Statistique et analyse des données

Jean-Claude Petit

Essai de présentation des tests non paramétriques à l'aide de la notion de statistique fortement distribution-free

Statistique et analyse des données, tome 1, n° 1 (1976), p. 68-78

<http://www.numdam.org/item?id=SAD_1976__1_1_68_0>

ESSAI DE PRESENTATION DES TESTS NON PARAMETRIQUES A L'AIDE DE LA NOTION DE STATISTIQUE FORTEMENT DISTRIBUTION-FREE.

Jean-Claude PETIT.
I.U.T. INFORMATIQUE - Université de Nancy II.

Janvier 1976.

I. INTRODUCTION

L'ensemble des observations effectuées lors d'une expérience aléatoire est représenté par le vecteur aléatoire $\vec{X} = (X_1,\ldots,X_N)$ à valeurs dans un espace $\mathfrak{X}$, où $\mathfrak{X} \subseteq \mathbb{R}^N$.

Les variables aléatoires X_i sont mutuellement indépendantes.

Le vecteur $\vec{X}$ et sa fonction de répartition H génèrent un espace probabilisé $(\mathfrak{X},\mathfrak{B},P_H)$ où $\mathfrak{B}$ est la tribu boulienne sur $\mathbb{R}^N$ et P_H la loi de probabilité associée à la fonction H.

On note Ω_1^* la classe des fonctions de répartition *continues sur* $\mathbb{R}$ et on envisage une classe Ω de fonctions de répartition sur $\mathbb{R}^N$, de la forme :

$$\Omega = \{H : \exists F_1,\ldots,F_N \in \Omega_1^* \text{ avec } H(x_1,\ldots,x_N) = \prod_{i=1}^{N} F_i(x_i)\}.$$

On considère une partition de Ω en deux classes Ω_H et Ω_K et l'on envisage les hypothèses :

(H) : la distribution de $\vec{X}$ appartient à Ω_H
(K) : la distribution de $\vec{X}$ appartient à Ω_K.

Remarque : les hypothèses (H) et (K) peuvent être très diverses, la seule restriction étant la continuité des fonctions de répartition.

On s'attachera dans cette étude à caractériser *les tests non paramétriques* de l'hypothèse (H) contre l'hypothèse (K), c'est-à-dire les tests fondés sur des statistiques dont la distribution sous l'hypothèse nulle est indépendante des distributions théoriques intervenant dans cette hypothèse [3].

De telles statistiques sont dites *distribution-free sur la classe* Ω_H :

DEFINITION 1 : une statistique T est dite distribution-free sur Ω_H s'il existe une fonction de répartition Q (sur $\mathbb{R}$) telle que :

$$\forall H \in \Omega_H, \quad P_H\left[T(\vec{X}) < t\right] = Q(t) \quad \text{pour tout } t \text{ dans } \mathbb{R}.$$

L'étude de l'existence de statistiques distribution-free a été effectuée par BELL (1). Cette existence est liée à celle d'une partition convenable de l'espace-échantillon.

Lorsque la classe Ω ne contient que des fonctions de répartition continues, il existe toujours au moins une statistique distribution-free sur Ω_H. (voir 5)

Les considérations théoriques développées ci-dessous, fondées sur le principe d'invariance de LAHMANN [4], se proposent de mettre en évidence les propriétés des statistiques distribution-free et d'en donner un procédé de construction. On s'attachera ensuite au cas particulier des statistiques de rang.

II - LES STATISTIQUES FORTEMENT DISTRIBUTION-FREE.

Ces statistiques constituent une extension des statistiques distribution-free ; elles conduisent au calcul de la puissance des tests qui leur sont associés.

II.1 - TRANSFORMATIONS SUR L'ESPACE ECHANTILLON.

Soit une transformation g bijective et mesurable définie sur l'espace échantillon $\mathfrak{X}$ et à valeurs dans $\mathfrak{X}$.
On note $\bar{g}(H)$ la fonction de répartition de la transformée $g(\vec{X})$; il vient alors :

$$\forall B \in \mathfrak{B}, \quad \underset{\bar{g}(H)}{P}\left[g(\vec{X}) \in B\right] = P_H\left[\vec{X} \in g^{-1}(B)\right]$$

Imposons à la transformation g de vérifier les conditions suivantes :

(c)

(i) $\forall H \in \Omega$, $\bar{g}(H) \in \Omega$

$\forall H \in \Omega$, $\exists H \in \Omega : H' = \bar{g}(H)$

ce qu'on note plus brièvement : $\bar{g}(\Omega) = \Omega$

(ii) $\bar{g}(\Omega_H) = \Omega_H$

Considérons un groupe de transformations bijectives et mesurables vérifiant les conditions (c) et noté G ; soit $\bar{G}$ le groupe des transformations sur Ω induit par G.

Il est alors possible de définir deux relations d'équivalences sur $\mathbb{R}^N$ et sur Ω respectivement :

(1) $\vec{x} \sim \vec{y} \pmod{G} \Longleftrightarrow \exists g \in G : \vec{y} = g(\vec{x})$ où $\vec{x}, \vec{y} \in \mathbb{R}^N$.

(2) $H \sim H' \pmod{\bar{G}} \Longleftrightarrow \exists g \in \bar{G} : H' = \bar{g}(H)$ où $H, H' \in \Omega$.

L'espace-échantillon et la classe Ω sont alors partitionnés en classes d'équivalence.

II.2 - ENSEMBLES CARACTERISTIQUES

Certains éléments de la tribu $\mathcal{B}$ jouent un rôle particulier pour définir la notion de statistique fortement distribution-free.

DEFINITIONS : Soit $A \in \mathcal{B}$

1. A *est dit distribution-free sur* Ω_H et de taille α si :

$$\forall H \in \Omega_H , P_H(A) = \alpha$$

2. A *est dit presque invariant sur* Ω pour le groupe G si :

$$\forall H \in \Omega, \forall g \in G \quad \text{on a : } P_H\left[A \, \Delta g(A)\right] = 0$$

3. A *est dit invariant en probabilité sur* Ω pour le groupe G si :

$$\forall H \in \Omega , \forall g \in G \quad \text{on a : } P_H(A) = P_H\left[g(A)\right].$$

Ces trois notions sont liées par les propositions suivantes :

Proposition 1 : Si A est presque invariant sur Ω pour G, alors A est invariant en probabilité sur Ω pour G. [2 ⇒ 3]

Démonstration : Puisque $P_H\left[A - g(A)\right] = P_H\left[g(A) - A\right] = 0$ par hypothèse, il s'ensuit que : $P_H\left[A \cap g(A)\right] = P_H(A)$ et $P_H\left[g(A) \cap A\right] = P_H\left[g(A)\right]$, d'où le résultat.

Proposition 2 : On suppose que $\mathfrak{A}_H$ est une classe d'équivalence relativement au groupe $\bar{G}$.
Si A est invariant en probabilité sur Ω pour G, alors A est distribution-free sur Ω_H. [3 ⇒ 1]

Démonstration : Soient H et H' deux éléments distincts dans Ω_H ; par hypothèse, il existe une transformation $\bar{g}$ dans $\bar{G}$ telle que : $H' = \bar{g}(H)$.
Il vient alors : $P_{H'}(A) = P_{\bar{g}(H)}(A) = P_H\left[g^{-1}(A)\right] = P_H(A)$, ce qui prouve que A est distribution-free sur Ω_H.

Proposition 3 : On suppose que la classe Ω est complète (i.e.: tout estimateur non biaisé de 0 est presque sûrement égal à 0). Si A est invariant en probabilité sur Ω pour G, alors A est presque invariant sur Ω pour G.

Démonstration : Soient I_A et $I_{g(A)}$ les variables indicatrices des ensembles A et g(A). A étant par hypothèse invariant en probabilité, il vient :

$$\forall H \in \Omega, \forall g \in G : \int\left[I_A - I_{g(A)}\right] dP_H = 0$$

Puisque Ω est supposée complète, il s'ensuit que :

$$\forall H \in \Omega, \forall g \in G : P_H\left[I_A - I_{g(A)} \neq 0\right] = 0$$

et par conséquent :

$$P_H\left[A \Delta g(A)\right] = 0$$

[2 ⇒ 3]

Ainsi, lorsque la classe $\mathfrak{A}$ est complète, les notions d'invariance en probabilité et de presque-invariance sont équivalents; (propositions 1 et 3).

II.3 - STATISTIQUES FORTEMENT DISTRIBUTION-FREE.

DEFINITION 1 : Une statistique T est dite fortement distribution-free sur Ω , vis à vis du groupe G si sa distribution est la même sur chacune des choses d'équivalence dans Ω induites par $\bar{g}$;

$$\forall H , K \in \Omega , \quad K \sim H \pmod{\bar{G}} \Rightarrow P_K\left[T < t\right] = P_H\left[T < t\right]$$

pour tout $t \in \mathbb{R}$

L'intérêt d'une telle propriété apparait clairement : si la classe Ω_H (associée à l'hypothèse nulle (H)) est une classe d'équivalence dans Ω , relativement au groupe $\bar{G}$, une statistique fortement distribution-free est alors distribution-free sur Ω_H et conduit à un test non paramétrique de (H) contre (K).

De plus, la puissance du test est la même à l'intérieur de chacune des classes d'équivalence sur Ω.

Caractérisons maintenant ces statistiques :

THEOREME 1 : Une condition nécessaire et suffisante pour qu'une statistique T soit fortement distribution-free sur Ω, vis à vis d'un groupe G, est que pour tout boulien B de $\mathfrak{B}$ l'ensemble $T^{-1}(B)$ soit invariant en probabilité sur Ω pour G.

Démonstration : Soit T fortement distribution-free ; $\forall H , K \in \Omega , H \sim K$, il vient :

$$P_K\left[T^{-1}(B)\right] = P_H\left[T^{-1}(B)\right] \quad \forall B$$

$$\text{or, } P_K\left[T^{-1}(B)\right] = P_{\bar{g}(H)}\left[T^{-1}(B)\right] = P_H\left[g^{-1} \circ T^{-1}(B)\right]$$

ce qui prouve que $T^{-1}(B)$ est invariant en probabilité.

La réciproque résulte immédiatement de l'égalité :

$$P_H\left[g \circ T^{-1}(B)\right] = P_{g^{-1}(H)}\left[T^{-1}(B)\right]$$

Les propriétés énoncées au paragraphe II.2 peuvent être traduites en termes de Statistiques.

DEFINITION : Une statistique T est dite presque invariante sur Ω pour G si :

$$\forall H \in \Omega \ , \forall g \in G \quad \text{on a :} \quad P_H\left[T = T \circ g\right] = 1$$

LEMME 1 : Une condition nécessaire et suffisante pour qu'une statistique T soit presque invariante sur Ω, pour un groupe G est que, pour tout boulien B de $\mathcal{B}$, les ensembles $T^{-1}(B)$ soient presque invariants sur Ω, pour G.

Démonstration : On a : $\{T^{-1}(B) \Delta\ g \circ T^{-1}(B)\} \subset \{T \neq T \circ g^{-1}\}$

Par suite : $\forall H \in \Omega$, $P_H\left[T^{-1}(B) \Delta\ g \circ T^{-1}(B)\right] \leq P_H\left[T \neq T \circ g^{-1}\right] = 0$

La condition est donc nécessaire.

Montrons qu'elle est suffisante :

$$P_H\left[T \neq T \circ g\right] = P_H\left[T > T \circ g\right] + P_H\left[T < T \circ g\right]$$

Un raisonnement simple sur les évènements permet de prouver que :

$$P_H\left[T \neq T \circ g\right] \leq \sum_{m=-\infty}^{+\infty} \sum_{k=1}^{+\infty} P_H\left[T^{-1}\left(-\infty ; \frac{m}{k}\right] \Delta\, g^{-1} \circ T^{-1}\left(\left]-\infty ; \frac{m}{k}\right]\right)\right]$$

Les possibilités intervenant au second membre de l'inégalité étant nulles par hypothèse, il s'ensuit que la statistique T est presque invariante.

On peut alors énoncer le théorème fondamental suivant :

THEOREME 2 :

(i) Si une statistique T est presque invariante sur Ω, pour G, alors T est fortement distribution-free sur Ω, pour G.

(ii) Si la classe Ω est complète, alors si T est fortement distribution-free sur Ω pour G, T est presque invariante.

Démonstration :

(i) T presque invariante $\Longrightarrow \forall B \in \mathcal{B}$, $T^{-1}(B)$ presque invariant (lemme 1)
$\Longrightarrow T^{-1}(B)$ invariant en probabilité (proposition 1 de II.2)
$\Longrightarrow$ T fortement distribution-free (théorème 1).

(ii) T fortement distribution-free $\Longrightarrow \forall B \in \mathcal{B}$, $T^{-1}(B)$ invariant en probabilité (théorème 1)
$\Longrightarrow$ T presque invariant (proposition 3 de II.2)

La notion de statistique presque invariante étant peu maniable dans la pratique, on la particularise :

DEFINITION 2 : Une statistique T est dite invariante pour un groupe de transformations G si elle est constante sur chacune des classes d'équivalence induites par G.

Il est clair que toute statistique invariante est presque invariante et donc fortement distribution-free.

DEFINITION 3 : Une statistique T^* est appelée un invariant maximal pour un groupe de transformations G si elle est invariante et si elle prend des valeurs distinctes sur chacune des classes d'équivalence induites par G.

THEOREME 3 : Une condition nécessaire et suffisante pour qu'une statistique T soit invariante pour G est qu'il existe une fonction h telle que :

$$\forall \vec{x} \in \mathbb{R}^N \quad , \quad T(\vec{x}) = h \circ T^*(x)$$

où T^* est un invariant maximal pour G.

Démonstration : Si $T(\vec{x}) = h \circ T^*(\vec{x})$ alors $T \circ g(\vec{x}) = h \circ T^* \circ g(\vec{x}) = h \circ T^*(x)$ $= T(\vec{x})$ pour tout g dans G ; T est donc invariante.

Réciproquement, si T est invariante et si $T^*(x_1) = T^*(x_2)$ alors il existe g dans G tel que : $x_2 = g(x_1)$ et par suite $T(x_2) = T(x_1)$.

Cette notion de statistique invariante, due à LEHMANN [4], conduit à celle de statistique fortement distribution-free ; elle peut être utilisée pour construire des tests non paramétriques (en choisissant le groupe de transformations adéquat) et de plus, elle fournit certains résultats concernant la puissance de ces tests.

Un exemple est donné au paragraphe suivant.

III - APPLICATION : LES TESTS DE RANG.

On introduit les statistiques de rang en tant que statistiques fortement distribution-free dans la situation particulière d'un test à deux échantillons.

Soit $x_1, x_2, \ldots, x_m$ un échantillon aléatoire non exhaustif extrait d'une variable aléatoire de fonction de répartition continue et strictement croissante, notée F.

Soit d'autre part $y_1, y_2, \ldots, y_m$ un échantillon aléatoire non exhaustif extrait d'une variable aléatoire de fonction de répartition continue et strictement croissante, notée G.

Les deux échantillons (x_i) et (y_i) sont supposés indépendants.

On se propose de tester l'hypothèse nulle : (H) : $F(x) = G(x)$, $\forall x \in \mathbb{R}$ contre l'hypothèse alternative : (K) : $F(x) \neq G(x)$ sur un ensemble de mesure non nulle.

En posant $N = m + n$ et $X_{m+i} = Y_i$ pour $i = 1,2,\ldots,n$, on est amené à considérer les classes de fonctions de répartition sur $\mathbb{R}^N$ définies ci-dessous :

$$\Omega = \{H : \exists F,G \text{ continues et strictement croissantes avec } H(\vec{x}) = \prod_{i=1}^{m} F(x_i) \prod_{i=m+1}^{N} G(x_i)\}$$

$$\Omega_H = \{H : \exists F \text{ continue et strictement croissante avec } H(\vec{x}) = \prod_{i=1}^{N} F(x_i)\}$$

Envisageons un groupe de transformations sur $\mathbb{R}^N$ de la manière suivante :

$$G = \{g = \vec{x} \to g(\vec{x}) = f(x_1), f(x_2), \ldots, f(x_N)$$

où f est une fonction à valeurs réelles continue et strictement croissante. $\}$

Il est immédiat que les bijections de G vérifient les conditions (C) et que Ω_H constitue une classe d'équivalence pour G.

(on a : $\bar{g}\,(H)_{(\vec{x})} = \prod_{i=1}^{m} F \circ f^{-1}(x_i) \prod_{i=m+1}^{N} G \circ f^{-1}(x_i)$)

Soit $\vec{x} = (x_1, x_2, \ldots, x_N)$ une réalisation du vecteur aléatoire $\vec{X}$; les distributions de $\mathcal{Q}$ étant continues par hypothèse, les valeurs x_i sont presque sûrement distinctes.

Notons $x^{(\cdot)} = (x^{(1)}, x^{(2)}, \ldots, x^{(N)})$ le point obtenu à partir de $\vec{x}$ en ordonnant les coordonnées par ordre croissant :

$$x^{(1)} < x^{(2)} < \ldots < x^{(N)}$$

Soit r_i le rang de x_i dans la suite $x^{(\cdot)}$; il vient $x_i = x^{(r_i)}$ pour $i = 1,2,\ldots,N$.

L'application T^* définie par :

$$T^*: \vec{x} = (x_1, x_2, \ldots, x_N) \longrightarrow T^*(\vec{x}) = \vec{r} = (r_1, r_2, \ldots, r_N)$$

est un *invariant maximal pour le groupe G* (à valeurs dans $\mathbb{N}^N$)

(en effet, $T^*(x) = T^*(x') \Rightarrow \vec{r} = \vec{r}' \Rightarrow \vec{x} \sim \vec{x}'$ (mod G)).

Toute statistique invariante pour G est alors de la forme : $T = h \circ T^*$.
On dispose donc de statistiques fortement distribution-free sur $\mathcal{Q}$ pour le groupe G : *ce sont les statistiques de rang.*

Ces statistiques ont été largement étudiées par HAJEK [2].

Ainsi, un test de (H) contre (K) fondé sur une statistique de rang T est *un test non paramétrique* et de plus, il est possible de déduire un résultat important concernant la puissance de ce test :

soient K_1 et K_2 deux éléments de $\mathcal{Q}_K = \mathcal{Q} - \mathcal{Q}_H$; posons :

$$K_1(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{m} F_1(x_i) \prod_{i=m+1}^{N} G_1(x_i)$$

et

$$K_2(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{m} F_2(x_i) \prod_{i=m+1}^{N} G_2(x_i)$$

Si K_1 et K_2 sont équivalents, (au sens de la relation d'équivalence induite par le groupe $\bar{G}$), il existe alors une bijection f telle que :

$$F_1 = F_2 \circ f^{-1} \text{ et } G_1 = G_2 \circ f^{-1}$$

Par conséquent : $G_1 \circ F_1^{-1} = G_2 \circ F_2^{-1}$.

Il est aisé de prouver la réciproque et par suite :

$$K_1 \sim K_2 \pmod{\tilde{G}} \iff G_1 \circ F_1^{-1} = G_2 \circ F_2^{-1}$$

En se fondant sur cette caractéristique des classes d'équivalence dans Ω_K et en utilisant le fait qu'une statistique de rang est fortement distribution-free sur Ω pour G, il apparaît que la distribution d'une statistique de rang sous l'hypothèse alternative et donc la *puissance d'un test de rang de (H) contre (K) ne dépend que de* $G \circ F^{-1}$.

On a utilisé cette propriété pour calculer la puissance de certains tests de rang pour différentes alternatives, avec un gain de calculs appréciable [5].

Remarque : Pour d'autres hypothèses (H) et (K) que celles envisagées ici, le groupe de transformations G sur l'espace échantillon sur lequel repose la construction des statistiques fortement distribution-free ne peut pas toujours être explicité.

IV CONCLUSION

La notion de statistique *"fortement distribution-free"* semble amener une présentation des tests non paramétriques plus large et plus élégante que la présentation classique fondée sur les tests de rang.
Elle apporte de plus une information sur la puissance de ces tests.

Cependant, cet essai nécessite l'hypothèse de continuité des fonctions de répartition et évite le cas des lois de probabilité discrètes.

REFERENCES

[1] BELL C.B. *"Some basic theorems of distribution-free statistics"* The Annals of Mathematical statistics. 35 - (1964) p. 150-156.

[2] HAJEK J. - SIDAK Z. *"Theory of Rank tests"* Academic Press New York (1967)

[3] KENDALL M.G. - SUNDRUM R.M. *"Distribution-free methods and order properties"* Review of the International Statistical Institute 21 p. 124-134 (1953)

[4] LEHMANN E.L *"Testing statistical hypothesis"* J. Wiley - New York 1959

[5] PETIT J.C. *"Essai de systématisation de la théorie des statistiques non paramétriques. Recherche de la puissance de certains tests de rang".* Thèse de troisième cycle. Université de Nancy I. (1974)