartition des villes selon la population pour un grand nombre de pays. Il a trouvé un tableau qui présente cet intérêt de donner aussi le nombre total R des habitants des villes de population supérieure

ou égale à x . Or, si $N = Ax^{-\alpha}$, on a $dR = x$, dN , d'où R , et par suite

$$\frac{R}{R - Nx} = \text{constante} = \alpha.$$

Il a donc fait calculer ce rapport pour les différentes valeurs de x du tableau; soit : pour $\alpha = 10.000$, 20.000 et 100.000 pour les pays du tableau, sauf quelques uns qui avaient donné d'autres valeurs de x ou dont les chiffres étaient indiqués comme « estimés ». En prenant, pour chaque pays restant, la médiane des trois valeurs calculées pour α , il a obtenu le tableau suivant :

Valeur de l'exposant α dans la répartition des villes suivant le nombre d'habitants au moyen de la formule $N = \frac{A}{x^\alpha}$.

Belgique.	1,65	France	1,34
Indes Britanniques.	1,51	Japon et Allemagne	1,23
Bulgarie:	1,48	Irlande	1,21
Hongrie.	1,43	Écosse.	1,20
Suisse et Pays Bas.	1,35		

Le petit nombre (trois) de valeurs de x employées ne permet pas d'attacher trop de signification à l'ordre des pays résultant de ces valeurs de α . Cependant, on vérifie que la valeur de α est en effet plus proche de l'unité pour l'Allemagne que pour la France. On voit, en tous cas, que toutes ces valeurs qui, *a priori*, pourraient s'étager de 1 jusqu'à l'infini, restent confinées dans des limites assez étroites de 1,13 à 1,65.

De même, pour chaque pays, les trois valeurs de α obtenues au lieu d'être égales, si la loi de Pareto était mathématiquement vérifiée, sont différentes comme on pouvait s'y attendre; mais leurs différences ne sont pas excessives.

Contentons nous de citer les deux bouts de l'échelle où l'on trouve :

	x	10.000	20.000	100.000
Belgique : α		1,57	1,65	2,14
Écosse : α		1,17	1,20	1,27

Il vaudrait la peine de reprendre ces calculs avec des statistiques plus complètes, telles que celle qu'a bien voulu me signaler M. HUBER : tome I du *Recensement de la France 1936* (relevé des populations des 837 villes de plus de 5.000 habitants).