REVUE DE STATISTIQUE APPLIQUÉE

M. D. INDJOUDJIAN

Notes sur les propriétés des fonctions de répartition empiriques

Revue de statistique appliquée, tome 18, nº 2 (1970), p. 5-29
<http://www.numdam.org/item?id=RSA_1970__18_2_5_0>

NOTES SUR LES PROPRIÉTÉS DES FONCTIONS DE RÉPARTITION EMPIRIQUES

M.D. INDJOUDJIAN

PLAN DE CES NOTES

Les propriétés des fonctions de répartition empiriques ont fait l'objet de très nombreuses études depuis une trentaine d'années. Les applications théoriques et pratiques en sont importantes.

Les notes qui suivent rassemblent en les ordonnant un grand nombre des résultats obtenus durant cette période et par des méthodes très diverses ; elles contiennent des références bibliographiques précises. L'annexe donne certaines des propriétés des "écarts" ω_m^2 (de Cramér - von Mises), W_m^2 (d'Anderson - Darling) et de certains écarts plus généraux entre fonction de répartition théorique et fonction de répartition empirique. Contrairement aux notes elles-mêmes, cette annexe fournit certaines démonstrations.

Diverses questions non examinées dans ces notes ont fait l'objet de travaux plus ou moins complets, par exemple les suivantes au sujet desquelles quelques indications bibliographiques sont données :

- Problème de k échantillons ($k > 2$)
(cf. [47] - [52 bis] - [57], etc...)

- Population-mère à deux dimensions (ou davantage)
(cf. [38],...)

- Cas où la fonction de répartition théorique n'est pas entièrement spécifiée. (Problèmes de test ou problèmes d'estimation).
(cf. [11],...)

- Cas où la fonction de répartition théorique n'est pas continue
(cf. [43] - [59] - [61],...

- Tests d'indépendance utilisant les fonctions de répartition empiriques à une et à deux dimensions.
(cf. [41],...).

- Etant donné la propriété élémentaire fondamentale

$$V\{F_m(x)\} = \frac{F(x)\,[1 - F(x)]}{m}$$

il devrait être intégressant - à notre connaissance, cela n'a pas été fait - de considérer la variate :

$$\Delta_m^*(\alpha, \beta) = \sup_{o<\alpha<x<\beta<1} \frac{|F_m(x) - F(x)|}{\sqrt{F(x)\,[1\ F(x)]}}$$

et d'étudier la loi (exacte pour m fini, ou limite pour $m \longrightarrow +\infty$ de

$$\sqrt{m}\,\Delta_m^*(\alpha, \beta)$$

1 - NOTATIONS PRINCIPALES

1.1

- Population-mère de fonction de répartition continue (sauf en 2.1) : F(x).

- Echantillon de m observations faites indépendamment sur cette population-mère : $x_1, \dots, x_m$. Ordonnées par valeurs croissantes : $x_1^*, \dots, x_m^*$.

- Autre échantillon indépendant, de taille n : mêmes notations, y remplaçant x.

- Fonction de répartition empirique $F_m(x)$ [resp. $G_n(x)$]: quotient par m [resp. n] du nombre h [resp. k] des observations $x_1, \dots, x_m$ (resp. y_1, $\dots y_n$) inférieures ou égales à x. En d'autres termes, pour $F_m(x)$ par exemple :

$$F_m(x) = \begin{cases} 0 & \text{si } x \leqslant x_1^* \\ \dfrac{h}{m} & \text{si } x_h^* < x \leqslant x_{h+1}^* \quad (h = 1, 2, \dots, m-1 \\ 1 & \text{si } x_m^* < x \end{cases}$$

1.2

- On peut définir bien des "écarts" ou "distances" entre la fonction de répartition empirique et la fonction de répartition théorique (ou, 1.3, entre deux fonctions de répartition empiriques) :

$$D_m^+ = \sup_{o<x<+\infty} [F_m(x) - F(x)] \quad , \quad D_m^- = \sup\,[F(x) - F_m(x)]\ ,$$

$$D_m = \sup_x |F_m(x) - F(x)| = \operatorname{Max}\{D_m^+, D_m^-\}$$

α, β et γ étant des paramètres ($0 < \alpha \leq \beta \leq 1$) et $\gamma \geq 1$) :

$$D_m^+(\alpha, \beta; \gamma) = \sup_{\alpha<F(x)<\beta} [F_m(x) - \gamma\, F(x)] \quad , \quad \text{généralisant } D_m^+ = D_m(0, 0\ ; 1)$$

$$\rho_m(\alpha,\beta\,;\gamma) = \sup_{\alpha<F(x)<\beta} \frac{F_m(x) - \gamma F(x)}{F(x)}\ ,\ \rho_m^+(\alpha) = \rho_m^+(\alpha, 1\,.\,1)\ ,\ \rho_m^+(\alpha,\beta) = \rho_m^+(\alpha,\beta\,;\,1)$$

$$\rho_m(\alpha,\beta\,;\gamma) = \sup_{\alpha<F(x)<\beta} \left| \frac{F_m(x) - \gamma F(x)}{F(x)} \right| \quad , \quad \rho_m(\alpha) = \rho_m(\alpha\,,\,1\,;1)\,.$$

$$\omega_m^2 = \int_{-\infty}^{+\infty} [F_m(x) - F(x)]^2\, dF(x) \quad , \quad W_m^2 = m \int \frac{[F_m(x) - F(x)]^2}{F(x)\,[1 - F(x)]}\, dF(x)$$

(cf. définition généralisant celles de ω_m^2 et de W_m^2 à l'Annexe).

Les écarts D ont été introduits par Kolmogorov, Smirnov et leurs élèves ; les écarts ρ par Renyi, ω_m^2 par Cramér et von Mises et W_m^2 par Anderson et Darling.

1.3

$$D_{m,n}^+ = \sup_{-\infty<x<+\infty} [F_m(x) - G_n(x)] \quad , \quad \Delta_n^+ = D_{n,m}^+$$

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)| \quad , \quad \Delta_n = D_{n,n}$$

2 - PROPRIETES DES FONCTIONS DE REPARTITION EMPIRIQUES

2.1

Quelle que soit la fonction de répartition théorique F(x) :

(a) $F_m(x) \xrightarrow{\mathcal{P}} F(x)$: loi <u>faible</u> des grands nombres — pour tout x

(b) $F_m(x) \xrightarrow{p.s.} F(x)$: loi <u>forte</u> des grands nombres — pour tout x

(c) $D_m \xrightarrow{p.s.} 0$: <u>théorème de Glivenko</u> (1933) : la convergence de la fonction de répartition empirique vers la fonction de répartition théorique est presque sûrement uniforme .

2.2

Les propriétés rappelées en 2.1 sont considérablement précisées et renforcées par les suivantes, quand F(x) est <u>continue</u>.

2.2.1

Les écarts définis en 1.2 et 1.3 suivent des lois de probabilité invariantes par rapport à toute transformation continue de la population-mère, c'est pourquoi on pourra supposer pour les démonstrations que X est distribuée uniformément sur [0, 1], c'est-à-dire que F(x) = x.

2.2.2

(a) $$\boxed{\lim_{m\to+\infty} \mathcal{P}r\,\{\sqrt{m}\, D_m \leqslant z\} = \lim_{m\to-\infty} \mathcal{P}r\,\{\sqrt{m}\, D_m \leqslant z\} = 1 - e^{-2z^2}}$$

Remarque : $4\ m(D_m^+)^2$ et $4\ m(D_m^-)^2$ suivent chacune asymptotiquement une loi du χ^2 à deux degrés de liberté.

$$\text{(b)} \quad \boxed{\lim_{m \to +\infty} \mathscr{P}r\ [\sqrt{m}\ D_m \leq z] = K(z)}$$

où la fonction de Kolmogorov :

$$K(z) = \begin{cases} 0 \ , & \text{si } z \leq 0 \\ \boxed{\sum_{j=-\infty}^{+\infty} (1)^j\ e^{-2j^2z^2}} & \text{si } z > 0 \end{cases}$$

$$\text{(c)} \quad \boxed{\lim_{m \to +\infty} \mathscr{P}r\ \{\sqrt{m}\ \rho_m^+(\alpha) \leq z\} = 2\ G\left(\sqrt{\frac{\alpha}{1-\alpha}}\ z\right) - 1} \ , \quad \text{si } z > 0$$

$$\text{"} \qquad = 0 \qquad \text{si } z \leq 0$$

où $0 < \alpha < 1$, $G(u)$ désignant la fonction de répartition de la loi de Laplace-Gauss réduite :

$$G(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} e^{-\frac{x^2}{2}}\ dx$$

$$\text{(d)} \quad \boxed{\lim_{m \to +\infty} \mathscr{P}r\ \{\sqrt{m}\ \rho_m(\alpha) < z\} = L\left(\sqrt{\frac{\alpha}{1-\alpha}}\ z\right)}$$

où la fonction $L(u)$, nulle pour $u < 0$, a, pour $u > 0$, l'expression

$$\boxed{L(u) = \frac{4}{\pi} \sum_{j=0}^{+\infty} (-1)^j\ \frac{e^{-\frac{(2j+1)^2\pi^2}{8u^2}}}{2j+1}}$$

égale, d'ailleurs, à

$$\sum_{r=-\infty}^{+\infty} (-1)^r\ \{G[(2r+1)u] - G[(2r-1)u]\}$$

(e) L'expression compliquée, donnée par Renyi [32] de $\lim_{m \to +\infty} \mathscr{P}r\ \{\sqrt{m}\ \rho_m(\alpha, \beta) < z\}$ pour $0 < \alpha < \beta < 1$, se simplifie pour $z = 0$ et l'on obtient

$$\boxed{\lim_{m \to +\infty} \mathscr{P}r\ \left\{\sup_{\alpha < F(x) < \beta}\ [F_m(x) - F(x)] < 0\right\} = \frac{1}{\pi}\ \text{Arcsin}\ \sqrt{\frac{\alpha(1-\beta)}{\beta(1-\alpha)}}}$$

Ainsi, la probabilité que la fonction de répartition empirique reste partout en-dessous de la fonction de répartition théorique tend vers zéro ssi $\alpha = 0$ et ou $\beta = 1$. La limite de cette probabilité est strictement positive si l'écart est considéré dans un intervalle $[\alpha, \beta]$ tel que $0 < \alpha < \beta < 1$.

(f) Les lois de $\sqrt{N}\,D^+_{m,n}$, $\sqrt{N}\,D^-_{m,n}$ et $\sqrt{N}\,D_{m,n}$ -où $N = \dfrac{mn}{m+n}$ - tendent, lorsque $m \longrightarrow +\infty$ et $n \longrightarrow +\infty$, m/n tendant vers une limite finie non nulle, vers les lois-limites de $\sqrt{m}\,D_m$, $\sqrt{m}\,D_m$ et $\sqrt{m}\,D_m$.

En particulier, si l'on considère deux échantillons de même taille n :

$$\lim_{n\to+\infty} \mathscr{P}r\left\{\sqrt{\frac{n}{2}}\,\Delta_n^+ \leqslant z\right\} = \lim_{n\to+\infty} \mathscr{P}r\left\{\sqrt{\frac{n}{2}}\,\Delta_n^- \leqslant z\right\} = 1 - e^{-2z^2}$$

(pour $z > 0$ - et 0 pour $z \leq 0$)

$$\text{(g)}\quad \boxed{\lim_{n\to+\infty} \mathscr{P}r\left\{\sqrt{\frac{n}{2}}\,\Delta_n^- \leq u, \sqrt{\frac{n}{2}}\,\Delta_n^+ \leq v\right\} = \sum_{r=-\infty}^{+\infty} e^{-2r^2(u+v)^2} - \sum_{r=1}^{+\infty}\left[e^{-2[ru+(r-1)v]^2} + e^{-2[(r-1)u+rv]^2}\right]}$$

pour $\min\{u, v\} > 0$ - et 0 pour $\min\{u, v\} \leqslant 0$

2.2.3

$$\text{(a)}\quad \boxed{\mathscr{P}r\left\{D_n^+ \leqslant \frac{j}{n}\right\} = \mathscr{P}r\left\{D_n^- \leqslant \frac{j}{n}\right\} = 1 - \sum_{s=1}^{n-j} \frac{1}{n-s}\, C_n^{j+s} \left(\frac{s}{n}\right)^{j+s}\left(1 - \frac{s}{n}\right)^{n-j-s}}$$

Pour $0 < x \leqslant 1$, $\mathscr{P}r\{n\,D_n^+ \leqslant x\} = \mathscr{P}r\{n\,D_n^- \leqslant x\}$ a l'expression précédente avec $j = [nx]$ (cf. L. Takacs [36].

(b) L'expression, assez compliquée, de la loi de $\rho_m^+(\alpha, \beta\ ; \gamma)$ est donnée dans L. Takacs [36].

(c) Pour $\dfrac{n}{m} = p$ entier positif,

$$\mathscr{P}r\left\{D^+_{m,n} < \frac{i}{n}\right\} = 1 - \frac{1}{C^n_{m+n}} \sum_{\frac{i}{p} < s < m} \frac{i}{n+i-sp}\, C^s_{s(p+1)-i}\, C^{m-s}_{m+n+i-1-s(p+1)}$$

(cf. V.S. Korolyuk [22]).

(d) Dans le cas particulier de deux échantillons de même taille m (p=1)

$$\boxed{\mathscr{P}r\left\{\Delta_m \quad \frac{i}{m}\right\} = 1 - \frac{C_{2m}^{m-i}}{C_{2m}^{m}}}$$

expression particulièrement simple (cf. Gnedenko et Korolyuk [16]) qui peut également s'écrire sous la forme

$$\Phi_m(z) = \mathscr{P}r\left\{\sqrt{\frac{m}{2}}\,\Delta_m^+ < z\right\} = \begin{cases} 0 \text{ pour } z < \dfrac{1}{\sqrt{2m}} \\ \text{l'expression précédente pour } i = [z\sqrt{2m}] \\ \qquad \text{et } \dfrac{1}{\sqrt{2m}} < z \leq \sqrt{\dfrac{m}{2}}. \\ 1 \text{ pour } z < \sqrt{\dfrac{m}{2}} \end{cases}$$

En outre V.V. Gnedenko [15] a montré qu'en posant

$$0 < z\sqrt{2m} - [z\sqrt{2m}] = \theta \leq 1,$$

$$\Phi_m^+(z) = 1 - e^{-2z^2}\left[1 + 2\sqrt{\frac{2}{m}}\theta z + \frac{1+4\theta^2}{m}z^2 - \frac{2}{zm}z^4 + \sigma\left(\frac{1}{n^2}\right) z^2\right]$$

Ainsi, l'erreur relative commise en assimilant $\Phi_m^+(z)$ à sa limite $1 - e^{-2z^2}$ est, pour $z \ll \sqrt{\frac{m}{2}}$ (c'est-à-dire $i \ll m$), approximativement égale à $\frac{2\theta}{e^{2z^2}-1} \cdot \frac{z}{\sqrt{\frac{m}{2}}}$.

(e) $$\boxed{\mathcal{P}r\left\{\Delta_m < \frac{i}{m}\right\} = \frac{1}{C_{2m}^m} \sum_{s=-\left[\frac{m}{i}\right]}^{\left[\frac{m}{i}\right]} (-1)^s \; C_{2m}^{m-si}}$$

(cf. Takacs 36 et Gnedenko-Korolyuk [16])

On écrit parfois ce résultat sous la forme suivante :

$$\Phi_m(z) = \mathcal{P}r\left\{\sqrt{\frac{m}{2}}\Delta_m < z\right\} = \begin{cases} 0 \text{ pour } z < \frac{1}{\sqrt{2m}} \\ \text{l'expression précédente pour } i = [z\sqrt{2}m] \\ \qquad \text{et } \frac{1}{\sqrt{2m}} z < \sqrt{\frac{m}{2}}. \\ 1 \text{ pour } z > \sqrt{\frac{m}{2}} \end{cases}$$

(f) $$\boxed{\mathcal{P}r\left\{\sqrt{\frac{n}{2}}\Delta_n^- < u, \sqrt{\frac{n}{2}}\Delta_n^+ < v\right\}}$$

est égale à 0, si

$$\min\{u, v\} \leq \frac{1}{\sqrt{2n}},$$

à 1, si

$$\min\{u, v\} > \sqrt{\frac{n}{2}}$$

et, si

$$\frac{1}{\sqrt{2n}} < \min\{u, v\} < \sqrt{\frac{n}{2}},$$

à l'expression suivante -où,

$$\alpha = [u\sqrt{2n}] \;, \quad \beta = [v\sqrt{2}n], s_0 = \left[\frac{n}{\alpha+\beta}\right], \quad s_1 = \left[\frac{n+\alpha}{\alpha+\beta}\right] \quad \text{et} \quad s_2 = \left[\frac{n+\beta}{\alpha+\beta}\right]$$

$$\boxed{\frac{1}{C_{2n}^{n}}\left(\sum_{s=-s_0}^{s_0} C_{2n}^{n-s(\alpha+\beta)} - \sum_{s=1}^{s_1} C_{2n}^{n+\alpha-s(\alpha+\beta)} - \sum_{s=1}^{s_2} C_{2n}^{n+\beta-s(\alpha+\beta)}\right)}$$

(cf. Gnedenko et Rvačeva [19])

(g) On démontre (cf. Annexe) que :

$$m\omega_m^2 = \frac{1}{12\,m} + \sum_{j=1}^{m} \left[F(x_j^*) - \frac{2\,j-1}{2\,m} \right]^2$$

et

$$W_m^2 = -\,m - \frac{1}{m} \sum_{j=1}^{m} \{(2\,j-1) \log F(x_j^*) + (2\,m-2\,j+1) \log[1 - F(x_j^*)]\}$$

et que

$$\boxed{E\,\{m\omega_m^2\} = \frac{1}{6}} \quad , \quad \boxed{V\,\{m\omega_m^2\} = \frac{4\,m-3}{180\,m} \longrightarrow \frac{1}{45}} \quad \text{et} \quad \boxed{E\,\{W_m^2\} = 1} \quad ;$$

mais les lois-limites de $m\,\omega_m^2$ et de W_m^2 ne sont pas normales. Toutefois la convergence en loi est rapide.

La première loi-limite a été établie par Smirnov [34], puis, plus explicitement, par Anderson et Darling [2-3] qui l'ont tabulée (*).

Les lois exactes ont été étudiées, notamment celle de ω_m^2 par Marshall [24] et par Pearson et Stephens [31], celle de W_m^2 par Lewis [23].

Anticipant sur le § 2.4, remarquons que le critère d'Anderson-Darling présente par rapport à ceux utilisant les écarts D_m ou ω_m^2 (respectivement de Kolmogorov-Smirnov et de Cramér-von Mises) l'avantage de ne pas être insensibles dans les régions de faible ou de forte probabilité ($F(x)$ très petit ou voisin de l'unité).

On montrera dans la même Annexe que ω_m^2 et $\frac{W_m^2}{m}$ sont des cas particuliers d'un "écart" ayant une définition plus large.

2.2.4

(a) La variate définie en 1.2, D_n^+ est égale à $\underset{1<h<n}{\text{Max}}\; D_n^+(h)$ ou $D_n^+(h) = F_n(x_h^*) - F(x_h^*)$ $(h = 1, 2, \ldots, n)$.

Pour l'étude de $D_n^+(h)$, on peut supposer $F(x) = x$ (et $0 < x < 1$), auquel cas $D_n^+(h) = \frac{h}{n} - x_h^*$.

- Désignons par r_n le nombre d'éléments non-négatifs parmi les $D_n^+(h)$ et par $\hat{r}_n$ la valeur de h, définie avec une probabilité égale à un, pour laquelle $D_n^+(h) = D_n^+$.

(*) On trouvera au § 2.5 les quantiles correspondants aux probabilités 0,95 et 0,99 des lois-limites de $\sqrt{m}\, D_m^+$ (ou $\sqrt{m}\, D_m^-$), $\sqrt{m}\, D_m$, $m\omega_m^2$ et W_m^2.

Théorème A

$$\boxed{\mathcal{P}r\ \{r_n = r\} = \mathcal{P}r\ \{\hat{r}_n = r\} = \frac{1}{n}\sum_{i=1}^{r}\frac{1}{i}\ C_n^{i-1}\left(\frac{i}{n}\right)^{i-1}\left(1-\frac{i}{n}\right)^{n-i}}$$

(cf. pour la première et la seconde loi respectivement : P. Cheng [8] et Z. Birnbaum and R. Pyke [5] qui ont établi également le résultat B suivant :

Théorème B

$$\boxed{\left(\frac{\hat{r}_n}{n} - D_n^+\right) \text{ suit une loi uniforme sur } [0,1]}$$

- L_n étant définie comme la longueur totale de la projection sur l'axe vertical des parties de la courbe représentative de la fonction de répartition théorique située au-dessus de la fonction de répartition empirique,

Théorème C

$$\boxed{\mathcal{P}r\left\{L_n = \frac{i}{n}\right\} = \frac{1}{n+1} \quad (i = 0,\ 1,\ \ldots,\ n)}$$

la loi de L_n est uniforme. cf. Gnedenko et Mihalevič [18]

(b) Posons :

$$N = \frac{m\,n}{m+n} \quad \text{et} \quad T_{m,n}(x,\ z) = G_n(x) + \frac{z}{\sqrt{N}} \quad , \quad \text{pour } z \geq 0$$

Nous considérerons que $F_m(x)$ coupe la courbe $T_{m,n}(x,\ z)$ qui correspond au paramètre z au point x_h si $F_m(x_h) < T_{m,n}(x\ ,\ z) < F_m(x_h + 0)$. Soit $\nu(z\ ;\ m,\ n)$ le nombre d'abscisses parmi $x_1, \ldots, x_m$ où une telle intersection se produit.

Alors, si $\frac{m}{n} = c^{te}$,

$$\boxed{\lim_{n\to+\infty} \mathcal{P}r\ \{\nu(z\ ;\ m,\ n) \leq 2\,t\sqrt{N}\}} = \begin{cases} 0 \text{ pour } t < 0 \\ \boxed{1 - e^{-2(z+t)^2}} \text{ pour } t \geqslant 0 \end{cases}$$

(cf. Smirnov [33]).

Corollaire (pour t = 0). On retrouve que, pour

$$z \geq 0, \ \lim_{n\to+\infty} \mathcal{P}r\ \{\sqrt{N}\ D_{m,n}^+ \leq z\} = 1 - e^{-2z^2}$$

- Dans le cas ou m = n, posons

$$\nu_n(z) = \nu(z\ ;\ n,\ n)$$

et soient h et k deux entiers non négatifs égaux à $[t\sqrt{2n}]$ et $[z\sqrt{2n}]$ respectivement :

$$\boxed{\mathcal{P}r\{\nu_n(z) < t\sqrt{2n}\}} = \begin{cases} 0 \ , \ \text{si } h < 0 \\ \boxed{1 - \dfrac{C_{2n}^{n-(h+k)}}{C_{2n}^{n}}} \quad \text{si } 0 < h \leq n-k. \\ 1 \ , \ \text{si } h + k > n \end{cases}$$

Ce résultat de Mihalevič [28] est équivalent au résultat de Gnedenko et Korolyuk [16], car Mihalevič démontre que

$$\mathcal{P}r\{\nu_n(z) < t\sqrt{2n}\} = \mathcal{P}r\{n\,\Delta_n^{+} < h + k\}$$

et, pour cela, que

$$\mathcal{P}r\{\nu_n(k) = h\} = \mathcal{P}r\{n\,\Delta_n^{+} = h + k\} \quad \text{pour } k > 0$$

- Nous définissons, lorsque $m = n$, le nombre de dépassements $C_n(z)$ de la courbe $T_{m,n}(x, z)$ par $F_m(x)$ comme le nombre des abscisses $x_1, x_2, \ldots x_m$ pour lesquelles $T_{m,n}(x_h, z) < F_n(x_h + 0)$

En particulier, $C_n(0)$ est le nombre des dépassements de $G_n(x)$ par $F_n(x)$:

Théorème :

$$\boxed{\mathcal{P}r\left\{C_n(0) = \frac{i}{n}\right\} = \frac{1}{n+1}} \quad i = 0, 1, \ldots, n)$$

La loi de $C_n(0)$ est uniforme

(cf. Gnedenko et Mihalevič [17], où se trouve aussi établi un résultat plus général correspondant à $m/n = p$ entier > 1).

Mihalevič [28] a également établi l'expression de la loi de $\mathcal{P}r\{C_n(z) < nt\}$.

2.2.5

Certaines des propriétés précédentes ont été étendues au cas où la taille m de l'échantillon est aléatoire et suit une loi de Poisson de paramètre λ :

$$\mathcal{P}r\{m = j\} = e^{-\lambda}\frac{\lambda^j}{j!} \quad (j = 0, 1, 2, \ldots)$$

et ce indépendamment de x. (cf. L. Takacs [36]).

En particulier M. Kac [20] a montré que, en posant :

$$\gamma(\lambda) = \sup_{-\infty < x < \infty} \left|\frac{m}{\lambda} F_m(x) - F(x),\right|$$

$$\boxed{\lim_{\lambda \to +\infty} \mathcal{P}r\{\sqrt{\lambda}\,\gamma(\lambda) \leq z\} = L(z)}$$

où L(z) est la fonction qui a été définie au § 2.2.2.d.

2.3

C'est la recherche de tests de l'hypothèse relative à la loi de la population-mère dont un échantillon est observé (ou de l'identité des populations-mères dont deux échantillons respectifs sont observés) - cf. § 2.4 ci-dessous - qui est à l'origine des études dont un certain nombre de résultats ont été rassemblés ci-dessus.

Lors des premiers travaux (Cramer, von Mises concernant ω_m^2 ou des variantes de ce critère ; Kolmogorov, Smirnov et leurs élèves concernant D_m, etc...), les méthodes étaient généralement compliquées et aucun principe fondamental ne s'en dégageait. L'accent était mis dans la plupart des cas sur la recherche de la loi-limite.

L'évolution des quelque vingt dernières années - mises à part les propriétés comparées des divers tests correspondants entre elles ou avec celles du test du χ^2 (cf. § 2.2.4) - a été marquée par une double tendance :

a) L'une vers la mise en évidence par des moyens aussi directs et aussi simples que possible des lois exactes des "écarts" pour des tailles finies d'échantillons. [cf. par exemple . [16] - [19] - [36] § § 39, 40 et 41, (probl. 11)].

Il est apparu de plus en plus nettement que les méthodes les plus profondes - et probablement les plus simples - étaient à cet égard de nature combinatoire, méthodes ayant d'ailleurs une portée beaucoup plus large - ce qui explique qu'elles soient disséminées dans le littérature scientifique et ne soient pas exposées de la manière systématique et élémentaire qui conviendrait à l'établissement de la plupart des résultats indiqués ci-dessus pour des tailles finies d'échantillons. Parmi les travaux les plus représentatifs de cette tendance, voir, outre ceux cités quelques lignes plus haut : [1] - [12] - [13] - [14] - [32] et [35].

Il est intéressant de noter qu'à partir de ces lois exactes pour des échantillons de taille finie, on peut souvent obtenir de façon claire les lois-limites - et même avec parfois l'avantage de faire apparaitre le degré d'approximation (cf. par exemple § 2.2.3.d. et [15]).

A titre d'exemple, montrons rapidement comment obtenir :

$$\lim_{m \to +\infty} \mathcal{P}r\left\{\sqrt{\frac{m}{2}}\,\Delta_m^+ < z\right\}$$

en partant de l'expression exacte (§ 2.2.3.d.)

$$\mathcal{P}r\left\{\sqrt{\frac{m}{2}}\,\Delta_m^+ < z\right\} = 1 - \frac{C_{2m}^{m-i}}{C_{2m}^{m}} \quad \text{pour } i = [z\sqrt{2m}] \quad \text{et} \quad \frac{1}{\sqrt{2m}} < z < \sqrt{\frac{m}{2}}$$

Pour cela, remarquons que

$$A_{m,i} = \frac{C_{2m}^{m}}{C_{2m}^{m-i}} = \frac{(m+i)\,!\;(m-i)\,!\;(2m)\,!}{(2m)\,!\;(m\,!)^2} \simeq \frac{e^{-(m+i)}\;(m+i)^{m+i}\;e^{-(m-i}\;(m-i)^{m-i}}{e^{-2m}\;m^{2m}},$$

d'après la formule de Stirling ; soit

$$A_{m,i} \sim \left(1+\frac{i}{m}\right)^{m+i} \left(1-\frac{i}{m}\right)^{m-i} \text{ et } \log A_{m,i} \sim (m+i)\left(\frac{i}{m}-\frac{i^2}{2m^2}+..\right)+(m-i)\left(-\frac{i}{m}-\frac{i^2}{2m^2}\right)$$

c'est-à-dire :

$$A_{m,i} \simeq \frac{z^2 \times 2m}{m} = 2z^2$$

Il en résulte que la limite de la loi cherchée est bien : $1 - e^{-2z^2}$.

b) l'autre tendance vers la recherche directe des lois-limites. A cet égard, les principes généraux se dégagent grâce à l'introduction explicite de processus stochastiques convenables. Après la lecture de l'article de Darling [9], qui donne une synthèse intéressante de travaux antérieurs à 1957, on lira l'article de Doob [10] qui a contribué à ouvrir cette nouvelle voie, ainsi que d'autres articles cités dans [9] ou plus récents mais non cités ici pour ne pas allonger exagénément la liste bibliographique. Il a été en particulier mis en évidence par Kac, Dempster, Donsker,... que les processus stochastiques s'introduisant de manière naturelle dans ces questions, par exemple dans l'étude du m ω_m^2 de Cramer-von Mises, ont des propriétés liées étroitement à celles de certaines équations intégrales dont le noyau est caractéristique du processus correspondant. C'est ainsi que la loi-limite de m ω_m^2 apparaît, par application du théorème de Mercer (de la théorie classique des équations intégrales à noyau positif, symétrique et continu), comme celle d'une variate égale à $\sum_{j=1}^{+\infty} \frac{Z_j^2}{j^2\pi^2}$, où $Z_1, \ldots, Z_j, \ldots$ est une suite infinie de variates normales réduites indépendantes. (On remarquera au passage la correspondance entre le fait que $\lim_{m\to+\infty} E\{m\,\omega_m^2\} = \frac{1}{6}$ et le fait que $\sum_{j=1}^{+\infty} \frac{1}{j^2} = \frac{\pi^2}{6}$).

2.4

Il est clair que les "écarts", parfois appelés critères, de Cramer-von Mises, de Kolmogorov-Smirnov, d'Anderson-Darling, etc... se prêtent particulièrement bien à tester (d'une manière bilatérale ou unilatérale) l'hypothèse selon laquelle une population-mère dont on observe un échantillon de taille m a une fonction de répartition (théorique) donnée, c'est-à-dire complètement spécifiée. Ces tests d'adéquation ont le double avantage d'avoir des propriétés indépendantes de la fonction de répartition théorique et de n'exiger aucun de ces groupements d'observations que suppose, avec l'arbitraire inhérent à une telle opération, l'application du test du χ^2 par exemple.

De plus, l'application des critères m ω_m^2, W_m^2, D_m^+, D_m, etc... est en pratique assez simple, d'autant plus que le recours à la loi limite est fréquemment suffisant (cf. § 2.5 : extrait de tables de lois-limites).

Les critères de Kolmogorov-Smirnov sont particulièrement aisés à appliquer graphiquement en portant, unilatéralement ou bilatéralement selon le cas, le quantile convenable de la loi correspondante : l'hypothèse est rejetée si la fonction de répartition sort de la "bande" ainsi obtenue.

On transpose aisément les remarques précédentes aux tests de l'identité entre les populations-mères dont sont issus deux échantillons d'observation.

Les tests de Kolmogorov-Smirnov sont corrects, mais non sans distorsion dans le cas d'une hypothèse alternative définie par un ensemble continu de fonctions de répartition théoriques. (cf. [25] - [26] - [27]).

L'étude de la puissance de ces tests et la comparaison de ces tests entre eux ou avec le test du χ^2 ont fait l'objet de nombreux travaux (par exemple : [7] - [11] - [21] - [25] et [27]).

Il existe un certain nombre de situations où le critère de Cramer-von Mises est plus puissant que celui du χ^2. Dans le cas d'alternatives très spécifiques, on devrait pouvoir trouver des tests plus puissants que ceux de Cramér-von-Mises ou de Kolmogorov-Smirnov ; mais dans le cas d'alternatives non spécifiées, ou définies de manière vague, ces tests semblent avoir une puissance assez bonne. Nous avons déjà dit (fin du § 2.2.3.g.) en quoi W_m^2 semblait devoir être préférable à $m\,\omega_m^2$.

Durbin [11] a indiqué un moyen ingénieux d'améliorer, au moins dans certains cas, l'application des tests qui nous intéressent.

Enfin, si la fonction de répartition théorique est inconnue, on peut l'estimer, par exemple au moyen du critère D_m de Kolmogorov-Smirnov, en traçant "une bande de confiance" autour de la fonction de répartition empirique. Toutefois, ce procédé n'est généralement pas le plus avantageux lorsqu'on n'est pas sans renseignements sur la fonction de répartition théorique, par exemple lorsque la loi théorique est spécifiée à la valeur près d'un ou plusieurs paramètres intervenant dans sa définition. En effet, on ne sait guère comment modifier les critères considérés pour les adapter à de telles situations.

2.5

Lois-limites de	Quantiles correspondant aux probabilités : 0,95	0,99	En pratique valables à partir de m (taille de l'échantillon) de l'ordre de :
$\sqrt{m}\,D_m^+$ ou $\sqrt{m}\,D_m^-$ (*)	1,224	1,518	50
$\sqrt{m}\,D_m$ (*)	1,358	1,628	100
$m\,\omega_m^2$	0,461	0,743	20
W_m^2	2,492	3,857	10

(cf. Tables dans : [4] - [29] - [30] - [32], notamment)

(*) Results applicables au cas de deux échantillons selon les indications du § 2.2.2.f.

3 - RENSEIGNEMENTS BIBLIOGRAPHIQUES

[1] ANDERSEN, E. SPARRE (1949 - "On the number of positive sums of random variables" [Skand. Aktuarietidskrift, 32, pp. 27-36].

[2] ANDERSON, T.W. and DARLING, D.A. (1952) - "Asymptotic theory of certain goodness of fit criteria based on stochastic processes". [Ann. Math. Stat., 23, pp. 193-212].

[3] ANDERSON, T.W. and DARLING, D.A. (1954) - "A test of goodness of fit" [J. Am. Stat. Ass., 49, pp. 765-769].

[4] BIRNBAUM, Z. (1952) - "Numerical tables of the distribution of Kolmogorov's statistics for finite sample size" [J. Am. Stat. Ass., 47, pp. 425-444].

[5] BIRNBAUM, Z. and PYKE, R. (1958) - "On some distributions related to the statistics D_n^+" [Ann. Math. Stat., 29, pp. 179-187].

[6] BIRNBAUM, Z. and TINGEY, F.H. (1951) - "One sided confidence contours for probability distribution functions" [Ann. Math. Stat., 22, pp. 592-596].

[7] CHAPMAN, D.G. (1958) - "A comparative study of several one-sided goodness of fit tests" [Ann. Math. Stat., 29, pp. 655-673].

[8] CHENG, P. (1958) - "Non negative jump points of an empirical distribution function relative to a theoretical distribution function" (en chinois, [traduit dans "Selected Transl. in Math. Stat. and Prob"., IMS and AMS*, vol. 3 (1962).]

[9] DARLING, D.A. (1957) - "The Kolmogorov-Smirnov, Cramér-von Mises tests" [Ann. Math. Stat., 28, pp. 823-838].

[10] DOOB, J.L. (1949) - "Heuristic approach to Kolmogoro-Smirnov theorems [Ann. Math. Stat., 20, pp. 393-403].

[11] DURBIN, J. (1961) - "Some methods of constructing exact tests" [Biometrika, 48, pp. 41-55].

[12] FELLER, W. (1959) - "On combinatorial methods in fluctuation theory" [in "Probability and Statistics", the Harald Cramér volume, edited by H. Grenander, J. Wiley, New York].

[13] FELLER, W. (1968) - "An introduction to probability theory and its applications" vol. 1, [J. Wiley, 3rd ed^{on}].

[14] FELLER, W. (1966) - "An introduction to probability theory and its applications" vol. 2, [J. Wiley].

[15] GNEDENKO, B.V. (1952) - "Some results on the maximum discrepancy between two empirical distributions" [Sel. Transl., vol. 1, 1961].

[16] GNEDENKO, B.V. and KOROLYUK, V.S. (1951) - "On the maximum discrepancy between two empirical distributions" [Sel. Transl., vol. 1, 1961].

(*) Ci-après en abrégé "Sel. Transl.".

[17] GNEDENKO, B.V. and MIHALEVIČ, V.S. (1952) - "On the distribution of the number of excesses of one distribution function over another" [Sel. transl., vol. 1, 1961)]

[18] GNEDENKO, B.V. and MIHALEVIČ, V.S. (1953) - "Two theorems on the behaviour of empirical distribution functions" [Sel. transl., vol. 1, 1961).]

[19] GNEDENKO, B.V. and RVAČEVA, E.L. (1952) - "On a problem of the comparison of two empirical distributions" [Sel. Transl., vol. 1, 1961].

[20] KAC, M. (1949) - "On deviations between theoretical and empirical distributions" [Proc. Nat. Acad. Sc., U.S.A., 35, pp. 252-257].

[21] KAC, M., KIEFER, J. and WOLFOWITZ, J. (1955) - "On tests of normality and other goodness of fit tests on distance méthods" [Ann. Math. Stat., 29, pp. 655-673].

[22] KOROLYUK, V.S. (1955) - "On the discrepancy of empirical distribution functions for the case of two independant samples" [Sel. transl., vol. 4, 1963].

[23] LEWIS, P.A.W. (1961) - "Distribution of the Anderson-Darling statistic" [Ann. Math. Stat., 32, pp. 1118-1123].

[24] MARSHALL, A.W. (1950) - "The samll sample distribution of $n\,\omega_n^2$" [Ann. Math. Stat., 29, pp. 307-309].

[25] MASSEY, F.J. (1950) - "A note on the power of a non-parametric test" [Ann. Math. Stat., 20, pp. 440-445].

[26] MASSEY, F.J. (1951) - "The Kolmogorov-Smirnov test of goodness of fit" [J. Ann. Stat. Ass., 46, pp. 68-78].

[27] MASSEY, F.J. (1952) - "Correction to Massey (1950)" [Ann. Math. Stat., 23, pp. 637-638].

[28] MIHALEVIČ, V.S. (1952) - "On the mutual discrepancy of two empirical distribution functions" [Sel. transl., vol. 1, 1961].

[29] MILLER, L.H. (1956) - "Tables of percentage points of Kolmogorov statistics" [J. Am. Stat. Ass., 51, pp. 111-121].

[30] OWEN, D.B. (1962) - "Handbook of statistical tables" [Addison Wesley, Reading (Mass.]

[31] PEARSON, E.S. and STEPHENS, M.A. (1962) - "The goodness of fit tests based on W and W . [Biometrika, 49, pp. 397-402].

[32] RENYI, A. (1966) - "Calcul des probabilités" [Dunod, Paris]

[33] SMIRNOV, N.V. (1939) - "On the estimation of the discrepancy between empirical curves of distributions for two independant samples" [Byull. Moskov. Gos. Univ. Sect. A2 - n° 2].

[34] SMIRNOV, N.V. (1949) - "On the Cramér-von Mises criterion" [Uspehi Matem. Nauk. (NS), 4, pp. 196-197].

[35] SPITZER, F. (1956) - "A combinatorial lemma and its applications to probability theory" [Transact. Amer. Math. Soc., 82, pp. 323-339].

[36] TAKACS, L. (1967) - "Combinatorial methods in the theory of stochastic processes" [J. Wiley, New York].

[37] BICKEL, P.J. (1968) - "Some contribution to the theory of order statistics" (A paraître).

[38] BICKEL, P.J. (1968) - "A distribution - free version of the smirnov Fwo-sample test in the p-variate case" (A paraître).

[39] BILLINGSLEY, P.C. (1968) - "Convergence of probability measures" [J. Wiley, New York].

[40] BLACKMAN, J. (1956) - "An extension of the Kolmogorov distribution Ann. Math. Stat., 27, pp. 513-520 + Correction :[Ann. Math. Stat., 29, pp. 318-324).]

[41] BLUM, J.R. KIEFER, J. and ROSENBLATT, M. (1961) - "Distribution free tests of independance based on the sample distribution function" [Ann. Math. Stat., 32, pp. 485-498].

[42] CAPON, J. (1961) - "Asymptotic efficiency of certain locally most powerful rank tests" [Ann. Math. Stat., 32, pp. 88-100]

[42 bis] CAPON, J. (1961) - "Asymptotic efficiency of certain locally most powerful rank tests" [J. Amer. Stat. Ass., 60, pp. 843-853].

[43] CARNAL, H. (1962) - "Sur les théorèmes de Kolmogorov et Smirnov dans le cas d'une distribution discontinue" [Comm. Math. Helveltici, 37, pp. 19-35].

[44] CARVALHO, P.E.O. (1959) - "On the distribution of the Kolmogorov-Smirnov Statistic" -[Ann. Math. Stat., 30, pp. 173-176].

[45] CHENTSOV, N.N. (1956) - "Weak convergence of stochastic processes whose trajectories have no discontinuities of the second king and the "heuristic" approach to the Kolmogorov-Smirnov tests" [Theor. Probability Appl., 1, pp. 140-144].

[46] CHIBISOV, D.M. (1965) - "An investigation of the asymptotic power of the tests of fit" [Theor. Probability Appl., 10, pp. 421-437].

[47] DAVID, H.T. (1958) - "A three-sample Kolmogorov-Smirnov test" [Ann. Math. Stat., 21, pp. 488-506].

[48] DONSKER, M.D. (1952) - "Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems" [Ann. Math. Stat., 22, pp. 277-281].

[49] DRION, E.F. (1952) - "Some distribution-free tests for the difference between two empirical distribution functions" [Ann. Math. Stat., 23, pp. 563-574].

[50] FISZ, M. (1960) - "On a result by M. Rosenblatt concerning the von-Mises Smirnov test" [Ann. Math. Stat., 31, pp. 427-429].

[51] FISZ. M. (1960) - "Some non-parametric tests for the k-sample problem" [colloquium math. 7, pp. 289-296].

[52] GNEDENKO, B.V. (1954) - "Tests of homogeneity of probability distributions in two independant samples" (en russe) [Math. Nachrichten, 12, pp. 29-66].

[52 bis] HAJEK, J. and SIDAK, Z. - "Theory of rank tests" [Acad. press , New York and Acad. publishing house of the Czechoslovak Academy of sc., Prague].

[53] HOEFFDING, W. (1951) - "A combinatorial central limit theorem" [Ann. Math. Stat., 22, pp. 558-566].

[54] ISHII, G. (1958) - 'Kolmogorov-Smirnov test in life test [Ann. Inst. Stat. Math. 10, pp. 37-38].

[55] KAMAT, A.R. (1956) - "A two-sample distribution free test" [Biometrika, 43, pp. 377-387].

[56] KARLIN, S. (1966) - "A first course in stochastic processes" [Acad. press. New York - Chapitre 9 -]

[57] KIEFER, J. (1959) - "k-sample analogues of the Kolmogorov-Smirnov and Cramér-von Mises tests" [Ann. Math. Stat., 30, pp. 42 à 447].

[58] MOTTO, M. (1957) - "On the Hoeffding's combinatorial central limit theorem" [Ann. Inst. Stat. Math., 8, pp. 145-154].

[59] NOETHER, G.E. (1963) - "Note on the Kolmogorov statistic in the discrete case" [Metrika, 7, pp. 115-116].

[60] ROSENBLATT, M. (1952) - "Limit theorems associated with variants of the von Mises Statistic" [Ann. Math. Stat. 23, pp. 617-623].

[61] SCHMID, P. (1958) - "On the Kolmogorov-Smirnov limit theorems for discontinuous functions" [Ann. Math. Stat., 29, pp. 1 011-1 027].

[62] TSAO, C.K. (1954) - "An extension of Massey's distribution of the maximum deviation between two sample cumulative step functions" [Ann. Math. Stat., 25, pp. 587-592].

[63] TUKEY, J.W. (1959) - "A quick, compact, Fwo-sample test to Duckworth's specifications" [Technometrics, 1, pp. 31-48].

ANNEXE

Propriétés de ω_m^2, de W_m^2 et d'"écarts" plus généraux

1/ Le changement de variable F(x) = p opéré sur l'intégrale de définition du ω_m^2 de Cramér-von Mises, montre, puisque $F_m(x)$ est une variate de Bernoulli (égale à h/m avec la probabilité

$$C_m^h [F(x)]^h [1 - F(x)]^{m-h}) \ (h = 0, 1, \dots, m),$$

que

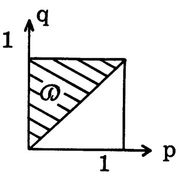
$$m^2 \omega_m^2 = \int_0^1 (u - mp)^2 \, dp \quad , \quad \text{avec} \quad \mathcal{P}r \{u = h\} = C_m^h p^h (1 - p)^h$$

Donc :

$$m^2 E\{\omega_m^2\} = \int_0^1 E\{(u - mp)^2\}\, dp = \int_0^1 V\{u\} dp = \int_0^1 mp(1 - p)\, dp = \frac{m}{6}$$

d'où

$$\boxed{E\{m\ \omega_m^2\} = \frac{1}{6}}$$



De même :

$$m^4\, \omega_m^4 = 2 \iint_{\omega} (u - mp)^2\ (v - mq)^2\ dp\, dq\ (0 < p < q < 1)$$

Or u, v - u et m - v suivent une loi multinomiale (m ; p, q - p, r), où $r = 1 - q$. De plus $-(v - mq) = m - v - m\,(1 - q) = w - mr$, en posant $w = m - v$.

Ainsi :

$$m^4\, \omega_m^4 = 2 \iint_{\omega} (u - mp)^2\ (w - mr)^2\ dp\,dq.$$

Posons

$$Z^{(\nu)} = Z\,(Z - 1) \ldots (Z - \nu + 1)$$

pour exprimer sous leur forme la plus simple les moments (factoriels) de la loi multinomiale faisant intervenir u et w :

$$E\,\{u^{(\alpha)}\, W^{(\beta)}\} = m^{(\alpha+\beta)}\, p^{\alpha}\, r^{\beta}.$$

En utilisant les identités :

$$(u - mp)^2 = u^{(2)} + (1 - 2\,mp)\ u + m^2p^2,$$
$$(w - mr)^2 = w^{(2)} + (1 - 2\,mr)\ w + m^2r^2,$$
$$\begin{aligned}(u - mp)^2\ (w - mr)^2 &= u^{(2)}\ w^{(2)} + (1 - 2\,mr)\, u^{(2)}\, w \\ &+ (1 - 2\,mp)^2 u\, w^{(2)} + m^2r^2\ u^{(2)} + m^2\ p^2\ w^{(2)} \\ &+ (1 - 2\ mp - 2\ mr + 4\ m^2\ pr)\ u\ w \\ &+ (1 - 2\ mp)\ m^2r^2\ u + (1 - 2\ mr)\ m^2p^2 w + m^4\, p^2\, r^2,\end{aligned}$$

on obtient aisément l'égalité :

$$E\,\{(u - mp)^2 (w - mr)^2\} = m\ [3\ (m - 2)\ p^2r^2 - (m - 2)\ (p^2r + pr^2) + m - 1)\ pr]$$

Or

$$\iint_{0<p<q<1} p^a r^b dp dq = \int_0^1 (1 - q)^b dq \int_0^q p^a dp = \frac{1}{a+1} \int_0^1 q^{a+1} (1 - q)^b\ dq = \frac{B(a+2,b+1)}{a+1} = \frac{a\,!\ b\,!}{(a+b+2)\,!}$$

Soit, pour

$$a = b = 2 \quad : \quad \frac{2\,!\,2\,!}{6\,!} = \frac{1}{180}$$

$$a = 2,\ b = 1 \text{ ou } a = 1,\ b = 2 : \frac{2\,!\,1\,!}{5\,!} = \frac{1}{60}$$

$$a = b = 1 \quad : \quad \frac{1\,!\,1\,!}{4\,!} = \frac{1}{24}$$

Il en résulte que

$$E\,\{m^4\,\omega_m^4\} = 2\,m\left[\frac{3\,(m-2)}{180} - \frac{2\,(2-2)}{60} + \frac{m-1}{24}\right] = \frac{m\,(3\,m-1)}{60}$$

D'où

$$V\,\{m^2\omega_m^2\} = E\,\{m^4\,\omega_m^4\} - [E\,\{m^2\,\omega_m^2\}]^2 = \frac{m\,(3\,m-1)}{60} - \frac{m^2}{36} = \frac{m\,(4\,m-3)}{180}$$

Soit enfin :

$$\boxed{V\,\{m\,\omega_m^2\} = \frac{4\,m-3}{180\,m}} \longrightarrow \frac{1}{45} \qquad \text{quand } m \longrightarrow +\infty$$

2/ Pour les applications pratiques, il convient généralement de calculer $m\,\omega_m^2$, non pas par la définition, mais par la première formule g. du § 2.2.3. Celle-ci s'établit de la façon suivante :

Puisque $F_m(x) = \frac{j-1}{m}$, pour $x^*_{j-1} < x \leqslant x^*_j$, on voit, en posant $x = -\infty$ et $x_{m+1} = +\infty$, que :

$$\omega_m^2 = \sum_{j=1}^{m+1}\int_{x^*_{j-1}}^{x^*_j}\left[\frac{j-1}{m} - F(x)\right]^2 d\,F(x) = \frac{1}{3}\sum_{j=1}^{m+1}\left\{\left[F(x) - \frac{j-1}{m}\right]^3\right\}_{x^*_{j-1}}^{x^*_j}$$

ou

$$3\omega_m^2 = \sum_{j=1}^{m+1}\left[F_j^3 - 3\,\frac{j-1}{m}\,F_j^2 + 3\left(\frac{j-1}{m}\right)^2 F_j - F_j^3 + 3\,\frac{j-1}{m}\,F_{j-1}^2 - 3\left(\frac{j-1}{m}\right)^2 F_{j-1}\right]$$

en posant : $F_j = F(x_j^*)$.

Toutes simplifications faites,

$$3\,\omega_m^2 = \sum_{j=1}^{m}\left[\frac{3j}{m}\,F_j^2 - 3\left(\frac{j}{m}\right)^2 F_j - 3\,\frac{j-1}{m}\,F_j^2 + 3\left(\frac{j-1}{m}\right)^2 F_j\right] + F_{m+1}^3\,,$$

soit, puisque

$$F_{m+1}^3 = F^3(+\infty) = 1 \quad , \quad 3\,\omega_m^2 = 1 + \frac{3}{m}\sum_{j=1}^{m}\left(F_j^2 - \frac{2j-1}{m}\,F_j\right)$$

Or :

$$\sum_{j=1}^{m}\left(F_j^2 - \frac{2j-1}{m}F_j\right) = \sum_{j=1}^{m}\left(F_j - \frac{2j-1}{2m}\right)^2 - \sum\left(\frac{2j-1}{2m}\right)^2$$

et, comme

$$\sum_{j=1}^{m}(2j-1)^2 = \frac{m(4m^2-1)}{3}, \quad \omega_m^2 = \frac{1}{3} - \frac{1}{m}\cdot\frac{m(4m^2-1)}{3}\cdot\frac{1}{4m^2} + \frac{1}{m}\sum_{j=1}^{m}\left(F_j - \frac{2j-1}{2m}\right)^2$$

soit enfin la formule g du § 2.2.3 du texte.

3/ On établit de façon analogue la seconde formule g. du § 2.2.3 relative au critère W_m^2 d'Aderson-Darling et on vérifie comme en 1 que $E\{W_m^2\} = 1$.

4/ $F_m(x)$ peut être considérée comme un cas particulier de la fonction

$$\boxed{H_m(x) = \frac{1}{m}\sum_{j=1}^{m} h(x - x_j),}$$

où h(x) est une fonction donnée.

Le cas de $F_m(x)$ correspond, en effet, à

$$h(x) = Y(x) = \begin{cases} 0, & \text{si } x \leq 0 \\ 1, & \text{si } x > 0 \end{cases}$$

$H_m(x)$ étant ainsi définie -en supposant que la densité de probabilité $f(x) = F'(x)$ existe - :

$$H_m(x) = \int_{-\infty}^{+\infty} h(x - \xi)\, dF_m(\xi)$$

$$E\{H_m(x)\} = \frac{1}{m}\cdot m\, E\{h(x - x_j)\} = \int_{-\infty}^{+\infty} h(x-\xi)\, dF(\xi)$$

c'est-à-dire :

$$\boxed{E_m\{H(x)\} = h(x) * f(x)}$$

[convoluée, on produit de composition, de h(x) et f(x)].

On démontre de façon analogue que :

$$\boxed{V\{H_m(x)\} = \frac{h^2(x) * f(x) - [h(x) * f(x)]^2}{m}}$$

Si l'on veut estimer la densité de probabilité f(x), on peut choisir $h(x) = \frac{Y(x+\varepsilon) - Y(x-\varepsilon)}{2\varepsilon}$ (ε positif petit donné), ou, ce qui revient au même, $H_m(x) = \frac{F_m(x+\varepsilon) - F_m(x-\varepsilon)}{2\varepsilon}$, auquel cas

$$E\{H_m(x)\} = \frac{F(x+\varepsilon) - F(x-\varepsilon)}{2\varepsilon} \sim f(x) + \frac{\varepsilon^2}{6} f''(x)$$

Donc $H_m(x)$ est un estimateur de la densité f(x), mais avec une distorsion qui est faible , si et seulement si $\varepsilon \ll \sqrt{\frac{6f(x)}{f''(x)}}$

Par ailleurs, en utilisant le fait que $\text{Cov}\{F_m(x), F_m(y)\} = \frac{F(x)[1-F(y)]}{m}$ pour $x < y$, on obtient en notant $F_{\pm} = F(x \pm \varepsilon)$:

$$4\varepsilon^2 m V\{H_m(x)\} = F_+(1-F_+) - 2F_-(1-F_+) + F_-(1-F_-) = \Delta(1-\Delta),$$

où

$$\Delta = F(x+\varepsilon) - F(x-\varepsilon) \sim 2\varepsilon f(x) + \frac{\varepsilon^3}{3} f''(x)$$

Donc :

$$V\{H_m(x)\} \simeq \frac{f(x)}{2m\varepsilon}$$

La variance de $H_m(x)$ comme estimateur de f(x) n'est donc faible que si, après avoir choisi ε assez petit pour limiter la distorsion, on choisit m assez grand pour que $2m\varepsilon$ soit grand.

On pourrait, par exemple, rendre minimale - pour m donné - la somme de la variance multipliée par un coefficient positif donné λ et du carré de la distorsion, soit

$$\omega = \frac{\varepsilon^4}{36}[f''(x)]^2 + \lambda \frac{f(x)}{2m\varepsilon}$$

On trouve aisément :

$$\varepsilon^5 = \frac{9\lambda}{2m} \cdot \frac{f(x)}{[f''(x)]^2} = \frac{1}{m} \frac{f(x)}{[f''(x)]^2}$$

si l'on choisit, pour fixér les idées, $\lambda = 2/9$.

Si l'on faisait tendre m vers l'infini en donnant à ε la valeur correspondante, $H_m(x)$ serait un estimateur correct de f(x). Bien entendu, ces remarques ont un caractère théorique et ne peuvent, dans les applications , que servir de guide.

5/ Pour généraliser l'écart de Cramér-von Mises, considérons, plutôt que la fonction de répartition empirique, la fonction $\Phi_m(x) = \sum_{i=0}^{m} u_i 1_i(x)$, où

$$1_i(x) = \begin{cases} 0 \ , \ \text{si} \ \ x < x_i^* \\ 1 \ , \ \text{si} \ \ x_i^* \leqslant x < x_{i+1}^* \\ 0 \ , \ \text{si} \ x_{i+1}^* < x \end{cases}$$

les paramètres $0 \leq u_0 < u_1 < u_2 < \ldots < u_{m-1} < u_m \leq 1$ étant choisis pour que

$$\Omega_m^2 = \int_{-\infty}^{+\infty} [\Phi_m(x) - F(x)]^2 \ \Psi[F(x)] \, dF(x) \ ,$$

où $\Psi(t)$ est une fonction positive donnée, ait pour m donné une espérance mathématique minimale.

On remarque que, pour $u_i = \frac{i}{m}$, $\Phi_m(x) = F_m(x)$; mais on constatera dans la suite que ce n'est pas pour le choix $\Psi(t) \equiv 1$ que ces u conduisent à Ω_m^2 d'espérance mathématique minimale.

Posant $F(x) = t$ et remarquant que :

$$1_i[F^{-1}(t)] \begin{cases} = 1 \text{ pour } t_i < t \leqslant t_{i+1} \text{ où } t_i = F(x_j^*) \\ = 0 \text{ ailleurs,} \end{cases}$$

$$\Omega_m^2 = \sum_{i=0}^{m} \int_{t_i}^{t_{i+1}} (u_i - t)^2 \, \Psi(t) \, dt \, .$$

La probabilité élémentaire de t_i et t_{i+1} étant

$$\varphi_{i,i+1} \, dt_i \, dt_{i+1} = \frac{m\,!}{(i-1)\,!\,(m-i-1)\,!} \ t_i^{i-1} (1 - t_{i+1})^{m-i-1} ,$$

$$E\{\Omega_m^2\} = \sum_{i=0}^{m} \iint_{0<t_i<t_{i+1}<1} \varphi_{i,i+1} \, dt_i \, dt_{i+1} \int_{t_i}^{t_{i+1}} (u_i - t)^2 \, \Psi(t) \, dt.$$

Or l'intégrale triple s'écrit, en changeant l'ordre des signes sommes :

$$\int_0^1 (u_i - t)^2 \, \Psi(t) \, dt \iint_{0<t_i<t_{i+1}<1} \varphi_{i,i+1} \, dt_i \, dt_{i+1} \ ;$$

mais la somme double est égale à

$$\frac{m\,!}{(i-1)\,!\,(m-i-1)\,!} \int_0^t t_i^{i-1} \, dt_i \int_t^1 (1 - t_{i+1})^{m-i-1} \, dt_{i+1}$$

$$= \frac{m\,!}{(i-1)\,!\,(m-i-1)\,!} \int_0^t \theta^{i-1} d\theta \int_0^{1-t} \tau^{m-i-1} d\tau$$

$$= \frac{m\,!\,(1-t)^{m-i}}{(i-1)\,!\,(m-i)\,!} \int_0^t \theta^{i-1} d\theta = \frac{m\,!}{(i-1)\,!\,(m-i)\,!} \ (1-t)^{m-i} \cdot \frac{t^i}{i}$$

$$= C_m^i \, t^i \, (1-t)^{m-i} \, .$$

L'intégrale triple est donc égale à

$$C_m^i \int_o^1 (u_i - t)^2 t^i (1 - t)^{m-i} \Psi(t)\, dt = C_m^i (A_i u_i^2 - 2 B_i u_i + C_i),$$

avec

$$A_i = \int_o^1 t^i (1 - t)^{m-i} \Psi(t)\, dt$$

$$B_i = \int_o^1 t^{i+1} (1 - t)^{m-i} \Psi(t)\, dt$$

$$C_i = \int_o^1 t^{i+2} (1 - t)^{m-i} \Psi(t)\, dt.$$

Donc

$$E\{\Omega_m^2\} = \sum_{i=o}^{m} C_m^i (A_i u_i^2 - 2 B_i u_i + C_i)$$

est minimale* pour :

$$u_i = \frac{B_i}{A_i} \quad (i = 1, 2, \ldots, m-1)$$

et égale à

$$\sum_{i=o}^{m} C_m^i \left(C_i - \frac{B_i^2}{A_i}\right)$$

Explicitons le cas $\Psi(t) = t^{\lambda-1} (1 - t)^{\mu-1}$ $(\lambda \geqslant 0, \mu \geqslant 0$.

Alors :

$A_i = B(i+\lambda, m-i+\mu)$, $B_i = B(i+\lambda+1, m-i+\mu)$ et $C_i = B(i+\lambda+2, m-i+\mu)$.

D'où

$$u_i = \frac{B_i}{A_i} = \frac{\Gamma(i+\lambda+1)\,\Gamma(m-i+\mu)\,\Gamma(m+\lambda+\mu)}{\Gamma(i+\lambda)\;\Gamma(m-i+\mu)\;\Gamma(m+\lambda+\mu+1)} = \frac{i+\lambda}{m+\lambda+\mu}$$

$$\sum_{i=o}^{m} C_m^i C_i = \sum_{i=o}^{m} C_m^i B(i+\lambda+2, m-i+\mu) = B(\lambda+2, \mu)^{**}$$

(*) En effet $A_i C_i - B_i^2 > 0$, car $A_i\lambda^2 + 2B_i\lambda + C_i = \int_o^1 (t+\lambda)^2 t^i (1-t)^{m-i} \Psi(t)\, dt > 0$.

(**) D'après l'identité $\sum_{i=o}^{m} C_m^i B(i+\alpha, m-i+\beta) = B(\alpha, \beta)$ obtenue en considérant la loi marginale d'une variate de Bernoulli dont p suivrait une loi de densité $\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)}$

et

$$\frac{B_i^2}{A_i} = u_i B_i = \frac{(i+\lambda)\,\Gamma(i+\lambda+i)\,\Gamma(m-i+\mu)}{(m+\lambda+\mu)\,\Gamma(m+\lambda+\mu+1)} \;;$$

mais $\quad (i+\lambda)\,\Gamma(i+\lambda+1) = \Gamma(i+\lambda+2) - \Gamma(i+\lambda+1),$

donc $\quad (m+\lambda+\mu)\, C_m^i \dfrac{B_i^2}{A_i} = (m+\lambda+\mu+1)\, C_m^i\, B(i+\lambda+2,\, m-i+\mu) - C_m^i\, B(i+\lambda+1,\, m-i+\mu)$

Par suite, une nouvelle application de la formule donnée en bas de page (**) donc :

$$(m+\lambda+\mu) \sum_{i=0}^{m} C_m^i \frac{B_i^2}{A_i} = (m+\lambda+\mu+1)\; B(\lambda+2,\, \mu) - B(\lambda+1,\, \mu),$$

d'où

$$(m+\lambda+\mu) \sum_{i=0}^{m} C_m^i \left(C_i - \frac{B_i^2}{A_i}\right) = [(m+\lambda+\mu) - (m+\lambda+\mu+1)]B(\lambda+2,\mu) + B(\lambda+1,\mu)$$

$$= B(\lambda+1,\, \mu) - B(\lambda+2,\, \mu), = B(\lambda+1,\, \mu+1),$$

d'où enfin :

$$\boxed{\min E\{\Omega_m^2\} = \frac{B(\lambda+1,\, \mu+1)}{m+\lambda+\mu}}$$

Un certain nombre de cas particuliers sont regroupés par le tableau suivant (voir p. 29).

Remarques :

a) Dans le dernier cas, que nous appellerons de Cramér-von Mises modifié, un calcul analogue à celui de la fin du § 1. de la présente annexe permet d'établir que :

$$V\,\{\Omega_m^2\} = \frac{m\,(4m+5)}{180\,(m+2)^4}$$

de sorte que

$$\underline{E\,\{m\,\omega_m^2\} = E\,\{(m+2)\,\Omega_m^2\} = \frac{1}{6},}$$

mais

$$V\,\{m\,\omega_m^2\} = \frac{4\,m-3}{180\,m} > V\,\{(m+2)\,\Omega_m^2\} = \frac{m\,(4m+5)}{180\,(m+2)^2}$$

b) Le calcul de $E\{\Omega_m^2\}$ pour le choix optimal $u_i = \dfrac{i+\lambda}{m+\lambda+\mu}$ peut se faire directement de manière simple comme au début du § 1 de la présente annexe.

$\Psi(t) = t^{\lambda-1}(1-t)^{\mu-1}$	$\frac{1}{t(1-t)}$	$\frac{1}{1-t}$	$\frac{1}{t}$	$\frac{1}{\sqrt{t(1-t)}}$	1
(λ, μ)	$(0, 0)$	$(1, 0)$	$(0, 1)$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$(1, 1)$
$u_i = \frac{i+\lambda}{m+\lambda+\mu}$	$\frac{i}{m}$	$\frac{i+1}{m+1}$	$\frac{i}{m+1}$	$\frac{i+1/2}{m+1}$	$\frac{i+1}{m+2}$
$\min E\{\Omega_m^2\} = \frac{B(\lambda+1, \mu+1)}{m+\lambda+\mu}$	$\frac{1}{m}$	$\frac{1}{2(m+1)}$	$\frac{1}{2(m+1)}$	$\frac{\pi}{8(m+1)}$	$\frac{1}{6(m+2)}$
Observations	Critère d'Anderson-Darling qui posent $W_m = m\,\Omega_m^2$ pour ce choix des u_i, c'est-à-dire $\Phi_m(x) = F_m(x)$.				Le critère ω_m^2 de Cramérvon Mises correspond à $\Psi(t) \equiv 1$, mais non au choix optimal ci-dessous $\left(u_i = \frac{i+1}{m+2}\right)$, puisque $\Phi_m(x) = F_m(x)$ correspond au choix $u_i = \frac{i}{m}$, de sorte que $E\{\omega_m^2\} = \frac{1}{6m}$ supérieur à $\frac{1}{6(m+2)}$