

REVUE DE STATISTIQUE APPLIQUÉE

J. P. GIVRY

Estimation de la loi de probabilité de durée

Revue de statistique appliquée, tome 13, n° 2 (1965), p. 5-28

http://www.numdam.org/item?id=RSA_1965__13_2_5_0

© Société française de statistique, 1965, 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/>

ESTIMATION DE LA LOI DE PROBABILITÉ DE DURÉE

J. P. GIVRY

Ingénieur Civil des Mines
Docteur-ingénieur
Chef du Laboratoire de Recherches
des Fabrications de la Compagnie Pechiney

TABLE DES MATIERES

	Pages
1 - Introduction.....	6
2 - Estimation de la probabilité de durée.....	6
3 - Utilisation du graphique de Henry.....	9
4 - Délimitation du domaine de confiance de la moyenne et de l'écart-type.....	12
5 - Utilisation d'une information a priori sur l'écart-type.....	18
6 - Conclusion.....	28

1 - INTRODUCTION

Dans un article précédent(*), en étudiant la gestion d'un parc à renouvellement aléatoire, nous avons mis en évidence l'intérêt qu'il y avait à estimer la durée d'un type nouveau d'équipements le plus tôt possible après sa mise en service. Cet intérêt était surtout économique, car la connaissance de la loi de probabilité des durées est indispensable pour asseoir les prévisions de renouvellement et la politique d'approvisionnement.

D'un point de vue plus technique, cette estimation est également importante, car on pratique rarement pendant des dizaines d'années un renouvellement à l'identique, soit que les idées évoluant on apporte des modifications volontaires à la conception des équipements, soit que les fournisseurs n'assurant plus la fabrication de telle nuance de matériau les équipements renouvelés se trouvent modifiés, involontairement cette fois et même quelquefois à l'insu de l'utilisateur. Dans tous les cas il est bon de connaître l'effet de ces modifications sur la durée. Sur une longue période un parc d'équipements à renouvellement aléatoire se présente donc comme une succession de lots, tous les équipements mis en place entre deux dates appartenant à un lot déterminé, mais les dates de transition n'étant pas forcément connues.

Statistiquement le problème est d'estimer la loi de probabilité à partir d'un échantillon tronqué. GUPTA en a présenté une solution purement numérique dans le cas où tous les équipements ont la même ancienneté et où les durées sont gaussiennes. Notre propos est de trouver une solution au cas général où les équipements n'ont pas forcément tous démarré au même moment. Comme on le verra, la méthode proposée est surtout graphique et à l'avantage de se prêter à une remise à jour facile, au fur et à mesure que le temps apporte de nouvelles informations ; la méthode permet aussi de vérifier la normalité de la distribution et de déceler les hétérogénéités.

Comme dans l'article précédent, nous ne traiterons que des lots gaussiens ; pour le statisticien ils ont évidemment l'avantage de limiter à deux les paramètres caractéristiques de la distribution, mais ce n'est pas la seule raison qui nous a fait nous y intéresser. Comme nous l'avons indiqué dans l'article cité, la loi de Gauss représente bien la distribution des durées dans de nombreuses applications.

2 - ESTIMATION DE LA PROBABILITE DE DUREE

Appelons $F(x)$ la probabilité d'atteindre la durée x .

Le problème général peut s'exprimer comme suit :

A l'instant t où l'on pratique l'estimation de la loi de probabilité des durées, on dispose des informations suivantes :

(*) Revue de statistique appliquée n° XIII-1

Les N équipements constituant le lot ont été mis en place à des instants différents t_i ; ils ont des anciennetés $a_i = t - t_i$; a_i est une fonction de t , $a_i(t)$. On distingue dans N deux classes, celle des équipements qui sont déjà morts, à des instants t'_i et pour lesquels on peut calculer une durée $d_i = t'_i - t_i$ et celle des équipements en service à l'instant t dont on sait seulement que la durée sera supérieure à l'ancienneté : $d_i \geq a_i(t)$.

Si on classe les N équipements en fonction non plus du temps absolu, t , comme sur la figure 1, mais en fonction de l'ancienneté $a_i(t)$ on obtient la figure 2, où l'on a indiqué par un segment blanc la période où l'équipement était en service et par un segment noir la période où il était mort. Les équipements déjà morts sont représentés par un segment blanc de longueur d_i , suivi d'un segment noir de longueur $a_i(t) - d_i$; les équipements en service sont représentés par un seul rectangle, blanc, de longueur $a_i(t)$.

En comptant les segments blancs et noirs en face de l'ancienneté a , on a une estimation de la loi de probabilité des durées $F(x)$ pour $x = a$. En effet les segments blancs représentent les équipements dont la durée a dépassé x et les segments noirs les équipements morts qui avaient une ancienneté suffisante pour atteindre éventuellement l'âge x . La somme des deux nombres

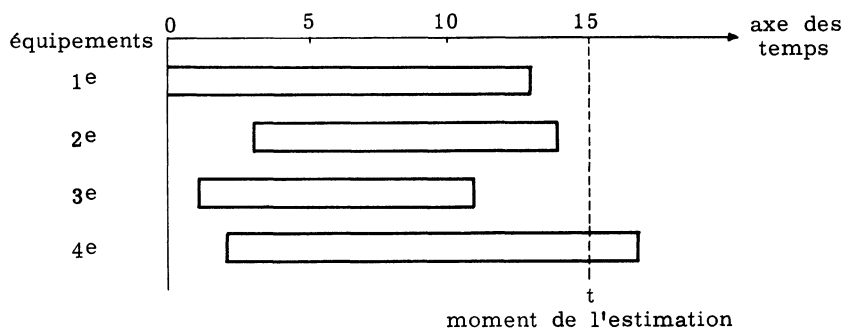


Fig. 1

de segments représente l'information totale disponible sur la durée x . A un moment, t , donné, cette information diminue le long de l'axe des x ; elle s'annule pour $x > t$; aucune estimation de $F(x)$ n'est possible pour $x > t$.

Le rapport blanc/blanc + noir est la fraction des équipements qui ont atteint la durée x parmi ceux qui ont une ancienneté au moins égale à x . Cette fraction tend en probabilité vers $F(x)$. En effet, en probabilité, chacun des segments individuels n'est pas partagé en deux tronçons, l'un blanc et l'autre noir, mais doit être considéré comme teinté de gris avec une densité égale à $F(x)$, donc variable de 1 = blanc, pour $x = 0$, à 0 = noir pour $x \infty$. La fraction blanc/blanc + noir n'est pas autre chose que la moyenne pondérée de n variables qui valent 1 tant que l'équipement est en vie et 0 quand il est mort. Chaque variable a une espérance mathématique égale à $F(x)$ et leur moyenne pondérée tend donc évidemment elle aussi vers $F(x)$.

La figure 3 montre comment, au fur et à mesure du temps, l'estimation de $F(x)$ s'améliore .

Aucune estimation de $F(x)$ n'est possible tant que $t \leq x$.

En résumé l'estimation de $F(x)$ est donnée, à l'instant t , par la fraction des équipements qui ont dépassé l'âge x parmi ceux qui ont une ancienneté supérieure ou égale à x .

Etat de l'information au temps $t = 15$

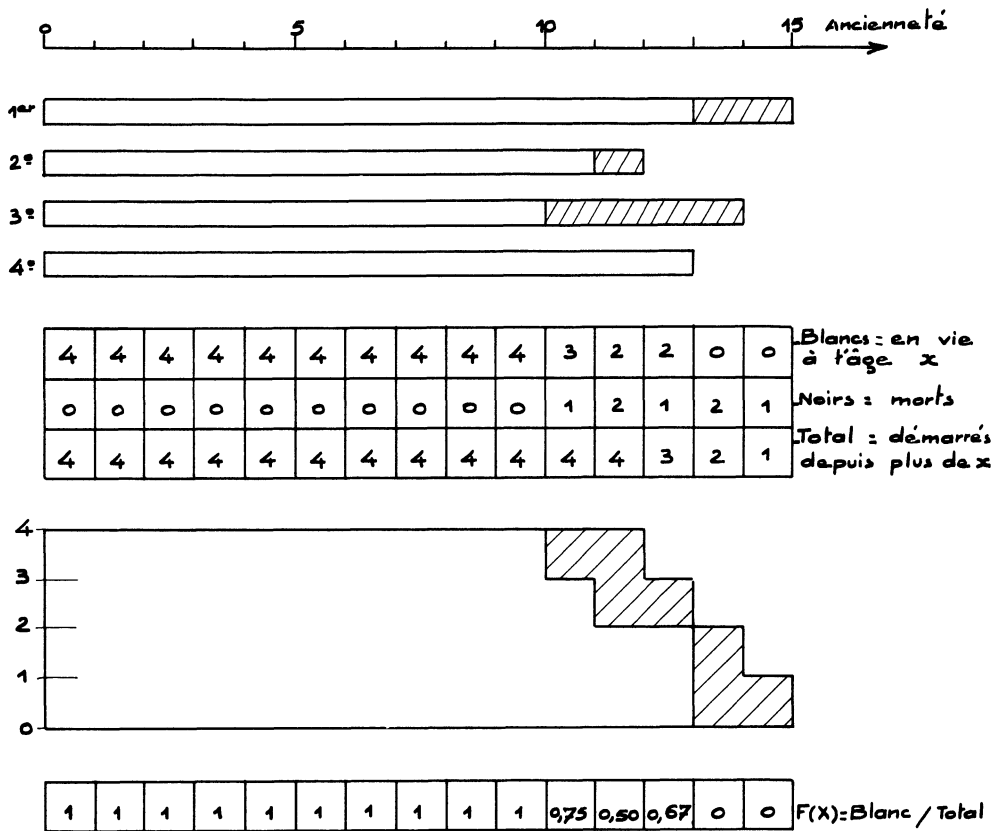


Fig. 2

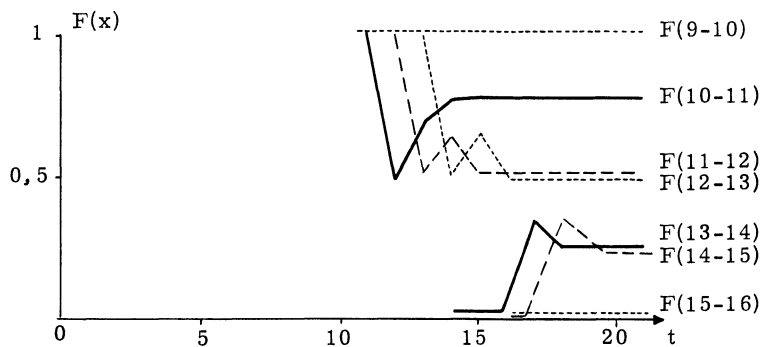


Fig. 3

3 - UTILISATION DES GRAPHIQUES DE HENRY

Si nous appliquons cette méthode d'estimation des probabilités $F(x)$ à des durées gaussiennes - il est tout naturel de pointer les points obtenus sur un graphique de Henry. Bien entendu le tracé sera incomplet tant que tous les équipements du lot ne seront pas morts et il n'y aura pas d'information pour tout x supérieur au temps t qui sépare le moment de l'estimation du démarrage du plus ancien des équipements étudiés. L'usage du graphique de Henry est cependant particulièrement précieux, D'une part les valeurs de $F(x)$, estimées à différents moments du temps, s'alignent en probabilité sur un segment de droite qui s'allonge progressivement sur la droite de moyenne m et d'écart-type s , rendant ainsi possible une estimation graphique de m et s , assez grossière certes, mais intéressante par sa simplicité, et d'ailleurs d'autant meilleure que le nombre d'équipements étudiés est plus grand et que le pourcentage des défaillances déjà enregistrées est plus élevé.

D'autre part le graphique de Henry permet de vérifier commodément la normalité des durées. Si celles-ci ne sont pas gaussiennes ou si elles résultent, comme cela arrive, du mélange de deux populations gaussiennes, les points $F(x)$ ne s'alignent plus sur une droite, mais dessinent une courbe dont l'aspect est caractéristique (figure 4).

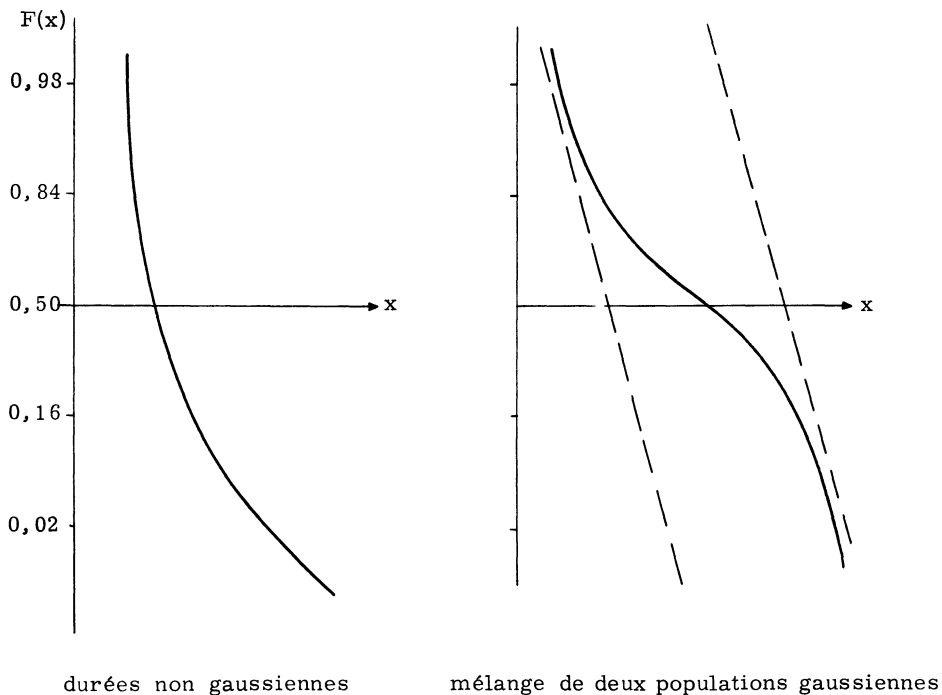


Fig. 4

Pour les durées sensiblement gaussiennes, il est tentant d'utiliser le graphique de Henry pour estimer graphiquement la moyenne et l'écart-type d'une population inconnue.

Considérons, à titre d'exemple, le cas d'un lot de 36 équipements, qui au moment de l'estimation est dans l'état suivant :

14 équipements démarrés il y a 21 mois :
 1 mort à 18 mois
 2 morts à 19 mois
 2 morts à 20 mois
 9 en vie

22 équipements démarrés il y a 18 mois :
 1 mort à 15 mois
 2 morts à 17 mois
 19 en vie

Le calcul des probabilités $F(x)$ s'effectue aisément, comme au §2 ci-dessus ; il permet de tracer la figure 5 qui conduit à attendre du lot une durée moyenne de 22 mois avec un écart-type de 3 mois.

Une telle conclusion serait illusoire, comme nous allons le montrer en testant une hypothèse sensiblement différente.

Testons l'hypothèse $m = 25, s = 6$ qui graphiquement s'écarte pas mal des valeurs empiriques de $F(x)$. Pour ce faire, il faut calculer l'état dans lequel se trouverait, en probabilité, le lot si l'hypothèse (m, s) était vraie ; puis il faut comparer les nombres d'équipements, que l'on rencontre dans les différentes classes, par un test de χ^2 .

Découpons donc le lot en 2 groupes, correspondant aux anciennetés 18 et 21 mois, puis chaque groupe en classes :

TABEAU 1

		Ancienneté 21 mois				Ancienneté 18 mois			
		Morts		en vie	Total	Morts		en vie	Total
		avant 18 mois	de 18 à 21 mois			avant 15 mois	de 15 à 18 mois		
Etat réel du lot		0	5	9	14	0	3	19	22
Etat théorique du lot dans l'hypothèse $m = 25, s = 6$	p_i	0,12	0,13	0,75	1	0,045	0,075	0,88	1
	np_i	1,68	1,82	10,50	14	0,99	1,65	19,36	22
n° de classe		1	2	3		4	5	6	

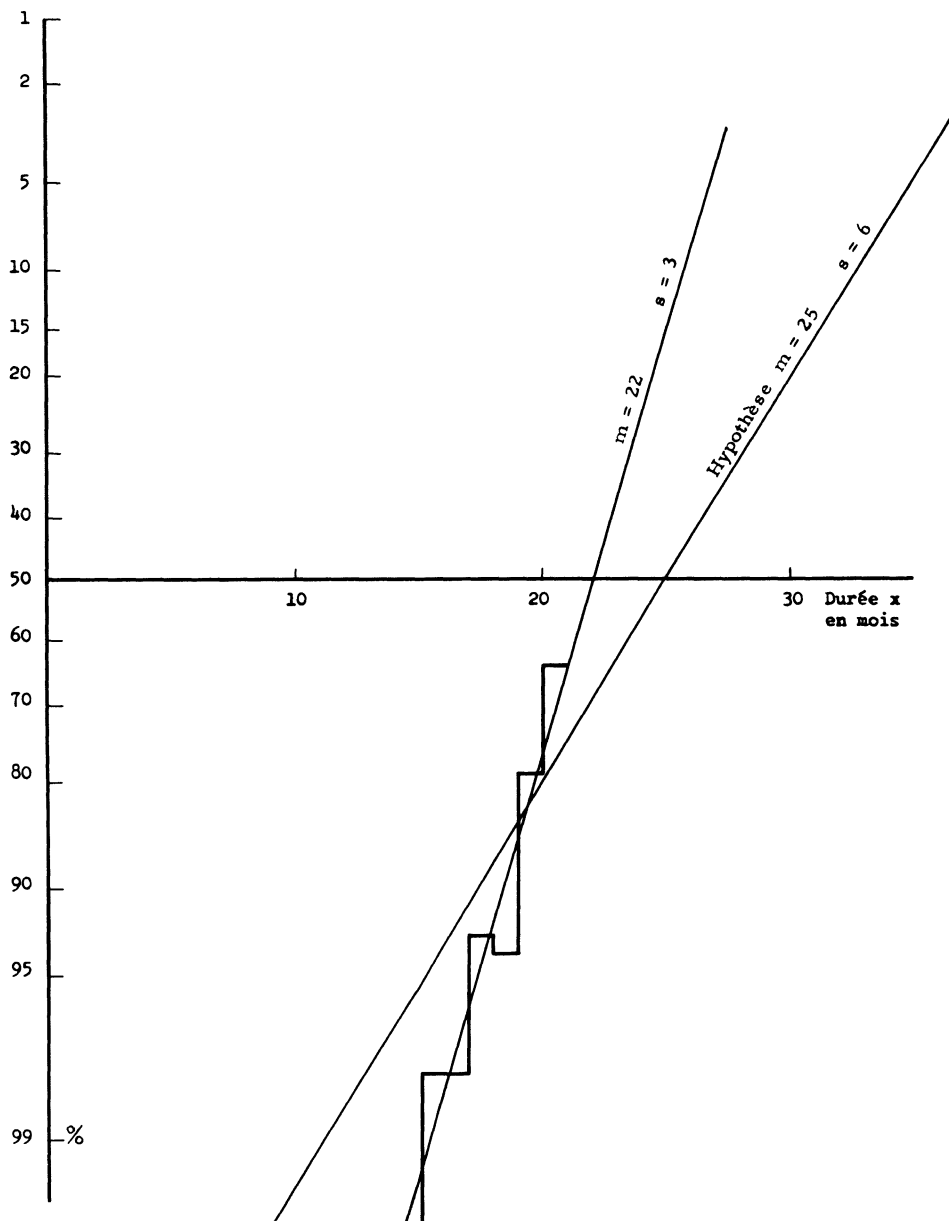


Fig. 5 - Probabilité de dépasser l'âge x : $F(x)$.

Pour appliquer le χ^2 , il faut que les classes contiennent un nombre théorique, np_i , au moins égal à 5, ce qui conduit à pratiquer des regroupements.

TABLEAU 2

Classes	Morts	en vie		Total
	1 + 2 + 4 + 5	3	6	
n_i	8	9	19	36
np_i	6,14	10,5	19,36	36
$\frac{(n_i - np_i)^2}{np_i}$	0,56	0,21	0,01	$\chi^2 = 0,78$

Une telle valeur de χ^2 , pour $v = 3 - 1 = 2$ degrés de liberté, a 73 % de chances d'être dépassée. L'hypothèse ($m = 25$, $s = 6$), bien qu'elle soit visuellement peu satisfaisante est cependant acceptable.

Nous avons ainsi recherché toutes les droites de Henry qui sont acceptables au seuil 5 %, soit par un test de χ^2 si la condition $np_i > 5$ permet de faire au moins deux classes, soit par un test binomial par "morts" et "vivants" dans le cas contraire.

Nous avons ainsi défini les deux familles de droites de Henry de la figure 6. Cette figure incite à une certaine prudence dans l'utilisation de la droite de Henry pour déterminer m et s à partir des premières valeurs observées de $F(x)$. Des tests plus rigoureux doivent être mis en oeuvre, qui donnent de m non pas l'estimation la plus probable (22 mois dans l'exemple ci-dessus) mais un véritable intervalle de confiance.

4- DELIMITATION DU DOMAINE DE CONFIANCE DE LA MOYENNE ET DE L'ECART TYPE.

En fait s est une inconnue au même titre que m et, à partir du moment où l'hypothèse de normalité des durées est admise, le problème est de déterminer les deux paramètres m et s qui sont nécessaires et suffisants pour caractériser complètement la loi de probabilité des durées.

Le premier travail est donc de délimiter le domaine de confiance de l'hypothèse dans le plan (m , s) : ultérieurement, et ce sera l'objet du paragraphe 5, on pourra utiliser une information a priori sur s pour améliorer l'estimation de m .

Une méthode rapide, presque entièrement graphique, est fondée sur l'emploi du test binomial.

Reprenons les valeurs de $F(x)$ des 36 équipements du tableau 1 et de la figure 5. Si nous appliquons le test binomial* à chaque tronçon de la courbe empirique $F(x)$, nous obtenons sur le graphique de Henry deux lignes de confiance à 95 % de part et d'autre de la courbe empirique (tableau 3 et figure 7).

* Voir abaque fig.11 dans l'article cité

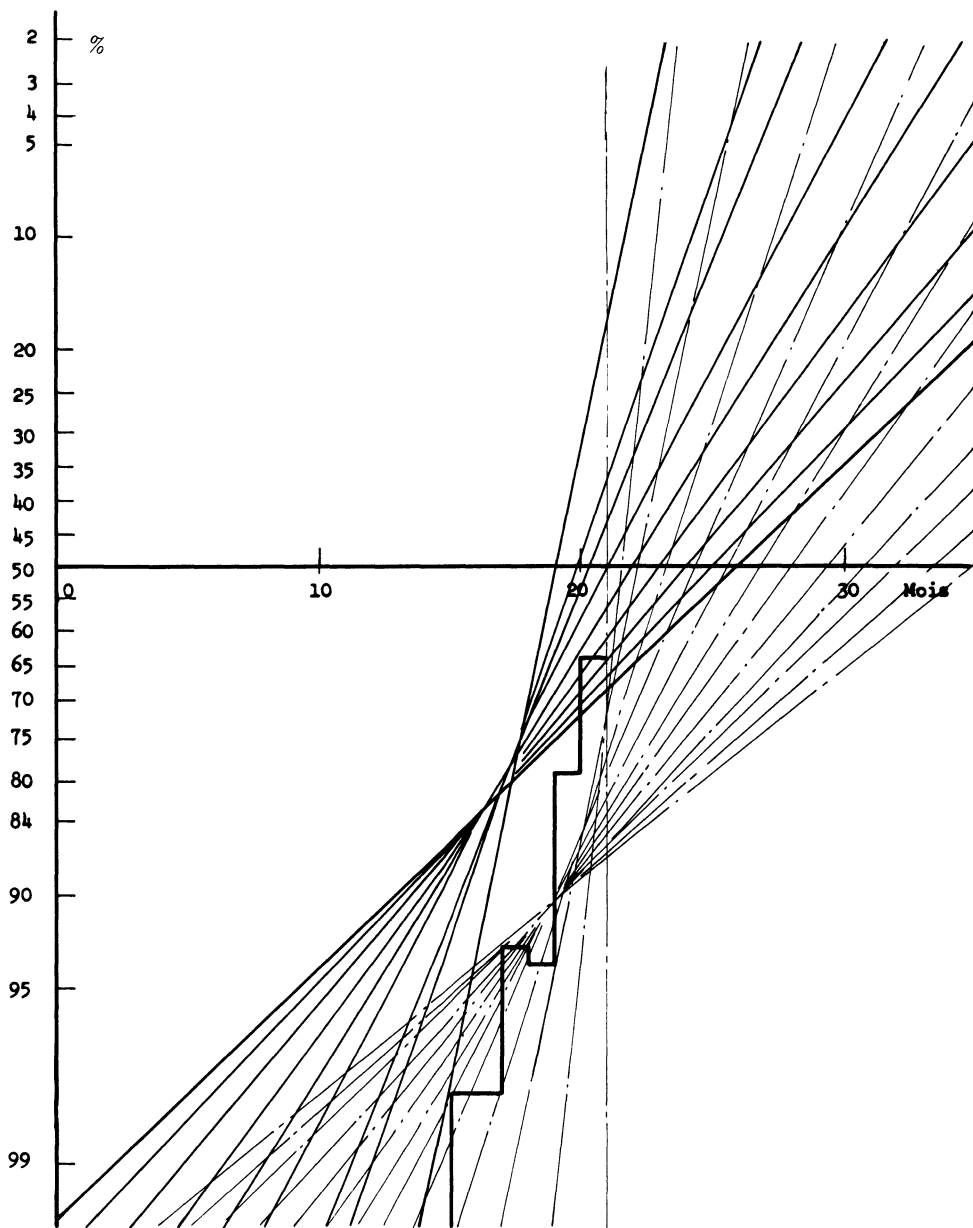


Fig. 6 - Familles de droites de Henry compatibles avec les observations au seuil 5 %.

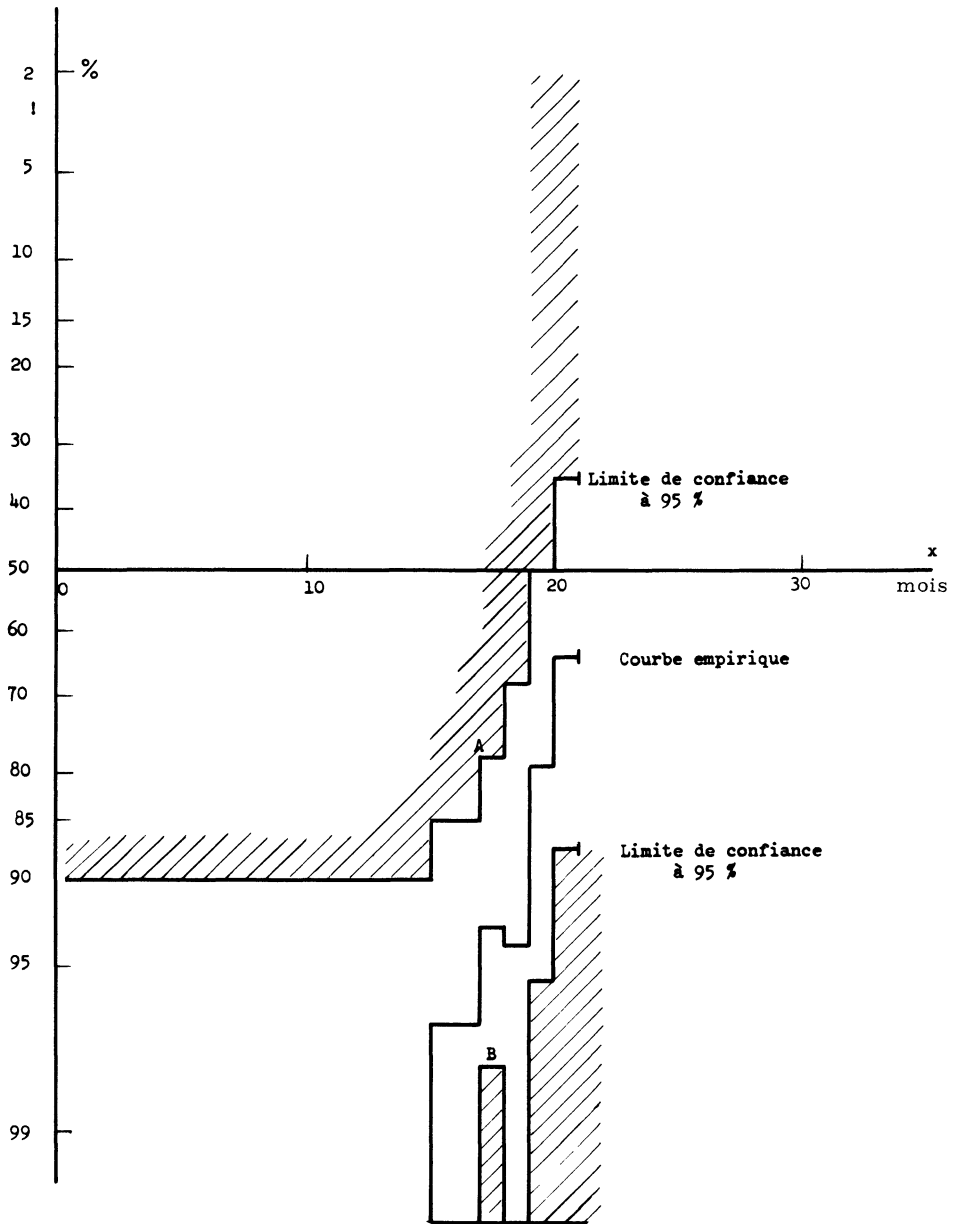


Fig. 7 - Probabilité de dépasser l'âge $x = F(x)$.

TABLEAU 3

Ancienneté x	0-15	15-16	16-17	17-18	18-19	19-20	20-21	21-22
éléments étudiés n	36	36	36	36	14	14	14	0
éléments ayant dépassé l'âge x	36	35	35	33	13	11	9	0
F(x)	1	0,97	0,97	0,92	0,93	0,79	0,64	?
Limite de) supér. confiance)	1	1	1	0,98	1	0,955	0,875	?
à 95 %) infér.	0,90	0,85	0,85	0,78	0,68	0,50	0,35	?

En tout point d'abscisse x intérieur au domaine de confiance 95 % de la figure 7 il est possible de faire passer une famille de droites de Henry compatibles avec l'observation empirique $F(x)$. Intéressons nous aux familles extrêmes, et prenons pour exemple $x = 17,5$ mois par exemple. La famille passant par le point $A(F(x) = 0,78)$ se représente dans le plan (m, s) par une droite. Il y a en effet une relation linéaire entre le m et le s de toutes les droites de Henry passant par A , puisque $F(x) = G\left(\frac{x-m}{s}\right)$ où G est l'intégrale de Gauss. Ainsi à tout point du graphique de Henry correspond une droite du plan (m, s) , de même qu'à tout point du plan (m, s) correspond une droite sur le graphique de Henry. Ecrire $G\left(\frac{x-m}{s}\right) = 0,78$ revient à poser $\frac{x-m}{s} = -0,77$, soit encore $m = x + 0,77s$. De la même manière le point $B(F(x) = 0,98)$ engendre une famille de droites de Henry représentées dans le plan (m, s) par la droite $m = x + 2,05s$ (fig. 8).

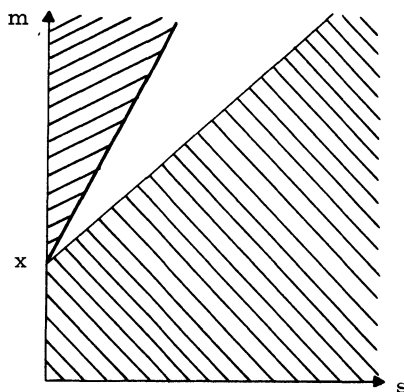


Fig. 8

Les deux droites de la figure 8 définissent une partie du plan (m, s) dans laquelle doit se trouver le point représentant la vraie population inconnue, pour être compatible, au seuil 5 %, avec l'information contenue au niveau d'ancienneté x .

Au début de l'utilisation du lot, aucune défaillance n'ayant encore été constatée, la limite de confiance supérieure de $F(x)$ n'a pas de sens et la pente de la droite $m = x + r s$ est infinie (figure 9). Ce n'est qu'à partir de la première défaillance que la limite supérieure prend un sens.

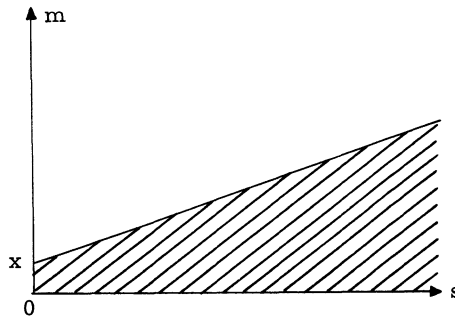


Fig. 9

Au contraire en fin d'utilisation, $F(x)$ est inférieur à 50 % et r est négatif (figure 10).

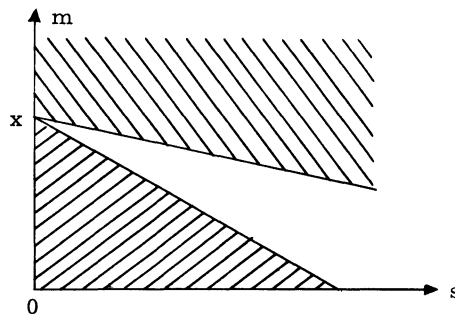


Fig. 10

Si l'on additionne les informations contenues à chaque niveau d'ancienneté x , on définit le domaine de confiance à 95 % du point (m, s) par la partie commune à tous les domaines relatifs aux différentes valeurs de x .

La figure 11 représente ainsi le domaine de confiance de m, s , pour le lot de 36 équipements pris comme exemple.

La construction graphique des droites $m = x + r s$ est facilitée par l'usage du graphique de Henry. On connaît évidemment le point $s = 0, m = x$;

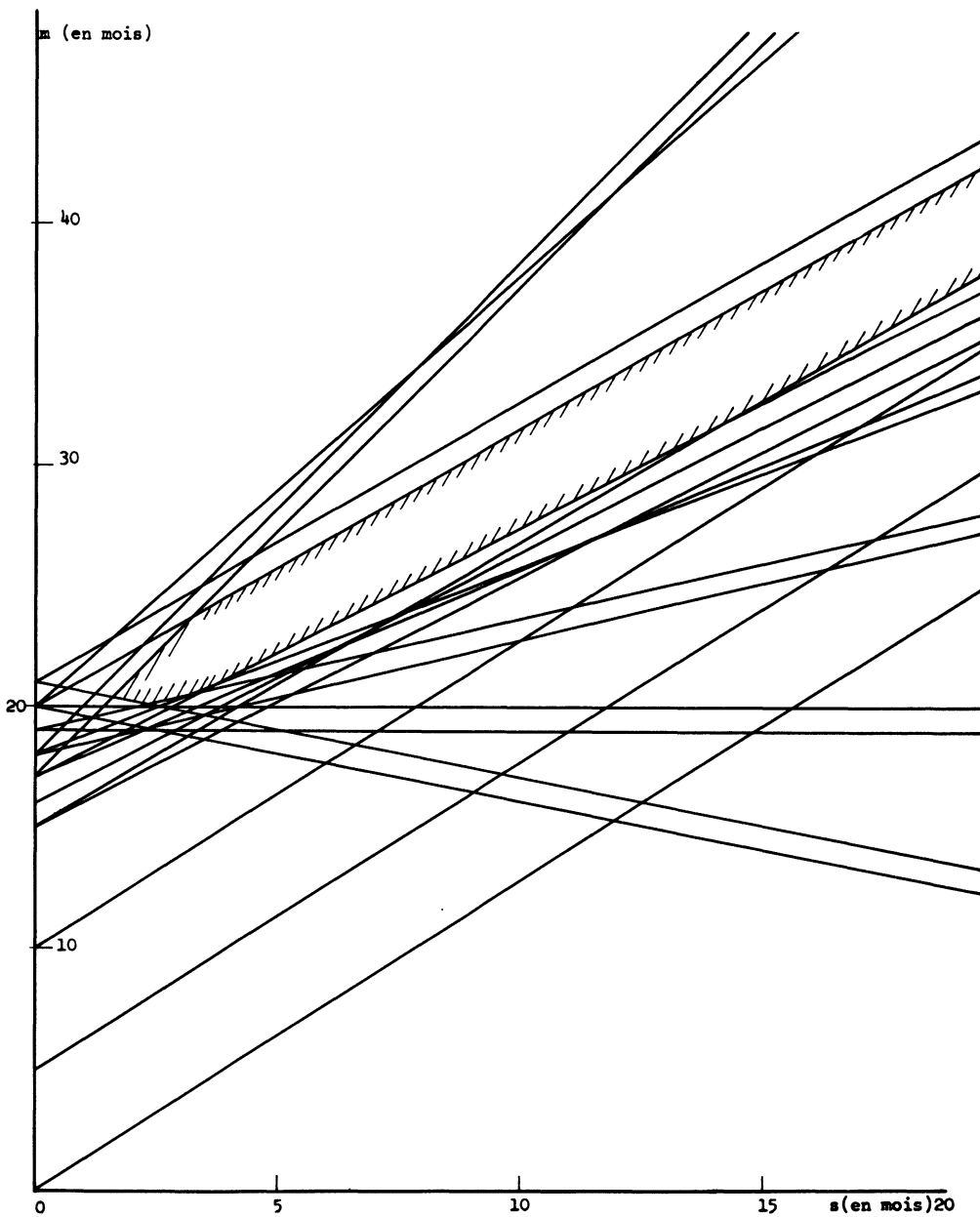


Fig. 11 - Domaine de confiance à 95 %.

il suffit d'y adjoindre un deuxième point que l'on obtient en faisant passer par le point $F(x)$ (A ou B sur la figure 7) une droite quelconque dont on mesure facilement le m et le s . On peut aussi remarquer que les ordonnées $F(x)$ du graphique de Henry sont proportionnelles à r et les graduer en r .

Pour montrer comment, le temps passant, l'information qui s'accumule améliore l'estimation de m et de s en resserrant le domaine de confiance autour de la vraie valeur, nous avons représenté sur les figures 13 et 14 différents tracés du domaine de confiance à 95 % de (m, s) pour le lot de 67 équipements, tous morts aujourd'hui, de la figure 12. Tout ce lot avait été démarré au même moment. Nous nous sommes fictivement replacés aux temps $t = 0, 24, 36, \dots, 72$ mois après le démarrage pour définir, par la méthode exposée dans ce paragraphe, le domaine de confiance du point (m, s) .

Le moment où le domaine de confiance se referme peut être précisé ; ce phénomène se produit quand la pente de la droite $m = x + r s$ du faisceau supérieur devient égale à la plus forte des pentes des droites du faisceau inférieur. Ainsi sur le graphique de Henry de la figure 12 ; pour $x = 39$ mois, la limite de confiance supérieure atteint le niveau 94,5 %. A ce moment la droite du faisceau supérieur a pour pente $r = 1,6$; elle est parallèle aux premières droites du faisceau inférieur (x compris entre 0 et 24 mois). A partir de ce moment le domaine se rétrécit rapidement et on peut donner à m un intervalle de confiance de plus en plus serré, alors que s reste largement indéterminé (Cf sur la figure 13 le graphique relatif à $t = 48$ mois).

Le domaine de confiance de (m, s) peut être vide, ce qui signifie qu'aucune droite de Henry n'est compatible avec les limites de confiance à 95 % de $F(x)$. Autant dire que les durées ne sont pas gaussiennes, comme sur la figure 4.

5 - UTILISATION D'UNE INFORMATION A PRIORI SUR s

Au début de l'exploitation d'un nouveau lot, il serait particulièrement précieux de disposer d'une information supplémentaire pour réduire l'indétermination existant sur m , en particulier tant que le domaine de confiance n'est pas fermé.

Le cas le plus simple est celui où l'on connaît à l'avance l'écart-type. Dans ce cas on utilise directement le graphique de Henry pour faire passer par les deux limites de confiance à 95 % de $F(x)$ les deux droites de pente s dont les intersections avec l'ordonnée $F(x) = 0,5$ donnent l'intervalle de confiance à 95 % de m (figure 15).

L'estimation de m est évidemment beaucoup plus rapide, comme le montre la figure 42, où nous avons représenté les étapes par lesquelles on serait passé pour estimer m , au fur et à mesure du temps, d'une part si on n'avait rien su de l'écart-type - et ce n'est alors qu'une présentation différente des graphiques des figures 13 et 14 - d'autre part si on avait su à l'avance que s valait 10 mois. Cette information supplémentaire aurait permis dans ce cas d'estimer la limite supérieure de m environ 20 mois plus tôt.

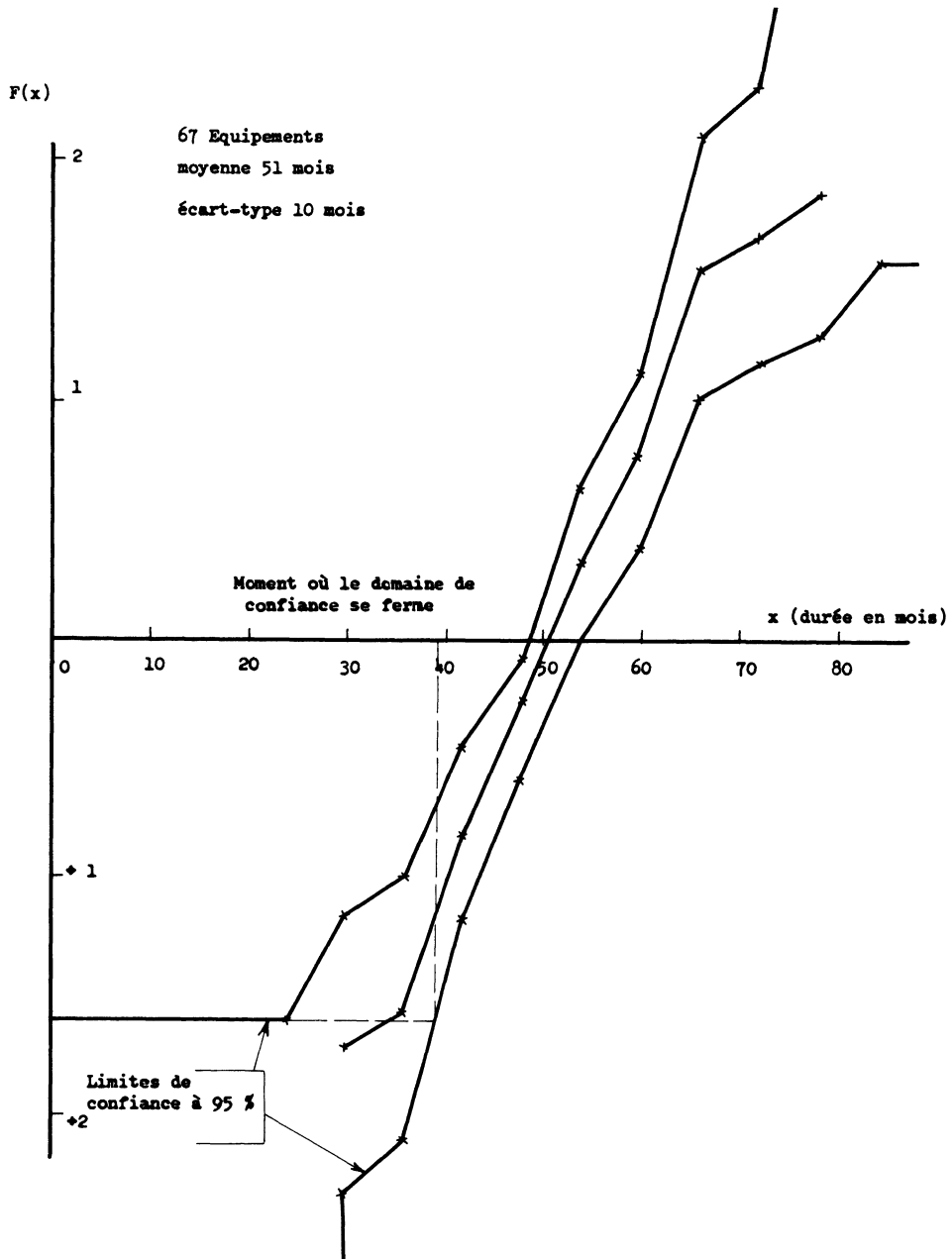


Fig. 12 - 67 équipements ; moyenne 52 mois ; écart-type 10 mois.

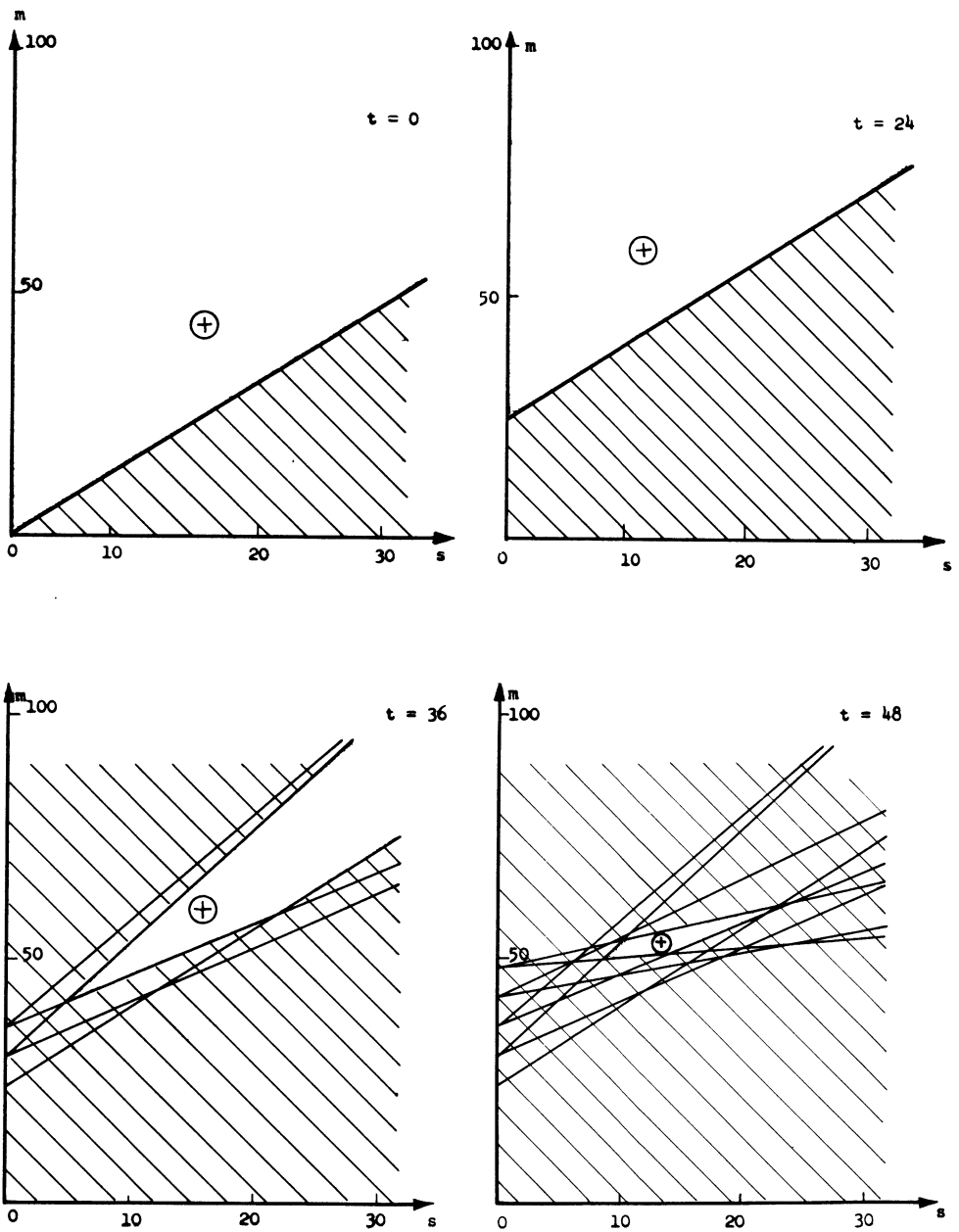


Fig. 13

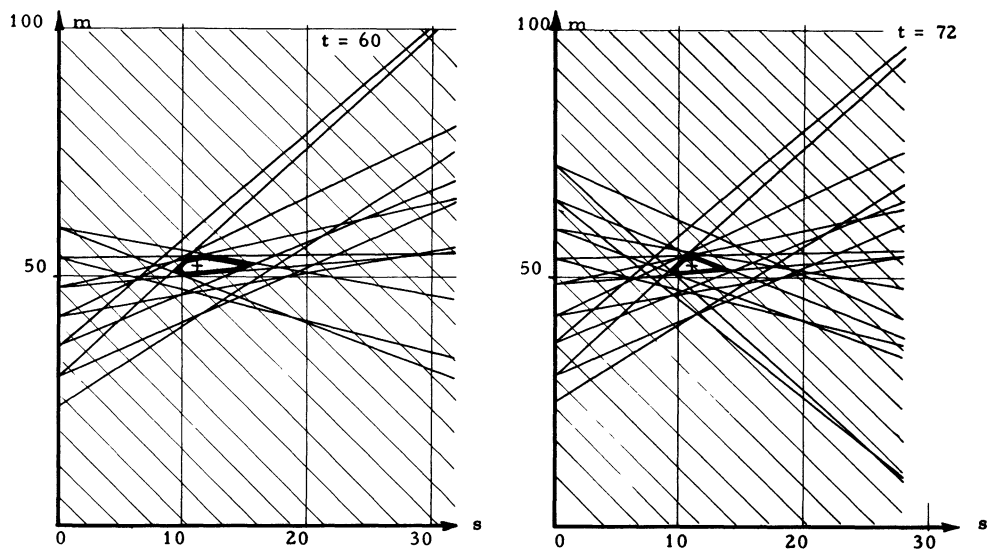


Fig. 14

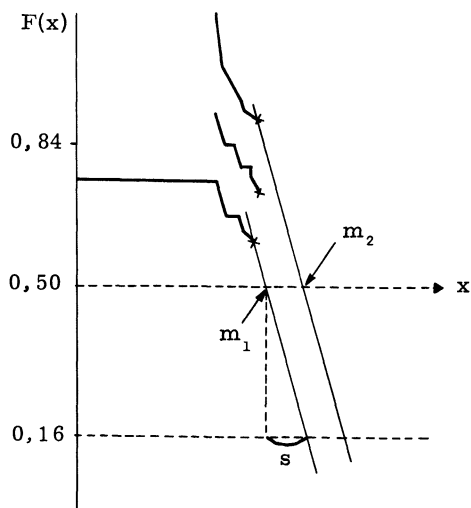
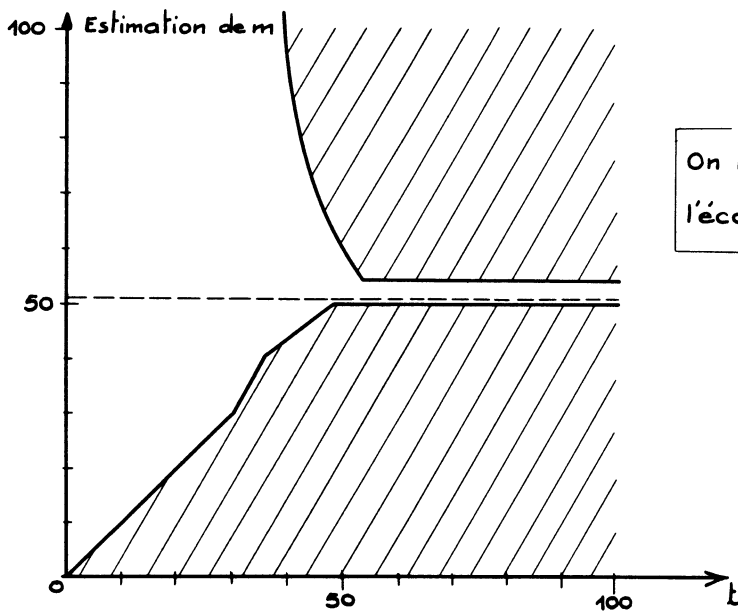
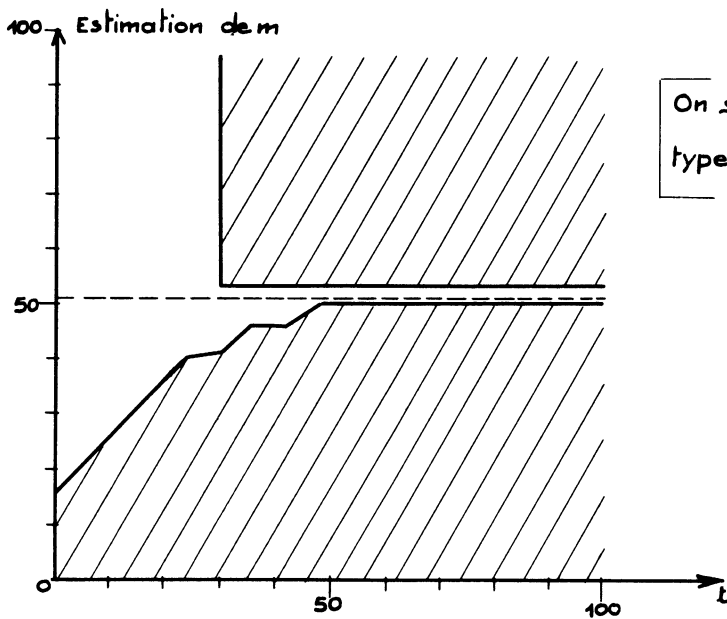


Fig. 15



On ne connaît pas
l'écart - type



On sait que l'écart -
type vaut 10

Fig. 16

S'il n'est pas possible de connaître a priori l'écart-type, il arrive souvent que l'étude de lots d'équipements analogues ait apporté une information sur un certain domaine de variation possible de s .

Dans le cas des cathodes des cellules d'électrolyse pour l'aluminium, nous avons ainsi recueilli une information sur s en étudiant 45 lots différents. Cette étude n'a pas fait apparaître, contrairement à notre attente, de relation entre la moyenne et l'écart-type, mais elle a permis de tracer l'histogramme des 45 valeurs s observées (figure 17). Notons au passage que ces écarts-types suivent sensiblement une loi de distribution log-normale (figure 18).

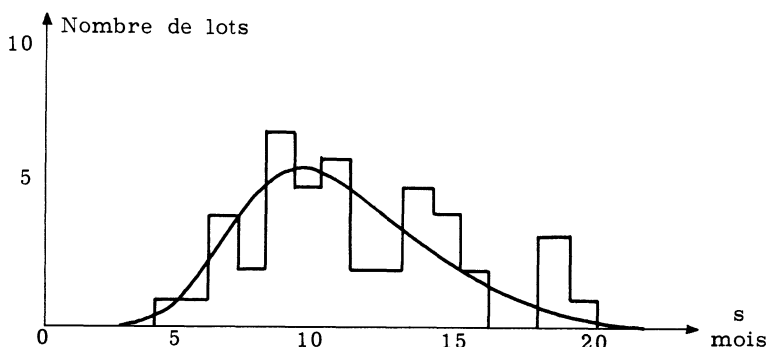


Fig. 17

Nous pouvons donc définir une fonction de répartition $\varpi(s)$ telle que $\varpi(s) ds$ soit la probabilité que s soit compris entre s et $s + ds$. Nous allons utiliser cette information a priori pour calculer l'intervalle de confiance de m .

A un instant t donné, on a constaté K défaillances parmi n . Calculons la vraisemblance de $m < m_1$. Si l'écart-type était s , la probabilité de défaillance serait $p = G\left(\frac{t - m_1}{s}\right)$ où G est l'intégrale de Gauss.

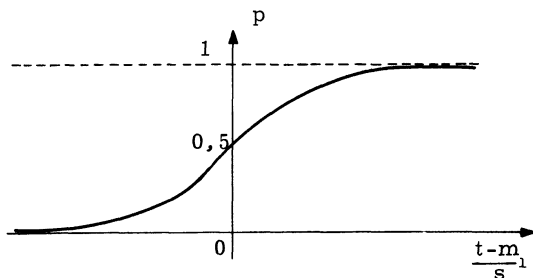


Fig. 19

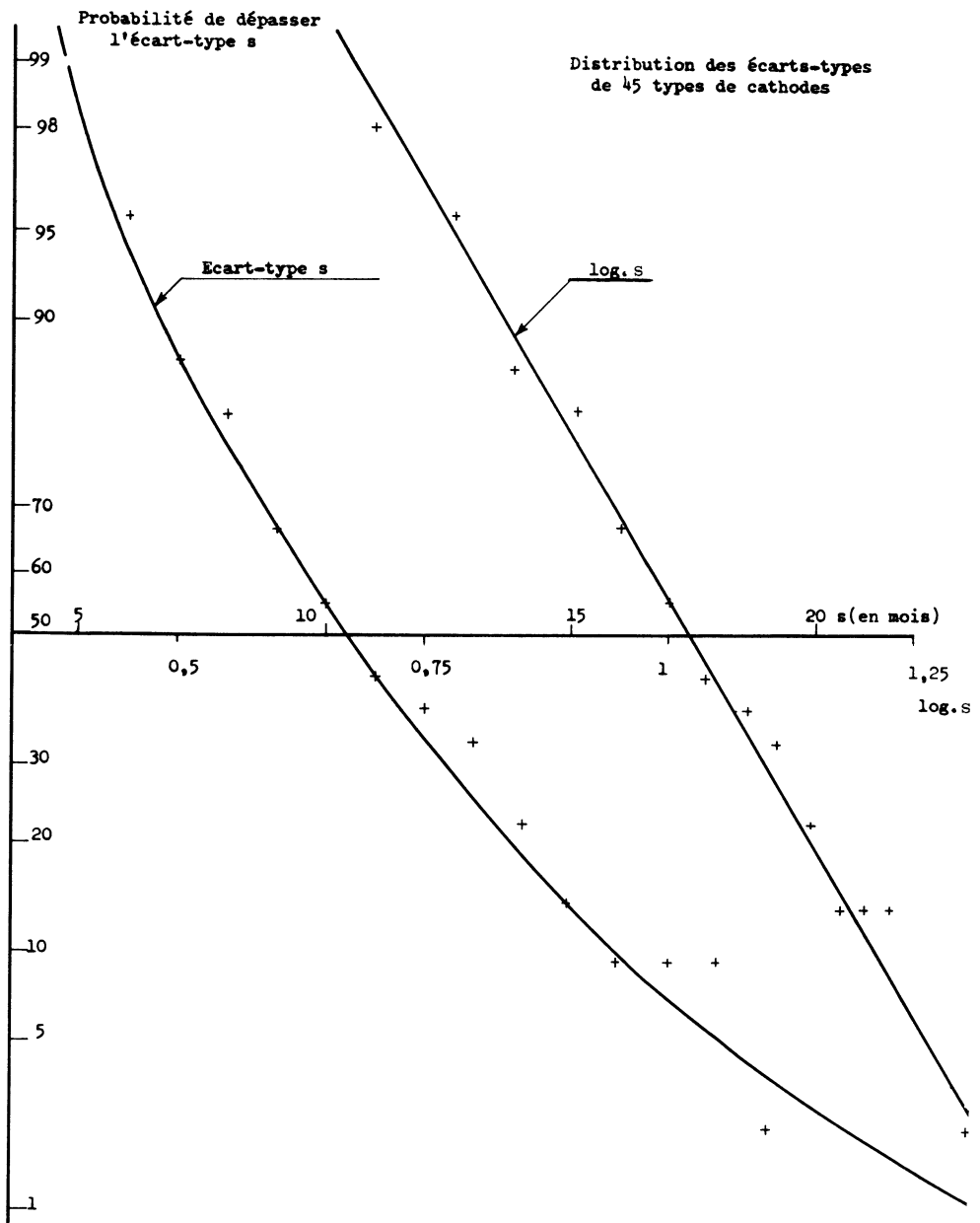


Fig. 18 - Distribution des écarts-types de 45 types de cathodes.

Si m_1 et s étaient vrais, la probabilité de constater de 0 à K défaillances avant le temps t serait $P(K, n, p)$, probabilité cumulée de la loi binomiale.

Si s était vrai, la vraisemblance que m soit inférieur à m_1 s'écrirait :

$$\text{Prob.}(m < m_1)_{s, s} = P \left[K, n, G \left(\frac{t - m_1}{s} \right) \right]$$

Compte tenu des probabilités a priori $\varpi(s)$, on écrit :

$$\text{Prob.}(m < m_1) = \int_0^\infty P \left[K, n, G \left(\frac{t - m_1}{s} \right) \right] \varpi(s) ds$$

Le tableau 5 donne un exemple de calcul ; les probabilités P y ont été calculées par la loi de Poisson.

TABLEAU 4

s	t = 30	K = 3		n = 67	m ₁ = 40	
	$\frac{t - m_1}{s} = r$	G(r)	n G	P(n G, K)	$\varpi(s) \Delta s$	P $\varpi(s) \Delta s$
0,5	- 20	0	0	1	0	0
1,5	- 6,67	0	0	1	0	0
2,5	- 4,00	0	0	1	0	0
3,5	- 2,86	0,0021	0,140	1	0,003	0,0030
4,5	- 2,22	0,0132	0,885	0,9866	0,017	0,0170
5,5	- 1,83	0,0336	2,250	0,7823	0,040	0,0313
6,5	- 1,54	0,0618	4,140	0,4085	0,070	0,0276
7,5	- 1,33	0,0918	6,150	0,1394	0,090	0,0125
8,5	- 1,18	0,1190	7,980	0,0424	0,120	0,0051
9,5	- 1,05	0,1469	9,850	0,0118	0,120	0,0013
10,5	- 0,95	0,1711	11,450	0,0036	0,110	0,0003
11,5	- 0,87	0,1921	12,800	0,0013	0,090	0,0001
12,5					0,080	
13,5					0,070	
14,5					0,056	
15,5					0,034	
16,5					0,030	
17,5					0,019	
18,5					0,012	
19,5					0,011	
20,5					0,007	
21,5					0,005	
22,5					0,004	
>23					0,012	
					1,000	0,0982 = Prob(m < 40)

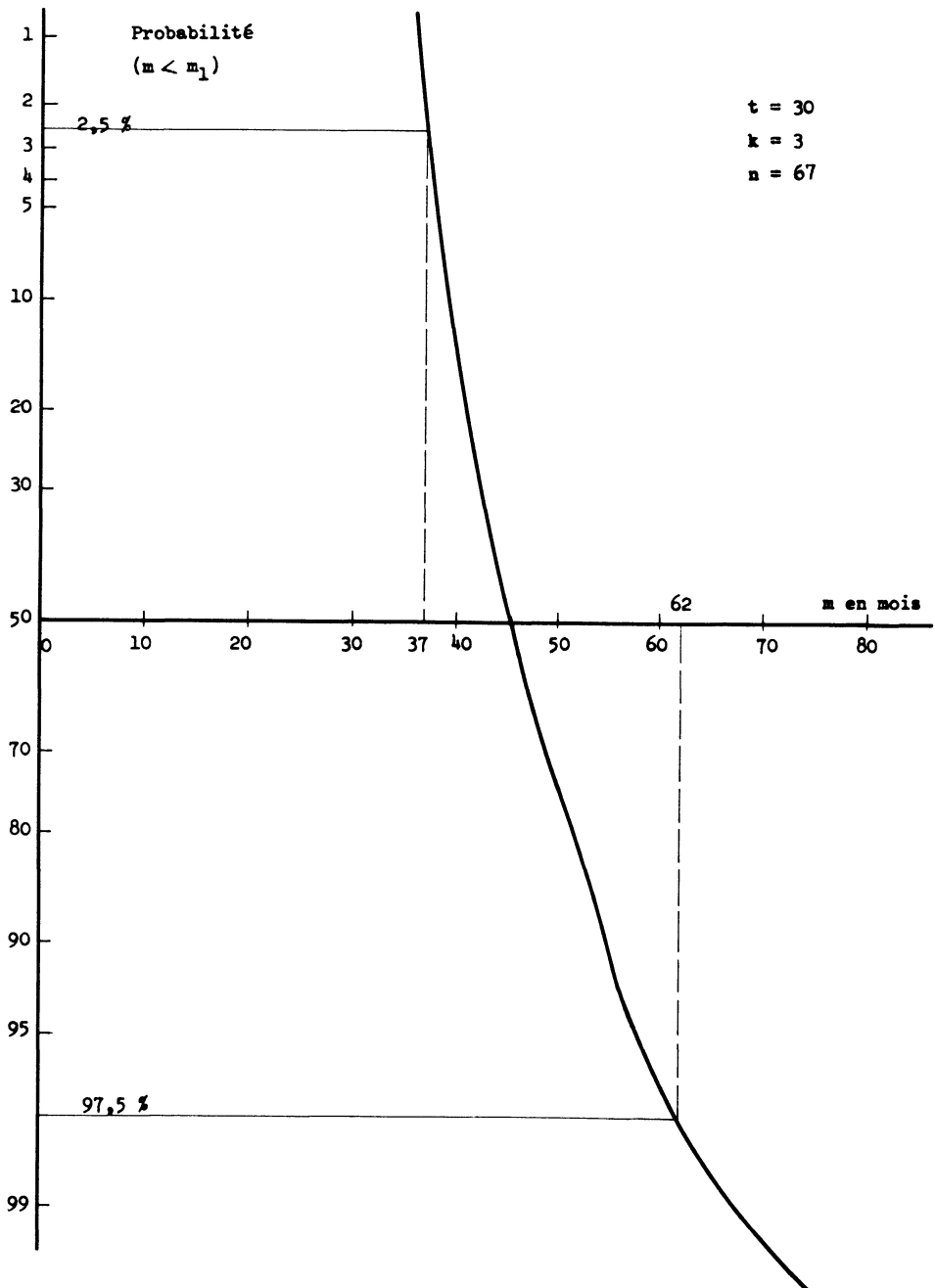


Fig. 20 - Probabilité ($m < m_1$).

En faisant le calcul pour différentes valeurs de m_1 et en reportant les vraisemblances obtenues sur un graphique comme celui de la figure 20, on obtient facilement les limites à 95 % de m .

Pour montrer l'intérêt de ce mode de calcul nous l'avons appliqué au lot de 67 équipements qui nous a déjà servi d'exemple et nous avons tracé la figure 21 où l'on voit comment l'estimation de m se précise rapidement au début de l'exploitation du lot, à condition d'utiliser l'information a priori que contient la loi de probabilité $\varpi(s)$: on obtient ainsi une estimation presque aussi rapide que si l'on connaissait à l'avance l'écart-type. Par contre il faut abandonner cette façon de faire au profit de la méthode graphique dès qu'assez de défaillances ont été constatées pour que l'écart-type vrai commence à se dégager, ce qui se produit dans notre exemple à 48 mois, quand 40 % des équipements sont morts.

Notons que, tant qu'aucune défaillance n'a encore été constatée, le calcul de la limite supérieure de m n'a évidemment pas davantage de sens dans cette méthode que dans l'autre.

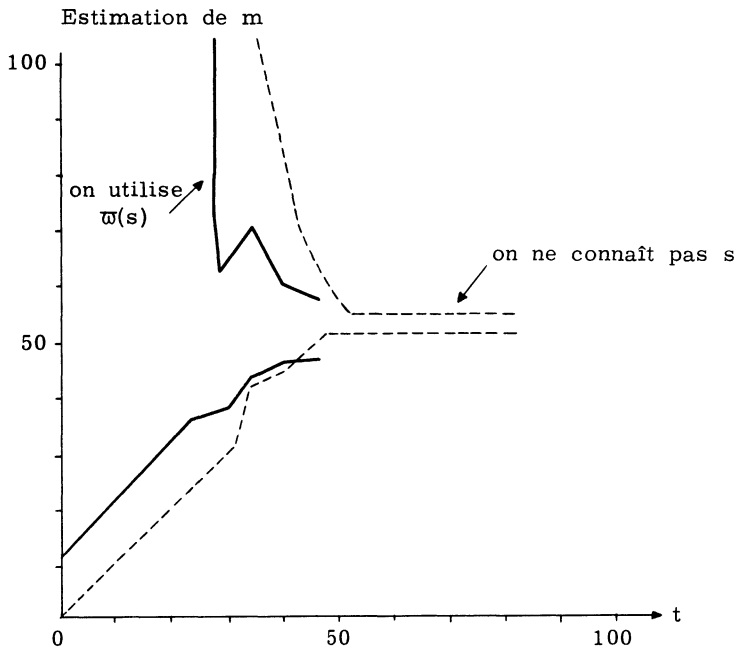


Fig. 21

6 - CONCLUSION

Les méthodes d'estimation, auxquelles est consacré cet article et que l'on pourrait qualifier de méthodes d'estimation dynamiques, ouvrent la porte sur ce secteur assez spécial de la statistique où le temps joue le rôle principal, en ne mettant l'information à notre portée qu'au fur et à mesure du déroulement d'une histoire. Elles montrent qu'il n'est pas nécessaire, comme on pourrait le craindre, d'attendre la fin de l'histoire pour en connaître l'essentiel. Ainsi la durée probable d'un lot peut elle être pratiquement connue alors que 50 % de ses éléments sont encore en vie.

BIBLIOGRAPHIE

- A.K. GUPTA - Estimation of the mean and standard deviation of a normal population from a censored sample.
- J.P. GIVRY - Gestion d'un parc à renouvellement aléatoire.