

REVUE DE STATISTIQUE APPLIQUÉE

H. PIN

Méthode graphique de séparation de deux populations

Revue de statistique appliquée, tome 5, n° 1 (1957), p. 77-83

http://www.numdam.org/item?id=RSA_1957__5_1_77_0

© Société française de statistique, 1957, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) 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/>

MÉTHODE GRAPHIQUE DE SÉPARATION DE DEUX POPULATIONS

par

H. PIN

Le problème de la « dissection » d'une courbe de fréquence représentant le résultat du mélange de deux distributions normales a évidemment — si cette hypothèse de normalité est fondée — une solution précise (1), tout au moins sur le plan théorique.

Mais la recherche de six inconnues — trois pour chaque distribution normale — conduit à écrire un système compliqué de six équations dont la résolution est laborieuse et qui a d'ailleurs l'inconvénient de faire appel aux valeurs numériques des moments d'ordre 1 à 5 de la distribution observée. Ces moments, calculés à partir d'un échantillon réduit peuvent être — particulièrement les moments d'ordre élevé — entachés d'erreurs importantes.

La méthode graphique proposée par l'auteur et dont le principe a déjà été utilisé sous une forme différente donne rapidement, si les circonstances ne sont pas trop défavorables, une solution approchée du problème.

Il arrive que l'on se trouve en présence d'une distribution que l'on soupçonne être constituée par deux populations normales, en partie confondues si bien que l'ensemble se trouve distribué suivant une loi de répartition inconnue.

Des considérations techniques font naître ce soupçon. On examine, par exemple, une production issue de très nombreuses machines (plusieurs milliers) de même type, non surveillées individuellement par carte de contrôle : c'est le cas dans l'industrie des textiles artificiels. Alors qu'on s'attendait à trouver une distribution normale habituellement observée pour des productions analogues, on constate une distribution qui s'en éloigne notablement, comme celle de la figure 1. On en vient à se demander si les machines ne donneraient pas deux types différents de production et s'il ne serait pas possible d'isoler ces deux populations, en essayant d'ajuster aussi bien que possible la courbe observée à deux courbes normales. Le problème serait résolu si l'on connaissait, pour chacune de ces deux populations, trois éléments : la moyenne m , l'écart-type σ , le nombre N d'individus.

(1) Strömgren - Tables and Diagrams for dissecting a frequency curve into components by the half invariant method - Skandinavisk Aktuarietidskrift 17. 1934 - 1-6.

K. Pearson - Philosophical Transactions - vol 185 - part I - p. 13 - Risser et Traynard - Les principes de la statistique mathématique - Gauthier Villars - 1933 - p. 96.

MÉTHODE BASSÉE SUR LES PROPRIÉTÉS DU GRAPHIQUE SEMI-LOGARITHMIQUE

En utilisant un graphique semi-logarithmique (ordonnées logarithmiques) on obtient le même tracé parabolique (1) pour toutes les courbes normales de même écart-type : c'est là une propriété bien connue de ces graphiques.

La méthode consiste à tracer successivement sur un calque, par tâtonnements, les deux courbes normales, en utilisant un graphique semi-logarithmique. L'exemple qui suit, figure 1, choisi particulièrement simple, permet de comprendre la façon d'opérer.

On utilise un graphique semi-logarithmique (figure 3) sur lequel sont tracées des courbes normales théoriques correspondant à diverses valeurs de σ (ces valeurs sont indiquées sur les courbes).

1) TRACÉ DE LA PREMIÈRE COURBE.

On place un calque sur une feuille semi-logarithmique, de même module que la figure 3, (en pratique, on utilise la figure 3). On y trace (figure 4) en utilisant les ordonnées logarithmiques, la courbe correspondant à la distribution observée de la figure 1 ; les valeurs des ordonnées à reporter sont indiquées à la première ligne de la figure 2 "courbe observée". On a soin, (figure 4) de tracer les lignes verticales des ordonnées et de repérer par des horizontales les valeurs l (du module) de façon à pouvoir toujours replacer correctement le calque sur le graphique.

On porte le calque (figure 4) sur le graphique (figure 3) et l'on cherche (figure 5) à ajuster le mieux possible, la première partie de la courbe observée avec l'une des courbes théoriques du graphique 3 ; quand on a trouvé la courbe qui convient le mieux on la trace entièrement en pointillé sur le calque. On y marque les points A, B, C sur les lignes verticales correspondant respectivement aux points D, E, F de la courbe observée (figure 5). Ainsi, on vient de tracer, à l'essai, une première courbe normale. Mais on ne connaît pas encore le nombre d'observations correspondant aux points A, B, C.

2) NOMBRE D'OBSERVATIONS A AFFECTER A LA PREMIÈRE COURBE

On porte le calque (figure 5) sur le graphique semi-logarithmique en l'ajustant correctement sur les graduations l (du module). On se trouve ainsi placé comme on l'était lorsqu'on a tracé la courbe observée (figure 4). On peut donc lire sur le graphique les valeurs correspondant aux points A, B, C, soit dans cet exemple : 7,2 - 2,9 - 0,75. On inscrit ces valeurs en les arrondissant à 7,3 et 1 à la deuxième ligne de la figure 2, de façon à obtenir toutes les ordonnées de la première courbe. On connaît alors donc complètement la première courbe normale. Il faut savoir maintenant si en retranchant cette courbe de la courbe observée, il reste encore une seconde courbe normale satisfaisante.

3) NOMBRE D'OBSERVATIONS A AFFECTER A LA DEUXIÈME COURBE

En soustrayant dans le tableau de la figure 2, les ordonnées de la 1ère courbe de celles de la courbe observée, on obtient les ordonnées de la 2ème courbe. On complète cette deuxième courbe sur le calque en y portant les points G, H, J (figure 6).

4) TRACÉ DE LA DEUXIÈME COURBE

Il reste alors, en opérant comme pour la première courbe, à chercher en tâtonnant, à ajuster cette deuxième courbe avec l'une des courbes théoriques de

(1) De la relation :

$$y_x = K e^{-\frac{(x-m)^2}{2\sigma^2}}$$

on déduit en effet :

$$\text{Log. } y_x = K' - \frac{(x - m)^2}{2\sigma^2}$$

la figure 3. On voit sur la figure 6 que, pour éviter de tronquer la 2ème courbe, à gauche, on aurait pu retrancher une observation de l'ordonnée L de la courbe observée. On aurait obtenu le point K pour la deuxième courbe, qui aurait donné un bon ajustement. Si cette deuxième courbe est jugée satisfaisante le travail est terminé. Il reste à relever les résultats.

5) PARAMÈTRES DES DEUX DISTRIBUTIONS.

Quand on a tracé (figure 5) la première courbe théorique on a obtenu immédiatement la valeur m de la moyenne. On peut de même obtenir graphiquement la valeur $m + \sigma$ ou $m + 2\sigma$, ce qui permet d'avoir σ . Mais on peut aussi obtenir σ en relevant sur la figure 3 la valeur de l'écart-type correspondant à la courbe normale choisie (7,5 dans cet exemple) et en multipliant cette valeur par l'intervalle de classe de la figure 5.

On peut procéder de la même façon pour la deuxième courbe. Dans cet exemple on trouve la même valeur de l'écart-type pour les deux courbes (courbe théorique 7,5) ce que, d'ailleurs, on attendait pour des raisons techniques.

En faisant coïncider le sommet de la deuxième courbe avec la graduation 100 de l'échelle logarithmique on lit (figure 6) la valeur 47 au sommet de la première courbe. On a ainsi l'importance relative des ordonnées maxima M. On en tire, si les écarts-types ne sont pas égaux, l'effectif N relatif de chaque courbe en utilisant la formule $N = 2,51 \sigma M$.

REMARQUES

a) Il peut arriver que l'on ne trouve, à la figure 3, aucune courbe théorique satisfaisante, la courbe observée se situant entre deux courbes théoriques. Il faut alors, ou bien avoir un jeu plus complet de graphiques de la figure 3, ou refaire le graphique observé avec une échelle légèrement agrandie.

b) Il peut arriver, notamment près des extrémités, qu'une fréquence de la courbe observée soit zéro (point rejeté à l'infini sur l'échelle logarithmique). On doit alors, soit sauter ce point dans le tracé en joignant directement au point suivant, soit ajouter arbitrairement une ordonnée très petite.

c) Bien entendu, on peut commencer à ajuster la courbe de droite avant celle de gauche si cela se présente mieux. On peut séparer deux courbes beaucoup plus confondues que celles de la figure 1 qui sont nettement distinctes. On peut même parfois séparer trois courbes normales confondues en extrayant d'abord celle de droite puis celle de gauche.

d) La même méthode peut être utilisée pour la loi de Poisson (fig. 7). Elle est notamment utile lors des réglages, assez longs, de machines (casses sur métiers de filature). On peut alors séparer deux distributions de Poisson ou une distribution normale, et une distribution de Poisson. La figure 7 permet aussi de vérifier très rapidement, sans utiliser le test χ^2 , si une distribution suit une loi de Poisson.

Si cette méthode a l'avantage d'être rapide, elle présente de sérieux inconvénients qu'il ne faut pas ignorer. Il n'est pas besoin de souligner ce qu'il peut y avoir d'arbitraire à chercher à séparer deux populations quand on n'a pas à cela de sérieuses justifications techniques. La méthode donne un résultat très imprécis, notamment en ce qui concerne l'écart-type et il est bien difficile de savoir à quel point l'ajustement est satisfaisant. Il faut remarquer notamment que, dans la soustraction opérée figure 2, la branche droite de la première courbe (figure 5) ne prend aucune variation aléatoire à son compte. Le fait que les courbes s'ajustent bien n'implique pas nécessairement qu'elles existent en réalité. On peut appliquer le test classique χ^2 en prenant pour valeurs théoriques la somme des ordonnées des deux courbes pointillées, mais l'application de ce test n'est que très grossièrement approximative, les conditions théoriques d'emploi n'étant pas satisfaites.

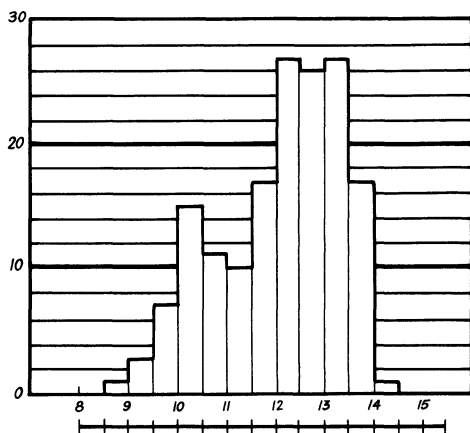


Fig. 1

	8	9	10	11	12	13	14	15					
COURBE OBSERVÉE	1	3	7	15	11	10	17	27	26	27	17	1	= 162
1 ^{ère} COURBE	1	3	7	15	11	7	3	1					= 48
						A	B	C					
2 ^{ème} COURBE					3	14	26	26	27	17	1		= 114
					G	H	J						

Fig. 2

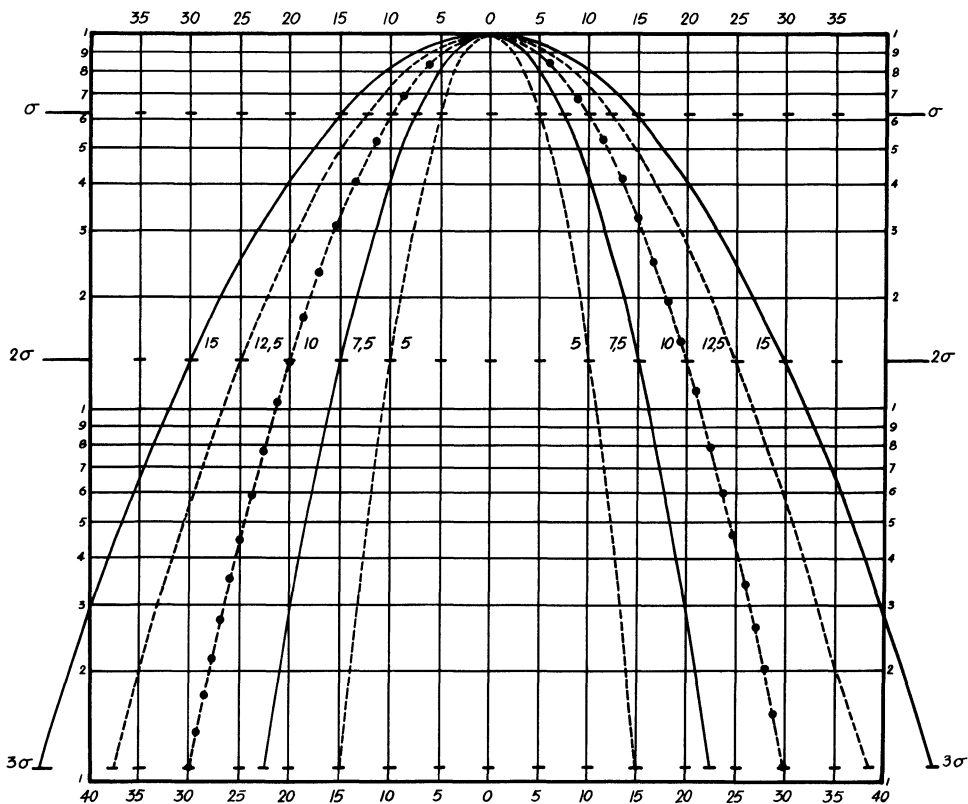


Fig. 3

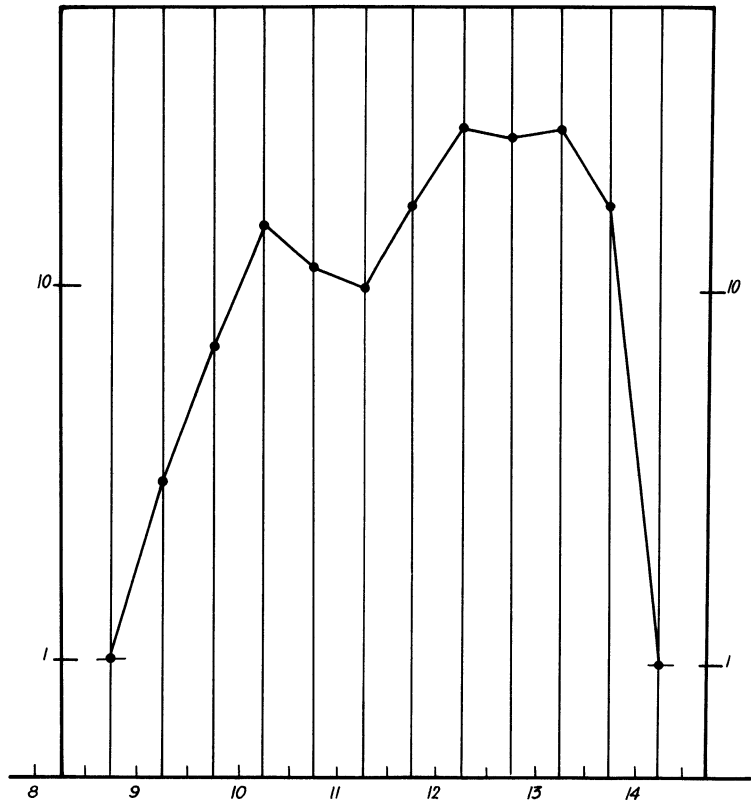


Fig. 4

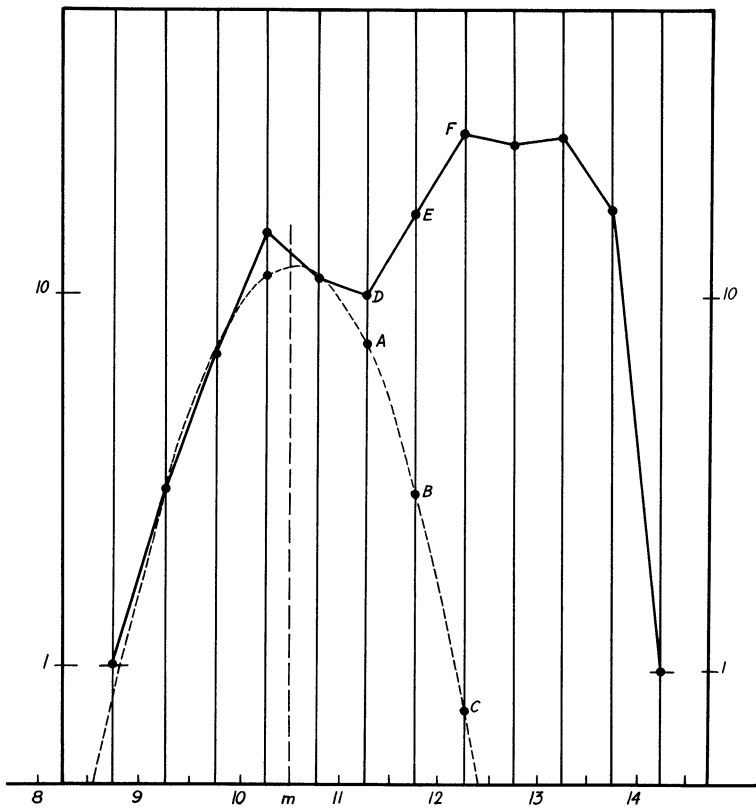


Fig. 5

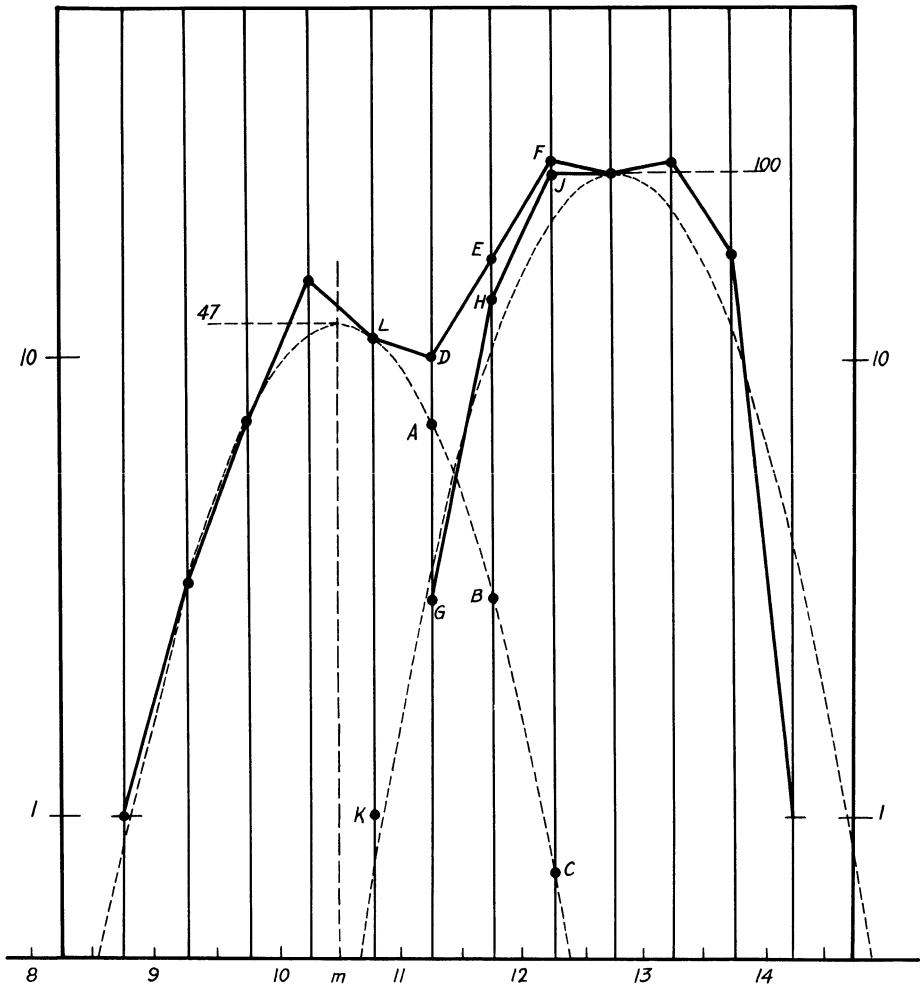


Fig. 6

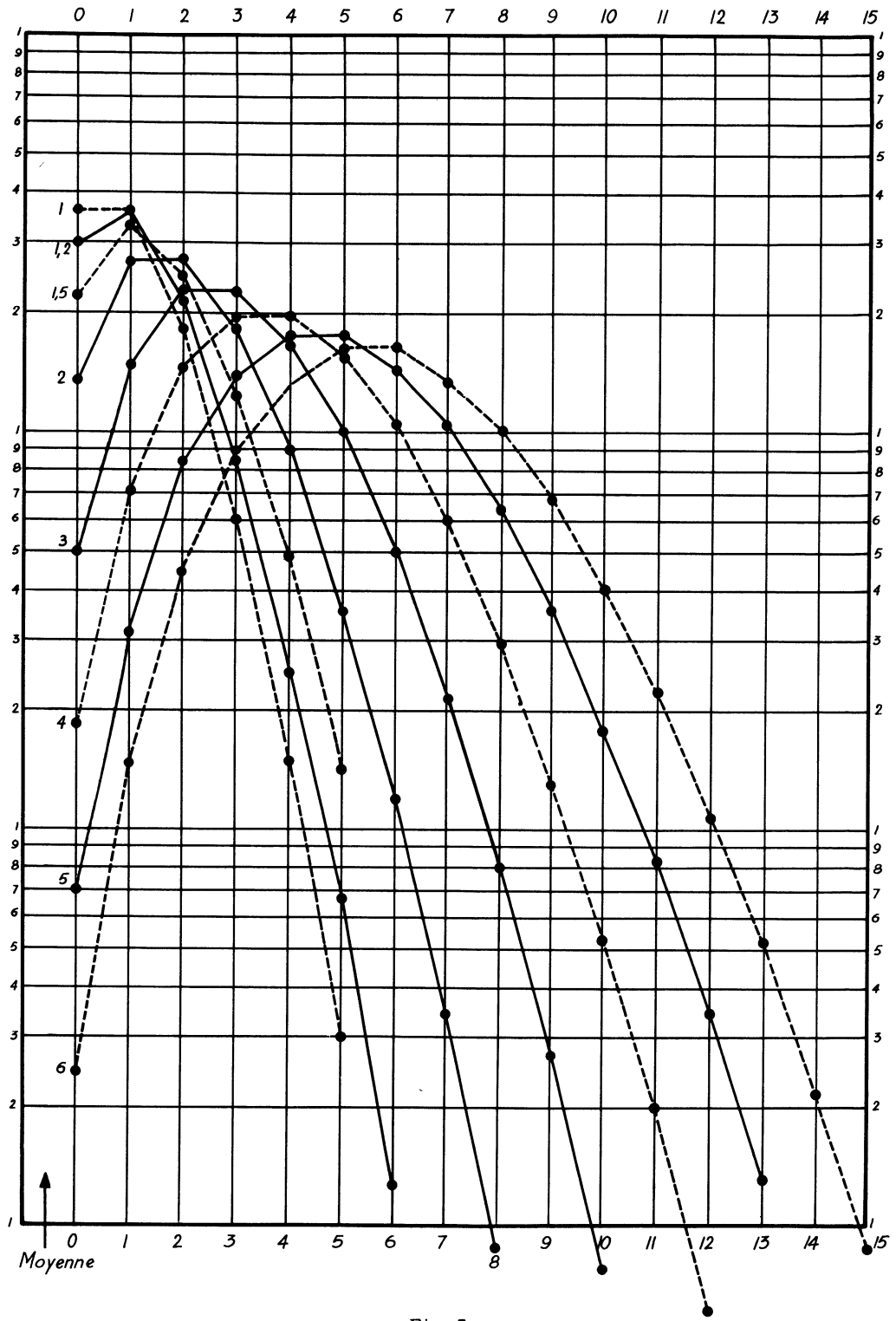


Fig. 7