Publ. Inst. Statist. Univ. Paris
16 (1967), 1-22

NOTION D'UN « ESPACE DE L'ÉCHANTILLON » TRONQUÉ, APPLIQUÉE A LA CONSTRUCTION D'UN TEST [1]

I.C. LERMAN

Introduction

Le point de départ de ce travail a été la recherche d'un test assez puissant entre deux hypothèses simples, caractérisées respectivement par les fonctions de répartition $F(x)$ et $G(x)$ au vu d'un petit nombre d'observations. Les fonctions de répartition $F(x)$ et $G(x)$ sont définies sur la droite réelle et sont supposées de formes analytiques quelconques.

On sait que le test le plus puissant pour un niveau de signification fixé est celui de Neyman et Pearson, basé sur le rapport des vraisemblances de l'échantillon dans chacune des hypothèses. Or, on ne connaît pas en général la loi, sous l'hypothèse à tester, de la statistique qui définit ce test ; de sorte qu'on ne peut pas déterminer en général la meilleure région critique pour un niveau de signification fixé. L'intérêt pratique du test de Neyman et Pearson s'en trouve limité.

Nous nous sommes préoccupés de chercher un test qui approche d'une manière satisfaisante le test le plus puissant. Nous avons été amenés à remplacer les fonctions de répartition $F(x)$ et $G(x)$ par leurs sauts respectifs sur les intervalles d'une même subdivision de la droite. Le test devient alors un test entre deux lois multinomiales.

Nous avons introduit l'ensemble des échantillons, d'effectif fixé, pour lesquels il n'y a pas plus d'une observation par intervalle d'une subdivision donnée σ. Nous avons nommé cet ensemble : "espace de l'échantillon tronqué par σ" ; nous en étudions les propriétés en premier lieu.

---

(1) Ce travail a fait l'objet d'une thèse de 3ème Cycle soutenue à l'Université de Paris le 10 Janvier 1966, le texte que nous publions en est un condensé.

Notre test est conditionnel. Les observations obtenues, nous déterminons une subdivision σ, pour laquelle il n'y a pas plus d'une observation par intervalle, nous nous plaçons alors dans l'espace de l'échantillon tronqué par σ pour déterminer la région critique.

Nous montrons au paragraphe C que notre méthode est efficace pour traiter certains tests non paramétriques connus.

A. Espace de l'échantillon tronqué par une subdivision

1. Introduction

Couple formé par une subdivision de la droite réelle $\mathbf{R}$ et par une fonction de répartition définie sur $\mathbf{R}$.

1. Définitions préliminaires

a) X étant une variable aléatoire réelle, la fonction de répartition $F(x)$ est définie par l'égalité

$$F(x) = \Pr\{X < x\}$$

b) Etant donné un nombre η, $o < \eta < 1$; nous appellerons $\xi = F^{-1}(\eta)$ la borne supérieure des nombres x tels que $F(x) < \eta$.

c) Soit la droite réelle $\mathbf{R}$; une subdivision $\sigma = (\xi_0, \xi_1, \ldots, \xi_{n-1}, \xi_n)$ de $\mathbf{R}$ est un ensemble de points de $\mathbf{R}$ rangés par ordre croissant tels que

$$-\infty = \xi_0 < \xi_1 < \ldots < \xi_{n-1} < \xi_n = +\infty$$

A σ correspond la partition de $\mathbf{R}$ en n intervalles

$$(-\infty, \xi_1[, [\xi_1, \xi_2[, \ldots, [\xi_{n-1}, \infty) ;$$

on dira que n est l'effectif de la subdivision.

2. Couple formé par la subdivision et la fonction de répartition $F(x)$.

Considérons les points $\eta_i = F(\xi_i)$; on a

$$o = \eta_0 \leqslant \eta_1 \leqslant \cdots \leqslant \eta_{n-1} \leqslant \eta_n,$$

Nous noterons $\Delta iF = \eta_i - \eta_{i-1} = p_i$; si $p_i \neq 0$ pour tout i ,

$$\omega = (\eta_0 , \eta_1 , \ldots , \eta_n)$$

est une subdivision d'effectif n de l'intervalle $[0 , 1]$, image de σ par F .

DÉFINITIONS . – a) Nous appellerons *module de la subdivision* σ *par rapport à* F la quantité $d_F(\sigma) = \sup_i \Delta iF$

b) Une subdivision σ' *sera dite plus fine que* σ si tout point de σ est un point de σ'

c) Une suite de subdivisions σ_n sera dite *tendant à devenir assez fine pour* F lorsque $n \longrightarrow \infty$ si $d_F(\sigma_n)$ tend vers 0 lorsque $n \longrightarrow \infty$

Réciproquement nous pouvons définir une subdivision de la droite réelle comme image réciproque, par une fonction de répartition, d'une subdivision de l'intervalle $[0 , 1]$.

Soit $\omega = (\eta_0 , \eta_1 , \ldots , \eta_{n-1} , \eta_n)$ une subdivision de l'intervalle $[0 , 1]$; on a $0 = \eta_0 < \eta_1 < \ldots < \eta_{n-1} < \eta_n = 1$. Nous définirons

$$\xi_i = F^{-1}(\eta_i) \quad , \quad i = 1 , 2 , \ldots , n - 1 .$$

On a

$$- \infty = \xi_0 < \xi_1 < \xi_2 < \ldots < \xi_{n-1} < \xi_n = \infty .$$

La subdivision $\sigma = (\xi_0 , \xi_1 , \ldots , \xi_n)$ est d'effectif n si on a

$$\forall i = 1 , 2 , \ldots , n \; ; \; F(\xi_i + 0) - F(\xi_i) < \eta_{i+1} - \eta_i = p_{i+1}$$

Une subdivision qui jouera un rôle important est celle définie par $\Delta iF = \frac{1}{n}$; la subdivision de la droite réelle correspondante est celle définie par les points $\xi_i = F^{-1}\left(\frac{i}{n}\right)$; $i = 0 , 1 , \ldots , n$; cette subdivision est d'effectif n si les sauts de la fonction de répartition F ne dépassent pas $\frac{1}{n}$.

2. Espace de l'échantillon tronqué par une subdivision

L'échantillon de N valeurs d'une variable aléatoire réelle de fonction de répartitions F sera noté $(x_1, x_2, \ldots, x_N)$ et on considèrera la subdivision $\sigma = (\xi_0, \xi_1, \ldots, \xi_n)$ de la droite réelle ; $n > N$.

Considérons l'ensemble des échantillons d'effectif N, de l'espace de l'échantillon pour lesquels il y a *au plus* une observation dans chaque intervalle $Ii = [\xi_{i-1}, \xi_i[$.

Cet ensemble constituera *l'espace de l'échantillon tronqué par la subdivision* σ. Géométriquement cet espace est $\mathbf{R}^N$ auquel on ôte les cylindres

$$\left\{ \begin{array}{l} (x_1, x_2, \ldots, x_N) : \xi_{i-1} \leqslant x_j < \xi_i \ , \ \xi_{i-1} \leqslant x_h < \xi_i \\ i = 1, 2, \ldots, n \quad ; \quad j, h = 1, 2, \ldots, N \quad ; \quad j \neq h \end{array} \right\}$$

Recherchons la probabilité de l'espace de l'échantillon tronqué par σ dans l'hypothèse F : la probabilité cherchée est celle que, pour un échantillon $(x_1, x_2, \ldots, x_N)$ il y ait exactement $(n - N)$ intervalles vides parmi les intervalles

$$Ii = [\xi_{i-1}, \xi_i[\quad i = 1, 2, \ldots, n .$$

Soit $i_1, i_2, \ldots, i_n$ une permutation de $1, 2, \ldots, n$;

$$Ii_1, Ii_2, \ldots, Ii_N \quad \text{et} \quad \sum_{N+1 \leqslant j \leqslant n} Ii_j$$

définissent, dans l'hypothèse F, une urne multinomiale à $(N + 1)$ catégorie ; la probabilité attachée à la catégorie Ii_j est $\Delta i_j F$.

La probabilité que les points x_j ; $j = 1, 2, \ldots, N$; tombent chacun dans un intervalle Ii_j est

$$N! \prod_{j=1}^{N} \Delta i_j F \tag{1}$$

La probabilité cherchée est

$$N! \sum_{J} \prod_{j=1}^{N} \Delta i_j F \ ; \ J = \{i_1, i_2, \ldots, i_N\} \tag{2}$$

la somme est étendue à toutes les combinaisons $J = \{i_1, i_2, \ldots, i_N\}$

de N indices parmi n ; la somme comporte $\binom{n}{N}$ termes.

PROPOSITION 1 . – *La probabilité de l'espace tronqué par une subdivision* σ' *est supérieure à celle de l'espace tronqué par* σ *si* σ' *est plus fine que* σ .

Il suffit de démontrer la proposition dans le cas où σ' se déduit de σ par l'adjonction d'un point. Ce nouveau point, pris dans l'intervalle I_i, entraîne le remplacement de I_i par les deux intervalles $I1_i$ et $I2_i$ disjoint dont I_i est la réunion. En appelant $\Delta 1_i F$ et $\Delta 2_i F$ les sauts respectifs de F sur $I1_i$ et $I2_i$ ($\Delta 1_i F + \Delta 2_i F = \Delta_i F$), on calcule l'accroissement de la probabilité de l'espace tronqué en remplaçant σ par σ', cet accroissement vaut

$$N!\ \Delta 1_i F \times \Delta 2_i F \sum_{H} \prod_{j=1}^{N-2} \Delta_{i_j} F ,$$

La somme est étendue à toutes les combinaisons $H = \{i_1, \ldots, i_{N-2}\}$ de $(N-2)$ indices parmi $1, 2, \ldots, n$. Cet accroissement est nul si $\Delta 1_i F \times \Delta 2_i F = 0$.

PROPOSITION 2 . – *La fonction de répartition* F *étant donnée, parmi les subdivisions* σ *d'effectif* n *celle qui réalise le maximum de l'espace tronqué est défini par*

$$\Delta_i F = \frac{1}{n} \quad ; \quad i = 1, 2, \ldots, n .$$

Posons $\Delta_i F = p_i$, une subdivision σ d'effectif n sera caractérisée par $(p_1, p_2, \ldots, p_n)$; la subdivision optimale est celle qui réalise le maximum de la quantité

$$S_N(p_1, \ldots, p_2) = \sum_{J} p_{i_1} p_{i_2} \cdots p_{i_N}$$

où $J = \{i_1, i_2, \ldots, i_N\}$

sous la contrainte $p_1 + p_2 + \cdots p_n - 1 = 0$.

Pour la rechercher nous utilisons la méthode des multiplicateurs de Lagrange ; considérons

$$S_N(p_1, \ldots, p_n) - \lambda(p_1 + p_2 + \cdots + p_n - 1)$$

La solution optimale est la solution des équations

$$\frac{\partial}{\partial p_i}\ [S_N(p_1, \ldots, p_n) - \lambda(p_1 + p_2 + \cdots + p_n - 1)] = 0$$

On obtient

$S_{N-1}(p_1, p_2, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n) = \lambda$ pour tout $i = 1, \ldots, n$;

Il en résulte que la solution optimale est celle définie par

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}$$

PROPOSITION 3. – *Etant données une fonction de répartition* F *continue et une suite de subdivision* σ_n *tendant à devenir assez fine pour* F ; *la probabilité de l'espace tronqué par la subdivision* σ_n *dans l'hypothèse* F *tend vers* 1 *lorsque* $n \longrightarrow \infty$.

Puisque la suite σ_n tend à devenir assez fine pour F, $\forall \nu$ entier positif, $\exists n_o$

$$n \geq n_o \Longrightarrow \sup_{\sigma_n} \Delta iF = d_F(\sigma_n) < \frac{1}{\nu}$$

Il suffit de démontrer que

$$d_F(\sigma_n) < \frac{1}{\nu} \leqslant \frac{1}{N} \Longrightarrow N!\ S_N \geq \frac{\nu!}{(\nu - N)!} \left(\frac{1}{\nu}\right)^N$$

puisque le second membre de l'inégalité tend vers 1 pour $\nu \longrightarrow \infty$

$$S_N = \sum_J \prod_{j=1}^{N} \Delta i_j F$$

où $J = \{i_1, \ldots, i_N\}$ est une combinaison de N indices parmi 1, 2, ..., n.

On démontre (1) par récurence sur N. Pour N = 1, (1) est évidente.

A un terme $\prod_{j=1}^{N} \Delta i_j F$ de S_N associons la quantité

$$B_J = B\{i_1, \ldots, i_N\} = \prod_{j=1}^{N} \Delta i_j F_x \left[\sum_{i=i_{N+1}}^{i_n} \Delta iF \right]$$

qui représente $(n - N)$ termes de

$$S_{N+1} = \sum_{\{i_1, \ldots, i_{N+1}\}} \prod_{j=1}^{N+1} \Delta i_j F$$

Considérons la somme $S = \sum_J B_J$; par raison de symétrie chacun des termes de S_{N+1} se retrouve dans S un même nombre K de fois : $K\, s_{N+1} = S$; le calcul nous permet de déterminer $K = N+1$ d'où

$$S_{N+1} = \frac{S}{N+1}$$

Pour $\quad d_F(\sigma_n) \leqslant \frac{1}{\nu} < \frac{1}{N+1}$; $B_J \geqslant \frac{\nu - N}{\nu} \prod_{j=1}^{N} \Delta i_j F$

d'où $\quad S = \sum_J B_J \geqslant \frac{\nu - N}{\nu} S_N$

et $\quad S_{N+1} \geqslant \frac{1}{N+1} \times \frac{\nu - N}{\nu} S_N$

Cette dernière inégalité permet de démontrer (1) pour $(N+1)$ dès que (1) est vraie pour N.

B. Test entre deux hypothèses simples

1. Définition du test

Etant donnée une subdivision σ de la droite réelle $\mathbf{R}$, nous caractériserons F et G fonctions de répartition sur $\mathbf{R}$ par les urnes multinomiales

$$(\Delta_1 F\ ,\ \Delta_2 F\ ,\ \ldots\ ,\ \Delta_n F) \quad \text{et} \quad (\Delta_1 G\ ,\ \Delta_2 G\ ,\ \ldots\ ,\ \Delta_n G)$$

où $\Delta_i F = F(\xi_i) - F(\xi_{i-1})$ et $\Delta_i G = G(\xi_i) - G(\xi_{i-1})$; $(\xi_0\ ,\ \xi_1\ ,\ \ldots\ ,\xi_n)$ définissant la subdivision σ, $I_i = [\xi_{i-1},\ \xi_i[$ est un intervalle de σ. A N intervalles $I_{i_1}\ ,\ I_{i_2}\ ,\ \ldots\ ,\ I_{i_n}$ parmi les n $(N < n)$ correspond le produit $\Delta_{i_1} F \times \Delta_{i_2} F \times \ldots \times \Delta_{i_N} F$ (resp. $\Delta_{i_1} G \times \Delta_{i_2} G \times \ldots \Delta_{i_N} G$) qui sera noté $\Delta^J F$ (resp. $\Delta^J G$) où $J = \{i_1\ ,\ i_2\ ,\ \ldots\ ,\ i_N\}$.

Nous nous proposons de définir un test de l'une des distributions multinomiales contre l'autre au vu des observations indépendantes $x_1, x_2, \ldots, x_N$. x_j tombant dans l'intervalle I_{i_j} de σ ; si les i_j sont mutuellement distincts

$$\Phi(\vec{x}) = 1 \quad \text{si} \quad \frac{\Delta^J G}{\Delta^J F} > C$$

$$= \varphi \quad \text{si} \quad \frac{\Delta^J G}{\Delta^J F} = C \quad ; \quad 0 < \varphi < 1 \qquad (1)$$

$$= 0 \quad \text{si} \quad \frac{\Delta^J G}{\Delta^J F} < C$$

$\Phi(\vec{x}) = \Phi(x_1, \ldots, x_N)$ désigne la probabilité de rejet de l'hypothèse nulle.

1. Caractéristiques d'un tel test dans l'espace plein

a) erreur de première espèce

$$\alpha = E_F[\Phi(x)] = N! \left[\sum_H \Delta^J F + \varphi \sum_K \Delta^J F \right]$$

b) puissance du test

$$1 - \beta = E_G[\Phi(x)] = N! \left[\sum_H \Delta^J G + \varphi \sum_K \Delta^J G \right]$$

$$H = \left\{ J \; ; \; \frac{\Delta^J G}{\Delta^J F} > C \right\} \quad , \quad K = \left\{ J \; ; \; \frac{\Delta^J G}{\Delta^J J} = C \right\}$$

c) probabilité de non conclusion peut-être calculée dans l'hypothèse F et dans l'hypothèse G. Cette probabilité est celle qu'il y ait au moins un intervalle qui contienne plus d'un point x_j.

2. Caractéristiques du test dans l'espace tronqué par σ

Appelons C_σ la condition : "Il n'y a pas plus d'une observation par intervalle de la subdivision σ" . Nous supposons savoir que C_σ est satisfaite ; l'erreur de première espèce α et la puissance du test $1 - \beta$ deviennent des espérances mathématiques conditionnelles de la fonction $\Phi(\vec{x})$

$$\alpha = E_F^{C_\sigma}[\Phi(\vec{x})] = \frac{\sum_{J\in H}\Delta^J F + \varphi \sum_{J\in K}\Delta^J F}{\sum_J \Delta^J F} \qquad (1)$$

$$1-\beta = E_G^{C_\sigma}[\Phi(\vec{x})] = \frac{\sum_{J\in H}\Delta^J G + \varphi \sum_{J\in K}\Delta^J G}{\sum_J \Delta^J G} \qquad (2)$$

α étant donnée, on détermine C et φ.

Pour réaliser notre test, en évitant la non conclusion, nous nous placerons dans un espace tronqué. Pour cela, les observations obtenues, nous déterminerons une subdivision σ pour laquelle il n'y a pas plus d'une observation par intervalle ; avec une telle subdivision la condition C_σ est satisfaite.

2. Propriétés optimales du test

Nous allons indiquer les propriétés optimales de notre test,

a) Dans l'espace tronqué par une subdivision

b) Dans l'espace plein.

a) THÉORÈME. – *Dans l'espace tronqué par la subdivision* σ, *le test défini par* $\Phi(x)$ *est le plus puissant pour tester l'hypothèse simple caractérisée par l'urne multinomiale* $(\Delta_1 F, \Delta_2 F, \ldots, \Delta_n F)$ *contre celle caractérisée par l'urne multinomiale* $(\Delta_1 G, \Delta_2 G, \ldots, \Delta_n G)$, *au vu des observations indépendantes* $x_1, x_2, \ldots, x_N$ *à un niveau de signification* α *fixé.*

La démonstration est analogue à celle du lemme de Neyman et Pearson ; on raisonnera dans X_σ, "espace de l'échantillon" $(x_1, x_2, \ldots, x_N)$ tronqué par la subdivision σ - [cf. par exemple [2 – 1]]

b) F et G sont supposées absolument continues et $F'(x) \neq 0$ pour tout x. Nous allons, dans ces conditions, nous occuper des *propriétés limites de notre test, pour une suite de subdivisions* σ_n, *pour lesquelles il n'y a pas plus d'une observation par intervalle* et devenant assez fine pour F, pour n tendant vers l'infini $F(x)$ étant continue et strictement croissante et $G(x)$ continue, la suite σ_n devient encore assez fine pour G.

Considérons une subdivision σ_n de la suite, et la région critique W_n, définie au §I par (1), qui lui est associée.

C et φ étant données, déterminons la forme limite, pour $n \longrightarrow \infty$, de W_n ; pour cela examinons la quantité associée à un échantillon $(x_1 , x_2 , \ldots , x_N)$ et à σ_n

$$\frac{\Delta^J G}{\Delta^J F} = \frac{\Delta_{i_1} G \cdot \Delta_{i_2} G \ \ldots \ \Delta_{i_N} G}{\Delta_{i_1} F \cdot \Delta_{i_2} F \ \ldots \ \Delta_{i_N} F} \tag{1}$$

Le rapport (1) peut-être écrit sous la forme

$$\frac{[G(x_1+h_1)-G(x_1-\eta_1)] \cdot [G(x_2+h_2)-G(x_2-\eta_2)] \cdot \ldots \cdot [G(x_N+h_N) - G(x_N-\eta_N)]}{[F(x_1+h_1)-F(x_1-\eta_1)] \cdot [F(x_2+h_2)-F(x_2-\eta_2)] \cdot \ldots \cdot [F(x_N+h_N) - F(x_N-\eta_N)]} \tag{2}$$

La continuité de F ou de G implique

$$\lim_{n \to \infty} \sup (h_1 , h_2 , \ldots , h_N , \eta_1 , \eta_2 , \ldots , \eta_N) = 0$$

$F'(x_i)$ étant différent de zéro pour tout $i = 1 , 2 , \ldots , N$, la limite du rapport (2) est

$$\frac{G'(x_1) \cdot G'(x_2) \cdot \ldots \cdot G'(x_N)}{F'(x_1) \ F'(x_2) \cdot \ldots \cdot F'(x_N)}$$

On montre alors que

$$\lim \text{ inf } W_n = \lim \text{ sup } W_n = W$$

où W est la région critique du test de Neyman - Pearson, dont le niveau de signification est déterminé à partir de C et de φ . On en déduit le

THÉORÈME . – *Etant données les fonctions de répartition absolument continues* F(x) , *telle que* $F'(x) \neq 0$ *pour tout* x , *et* G(x) ; *étant donnée la suite de subdivisions* (σ_n) *tendant à devenir assez fine pour* F *lorsque* $n \longrightarrow \infty$; W *désignant la région critique du test* (1) *adopté, de* F(x) *contre* G(x) , *au vu des observations indépendantes* $x_1 , x_2 , \ldots , x_N$ *dans "l'espace de l'échantillon" tronqué par* σ_n , *la région critique* W_n *est optimale à la limite pour* $n \longrightarrow \infty$; *dans le sens suivant*

$$\lim \text{ inf } W_n = \lim \text{ sup } W_n = W \ .$$

où W *est la région critique relative au test dont la puissance est maximum.*

3. Généralisation du test d'une variable multidimensionnelle

Considérons le cas de deux dimensions ; nous regarderons (x_1, y_1) , (x_2, y_2) , . . . , (x_N, y_N) ; comme un échantillon d'une variable aléatoire bidimensionnelle de fonction de répartition $F(x, y)$ dans l'hypothèse nulle et $G(x, y)$ dans l'alternative à l'hypothèse nulle. On introduit une "fonction d'ordre" $t(x, y)$, [cf 13 - 1]] , telle que : $Z = t(X, Y)$ est une variable réelle de fonction de répartition : $T(z)$ dans l'hypothèse nulle et $S(z)$ l'alternative à l'hypothèse nulle. On substitue à l'échantillon

$$(x_1, y_1) , (x_2, y_2) , \ldots , (x_N, y_N) ,$$

l'échantillon $(z_1, z_2, \ldots, z_N)$ où $z_i = t(x_i, y_i)$ et on considère le test de la fonction de répartition $T(z)$ contre la fonction de répartition $S(z)$, comme précédemment, au vu des observations $z_1, z_2, \ldots, z_N$.

Le choix de la fonction d'ordre dépend du problème statistique particulier posé. Le cas de K dimensions est exactement analogue au cas de 2 dimensions.

4. Exécution du test dans un espace tronqué particulier

L'exécution du test suppose que les observations $x_1, x_2, \ldots, x_N$, sont toutes distinctes. Considérons la subdivision σ_n définie à partir de la fonction de répartition F par

$$\xi_i = F^{-1}\left(\frac{i}{n} \right) , \quad i = 1, 2, \ldots, n$$

n est choisi assez grand pour qu'il n'y ait pas plus d'une observation x_j par intervalle $I_i = [\xi_{i-1}, \xi_i[$. On a de cette façon

$$\Delta_i F = F(\xi_i) - F(\xi_{i-1}) = \frac{1}{n}$$

et les différents produits notés $\Delta^J F$ sont tous égaux à $\left(\frac{1}{n}\right)^N$.

Nous allons réaliser le test dans l'espace tronqué par la subdivision σ_n ; la région critique est définie par

$$\begin{aligned} \Phi(\vec{x}) &= 1 \quad \text{si} \quad \Delta^J G > C \\ &= \varphi \qquad\qquad = C \\ &= 0 \qquad\qquad < C \end{aligned} \tag{1}$$

les constantes C et φ, $0 < \varphi < 1$, sont déterminées à partir du niveau de signification α du test.

Ordonnons les différentes quantités $\Delta^J G$ par valeurs croissantes

$$\Delta^1 G \leqslant \Delta^2 G \leqslant \ldots \leqslant \Delta^K G < \ldots \leqslant \Delta^m G ,$$

où $m = \binom{n}{N}$. La constante C est précisément la plus petite valeur de $\Delta^J G$, soit $\Delta^K G$, telle que la proportion p des valeurs de $\Delta^J G$ strictement plus grandes que $\Delta^K G$ soit inférieure à α.

$\varphi = (\alpha - p)/p'$, où p' est la proportion des valeurs $\Delta^J G$ égales à $\Delta^K G$. Remplaçons le niveau α par un niveau très voisin $a = \frac{[\alpha m]}{m}$, où $K_\alpha = [\alpha m]$ désigne la partie entière de αm ; dans le cas général où dans la suite ordonnée (2) on a $\Delta^{m-K_\alpha} G < \Delta^{m-K_\alpha+1} G$, la région critique du test devient

$$\begin{aligned} \Phi(\vec{x}) &= 1 \quad \text{si} \quad \Delta^J G \geqslant \Delta^{m-K_\alpha+1} G \\ &= 0 \qquad\quad < \end{aligned} \tag{3}$$

Nous nous placerons dans ce cas ; la réalisation du test nécessite alors la détermination de $\Delta^{m-K_\alpha+1} G$ qui est la quantile $(1 - a)$ de la distribution des valeurs de $\Delta^J G$; un tel calcul ne peut être envisagé, $m = \binom{n}{N}$ est très important même pour N assez petit. On est alors conduit à *estimer* le quantile $(1 - a)$ de la population des valeurs de $\Delta^J G$; soit C' une telle estimation. La donnée de C' nous permet de remplacer la région critique inconnue W par une région critique W' que nous appellerons *région critique estimée* définie par

$$\begin{aligned} \Phi(\vec{x}) &= 1 \quad \text{si} \quad \Delta^J G \geqslant C' \\ &= 0 \qquad\quad < \end{aligned} \tag{4}$$

1. Estimation non paramétrique pour la région critique

L'estimation de la région critique revient, comme nous venons de le voir, à celle de $\Delta^{m-K_\alpha+1} G$; quantile de rang $(1 - a)$ de la population finie dont les valeurs sont les nombres $\Delta^J G$; J parcourant toutes les combinaisons de N indices parmi $1, 2, \ldots, n$. Posons, pour alléger les notations, $\varpi^J = \Delta^J G$, $S = m - K_\alpha + 1$ et supposons, pour fixer les idées, les quantités ϖ^J distinctes. La suite ordonnée, (2) ci-dessus, peut-être écrite sous la forme

$$\varpi^1 < \varpi^2 < \ldots < \varpi^s < \ldots < \varpi^m \ , \quad m = \binom{n}{N} \ . \tag{5}$$

Nous estimerons ϖ^s à partir d'un échantillon de M valeurs tirées au hasard (tirage exhaustif) parmi les valeurs ϖ^J

$$\varpi^{(1)} < \varpi^{(2)} < \ldots < \varpi^{(K)} < \ldots < \varpi^{(M)} \ ; \tag{6}$$

suite ordonnée des valeurs dans l'échantillon.

Il est évident d'après un raisonnement d'analyse combinatoire que

$$\Pr \{\varpi^{(K)} = \varpi^s\} = \frac{\binom{s-1}{K-1} \cdot \binom{m-s}{M-K}}{\binom{m}{M}} \tag{7}$$

où $s = K, K+1, \ldots, m - M + K$

s étant donné, la valeur h de K, pour laquelle la probabilité (7) est maximum fournit l'estimateur $\varpi^{(h)}$ de ϖ^s que nous adopterons. Une telle valeur de h est donnée par $h = \left[\frac{s \quad M}{m+1} + 1\right]$; le crochet désignant la partie entière. Le même traitement s'applique si les inégalités de (5) ne sont pas toutes strictes, ce qui importe c'est le *rang* de ϖ^s dans la suite (5).

La "région critique estimée" que nous adopterons est

$$\begin{aligned} \Phi(x) &= 1 \quad \text{si} \quad \Delta^J G \geqslant \varpi^{(h)} \\ &= 0 \qquad\quad < \end{aligned} \tag{8}$$

Le niveau de signification A de cette région critique est exactement la proportion des valeurs ϖ^i de la suite (5) supérieures ou égales à $\varpi^{(h)}$. A est une variable aléatoire définie à partir du tirage de l'échantillon (6). L'étude de la loi de probabilité de A cf [7]] montre que pour M assez grand on peut admettre les égalités $\hat{A} = a$; $\hat{A}$ désigne la valeur la plus probable de A.

$$E(A) = a$$

et

$$\sigma^2(A) \leqslant \frac{a}{M} + \frac{1}{M^2} \ ;$$

$\sigma^2(A)$ variance de A.

Nous allons proposer une estimation de la puissance du test.

La puissance du test exact (3) est égale à $\sum_{i=s}^{m} \varpi^i \Big/ \sum_{i=1}^{m} \varpi^i$. Remplaçons dans cette expression ω^i par son estimation $\varpi^{(K)}$ où $K = \left[\frac{i\ M}{m+1} + 1\right]$. L'expression de cette estimation peut-être utilement simplifiée et devient :

$$\sum_{i=r}^{M} \varpi^{(i)} \Big/ \sum_{i=1}^{M} \varpi^{(i)}$$

où $r = [M(1 - a) + 1]$, (cf 7).

2. Estimation de la région critique valable pour N assez grand.

Posons $\text{Log } \Delta_i G = \delta_i$, $i = 1, 2, \ldots, n$ et soit $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} \delta_i$, $S^2 = \frac{1}{n} \sum_{i=1}^{n} (\delta_i - \bar{d})^2$ la moyenne et variance de la distribution des nombres δ_i. Pour ce qui concerne la détermination de la région critique, au lieu de considérer la distribution de $\Delta^J G$, nous pouvons, d'une manière équivalente, considérer la distribution des valeurs de $\text{Log } \Delta^J G$. On a

$$\text{Log } \Delta^J G = \text{Log} \prod_{j=1}^{N} \Delta_{i_j} G = \sum_{j=1}^{N} \text{Log } \Delta_{i_j} G = \sum_{j=1}^{N} \delta_{i_j} ;$$

La distribution des valeurs de $\sum_{j=1}^{N} \delta_{i_j}$ est celle de sommes portant sur un échantillon d'effectif N, tiré sans remise de la population finie : $\delta_1, \delta_2, \ldots, \delta_n$. Si $\Delta_{(1-a)}$ désigne le quantile de rang $(1 - a)$ de la distribution de combinaison des valeurs de $\sum_{j=1}^{N} \delta_{i_j}$, la région critique adoptée est définie par

$$W = \left\{ (x_1, \ldots, x_N) \ ; \ x_j \in I_{i_j} \quad \text{et} \quad \sum_{j=1}^{N} \delta i_j \geqslant \Delta_{(1-a)} \right\}$$

Si n est grand, ce qui est en général le cas pour N assez grand, si $\frac{N}{n}$ n'est pas très petit et si $\frac{1}{n} \sum_{i=1}^{n} (\delta_i - \bar{d})^r \Big/ \left[\frac{1}{n} \sum_{i=1}^{n} (\delta_i - \bar{d})^2\right]^{\frac{r}{2}} = 0(1)$ pour $r = 3, 4, \ldots$;

Alors la distribution de $\sum_{j=1}^{N} \delta i_j$ *est approximativement normale de moyenne* $N\bar{d}$ *et de variance* $N\left(1 - \frac{N}{n}\right) S^2$.

Le résultat énoncé se déduit d'un résultat plus général établi par Wald et Wolfowitz [cf [12]] .

Pour se rendre compte de la valeur de l'approximation normale, on peut calculer les coefficients B_1 et B_2 de Pearson en fonction de ceux β_1 et β_2 de la distribution des valeurs δ_i . Le calcul des moments de $\sum_{j}^{N} \delta_{i_j}$ est celui des moments d'une moyenne d'une population finie, [cf [4]] .

On obtient

$$B_1 = \beta_1 \times \frac{(1 - 2\varepsilon)^2}{\varepsilon (1 - \varepsilon)} \times \frac{1}{n^2} \quad \text{où} \quad \varepsilon = \frac{N}{n}$$

$$B_2 = 3 + \beta_2 \times \left[\frac{1}{\varepsilon(1 - \varepsilon)} - 6 \right] \times \frac{1}{n}$$

L'approximation normale est en général satisfaisante. $\Delta(1 - a)$ sera estimé par le quantile de rang $(1 - a)$ de la loi normale

$$\mathcal{N}\left(N \bar{d} \, , \; S \sqrt{N \left(1 - \frac{N}{n}\right)}\right)$$

Il résulte, dans le cadre de notre approximation, une estimation de la puissance du test. La distribution de combinaison de ϖ^J est approximativement log-normale ($\varpi^J = \Delta^J G$) , la puissance du test exact est égale à $\sum_{i=s}^{m} \varpi^i \Big/ \sum_{i=1}^{m} \omega^i$ où $s = m - K_\alpha + 1$, (se référer aux notations du paragraphe précédent) elle peut être écrite sous la forme

$$a \times \frac{\frac{1}{K_\alpha} \sum_{i=s}^{m} \varpi^i}{\frac{1}{m} \sum_{i=1}^{m} \varpi^i}$$

le dénominateur est la moyenne de toutes les valeurs de ϖ^J , on remplacera cette quantité par la moyenne de la loi log-normale ; $\frac{1}{K_\alpha} \sum_{i=s}^{m} \varpi^i$ est la moyenne des K_α plus grandes valeurs de ϖ^J on remplacera cette moyenne par celle de la loi log-normale tronquée pour les valeurs de la variable inférieure à $u(1 - a)$, $u(1 - a)$ désignant le quantile de rang $(1 - a)$ de la loi log-normale.

5. Aperçu sur les calculs relatifs à l'application pratique du test

Nous allons exposer cet aperçu en nous aidant d'un exemple ; soit le test de l'hypothèse simple $F(x) = 1 - e^{-x}$, contre l'hypothèse simple $G(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t}{2}} . t^{-\frac{1}{2}} dt$, (lois du type Γ) au vu des observations indépendantes

$x_1 = 0,052$; $x_2 = 0,068$; $x_3 = 0,165$; $x_4 = 0,195$; $x_5 = 0,459$;
$x_6 = 0,772$; $x_7 = 1,71$; $x_8 = 1,98$; $x_9 = 4,0$. $(N = 9)$

La subdivision σ , définie à partir de la fonction de répartition par $\Delta_i F = 0,02$, $(n = 50)$, est telle qu'il n'y a pas plus d'une observation par intervalle, [cf [7]] , où cet exemple a été traité]. Considérons les nombres $\delta_i = \text{Log } \Delta_i G$ ordonnés par valeurs non décroissantes

$$\delta_{00} \leqslant \delta_{01} \leqslant \delta_{02} \leqslant \cdots \leqslant \delta_{48} \leqslant \delta_{49} \ .$$

1. Méthode d'estimation non paramétrique

D'après ce que nous avons vu si

$$\varpi^{(1)} \leqslant \varpi^{(2)} \quad \ldots \leqslant \varpi^{(r)} \leqslant \ldots \leqslant \varpi^{(M-1)} \leqslant \varpi^{(M)}$$

est un échantillon de M valeurs tirées au hasard (tirage exhaustif) parmi les valeurs de $\varpi^J = \Delta^J G = \prod_{j=1}^{9} \Delta_{i_j} G$, la région critique estimée de niveau a est définie par $\{\Delta^J G \geqslant \varpi^{(r)}\}$ où $r = [M(1-a)+1]$.

Au lieu de considérer les valeurs de $\Delta^J G$ on considèrera leurs logarithmes. Dans ces conditions la pratique du test nécessite

a) Le calcul de $\sum_{j=1}^{9} \delta_{i_j}$ où i_j, $j = 1, 2, \ldots, 9$ sont les indices des intervalles où il y a une observation.

b) Le tirage de l'échantillon d'effectif M

A chaque combinaison : $i_1, i_2, \ldots, i_9$ de $N = 9$ indices parmi les indices $00, 01, \ldots, 48, 49$; correspond une des valeurs de $\text{Log } \Delta^J G = \sum_{j=1}^{9} \delta_{i_j}$. Pour tirer au hasard d'une manière exhaustive un échantillon de M valeurs parmi les $\binom{n}{N}$ valeurs de $\text{Log } \Delta^J G$ il s'agit de tirer au hasard d'une manière exhaustive M combinaisons de $N = 9$ indices parmi $n = 50$. Cela se fait à l'aide d'une

table de nombres au hasard, nous allons expliquer comment, sur notre exemple. Dans une table de nombres au hasard, nous considérons la suite des nombres de 2 chiffres en prenant le complément à 50 de tout nombre dépassant 49. On retiendra successivement tout ensemble, de nombres de 2 chiffres, successifs, comportant chacun exactement $N = 9$ nombres distincts ; les 9 nombres de chacun de ces ensembles représentent une combinaison de 9 indices parmi $00, 01, \ldots, 48 ; 49$, tirée au hasard.

Le tirage effectué comme il vient d'être décrit est un tirage au hasard *non exhaustif ;* on le rend exhaustif en ne retenant pas toute combinaison déjà tirée.

2. Méthode d'estimation de la région critique, valable pour N assez grand

L'application du test suppose

a) Le calcul de $\sum_{j=1}^{N} \delta_{i_j}$ sachant que l'observation x_j appartient à l'intervalle I_{i_j}.

b) Les calculs de $\bar{d}$ et S^2, moyenne et variance de la distribution des nombres δ_i.

c) La détermination de $\Delta(1 - a)$ quantile de rang $(1 - a)$ de la loi normale

$$\mathfrak{N}\left(N\,\bar{d}\,,\ S\ N\left(1 - \frac{N}{n}\right)\right)$$

Le rejet de l'hypothèse nulle correspond à

$$\sum_{j=1}^{N} \delta_{i_j} \geqslant \Delta(1 - a)$$

C. Tests non paramétriques

1. Reprise d'un test classique (cf. [2-2])

Soient $x_1, x_2, \ldots, x_N$; N observations indépendantes d'une variable aléatoire, X, de fonction de répartition F(x) continue, dont on désigne par x_p le pième quantile $p = F(x_p)$. Le test étudié est le suivant

$$\begin{array}{l}\text{hypothèse} \quad H : x_p = x_0 \\ \text{contre} \\ \text{hypothèse} \quad K : x_p > x_0\end{array} \qquad (1)$$

pour raison de simplicité des notations nous prendrons $x_0 = 0$.

Considérons une hypothèse particulière de H caractérisée par la fonction de répartition continue F ; on a $F(0) = p$.

Soit la subdivision de la droite, *assez fine pour qu'elle ne contienne pas plus* d'un point x_1 par intervalle, définie comme suit .

Les points de la subdivision sur le demi-axe réel négatif sont

$$\xi_i = F^{-1}\left(\frac{i\,p}{n}\right) \; ; \; i = 1 \; , \; 2 \; , \ldots , \; n$$

et ceux sur le demi-axe positif

$$\xi_{n+j} = F^{-1}\left(p + \frac{j\,q}{m}\right)$$

où $j = 1 \; , \; 2 \; , \ldots , \; m$; $q = 1 - p$; il y a $(n+m-1)$ points de subdivision. On a

$$\Delta_1 F = \Delta_2 F = \cdots = \Delta_n F = \frac{p}{n} \; ; \; \Delta_{n+1} F = \Delta_{n+2} F = \cdots = \Delta_{n+m} F = \frac{q}{m} \qquad (2)$$

$F(0) = \sum_{i=1}^{n} \Delta_i F = p$, le point 0 est donc un point de la subdivision.

Remplaçons l'hypothèse composite K par l'hypothèse simple $G(x)$ de K , "la plus difficile à distinguer pour le statisticien" de $F(x)$. On a $G(0) = p^* < p$ et $G(x)$ est définie par

$$\Delta_1 G = \Delta_2 G = \cdots = \Delta_n G = \frac{p^*}{n} \; , \; \Delta_{n+1} G = \Delta_{n+2} G = \cdots = \Delta_{n+m} G = \frac{q^*}{m} , \qquad (3)$$

$q^* = 1 - p^*$, définition incomplète, mais suffisante pour notre problème de G .

Le meilleur test dans l'espace tronqué par la subdivision σ est défini comme suit

$$\begin{array}{lll}\Phi(\vec{x}) = 1 & \text{si} & \dfrac{\Delta^J G}{\Delta^J F} > C \\ = \varphi & & = C \\ = 0 & & < C\end{array} \qquad (4)$$

Le rapport $\frac{\Delta^J G}{\Delta^J F}$ peut-être écrit, compte tenu de (2) et de (3), sous la forme

$$\left(\frac{p^*}{p}\right)^{N-I(x_1,\ldots,x_N)} \times \left(\frac{q^*}{q}\right)^{I(x_1,\ldots,x_N)}$$

soit

$$\left(\frac{pq^*}{p^*q}\right)^{I(x_1,\ldots,x_N)} \times \left(\frac{p^*}{p}\right)^N$$

où $(x_1, \ldots, x_N)$ désigne le nombre d'observations positives. Puisque pq^*/p^*q est plus grand que 1, le rapport des vraisemblances est une fonction monotone croissante de $I(x_1, \ldots, x_N)$ et le test (4) devient

$$\begin{aligned} \Phi(\vec{x}) &= 1 \quad \text{si} \quad I(x_1, \ldots, x_N) > C \\ &= \varphi \qquad\qquad\quad = C \\ &= 0 \qquad\qquad\quad < C \end{aligned} \tag{5}$$

C et φ étant déterminés à partir du niveau du test.

Nous montrons ainsi que l'hypothèse d'absolue continuité de la variable aléatoire X (existence d'une densité $f(x)$) habituellement supposée pour asseoir la construction théorique du test, n'est pas nécessaire ; seule l'hypothèse de continuité de la fonction de répartition a été utilisée.

On peut par la même technique étudier un test analogue qui généralise le test précédent, soit

$$\begin{aligned} H &: F(x_0 + h) - F(x_0) = q \\ K &: F(x_0 + h) - F(x_0) = q^* > q \end{aligned} \tag{6}$$

au vu des observations indépendantes $x_1, \ldots, x_N$ de la variable aléatoire de fonction de répartition $F(x)$. x_0 est une valeur donnée de x et h un accroissement fixé ; pour le test précédent on avait $x_0 = 0$ et $h = \infty$.

2. Proposition d'un test

$$\begin{aligned} &\text{hypothèse} \quad H : G(x) = F(x) \quad \text{pour tout} \quad x \\ &\text{contre} \\ &\text{hypothèse} \quad K : G(x) > F(x) \quad \text{" " "} \end{aligned} \tag{1}$$

$G(x)$ étant la fonction de répartition inconnue d'une variable aléatoire réelle X dont on possède N observations indépendantes. $F(x)$ est une fonction de répartition continue donnée.

L'hypothèse composite K peut encore s'exprimer (raison de continuité)

$$G(x) = F\,[x + \delta(x)] \tag{2}$$

où on a uniformément par rapport à x

$$\delta(x) \geqslant d > 0 .$$

Nous remplacerons l'hypothèse composite K par l'hypothèse $G(x)$ de K, "la plus difficile à distinguer, pour le statisticien, de $F(x)$" soit

$$G(x) = F(x + d) \tag{3}$$

Considérons alors la subdivision σ de la droite réelle assez fine pour qu'elle ne contienne pas plus d'un point par intervalle définie comme suit

$$\xi_i = F^{-1}\left(\frac{i}{n}\right) \quad i = 1\ ,\ 2\ ,\ \ldots\ ,\ n \tag{4}$$

on a

$$\Delta_1 F = \Delta_2 F = \ldots = \Delta_h F = \frac{1}{n}$$

Les sauts correspondants de la fonction G sur chacun des intervalles de la subdivision sont donnés par

$$\Delta_1 G = F\left[F^{-1}\left(\frac{1}{n}\right) + d\right]$$

$$\Delta_2 G = F\left[F^{-1}\left(\frac{2}{n}\right) + d\right] - F\left[F^{-1}\left(\frac{1}{n}\right) + d\right]$$

$$\text{----------}$$

$$\Delta_n G = 1 - F\left[F^{-1}\left(\frac{n-1}{n}\right) + d\right]$$

On effectue alors le test entre les deux hypothèses simples dans l'espace tronqué par la subdivision σ.

Conclusion

Notre test emprunte certaines techniques classiques. Il est en effet de pratique assez générale, de remplacer dans un test, comme nous l'avons fait, une distribution par une urne multinomiale ; l'exemple le plus célèbre est celui du test de χ^2, un autre exemple est le "test des cases vides" proposé par David [cf [1]] .

Notre test est basé sur la méthode du rapport des vraisemblances ; il est défini à partir du test de Neyman et Pearson. Sous cet aspect il s'apparente au test de Wilks [cf [14]] et au test basé sur la statistique "Information de Kullback" [cf 5], [3]] .

Considéré comme un test de la qualité d'un ajustement, notre test n'est praticable que si l'on est en mesure de préciser une alternative simple à l'hypothèse nulle.

Parmi les tests entre deux hypothèses simples, celui que nous proposons est presque aussi puissant que le test le plus puissant [cf § B, II]. Notre test est d'application assez générale ; il suffit par exemple que la fonction de répartition F de l'hypothèse nulle soit continue. En particulier il s'applique utilement dans le cas où les observations $x_1, \ldots, x_N$ n'ont pas de résumé exhaustif pour la fonction de répartition F .

Bibliographie

[1] DAVID (F.N.).- Two combinatorial tests of whether a sample has come from a given population, *Biometrika, 37* (1950), 97-110.

[2] FRASER (D.A.S.). - *Nonparametric Methods in Statistics*, John Wiley, New York, (1957), 2.I, 73-75, 2.II, 167-171.

[3] HOEFFDING (W.).- Asymptotically optimal tests for multinomial distributions, *Ann. Math. Statist., 36* (1965), 369-400.

[4] KENDAL (M. G.) & STUART (A.). - *The advanced Theory of Statistics*, Charles Griffin, London, (1963), 301-305.

[5] KULLBACK (S.). - *Information Theory and Statistics*, John Wiley, New York, (1959).

[6] LEHMANN (E.L.).- *Testing Statistical Hypotheses*, John Wiley, New York, (1959).

[7] LERMAN (I.C.). - Thèse de 3ème cycle, Statistique Mathématique, Université de Paris, (1966).

[8] NEYMAN (J.) & PEARSON (E.S.). - On the use and interpretation of certain test criteried for purposes of statistical inference, *Biometrika, 20 A* (1928), I, 175-240, II, 263-294.

[9] NEYMAN (J.) & PEARSON (E.S.). - On the problem of the most efficient tests of statistical hypotheses, *Trans. Roy. Soc. London, 231* (1933), 298-337.

[10] WALD (A.). - Asymptotically most powerful tests, *Ann. Math. Statist., 12* (1941), 1-19.

[11] WALD (A.). - Some examples of asymptotically most powerful tests, *Ann. Math. Statist., 12* (1941), 396-408.

[12] WALD (A.) & WOLFOWITZ (J.). - Statistical tests based on permutations of the observations, *Ann. Math. Statist., 15* (1944), 358-372.

[13] WILKS (S.S.). - *Mathematical Statistics*, John Wiley, New York, (1962), 237-245.

[14] WILKS (S.S.). - The large sample distribution of the likelihood for lesting composite hypotheses, *Ann. Math. Statist., 9* (1938), 60-62.

Reçu le 15 décembre 1966

Statistique Mathématique
Université de Paris