JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

LUCIEN MARCH

Quelques exemples de distribution des salaires

Journal de la société statistique de Paris, tome 39 (1898), p. 241-248
<http://www.numdam.org/item?id=JSFS_1898__39__241_0>

IV.

QUELQUES EXEMPLES DE DISTRIBUTION DES SALAIRES (*fin*) [1].

Les différences entre les logarithmes calculés et les logarithmes exacts sont assez sensibles. En valeur absolue, la différence moyenne représente 17 p. 100 de la valeur moyenne du logarithme exact. Cela indique que la ligne des logarithmes n'est pas très convenablement interpolée par la courbe (2). Mais, au point de vue de la ligne des nombres y, l'ajustement est dépourvu de toute valeur. C'est qu'en effet, en rendant minimum la somme des carrés des différences entre les logarithmes exacts et les logarithmes calculés, on n'abaisse pas du tout au minimum la somme des carrés des différences entre les nombres.

Dans un numéro récent de ce Journal [2], M. Vilfredo Pareto proposait, pour améliorer le résultat, d'appliquer la méthode des moindres carrés à l'équation aux logarithmes dont les deux nombres auraient, au préalable, été multipliés par y. Nous ferons un peu plus loin une application de cette idée. Dans le cas présent, les calculs à faire seraient notablement accrus, et il nous paraît bien préférable d'appliquer tout simplement la méthode à un nombre limité d'observations, à celles qui ont le plus de poids. C'est ce que nous allons faire maintenant.

L'application de la méthode des moindres carrés aux équations de la forme (2) établies en donnant à x les valeurs allant de 4 à 15, conduit aux valeurs suivantes pour les constantes :

$$(4) \qquad \begin{cases} A = -\,0,2130 \\ B = 9,8926 \\ C = -\,0,5552 \end{cases}$$

On remarquera que ces valeurs sont très différentes de celles trouvées précédemment, ce qui confirme l'observation faite plus haut touchant l'emploi de la méthode d'approximation ; il semblerait, au premier abord, que les valeurs trouvées pour les constantes, lorsqu'on se borne à une partie notable des observations, devraient pouvoir être considérées comme des valeurs approchées de celles calculables pour l'ensemble des observations : il n'en est rien, comme on le voit.

Calculons maintenant les valeurs des logarithmes et des nombres au moyen de la nouvelle formule.

x.	log y calculé.	δ.	y calculé.	Δ.
1	$\bar{1},232$	+ 2 637	0,17	+ 74
5	3,926	+ 147	8 420	+ 3 425
10	4,127	— 118	13 410	— 3 195
15	3,093	+ 142	1 240	+ 480
20	1,556	+ 1 016	36	+ 337
25	$\bar{1},732$	+ 2 777	0,54	+ 323
30	$\bar{3},743$	+ 3 536	0,05	+ 19
35	$\bar{5},620$	+ 5 516	0,00	+ 14
40	$\bar{7},432$	+ 7 682	0,00	+ 13
45	$\bar{9},158$	+ 9 883	0,00	+ 11

(1) Voir *Journal de la Société de statistique de Paris*, juin 1898, page 193.

(2) *Journal de la Société de statistique de Paris*, novembre 1877, page 372.

Sans qu'il soit nécessaire de calculer le module des écarts dans chaque cas, il est visible que la ligne des logarithmes est bien plus mal représentée par la formule (4) que par la formule (3), tandis qu'au contraire la formule (4) fournit des valeurs de y plus approchées des valeurs exactes que la formule (3) ; la moyenne des valeurs absolues des écarts est égale à 789 au lieu de 1 736. Cela s'explique aisément.

B. En appliquant la méthode de Cauchy aux équations (2) pour les valeurs de x allant de 4 à 15, nous avons obtenu les valeurs suivantes des constantes :

$$(5) \qquad \begin{cases} A = -\ 0,3057 \\ B = \ \ 10,3822 \\ C = -\ 0,5941 \end{cases}$$

La même méthode appliquée à l'équation obtenue en multipliant par y les deux membres de chacune des équations (2), conduit aux valeurs

$$(6) \qquad \begin{cases} A = \ \ 1,4963 \\ B = \ \ 7,353 \\ C = -0,4782 \end{cases}$$

et le tableau des valeurs calculées, dans les deux cas, s'établit ainsi, pour les valeurs particulières de x que nous avons déjà choisies.

x.	Formule (5).				Formule (6).			
	log y.	δ.	y.	Δ.	log y.	δ.	y.	Δ.
1	$\bar{1}$,100	+ 2 768	0,12	+ 74	1,018	+ 851	10	+ 64
5	3,980	+ 93	9 560	+ 2 285	4,245	— 172	17 600	— 5 755
10	4,135	— 126	13 650	— 3 435	4,067	— 58	11 670	— 1 455
15	2,993	+ 242	985	+ 735	2,971	+ 264	936	+ 784
20	1,320	+ 1 252	21	+ 352	1,499	+ 1 073	31	+ 342
25	$\bar{1}$,355	+ 3 154	0,22	+ 323	$\bar{1}$,820	+ 2 689	0,66	+ 321
30	$\bar{3}$,207	+ 4 072	0,01	+ 19	$\bar{3}$,985	+ 3 294	0,01	+ 19
35	$\bar{6}$,931	+ 6 213	0,00	+ 14	$\bar{4}$,113	+ 5 033	0,00	+ 14
40	$\bar{8}$,562	+ 8 552	0,00	+ 13	$\bar{7}$,722	+ 7 392	0,00	+ 13
45	$\overline{10}$,123	+ 10 917	0,00	+ 11	$\bar{8}$,133	+ 8 908	0,00	+ 11

Les résultats de l'application simple de la méthode de Cauchy sont tout à fait analogues à ceux que nous avait fournis l'application de la méthode des moindres carrés pour les valeurs de x allant de 4 à 15 ; les valeurs de y calculées au moyen de cette méthode sont même probablement plus approchées ; la moyenne des valeurs absolues des écarts, pour les 10 valeurs particulières de x considérées, est en effet égale à 726 au lieu de 789.

L'application de la méthode de Cauchy, après multiplication par y des deux nombres de l'équation (2), fournit des valeurs meilleures des logarithmes, mais des valeurs moins satisfaisantes pour les nombres.

C. Appliquons maintenant une troisième méthode préconisée par le professeur Pearson [1], qui en a fait des applications nombreuses et des plus intéressantes. Elle consiste à identifier la somme des moments par rapport à l'axe des y, soit des

(1) *Philosophical transactions of the Royal Society of London*, 1895 : « Contribution to the mathematical theory of evolution. — II. Skew variations in homogeneous material. »

différents points de la ligne des observations, soit des différents points de la courbe définie par les constantes à calculer.

Ces moments sont de différents ordres : Σy est le moment d'ordre 0, $\Sigma y x$ est du premier ordre, $\Sigma y x^2$ du second, $\Sigma y x^3$ du troisième, etc. On calcule autant de moments d'ordre différent qu'il y a de constantes à déterminer, et il est évident que plus on prend des moments d'ordre élevé, plus on donne d'importance aux observations éloignées de l'axe des moments choisi.

Dans l'exemple qui nous occupe, si l'on cherche à déterminer les trois constantes au moyen des moments d'ordre allant jusqu'au troisième, pris par rapport à l'axe des y, l'influence des observations relatives aux grandes valeurs de x domine complètement celle des observations relatives aux petites valeurs de x, en sorte que l'on obtient une courbe assez bien ajustée à une extrémité de la ligne des observations et mal ajustée à l'autre. Ainsi, on est conduit à attribuer à B une valeur négative, c'est-à-dire que la courbe, au lieu de partir de l'origine, comme la ligne des observations, est asymptote de l'axe des y.

Il convient donc de laisser de côté les moments d'ordre supérieur au second et de calculer les constantes au moyen des moments d'ordre 0, 1 et 2.

Les valeurs alors obtenues sont :

$$\left\{\begin{array}{l} A = 2,3331 \\ B = 4,5956 \\ C = -0,2852 \end{array}\right. \tag{7}$$

Voici le tableau des logarithmes et des nombres calculés pour les valeurs de x précédemment considérées:

x	$\log y$	δ.	y	Δ
1	2,048	— 179	112	— 38
5	4,119	— 46	13 160	— 1 315
10	4,077	— 68	11 950	— 1 735
15	3,460	— 225	2 887	— 1 167
20	2,609	— 37	407	— 34
25	1,628	+ 881	43	+ 280
30	0,566	+ 713	4	+ 15
35	$\bar{1}$,448	+ 1 698	0,3	+ 14
40	$\bar{2}$,189	+ 2 825	0,0	+ 13
45	$\bar{3}$,098	+ 3 943	0,0	+ 11

Sans calculs plus développés, on voit à première vue que les logarithmes calculés par cette méthode sont bien moins approchés que ceux calculés par la formule (3), mais ils sont plus approchés que ceux obtenus à l'aide des autres formules précédentes.

En ce qui concerne les nombres, il est vraisemblable que les résultats fournis par la formule (7) sont plus voisins de l'exactitude que tous ceux obtenus précédemment. Prenons comme indice d'exactitude la moyenne des valeurs absolues des écarts et nous constatons que cette moyenne est, dans les différents cas :

Formule (3). . . .	1 736
— (4). . . .	789
— (5). . . .	726
— (6) . . .	878
— (7). . . .	462

Sans prétendre aucunement que ces coefficients calculés pour 10 valeurs seulement de x caractérisent la précision relative des formules, on doit néanmoins reconnaître, vu la régularité suffisante de la ligne des observations, que la dernière est sans aucun doute celle qui conduit à la courbe la plus voisine de cette ligne, dans son ensemble. Ajoutons encore une indication intéressante.

Si l'on calcule la surface de chacune des courbes déterminées par nos formules [1], l'unité de mesure des longueurs étant la différence de deux valeurs consécutives de x, sur le tableau de distribution des salaires, on arrive aux résultats suivants :

		Surface
Ligne des observations (nombre d'ouvriers observés)		134 487
Courbe correspondant a la formule	(3)	41 800
	(4)	116 400
	(5)	125 500
	(6)	139 200
	(7)	134 487

Remarquons que la formule (7) devait nécessairement conduire à une surface rigoureusement égale à celle de la ligne des observations, car, dans cette formule, A est calculé au moyen du moment d'ordre 0, $\Sigma\, y \times x^0$, qui n'est autre chose que le nombre des ouvriers observés.

Supposons qu'un dessinateur cherche à tracer une courbe au sentiment épousant le mieux possible la forme générale de la ligne des observations, il s'efforcera de répartir la surface comprise entre cette ligne et la courbe à tracer, de manière à établir une compensation aussi parfaite que possible, des surfaces intérieures et extérieures à la courbe dans les diverses parties du diagramme.

Un mathématicien qui voudrait appliquer ce procédé au moyen de formules devrait déterminer les constantes de la courbe par la condition que les surfaces comprises entre certaines ordonnées déterminées et, d'une part la ligne des observations, d'autre part la courbe, soient égales. Cette méthode serait peut-être applicable à une courbe représentée par une équation algébrique, mais elle ne peut convenir aux équations de la forme (1), à cause de l'impossibilité d'effectuer les intégrations ou de l'extrême complication des calculs d'approximation auxquels il faudrait se livrer.

Mais, si, entre plusieurs courbes de cette forme, on choisit celle dont la surface totale se rapproche le plus de celle de la ligne des observations et qui, en même temps, est telle que la surface interposée entre elle et la ligne des observations est moyennement la plus petite, on aura évidemment choisi, parmi ces courbes, celle qui sera la plus conforme à la courbe au sentiment du dessinateur.

Pour cette raison, la formule (7), obtenue par la méthode des moments du professeur Pearson, sera théoriquement préférée et à celles qui résultent de l'application, soit de la méthode de Legendre, soit de la méthode de Cauchy, à l'équation aux logarithmes (2).

La méthode des moments est d'ailleurs la plus simple à appliquer, parce que

(1) La surface de chaque groupe est donnee par la formule

$$\log S = A - (B+1)\log - \left(\frac{c}{\log e}\right) + \log \Gamma\,(B+1).$$

les calculs à faire sont, pour la plupart, purement mécaniques, grâce en partie aux formules établies par le professeur Pearson. Elle fournit donc aussi la courbe qui, *pratiquement*, représente le mieux possible l'allure générale de la ligne des observations.

De nombreux exemples, donnés par le professeur Pearson dans ses mémoires, montrent que l'ajustement obtenu par sa méthode est presque toujours très satisfaisant lorsque l'on peut employer les moments du 3e et du 4e ordre. Dans le cas présent, on n'a pu employer que les moments d'ordre 0, 1 et 2; aussi l'ajustement de la courbe représentée par la formule (7) à une partie seulement de la ligne des observations est-il tout à fait insuffisant.

2. — *Ajustement d'une partie de la courbe.*

Pour comparer les diverses méthodes dont il vient d'être question au point de vue de l'ajustement d'une portion de courbe, considérons la partie comprise entre les ordonnées correspondant à $x=4$ et $x=15$, et calculons les valeurs de log y fournies par l'application des diverses formules ; en voici le tableau, ainsi que celui des écarts δ, calculés en millièmes, par rapport aux valeurs exactes de log y :

x	Valeurs exactes de log y	Formule (3).		Formule (4).		Formule (5)		Formule (6).		Formule (7).	
		log y	δ	log y	δ	log y	δ	log y.	δ.	log y	δ.
4	3,179	3,366	— 187	3,523	— 344	3,568	— 389	4,010	— 831	3,959	— 780
5	4,073	3,442	+ 631	3,926	+ 147	3,980	+ 93	4,245	— 172	4,119	— 46
6	4,422	3,482	+ 940	4,154	+ 268	4,209	+ 213	4,348	+ 74	4,198	+ 224
7	4,416	3,497	+ 919	4,261	+ 155	4,307	+ 109	4,362	+ 54	4,221	+ 195
8	4,295	3,493	+ 802	4,279	+ 16	4,318	— 23	4,311	— 16	4,202	+ 93
9	4,174	3,476	+ 698	4,230	— 56	4,255	— 81	4,209	— 35	4,183	— 9
10	4,009	3,448	+ 552	4,128	— 119	4,135	— 126	4,067	— 58	4,077	— 68
11	3,949	3,411	+ 538	3,982	— 33	3,971	— 22	3,895	+ 54	3,982	— 33
12	3,701	3,366	+ 335	3,801	— 100	3,769	— 68	3,693	+ 08	3,870	— 169
13	3,537	3,315	+ 222	3,589	— 52	3,535	+ 2	3,471	+ 66	3,745	— 208
14	3,331	3,260	+ 71	3,352	— 21	3,229	+ 102	3,229	+ 102	3,608	— 277
15	3,235	3,199	+ 36	3,093	+ 142	2,993	+ 242	2,971	+ 264	3,460	— 225

La moyenne des valeurs absolues des écarts est, dans chaque cas :

$$\text{Formule}\left\{\begin{array}{ll}(3)\ldots & 494\\ (4)\ldots & 121\\ (5)\ldots & 122{,}5\\ (6)\ldots & 144{,}5\\ (7)\ldots & 177\end{array}\right.$$

Ici, les valeurs calculées les plus exactes sont, ainsi que cela devait être, celles calculées par l'application de la méthode des moindres carrés pour les valeurs de x allant de 4 à 15 (formule 14), mais la méthode de Cauchy, appliquée à l'équation (2) sous sa forme simple (formule 5), fournit des valeurs moyennement presque aussi exactes. La méthode de Cauchy, donnant lieu à des calculs moins longs et susceptibles de vérifications multipliées, semble donc en pratique devoir être souvent préférée, comme méthode d'ajustement, à celle de Legendre.

On prend généralement pour mesure de la précision l'expression $\frac{1}{2h^2}$, h étant égal à $\sqrt{\frac{\Sigma\Delta^2}{n-3}}$. Cette dernière expression est égale, dans le cas de la formule (4), à 56,5, et, dans le cas de la formule (5), à 59. La précision de la méthode des moindres carrés est donc à celle de la méthode de Cauchy comme 59 est à 56,5.

La différence n'est pas grande, quoique proportionnellement plus importante que celle trouvée ci-dessus en comparant les moyennes des valeurs absolues des écarts. Il nous paraît que la moyenne des valeurs absolues, laquelle a simplement le défaut de ne pouvoir s'exprimer algébriquement, donne une idée meilleure de la précision de l'ajustement que la moyenne des carrés; celle-ci, en effet, exagère l'importance des grands écarts, les moins nombreux.

Voyons maintenant ce que sont les nombres dont on a donné ci-dessus le tableau des logarithmes. (Il n'est pas utile de calculer ceux relatifs à l'emploi de la formule [3] trop peu approchée.)

x	Valeurs exactes de y	Formule (4).		Formule (5)		Formule (6)		Formule (7).	
		y	Δ	y	Δ	y	Δ.	y	Δ
4	1 509	3 540	— 2 031	3 700	— 2 191	10 240	— 8 731	9 100	— 7 591
5	11 845	8 440	+ 3 405	9 550	+ 2 295	17 580	— 5 735	13 160	— 1 315
6	26 404	14 260	+ 12 144	16 200	+ 10 204	22 290	+ 4 114	15 780	+ 10 624
7	26 054	18 240	+ 7 814	20 280	+ 5 774	23 020	+ 3 034	16 640	— 9 414
8	19 730	19 020	+ 710	20 800	— 1 070	20 470	— 740	15 920	+ 3 810
9	14 922	17 000	— 2 078	18 000	— 3 078	16 190	— 1 268	15 250	— 328
10	10 215	13 430	— 3 215	13 650	— 3 435	11 670	— 1 455	11 940	— 1 725
11	8 894	9 600	— 706	9 360	— 466	7 860	+ 1 034	9 600	— 706
12	5 020	6 330	— 1 310	5 880	— 860	4 930	+ 90	7 415	— 2 395
13	3 442	3 880	— 438	3 430	+ 12	2 960	+ 482	5 560	— 2 118
14	2 143	2 250	— 107	1 695	+ 448	1 695	+ 448	4 056	— 1 913
15	1 720	1 240	+ 520	985	+ 735	936	+ 784	2 885	— 1 165

La moyenne des valeurs absolues des écarts est, pour chacune des formules :

Formule	(4) . .	2 873
	(5). .	2 547
	(6) . .	3 076
	(7). . .	3 592

Ainsi, la méthode de Cauchy, appliquée à l'équation aux logarithmes (2), donne pour les nombres des valeurs plus approchées que la méthode des moindres carrés appliquée à la même équation. Ce résultat avait déjà été constaté ci-dessus, pour l'ensemble des observations; il avait été signalé, pour un autre exemple, par M. Vilfredo Pareto.

Résumons les conclusions qui résultent des calculs précédents.

1° Pour obtenir une courbe caractérisant le mieux possible l'allure générale d'une ligne d'observations statistiques, la méthode la plus pratique, et celle qui conduit aux résultats les plus satisfaisants quand on ne peut connaître de valeurs approchées des constantes à déterminer, semble être la méthode des moments du professeur Pearson, et ce dernier a justifié ce choix par de nombreux exemples dans les cas où l'on peut utiliser des moments d'ordre supérieur au second.

L'exemple actuel montre qu'il en est encore ainsi dans un cas où l'on doit se contenter des moments d'ordre 0, 1 et 2.

2° S'il s'agit d'obtenir un ajustement aussi parfait que possible à une fraction de la ligne des observations, la méthode de Cauchy apparaît la plus pratique, et elle semble même la plus exacte lorsque l'on est forcé de passer par les logarithmes des coordonnées.

3° Dans l'application de la méthode des moindres carrés ou de la méthode de Cauchy à un ensemble d'observations statistiques dont l'on veut représenter la tendance générale par une formule, au lieu de donner aux équations formées pour l'application de l'une de ces méthodes un poids proportionnel au nombre des observations, il est préférable d'abandonner un certain nombre de ces équations. On pourrait prendre pour règle d'abandonner toutes celles qui correspondent à la partie de la courbe en dehors des points de changement de courbure.

3. — *Détermination de la valeur normale du salaire.*

La ligne des observations ayant été remplacée par une courbe dont l'équation a la forme

$$y = \alpha x^{\beta} e^{-\gamma x},$$

il est facile de voir que cette courbe présente un sommet et que ce sommet correspond à une abcisse égale à $\frac{\beta}{\gamma}$ ou, en fonction de B et C, à $\frac{B}{C}\log e$.

Si la distribution des salaires est effectivement représentée par cette courbe, il est naturel de regarder l'abcisse du sommet de la courbe comme représentant le salaire normal, car une courbe de ce genre n'a qu'un seul maximum.

En adoptant cette manière de voir, la valeur normale du salaire correspondant aux diverses formules que nous avons envisagées est, dans chaque cas :

Formule	(3). . .	7,28
	(4). . .	7,73
	(5). . .	7,59
	(6) . .	6,67
	(7). . .	7,00

Ces valeurs ne sont pas très différentes. Bien que, dans les différentes formules, la valeur de B passe de 2,03 à 10,38, la position du maximum déterminée par la relation

$$x = \frac{B}{C}\log e$$

ne varie pas beaucoup.

Néanmoins, il faut choisir entre ces valeurs, et il nous paraît naturel de prendre celle qui correspond à l'emploi de la méthode des moments. En voici les raisons.

D'abord, nous avons vu que cette méthode est d'application simple. En second lieu, toutes les fois (et c'est le cas le plus général en statistique) qu'on ne peut appliquer directement les autres méthodes, ni connaître d'avance des valeurs approchées des constantes, elle paraît la plus exacte si l'on veut tenir compte de toutes les observations. Or, il semble tout à fait rationnel et conforme à l'idée qui s'attache au mot normal ou type, de tenir compte de toutes les observations.

Il n'est d'ailleurs pas nécessaire, pour déterminer la position du sommet de la courbe de distribution représentée par la formule (7), d'en calculer les constantes. On démontre aisément que l'ordonnée maxima est à une distance de l'ordonnée qui passe par le salaire moyen égale à $\frac{\Sigma y X^2}{\Sigma y x}$, X représentant l'écart entre chaque taux de salaire considéré et le salaire moyen, y le nombre des ouvriers à chaque taux de salaire et x le taux du salaire considéré.

$\Sigma y x$ est le salaire total distribué, et $\Sigma y X^2$ se calcule aisément au moyen de $\Sigma y x$ et de $\Sigma y x^2$, à l'aide de la formule $\Sigma y X^2 = \Sigma y x^2 - (\Sigma y x)^2 \times \frac{1}{\Sigma y}$.

Lucien MARCH,
Ingénieur à l'Office du travail.