Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(3), p. 163-172.

DOI: 10.20982/tqmp.08.3.p163

Un indice général d'association entre deux variables continues
A general non-linear index of association for two continuous variables

Louis Laurencelle
Université du Québec à Trois-Rivières

Measuring and assessing the degree of association between two continuous variables, say *X* and *Y*, has heretofore been restricted by the mandatory specification of a parametric model, be it linear (simple or polynomial), cyclic, autoregressive, or other. We propose as a new quantifying principle the idea that, if a variable *Y* is in some way linked to a variable *X*, values of *Y* immediately neighbouring on *X* should differ less than non-neighbouring ones, so that the "permutative variance" (i.e. variance of successive differences) of the *Y* concomitants of *X* should be low. Two indices, one asymmetrical (*Y* on *X*), the other symmetrical (*Y* cum *X*), are explored and exemplified, and their appropriate critical values, power characteristics and relative merits are established.

Article first published in Lettres Statistiques, *2005, vol. 12, p. 81-97.*

Il existe, pour détecter et quantifier le degré d'association entre deux variables statistiques X et Y, un certain nombre d'indices, le plus connu étant sans doute le coefficient de corrélation linéaire, dit de Bravais-Pearson, dont l'élaboration ainsi que le symbole « *r* » seraient attribuables à Francis Galton et la formule d'estimation à Karl Pearson (Stigler 1986). La « corrélation de rangs » de Spearman (parfois notée ϱ_S) est une proche parente de *r*, puisqu'elle reprend les calculs du *r* à partir des rangs respectifs des valeurs X et Y considérées, tandis que pour le τ (« tau ») de Kendall, la relation monotone entre X et Y est repensée et évaluée sur un mode combinatoire. Ainsi, pour une série bivariée dans laquelle Y varie de façon monotone selon X, les coefficients ρ_S et τ sont parfaits, c'est-à-dire égaux à +1, et le coefficient *r* est lui-même proche de +1. Les autres indices, soit le coefficient de concordance de Kendall, la corrélation point-bisériale, la corrélation bisériale, la corrélation ϕ (« phi »), la corrélation tétrachorique, sont des généralisations ou des cas particuliers de *r*, applicables par exemple lorsque l'une des variables en jeu ou les deux sont réduites à deux valeurs possibles. Le lecteur intéressé peut se documenter sur ces indices dans les manuels de statistique descriptive, par exemple Howell (1998), Siegel et Castellan (1988) et d'autres.

Si nous étudions deux variables continues, par exemple la température ambiante dans une usine (X) et la productivité de ses travailleurs (Y), le coefficient *r* et ses avatars ne mettent à notre disposition qu'un outil qui reflète le degré d'association linéaire, voire monotone, entre celles-ci. L'obtention d'un coefficient *r* élevé, tel *r* = 0,95, nous conforterait dans l'hypothèse d'une association forte et linéaire entre X et Y, mais, d'un autre côté, l'obtention d'un coefficient *r* faible ou nul, tel *r* = 0,05, nous confirmerait seulement qu'il n'y a pas de relation *linéaire* entre les variables, sans nous informer sur la présence possible d'une autre forme de relation.

Il existe en fait d'autres méthodologies statistiques permettant l'estimation d'une relation d'ordre général entre des variables X et Y. Dans le domaine des variables catégorielles, le coefficient de contingence *C*, basé sur la statistique du Khi-deux de Pearson pour un tableau de fréquences à deux dimensions, fait très bien l'affaire, avec ses dérivés tels que l'indice « V » de Cramer. Pour l'étude de variables continues, nous trouvons le « rapport de

Tableau 1: Données prises à 50 moments

Y_1 à Y_{10} =	-0,235	1,027	-0,805	-0,365	1,951	0,825	1,420	1,931	0,488	-1,540
Y_{11} à Y_{20} =	-0,724	-0,148	-1,820	-1,508	-0,530	-1,439	-1,022	0,485	-1,918	1,043
Y_{21} à Y_{30} =	0,019	1,438	1,005	2,774	1,224	-0,154	1,463	0,680	1,491	0,726
Y_{31} à Y_{40} =	-0,853	-1,357	-2,449	-1,009	-2,200	0,468	1,085	-1,183	-0,781	0,603
Y_{41} à Y_{50} =	0,026	-0,656	-0,320	2,875	1,986	1,546	0,840	1,358	1,357	1,286

corrélation », symbolisé η (« êta ») ou η^2, qui équivaut au coefficient de détermination R^2 mais est basé sur un calcul typique de l'analyse de variance. En fait, le coefficient η^2 suppose qu'on catégorise d'abord l'une des variables, disons la variable X jalonnée en *k* catégories, puis qu'on mesure la portion de variance de Y exprimée par la répartition des Y entre les catégories. La distribution de l'indice η^2, dérivé de l'ANOVA, suit celle du quotient *F*. Excellente statistique, mais qui suppose donc une catégorisation arbitraire d'une des variables et ne constitue donc pas un indice d'association général pour deux variables continues.

Les auteurs, pourtant, se sont penchés sur la question, mais les propositions faites ont concerné les fonctions de probabilité bivariées plutôt que les séries statistiques comme telles. Steffensen, en 1941, s'inspirant d'une mesure antérieure conçue pour variables discrètes, propose un indice d'association général « ω » pour une fonction $f(X,Y)$ en variables continues, la mesure étant basée sur la différence « $f(X,Y) - f(X) \times f(Y)$ »; Silvey réédite cette proposition sous une forme comparable en 1964. Scarsini et Venetoulias (1993) élaborent des fonctions bivariées continues à relations de dépendance non linéaires, sans se soucier toutefois d'une mesure d'association globale entre les variables. Enfin, dans une approche sans doute inspirée de Tukey (1977), Fisher et Switzer (1985) avancent une méthodologie graphique destinée à révéler la dépendance mutuelle entre X et Y dans une série statistique. Le procédé, laborieux et assez complexe, est encore basé sur l'égalité théorématique « $f(X,Y) = f(X) \times f(Y)$ » et revient à convertir le graphe $\{X_i,Y_i\}$ original en un autre ; l'indépendance entre X et Y apparaît alors comme un arrangement rectangulaire (plutôt que sphérique) des points (X_i,Y_i), et toute forme de dépendance (linéaire ou non) se répercute dans un arrangement débordant. Le graphe se construit au moyen de

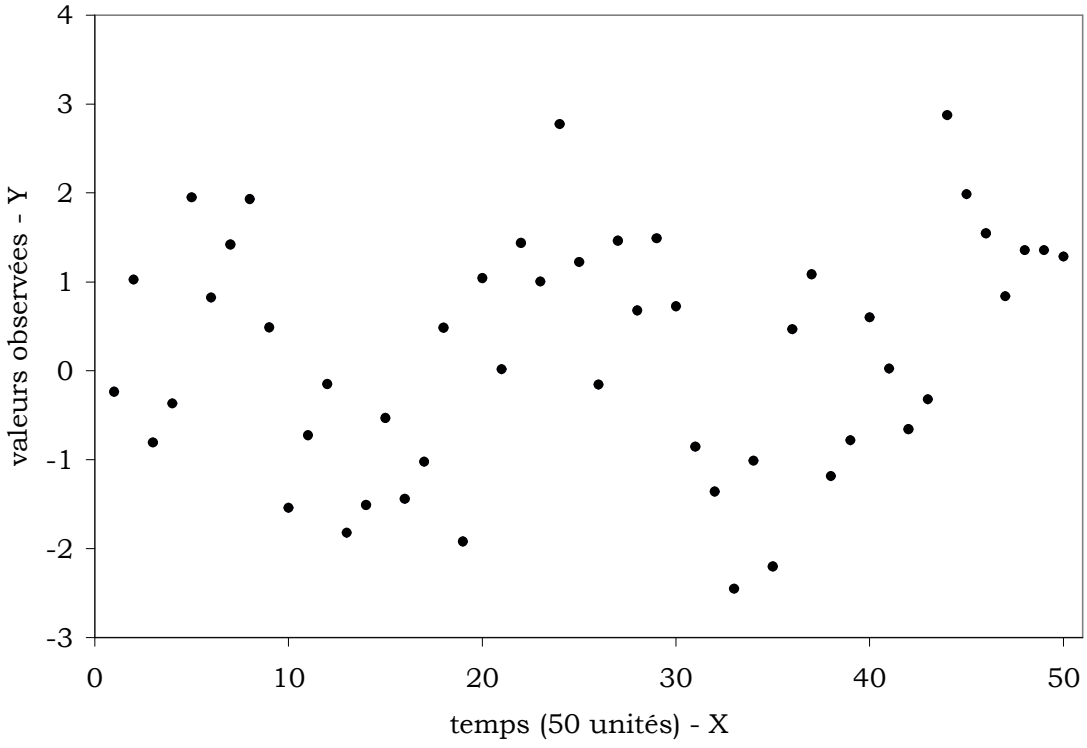

Figure 1. Échantillon d'un modèle sinusoïdal de *n* = 50 valeurs, bruitées à 50 % (R^2 = 0,50)

transformations non linéaires des données. Les auteurs ne proposent pas d'indice global reflétant la non-indépendance de la série $\{X_i, Y_i\}$.

L'indice "A" et un exemple

Prenons l'exemple, fictif, d'un paramètre tel que température ou humidité, échantillonné à intervalle régulier pour une certaine période. Le tableau 1 (en page 35) fournit les 50 mesures successives Y_i obtenues ; les données de temps X_i, elles, sont équidistantes (p. ex. 1, 2, 3, ..., 50).

La figure 1 fait voir la variation du paramètre (Y) en fonction du temps (X).

Le graphe présenté à la figure 1 nous conduit à au moins trois constatations : 1) la relation de Y à X, si elle existe, est bruyante, stochastique ; (2) il semble y avoir une relation cyclique, de type sinusoïdal, entre Y et X, et (3) on ne perçoit pas de tendance monotone évidente, ni à l'accroissement, ni à la diminution. Cette dernière constatation se confirme par le calcul du coefficient *r*, qui vaut ici 0,160 et n'est pas significatif au seuil bilatéral de 5 %.

Les valeurs X étant ici naturellement ordonnées (en ordre croissant), on remarque que les valeurs de Y forment un tracé approximatif, une ligne de fonction, dont l'existence supposée agit en rapprochant les Y_i successifs l'un de l'autre; or, s'il n'existait aucune dépendance entre Y et X, les différences observées entre les valeurs Y_i successives seraient statistiquement du même ordre que les ${}_nC_2$ différences possibles parmi les *n* valeurs Y_i de la série. Ainsi, la dépendance de Y sur X se reflète par une réduction systématique des différences successives.

Les indices A et A_S. De même que la variance, s^2, mesure l'importance des différences entre les mesures d'une série statistique, la variance permutative, *p*, évalue quant à elle l'importance des différences successives, selon :

$$p_Y = \frac{\sum_{i=1}^{n-1}(Y_i - Y_{i+1})^2}{2(n-1)} \qquad (1)$$

Laurencelle (1983) examine cette statistique peu connue, dont la première étude systématique remonterait à Von Neumann et collaborateurs (1941). L'auteur rapporte aussi les moments simples et composés suivants, applicables pour des échantillons aléatoires normaux de *n* variables :

$$\mathrm{E}\{\ p_Y\ \} = \sigma^2 \qquad (2a)$$

$$\mathrm{var}\{\ p_Y\ \} = \frac{3n-4}{(n-1)^2}\sigma^4 \qquad (2b)$$

$$\mathrm{E}\{\ p_Y / s^2\ \} = 1 \qquad (3a)$$

$$\mathrm{var}\{\ p_Y / s^2\ \} = \frac{n-2}{(n-1)(n+1)} \qquad (3b)$$

Dans une série bivariée $\{ X_i, Y_i \}$, nous définissons la variance permutative conditionnelle de Y sur X, symbolisée $p_{Y.X}$, comme la variance permutative des concomitantes $Y_{[i]}$, après le tri en ordre croissant des X_i ; nous aurons aussi la variance permutative conditionnelle des X sur Y, notée $p_{X.Y}$. Dans ce contexte, l'indice d'association général proposé est :

$$\mathrm{A} = \frac{p_{Y.X}}{s_Y^2} \ . \qquad (4)$$

Sous la condition d'indépendance statistique des X et Y, l'indice A possède les deux premiers moments indiqués ci-dessus ainsi qu'une distribution de probabilité (voir plus bas) tendant vers la distribution normale pour *n* élevés. Une dépendance stochastique quelconque entre X et Y aura pour effet de *réduire* $p_{Y.X}$ et donc d'abaisser l'indice A vers zéro.

D'autre part il arrive en sciences humaines comme en général, que le chercheur soit placé devant des données (X_i, Y_i) logiquement symétriques, c.-à-d. pour lesquelles aucune des deux variables ne joue un rôle générateur ou déterminant. Pour exemples, prenons le cas de la covariation des mesures de glycémie et de température corporelle chez des patients, des scores de motivation et de réussite scolaires, du rendement per capita d'une usine et du salaire des employés, etc. Pour des cas semblables, le coefficient de corrélation *r*, qui est naturellement symétrique, n'oblige pas le chercheur à décider, pour son calcul, du rôle possible, causal ou générateur, des variables en jeu. Nous proposons donc aussi un indice A symétrisé, symbolisé "A_S", et obtenu par :

$$\mathrm{A_S} = \sqrt{\frac{p_{X.Y}\,p_{Y.X}}{s_X^2 s_Y^2}} \ . \qquad (5)$$

Exemple de calcul de l'indice A. Les données présentées en tableau ci-dessus et illustrées à la figure 1 répondent au critère d'une relation asymétrique, une quantité mesurée étant échantillonnée en fonction du temps, et s'adressent donc à l'indice A. La variance des $n = 50$ mesures Y est $s_Y^2 \approx 1{,}7167$. Pour la variance permutative, d'après (1), nous sommons d'abord les carrés des différences successives, $(-0{,}235 - 1{,}027)^2 + (1{,}027 - (-0{,}805))^2 + \ldots + (1{,}357 - 1{,}286)^2 \approx 91{,}4159$, puis $p_Y = 91{,}459 / (2 \times 49) \ . \ 0{,}9328$. Enfin, $A = p_Y / s_Y^2 \approx 0{,}543$.

Nous verrons plus bas que cette valeur, A = 0,543 basée sur $n = 50$, déborde (par le bas) le seuil de signification à 0,01, la valeur critique appropriée étant d'environ 0,681.

Exemple de calcul de l'indice A_S. Notre second exemple fictif est constitué par un échantillon de $n = 15$ données $\{ X_i, Y_i \}$, présentées dans la partie gauche du tableau 2.

Tableau 2: 15 données fictives

X_i	Y_i	$X_{(i)}$	$Y_{[i]}$	$X_{[i]}$	$Y_{(i)}$
1,137	0,101	0,048	0,884	1,135	-0,048
0,824	0,066	0,081	0,786	1,492	-0,043
0,782	0,068	0,249	0,571	1,086	0,014
0,613	0,099	0,596	0,190	0,824	0,066
1,014	0,101	0,613	0,099	0,782	0,068
0,249	0,571	0,634	0,080	1,209	0,078
1,492	-0,043	0,782	0,068	0,634	0,080
1,209	0,078	0,824	0,066	0,889	0,082
0,634	0,080	0,889	0,082	0,613	0,099
0,596	0,190	1,014	0,101	1,137	0,101
0,048	0,884	1,086	0,014	1,014	0,101
1,135	-0,048	1,135	-0,048	0,596	0,190
1,086	0,014	1,137	0,101	0,249	0,571
0,081	0,786	1,209	0,078	0,081	0,786
0,889	0,082	1,492	-0,043	0,048	0,884

On obtient tout de go les quantités $s_X^2 \approx 0{,}17831$ et $s_Y^2 \approx 0{,}08660$. Pour établir $p_{Y.X}$, la variance permutative des Y conditionnée sur les X, il s'agit de trier (en ordre croissant) la série des X, obtenant ainsi les statistiques d'ordre des X, soit les $X_{(i)}$, ainsi que leurs concomitantes Y, les $Y_{[i]}$ soit ces deux séries apparaissent dans les colonnes centrales du tableau. On calcule alors les différences successives des Y, soit $(0{,}884 - 0{,}786)^2 + (0{,}786 - 0{,}571)^2 +$ etc., obtenant enfin $p_{Y.X} \approx 0{,}00924$. Le tri des données selon Y, puis la sommation des différences successives, tel qu'illustré dans les colonnes de

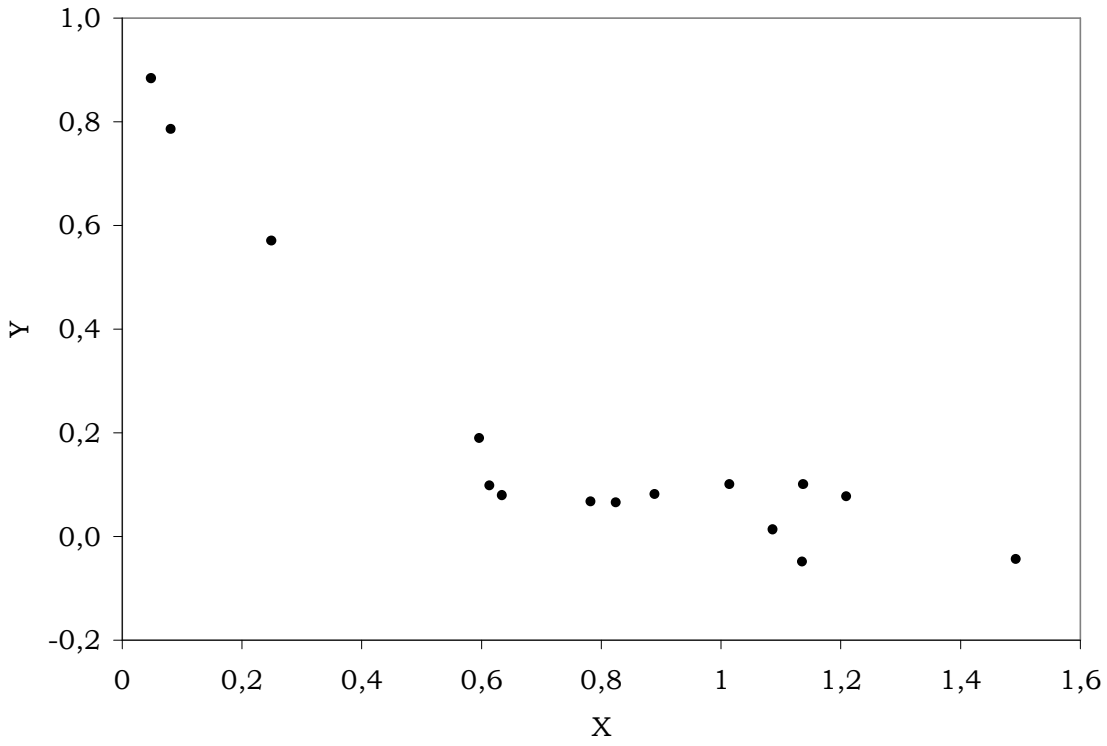

Figure 2. Graphique d'un échantillon de $n = 15$ données $X_i : Y_i$

droite, fournissent à leur tour $p_{X.Y} \approx 0{,}05833$. Finalement, en appliquant la formule (5), nous obtenons $A_S = [0{,}00924 \times 0{,}05833 / 0{,}17831 \times 0{,}08660]^{1/2} \approx 0{,}187$: la valeur critique (approximative) au seuil de 1 % étant de 0,544, nous enregistrons ici un indice qui confirme la présence d'une relation régulière entre X et Y. La figure 2 illustre cette relation entre X et Y, relation qui pourrait être de type exponentiel ou logarithmique : le modèle linéaire ne convient probablement pas. Toutefois, si calculé, le coefficient de corrélation donne ici la valeur $r \approx -0{,}837$, significative elle aussi au seuil bilatéral de 1 %.

Distribution et moments des indices A et As

La statistique A. L'indice A a été étudié sous diverses dénotations par queques auteurs, notamment Von Neumann (1941) avec l'indice η, équivalant à $A \times 2n/(n-1)$, et Young (1941) avec l'indice C, équivalent à 1 - A. Aux deux premiers moments fournis plus haut, en (3a) et (3b), Young (1941) ajoute les deux suivants :

$$\gamma_1 = 0 \qquad (6a)$$

$$\gamma_2 = \frac{-6(n^3 - 13n + 24)}{(n-2)^2(n+3)(n+5)}. \qquad (6b)$$

La distribution apparaît donc symétrique avec un coefficient de voussure négatif, presque égal à $-6/(n+4)$.

Young (1941) exploite les moments trouvés pour établir une approximation par une distribution de Pearson de type II (cf. Kendall et Stuart 1977) ; Von Neumann (1941) propose aussi une méthode de calcul, que Hart (1942) utilise afin de présenter des valeurs critiques. Nous proposons à notre tour une approximation par une loi *Bêta* symétrique, $\beta(p,p)$, le paramètre p étant fixé par le coefficient γ_2 et valant à peu près ½ (n+1). Soit b, une variable *Bêta* ainsi déterminée ; alors la quantité A lui est reliée par :

$$A \approx 1 + (2b-1)\sqrt{\frac{n^2-4}{n^2-1}} \qquad (7)$$

la forme et les quantiles de ce modèle *Bêta* épousent étroitement ceux des distributions Monte Carlo de A correspondantes.

La statistique As. La forme symétrisée du quotient de variance permutative, que nous dénotons A_S, n'est pas répertoriée dans la documentation. Notre étude, à partir d'un lourd échantillonnage Monte Carlo sur des séries binormales, fournit les indications suivantes.

1. L'espérance est légèrement déficiente. En fait, $E\{ (A_S)^2 \}$ déborde 1, l'excès allant en diminuant selon n, alors que $E\{ A_S \} < 1$, la valeur rejoignant doucement l'asymptote. La fonction suivante est descriptive :

$$A_S \approx 0{,}98 + 0{,}02 \times [(n-2)/(n-1)]^{13{,}5}.$$

L'examen attentif montre que les facteurs apparaissant au numérateur de A_S, $p_{X.Y}$ et $p_{Y.X}$, sont en corrélation positive, ce qui explique sans doute le biais de l'espérance.

Pour des échantillons de séries de n = 3, 4 et 5 données binormales, les valeurs de corrélation trouvées sont 0,068, 0,082 et 0,081 ; cette faible corrélation persiste jusqu'à des longueurs n de 30, voire de 50. Par curiosité, nous avons étudié les distributions permutationnelles complètes des données rectangulaires X_i, Y_i = 1, 2, 3, ..., n (en exploitant leurs n! permutations). Pour n = 3, 4 et 5, nous obtenons maintenant ϱ = 1, 0,702 et 0,538, la corrélation étant encore à 0,240 pour n = 10.

L'explication de cette corrélation positive tient, croyons-nous, à la dépendance mutuelle des variances permutatives $p_{X.Y}$ et $p_{Y.X}$. En effet, dans toutes les permutations possibles des séries X et Y, il en existe deux pour lesquelles l'ordre des deux séries concorde (directement ou inversement) : dans ce cas, le tri de X va produire en même temps le tri de Y, et les quantités $p_{Y.X}$ et $p_{X.Y}$ seront conjointement minimales ; le cas des permutations contenant de petites perturbations de l'ordre parfait tombe aussi dans cet argument. Bien entendu, ces permutations qui réduisent conjointement $p_{Y.X}$ et $p_{X.Y}$ deviennent peu à peu minoritaires parmi les n! permutations qui contribuent au calcul des espérances, de sorte que corrélation et biais finissent par s'évanouir.

2. La variance est approximativement égale à $1 / 2n$.

3. Une légère asymétrie négative est présente, démarrant à $\gamma_1 \approx -0{,}20$ pour n = 3, et qui tombe sous -0,10 à n = 20 et sous -0,03 à n = 50.

4. Il y a une légère platykurtose, sauf à n = 3 ($\gamma_2 \approx -1{,}41$), inférieure à -0,30 pour n = 4, et tombant sous -0,1 à n = 10 et sous -0,03 à n = 50.

Nous n'avons pas tenté de modéliser cette distribution.

Le test de corrélation générale et les valeurs critiques

Les considérations ci-dessus de même que les études Monte Carlo que nous avons menées permettent de proposer un test statistique qui aide le chercheur à décider s'il y a ou non un lien de dépendance quelconque entre ses mesures X et Y.

Le tableau 3, à la page suivante, présente des listes de valeurs critiques pour des tests unilatéraux aux seuils de 0,05 et 0,01, pour les indices A et A_S. La valeur obtenue est réputée significative si elle est *inférieure* ou égale à la valeur critique : ce cas indique que les valeurs X et Y ont tendance à évoluer conjointement, et qu'il est opportun et intéressant de chercher un modèle décrivant leur dépendance réciproque.

Les valeurs données pour la statistique asymétrique A découlent du modèle de distribution *Bêta* présenté plus haut; font exception les séries de n = 3 et n = 4 données, pour lesquelles les centiles Monte Carlo, basés sur 10^6

Tableau 3. Valeurs critiques des indices A et As

N	Indice A		Indice As	
	P 0,05	0,01	0,05	0,01
3	0,506	0,500	0,526	0,505
4	0,390	0,313	0,402	0,333
5	0,419	0,262	0,446	0,287
6	0,443	0,283	0,484	0,313
7	0,468	0,307	0,518	0,346
8	0,491	0,331	0,551	0,379
9	0,512	0,354	0,578	0,408
10	0,531	0,376	0,601	0,436
11	0,548	0,396	0,621	0,463
12	0,564	0,414	0,638	0,486
13	0,578	0,431	0,652	0,508
14	0,591	0,447	0,665	0,527
15	0,603	0,461	0,677	0,544
16	0,614	0,475	0,688	0,559
17	0,624	0,487	0,698	0,573
18	0,633	0,499	0,706	0,586
19	0,642	0,510	0,715	0,597
20	0,650	0,520	0,722	0,609
21	0,657	0,530	0,729	0,619
22	0,665	0,539	0,736	0,628
23	0,671	0,548	0,743	0,638
24	0,678	0,556	0,748	0,645
25	0,684	0,564	0,753	0,653
26	0,689	0,571	0,759	0,659
27	0,695	0,578	0,763	0,667
28	0,700	0,585	0,768	0,673
29	0,705	0,591	0,771	0,678
30	0,709	0,598	0,776	0,685
31	0,714	0,603	0,779	0,689
32	0,718	0,609	0,784	0,695
33	0,722	0,614	0,787	0,700
34	0,726	0,620	0,790	0,705
35	0,729	0,624	0,793	0,709
36	0,733	0,629	0,797	0,715
37	0,736	0,634	0,799	0,718
38	0,740	0,638	0,802	0,722
39	0,743	0,643	0,805	0,726
40	0,746	0,647	0,807	0,730
41	0,749	0,651	0,810	0,733
42	0,752	0,655	0,813	0,737
43	0,755	0,658	0,815	0,739
44	0,757	0,662	0,816	0,743
45	0,760	0,665	0,819	0,746
46	0,762	0,669	0,821	0,749
47	0,765	0,672	0,823	0,752
48	0,767	0,675	0,825	0,754
49	0,770	0,678	0,827	0,757
50	0,772	0,681	0,828	0,759

échantillons, sont fournis : ces valeurs correspondent à celles dans Hart (1942), mutatis mutandis. Au-delà de n = 50, l'approximation normale convient très bien, selon :

$$A_{[P]} = 1 + z_{[p]}\sqrt{\frac{n-2}{(n+1)(n-1)}} \qquad (8)$$

Par exemple, pour n = 50 et P = 0,05, nous avons $z_{[0,05]}$ = -1,6449 et $A_{[0,05]} \approx 0{,}772$.

Pour l'indice symétrisé As, en l'absence d'un modèle de distribution, nous présentons les valeurs critiques obtenues par simulation Monte Carlo (basée sur 10^6 échantillons). L'approximation normale est ici aussi possible mais elle se révèle peu précise. Pour obtenir les valeurs critiques au-delà de n = 50, nous avons mis au point les fonctions d'approximation suivantes :

$$A_{S[0,05]} = 1 - \frac{1,48}{(n+1)^{0,5469}} \qquad (9a)$$

et :

$$A_{S[0,01]} = 1 - \frac{2,0671}{(n+1)^{0,5456}}, \qquad (9b)$$

lesquelles sont précises à ± 0,001 à partir de n = 50.

Études sur la puissance des indices A et As

On ne trouve pas de compétiteurs, à proprement parler, pour nos indices de corrélation générale A et As, de sorte qu'il est impossible de leur rendre justice par le moyen habituel, c.-à-d. par des études comparatives de puissance, voire des études de puissance relative. Nous nous sommes replié sur une autre approche, qui consiste à illustrer la puissance de ces indices comme outils de détection d'un lien de dépendance entre les séries X et Y : nous montrerons à l'occasion le rendement d'autres indices, tel le simple r de Bravais-Pearson. L'illustration se fait en trois exemples : le modèle linéaire simple (du premier degré), le modèle quadratique (ou polynomial du second degré), le modèle sinusoïdal.

Le modèle linéaire : $Y' = b\,X + \varepsilon$

La dépendance linéaire entre X et Y constitue la forme la plus simple et la plus connue de dépendance, et une pour laquelle le coefficient de corrélation r a été conçu. L'étude de puissance a consisté à générer des séries binormales de n

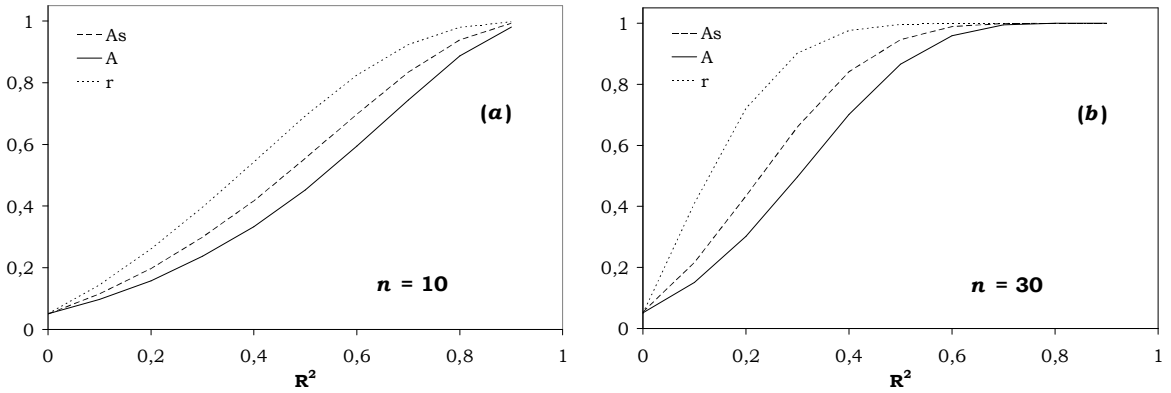

Figure 3. Puissance des indices As, A et *r* (bilatéral) pour un modèle linéaire simple au seuil de signification de 0,05 en fonction de R^2 (= ϱ^2), pour n = 10 (*a*) et 30 (*b*).

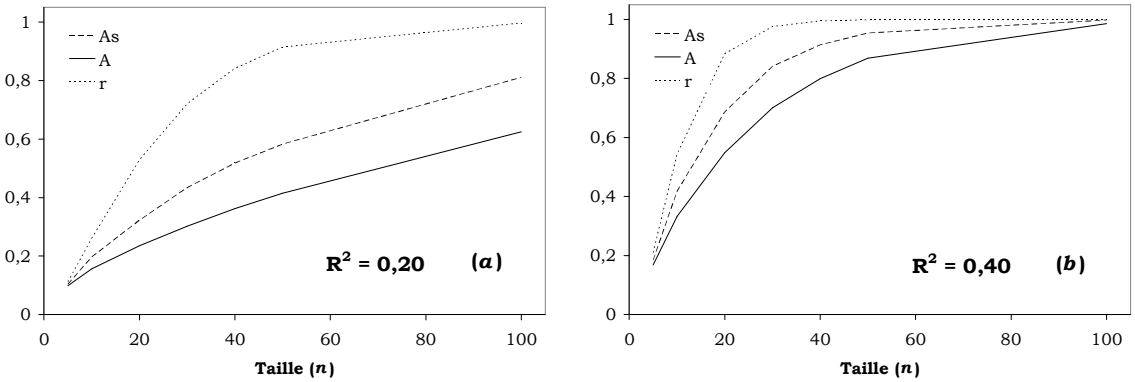

Figure 4. Puissance des indices As, A et *r* (bilatéral) pour un modèle linéaire simple au seuil de signification de 0,05 en fonction de n, pour R^2 (= ϱ^2) = 0,20 (*a*) et 0,40 (*b*).

couples (X_i, Y_i) à lien de dépendance linéaire ρ ($\rho > 0$). La technique de génération est indiquée par l'algorithme suivant, z_1 et z_2 étant des variables aléatoires de distribution normale standard :

Soit i variant de 1 à n : (10)

Obtenir $X_i \leftarrow z_1$;

Produire $Y_i \leftarrow \rho\, z_1 + \sqrt{1-\rho^2}\, z_2$.

Nous avons croisé différentes tailles n et valeurs de R^2 (= ϱ^2), en générant chaque fois 10^5 échantillons Monte Carlo, et en vérifiant la proportion de cas fournissant des valeurs significatives des indices comparés : cette proportion constitue notre statistique de puissance.

Les graphiques suivants montrent les variations comparatives de puissance de nos indices As, A et de *r*, soit selon R^2 (figure 3) ou selon n (figure 4), ce au seuil de 0,05.

Le coefficient de corrélation *r*, conçu spécifiquement pour ce modèle, se montre facilement supérieur en puissance, ce davantage pour des relations plus bruitées (à R^2 faible) que moins bruitées (à R^2 fort). De plus, dans ce modèle de dépendance linéaire simple, l'indice symétrisé As se révèle systématiquement plus puissant, quoique de peu, que l'indice A.

Le modèle quadratique : $Y' = b_1 X + b_2 X^2/\sqrt{2} + \varepsilon$

Le modèle polynomial du deuxième degré constitue un bel exemple d'un lien de dépendance non linéaire. Pour autant que X et ε soient des variables normales standards, les trois composantes X, $X^2/\sqrt{2}$ et ε ont chacune une variance unitaire et des covariances nulles ; par conséquent, le lien de dépendance non linéaire global (R^2) entre X et Y est proportionnel à la somme $b_1^2 + b_2^2$, et la variance expliquée peut alors être équitablement partagée en une part linéaire et une part quadratique. L'algorithme suivant détaille la technique de génération, qui exploite deux paramètres de dépendance : le lien global (R^2) et la « courbure » (part_b_2) indiquée par la portion de variance associée au paramètre b_2:

Soit R^2, le lien de dépendance globale entre X et Y ; (11)

part_b_2, la portion de variance du lien entre X et Y affectée au coefficient b_2. Par exemple, si R^2 égale 0,50 et part_b_2 est de 0,60, alors $b_2 \leftarrow \sqrt{[\text{part\_}b_2 \times R^2/(1-R^2)]}$ et $b_1 \leftarrow \sqrt{[R^2/(1-R^2) - b_2^2]}$;

i variant de 1 à n :

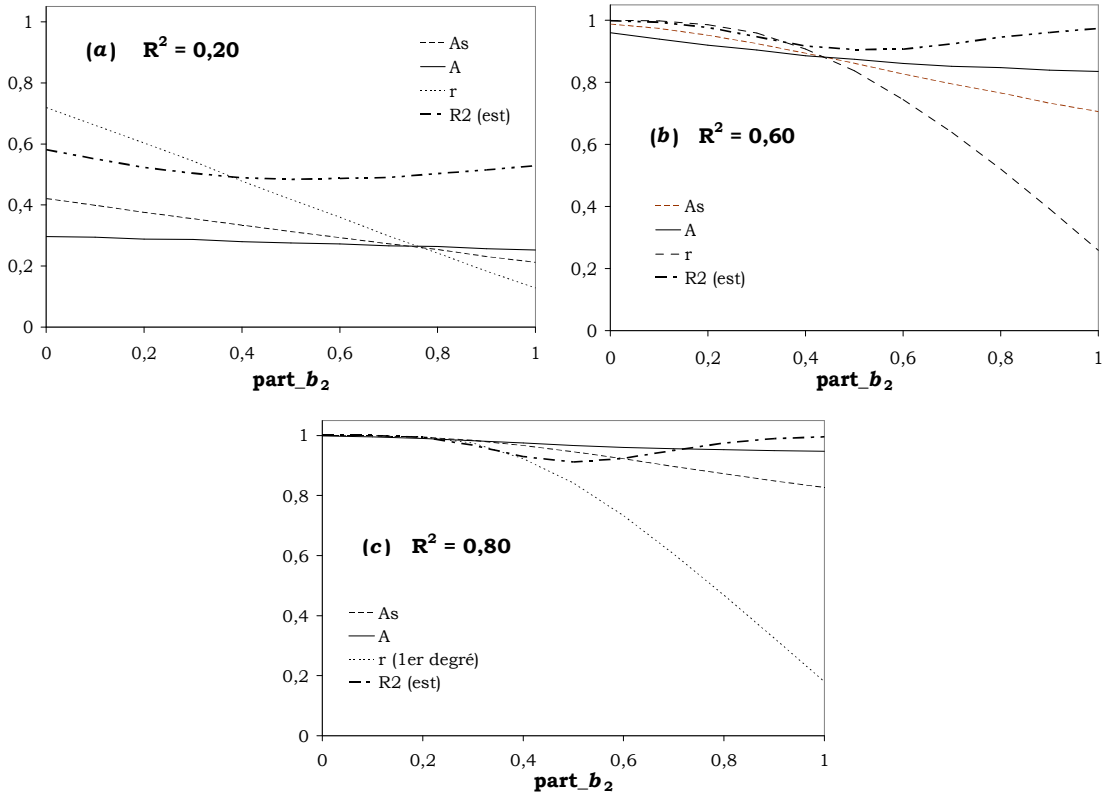

Figure 5. Puissance des indices As, A, *r* et R2_est (associé au modèle de régression polynomial de degré 2) pour un modèle quadratique au seuil de signification de 0,05, pour $n = 30$ et des valeurs de R^2 de 0,20 (*a*), 0,60 (*b*) et 0,80 (*c*).

Obtenir $X_i \leftarrow z_1$;
Produire $Y_i \leftarrow b_1 X_i + b_2 X_i^2/\sqrt{2} + z_2$.

Différentes combinaisons de tailles (n) et de forces de lien (R^2) ont donné lieu à des échantillonnages Monte Carlo, chaque fois avec 10^5 échantillons. Nous en retenons trois, dont les résultats sont présentés à la figure 5, soit une taille de $n = 30$ paires X : Y avec $R^2 = 0{,}20$ (fig. 5a), $R^2 = 0{,}60$ (fig. 5b) et $R^2 = 0{,}80$ (fig. 5c). En plus des indices A et As et du coefficient linéaire *r*, nous avons aussi obtenu le coefficient de détermination estimé (R2 est) associé à la régression polynomiale de degré 2 ajustée par les moindres carrés, et dont le test de signification correspond ici à une distribution $F_{2\,;\,27}$: dans tous les cas, la puissance est estimée en regard d'un seuil de signification de 0,05.

Dans la portion gauche des graphiques, là où la fonction de dépendance est purement ou surtout linéaire (avec « part_b_2 » ≤ 0,4), le coefficient linéaire *r* domine en puissance mais il cède ensuite la place au modèle polynomial de degré 2 (R2 est) lorsque la courbure de la fonction s'accentue. Évidemment, si le chercheur sait d'avance ou constate qu'il a affaire à une fonction polynomiale de degré 2, il optera d'emblée pour le modèle correspondant et bénéficiera ainsi d'une puissance quasi optimale.

Quant aux indices généraux A et As, ils s'en tirent assez bien, et notamment l'indice asymétrique A qui tend vers la puissance maximale disponible lorsque R^2 augmente, comme le montre la figure 5c.

Le modèle sinusoïdal : $Y' = b_s K \sin(2{,}5\,x) + \varepsilon$, $x = 0 \,..\, 2\pi$

Comme dernier modèle de dépendance, nous avons choisi un modèle sinusoïdal, dénoté par la fonction « sin(2,5 *x*) », plus ou moins bruité par l'adjonction d'une variable normale standard indépendante (ε). En fait, dans le domaine $(0 \,..\, 2\pi)$ de *x*, la variance propre de "sin(2,5*x*)" est de K^{-2}, où $K = 5\pi\sqrt{2} / \sqrt{[(5\pi)^2 - 8]}$; ainsi, la fonction « $K \cdot \sin(2{,}5x)$ » a une variance unitaire, et la force du lien de dépendance dans le modèle bruité est encore mesurée par $R^2 = b_s^2 / (b_s^2 + 1)$. L'algorithme suivant a servi à produire les séries de données:

Soit R^2 et $b_s \leftarrow R / \sqrt{(1 - R^2)}$; (12)
i variant de 1 à *n* :
Définir $X_i = x \leftarrow (i - 0{,}5)/n \times 2\pi$;
Produire $Y_i \leftarrow b_s K \cdot \sin(2{,}5x) + z$.

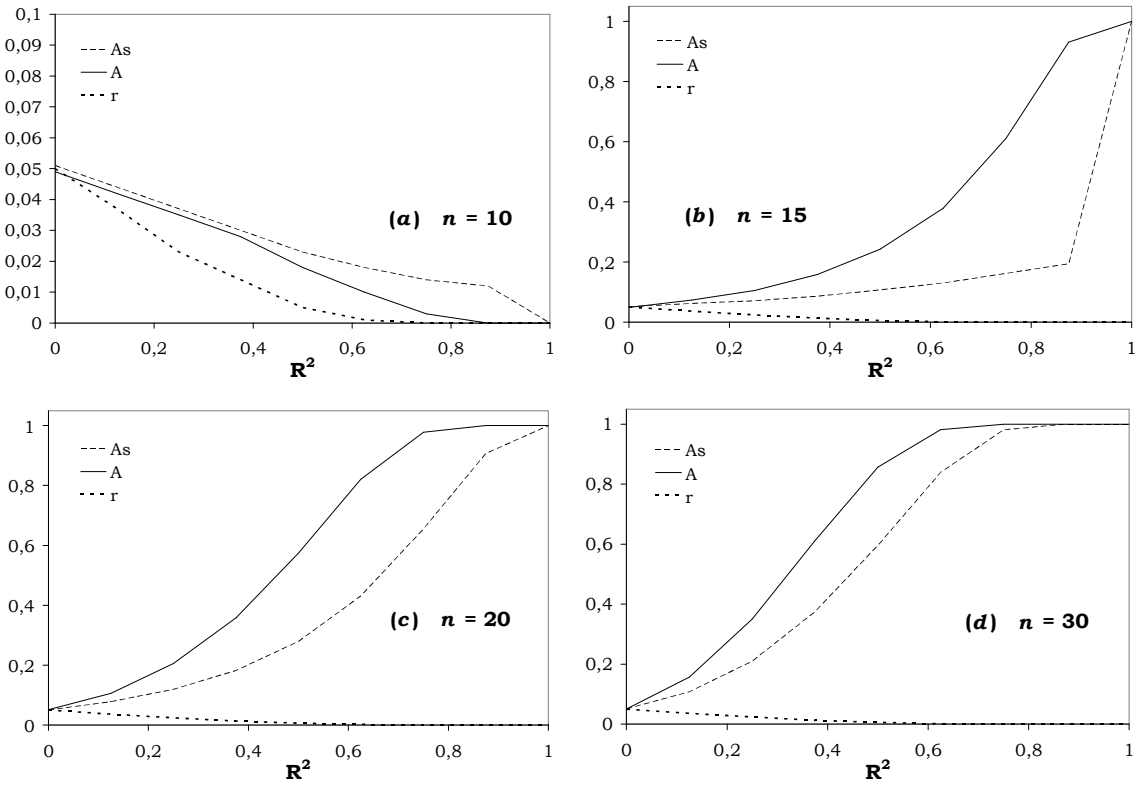

Figure 6. Puissance des indices As, A et *r* (bilatéral) pour un modèle sinusoïdal au seuil de signification de 0 ,05 en fonction de R^2, pour *n* = 10 (*a*), 15 (*b*), 20 (*c*) et 30 (*d*).

Le tracé de la sinusoïde, qui comprend ici deux cycles et demi, réclame un minimum de points pour la dessiner : la figure 1 illustre un tel modèle étalé sur *n* = 50 points, avec R^2 = 0,50. Les graphiques de la figure 6, illustrant la puissance des indices A_S, A et *r* au seuil de signification de 0,05, correspondent à des séries de *n* = 10, 15, 20 et 30 points respectivement.

La figure 6*a* permet de constater que 10 points ne suffisent pas à traquer une sinusoïde de 2,5 cycles, ce d'autant moins que R^2 augmente et que la sinusoïde est moins bruitée. Dès *n* = 15 cependant (fig. 6b), les tests As et A y parviennent, avec un net avantage pour l'indice asymétrique A qui domine partout (fig. 6c et 6d). Le coefficient *r* ne détecte pas du tout la dépendance cyclique, tel qu'attendu.

Discussion et conclusion

Au contraire du coefficient *r*, qui invoque un modèle linéaire (ou proportionnel), ou par exemple du coefficient de détermination attaché à une régression polynomiale d'un degré donné, les indices As et A n'invoquent aucun modèle ; leur calcul suppose seulement que, s'il y a dépendance entre X et Y, cette dépendance aura pour effet de rapprocher les concomitantes réciproques des séries { X_i } et { Y_i }, entraînant une réduction de leurs variances permutatives conditionnelles et l'abaissement de l'indice. On peut donc parler ici d'indices de corrélation non linéaires, au sens fort.

Les indices A et As ne sont pas des compétiteurs du coefficient de corrélation linéaire *r*, pour diverses raisons. D'abord, dans plusieurs cas évidents, le modèle linéaire reste le seul modèle raisonnable, et son calcul de même que la puissance de l'indice *r* en rendent incontournable l'utilisation. Nos études de puissance ont néanmoins montré que le coefficient de corrélation *r* n'est pas vraiment un indice universel, et qu'il s'accommode mal à la fois de la curvilinéarité des fonctions ou de leur cyclicité : en fait, le coefficient *r* se révèle plus ou moins aveugle dans ces cas. Par contraste, comme le montrent nos figures 3 à 6, les indices As et A gagnent en puissance absolue et relative à mesure que le modèle est mieux défini, ce *quel que soit le modèle*. Cet avantage particulier nous conduit à proposer au chercheur la démarche suivante, lorsque confronté à une série { X_i, Y_i } dont il veut déterminer le lien de dépendance : Si le coefficient *r* est significatif, vérifier sur graphique la pertinence du modèle linéaire (versus un autre modèle monotone, non linéaire) ou, si le coefficient *r* s'avère non significatif, vérifier la présence d'un lien non monotone par le test des indices A et As et, le cas échéant, s'engager dans la recherche d'un modèle adéquat.

Références

Fisher, N. I., Switzer, P. (1985). Chi-plots for assessing dependence. *Biometrika*, *72*, 253-265.

Hart, B. I. (1942). Significance levels for the ratio of the mean square successive difference to the variance. *Annals of mathematical statistics*, *13*, 445-447.

Howell, D. C. (1998). *Méthodes statistiques en sciences humaines*. Paris : De Boeck Université.

Kendall, M. G., Stuart, A. (1977). *The advanced theory of statistics. Volume 1 : Distribution theory*. New York : Macmillan.

Laurencelle, L. (1983). La variance permutative. *Lettres Statistiques*, *7*, 22 p.

Mari, D. D., Kotz, S. (2001). *Correlation and dependence*. Singapour : Imperial College Press.

Scarsini, M., Venetoulias, A. (1993). Bivariate distributions with nonmonotone dependence structure. *Journal of the American Statistical Association, 88*, 338-344.

Siegel, S., Castellan, N. J. Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2e édition). New York, McGraw-Hill.

Silvey, S. D. (1964). On a measure of association. *Annals of mathematical statistics*, *35*, 1157-1166.

Steffensen, J. F. (1941). On the *w* test of dependence between statistical variables. *Skandinavisk aktuarietidskrift*, *24*, 13-33.

Stigler, S. M. (1986). *The history of statistics. The measurement of uncertainty before 1900*. Cambridge (Mass) ; Belknap Press.

Tukey, J. W. (1977) . *Exploratory data analysis.* Reading (Mass) : Addison-Wesley.

Von Neumann, J., Kent, R. H., Bellinson, H. R., Hart, B. I. (1941). The mean square successive difference. *Annals of mathematical statistics*, *12*, 153-162.

Von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Annals of mathematical statistics*, *12*, 367-395.

Young, L. C. (1941). On randomness in ordered sequences. *Annals of mathematical statistics*, *12*, 293-300