AJUSTEMENT D'UNE COURBE DE DISTRIBUTION DU VOCABULAIRE

Si, dans une oeuvre ou un fragment d'oeuvre, on classe les différents lemmes par ordre croissant du nombre d'occurrences, on constate toujours que la classe la plus nombreuse est celle des lemmes employés une seule fois. Au fur et à mesure que croît le nombre d'occurrences par lemme, le nombre de ceux-ci diminue très rapidement, d'abord de façon régulière, puis avec des alternances d'augmentation et de diminution, pour finalement atteindre la valeur 1, tandis que certains nombres d'occurrences ne sont pas représentés.

On peut figurer semblable distribution par un graphique "en bâtonnets" ou encore, en reliant les sommets de ces "bâtonnets" par des segments de droite, par une ligne brisée qui se rapproche très fort de la verticale au début, puis s'infléchit pour devenir une horizontale interrompue aux valeurs élevées de x (si l'on admet que x représente le nombre d'occurrences, et y le nombre de lemmes).

On s'accorde généralement à reconnaître que les lemmes employés très peu souvent sont chargés de signification, tandis que les plus fréquents ne sont que des mots-outils dont l'usage est imposé à l'auteur par les exigences de la syntaxe. De plus, certains mots "signifiants" acquièrent sans doute une valeur particulière du fait même de leur répétition; on ne peut donc identifier dans l'absolu la catégorie des "signifiants" et celle des *semel dicta.*

Pour établir cette distinction sur une base statistique, il faudrait disposer d'une courbe aussi proche que possible sans doute de la distribution réelle, mais qui représenterait une distribution aléatoire et servirait de point de comparaison, en mettant en évidence au moyen d'un test de χ^2, les irrégularités de la courbe.

Or, la forme générale de la distribution des lemmes par nombre d'occurrences fait pènser à la courbe mathématique de l'hyperbole équilatère de formule $x \cdot y = a$ ou, sous une autre forme, $y = \frac{a}{x}$.

Pour déterminer aussi exactement que possible l'expression algébrique qui donnerait les résultats les plus proches des valeurs observées, j'ai considéré d'abord que chaque valeur de y représentait une fraction de l'effectif total diminué du nombre de lemmes déjà répartis, soit, en représentant par N le nombre total d'occurrences, pour chaque valeur de x :

$N - O$	pour	$x = 1$
$N - y_1$		$x = 2$
$N - (y_1 + y_2)$		$x = 3$
$N - (y_1 + y_2 + y_3)$		$x = 4$
$N - (y_1 + y_2 + y_3 + y_4)$		$x = 5$ etc.

En raison du décroissement rapide des valeurs observées, le dénominateur de l'expression envisagée devrait comporter un facteur en x^2 et, afin de tenir compte également de l'allongement de la courbe vers la droite du graphique (c'est-à-dire vers les valeurs élevées de x), un second facteur constitué par le

nombre moyen d'occurrences par lemme, soit $\bar{x} = \frac{N}{n}$ où n représente le nombre total de lemmes différents.

Si l'on compare les résultats obtenus à partir de la formule ainsi définie, soit

$$y = \frac{N - \Sigma_0^{x-1} y_x}{x^2 \cdot \bar{x}} \quad (1)$$

on constate que les valeurs de y ainsi obtenues sont encore trop élevées; on complète donc le dénominateur de l'expression (1) par un troisième facteur k, qui paraît significatif de chaque distribution observée. On arrive ainsi à la fonction suivante :

$$y' = \frac{N - \Sigma_0^{x-1} y_x}{x^2 \cdot \bar{x} \cdot k} \quad (2)$$

Il reste à déterminer k pour chaque distribution considérée. De (2) on extrait cette valeur en fonction de N, de $\bar{x}$ et de y_x, soit :

$$k_x = \frac{N - \Sigma_0^{x-1} y_x}{x^2 \cdot \bar{x} \cdot y_x} \quad (3)$$

Comme la distribution réelle fait apparaître une régularité évidente pour les valeurs faibles de x, on peut imaginer d'ajuster la courbe théorique sur cette partie de la courbe réelle, et d'observer les différences entre les valeurs de y et de y'.

Il paraît toutefois prudent de ne pas considérer la valeur de y correspondant à x = 1, car le nombre de lemmes employés une seule fois est limité par l'étendue du vocabulaire de l'auteur, alors que la courbe de l'hyperbole équilatère a pour limite l'infini lorsque x tend vers 0.

On calcule donc k à partir de l'expression (3) pour les valeurs successives de x depuis 2 jusqu'à la valeur qui précède la première irrégularité de la ligne des y. Cette valeur, qui correspond soit à un accroissement, soit à une égalité, soit encore à une diminution trop marquée de y, est déterminée par les relations :

$$y_x \geqslant y_{x-1} \qquad ou \qquad y_{x-1} - y_x > y_{x-2} - y_{x-1} .$$

La valeur générale de k est alors obtenue en faisant la moyenne des valeurs partielles ainsi obtenues. On peut alors calculer y' à partir de l'expression (2).

Comme la distribution réelle ne comporte nécessairement que les valeurs entières de x et de y, on estime qu'il est judicieux de calculer les valeurs de y' à l'unité près, en forçant ou en négligeant la partie décimale selon qu'elle atteint ou non 0,5.

En reportant les valeurs de y' ainsi obtenues sur le graphique de la distribution réelle, on observe que la position relative des deux lignes y et y' détermine trois zones : la première, correspondant aux faibles valeurs de x, où y' est continuellement inférieur à y, sauf pour x = 1, éventuellement x = 2;

et cela s'explique, comme je l'ai dit, par la limitation de richesse du vocabulaire; une seconde partie où les lignes y et y' se recoupent plusieurs fois; et une troisième partie, correspondant aux valeurs élevées de x, où y' se confond avec y pour lui redevenir continuellement inférieur. Cette troisième partie est en général la même que celle où $y' \leqslant 2$.

Il me semble qu'à partir de là, on peut répartir l'ensemble du vocabulaire en lemmes "signifiants" et "non-signifiants", en définissant l'ensemble des "signifiants" par S et l'ensemble des "non-signifiants" par $\overline{S}$ dans les expressions suivantes :

$$S = \{ x \xi y' < y \} \cup \{ x \xi x < 3 \}$$
$$\overline{S} = \{ x \xi y' > y \quad \text{ssi} \quad x > 3 \} \cup \{ x \xi y' \leqslant 2 \}.$$

Dans ces conditions, les lemmes employés une ou deux fois sur l'ensemble de l'oeuvre ou du fragment d'oeuvre sont nécessairement considérés comme "signifiants"; il convient en outre d'examiner particulièrement les lemmes dont la fréquence d'emploi est égale aux valeurs de x pour lesquelles y' est inférieur à y; on devrait découvrir, parmi ces lemmes, un certain nombre de "signifiants" dont la valeur est renforcée par leur répétition.

Si l'on calcule la valeur de χ^2, on peut, d'une part, déterminer l'ampleur de la divergence entre les deux courbes et, d'autre part, relever les valeurs de x où se manifeste un excès important du nombre de lemmes. Ces divergences peuvent résulter soit d'une intention, soit d'une obsession de l'auteur, qui

l'entraîne à l'emploi particulièrement fréquent de tel ou tel mot.

Afin de déterminer dans quelle mesure la courbe théorique calculée de cette manière pouvait être considérée comme une distribution aléatoire, j'ai appliqué le test suivant : à partir des valeurs théoriques obtenues par l'application de l'expression (2) à une distribution de vocabulaire, j'ai établi une nouvelle courbe théorique, en considérant les premières valeurs obtenues comme celles d'une distribution réelle. Le calcul de χ^2, dans cette seconde opération d'ajustement, donne une valeur de 2,6351 ce qui, pour $\nu = 21$, correspond à une probabilité supérieure à 0,99999. On peut donc considérer cette méthode d'ajustement comme suffisamment précise, puisque les valeurs obtenues à partir d'une distribution théorique ne présentent que des différences tout à fait négligeables.

J'ai appliqué cette méthode d'investigation au vocabulaire de l'*Electre* d'EURIPIDE, en considérant tout d'abord l'ensemble de l'oeuvre, puis en examinant le vocabulaire des Episodes, celui des Choeurs, des Mélodrames, et enfin celui du rôle d'Electre, d'Oreste et de Clytemnestre dans les Episodes. On peut relever les points suivants :

a) la distribution de l'ensemble de l'oeuvre donne une valeur de $\chi^2 = 18,25$ pour $\nu = 22$ soit une probabilité de 0,61;

b) la distribution du rôle d'Electre (dans les Episodes) donne une valeur de $\chi^2 = 22,71$ pour $\nu = 11$, soit une probabilité très faible : 0,01768 résultant évidemment des valeurs de y_2, y_3 et y_{15}, y_{16} (ces valeurs étant regroupées pour le calcul de χ^2);

c) les autres distributions correspondent à des probabilités comprises entre 0,52 et 0,26.

On peut en déduire que la distribution du vocabulaire sur l'ensemble de l'oeuvre n'est pas significative, contrairement à celle du rôle d'Electre dans les Episodes, où il paraît y avoir une recherche particulière de la variété, qui se marque également dans les mots employés deux ou trois fois (peut-être dans une intention de répétition significative). Quant aux autres distributions, leur répartition mérite sans doute un examen plus approfondi portant, dans les Episodes, sur les valeurs de y_5 et y_7; dans les Choeurs, sur la valeur de y_2; dans les Mélodrames, sur la valeur de y_4; pour le rôle d'Oreste, sur les valeurs de y_2 et de y_4, et pour le rôle de Clytemnestre, sur les valeurs de y_4 et de y_5.

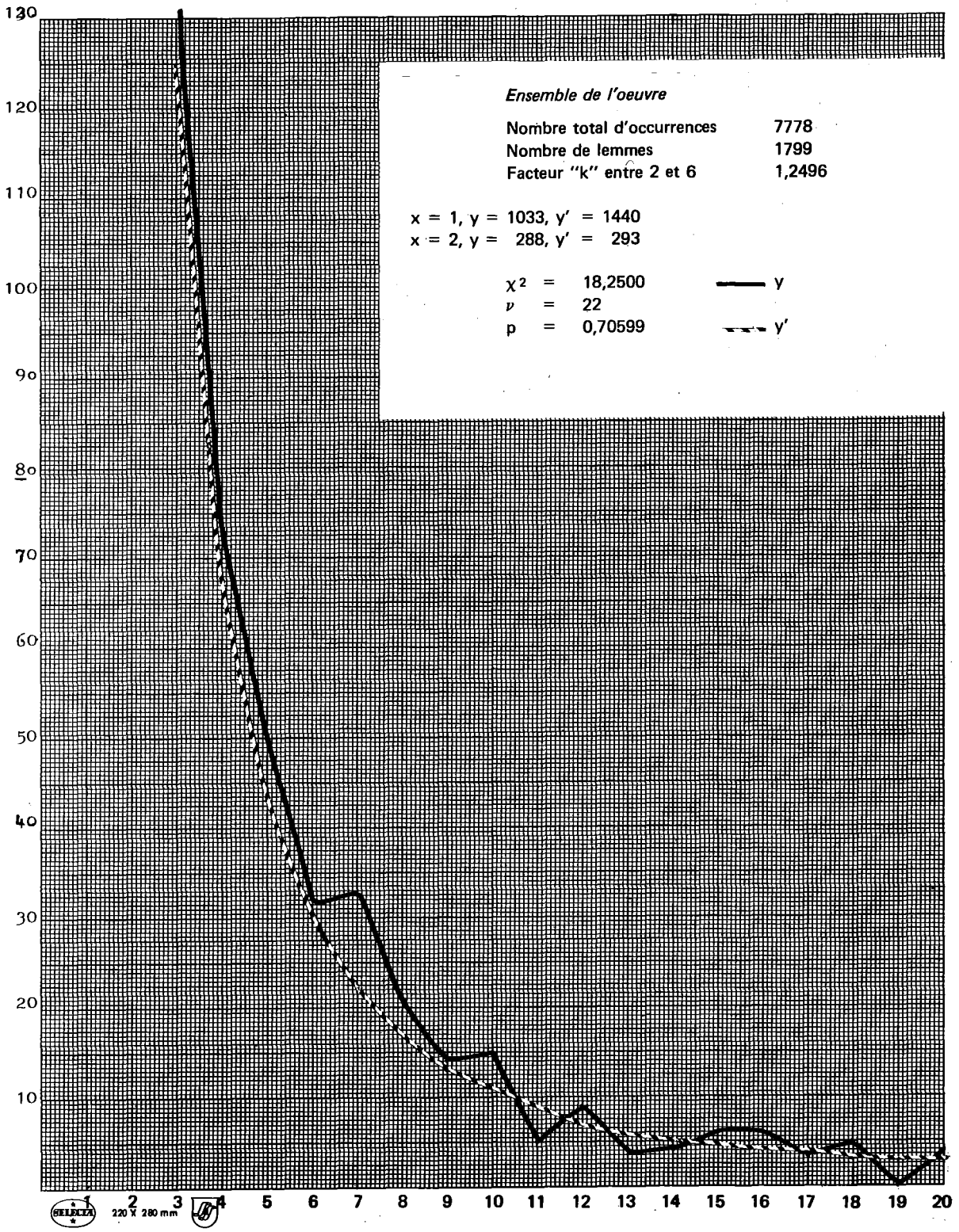
Cet examen devrait concerner spécifiquement la valeur sémantique des lemmes apparaissant avec ces fréquences-là dans les parties considérées. Mais cette étude doit reposer sur un procédé statistique qui reste à déterminer pour l'ensemble de l'oeuvre. Aussi je me propose de revenir sur cette question en abordant notamment le problème des regroupements de synonymes par référence à un système de catégories sémantiques.

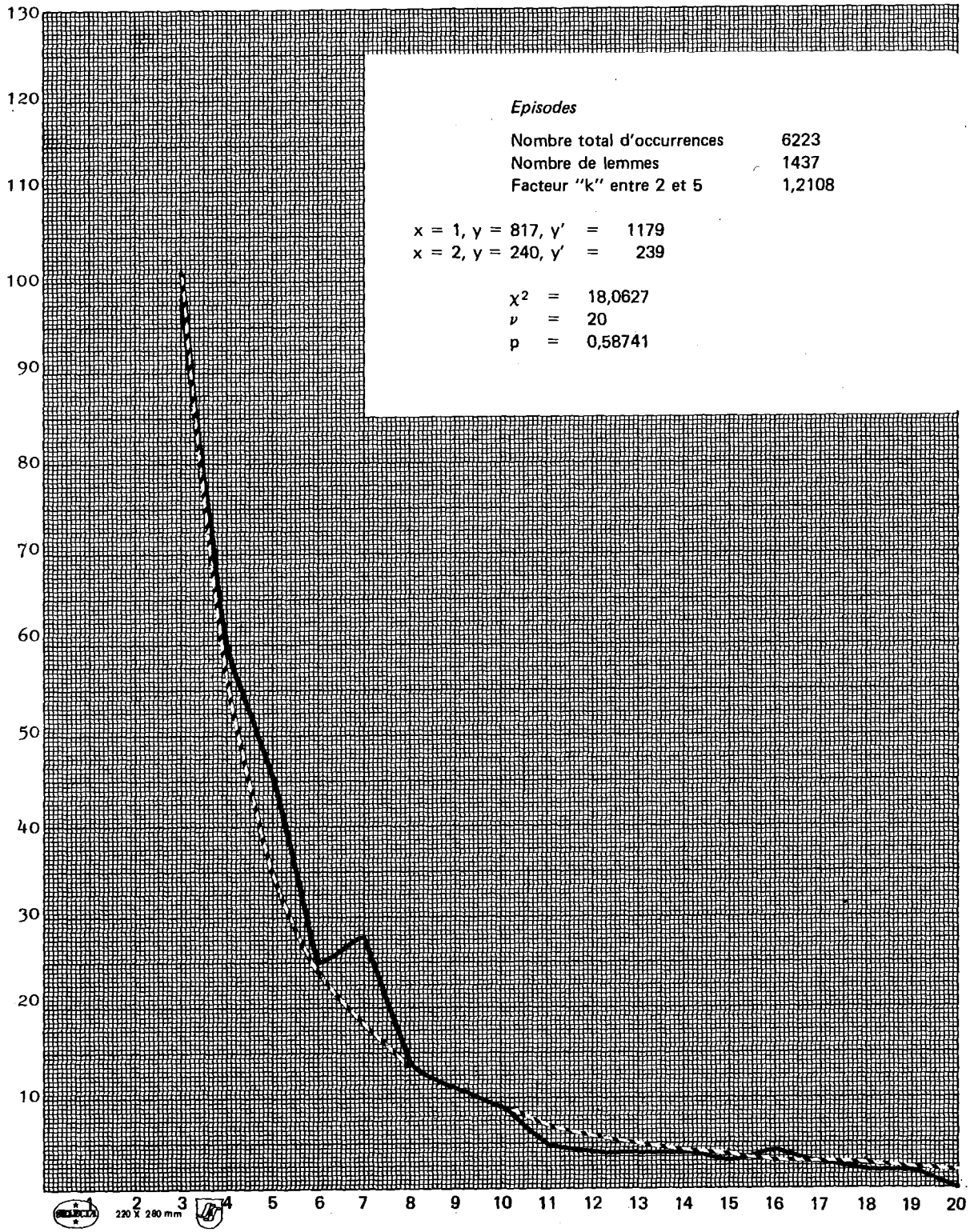
C'est pourquoi le texte que l'on vient de lire ne se proposait comme but que de présenter un moyen d'apprécier la régularité d'une courbe de distribution du vocabulaire, afin de mettre en relief les éléments significatifs du lexique d'un auteur ou d'une oeuvre déterminée.

Université de Liège G. RIGO.

Nombre d'occurrences	ENSEMBLE DE L'OEUVRE			EPISODES			CHOEURS			MELODRAMES			ELECTRE (Epis.)			ORESTE (Epis.)			CLYTEMNESTRE		
	Réel	Théor.	χ^2	Réel	Théor.	χ^2	Réel	Théor.	χ^2	Réel	Théor.	χ^2	Réel	Théor.	χ^2	Réel	Théor.	χ^2	Réel	Théor.	χ^2
1	1033	1440	-	817	1179	-	443	534	-	150	124	-	456	467	-	283	376	-	170	242	-
2	288	293	0,0853	240	239	0,0041	96	76	5,2631	22	19	0,4736	127	93	12,4301	83	64	5,6406	34	31	0,2903
3	132	124	0,5161	98	101	0,0891	35	30	0,8333	5	8	1,1250	50	39	3,1025	31	26	0,9615	12	12	-
4	74	68	0,5294	60	56	0,2857	14	16	0,2500	8	4	3,5714	24	22	0,1818	18	14	1,1428	9	6	1,5000
5	50	43	1,1395	46	35	3,4571	10	10	-	4	3		12	14	0,2857	10	9	0,1111	6	4	1,2857
6	32	30	0,1333	25	24	0,0416	8	7	0,1428	0	2		8	9	0,1111	5	6	0,1666	4	3	
7	33	22	5,5000	28	18	5,5555	7	5	0,8000	1	1		5	7	0,5714	6	4	0,5714	3	2	
8	21	17	0,9411	14	14	-	1	4	1,2857	1	1	-	6	5	0,2000	3	3		2	1	-
9	14	13	0,0769	11	11	-	3	3		1	1		5	4		1	3	0,2000	0	1	
10	15	11	1,4545	9	9	-	1	2		2	1		4	3	0,1666	5	2		0	1	
11	5	9	1,7777	5	7	0,5714	3	2	0,1666	0	0		3	3		1	2		0	1	
12	9	7	0,5714	4	6	0,6666	3	2		1	0		1	2		2	1		0	1	
13	4	6	0,6666	4	5	0,2000	0	1					2	2	0,6666	3	1	0,8000	2	1	-
14	4	5	0,2000	4	4	0,1250	2	1					1	2		1	1		1	1	
15	6	5	0,2000	3	4		0	1	0,2000				4	1		0	1		0	0	
16	6	4	0,1250	4	3	0,1666	1	1					4	1		2	1		1	0	
17	3	4		3	3		1	1					0	1	3,2000	0	1	-			
18	5	3	0,1666	2	3	0,2000	1	1					1	1		1	1				
19	0	3		2	2		0	1					0	1		2	1				
20	4	3	0,2000	0	2	1,5000	0	1					5	1		0	1				
21	2	2		6	2		0	1					2	1		1	0				
22	2	2		3	2		6	0					1	1	1,8000	0	0				
23	1	2	0,1666	2	2								0	1							
24	4	2		2	1	0,2000							0	1							
25	2	2	-	1	1								1	1							
26	1	2		1	1								2	0							
27	4	1		2	1								0	0							
28	2	1		3	1																
29	1	1	1,8000	1	1	3,2000															
30	3	1		2	1																
31	1	1		1	1																
32	1	1		1	1	-															
33	2	1		0	1																
34	1	1		1	1																
35	0	1		0	1	1,8000															
36	2	1	0,2000	0	1																
37	0	1		0	1																
38	1	1		2	1																
39	0	1		1	1																
40	0	1		2	1																
41	1	1		0	1																
42	0	1	1,8000	0	0																
43	1	1		1	0																
44	1	1		1	0																
45	0	1		0	0																
Total *	(1799)		18,2500	(1437)		18,0627	(632)		8,9415	(195)		5,1700	(736)		22,7158	(464)		9,5940	(245)		3,0760

* Le calcul de l'effectif théorique n'a pas été poursuivi au-delà de la valeur de x où le résultat est inférieur à 0,5. On constatera, dans la colonne "effectif réel", que certaines valeurs inférieures de x ne sont déjà plus représentées. Il ne convenait pas, semble-t-il, de faire porter les observations sur les valeurs de x supérieures à la limite ainsi définie.



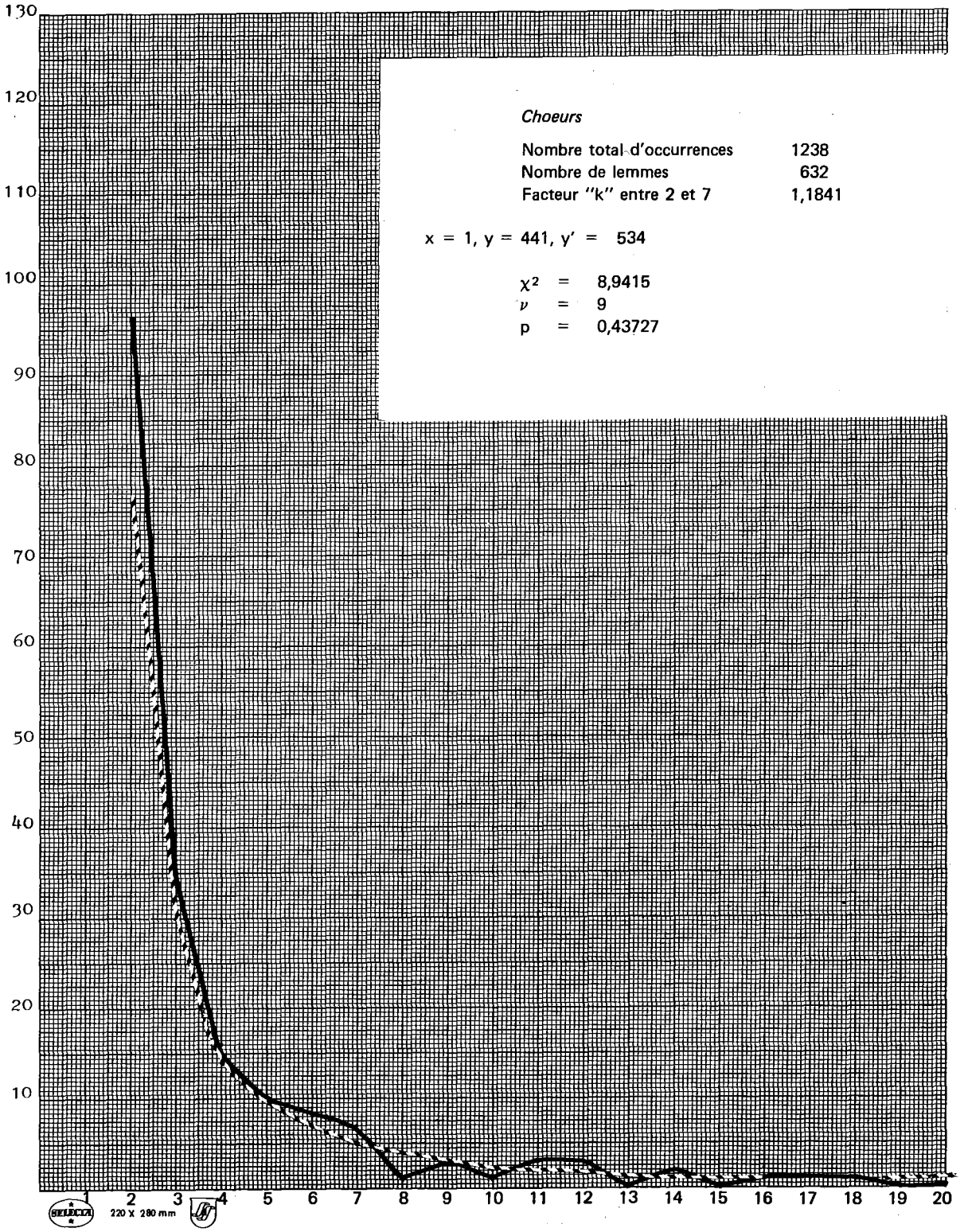


Episodes

Nombre total d'occurrences	6223
Nombre de lemmes	1437
Facteur "k" entre 2 et 5	1,2108

$x = 1,\ y = 817,\ y' = 1179$
$x = 2,\ y = 240,\ y' = 239$

$\chi^2 = 18{,}0627$
$\nu = 20$
$p = 0{,}58741$

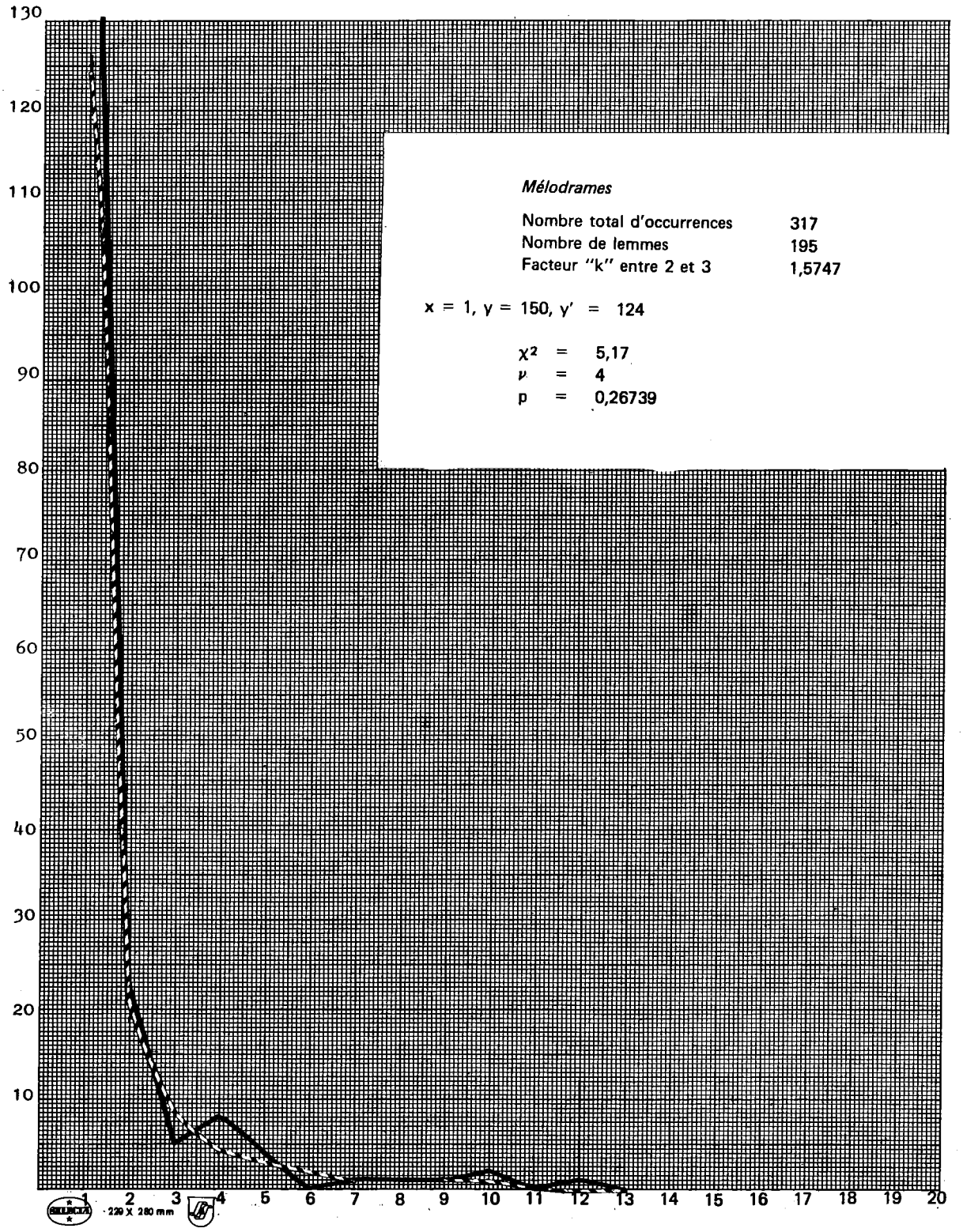


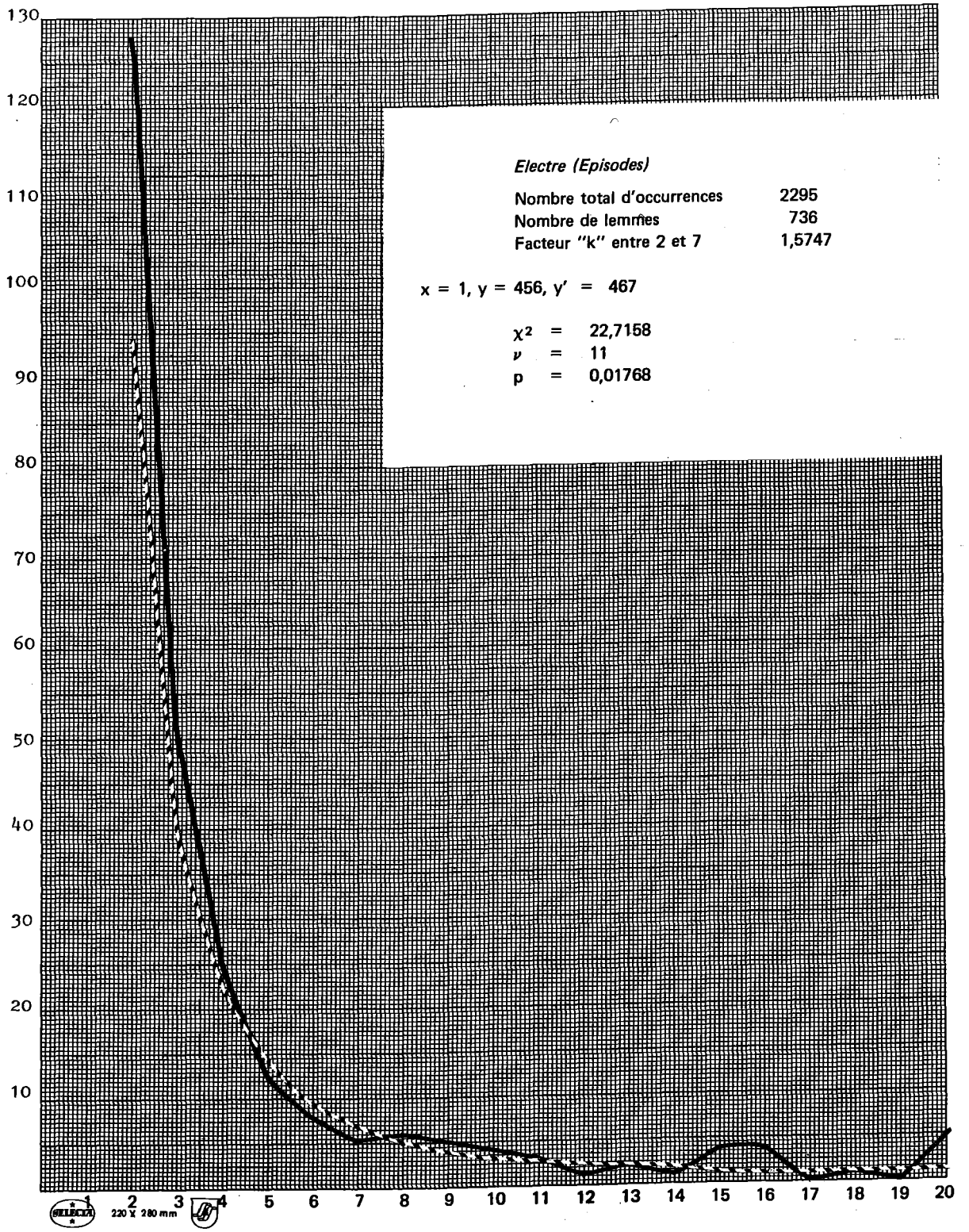
Choeurs

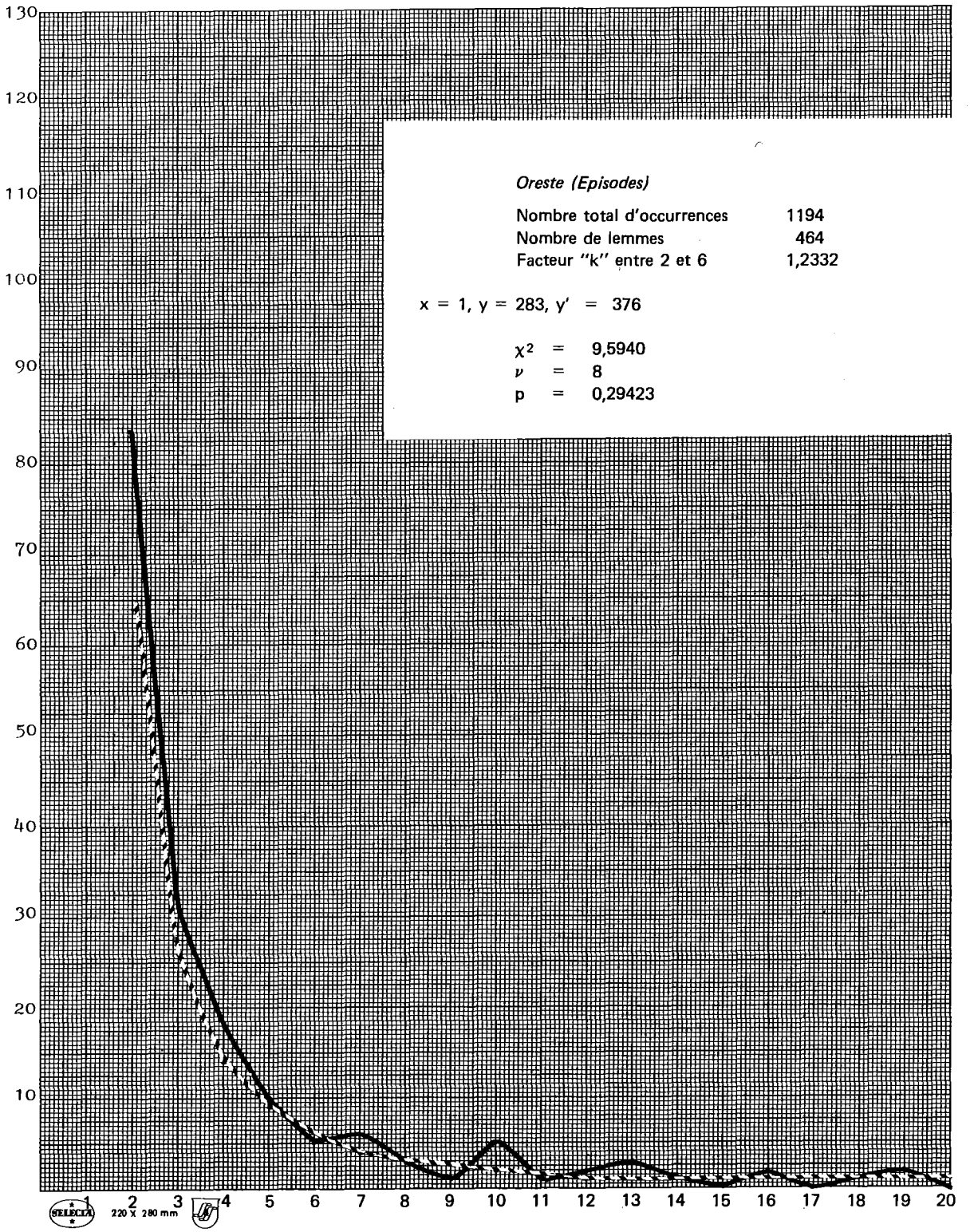
Nombre total d'occurrences	1238
Nombre de lemmes	632
Facteur "k" entre 2 et 7	1,1841

x = 1, y = 441, y' = 534

χ^2 = 8,9415
ν = 9
p = 0,43727







Oreste (Episodes)

Nombre total d'occurrences	1194
Nombre de lemmes	464
Facteur "k" entre 2 et 6	1,2332

$x = 1$, $y = 283$, $y' = 376$

$\chi^2 = 9{,}5940$
$\nu = 8$
$p = 0{,}29423$