REVUE DE STATISTIQUE APPLIQUÉE

P. FERIGNAC

Test de Kolmogorov-Smirnov sur la validité d'une fonction de distribution

Revue de statistique appliquée, tome 10, nº 4 (1962), p. 13-32
<http://www.numdam.org/item?id=RSA_1962__10_4_13_0>

TEST DE KOLMOGOROV-SMIRNOV
SUR LA VALIDITÉ D'UNE FONCTION DE DISTRIBUTION

P. FERIGNAC
Statisticien

I - LE PROBLEME -

La solution d'un grand nombre problèmes statistiques implique la connaissance de la distribution de la variable aléatoire X soumise au contrôle.

La distribution de X est fondamentale dans de nombreux tests d'hypothèses. Par exemple, le test t de Student, les cartes de contrôle de la moyenne et de la dispersion d'une caractéristique mesurable, le test F de Snédécor, le test T d'Hotelling,... supposent que les variables aléatoires obéissent à une loi normale.

Les problèmes sur la sûreté de fonctionnement d'un matériel sont basés sur la probabilité de durée de vie de ses éléments qu'il est nécessaire de connaître en vue de chiffrer la fiabilité d'un ensemble qui compte souvent plusieurs milliers de pièces élémentaires.

On a, en général, des raisons théoriques ou une expérience suffisante qui conduisent à l'adoption d'une forme déterminée pour la fonction de distribution $F(x)$ dont on sait estimer les paramètres. Dans ce qui suit nous supposons que $F(x) = \text{Prob}(X \leqslant x)$ est connue. Alors, ayant extrait un échantillon aléatoire d'effectif n, nous nous proposons de vérifier si la distribution de x dans l'échantillon est compatible avec la fonction de distribution théorique $F(x)$ ou si quelque cause contrôlable est intervenue qui infirme la validité de cette fonction pour la description des observations.

On juge habituellement l'ensemble des écarts entre les distributions théorique et observée à l'aide du test de χ^2 de K. Pearson. Ce test, indépendant de la forme de $F(x)$, demande un assez grand nombre d'observations puisqu'aucune des classes ne doit avoir un effectif trop faible et que, d'autre part, le groupement des résultats entraîne la perte d'une partie de l'information. Le test de Kolmogorov-Smirnov, qui n'est pas astreint à ces restrictions, et dont l'application est facile peut rendre de grands services dans la solution du problème posé.

II - DESCRIPTION DU TEST -

On considère la fonction de distribution $F(x)$ d'une <u>variable aléatoire continue</u> et les fréquences cumulées $S_n(x)$ d'un échantillons d'effectif n. Si les valeurs observées de x proviennent de la population décrite par $F(x)$, les écarts entre les probabilités à priori et les fréquences cumulées dépendent des fluctuations aléatoires de l'échantillonnage et, en vertu de la loi des grands nombres, tendent vers zéro lorsque n croit indéfiniment. Il est intuitif de caractériser la divergence entre $S_n(x)$ et $F(x)$ par le plus grand des écarts observés D_n, soit :

$$D_n = \text{plus grande valeur de } \left| S_n(x) - F(x) \right| .$$

Connaissant la loi de répartition de la grandeur aléatoire D_n, on peut alors, avec un risque fixé à l'avance, conclure si l'écart maximum entre $F(x)$ et $S_n(x)$ peut être imputé au hasard de l'échantillonnage ou, au contraire, est dû au fait que l'échantillon n'est pas extrait d'une population décrite par la fonction de distribution $F(x)$.

<u>La répartition de D_n est indépendante de $F(x)$</u>, le test d'hypothèse est donc un test non paramétrique.

La loi limite de D_n a été étudiée par Kolmogorov lorsque n tend vers l'infini [1](1) et par F. J. Massey Jr pour les petites valeurs de n [3] et [4].

<u>Exemple</u> :

Supposons que nous voulions tester l'hypothèse que 5 nombres 0,52 ; 0,65 ; 0,13 ; 0,71 et 0,58 ont été choisis au hasard dans une population de nombres uniformément répartis entre 0 et 1. La fonction de distribution est alors $F(x) = x$. Si x_k est une valeur de x, telle que parmi les n valeurs de l'échantillon classées selon l'ordre de grandeur croissante, k soient inférieures ou égales à x_k, on a :

$$S_n\,(x_k) = \frac{k}{n} \text{ pour } x_k \leqslant x < x_{k+1},$$

(x_k et x_{k+1} sont les valeurs observées de rangs k et k + 1)

d'où le graphique de la figure 1.

$S_5(x)$ est une ligne polygonale en escalier qui présente 5 discontinuités, quand x franchit la valeur 0,52, $S_5(x)$ passe de 0,20 à 0,40, et la distance maximum D_5 réalisée pour x = 0,52 est égale à 0,52 - 0,20 = 0,32.

Si nous nous reportons à la table II on lit qu'on a 20 chances sur 100 de trouver du fait du hasard, une valeur de D_5 supérieure à 0,446 pour n = 5. Nous concluons que l'écart 0,32 n'est pas significatif. Nous pouvons donc considérer comme raisonnable l'hypothèse de la distribution uniforme de la population-mère d'où nous avons tiré l'échantillon.

---

(1) Les chiffres entre crochets sont des renvois aux références citées à la fin de l'article.

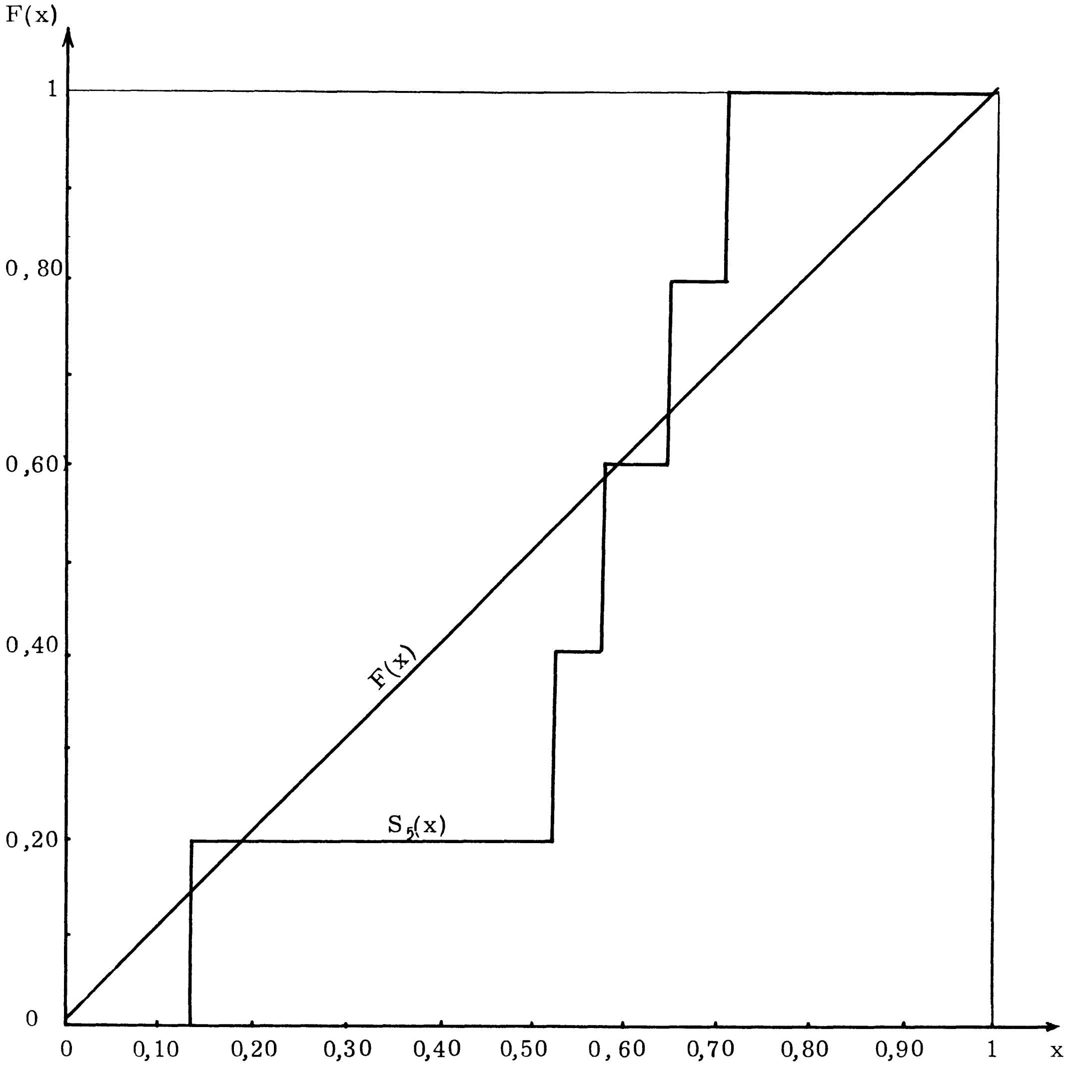

Figure 1

III - ETUDES THEORIQUES FONDAMENTALES -

1 - Echantillons de grande taille.

Kolmogorov a étudié la distribution de D_n dans le cas où n tend vers l'infini ("Bulletin de l'Académie des Sciences de l'U.R.S.S., 1933).

Théorème de Kolmogorov.

Soient $F(x)$ *une fonction de distribution continue et* D_n *une variable aléatoire définie par*

$$D_n = \textit{borne supérieure } \left| S_n(x) - F(x) \right| .$$

Pour toute valeur de $z \geqslant 0$, *si* $n \longrightarrow \infty$ *on a :*

$$\Pr(D_n \leqslant z n^{-\frac{1}{2}}) \longrightarrow L(z)$$

où $L(z)$ *est la fonction de distribution cumulée donnée par*

$$L(z) = 1 - 2 \sum_{\nu=1}^{\infty} (-1)^{\nu-1} e^{-\nu^2 z^2}$$

Pour $z < 0$ *on a évidemment* $L(z) = 0$.

W. Feller, a donné de ce théorème une démonstration qui, bien que plus simple que celle de Kolmogorov, reste assez difficile ; nous ne la reproduisons pas ici. Le lecteur qui désire approfondir cette question pourra se reporter à la référence [1].

N. Smirnov a calculé la table de L(z) (publiée en 1939 dans le "Bulletin Mathématique de l'Université de Moscou") que nous reproduisons d'après Annals of Mathematical Statistics [6].

Un problème annexe du problème posé au § I est celui qui consiste à comparer les fréquences cumulées observées dans deux échantillons afin de tester l'hypothèse de leur extraction de la même population-mère. Le critère de vraisemblance de cette hypothèse est fourni par le théorème de Smirnov. [1]

Théorème de Smirnov.

Soient $(x_1\ x_2 \ldots\ldots x_m)$ *et* $(y_1\ y_2 \ldots\ldots y_n)$ *deux échantillons de variables aléatoires indépendantes continues ayant la même fonction de distribution* $F(x)$.

Si l'on représente par $S_m(x)$ *et* $T_n(x)$ *les fréquences cumulées dans chacun des échantillons on définit la variable aléatoire*

$$D_{m,n} = \textit{borne supérieure} \left| S_m(x) - T_n(x) \right|$$

Posons

$$N = \frac{mn}{m + n}$$

et supposons que $m \longrightarrow \infty$, $n \longrightarrow \infty$ *de manière que* $\frac{m}{n} \longrightarrow a$ *où a est une constante. Dans ces conditions, pour toute valeur de* $z \geqslant 0$ *on a :*

$$\Pr.\ (D_{m,n} \leqslant z.\ N^{-\frac{1}{2}}) \longrightarrow L(z)$$

où $L(z)$ *est la fonction définie par le théorème de Kolmogorov.*

Une démonstration de ce théorème est donnée par W. Feller [1].

La table I peut être utilisée pour la comparaison de deux distributions expérimentales lorsque m et n sont grands.

TABLE I

Table de la fonction L(z)

z	L(z)	z	L(z)	z	L(z)
0,28	0,000 001	0,69	0,272 189	1,10	0,822 282
0,29	0,000 004	0,70	0,288 765	1,11	0,829 950
0,30	0,000 009	0,71	0,305 471	1,12	0,837 356
0,31	0,000 021	0,72	0,322 265	1,13	0,844 502
0,32	0,000 046	0,73	0,339 113	1,14	0,851 394
0,33	0,000 091	0,74	0,355 981	1,15	0,858 038
0,34	0,000 171	0,75	0,372 833	1,16	0,864 442
0,35	0,000 303	0,76	0,389 640	1,17	0,870 612
0,36	0,000 511	0,77	0,406 372	1,18	0,876 548
0,37	0,000 826	0,78	0,423 002	1,19	0,882 258
0,38	0,001 285	0,79	0,439 505	1,20	0,887 750
0,39	0,001 929	0,80	0,455 857	1,21	0,893 030
0,40	0,002 808	0,81	0,472 041	1,22	0,898 104
0,41	0,003 972	0,82	0,488 030	1,23	0,902 972
0,42	0,005 476	0,83	0,503 808	1,24	0,907 648
0,43	0,007 377	0,84	0,519 366	1,25	0,912 132
0,44	0,009 730	0,85	0,534 682	1,26	0,916 432
0,45	0,012 590	0,86	0,549 744	1,27	0,920 556
0,46	0,016 005	0,87	0,564 546	1,28	0,924 505
0,47	0,020 022	0,88	0,579 070	1,29	0,928 288
0,48	0,024 682	0,89	0,593 316	1,30	0,931 908
0,49	0,030 017	0,90	0,607 270	1,31	0,935 370
0,50	0,036 055	0,91	0,620 928	1,32	0,938 682
0,51	0,042 814	0,92	0,634 286	1,33	0,941 848
0,52	0,050 306	0,93	0,647 338	1,34	0,944 872
0,53	0,058 534	0,94	0,660 082	1,35	0,947 756
0,54	0,067 497	0,95	0,672 516	1,36	0,950 512
0,55	0,077 183	0,96	0,684 636	1,37	0,953 142
0,56	0,087 577	0,97	0,696 444	1,38	0,955 650
0,57	0,098 656	0,98	0,707 940	1,39	0,958 040
0,58	0,110 395	0,99	0,719 126	1,40	0,960 318
0,59	0,122 760	1,00	0,730 000	1,41	0,962 486
0,60	0,135 718	1,01	0,740 566	1,42	0,964 552
0,61	0,149 229	1,02	0,750 826	1,43	0,966 516
0,62	0,163 225	1,03	0,760 780	1,44	0,968 382
0,63	0,177 753	1,04	0,770 434	1,45	0,970 158
0,64	0,192 677	1,05	0,779 794	1,46	0,971 846
0,65	0,207 987	1,06	0,788 860	1,47	0,973 448
0,66	0,223 637	1,07	0,797 636	1,48	0,974 970
0,67	0,239 582	1,08	0,806 128	1,49	0,976 412
0,68	0,255 780	1,09	0,814 342	1,50	0,977 782

TABLE I (Suite)

z	L(z)	z	L(z)	z	L(z)
1,51	0,979 080	1,90	0,998 536	2,29	0,999 944
1,52	0,980 310	1,91	0,998 644	2,30	0,999 949
1,53	0,981 476	1,92	0,998 744	2,31	0,999 954
1,54	0,982 578	1,93	0,998 837	2,32	0,999 958
1,55	0,983 622	1,94	0,998 924	2,33	0,999 962
1,56	0,984 610	1,95	0,999 004	2,34	0,999 965
1,57	0,985 544	1,96	0,999 079	2,35	0,999 968
1,58	0,986 426	1,97	0,999 149	2,36	0,999 970
1,59	0,987 260	1,98	0,999 213	2,37	0,999 973
1,60	0,988 048	1,99	0,999 273	2,38	0,999 976
1,61	0,988 791	2,00	0,999 329	2,39	0,999 978
1,62	0,989 492	2,01	0,999 380	2,40	0,999 980
1,63	0,990 154	2,02	0,999 428	2,41	0,999 982
1,64	0,990 777	2,03	0,999 474	2,42	0,999 984
1,65	0,991 364	2,04	0,999 516	2,43	0,999 986
1,66	0,991 917	2,05	0,999 552	2,44	0,999 987
1,67	0,992 438	2,06	0,999 588	2,45	0,999 988
1,68	0,992 928	2,07	0,999 620	2,46	0,999 989
1,69	0,993 389	2,08	0,999 650	2,47	0,999 990
1,70	0,993 823	2,09	0,999 680	2,48	0,999 991
1,71	0,994 230	2,10	0,999 705	2,49	0,999 992
1,72	0,994 612	2,11	0,999 728	2,50	0,999 9925
1,73	0,994 972	2,12	0,999 750	2,55	0,999 9956
1,74	0,995 309	2,13	0,999 770	2,60	0,999 9974
1,75	0,995 625	2,14	0,999 790	2,65	0,999 9984
1,76	0,995 922	2,15	0,999 806	2,70	0,999 9990
1,77	0,996 200	2,16	0,999 822	2,75	0,999 9994
1,78	0,996 460	2,17	0,999 838	2,80	0,999 9997
1,79	0,996 704	2,18	0,999 852	2,85	0,999 99982
1,80	0,996 932	2,19	0,999 864	2,90	0,999 99990
1,81	0,997 146	2,20	0,999 874	2,95	0,999 99994
1,82	0,997 346	2,21	0,999 886	3,00	0,999 99997
1,83	0,997 533	2,22	0,999 896		
1,84	0,997 707	2,23	0,999 904		
1,85	0,997 870	2,24	0,999 912		
1,86	0,998 023	2,25	0,999 920		
1,87	0,998 145	2,26	0,999 926		
1,88	0,998 297	2,27	0,999 934		
1,89	0,998 421	2,28	0,999 940		

2 - Petits échantillons.

Le théorème de Kolmogorov est un théorème limite qui s'applique lorsque $n \longrightarrow \infty$. Nous verrons un peu plus loin à partir de quelle valeur de n on peut utiliser la fonction L(z) avec une approximation suffisante pour les besoins usuels. F.J. Massey a étudié la distribution de D_n pour des échantillons de faible effectif [3] et publié la table II que nous reproduisons [4]. Cette table donne les valeur D_n qu'on a la probabilité α de dépasser du fait du hasard. Si $S_n(x)$ représente les fréquences cumulées et F(x) la fonction de distribution on a :

$$\text{Pr. borne supérieure } \left|S_n(x) - F(x)\right| > D_{n,\alpha} = \alpha ,$$

α représente donc le risque d'erreur de 1ère espèce.

TABLE II

Taille de l'échantillon	Risque de 1ère espèce (α)				
	0,20	0,15	0,10	0,05	0,01
1	0,900	0,925	0,950	0,975	0,995
2	0,684	0,726	0,776	0,842	0,929
3	0,565	0,597	0,642	0,708	0,828
4	0,494	0,525	0,564	0,624	0,733
5	0,446	0,474	0,510	0,565	0,669
6	0,410	0,436	0,470	0,521	0,618
7	0,381	0,405	0,438	0,486	0,577
8	0,358	0,381	0,411	0,457	0,543
9	0,339	0,360	0,388	0,432	0,514
10	0,322	0,342	0,368	0,410	0,490
11	0,307	0,326	0,352	0,391	0,468
12	0,295	0,313	0,338	0,375	0,450
13	0,284	0,302	0,325	0,361	0,433
14	0,274	0,292	0,314	0,349	0,418
15	0,266	0,283	0,304	0,338	0,404
16	0,258	0,274	0,295	0,328	0,392
17	0,250	0,266	0,286	0,318	0,381
18	0,244	0,259	0,278	0,309	0,371
19	0,237	0,252	0,272	0,301	0,363
20	0,231	0,246	0,264	0,294	0,356
25	0,21	0,22	0,24	0,27	0,32
30	0,19	0,20	0,22	0,24	0,29
35	0,18	0,19	0,21	0,23	0,27
supérieur à 35	$\frac{1,07}{\sqrt{n}}$	$\frac{1,14}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Comparaison des tables I et II.

Dans la table II F.J. Massey admet qu'à partir d'un échantillon de $n = 36$ observations on peut utiliser la table I qui donne les valeurs z aux seuils de confiance 0,80 ; 0,85 ; 0,90 ; 0,95 et 0,99. Il prend, en effet, pour D_n la valeur $\frac{z}{\sqrt{n}}$, z étant tiré de la table I. Il est intéressant de voir la rapidité de la convergence de la distribution exacte de D_n pour de petits échantillons vers les valeurs limites pour n infini, déduites de la table I. Pour cela on a récapitulé dans la table III les seuils de confiance calculés d'après la table I pour $n \longrightarrow \infty$ correspondant aux valeurs de D_n de la table II, pour $1 - \alpha = 0,80$; 0,85 ; 0,90 ; 0,95 et 0,99.

TABLE III

$P = Pr \ |S_n(x) - F(x)| \leqslant D_n$ d'après la table II

n = effectif de l'échantillon

Les probabilités dans le corps de la table sont à 0,005 près

n \ P	0,80	0,85	0,90	0,95	0,99
4	0,73	0,78	0,84	0,91	0,97
9	0,75	0,81	0,86	0,93	0,98
16	0,76	0,82	0,88	0,94	0,98
25	0,78	0,82	0,89	0,95	0,99
35	0,79	0,84	0,91	0,95	0,99

Nous constatons que les probabilités calculées à l'aide de la loi limite de la table I se rapprochent de leur vraie valeur quand n croît et que pour $n \geqslant 36$ l'approximation fournie par la table I est pratiquement suffisante. Donc n = 36 est un grand nombre à partir duquel le théorème de Kolmogorov s'applique au moins pour $\alpha \leqslant 0,20$.

IV - PUISSANCE DU TEST -

Considérons que nous appliquons le test de Kolmogorov - Smirnov en comparant $S_n(x)$ et $F_o(x)$. Nous pouvons commettre dans notre jugement une erreur de 1ère espèce que nous savons chiffrer : nous avons la probabilité α de rejeter à tort la fonction de distribution $F_o(x)$.

Le test peut comporter une erreur de 2ème espèce qui consiste à admettre la fonction $F_o(x)$ alors que l'échantillon provient d'une distribution décrite par $F_1(x)$. Si β est la probabilité de cette alternative, β caractérise le risque de 2ème espèce. La puissance du test $\pi = 1 - \beta$ mesure sa faculté de discrimination entre les deux distributions $F_o(x)$ et $F_1(x)$.

F.J. Massey [4] a étudié la puissance du test de Kolmogorov pour les grands échantillons (n > 36). Il introduit à cet effet, l'écart-maximum entre $F_o(x)$ et $F_1(x)$, soit : Δ = borne supérieure $|F_o(x) - F_1(x)|$ et démontre que la <u>puissance du test n'est jamais inférieure</u> à

$$\pi = 1 - \frac{1}{\sqrt{2\pi}} \int_{\lambda_1}^{\lambda_2} e^{-\frac{t^2}{2}} dt$$

avec

$$\lambda_1 = 2\,(\Delta\sqrt{n} - z), \qquad \lambda_2 = 2\,(\Delta\sqrt{n} + z),$$

n = effectif de l'échantillon supposé de grande taille,

$z = \frac{D_n}{\sqrt{n}}$, D_n correspondant au risque de 1ère espèce α adopté.

λ suit donc la loi normale N(0,1) dont la fonction de distribution est $\Phi(\lambda)$. Si l'on admet pour α les niveaux courants 0,05 et 0,01, on prendra pour z respectivement 1,36 et 1,63, d'où :

$$\pi_1 = 1 - \{\Phi\ [2(\Delta\sqrt{n} + 1,36)] - \Phi\ [2(\Delta\sqrt{n} - 1,36)]\}\ \text{pour}\ \alpha = 0,05$$

$$\pi_2 = 1 - \{\Phi\ [2(\Delta\sqrt{n} + 1,63)] - \Phi\ [2(\Delta\sqrt{n} - 1,63)]\}\ \text{pour}\ \alpha = 0,01$$

Les risques de 2ème espèce correspondants β_1 et β_2 sont évidemment

$$\Phi\ [2(\Delta\sqrt{n} + 1,36)] - \Phi\ [2(\Delta\sqrt{n} - 1,36)] \quad \text{et}$$

$$\Phi\ [2(\Delta\sqrt{n} + 1,63)] - \Phi\ [2(\Delta\sqrt{n} - 1,63)]$$

Ainsi lorsque $\Delta = 0,10$; $n = 400$ on a $\Delta\sqrt{n} = 2$

- pour $\alpha = 0,05$

$$\pi_1 = 1 - [\Phi(6,72) - \Phi(1,28)] \sim \Phi(1,28) \sim 0,90$$

$$\beta_1 \sim 0,10,$$

- pour $\alpha = 0,01$

$$\pi_2 = 1 - [\Phi(7,26) - \Phi(0,74)] \sim \Phi(0,74) \sim 0,77$$

$$\beta_2 \sim 0,23$$

Ces résultats montrent que lorsque le risque de 1ère espèce décroît, le risque de 2ème espèce croît, il faut donc se fixer à priori les 2 risques, eu égard au but que l'on se propose, et déterminer alors l'effectif de l'échantillon à prélever.

Les courbes de la puissance du test π en fonction de $\Delta\sqrt{n}$ sont représentées pour $\alpha = 5\ \%$ et $1\ \%$ sur la figure 2 qui permet de répondre rapidement à quelques questions telles que la suivante :

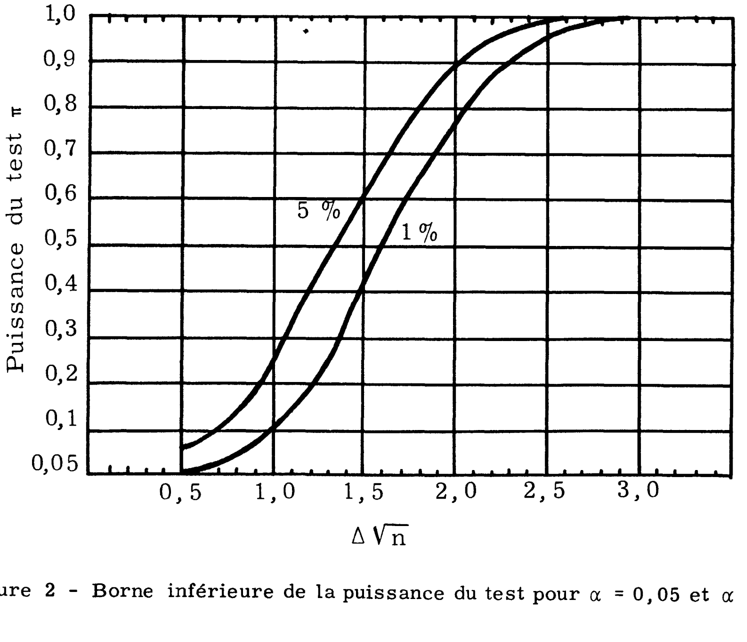

Figure 2 - Borne inférieure de la puissance du test pour $\alpha = 0,05$ et α 0,01.

Supposons que la différence maximum en valeur absolue de $F_0(x)$ et $F_1(x)$ est 0,2 et qu'on admette les risques α = 0,05 et β = 0,50, quelle est la taille de l'échantillon à prélever ? On lit sur le graphique l'abscisse 1,36 à l'intersection de la droite π = 0,50 et de la courbe 5 %.

Donc : $\Delta\sqrt{n} = 1{,}36$ et $n = \left(\frac{1{,}36}{0{,}2}\right)^2 \sim 46{,}24$. Un échantillon de 47 unités sera donc nécessaire.

Remarque :

La figure 2 donne la limite inférieure de la puissance π du test, celui-ci est donc en général plus puissant que ne l'indique une lecture sur le graphique. De même, le risque de 2ème espèce β est en général inférieur au risque qui s'en déduit ($\beta = 1 - \pi$).

V - APPLICATIONS NUMERIQUES -

1 - Observations individuelles.

a) On essaye pour un matériel une loi de survie exponentielle avec une valeur moyenne de 170 heures, la fonction de distribution des pannes à l'âge x, exprimé en heures, est : $F(x) = 1 - e^{-x/170}$ [2]

Cette fonction de distribution théorique est représentée sur la figure 3.

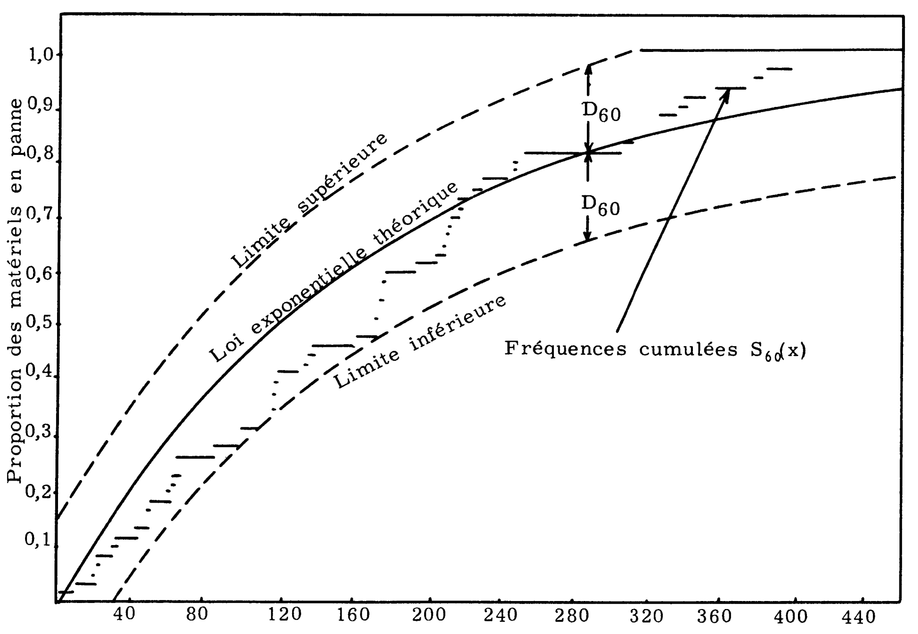

Figure 3 - Durée de service en heures, x.

Dans l'expérience décrite par J. A. Foster, [2] on prélève un échantillon de 60 unités pour lesquels on a noté les durées de vie ci-dessous classées par ordre de grandeur croissante.

Les fréquences cumulées observées S_{60} correspondantes sont représentées par la courbe en escalier de la figure 3.

TABLEAU I

1	50	114	171	212	314
9	60	116	174	213	327
18	62	116	174	219	327
21	63	117	175	220	339
23	67	118	175	225	343
29	67	133	176	230	356
34	84	135	193	247	376
43	100	139	204	248	384
48	102	163	209	250	409
48	111	171	210	310	410

Si l'on admet le risque de 1ère espèce $\alpha = 0,10$, d'après la table II on a : $D_{60} = \frac{1,22}{\sqrt{60}} = 0,1575$. Portant cette valeur de part et d'autre de la fonction de distribution théorique on obtient les deux limites en trait pointillé. La courbe en escalier ne doit pas franchir ces limites pour que la fonction de distribution exponentielle F(x) soit acceptable. On admet qu'une fois sur dix on la rejettera à tort ($\alpha = 0,10$).

Les courbes des fréquences cumulées touche 2 fois la limite inférieure de confiance :

- un peu avant 100 h,
- un peu avant 111 h.

Nous concluons donc au rejet de la fonction de distribution

$$F(x) = 1 - e^{-x/170}$$

A titre de vérification on calcule la distance de F(x) aux fréquences cumulées pour :

- $x = 99 : F(99) = 1 - e^{-99/170} = 0,442$

  $S_{60}(99) = \frac{17}{60} = 0,283$

  $d_{60} = 0,442 - 0,283 = 0,159 > 0,1575$

- $x = 110 : F(110) = 1 - e^{-110/170} = 0,476$

  $S_{60}(110) = \frac{19}{60} = 0,317$

  $d_{60} = 0,476 - 0,317 = 0,159 > 0,1575$

Les deux distance d_{60} sont supérieures à $D_{60} = 0,1575$ ce qui confirme la conclusion précédente.

b) Nous appliquons le test de Kolmogorov à la fonction de distribution théorique normale de moyenne 170 h et d'écart-type 110 h qui est représenté en trait plein sur la figure 4. En procédant comme dans le cas

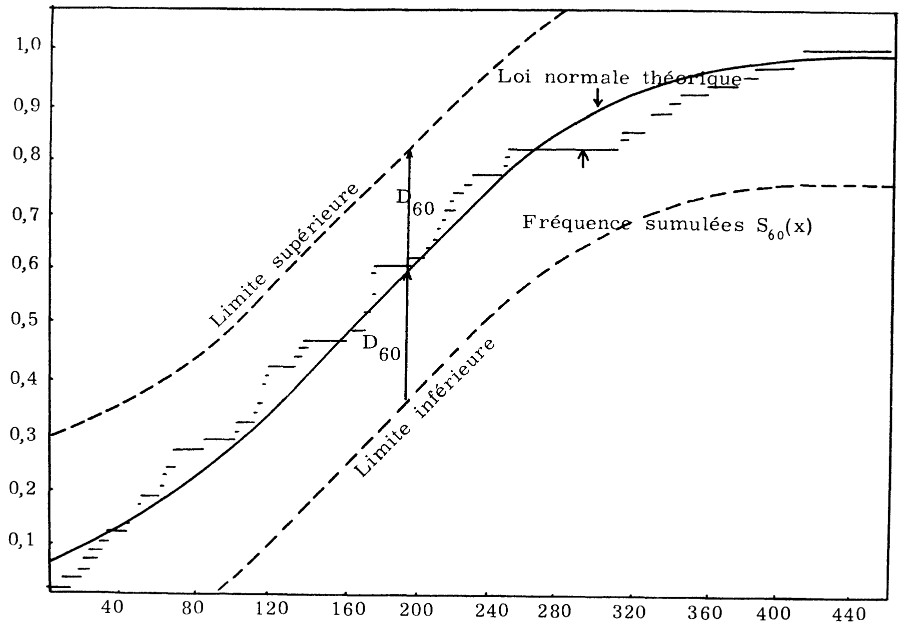

Figure 4 - Durée de service en heures, x.

a nous constatons que la courbe en escalier des fréquences cumulées observées reste comprise entre les deux limites de confiance. La distribution normale N(170 ; 110) peut donc être retenue pour décrire la durée de vie du matériel d'après l'information expérimentale dont nous disposons.

c) Puissance du test pour la discrimination entre les deux lois précédentes : exponentielle et normale.

Nous avons vu au § IV qu'il était possible de calculer une valeur minorée de la puissance du test.

Nous faisons le calcul en admettant l'hypothèse d'une distribution normale des durées de vie contre l'hypothèse d'une distribution exponentielle.

Avec les notions du § IV on mesure sur les figures 3 et 4 la distance maximale entre les deux fonctions de distribution théoriques : on a $\Delta = 0,17$.

$$\Delta \sqrt{n} = 0,17 \sqrt{60} = 1,317$$

$$z = 1,22$$

$$\lambda_1 = 2(1,317 + 1,22) = 5,074$$

$$\lambda_2 = 2(1,317 - 1,22) = 0,194,$$

d'où

$$\pi = 1 - [\Phi(5,074) - \Phi(0,194)] \sim 0,57$$

La puissance du test π n'est jamais inférieure à 0,57 ou encore, le risque de 2ème espèce β est inférieur ou égal à 0,43. En d'autres termes, en admettant la distribution normale au lieu de la distribution exponentielle nous nous tromperons au plus 43 fois sur 100 sans pouvoir mieux préciser ce risque.

Si nous admettions pour β la valeur maximale, 0,10 nous devrions avoir approximativement :

$$\Phi(\lambda_1) = 0,90$$

$$\lambda_1 = 1,28$$

d'où : $$2(0,17\sqrt{n} - 1,22) = 1,28$$

soit : $$n = 121,$$

il faudrait donc un échantillon de 121 observations du durée de vie pour lequel l'écart maximal, D_{121}, entre la fonction de distribution théorique et les fréquences cumulées serait $D_{121} = \frac{1,22}{11} = 0,1109$ (au lieu de $D_{60} = 0,1575$) correspondant à un risque de 1ère espèce $\alpha = 0,10$.

2 - Observations groupées.

Nous empruntons à F.J. Massey [4] l'exemple d'un échantillon de 511 unités provenant d'une population normale. Les données sont relevées dans la tableau 2.

La plus grande différence, en valeur absolue, entre les effectifs observés et théoriques est 12,41 qui correspond à une différence de $F(x)$ et $S_{511}(x)$ de $\frac{12,41}{511} = 0,024$. De la table II on tire la valeur de D_{511} correspondant au risque de 1ère espèce $\alpha = 0,05$: $D_{511} = \frac{1,36}{\sqrt{511}} = 0,060$. Nous concluons donc que l'écart n'est pas significatif au seuil habituel 0,05 et que l'échantillon peut provenir d'une population normale, les écarts entre les distributions théorique et observée étant imputables au hasard.

Comme nous avons un grand échantillon nous pouvons appliquer le théorème limite de Kolmogorov traduit par la table I. La valeur de z correspondant à l'écart maximal observé 0,024 est $z = 0,024\sqrt{511} \sim 0,54$. De la table I de $L(z)$ on tire :

$$\Pr(z \leqslant 0,54) = 0,067497,$$

on a donc environ 93 chances sur 100 d'observer, du fait du hasard de l'échantillonnage, un écart D_{511} supérieur à 0,024, l'accord est donc étonnament bon.

TABLEAU 2

Comparaison des distributions théorique et observée pour un échantillon d'effectif 511 extrait d'une population normale.

Limite supérieure de classe	Nombres d'observations cumulées		
	Observé	Théorique	Différence en valeur absolue
39,5	511	511,00	0
38,5	510	509,16	0,84
37,5	510	506,45	3,55
36,5	505	500,83	4,17
35,5	493	490,05	2,95
34,5	469	471,45	2,45
33,5	447	442,43	4,57
32,5	402	401,29	0,71
31,5	356	349,22	6,78
30,5	300	287,59	12,41*
29,5	228	223,36	4,64
28,5	162	161,73	0,27
27,5	114	109,66	4,34
26,5	73	68,52	4,48
25,5	43	39,50	3,50
24,5	24	20,90	3,10
23,5	14	10,12	3,88
22,5	9	4,50	4,50
21,5	2	1,79	0,21
20,5	2	0,61	1,39
19,5	1	0,20	0,80

Remarques :

1/ Le groupement des observations en classes tend à diminuer la valeur de D_n. En conséquence, le risque de 1ère espèce, α, est donc inférieur à celui qui est donné dans la table II. Toutefois, pour de grands échantillons le groupement n'entraine qu'une erreur négligeable si l'intervalle de classe n'est pas trop large et le nombre de classes pas trop faible (par exemple supérieur à 10). L'exemple ci-dessus est dans ce cas. De toute façon, grouper dans un très petit nombre de classes peut entraîner des conclusions erronées même pour de grands échantillons.

2/ Le test de Kolmogorov suppose que la fonction de distribution théorique est donnée à priori indépendamment de l'information fournie par l'échantillonnage. La distribution de la déviation maximale D_n n'est pas connue quand certains paramètres de la loi théorique sont estimés d'après les données de l'échantillon. Il est vraisemblable que l'estimation de certains paramètres en utilisant les observations de l'échantillon tend à diminuer la valeur critique de D_n correspondant au risque adopté. Donc, si la distance trouvée est alors supérieure à D_n de la table II on peut sûrement conclure que l'écart est significatif et rejeter la distribution théorique soumise au test.

F. J Massey a vérifié expérimentalement cette propriété [4]. Sur 100 échantillons de 10 unités extraits d'une population normale, dont les paramètres sont estimés d'après les échantillons, il a trouvé : le 95è percentile de D_{10} à 0,29 au lieu de 0,41 (de la table II) et le 90è percentile à 0,25 au lieu de 0,37 (de la table II) et aucune observation au-dessus de la valeur 0,37.

VI - JUGEMENT SUR LA VALIDITE D'UNE FONCTION DE DISTRIBUTION THEORIQUE FONDE SUR LA LOI BINOMIALE -

1 - Principe.

Connaissant la fonction de distribution F(x) d'une variable aléatoire X, la probabilité d'observer une valeur inférieure ou égale à x est F(x). Lorsque l'épreuve est répétée n fois la proportion des résultats inférieurs ou égaux à x, soit $S_n(x)$, est une variable aléatoire qui suit une loi binomiale B [n, F(x)]. Nous connaissons donc les limites des fluctuations aléatoires de $S_n(x)$ à un seuil de confiance fixé à l'avance. Pour toute valeur de x, indépendamment des conclusions en d'autres points de la distribution, $S_n(x)$ doit donc demeurer entre les limites ci-dessus. Dans cette éventualité, et dans cette éventualité seulement, nous dirons que F(x) est compatible avec les données de l'échantillon.

2 - Application.

Le test binomial étant indépendant de F(x) nous prendrons pour celle-ci F(x) = x avec $0 \leqslant x \leqslant 1$.

Dans les figures 5, 6 et 7 nous représentons les limites de confiance correspondant à la fonction de distribution F(x), au seuil de probabilité 0,95 respectivement pour n = 5, 20 et 100. Les limites K du test de Kolmogorov-Smirnov sont tirées de la table II, les limites B du test binomial sont lues sur les abaques de la distribution binomiale.(1)

Nous constatons que l'intervalle de confiance du test binomial B n'est pas uniforme comme pour le test, K de Kolmogorov - Smirnov, que les premières limites sont plus serrées que les secondes et que lorsque n croît les limites B se rapprochent des limites K.

Notons que le test binomial s'applique que la variable aléatoire X soit continue ou discrète.

D'autre part, l'intervalle de confiance maximal se produit pour F(x) = 0,50, on pourrait donc, dans un but de simplification, admettre que cet intervalle est constant au détriment de la sévérité du test qui demeure néanmoins plus sévère que le test de Kolmogorov.

3 - Applications numériques.

Nous reprenons les exemples numériques du § V auxquels nous appliquons le test binomial.

----

(1) Tables statistiques du Centre de Formation aux Applications Industrielles de la Statistique - (Institut de Statistique de l'Université de Paris).

Comparaison des limites de
Confiance de Kolmogorov K
et Binomiale B

n = 5 , Seuil de confiance = 0,95

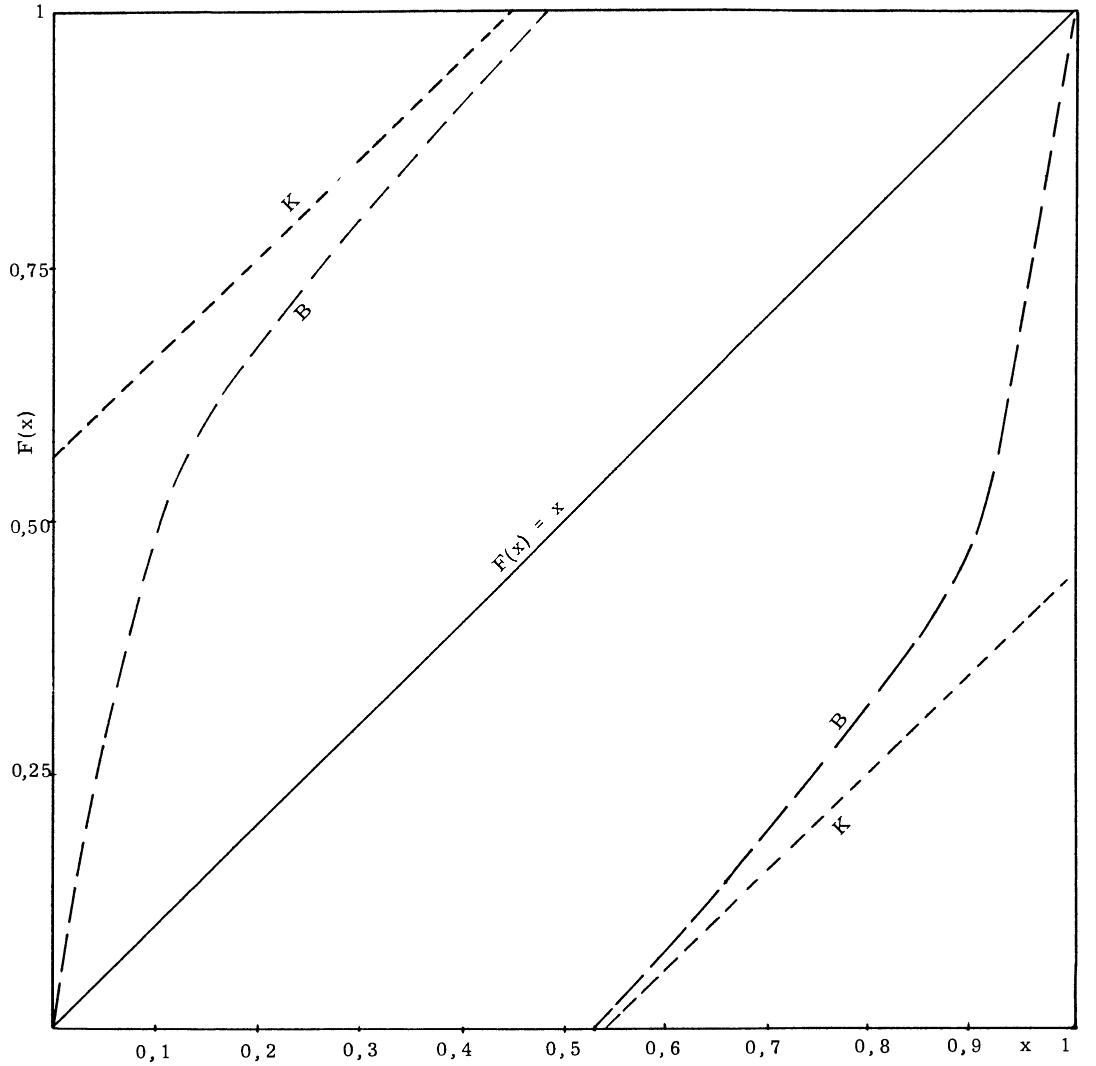

Figure 5

n = 20 , Seuil de confiance = 0,95

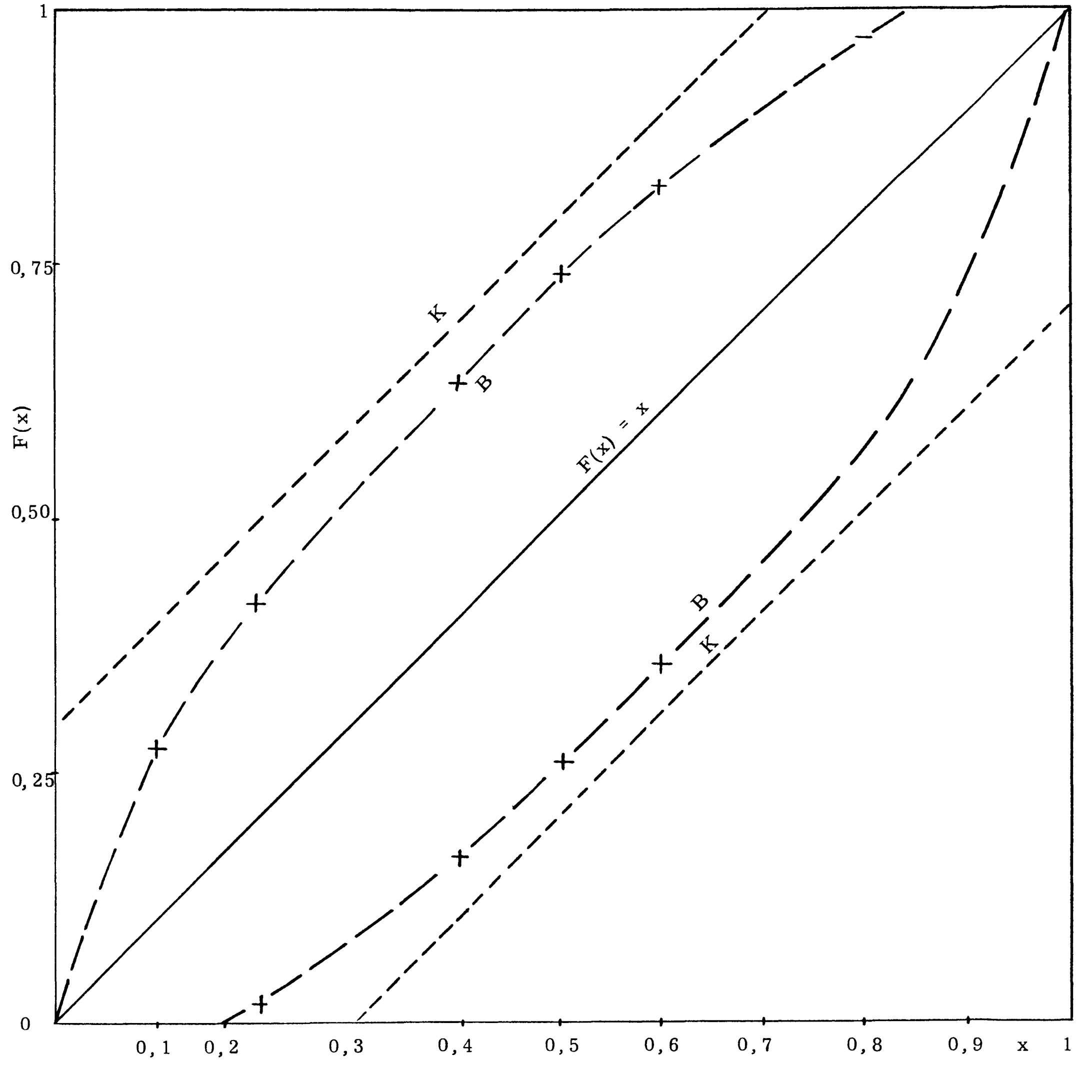

Figure 6

n = 100, Seuil de confiance = 0,95

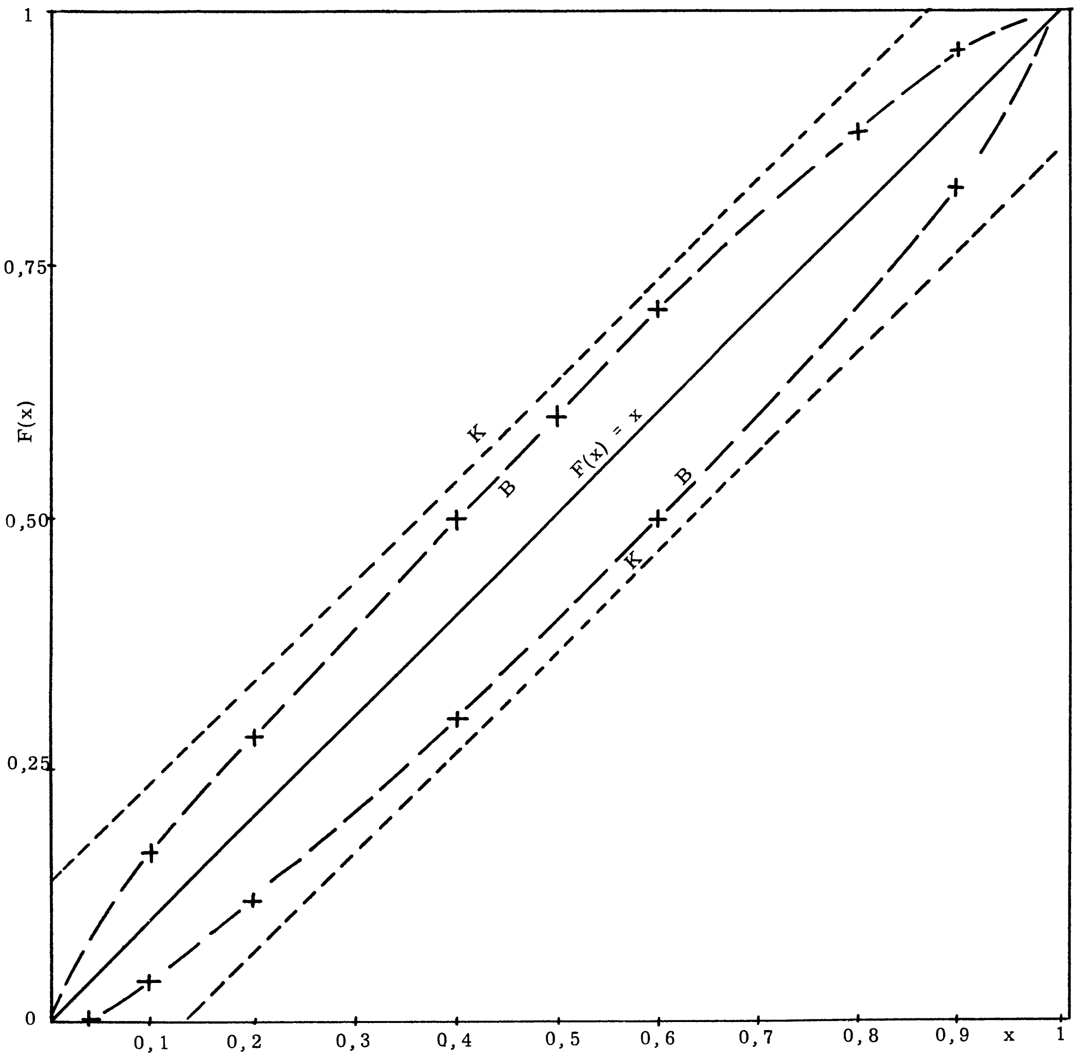

Figure 7

- Pour la durée de vie d'un matériel nous allons vérifier la validité de $F(x) = 1 - e^{-x/170}$ pour x = 99. On a :

$$F(99) = 0,442,$$

$$S_{60}(99) = 0,283.$$

Les limites de confiance de la loi binomiale B(60 ; 0,442) au seuil 0,90 sont : 0,33 - 0,56 et $S_{60}(99)$ est hors de ces limites, nous rejetons donc la fonction de distribution exponentielle ci-dessus.

- Appliquons le test à la distribution théorique normale N(170, 110) au point d'abscisse x = 99. On a : F(99) = 0,26 ; les limites de confiance de la loi binomiale B(60 ; 0,26) au seuil 0,90 sont 0,16 - 0,32. Cet intervalle contenant la fréquence dans l'échantillon 0,283, nous acceptons l'hypothèse d'une distribution normale avec un risque de 1ère espèce $\alpha = 0,10$.

- Dans l'exemple donné au tableau 2 nous testons le gros écart 12,41 entre les distributions théorique et observée par le test binomial $B(511 ; 0,56)(\frac{287,59}{511} = 0,56)$.

L'échantillon est de grande taille, la probabilité théorique n'est pas voisine de zéro, nous admettons donc pour les limites de confiance de la loi binomiale au seuil 0,95 les limites à 2 écarts-types autour de 0,56. On a :

$$\sigma_s = \sqrt{\frac{0,56 \times 0,44}{511}} \sim 0,022,$$

et l'écart maximal qu'on peut permettre à S_{511} est $0,022 \times 2 = 0,044$, d'où l'on conclut en faveur de l'adoption de la fonction de distribution normale pour la population-mère.

Dans les trois exemples numériques que nous avons examinés on aboutit donc aux mêmes conclusions par l'application soit du test de Kolmogorov soit du test binomial.

REFERENCES

[1] FELLER, W - "On the Kolmogorov-Smirnov limit theorems for empirical Distributions" - Annals of Mathematical Statistics - 19 - 1948 - pp 177 - 189.

[2] FOSTER, J.A. - "Kolmogorov-Smirnov test for goodness of fit : What it is - How to apply it" - Industrial Quality Control - Vol 18 - N° 7 - 1962 - pp 4-8.

[3] MASSEY, F.J., Jr - "A note on the estimation of distribution function by confidence limits" - Annals of Mathematical Statistics - 21 - 1950 - pp 116 - 119.

[4] MASSEY, F. J., Jr - "The Kolmogorov-Smirnov test for goodness of fit" - Journal of the American Statistical Association - 253 - Mars 1951 - pp 68 - 78.

[5] MILLER, L.H. - "Table fo percentage points of Kolmogorov-statistics" - Journal of the American Statistical Association - 273 - Mars 1956 - pp 111 - 121.

[6] SMIRNOV, N. - "Table for estimating the goordness of fit of empirical distributions" - Annals of Mathematical Statistics - 19 - 1948 - pp 279 - 281.

[7] E. de WILDE - "Application du test de Kolmogorov-Smirnov dans la pratique industrielle. Congrès international. Organisation Européenne pour le contrôle de qualité. Bruxelles 1959.