

STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

Publicatie

102

Sur les ensembles de confiance généraux
et les méthodes dites non paramétriques

D. van Dantzig



1954

**Sur les ensembles de confiance généraux
et les méthodes dites non paramétriques**

PAR

M. D. VAN DANTZIG

CBRM

COLLOQUE SUR L'ANALYSE STATISTIQUE

Tenu à Bruxelles les 15, 16 et 17 décembre 1954

Sur les ensembles de confiance généraux et les méthodes dites non paramétriques

par M. D. VAN DANTZIG (Amsterdam)

SOMMAIRE

On discute une méthode pour estimer des ensembles de confiance par rapport à l'ensemble de *toutes* les alternatives d'un test, où les variables aléatoires sont indépendantes et continues, isomoires ou non.

D'ailleurs on donne une méthode pour étudier d'une manière unifiée quelques tests non paramétriques récents, généralisant celles de M. G. Kendall (1948), M. Friedman (1937) et F. Wilcoxon (1945). On considère le cas, où les « sondes » (« test-statistics ») peuvent s'exprimer en fonctions linéaires, bilinéaires, quadratiques ou biquadratiques des « signes » (« scores ») $x_{ij} = \text{sgn}(x_i - x_j)$, les x_i étant des nombres aléatoires, le plus souvent supposés indépendants, continu et isomoires sous l'hypothèse essayée.

Considérons une « catégorie d'épreuves » selon la terminologie de M. Fréchet, et soient $\underline{x}_1, \dots, \underline{x}_n$ des nombres aléatoires définis sur elle. Ici le soulignement des x_i indique leur caractère aléatoire. De préférence, mais point du tout nécessairement, un nombre aléatoire et une valeur qu'il peut prendre (ou prend actuellement dans une expérience) seront dénotés par la même lettre. Supposons qu'on sache que la fonction de répartition H de l'ensemble des x_i appartient en tout cas à un ensemble Ω de telles fonctions. Pour tester une fonction particulière $H_0 \in \Omega$ on utilise un « test statistic », c'est-à-dire une fonction univoque des valeurs x_1, \dots, x_n observées (qui peut donc être calculée indépendamment de la fonction de répartition inconnue, mais qui peut dépendre de l'ensemble Ω), et telle qu'on rejette l'hypothèse essayée lorsque sa valeur dépasse une valeur déterminée qui ne dépend que du « seuil de méfiance » (*level of significance, onbetrouwbaarheidsdrempel*) admis, par exemple 0,05. Accepter un seuil de méfiance α déterminé veut dire admettre qu'une fraction α des énoncés statistiques (réjections d'hypothèses, prédictions, etc.) au plus

pourront être faux dans une suite prolongée infiniment de tels énoncés indépendants. Puisque cette fraction mesure le degré de méfiance qu'on donne à ces énoncés et α est le seuil qu'il ne doit pas surpasser, nous avons choisi le terme « seuil de méfiance » qui explique plus clairement la signification de ce concept que le terme anglais « level of significance ». Nous proposons la dénomination « sonde » pour une telle fonction des nombres aléatoires. Soit $t = t(x_1, \dots, x_n)$ cette sonde. En employant t seulement pour tester H_0 , on détermine le nombre t_α^0 le plus petit tel que $P\{t \geq t_\alpha^0 | H_0\} \leq \alpha$ et l'on rejette l'hypothèse $H = H_0$ lorsque l'observation donne un nombre $t \geq t_\alpha^0$, et on ne la rejette pas lorsque $t < t_\alpha^0$.

Toutefois il serait beaucoup plus important de connaître l'ensemble ω de toutes les $H \in \Omega$ tels que $t < t_\alpha$ où t_α est le plus petit nombre tel que

$$P\{t \geq t_\alpha | H\} \leq \alpha$$

c'est-à-dire où

$$P\{t \geq t | H\} > \alpha. \quad (1)$$

L'ensemble ω de tous ces H est un *ensemble de confiance* par rapport au seuil de méfiance α pour l'hypothèse H .

En général il ne sera pas possible de déterminer l'ensemble ω dans une forme raisonnablement simple pour un t et un α (et une sonde t et un ensemble Ω) donnés. Par exemple cela ne sera pas possible lorsque H est le produit des fonctions H_i continues des x_i séparément, c'est-à-dire lorsqu'on sait seulement que les aléatoires x_i sont indépendantes (et continues). Dans ce cas H dépend d'une infinité de paramètres inconnus et (1) exprime une inégalité entre ces paramètres. Pour cette raison le terme « méthodes non paramétriques », que l'on utilise dans ce cas, n'est pas très approprié.

Or ce qu'on peut faire souvent, le nombre t , de même que Ω , la sonde t et α , étant donnés, c'est déterminer deux ensembles ω_* et ω^* , tels que l'on a

$$\omega_* \subset \omega \subset \omega^*. \quad (2)$$

C'est-à-dire l'on sait que l'ensemble de confiance $\text{spr } \alpha$ contient ω_* et qu'il est contenu dans ω^* . Ici $\text{spr } \alpha$ exprime que l'ensemble de confiance, qui est aléatoire puisqu'il dépend de la valeur t dérivée des observations, contient la fonction de répartition inconnue, sauf une probabilité $\leq \alpha$ (« *salva probabilitate α* »). En général on a donc pour un énoncé aléatoire $\underline{\mathcal{A}}$ quelconque

$$\underline{\mathcal{A}} \text{ spr } \alpha \iff P\{\underline{\mathcal{A}}\} \geq 1 - \alpha.$$

Donc «spr 0» exprime ce qu'on appelle souvent «presque certainement».

Notamment on peut déterminer ω_* et ω^* lorsqu'on sait calculer les moments des deux premiers ordres de \underline{t} pour chaque $H \in \Omega$, supposés finis. En effet, posons

$$\theta(H) \stackrel{\text{def}}{=} \mathbb{E} \{ \underline{t} | H \} \quad (4)$$

$$\sigma(H)^2 \stackrel{\text{def}}{=} \text{var} \{ \underline{t} | H \}, \quad \sigma(H) \geq 0. \quad (5)$$

Alors on aura, d'après Cantelli ⁽²⁾,

$$\mathbb{V}_H^Q \mathbb{V}_c^P P \{ \underline{t} - \theta(H) \geq c | H \} \leq \frac{\sigma(H)^2}{c^2 + \sigma(H)^2}, \quad (6)$$

où le symbole \mathbb{V} signifie «pour tou(te)s les», et notamment \mathbb{V}_x^P «pour tous les x qui sont éléments de l'ensemble E ». D'ailleurs P est l'ensemble des nombres positifs. Donc si $t \geq \theta(H)$

$$\begin{aligned} P \{ \underline{t} \geq t | H \} &= P \{ \underline{t} - \theta(H) \geq t - \theta(H) | H \} \\ &\leq \frac{\sigma(H)^2}{\{ t - \theta(H) \}^2 + \sigma(H)^2}. \end{aligned}$$

Cette probabilité est donc certainement $\leq \alpha$ si le membre droit est $\leq \alpha$. Donc, en définissant ω_1 comme l'ensemble de toutes les H telles que

$$\theta(H) \leq t \text{ et } \sigma(H)^2 \leq \alpha [\{ t - \theta(H) \}^2 + \sigma(H)^2]$$

on sait que chaque $H \in \omega_1$ devra être rejetée. En définissant donc ω^* comme le complément de ω_1 , c'est-à-dire comme l'ensemble de toutes les H telles que

$$\theta(H) + \left(\frac{1 - \alpha}{\alpha} \right)^{\frac{1}{2}} \sigma(H) > t \quad (\omega^*) \quad (7)$$

on aura donc $\omega \subset \omega^*$ (cf. fig. 1).

D'autre part on tire de (6) que, si $t < \theta(H)$,

$$\begin{aligned} P \{ \underline{t} < \underline{t} | H \} &= P \{ \theta(H) - \underline{t} > \theta(H) - t | H \} \\ &\leq \frac{\sigma(H)^2}{\{ t - \theta(H) \}^2 + \sigma(H)^2}, \end{aligned}$$

⁽¹⁾ Le symbole $\stackrel{\text{def}}{=}$ désigne une égalité, définissant le membre gauche.

⁽²⁾ L'amélioration des bornes (7) et (8), obtenue par l'emploi de l'inégalité de Cantelli au lieu de celle de Bienaymé, que j'avais utilisée d'abord, est due à M. R. Doornbos, qui a aussi bien voulu m'aider dans la rédaction de cette conférence.

donc certainement $<1-\alpha$ si le membre droit l'est. Dans ce cas on aura donc

$$P\{\underline{t} \geq t | H\} > \alpha$$

et ces hypothèses-ci ne seront pas rejetées. On aura donc $\omega_* \subset \omega$ si ω_* est l'ensemble de toutes les H telles que $\theta(H) > t$ et

$$\sigma(H)^2 < (1-\alpha)[\{\underline{t} - \theta(H)\}^2 + \sigma(H)^2]$$

ou

$$\theta(H) - \left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{2}} \sigma(H) > t \quad (\omega^*) \quad (8)$$

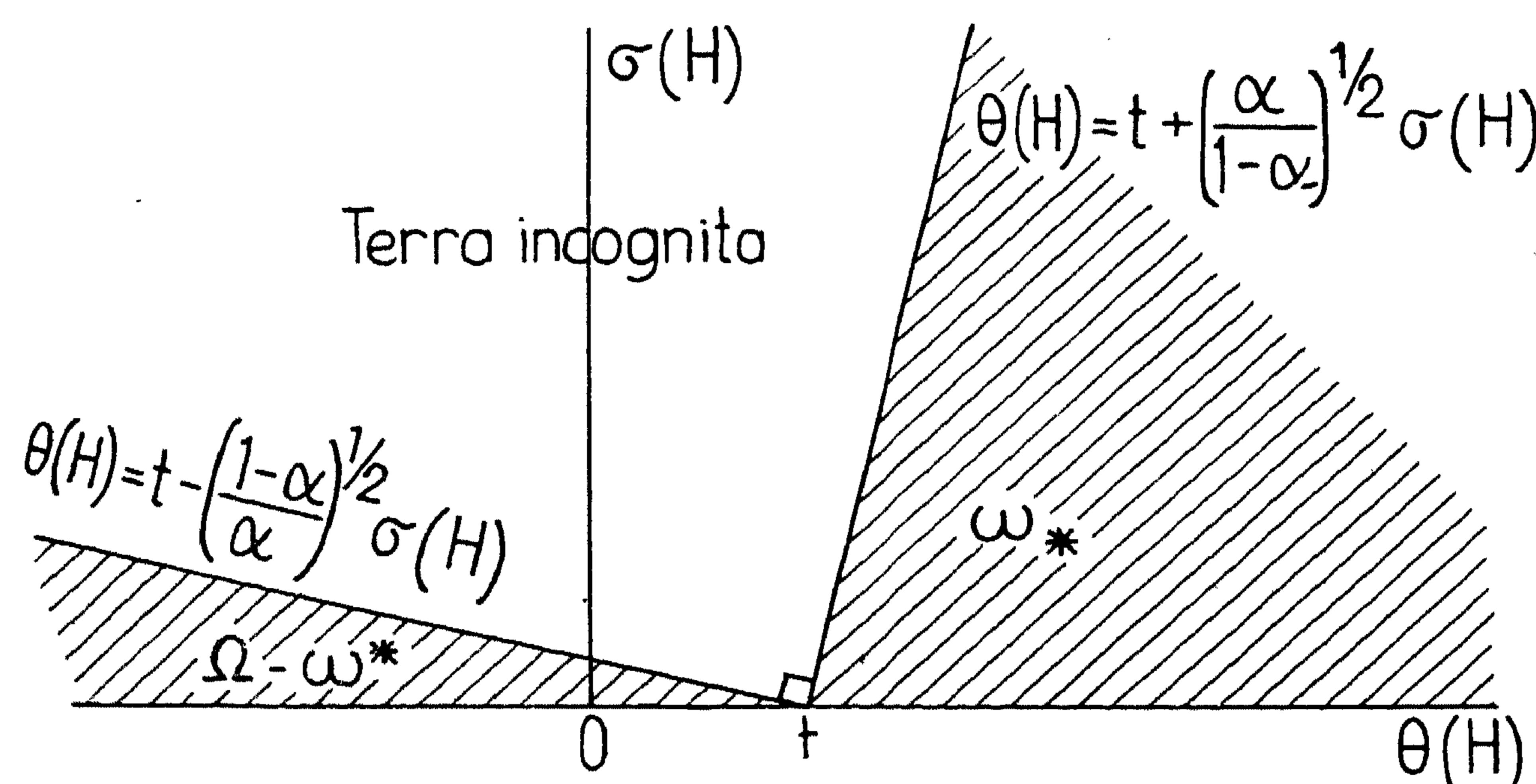


FIG. 1.

Il va sans dire que l'ensemble d'indécision $\omega^* - \omega_*$ peut être diminué si l'on peut calculer quelque moment absolu d'ordre supérieur de \underline{t} pour chaque $H \in \Omega$, ou si l'on peut démontrer que (par exemple) la distribution de \underline{t} est pour chaque $H \in \Omega$ asymptotiquement normale et que l'on estimer la différence entre la fonction de répartition de \underline{t} et son approximation normale.

Si, par exemple, on peut estimer la différence entre la fonction de Lévy⁽³⁾ de \underline{t} (réduit et standardisé) pour chaque H pour tous les arguments u suffisamment petits et $e^{-u^2/2}$, on en peut déduire une estimation

$$|P\{\underline{t} - \theta(H) \leq x\sigma(H) | H\} - \Phi(x)| \leq \varepsilon(H) \quad (9)$$

pour tous les x suffisamment grands, disons pour $x \geq c(H)$, $\Phi(x)$ étant la fonction de répartition normale. Donc, $\xi(\beta)$

⁽³⁾ Nous préférons ce nom à celui de « fonction caractéristique », terme très peu caractéristique.

étant pour un $\beta > 0$ la fonction inverse de $1 - \Phi(\xi)$, c'est-à-dire la solution de l'équation

$$\Phi[\xi(\beta)] = 1 - \beta$$

on aura

$$P\{t \geq t | H\} \leq \varepsilon(H) + 1 - \Phi\left(\frac{t - \theta(H)}{\sigma(H)}\right) \leq \alpha$$

si $\varepsilon(H) \leq \alpha$ et

$$t - \theta(H) \geq \sigma(H) \xi[\alpha - \varepsilon(H)]$$

et l'on peut prendre pour ω^* l'ensemble de toutes les H telles que

$$t - \theta(H) < \sigma(H) \xi[\alpha - \varepsilon(H)] \quad (10)$$

ou $\varepsilon(H) > \alpha \quad (\omega^*)$.

De même on aura

$$P\{t \geq t | H\} \geq -\varepsilon(H) + 1 - \Phi\left(\frac{t - \theta(H)}{\sigma(H)}\right) > \alpha$$

si $\varepsilon(H) < 1 - \alpha$ et

$$t - \theta(H) < \sigma(H) \xi[\alpha + \varepsilon(H)] \quad (\omega_*) \quad (11)$$

Dans ce cas ces H -ci ne seront pas rejetées. En utilisant cette méthode-ci on pourra souvent réduire considérablement l'ensemble d'indécision $\omega^* - \omega_*$ (cf. fig. 2).

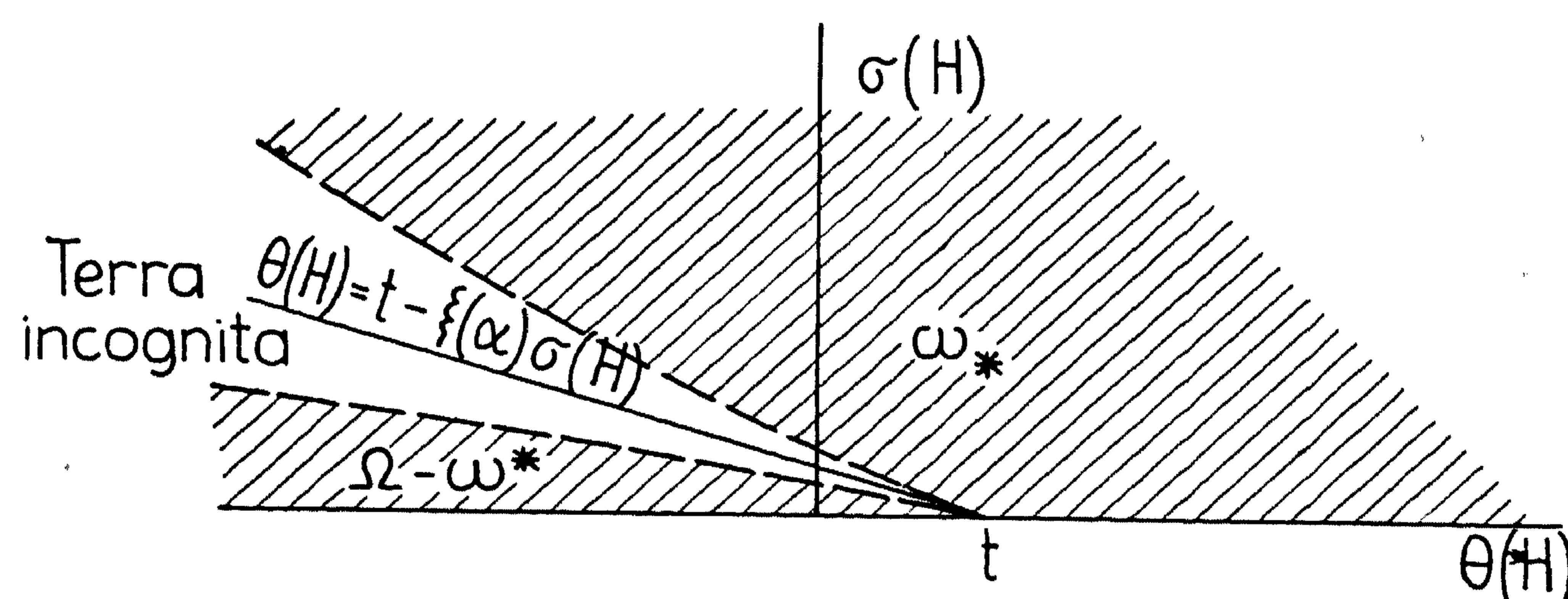


FIG. 2.

Lorsqu'on peut estimer et rendre petite la différence entre la fonction de répartition de t et une autre telle fonction (par exemple celle de χ^2 pour un nombre déterminé de degrés de liberté) on pourra évidemment employer la même méthode, en remplaçant $\xi(\beta)$ par la fonction inverse de cette autre fonc-

tion de répartition. Ces méthodes-ci se laissent appliquer à bien des tests, paramétriques ou non.

Quoique, de cette manière, on n'obtiendra pas en général des résultats *optimaux*, ceci ne diminue pas beaucoup l'application de la méthode esquissée. Car, on ne doit pas oublier qu'en général l'optimalité d'une méthode de test ou d'estimation ne vaut que sous des conditions bien idéalisées. Et, ce qui est optimal sous des conditions idéales peut être loin de l'être sous des conditions réelles.

Nous mentionnons maintenant quelques-unes de ces méthodes non paramétriques.

Considérons une suite finie de n nombres x_1, \dots, x_n . Dans la plupart des applications ces nombres représenteront des valeurs que n aléatoires x_1, \dots, x_n prennent dans une observation (ou dans plusieurs observations).

Pour commencer abandonnons les aléatoires. Soit

$$\mathcal{J} = \{1, \dots, n\}$$

l'ensemble des indices. Lorsque les $x_i (i \in \mathcal{J})$ sont remplacées par $y_i = f(x_i)$, où $f(x)$ est une fonction monotone croissante, les n^2 quantités, que nous appellerons signes (anglais « scores »)

$$x_{ij} \stackrel{\text{def}}{=} \text{sgn}(x_i - x_j)$$

où

$$\text{sgn}(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \\ -1 & \text{si } x < 0 \end{cases} \quad (13)$$

restent les mêmes, et on peut prouver que chaque quantité, fonction des x_i , qui possède cette invariance par rapport à toutes les transformations monotones se laisse exprimer en fonction des x_{ij} .

Les n^2 tailles satisfont aux conditions

$$\nabla_i^{\mathcal{J}} \nabla_j^{\mathcal{J}} x_{ij} = x_{ij}, \quad (14)$$

$$\nabla_i^{\mathcal{J}} \nabla_j^{\mathcal{J}} x_{ij} + x_{ji} = 0, \quad (15)$$

$$\nabla_i^{\mathcal{J}} \nabla_j^{\mathcal{J}} \nabla_k^{\mathcal{J}} x_{ij} x_{jk} x_{ki} + x_{ij} + x_{jk} + x_{ki} = 0. \quad (16)$$

Plus généralement on a

$$\nabla_h^{\mathcal{J}} \nabla_i^{\mathcal{J}} \nabla_j^{\mathcal{J}} \nabla_k^{\mathcal{J}} x_{hi} x_{jk} + x_{ij} x_{hk} + x_{jh} x_{ik} = x_{hi} x_{hj} x_{hk} x_{ij} x_{ik} x_{jk}, \quad (17)$$

identité qui se déduit de (16). On démontre facilement que chaque matrice x_{ij} ayant les propriétés (14), (15), (16) se

laisse exprimer dans la forme (12) au moyen de n nombres x_i convenablement choisis. Notamment on peut choisir $x_i = s_i$, où

$$s_i = \sum_{j \in \mathcal{J}} x_{ij} . \quad (18)$$

L'identité (16) se laisse remplacer par

$$x_{ij}x_{ik} + x_{jk}x_{ji} + x_{ki}x_{kj} = 1 - \delta_{ijk} , \quad (19)$$

où δ_{ijk} est un symbole de Kronecker généralisé, à savoir = 1 si $x_i = x_j = x_k$ et = 0 dans tout autre cas. Si les x_i sont différents les s_i sont intimement liés aux « rangs » r_i des valeurs observées x_i , lorsque celles-ci sont rangées par ordre de grandeur croissante :

$$s_i = 2r_i - (n + 1) . \quad (20)$$

Dans le cas général, (20) servira comme *définition* des r_i . Lorsque quelques-uns des x_i sont égaux on appellera un « lien » (anglais : « tie », hollandais « knoop ») un sous-ensemble de \mathcal{J} tel que les x_i correspondants soient égaux et qui n'est pas contenu dans un sous-ensemble plus large ayant la même propriété. On attribue selon M. G. Kendall (1948) à chaque tel x_i comme rang r_i , la moyenne arithmétique des rangs qu'ils auront lorsqu'ils sont assujettis à des changements de valeur suffisamment petits, mais tels qu'ils deviennent tous différents. On voit aisément que la définition de Kendall (1948) est d'accord avec (20). Au lieu des r_i nous emploierons toujours les s_i , qu'on pourrait appeler les « doubles rangs réduits », mais que nous appellerons brièvement « rangs » dans les descriptions verbales.

Les signes et les rangs entrent dans plusieurs tests non paramétriques d'hypothèses statistiques. Nous en énumérons les plus importantes en introduisant entre temps quelques abréviations utiles.

Considérons d'abord le *test de deux échantillons* (« two sample test ») de Wilcoxon (1945), dans la forme générale développée par H. B. Mann et D. R. Whitney (1947). Soient donnés deux échantillons $x_1, \dots, x_m ; y_1, \dots, y_n$ qui sont les valeurs prises par $m + n$ variables aléatoires, dont on sait (Ω) qu'elles sont toutes indépendantes, et que les x_i , de même que les y_j , sont « isomoires », c'est-à-dire qu'elles ont toutes la même fonction de répartition, qui d'ailleurs est continue, par exemple $\mathcal{F}(x)$ pour les x_i , $\mathcal{G}(y)$ pour les y_j . Afin de tester si $\mathcal{F}(x)$ et $\mathcal{G}(x)$ sont identiques, on utilise la sonde, égale au nombre de fois qu'un des x_i , $i \in \{1, \dots, m\}$ surpassé un

des y_j , $j \in \{1, \dots, n\}$ pourvu qu'aucun des x_i ne soit égal à un y_j . En introduisant la fonction de saut unitaire

$$\iota(x) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases} \quad (21)$$

cette quantité est donc égale à

$$W = \sum_1^m \sum_1^n \iota(x_i - y_j). \quad (22)$$

Or, excepté pour la valeur $x=0$, on a

$$\iota(x) = \frac{1}{2} + \frac{1}{2} \operatorname{sgn}(x). \quad (23)$$

Dans le cas général, où des valeurs égales sont admises, on peut définir

$$W \stackrel{\text{def}}{=} \frac{1}{2} \sum_1^m \sum_1^n \operatorname{sgn}(x_i - y_j) + \frac{1}{2} mn. \quad (24)$$

Au lieu de W on peut aussi bien utiliser $t_w = 2W - mn$ comme sonde. Changeons les notations. Mettons $n_A = m$, $n_B = n$ et écrivons n pour $m+n$. Rangeons maintenant les $n_A + n_B$ valeurs observées dans une seule suite, dans un ordre arbitraire, que nous appelons x_1, \dots, x_n . Le premier échantillon est déterminé par l'ensemble A des indices i , tels que x_i lui appartient ; de même B dénotera l'ensemble des indices des valeurs appartenant au deuxième échantillon, de même que $A+B = \{1, \dots, n\}$. Avec ces notations on aura

$$t_w \stackrel{\text{def}}{=} \sum_{i \in A} \sum_{j \in B} \operatorname{sgn}(x_i - y_j) = \sum_{i \in A} \sum_{j \in B} x_{ij}. \quad (25)$$

Puisque nous aurons souvent à sommer un nombre de quantités, dénotées par une lettre avec un indice, par exemple u_i , lorsque l'indice parcourt un ensemble (fini), par exemple A , il est utile de dénoter cette somme par la lettre avec le symbole de l'ensemble comme indice. Donc

$$u_A \stackrel{\text{def}}{=} \sum_{i \in A} u_i. \quad (26)$$

De même avec plusieurs indices, par exemple

$$u_{A, B} \stackrel{\text{def}}{=} \sum_{i \in A} \sum_{j \in B} u_{ij}. \quad (27)$$

Avec ces notations la sonde de Wilcoxon (1945) (doublée et réduite) s'écrit

$$t_w = x_{A, B} . \quad (28)$$

La fonction de répartition sous l'hypothèse que tous les x_i sont indépendants, continus et isomorphes est connue, et asymptotiquement normale ; on rejette cette hypothèse lorsque $|x_{A, B}|$ dépasse la valeur correspondante avec le seuil de méfiance prescrit α .

Or, on peut considérer les sommes $x_{A, B}$ lorsque A et B sont des sous-ensembles de \mathcal{J} arbitraires, non nécessairement complémentaires. Alors $x_{A, B}$ est une fonction additive de A (de même que de B), c'est-à-dire

$$x_{A_1 + A_2, B} = x_{A_1, B} + x_{A_2, B} ; \quad (29)$$

pourvu que les ensembles A_1 et A_2 soient *disjoints*. D'ailleurs on a

$$x_{A, B} = -x_{B, A} , \quad (30)$$

donc spécialement $x_{A, A} = 0$. D'ailleurs, lorsque A et B sont disjoints

$$x_{A, A+B} = x_{A, A} + x_{A, B} = x_{A, B} \quad (31)$$

et lorsque A et B ont l'intersection $C = A \cap B$

$$\begin{aligned} x_{A, B} &= x_{A-C, B} + x_{C, B} = x_{A-C, B} + x_{C, B-C} \\ &= x_{A-C, B-C} + x_{A-C, C} + x_{C, B-C} \end{aligned} \quad (32)$$

puisque $x_{C, C} = 0$, où $A-C$ et $B-C$ dénotent les compléments de C dans A et B respectivement. Par (32) $x_{A, B}$ pour A et B non disjoints est exprimé au moyen de telles quantités se rapportant aux ensembles disjoints.

En prenant pour A l'ensemble qui n'existe que d'un élément i , et dénotant cet ensemble par i simplement (au lieu de $\{i\}$) et l'ensemble complémentaire par $\mathcal{J}-i$, et pour A l'ensemble \mathcal{J} , on obtient

$$x_{i, \mathcal{J}} = x_{i, \mathcal{J}-i} = \sum_{j \in \mathcal{J}} x_{ij} ,$$

c'est-à-dire le « rang » (double rang réduit) s_i . Celui-ci est donc un cas spécial du wilcoxonien $x_{A, B}$.

Passons maintenant aux aléatoires. Le plus souvent on teste l'hypothèse H_0 , exprimant que les aléatoires

$$\underline{x}_i (i \in \mathcal{J} = A + B)$$

supposées continues, sont indépendantes et isomorphes. Cette hypothèse implique l'hypothèse, dite H_{00} , disant que la distribution simultanée des x_i est invariante par rapport au groupe de toutes les permutations des indices i . Restreignons-nous d'abord au cas où les distributions des aléatoires x_i sont d'ailleurs continues, de même qu'il n'y a pas de liens (c'est-à-dire que tous les liens ont la longueur 1), c'est-à-dire que tous les x_i sont différents de 0.

Sous l'hypothèse H_{00} , ou aussi sous H_0 et dans le cas de continuité non seulement l'espérance

$$\mathcal{E}_0 \underline{W} = \frac{1}{2} mn$$

et la variance,

$$\mathcal{E}_0 \left(\underline{W} - \frac{1}{2} mn \right)^2 = \frac{1}{12} mn(m+n+1),$$

mais aussi la fonction de répartition de \underline{W} est connue depuis Mann et Whitney (1947). On sait d'ailleurs qu'elle est asymptotiquement normale.

Plus généralement on prouve facilement (toujours dans le cas de continuité)

$$\mathcal{E}_0 \underline{x}_{ij} = 0, \quad (33)$$

$$\begin{cases} \mathcal{E}_0 \underline{x}_{hi} \underline{x}_{jk} = 0 & \text{si } h, i, j, k \text{ sont inégaux,} \\ \mathcal{E}_0 \underline{x}_{hi} \underline{x}_{ji} = \frac{1}{3} & \text{si } h, i, j \text{ sont inégaux,} \\ \mathcal{E}_0 \underline{x}_{hi}^2 = 1 & \text{si } h, i \text{ sont inégaux.} \end{cases} \quad (34)$$

ou généralement,

$$\begin{aligned} \mathcal{E}_0 \underline{x}_{hi} \underline{x}_{jk} &= \frac{1}{3} (\delta_{hj} - \delta_{ij} - \delta_{hk} + \delta_{ik}) + \frac{1}{3} (\delta_{hj} \delta_{ik} - \delta_{ij} \delta_{hk}) \\ &= \{(1 + \delta_{hj})(1 + \delta_{ik}) - (1 + \delta_{ij})(1 + \delta_{hk})\}, \end{aligned} \quad (35)$$

où

$$\delta_{hj} = \begin{cases} 1 & \text{si } h = j, \\ 0 & \text{si } h \neq j, \end{cases} \quad (36)$$

est le symbole de Kronecker. La relation (35) montre clairement la symétrie par rapport aux deux paires d'indices hi et jk , et l'antisymétrie par rapport à chacune de ces paires.

Donc, en considérant la forme linéaire générale

$$\underline{L} = \frac{1}{2} \sum^{i,j} l_{ij} \underline{x}_{ij}, \quad (37)$$

où i et j parcouruent \mathcal{J} , et où l'on peut supposer sans restriction

$$l_{ij} = -l_{ji}, \quad (38)$$

on obtient immédiatement

$$\mathcal{E}_0 \underline{L} = 0 \quad (39)$$

$$\text{var}_0 \underline{L} = \mathcal{E}_0 \underline{L}^2 = \frac{1}{6} \sum_{h,i} l_{hi}^2 + \frac{1}{3} \sum_h l_{h,\mathcal{J}}^2, \quad (40)$$

où, comme auparavant

$$l_{h,\mathcal{J}} = \sum_{j \in \mathcal{J}} l_{hj}.$$

L'espérance et la variance de \underline{W} se déduisent directement de celles de $\underline{x}_{A,B}$, ce qui est un cas spécial de \underline{L} , qu'on obtient en posant

$$l_{ij} = I_A^i I_B^j - I_B^i I_A^j, \quad (41)$$

où, pour un sous-ensemble donné de S et \mathcal{J} , I_s^i est la «fonction caractéristique» de cet ensemble, c'est-à-dire $=1$ lorsque $i \in S$ et $=0$ si non. Evidemment I_s^i est une généralisation du δ de Kronecker qu'il devient pour $n_s=1$. Alors on obtient

$$l_{i,\mathcal{J}} = n_B I_A^i + n_A I_B^i,$$

où n_s est le nombre d'éléments de S , tandis que $n_{\mathcal{J}}=n$. Donc pour $A \cap B=0$. (Cf. Mann and Whitney.)

$$\mathcal{E}_0 \underline{x}_{A,B} = 0 \quad \mathcal{E}_0 \underline{x}_{A,B}^2 = \frac{1}{3} n_A n_B (n_A + n_B + 1). \quad (42)$$

Dans le cas plus général où A et B peuvent avoir n_c éléments communs l'additivité (31) ou (32) donne au lieu de la dernière expression dans (42)

$$\frac{1}{3} n_A n_B (n_A + n_B - 2 n_c + 1) - \frac{1}{3} n_c^2.$$

En prenant en particulier $A=i(=\{i\})$, $B=\mathcal{J}$, donc

$$n_A = n_c = 1, \quad n_{\mathcal{J}} = n, \quad \underline{x}_{A,B} = \underline{x}_{i,\mathcal{J}} = \underline{S}_i :$$

$$\mathcal{E}_0 \underline{S}_i^2 = \frac{1}{3} (n^2 - 1), \quad (43)$$

identité qu'on peut aussi dériver directement.

Remarquons que les relations (34) se laissent généraliser en remplaçant les signes par des wilcoxoniens. En effet, en sommant les deux membres de (35) par rapport à $h \in A$, $i \in B$, $j \in C$, $k \in D$ où A , B , C , D sont des sous-ensembles, disjoints ou non, de S , on obtient

$$\mathcal{E}_0 x_{A,B} x_{C,D} = \frac{1}{3} (n_{A \cap C} n_B n_D - n_{B \cap C} n_A n_D - n_{A \cap D} n_B n_C + n_{B \cap D} n_A n_C + n_{A \cap C} n_{B \cap D} - n_{A \cap D} n_{B \cap C}). \quad (44)$$

Spécialement, si $A \cup B$ et $C \cup D$ sont disjoints, $\mathcal{E}_0 x_{A,B} x_{C,D} = 0$, et, si A , B , D sont disjoints,

$$\mathcal{E}_0 x_{A,B} x_{A,D} = \frac{1}{3} n_A n_B n_D. \quad (45)$$

Les relations (44) se laissent écrire d'une manière plus simple lorsqu'on introduit au lieu des sommes

$$u_A = \sum_{i \in A} u_i$$

les *moyennes* correspondantes, que nous dénotons par un suffixe en haut, qui indique l'ensemble sur lequel la moyenne est prise

$$u^A \stackrel{\text{def}}{=} \frac{1}{n_A} \sum_{i \in A} u_i = \sum_{i \in A} u_i / \sum_{i \in A} 1. \quad (46)$$

Lorsque $n_A = 0$ nous définissons u^A comme 0, quels que soient les nombres u_i (finis et en nombre fini).

Spécialement on peut introduire les quantités

$$x^{A,B} \stackrel{\text{def}}{=} x_{A,B} / n_A n_B, \quad (47)$$

qui ont le caractère de « signes moyennes pondérées ». Elles sont des moyennes de signes, donc toujours ≤ 1 en valeur absolue, et aussi antisymétriques par rapport aux deux indices. Par définition $x^{A,B} = 0$ si A ou B est vide.

Remarquons que l'identité correspondante avec (31) est un peu différente pour les moyennes

$$x^{A,A+B} = \frac{n_B}{n_{A+B}} x^{A,B}, \quad (48)$$

où $n_{A+B} = n_A + n_B$ si A et B sont disjoints.

Au moyen des notations (47) et

$$\hat{x}^{A,B} \stackrel{\text{def}}{=} \frac{n_{A \cap B}}{n_A n_B} \quad (49)$$

si $n_A n_B \neq 0$ (et $\delta^{A,B} = 0$ si $n_A n_B = 0$), ce qui est une autre généralisation des Kronecker δ , dans laquelle elle passe pour $n_A = n_B = 1$, (44) devient

$$\begin{aligned} \mathcal{E}_0 \underline{x}^{A,B} \underline{x}^{C,D} &= \frac{1}{3} (\delta^{A,C} - \delta^{A,D} - \delta^{B,C} + \delta^{B,D}) \\ &\quad + \frac{1}{3} (\delta^{A,C} \delta^{B,D} - \delta^{A,D} \delta^{B,C}) \\ &= \frac{1}{3} \{ (1 + \delta^{A,C}) (1 + \delta^{B,D}) - (1 + \delta^{A,D}) (1 + \delta^{B,C}) \}, \end{aligned} \quad (50)$$

ce qui est une généralisation immédiate de (35), dans laquelle les effectifs n_A, \dots, n_D ne paraissent plus explicitement.

Spécialement, lorsque A, B et D sont disjoints et non vides,

$$\begin{aligned} \mathcal{E}_0 \underline{x}^{A,B} \underline{x}^{A,D} &= \frac{1}{3} \frac{1}{n_A}, \\ \text{var}_0 \underline{x}^{A,B} &= \mathcal{E}_0 (\underline{x}^{A,B})^2 = \frac{1}{3} \left(\frac{1}{n_A} + \frac{1}{n_B} + \frac{1}{n_A n_B} \right), \end{aligned}$$

ce qui se déduit aussi de (42) en divisant les deux membres par $n_A n_B$.

Avant de discuter les usages plus amples qu'on peut faire de la sonde de Wilcoxon, nous en voulons introduire quelques autres.

La forme la plus simple du coefficient dit de « corrélation de rang » de M. G. Kendall (1948) (introduit avant lui par R. Greiner (1909) et F. Esscher (1924) comme estimateur du coefficient de corrélation ordinaire d'une distribution binormale) s'écrit, à part d'un facteur constant,

$$\underline{t}_{0,K} = \sum_{i>j}^{i,j} \underline{x}_{ij} = \frac{1}{2} \sum_{i>j}^{i,j} \text{sgn}(i-j) \underline{x}_{ij}. \quad (51)$$

C'est donc aussi une fonction linéaire des signes, avec, selon (37), $\underline{l}_{ij} = \text{sgn}(i-j)$, donc $\underline{l}_{i,j} = 2i - n - 1$ et

$$\begin{aligned} \mathcal{E}_{\underline{t}_{0,K}} &= 0, \quad \text{var}_0 \underline{t}_{0,K} = \frac{1}{6} n(n-1) \\ &\quad + \frac{1}{3} \sum^h (2h - n - 1)^2 = \frac{1}{18} n(n-1)(2n+5) \end{aligned} \quad (52)$$

Cette sonde sert à tester s'il y a un progrès (anglais : « trend », hollandais : « verloop ») (positif ou négatif) dans la suite des x_i .

Ce test a été généralisé par T. J. Terpstra (1952a) pour

le cas où l'on veut savoir s'il y a un progrès dans une suite de groupes d'observations, les observations d'un même groupe étant toujours isomorphes. Lorsque, comme auparavant, \mathcal{I} désigne l'ensemble de toutes les observations, tandis que $A_\lambda (\lambda \in \Lambda = \{1, \dots, k\})$ sont les groupes (disjoints) d'observations, on peut utiliser, en remplaçant la sonde T de Terpstra par ⁽⁴⁾

$$\begin{aligned} t_{0, T} &= 2 T - \left(n^2 - \sum_{\lambda} n_{A_\lambda}^2 \right) \\ t_{0, T} &= \frac{1}{2} \sum_{\lambda, \mu} \text{sgn}(\lambda - \mu) x_{A_\lambda, A_\mu}, \end{aligned} \quad (53)$$

expression qui montre immédiatement l'analogie avec (51). Pour le cas spécial, où les aléatoires ne prennent que deux valeurs et où par conséquent chaque groupe d'observations existe de deux liens, M^{me} van Eeden et M. J. Hemelrijck (1955) ont démontré qu'il est préférable de remplacer la somme (53) par une somme pesée

$$t_{EH} = \frac{1}{2} \sum_{\lambda, \mu} \text{sgn}(\lambda - \mu) \frac{x_{A_\lambda, A_\mu}}{n_{A_\lambda} n_{A_\mu}}$$

c'est-à-dire

$$t_{EH} = \frac{1}{2} \sum_{\lambda, \mu} \text{sgn}(\lambda - \mu) x_{A_\lambda, A_\mu}.$$

Avec cette méthode-ci on obtient que la classe d'hypothèses contre lesquelles le test est consistant ne dépend pas des nombres n_{A_λ} .

Puisque la sonde (53) aussi est linéaire, elle est de la forme (37) avec

$$l_{ij} = \frac{1}{2} \sum_{\lambda, \mu} \text{sgn}(\lambda - \mu) I_{A_\lambda}^i I_{A_\mu}^j = \left(\sum_{\lambda > \mu} \text{sgn}(\lambda - \mu) I_{A_\lambda}^i I_{A_\mu}^j \right). \quad (54)$$

Tandis qu'évidemment $\mathcal{E}t_{0, T} = 0$, on trouve au moyen de (44) avec un peu d'algèbre

$$\text{var}_0 t_{0, T} = \frac{1}{9} \left(n^3 - \sum_{\lambda} n_{A_\lambda}^3 \right) + \frac{1}{6} (n^2 - n_{A_\lambda}^2), \quad (55)$$

résultat qui s'accorde avec celui trouvé par Terpstra. En effet, $t_{0, T}$ est équivalent avec $t_{0, K}$ dans le cas où les x_i entrant dans cette sonde-ci ne sont pas tous différents, mais forment les liens A_λ .

Jusqu'ici nous n'avons considéré que des fonctions linéaires des signes. Or, plusieurs sondes sont des fonctions

⁽⁴⁾ Veuillez lire A_λ et A_μ au lieu de $A\lambda$ et $A\mu$ dans les formules suivantes.

bilinéaires ou quadratiques. Remarquons d'abord que la forme générale de la sonde $t_{0,k}$ de Kendall (1948) s'obtient lorsqu'on remplace les nombres donnés i dans $\text{sgn}(i-j)$ par un second système de nombres aléatoires $\underline{y}_{ij} = \text{sgn}(\underline{y}_i - \underline{y}_j)$

$$\underline{t}_k = \frac{1}{2} \sum \underline{x}_{ij} \underline{y}_{ij}. \quad (56)$$

Ici la classe d'hypothèses admises consiste de toutes les distributions où les n paires $(\underline{x}_1, \underline{y}_1), \dots, (\underline{x}_n, \underline{y}_n)$ sont indépendantes (mais pour chaque i , \underline{x}_i et \underline{y}_i peuvent être dépendantes), et l'hypothèse testée H_0 exprime qu'elles sont isomorphes et que les \underline{y}_i sont indépendants des \underline{x}_i . Dans le cas où les fonctions de répartition communes des \underline{x}_i et des \underline{y}_i restent continues, la valeur de la variance (et naturellement aussi celle de la moyenne) reste la même, donnée par (52). Dans le cas des distributions discontinues on a encore $\mathcal{E}_0 \underline{t}_k = 0$ et la variance a été calculée par M. G. Kendall (1948).

La sonde \underline{t}_k se laisse généraliser pour le cas où \mathcal{J} est la somme de m ensembles disjoints $A_\lambda (\lambda \in \{1, \dots, m\})$, et qu'on sait que les aléatoires appartenant à chaque A_λ sont isomorphes. On pourrait alors remplacer les signes $\underline{x}_{ij}, \underline{y}_{ij}$ dans (56) par leurs généralisations $\underline{x}_{A\lambda, A\mu}; \underline{y}_{A\lambda, A\mu}$, ce qui a été proposé récemment par T. J. Terpstra (1954a), qui utilise donc

$$\underline{t}_{1,T} = \frac{1}{2} \sum^{\lambda, \mu} \underline{x}_{A\lambda, A\mu} \underline{y}_{A\lambda, A\mu}. \quad (57)$$

Ici aussi on soupçonne qu'un dénominateur dans chaque terme comme $n_{A\lambda} n_{A\mu}$ pourrait être utile.

Après ces deux cas de fonctions bilinéaires nous discutons enfin quelques fonctions quadratiques. Pour le problème de k échantillons, indépendamment l'un de l'autre, W. H. Kruskal (1952), T. J. Terpstra (1952b) et (1954b) et P. J. Rijkoort (1952) ont proposé un test. En supposant que λ parcourt les nombres $1, \dots, k$, que les indices des observations du λ^{me} échantillon forment l'ensemble A_λ , de même que les A_λ sont des sous-ensembles disjoints de \mathcal{J} , tel que $\mathcal{J} = \sum A_\lambda$, la sonde de Rijkoort (1952) s'obtient comme suit. Supposons toutes les observations arrangeées selon grandeur croissante, et soit r_i le rang de \underline{x}_i et S_λ la somme des rangs appartenant au λ^{me} échantillon. Alors la sonde de Rijkoort (1952) est

$$\underline{S} = \sum^\lambda \left(S_\lambda - \frac{1}{2} (n+1) n_{A\lambda} \right)^2 \quad (58)$$

Elle se laisse exprimer facilement au moyen des wilcoxoniens.

On trouve pour $4\bar{S}$

$$\bar{t}_R = \sum_{\lambda}^{\lambda} \bar{x}_{A\lambda}^2, \mathcal{J} . \quad (59)$$

La sonde de Kruskal s'écrit

$$\bar{t}_{Kr} = \sum_{\lambda}^{\lambda} \frac{\bar{x}_{A\lambda}^2, \mathcal{J}}{n_{A\lambda}} . \quad (60)$$

Le premier test de Terpstra (1952b) coïncide avec (60), le second est

$$\bar{t}_{2,T} = \sum_{\lambda < \mu} \frac{\bar{x}_{A\lambda}^2, A\mu}{n_{A\lambda} n_{A\mu}} - \frac{1}{n+1} \sum_{\lambda}^{\lambda} \frac{\bar{x}_{A\lambda}^2, \mathcal{J}}{n_{A\lambda}} . \quad (61)$$

Une autre sonde, pour le problème des deux échantillons, introduit récemment par A. Rényi (1954), et servant donc le même but que celui de Wilcoxon (mais étant consistant contre une classe plus étendue d'hypothèses alternatives), est

$$\begin{aligned} \bar{t}_{Re} = & \frac{1}{2 n_A n_B (n_B - 1)} \sum_{i \in A} \bar{x}_{i,B}^2 + \frac{1}{2 n_A n_B (n_A - 1)} \sum_{j \in B} \bar{x}_{A,j}^2 \\ & + \frac{n_B - 2}{2(n_B - 1)} + \frac{n_A - 2}{2(n_A - 1)} . \end{aligned} \quad (62)$$

La sonde de Kendall (1948) (56) pourrait être utilisée pour tester l'hypothèse, impliquant aussi l'hypothèse H_{00} , que les y_i sont indépendants des x_i . Il a été démontré cependant par W. Hoeffding (1948a) que ce test n'est pas consistant contre toutes les hypothèses où les fonctions de répartition des x_i et des y_i sont continues. Un test ayant cette propriété a été développé par W. Hoeffding (1948a), qui utilise la sonde

$$\bar{t}_H = \sum_{(h, i, j, k, l) \neq} (\bar{x}_{hi} - \bar{x}_{hj})(\bar{y}_{hi} - \bar{y}_{hj})(\bar{x}_{hk} - \bar{x}_{hl})(\bar{y}_{hk} - \bar{y}_{hl}) ,$$

fonction biquadratique des signes. La notation $(\alpha_1, \dots, \alpha_n) \neq$ veut dire que $\alpha_i \neq \alpha_j$ dès que $i \neq j$.

Pour la corrélation de rang il existe encore un autre coefficient, dû à C. Spearman (1904), coefficient qui est une fonction linéaire de

$$\bar{t}_S = \sum_i (\bar{x}_{i,J} - \bar{y}_{i,J})^2 ,$$

ce qui, sous l'hypothèse H_{00} (donc aussi sous H_0) est équivalent, sauf une transformation linéaire, avec

$$\underline{t}_s' = \sum_i \underline{x}_{i,\mathcal{J}} \underline{y}_{i,\mathcal{J}}$$

et aussi, pour le cas où $y_i = i$, avec

$$\underline{t}_s'' = \sum_i (i - j) \operatorname{sgn} (x_i - x_j).$$

Une généralisation de cette méthode-ci est la méthode des m groupements (« m rankings ») de M. Friedman (1937). Dans ce cas on a m groupements de la même longueur n . On veut essayer l'hypothèse que pour chaque groupement toutes les permutations des rangs ont la même probabilité. Désignons les observations $x_1^{(1)}, \dots, x_n^{(1)}; x_1^{(2)}, \dots, x_n^{(2)}; \dots; x_1^{(m)}, \dots, x_n^{(m)}$. Alors, la sonde de Friedman s'écrit, à part d'un facteur constant,

$$\underline{t}_F = \sum_1^n \left\{ \sum_1^m \underline{x}_{j,\mathcal{J}}^{(i)} \right\}^2.$$

La méthode de Friedman (1937) se laisse appliquer seulement au cas d'un schéma rectangulaire de cases, avec précisément une observation dans chaque case. La méthode a été généralisée partiellement par J. Durbin (1951), et à peu près complètement par A. Benard et Ph. van Elteren (1953), qui admettent des nombres d'observations arbitraires, assujettis à des restrictions faibles. Leur méthode contient plusieurs autres comme cas spéciaux, par exemple le test de Friedman (1937), la généralisation de Durbin (1951), les tests pour le problème des k échantillons de Kruskal (1952) et de Terpstra (1952b), et le test de Wilcoxon (1945).

La sonde est une fonction quadratique des

$$\underline{u}_j \stackrel{\text{def}}{=} \sum_1^m \underline{x}_{A_{ij}, \mathcal{J}}$$

en supposant que les indices des observations du i^{me} groupement dans la j^{me} case forment l'ensemble A_{ij} . Si le nombre d'observations avec indices appartenant à l'ensemble A_{ij} est désigné par k_{ij} et le nombre d'observations du i^{me} groupement par

$$k_i \left(k_i = \sum_1^n k_{ij} \right),$$

on obtient pour la covariance de \underline{u}_j et \underline{u}_l

$$\sigma_{j,l} = \frac{1}{12} \sum_1^m k_{ij} (\delta_{jl} k_i - k_{il}) (k_i + 1),$$

où δ_{jl} est le symbole de Kronecker défini par (36). Alors la sonde de Benard et van Elteren (1953) est égale à

$$t_{BE} = \begin{vmatrix} \sigma_{11} & \dots & \sigma_{1,n-1} & \underline{u}_1 \\ \vdots & & & \\ \sigma_{n-1,1} & \dots & \sigma_{n-1,n-1} & \underline{u}_{n-1} \\ \underline{u}_1 & \dots & \underline{u}_{n-1} & 0 \end{vmatrix} \cdot \begin{vmatrix} \sigma_{11} & \dots & \sigma_{1,n-1} \\ \vdots & & \\ \sigma_{n-1,1} & \dots & \sigma_{n-1,n-1} \end{vmatrix}^{-1}$$

Puisqu'il est facile de calculer la moyenne et la variance d'une fonction *linéaire* des x_{ij} pour toute la classe Ω d'hypothèses H où les x_i sont indépendantes, isomores ou non, continues ou non, la première méthode à déterminer des bornes d'un ensemble de confiance pour l'hypothèse H , décrite plus haut, s'y laisse toujours appliquer. Quant aux fonctions quadratiques, le calcul de la moyenne dans ce cas général reste encore simple, mais celui de la variance devient déjà un peu compliqué.

L'asymptotique normale a été démontrée dans plusieurs cas par les recherches de W. Hoeffding (1948b), E. Lehmann (1951) et d'autres.

Récemment D. J. Stoker (1955) a examiné une classe de sondes pour le problème des deux échantillons, proposée par D. van Dantzig et J. Hemelrijck (1953). Un cas spécial de cette classe est entre autres la sonde de Wilcoxon (1945). Stoker a démontré que sous des conditions assez générales les sondes de la classe considérée sont asymptotiquement distribuées selon la loi normale. Ici le cas où les fonctions de répartition ne sont pas continues a été recherché spécialement.

De plus Stoker a examiné la déviation de la normalité de la sonde de Wilcoxon. Il a dérivé une borne supérieure pour la déviation extrême de normalité de la fonction de répartition de la sonde sous une hypothèse générale, où les fonctions de répartition des observations peuvent être discontinues. Au moyen de ce résultat-ci la démonstration de la consistance du test de Wilcoxon par D. van Dantzig (1951) a été généralisée pour des fonctions de répartition discontinues. Aussi des ensembles de «surconfiance» ω^* par rapport à la classe des hypothèses alternatives ont été dérivés comme application de la théorie décrite dans D. van Dantzig (1951) et D. van Dantzig et J. Hemelrijck (1953).

Bibliographie

- BENARD, A. and VAN ELTEREN, Ph. (1953), *A generalization of the method of m rankings* (*Indagationes Mathematicae*, **15**, 358-369).
- VAN DANTZIG, D. (1951), *On the consistency and the power of Wilcoxon's two sample test* (*Indagationes Mathematicae*, **13**, 1-8).
- VAN DANTZIG, D. and HEMELRIJK, J. (1953), *Statistical Methods based on few assumptions*, 28th Session, International Statistical Institute, Rome, 6-12 Sept. (*Bulletin de l'Institut International de Statistique*, tome **34**, 2^e livraison, 239-267).
- DURBIN, J. (1951), *Incomplete blocks in ranking experiments* (*British Journal of Psychology*, **4**, 85-90).
- VAN EEDEN, C. and HEMELRIJK, J. (1955), *A test for the equality of probabilities against a class of specified alternative hypotheses, including trend*, Mathematical Centre, Statistical Department, Report S 157 (VP 3) (*Indagationes Mathematicae*, **17**, 191-198, 301-308).
- ESSCHER, F. (1924), *On a method of determining correlation from the ranks of variates* (*Skandinavisk Aktuariertidsskrift*, **7**, 201-219).
- FRIEDMAN, M. (1937), *The use of ranks to avoid the assumption of normality implicit in the analyses of variance* (*Journal of the American Statistical Association*, **32**, 675-699).
- GREINER, R. (1909), *Über das Fehlersystem der Kollektivmasslehre* (*Zeitschrift für Mathematik und Physik*, **57**, 121, 225, 337).
- HOEFFDING, W. (1948a), *A non-parametric test of independence* (*Annals of Mathematical Statistics*, **19**, 546-557).
- (1948b), *A class of statistics with asymptotically normal distribution* (*Annals of Mathematical Statistics*, **19**, 293-325).
- KENDALL, M. G. (1948), *Rank Correlation Methods*, London.
- KRUSKAL, W. H. (1952), *A non-parametric test for the several sample problem* (*Annals of Mathematical Statistics*, **23**, 525-539).
- LEHMANN, E. L. (1951), *Consistency and unbiasedness of certain non-parametric tests* (*Annals of Mathematical Statistics*, **21**, 165-179).
- MANN, H. B. and WHITNEY, D. R. (1947), *On a test whether one of two random variables is stochastically larger than the other* (*Annals of Mathematical Statistics*, **18**, 50-60).
- RIJKOORT, P. J. (1952), *A generalization of Wilcoxon's test* (*Indagationes Mathematicae*, **14**, 394-403).
- RÉNYI, A. (1954), *On the theory of order statistics* (*Proceedings of the International Congress of Mathematicians*, 1954, vol. 1, Amsterdam).
- SPEARMAN, C. (1904), *The proof and measurement of association between two things* (*American Journal of Psychology*, **15**, 88).
- STOKER, D. J. (1955), *Oor 'n klas van toetsingsgrootede vir die probleem van twee stekaproewe*, dissertation, Amsterdam.
- TERPSTRA, T. J. (1952a), *Een parametervrije toets tegen trend voor groepen waarnemingen*, Mathematisch Centrum, Statistische Afdeling, Rapport S 73 (M 28).
- (1952b), *A non-parametric k sample test and its connection with the H-test*, Mathematical Centre, Statistical Department, Report S 92 (VP 2).
- (1954a), *Een generalisatie van Kendall's rangcorrelatietoets*, Mathematisch Centrum, Statistische Afdeling, Rapport S 145 (M 49).
- (1954b), *A non-parametric test for the problem of k samples* (*Indagationes Mathematicae*, **16**, 505-521).
- WILCOXON, F. (1945), *Individual comparisons by ranking methods* (*Biometrics Bulletin*, **1**, 80-83).