REVUE CONGOLAISE DES SCIENCES HUMAINES ET SOCIALES

Site web: http://recosh.cresh-rdc.org

ISSN :2957-6385 (print), ISSN: 2957-6393 (online)
https:/www. doi.org/10.59189/crsh20260411

Analyse de corrélation : fondements, démarche, intérêt et limites

Correlation analysis: fundamentals, methodology, benefits ans limitations

José Mangalu Mobhe Agbada[1] et Eugenie Kabali Hamuli

Résumé. *Dans cet article, les auteur.e.s examinent la relation entre deux variables quantitatives à l'aide du test de corrélation, une méthode couramment utilisée mais souvent mal appliquée et mal interprétée. Ils proposent de revisiter, sous un éclairage nouveau et contextualisé, non seulement, les présupposés théoriques et les conditionnalités relatives à l'application de ce test, mais aussi, l'intérêt et les pièges à éviter dans son utilisation et dans l'interprétation des résultats qui en sont issus les bases théoriques. L'analyse de corrélation cherche à vérifier s'il existe un lien linéaire entre des variables et si ce lien est suffisamment fort pour ne pas être dû au hasard. L'article présente également des exemples concrets pour permettre aux lecteurs de mettre en pratique les concepts, soulignant l'importance de maîtriser cette technique, souvent préalable à l'utilisation de tests statistiques plus avancés, comme la régression linéaire.*

Mots-clés : Corrélation, variables quantitatives, normalité, linéarité, diagramme de dispersion, covariance, coefficient de corrélation, coefficient de détermination, intervalle de confiance.

Abstract. *In this paper, the authors examine the relationship between two quantitative variables using the correlation test, a method that is commonly used but often misapplied and misinterpreted. They propose to revisit, in a new and contextualized light, not only the theoretical presuppositions and conditionalities relating to the application of this test, but also the interest and pitfalls to be avoided in its use and in the interpretation of the results resulting from it. Correlation analysis seeks to check whether there is a linear link between variables and whether this link is strong enough not to be due to chance. The article also presents concrete examples to allow readers to put the concepts into practice, highlighting the importance of mastering this technique, which is often a prerequisite for the use of more advanced statistical tests, such as linear regression.*

Key words: Correlation, quantitative variables, normality, linearity, scatter plot, covariance, correlation coefficient, coefficient of determination, confidence interval.

OPEN ACCESS

Received: 02 july 2025

Accepted: 15 August 2025

Available online: 11 nov. 2025

[1] Professeur à l'Ecole des Sciences de la Population et du Développement Université de Kinshasa
e-mail : jose.mangalu@gmail.com, Tél. : +243999917013

Introduction

Dans cet article, il sera question de décrire la relation entre deux variables quantitatives à l'aide de l'analyse de corrélation. Comme l'analyse de Khi-carré, l'analyse de corrélation se propose de vérifier l'existence d'un lien (linéaire) entre les variables quantitatives prises deux à deux sur un échantillon et si ce lien est assez fort pour que l'on soit sûr qu'il n'a pas pu être observé par hasard. En plus de déterminer l'existence de ce lien, la corrélation indique également l'intensité, le sens et la significativité de ce lien. En d'autres termes, l'analyse de corrélation permet de voir comment deux variables quantitatives co-varient ou varient simultanément. En somme, les relations entre variables quantitatives prises deux à deux peuvent être repérées de plusieurs façons : leur description à l'aide du diagramme de dispersion, la mesure de leur intensité via le coefficient de corrélation r de Bravais-Pearson et le coefficient de détermination R^2 et la régression linéaire simple, etc. Costa (2013, p. 119). La corrélation est particulièrement utile lorsque les variables analysées expriment des quantités de temps (âge, durée entre deux événements, etc.), des quantités monétaires (revenus, patrimoine, dépenses, consommation, etc.), des fréquences ou des indicateurs synthétiques construits pour les besoins de l'enquête (Martin, 2009). Autant que l'analyse de Khi-carré, la corrélation constitue également une analyse descriptive ou exploratoire à d'autres analyses plus élaborées, notamment l'analyse de la régression linéaire.

Dans la suite de cet article, il sera question de présenter les fondements, la démarche, l'intérêt et les limites de la corrélation linéaire. Concrètement, on abordera les points suivants : les conditions d'application du test de corrélation, les hypothèses associées à ce test, la construction du diagramme de dispersion, les étapes du calcul de coefficient de corrélation, la détermination des intervalles de confiance associé, le calcul du coefficient de détermination, le test de signification du coefficient de détermination, l'interprétation et la présentation des résultats de la corrélation ainsi que l'intérêt et les limites du diagramme de dispersion et du coefficient de corrélation de Bravais-Pearson.

Démarche de réalisation d'une analyse de corrélation

La démarche pour réaliser une analyse de corrélation passe par les étapes suivantes :

1. Vérification des conditions d'application de la méthode ;
2. Pose des hypothèses statistiques ;
3. Construction du diagramme de dispersion et l'examen visuel de son allure ;
4. Calcul du coefficient de corrélation linéaire (r) de Bravais-Pearson ;
5. Le cas échéant, détermination d'un intervalle de confiance pour r ;
6. Calcul du coefficient de détermination (R^2) ;
7. Réalisation du test d'hypothèse sur r ou sur R^2 ;
8. Interprétation et présentation des résultats.

Conditions d'application de l'analyse de corrélation

A l'instar d'autres tests d'hypothèse, le recours à la corrélation suppose l'observance de certaines conditions d'application, au nombre desquelles il y a lieu de citer : la linéarité de la relation, le caractère aléatoire de l'échantillon, la normalité de la distribution des variables parentes[2] et l'absence de valeurs aberrantes ou des *outliers* ou des valeurs atypiques.

Deux variables entretiennent une relation linéaire lorsque la variation relative de l'une est automatiquement associée à une variation constante de l'autre. En d'autres termes, les variables X et Y entretiennent une relation linéaire si une variation de p % de X entraîne toujours une variation de q % de Y, la valeur de q pouvant être positive ou négative.

Le caractère aléatoire de l'échantillon postule que les observations tirées d'un échantillon doivent être indépendantes de celles tirées de l'autre. En effet, deux observations sont dites indépendantes lorsque la connaissance de l'une ne conditionne pas ou n'implique pas la connaissance de l'autre. Cette condition n'est pas remplie lorsque par exemple on compare les données Y en fonction du temps X (données de la vielle ne sont pas indépendantes de celles du lendemain) ou encore la température de la nuit n'est pas indépendante de la température prévalue la journée. Dans ce cas, les données sont corrélées et il faut faire appel à d'autres techniques d'analyse, notamment de séries chronologiques.

La normalité de la distribution implique que dans la population-mère d'où est tiré l'échantillon, les distributions conditionnelles de Y liées à chaque valeur de X doivent être normales et de variances égales et symétriquement, chaque distribution conditionnelle de X liée à chaque valeur de Y doit être normale et de variance égale. En d'autres termes, la plupart des valeurs des observations doivent se regrouper autour de la moyenne et les autres valeurs s'en écartent symétriquement des deux côtés. Cette condition est difficile, voire impossible à vérifier dans la pratique même si elle est souvent vraie. Dans la réalité, on estime que plus la taille de l'échantillon est importante ($\geq$ 30), plus la distribution a tendance à être normalement distribuée. Ainsi, cette condition n'est d'application que dans les cas où la taille de l'échantillon est inférieure à 30.

Hypothèses statistiques associées à la corrélation

Autant que les autres tests statistiques, l'analyse de la corrélation exige également que l'on spécifie les hypothèses statistiques : l'hypothèse nulle (H_0) et l'hypothèse alternative (H_1). L'H_0 postule l'absence de relation entre les deux variables ($r = 0$). Son rejet implique l'acceptation de l'H_1 postulant l'existence d'un lien entre les deux variables ($r \neq 0$).

Toutefois, il n'est pas sans intérêt de rappeler que toute recherche des liens statistiques entre variables doit absolument être précédée de l'examen de l'existence des liens logiques entre elles. En effet, il est hasardeux de se lancer dans des démonstrations statistiques sophistiquées si on n'a pas pris soin de démontrer, notamment en recourant à la littérature et à l'expérience, l'existence des liens logiques entre phénomènes. C'est seulement lorsque l'on se rassure de l'existence de ces liens logiques que

[2] On appelle variable parente, toute variable étudiée et considérée sur l'ensemble de la population de référence et non seulement sur l'échantillon.

l'on pourrait se lancer dans des calculs statistiques sophistiqués, notamment pour en déterminer l'ampleur. C'est l'intérêt de la corrélation.

Construction du diagramme de dispersion

Dans le cas de la corrélation et même de la régression linéaire, l'un des instruments à utiliser pour visualiser les liens logiques, voire statistiques entre variables est sans conteste le diagramme de dispersion. Celui-ci se présente sous forme d'un plan cartésien où, par convention, la variable dépendante ou supposée telle (notée Y)[3] est placée sur l'axe vertical (axe des ordonnées) et la variable indépendante ou supposée telle (notée X) est placée sur l'axe horizontal (axe des abscisses), chaque observation est représentée par un point, qui est l'intersection de ses scores sur ces deux variables. Ces différents points forment une espèce de nuage sur ce plan. L'orientation que prennent ces points sur l'axe cartésien que représente ce diagramme de dispersion donne une indication sur l'existence ou non du lien, sur la forme de ce lien (linéaire, curvilinéaire, polynomiale, logarithmique ou autre), sur son intensité (faible, moyen ou fort) et sur son sens (positif ou négatif). Les Figures 1 à 6 ci-dessous représentent les différentes formes de nuages de points sur un plan défini par les deux variables Y et X.

Par exemple, la direction de nuage des points sur la Figure 1 est caractéristique d'une absence de relation (linéaire) entre la variable X (nombre d'enfants des femmes) et la variable Y (poids des femmes), ainsi chaque valeur de la variable de X pouvant être ou étant associée à des valeurs faibles ou élevées de la variable Y et vice-versa. En revanche, sur les 5 autres Figures (2 à 6), les nuages des points prennent chaque fois une forme particulière. Les positions de points suivent une certaine logique par rapport à la variable X et à la variable Y, ce qui laisse transparaitre une certaine relation entre les deux variables.

Outre l'existence ou non d'une relation entre variables, l'agencement des points sur l'espace cartésien indique également la forme et le sens de la relation entre les variables. On peut globalement distinguer 4 formes de relation : une relation linéaire positive parfaite (Figure 2), une relation linéaire positive imparfaite (Figure 3), une relation linéaire négative parfaite (Figure 4), une relation linéaire négative imparfaite (Figure 5) et une relation non-linéaire. Ainsi, les Figures 2 à 4 sur lesquelles les points s'agencent sur une certaine direction et se concentrent autour d'une droite traduisent une relation linéaire alors que sur la Figure 6, les points se concentrent également mais forment une figure qui s'éloigne d'une droite, cela dénote une relation non-linéaire (assimilée souvent à l'absence de relation).

La Figure 2 qui, en 2020, met en relation l'évolution de l'âge d'un père (la variable X) qui a eu son enfant en 2005 à l'âge de 30 et l'âge de l'enfant (la variable Y) dénote une relation positive parfaite. Les relations positives se lisent du coin inférieur gauche du diagramme à son coin supérieur droit. Il y a une relation positive entre variables lorsque toutes les deux évoluent dans le même sens ; c'est-à-dire aux valeurs élevées d'une variable (X) correspondent également les valeurs

[3] A noter que contrairement à la régression qui exige de spécifier une variable dépendante et une ou plusieurs variables indépendantes, la corrélation n'impose pas cette spécification, les positions ou les fonctions des variables étant interchangeables. La corrélation est donc une analyse symétrique. Toutefois, ainsi qu'il a déjà été dit, étant donné que la corrélation est souvent préliminaire à d'autres types d'analyse, la spécification des positions ou des fonctions des variables est souhaitable.

élevées de l'autre variable (Y) ou inversement, aux valeurs faibles de X correspondent également les valeurs faibles de Y. Une relation positive est dite parfaite lorsque tous les couples des points représentant les observations s'alignent tous parfaitement sur une droite imaginaire. Dans ce cas, on peut, à chaque valeur de X associée avec certitude la valeur correspondante de Y. Signalons tout de même qu'il n'existe pas, dans la réalité sociale, plusieurs situations où deux variables soient positivement et parfaitement corrélées, sauf si l'on croise une variable par rapport à elle-même.

Dans une relation positive imparfaite, les deux variables évoluent dans le même sens, mais les points ne sont plus parfaitement alignés sur la droite. C'est le cas de la Figure 3 qui met en relation l'âge des jeunes enfants et leurs poids. On y voit que plus l'âge (en mois) augmente, plus également le poids augmente, mais la relation n'est plus parfaite comme dans la Figure 2.

La Figure 4 met en relation le nombre de pensionnaires ayant déjà été servis dans un hospice pour vieillards (X) et le nombre de repas restants (Y). Cette Figure dénote une relation négative parfaite. Les relations négatives se lisent du coin supérieur gauche du diagramme à son coin inférieur droit. Il y a une relation négative entre variables lorsque les deux évoluent dans le sens opposé ; c'est-à-dire aux valeurs élevées d'une variable (X) correspondent les valeurs faibles de l'autre variable (Y) ou inversement, aux valeurs faibles de X correspondent les valeurs élevées de Y. Une relation négative est dite parfaite lorsque tous les couples des points représentant les observations s'alignent tous parfaitement sur une droite imaginaire. Dans ce cas, on peut, à chaque valeur de X associée avec certitude la valeur correspondante de Y. Autant que pour les relations positives parfaites, il n'existe pas, dans la réalité sociale, plusieurs situations où deux variables soient négativement et parfaitement corrélées.

Dans une relation négative imparfaite, les deux variables évoluent dans le sens opposé, mais les points ne sont plus parfaitement alignés sur la droite. C'est le cas de la Figure 5 qui met en le nombre d'années de scolarité des femmes et le nombre d'enfants qu'elles ont eus. On y observe globalement que plus le nombre d'années de scolarité augmente, plus le nombre d'enfants (la fécondité) diminue, mais la relation n'est plus parfaite comme dans la Figure 4.

Lorsque l'allure que dégage le nuage des points sur un diagramme de dispersion ne permet pas de tracer une ligne, il faut se garder de conclure à la hâte à l'absence des liens entre variables. D'abord s'assurer que l'ajustement avec une autre forme, comme un U inversé ou une courbe (relation curvilinéaire), n'est pas possible avant de conclure en l'absence des liens. Pour rappel, toute relation entre variables n'est pas forcément linéaire. Pour des relations non-linéaires, on pourrait être amené à forcer la linéarisation à travers la transformation des variables initiales, en logarithmes, moyennes mobiles, etc.. En effet, la corrélation ne mesure que des liens linéaires. Encore une fois, c'est l'examen des nuages des points que forme le diagramme de dispersion qui permet de soupçonner la forme de liens et de choisir le type d'analyse approprié.

Figure 1 : Relation entre la taille des femmes et le nombre d'enfants qu'elles ont eus

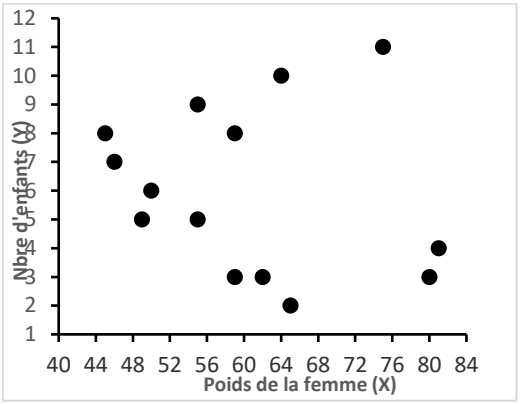


Figure 2 : Relation entre l'âge du père et l'âge de son enfant

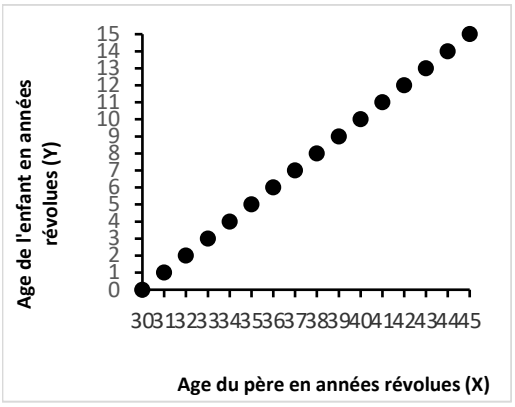


Figure 3 : Relation entre l'âge des enfants (en mois) et leurs poids (en Kgs)

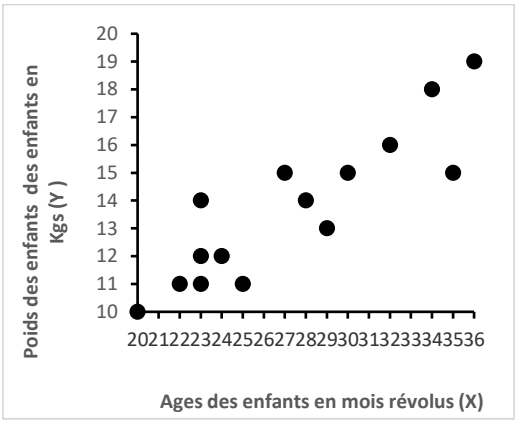


Figure 4 : Relation entre le nombre de pensionnaires servis et le nombre de plats restants

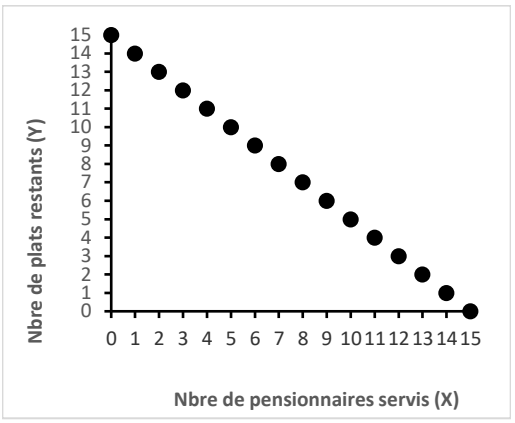


Figure 5 : Relation entre nombre d'années de scolarité des femmes et leur fécondité

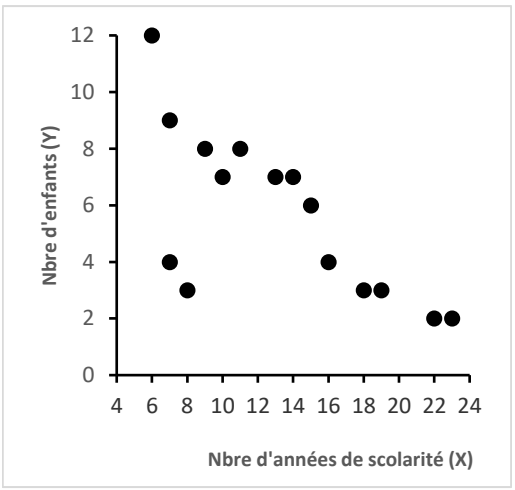
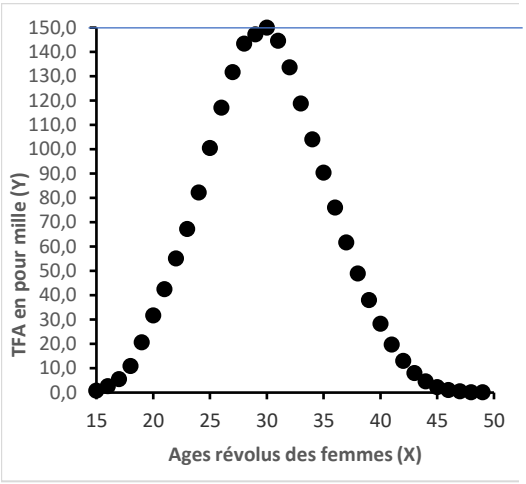


Figure 6 : Relation entre l'âge des femmes et les Taux de Fécondité par âgé



Coefficient de corrélation linéaire (r) de Bravais-Pearson

Une autre indication que l'on peut tirer des liens entre deux variables est la force ou l'intensité de la relation entre elles. En effet, qu'elle soit positive ou négative, la relation entre deux variables sera d'autant plus forte que les points sont peu dispersés autour de la droite (imaginaire) qui traverse le nuage des points en respectant sa direction principale et d'autant plus faible que les points sont très dispersés par rapport à cette droite. Ainsi, si tous les points s'alignent exactement sur cette droite, la relation entre les deux variables est parfaite, ce qui équivaut à un coefficient de corrélation en valeur absolue de 1.

La corrélation linéaire de Bravais-Pearson est mesure standardisée, il est noté par la lettre r et varie de -1 à +1. Il permet de quantifier la force ou l'intensité d'une relation linéaire entre deux variables, d'en identifier le sens (positif ou négatif) et de tester si elle est ou non significative. En d'autres termes, il constitue une mesure du degré de concentration des observations le long de la droite de régression. Toutefois, même si le coefficient de corrélation de Bravais-Pearson (r) est le plus utilisé, d'autres statistiques mesurant la force de relation entre variables quantitatives existent également comme êta carré (η^2), le V de Cramer et le coefficient de corrélation de Spearman.

Pour le calcul effectif de coefficient de corrélation, suivre les étapes suivantes :

1. Disposer de deux séries des données quantitatives des variables X et Y ;
2. Calculer la moyenne arithmétique de chacune des variables ;
3. Calculer la variance de chacune de séries des variables ;
4. Calculer l'écart-type de chacune de séries des variables ;
5. Calculer la covariance de deux variables ;
6. Calculer le coefficient de corrélation.

Le coefficient de corrélation linéaire de Bravais-Pearson (r) s'obtient en divisant la covariance des variables concernées par le produit de leurs écarts-types respectifs, à travers la relation suivante :

$$r_{xy} = \frac{S_{xy}}{S_y * S_x}$$

Où :

r_{xy}= coefficient de corrélation linéaire entre les variables X et Y

S_{xy}= covariance des variables X et Y

S_y= écart type de la variable Y

S_x = écart-type de la variable X

Pour trouver la corrélation, on passe par le calcul de la covariance dont la formule est :

$$S_{xy} = \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{N-1}$$

Où :

$\bar{y}$ représente la moyenne arithmétique de la variable Y

$\bar{x}$ représente la moyenne arithmétique de la variable X

y_i = score ou valeur de la variable Y

x_i = score ou valeur de la variable X

N = la taille de l'échantillon.

De façon concrète, pour calculer la corrélation, remplir les différentes colonnes du Tableau 1 suivant :

Tableau 1 : Calcul du coefficient de corrélation r

N°	Variable Y (y_i)	Variable X (xi)	yi-$\bar{y}$	(yi-$\bar{y}$) ²	xi-$\bar{x}$	(xi-$\bar{x}$) ²	(yi-$\bar{y}$) (xi-$\bar{x}$)
1	$\bar{y}=\frac{1}{N}\sum(y_i)$	$\bar{x}=\frac{1}{N}\sum(x_i)$		$\sum(y_i-\bar{y})^2$		$\sum(x_i-\bar{x})^2$	$\sum(y_i-\bar{y})(x_i-\bar{x})$
2				$S_y^2=\frac{\sum(yi-\bar{y})^2}{N}$		$S_x^2=\frac{\sum(xi-\bar{x})^2}{N}$	$S_{xy}=\frac{\sum(yi-\bar{y})(xi-\bar{x})}{N-1}$

Une fois que toutes les composantes de la formule calculées, appliquer simplement la formule de coefficient de corrélation, soit :

$$r_{xy}=\frac{S_{xy}}{S_y * S_x}$$

Intervalles de confiance

Pour des besoins d'interpolation des résultats obtenus de l'échantillon à l'ensemble de la population-mère, il peut s'avérer important de déterminer les intervalles de confiance associés au coefficient r. Pour cela, on peut se référer utilement à la procédure proposée par Rosnow et Rosenball (1996). Cette procédure se décline en 6 principales étapes que voici :

1. Se fixer un niveau de confiance ou une marge d'erreur. En sciences sociales, le niveau de confiance est généralement fixé à 95 %, laissant une marge d'erreur de 5 % ;
2. Consulter la table pour transformer le coefficient r en score Z de Fisher. A défaut, utiliser la relation suivante pour transformer le coefficient r calculé de Fisher :

   $Z_r=\frac{1}{2}\ln\left(\frac{(1+r)}{(1-r)}\right)$;
3. Multiplier $1/\sqrt{(n-3)}$ par 1,96 ;
4. La limite inférieure de l'intervalle de confiance est donnée par la différence entre le nombre trouvé en 2 et celui trouvé en 3 ;
5. La limite supérieure est donnée par la somme des nombres trouvés en 2 et en 3 ;
6. Consulter à nouveau la table F de Fisher pour retrouver les *r* correspondants aux valeurs de Z_r (Z de Fisher) calculées ou à défaut, utiliser la relation suivante pour convertir les Z_r en r :

   $r=\frac{e^{2*z_r}-1}{e^{2*z_r}+1}$

Coefficient de détermination

Le coefficient de détermination ou la part de la variance expliquée, symbolisé par la lettre R^2 (R carré), mesure la part de la variation d'une variable (généralement la variable supposée dépendante, Y) qui est expliquée par la variation de l'autre variable (généralement le variable supposée indépendante, X). Le coefficient de détermination permet ainsi d'apprécier, quantitativement, la force d'une explication et d'une relation et donc, inversement, la part de la variation restant à expliquer. Pour obtenir le R^2, rapporter le carré de la co-variation du coefficient de corrélation r des deux variables au produit de leurs variances, soit la relation suivante :

$$R^2=(r)^2=\frac{S_{xy}^2}{S_x^2 * S_y^2}$$

On peut une fois de plus se servir utilement du Tableau 1 pour compléter le calcul du coefficient de détermination.

Test de signification de r ou de R^2

Comme d'ordinaire, les données utilisées pour des analyses proviennent généralement des échantillons, il convient par la suite de tester la signification de r ou de R^2 pour évaluer s'il s'écarte significativement de 0, correspondant à l'hypothèse nulle (H_0). Deux tests sont généralement utilisés : le test *F* ou le test *t*. Toutefois, le test *F* est plus utilisé, il est réalisé à travers la relation suivante :

$$F = \frac{R^2(n-2)}{1-R^2}$$

Où

R^2 = Coefficient de détermination et n= Taille de l'échantillon

L'hypothèse nulle postulant l'absence de relation entre les deux variables, soit $r=R^2=0$. Si le test *F* indique un niveau de signification inférieur à 5 %, on rejette l'H_0 et en accepte l'H_1.

Les degrés de liberté pour ce test sont 1 et N-2. Une fois que le ratio *F* et ses degrés de liberté sont calculés, se référer à la table de distribution théorique de F au seuil de 5 %. Nous trouvons le degré de liberté dans la première colonne des valeurs de *F* du tableau. N-2 dans trouve dans les degrés de liberté associés aux rangées de la table de distribution théorique de *F*. Si *F* calculé est plus grand que la valeur correspondante de la table de distribution théorique de *F,* rejeter l'H_0 et accepter l'H_1.

Matrice de corrélation

L'analyse de corrélation est par définition une analyse bivariée, portant sur des variables prises deux à deux. Lorsque l'on dispose de plus de deux variables, on construit généralement une matrice de corrélation. Dans cette matrice, les mêmes variables figurent simultanément en lignes et en colonnes. C'est à l'intersection de chaque ligne et de chaque colonne que se trouve le coefficient de corrélation (r) du couple formé par la variable définie en ligne et celle définie en colonne ; la diagonale principale de la matrice (qu'on ne présente pas et dont le r est égal à 1) représentant le r croisant une variable à elle-même. Cette diagonale principale coupe la matrice de corrélation en deux parties, appelées triangles. On a donc un triangle supérieur au-dessus de la diagonale principale et un triangle inférieur en dessous de la diagonale principale. Chacun de ces triangles reprend exactement les mêmes résultats (coefficients r, …), étant donné que la symétrie de la corrélation. Ce qui implique in fine qu'on ne présente et n'interprète que l'un de ces triangles.

Interprétation du coefficient de corrélation r et présentation des résultats

Pour interpréter le coefficient de corrélation r, on se réfère à la fois à sa valeur et à son signe. La valeur de r varie toujours entre -1 et +1. La valeur absolue de ce coefficient indique l'intensité ou la force de la relation linéaire entre les deux variables, la valeur zéro (0) indique une absence totale de relation entre les deux variables (cas très rare) et la valeur 1 indique une relation parfaite (cas très rare). Plus cette valeur (en absolue) s'approche de 1, plus la relation est forte entre les deux variables. Pour les valeurs intermédiaires de r autres que 0 et 1, se référer au Tableau 2 ci-dessous. Le signe associé au coefficient r indique le sens de la relation (positif ou négatif) ; le signe plus (+) pour une relation linéaire positive et le signe moins (-) pour une relation linéaire négative.

Tableau 2 : Règles d'interprétation de coefficient r

Valeur absolue de r	Interprétation
1	Lien parfait
0,8 à 0,9	Lien très fort
0,6-0,7	Lien fort
0,4 à 0,5	Lien modéré
0,1 à 0,3	Lien faible
0	Lien nul (absence de lien)

Toutefois, au niveau agrégé, on considère comme étant fortes, les corrélations dont la valeur absolue se situe dans l'intervalle (0,50 à 0,69), comme très fortes, celles dont la valeur absolue se situe dans l'intervalle (0,70 à 0,89) ; des r supérieurs ou égaux à 0,90 indiquant une relation quasi-parfaite entre les deux variables. En revanche, au niveau individuel, les corrélations sont généralement plus faibles. Cela est notamment dû à la plus grande variabilité des mesures réalisées au niveau individuel. En effet, il est établi que plus l'on agrège une variable, plus on fait disparaître les valeurs extrêmes et partant, les variabilités au niveau individuel. Les mesures agrégées tendent à s'approcher des comportements moyens.

Pour la présentation des résultats de l'analyse de la corrélation dans un rapport scientifique, on pourrait, en reprenant l'exemple de la Figure 3 sur le lien entre l'âge des enfants (nourrissons) et leurs poids, simplement écrire que le poids des enfants est positivement et fortement corrélé à leurs âges (n=15, r=0,79, p <0,001, R^2 =0,62). On peut également dire que lorsque l'âge des enfants augmente, leurs poids augmentent aussi. Le coefficient de détermination (R^2 =0,62) montre que l'âge des enfants explique à concurrence de 62 % les variations observées dans leurs poids. La signification p (0,001) induit que cette corrélation peut être généralisée à l'ensemble de la population d'enfants d'où est tiré l'échantillon.

Application de la corrélation et source des données

Questions de recherche et source des données

Dans cette étude, il est question de répondre aux questions suivantes :

Existe-t-il une relation entre le niveau de vieillissement démographique d'un pays et son niveau de développement socio-économique ? Quel est le sens de ce lien ? Quels en sont les facteurs les plus déterminants ? Peut-on désormais utiliser le niveau de vieillissement démographique comme un indicateur de développement socio-économique des Nations ?

Pour répondre à ces questions, une analyse agrégée a été réalisée sur 109 pays du Sud autour de l'année 2010, en utilisant à la fois les indicateurs de vieillissement démographique et du développement économique. Les données en rapport avec le vieillissement démographique ont été tirées des estimations et projections de population mondiale réalisées par la Division de Population des Nations Unies, la révision de 2010. Ces estimations et projections reposent sur l'observation des tendances des phénomènes démographiques majeurs (fécondité, mortalité et migration) et sont élaborées sur base de trois hypothèses (faible, moyenne et haute) associées à l'évolution de ces trois principales composantes de la dynamique démographique. Seules les données tirées de l'hypothèse moyenne ont été utilisées.

Pour ce qui est du choix des indicateurs de vieillissement démographique, il dépend à la fois de la façon dont le vieillissement démographique a été conceptualisé et de la fixation de l'âge à partir duquel une personne

peut être considérée comme âgée. Le vieillissement démographique a été conceptualisé comme l'augmentation du nombre et de la proportion des personnes âgées au sein d'une population. A quel âge alors une personne est considérée comme âgée ? Si pendant longtemps, dans la plupart des pays, on considérait l'âge de 60 ans comme marquant le début du vieillissement, depuis un certain temps, notamment grâce à l'allongement de l'espérance de vie à la naissance, de voix s'élèvent pour réclamer de porter ce seuil à 65 ans. Si ce nouveau seuil peut facilement se justifier dans les pays développés où l'espérance de vie à la naissance a dépassé les 70 ans, il est difficilement justifiable dans la plupart des pays en développement, notamment en Afrique, où l'espérance de vie à la naissance tourne autour de 50 ans. C'est pourquoi dans le cadre de cette étude, le seuil de 60 ans a été retenu comme marquant le début du vieillissement.

Ainsi, le premier indicateur auquel on pense pour mesurer le niveau de vieillissement d'un pays est la proportion des personnes de 60 ans et plus. Si cet indicateur traduit l'augmentation relative de la part des personnes âgées au sein d'une population, on peut noter que l'augmentation de la proportion des personnes de 60 ans et plus ne résulte pas seulement de l'augmentation du nombre de personnes âgées au sein d'une population, elle peut aussi résulter de la baisse du nombre d'enfants de moins de 15 ans et/ou des adultes de 15-59 ans. C'est pourquoi, dans le cadre de cette étude, pour appréhender le niveau de vieillissement démographique d'un pays, la proportion des personnes de 60 ans et plus a été associée à d'autres indicateurs, notamment la proportion des enfants de moins de 15 ans, la proportion des adultes de 15-49 ans, l'âge médian et le rapport de dépendance économique.

Pour ce qui du développement des pays, pendant longtemps, il a été mesuré par le Produit National Brut (PNB) et sa déclinaison, le Produit Intérieur Brut (PIB). Le PIB mesure la production totale de biens et services réalisée à l'intérieur d'un pays pendant une période donnée. Dans ce sens, il mesure la santé économique des pays. Cet indicateur a reçu beaucoup de critiques de la part de spécialistes. En effet, le PIB n'est qu'une mesure globale, une moyenne et comme toute moyenne, il est très sensible aux valeurs extrêmes. De même, il ne permet d'appréhender ni les inégalités sociales ni leur évolution. On peut très bien avoir un PIB moyen qui augmente alors que les revenus (qu'il est censé mesurer) diminuent pour une majorité de la population et augmentent fortement pour une minorité, ce qui renforce les inégalités. Par ailleurs, il ne mesure pas non plus tout ce qui est qualitatif, comme le bien-être, la sécurité, l'éducation, la liberté. Toutefois, malgré toutes ces imperfections, il a été résolu d'utiliser le PIB/Hab. comme indicateur du niveau de développement économique des pays. Les différents indicateurs de vieillissement démographiques et du développement économique utilisés sont repris dans le Tableau 3 ci-après :

Tableau 3 : Indicateurs de vieillissement démographique et de développement économique utilisés

Indicateurs du vieillissement démographique	Indicateur du développement économique
Proportion de moins de 15 ans (Prop_15ans) Proportion de 15 à 59 ans (Prop15_59ans) Proportion de 60 ans et + (Prop60ans) Age médian (Age_Median) Rapport de dépendance économique (RDE)	Produit intérieur brut par habitant (PIB/Hab.).

Le type de recherche envisagée, décliné ici par les questions de recherche et la nature des variables en présence (toutes quantitatives) ont motivé le choix de la corrélation linéaire.

Illustration de la corrélation linéaire

Vérification des conditions d'application de la corrélation linéaire

La première chose à faire lorsque l'on se propose de recourir à l'analyse de la corrélation, ou à toute autre méthode d'analyse, est de s'assurer que les conditions d'application de la méthode sont remplies. Pour ce qui est de la corrélation, il s'agit notamment de la normalité de la distribution, de l'absence des valeurs aberrantes et de la linéarité de la relation. En rapport avec la normalité de la distribution, étant donné que le nombre d'observations (n=109) dépasse largement le minimum attendu (30), on peut supposer que cette condition est remplie. Pour détecter la forme de la relation entre chacun des indicateurs de vieillissement démographique et le niveau de développement économique, il a été fait recours à l'examen visuel des diagrammes de dispersion.

Construction des diagrammes de dispersion

Pour la construction des diagrammes de dispersion, le PIB/Hab. a été considéré comme variable dépendante et chacun des indicateurs du vieillissement démographique a été pris comme variable indépendante. Aussi, il sera construit autant de diagrammes de dispersion que les couples formés par chacune des variables indépendantes et la variable dépendante.

Ces diagrammes de dispersion sont représentés sur les Figures 7 à 11. On y note globalement que la relation entre chacun des indicateurs du vieillissement démographique, d'une part, et le PIB/Hab., d'autre part, se présente sous forme linéaire. Pour certains, la relation est positive et alors qu'elle est négative pour d'autres. Par exemple, sur la Figure 7, on observe une relation négative entre la proportion de moins de 15 ans et le PIB/Hab. L'augmentation de la proportion des enfants de moins de 15 ans s'accompagne d'une baisse du PIB/Hab. Ceci se comprend intuitivement dans la mesure où les enfants ne participent pas à la production de la richesse nationale que représente le PIB/Hab. Sur la Figure 8, on note une relation positive entre la proportion d'adultes (15 à 59 ans) et le PIB/Hab. Une autre relation négative s'observe sur la Figure 11 entre le rapport de dépendance économique, c'est-à-dire le nombre d'inactifs (enfants de moins de 15 ans et personnes de 60 ans et plus) pris en charge par chaque actif (personne de 15-49 ans), plus cette charge augmente, moins il y a production des richesses.

Figure 7 : Relation entre Proportion de moins de 15 ans et PIB/Hab.

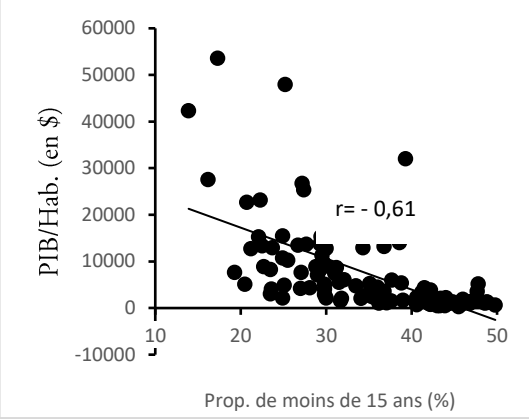


Figure 8 : Relation entre Proportion de 15-59 ans PIB/Hab.

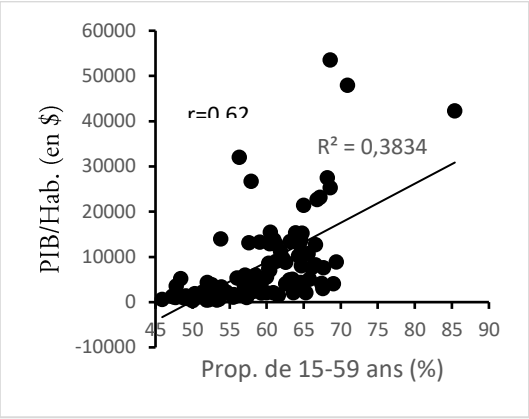


Figure 9 : Relation entre Proportion de 60 ans et + et PIB/Hab.

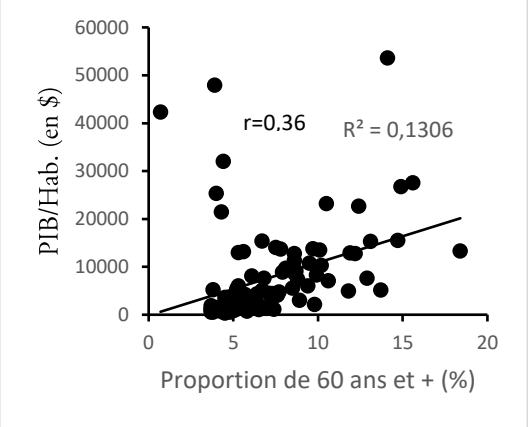


Figure 10 : Relation entre l'âge médian et PIB/Hab.

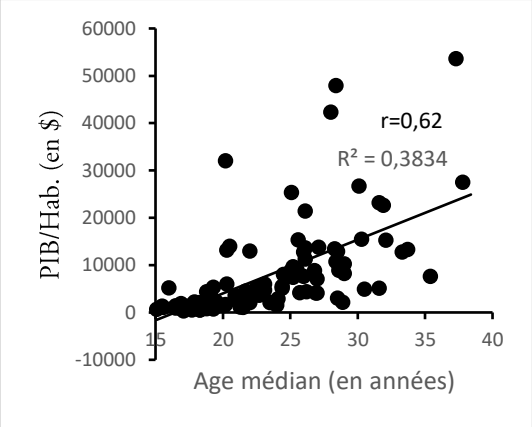
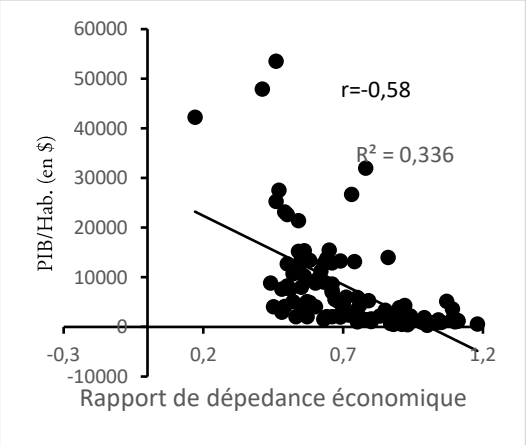


Figure 11 : Relation entre Rapport de dépendance économique et PIB/Hab.



Fixation des hypothèses statistiques

Etant donné que la corrélation se calcule en prenant les variables deux à deux, on peut postuler ici des hypothèses nulles entre chacun des indicateurs du vieillissement démographique et le PIB/Hab. Ainsi, on peut par exemple poser qu'il n'y a pas de lien entre la proportion de moins de 15 ans et le PIB/Hab. et ainsi que suite. Les hypothèses alternatives seront justes le contraire des hypothèses nulles.

Calcul des coefficients de corrélation

Pour le besoin de cette illustration, deux procédé de calcul du coefficient de corrélation seront utilisés : d'abord manuellement et ensuite en recourant à l'application SPSS. Les calculs manuels porteront sur l'exemple de la Figure 5 portant sur la relation entre le nombre d'années de scolarité des femmes et le nombre de leurs enfants. Pour la partie SPSS, l'exemple sur la relation entre le vieillissement démographique et le développement économique sera utilisé.

Pour l'illustration manuelle, il sera fait usage du Tableau 1 (Tableau de calcul du coefficient de corrélation). Les données à utiliser sont reprises dans le Tableau 4 qui suit :

Tableau 4 : Répartition des femmes enquêtées selon le nombre d'années de scolarité des femmes et le nombre de leurs enfants

Nombre d'années de scolarité	Nombre d'enfants
6	12
7	9
7	4
8	3
9	8
10	7
11	8
13	7
14	7
15	6
16	4
18	3
19	3
22	2
23	2

Tableau 4 : Calcul de coefficient de corrélation entre le nombre d'années de scolarité des femmes et le nombre de leurs enfants

N°	Variable Y (y_i)	Variable X (xi)	yi-$\bar{y}$	(yi-$\bar{y}$) ²	xi-$\bar{x}$	(xi-$\bar{x}$) ²	(yi-$\bar{y}$) (xi-$\bar{x}$)
1	6	12	6,33	40,11	-7,2	51,84	-45,60
2	7	9	3,33	11,11	-6,2	38,44	-20,67
3	7	4	-1,67	2,78	-6,2	38,44	10,33
4	8	3	-2,67	7,11	-5,2	27,04	13,87
5	9	8	2,33	5,44	-4,2	17,64	-9,80
6	10	7	1,33	1,78	-3,2	10,24	-4,27
7	11	8	2,33	5,44	-2,2	4,84	-5,13
8	13	7	1,33	1,78	-0,2	0,04	-0,27
9	14	7	1,33	1,78	0,8	0,64	1,07
10	15	6	0,33	0,11	1,8	3,24	0,60
11	16	4	-1,67	2,78	2,8	7,84	-4,67
12	18	3	-2,67	7,11	4,8	23,04	-12,80
13	19	3	-2,67	7,11	5,8	33,64	-15,47
14	22	2	-3,67	13,44	8,8	77,44	-32,27
15	23	2	-3,67	13,44	9,8	96,04	-35,93
Total	198	85	-	121,33	-		
Moyenne	$\bar{y}=\frac{1}{N}\sum(y_i)$	$\bar{x}=\frac{1}{N}\sum(x_i)$		$\sum(y_i-\bar{y})^2$		$\sum(x_i-\bar{x})^2$	$\sum(y_i-\bar{y})(x_i-\bar{x})$
	13,2	5,67		121,33		430,4	-161
Variances et co-variance				$S_y^2=\frac{\sum(y_i-\bar{y})^2}{N-1}$		$S_x^2=\frac{\sum(x_i-\bar{x})^2}{N-1}$	$S_{xy}=\frac{\sum(yi-\bar{y})(xi-\bar{x})}{N-1}$
				8,67		30,74	-11,5
Ecart-types (racine carrée de la variance)			s_y	2,94	s_x	5,54	
Produit des écart-types							16,32
Coefficient de corrélation						$r_{xy}=\frac{S_{xy}}{S_x * S_y}$	$=\frac{-11,5}{16,32}$: -0,7045

Au vu du coefficient de corrélation, on peut donc conclure à une relation négative forte entre le nombre entre le nombre d'années de scolarité des femmes et le nombre d'enfants qu'elles ont eus. Ainsi, plus le nombre d'années de scolarité augmente, moins les femmes ont des enfants. Reste à savoir si cette relation trouvée au niveau de l'échantillon a la chance de se reproduire au niveau de l'ensemble de la population. Pour cela, il faut d'abord construire l'intervalle de confiance de r au sein de la population et par la suite, procéder au test de ce coefficient.

Détermination des intervalles de confiance

1° La marge d'erreur a été fixée à 5 %,

2° La transformation de r en Z score de Fisher, soit la relation suivante :

$$Z_r = \frac{1}{2}\ln\left(\frac{(1+(-0{,}7045))}{(1-(-0{,}7045))}\right) : \frac{1}{2}\ln\left(\frac{0{,}2955}{1{,}7045}\right) : \frac{1}{2}\ln(0{,}1733646) : \frac{1}{2}(-1{,}7524) : -0{,}8762$$

3° Multiplier $1/\sqrt{(n-3)}$ par 1,96 ;

$\sqrt{n-3} : \sqrt{15-3}$ = 3,4641 ; $\frac{1}{3{,}4641} = 0{,}2887$; $0{,}2887 * 1{,}96 = 0{,}5658$;

4° La limite inférieure est égale à : -0,8762 – 0,5658 = -1,442 et la limite supérieure est égale à : -0,8762 + 0,5658 = -0,3104 ;

5° Consulter à nouveau la table F de Fisher ou utiliser la relation suivante $r = \frac{e^{2*z_r}-1}{e^{2*z_r}+1}$ pour retrouver les *r* correspondants aux valeurs de Z_r (*Z* de Fisher) calculées. Soit :

$$r_{inf} = \frac{e^{2*(-1{,}442)}-1}{e^{2*(-1{,}442)}+1} : \frac{e^{(-2{,}884)}-1}{e^{(-2{,}884)}+1} : \frac{0{,}0559-1}{0{,}0559+1} : \frac{-0{,}9441}{1{,}0559} : -0{,}8941$$

$$r_{sup} = \frac{e^{2*(-0{,}3104)}-1}{e^{2*(-0{,}3104)}+1} : \frac{e^{(-0{,}6208)}-1}{e^{(-0{,}6208)}+1} : \frac{0{,}5375-1}{0{,}5375+1} : \frac{-0{,}4625}{1{,}5375} : -0{,}3008$$

La vraie valeur de coefficient de corrélation dans la population se retrouve, au niveau de confiance de 95 %, dans l'intervalle ci-après :

-0,8941 < r > -0,3008

Rappelons qu'actuellement, tous ces calculs se réalisent automatiquement sur l'ordinateur à l'aide des applications spécifiques. Le sens du développement manuel qui précède consistait à exposer la démarche pour permettre aux lecteurs de savoir comment les choses se réalisent à partir des applications statistiques spécialisées.

Calcul du coefficient de détermination

En se référant au Tableau 4, le calcul de coefficient de détermination revient à élever la covariance au carré et à la diviser par le produit de variance de chacune de deux variables ou simplement à élever le coefficient de corrélation r au carré, soit la formule suivante :

$$R^2 : (r)^2 = \frac{S_{xy}^2}{S_x^2 * s_y^2} = \frac{(-11{,}5)^2}{8{,}67 * 30{,}74} = 0{,}4963$$

On note ici que le nombre d'enfants que les femmes ont eus est expliqué à près de 50 % par le nombre d'années qu'elles ont passé à l'école, les autres facteurs non pris en compte ici se partagent les 50 % restants. En d'autres termes, près de 50 % de la variation du nombre d'enfants sont expliqués par le nombre d'années de scolarité des femmes et 50 % restants de la variation du nombre d'enfants sont expliqués par des facteurs non pris en compte ici.

Test de signification de r ou de R^2

On recourt pour cela au test F, dont la formule est la suivante :

$$F = \frac{R^2(n-2)}{1-R^2}$$

L'hypothèse nulle (H_0) à tester ici pose l'absence de relation entre le nombre d'années de scolarité des femmes et le nombre de leurs enfants, soit $r=R^2=0$. Si le test *F* indique un niveau de signification inférieur à 5 %, on rejette l'H_0 et en accepte l'H_1.

$$F = \frac{(0{,}4963)^2\ (15-2)}{1-0{,}4963} = 12{,}84$$

C'est donc ce F_c (Coefficient de Fisher calculé) que l'on doit comparer avec le F_t de la table. Pour entrer dans la table, on a besoin de 3 repères : la marge d'erreur, généralement fixée à 5 % (0,05) ; les degrés de liberté. Les degrés de liberté pour ce test sont 1 (pour la première colonne) et N-2 (pour la ligne) (Dans cet exemple, N-2=13, soit 15-2). Avec ces entrées, la table de Fisher (en annexe 1) indique un *F* théorique de 4,67. Si F_c calculé est plus grand que la valeur correspondante de la table de distribution théorique de *F,* rejeter l'H_0 et accepter l'H_1. Etant donné que F_c (12,84) est > au F_t (4,67), l'H_0 ne peut être acceptée.

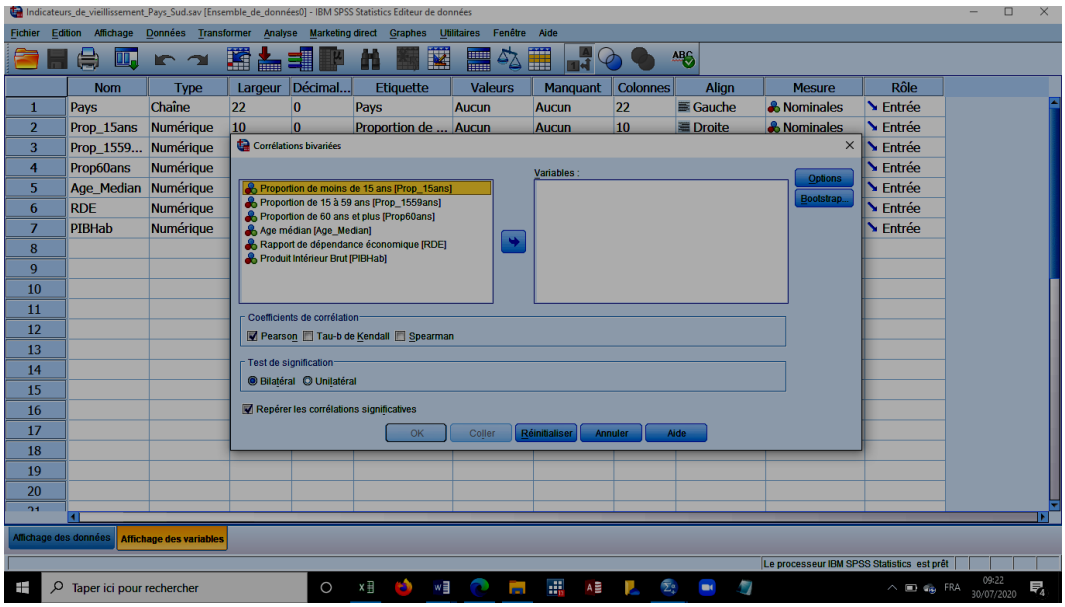
Construction de la matrice de corrélation

Nous avons déjà indiqué que l'on recourt à la matrice de corrélation lorsque l'on dispose de plus de deux variables et que l'on voudrait se faire une idée sur la relation entre elles. Plutôt que de répéter la même opération plusieurs, on doit lancer une seule fois toutes les variables et obtenir les différents coefficients de corrélation de variables prises deux à deux, le nombre d'observations, le test de signification et les intervalles de confiance. Il est quasi-impossible de construire la matrice de corrélation manuellement, il sera dès lors fait recours à l'application SPSS, les données utilisées sont celles relatives à quelques indicateurs de vieillissement démographique et de développement de quelques pays du Sud autour de l'année 2010 (Voir Tableau des données, en annexe 2).

La démarche consiste d'abord à entrer ces données sur l'éditeur de SPSS. Une fois que les données sont saisies, procéder comme suit :

1° Cliquer sur Analyse→ Corrélation→ Bivariée ; la boîte de dialogue ci-dessous s'affiche :

Figure 12 : Boîte de dialogue Corrélation bivariée



2° Sélectionner toutes les variables pour lesquelles vous souhaitez obtenir des coefficients de corrélation et les faire basculer dans l'espace approprié à droite ;

3° Eventuellement, cliquer sur le bouton Option pour demander certains statistiques comme la moyenne et l'écart-type de chacune de variables et gérer les données manquantes, cliquer ensuite sur Poursuivre pour valider tous les choix ;

4° Eventuellement, cliquer sur le bouton Bootstrap pour appliquer ce test (cette opération n'est pas indispensable) et cliquer sur Poursuivre ;

5° Enfin, cliquer sur OK pour afficher la matrice de corrélation qui suit

Tableau 5 : Matrice de corrélation entre les indicateurs de vieillissement démographique et du développement économique

Corrélations

		Prop_15ans	Prop_1559ans	Prop60ans	Age_Median	RDE	PIBHab
Prop_15ans	Corrélation de Pearson	1	-,948**	-,742**	-,954**	,957**	-,611**
	Sig. (bilatérale)		,000	,000	,000	,000	,000
	N	109	109	109	109	109	106
Prop_1559ans	Corrélation de Pearson	-,948**	1	,490**	,835**	-,989**	,619**
	Sig. (bilatérale)	,000		,000	,000	,000	,000
	N	109	109	109	109	109	106
Prop60ans	Corrélation de Pearson	-,742**	,490**	1	,854**	-,537**	,361**
	Sig. (bilatérale)	,000	,000		,000	,000	,000
	N	109	109	109	109	109	106
Age_Median	Corrélation de Pearson	-,954**	,835**	,854**	1	-,848**	,619**
	Sig. (bilatérale)	,000	,000	,000		,000	,000
	N	109	109	109	109	109	106
RDE	Corrélation de Pearson	,957**	-,989**	-,537**	-,848**	1	-,580**
	Sig. (bilatérale)	,000	,000	,000	,000		,000
	N	109	109	109	109	109	106
PIBHab	Corrélation de Pearson	-,611**	,619**	,361**	,619**	-,580**	1
	Sig. (bilatérale)	,000	,000	,000	,000	,000	
	N	106	106	106	106	106	106

**. La corrélation est significative au niveau 0.01 (bilatéral).

Cette matrice fournit les informations suivantes : les différents coefficients de corrélation entre les couples de variables placées en lignes et celles placées en colonnes, l'indication de la signification statistique associée à chaque coefficient et le nombre d'observations prises en compte pour le calcul de chaque coefficient. On aurait pu obtenir aussi, si on l'avait demandé, les intervalles de confiance associés à chaque coefficient de corrélation. Comme c'était déjà indiqué, les coefficients de corrélation sont calculés en combinant les variables deux à deux. C'est à l'intersection de chaque ligne et de chaque colonne que se trouve le coefficient de corrélation (r) du couple formé par la variable définie en ligne et celle définie en colonne ; la diagonale principale de la matrice représentant le r croisant une variable à elle-même. Cette diagonale, disions-nous, coupe la matrice en deux triangles identiques. Ainsi, dans la présentation finale des résultats, il faut supprimer la diagonale et ne garder qu'un des triangles. La matrice à présenter dans un rapport scientifique pourrait donc se ressembler à ce qui suit :

Tableau 5 bis : Matrice de corrélation entre les indicateurs de vieillissement démographique et du développement économique

Variables	Prop_15ans	Prop1559ans	Prop60ans+	Age_Médian	RDE
Prop1559ans	-0,95***				
	(109)				
Prop60ans+	-0,74***	0,49***			
	(109)	(109)			
Age_Médian	-0,95***	0,84***	0,85***		
	(109)	(109)	(109)		
RDE	0,95***	-0,99***	-0,54***	-0,85***	
	(109)	(109)	(109)	(109)	
PIB/Hab.	-0,61***	0,62***	0,36***	0,62***	-0,58***
	(106)	(106)	(106)	(106)	(106)

*** : Significatif à 1 %

Les chiffres entre parenthèses représentent le nombre d'observations prises en compte dans le calcul de chaque coefficient r. Si la taille de l'échantillon est la même pour tous les coefficients, il serait inutile de présenter ce chiffre dans la matrice.

On peut procéder par la suite à l'interprétation de ces différents coefficients r comme indiqué précédemment. Par exemple, on note que le lien entre la proportion de moins de 15 ans et le PIB/hab. est forte et négative (r = -0.61). Ainsi, si la proportion des enfants de moins de 15 ans augmente, le PIB/Hab. a tendance a baissé, ce résultat est statiquement signification au seuil de 1 %. Ce qui revient à dire que l'on a 99 % de chance que ce résultat soit également observé au niveau de la population dans son ensemble. Pour trouver les coefficients de détermination correspondants, il suffit d'élever chaque coefficient de corrélation au carré.

Intérêts et limites du diagramme de dispersion et de la corrélation linéaire

Le diagramme de dispersion est très utile pour se faire une idée sur la forme et le sens de relation entre deux variables quantitatives, il en offre un aperçu visuel. Mais sa lecture devient fastidieuse lorsque le nombre d'observations est trop grand

et que les scores de variables ont un nombre limité des valeurs. Dans ce cas, les points du diagramme s'empilent les uns et les autres. C'est ainsi que le diagramme de dispersion est plus utile pour des données agrégées que pour des données désagrégées ou individuelles. Par ailleurs, le diagramme de dispersion ne permet pas en lui-même de se faire une idée précise sur l'intensité ou la force de cette relation. C'est pourquoi son examen doit être complété par le calcul du coefficient de corrélation (linéaire) (r) de Bravais-Pearson.

En rapport avec le coefficient de corrélation, un de ses intérêts est justement le fait qu'en plus de détecter l'existence ou non du lien entre deux variables, il permet de quantifier cette relation et d'en indiquer le sens. Le coefficient de corrélation permet ainsi de hiérarchiser les relations entre variables. Sa principale limite tient du fait qu'il permet de n'étudier que des relations linéaires. Il n'est pas possible, avec le coefficient de corrélation, d'étudier de relation de nature autre que linéaire (en forme de U inversée, curvilinéaire, etc.). Ainsi, l'absence de relation linéaire entre variables induite par le coefficient de corrélation ne signifie pas forcément que les variables en présence n'entretiennent aucune relation entre elles. C'est pourquoi, il est conseillé de faire précéder le calcul de coefficient de corrélation par la visualisation de relation entre variables à travers le diagramme de dispersion.

Une autre limite de coefficient de corrélation, qu'il partage avec les autres méthodes et techniques descriptives, est que la confirmation d'une relation linéaire entre deux variables n'induit pas nécessairement une causalité entre elles. C'est ainsi que même dans le cas d'une relation positive parfaite comme dans l'exemple de la Figure 2, on ne peut pas forcément déduire à une causalité. En effet, même si le lien entre l'âge du père (X) et celui de son l'enfant (Y) est positif et parfait, on ne peut pas déduire que c'est l'augmentation de l'âge du père qui fait augmenter aussi l'âge de son enfant. Il y a probablement une troisième variable (la durée de vie ou le temps écoulé) qui agit simultanément et de la même façon à la fois sur l'âge du père et sur celui de son enfant. La non-causalité tient notamment du fait que le coefficient de corrélation r est une mesure symétrique en ce sens qu'il ne distingue la variable dépendante de la variable indépendante. En effet, la corrélation de X avec Y (r_{xy}) est identique à la corrélation de Y avec X (r_{yx}). Enfin, le coefficient r est sensible à la taille de l'échantillon. Plus cette taille est grande, plus le r est significatif, même si sa valeur absolue est faible. On considère qu'échantillon est de grande taille lorsqu'elle atteint au moins 30 unités. A l'inverse, des échantillons de petite taille produisent de r élevés mais non significatifs.

Conclusion

Cet article a porté sur l'analyse de la corrélation en tant que méthode permettant de mesurer le lien entre variables quantitatives prises deux à deux. La mesure de ce lien est évaluée à partir d'une statistique appelée « Coefficient de corrélation », proposée par des statisticiens Bravais et Pearson. Au-delà de la force de lien entre variables, le coefficient de corrélation indique également le sens ou la direction de cette relation.

C'est une mesure facile à calculer et d'interprétation relativement aisée. Toutefois, son utilisation obéit à un certain nombre de conditions dont la normalité de la distribution, l'absence des valeurs aberrantes, le caractère aléatoire de l'échantillon et le caractère linéaire de la relation attendue. C'est ainsi que pour être correct, on devrait plutôt parler du coefficient de corrélation linéaire de Bravais-Pearson. Si la condition de normalité est difficile à vérifier dans la pratique, on recourt à l'examen visuel du diagramme de dispersion pour s'assurer de la condition de linéarité. L'absence de linéarité ne doit pas conduire nécessairement à l'absence de relation entre variable, elle doit plutôt permettre

d'envisager (selon la forme du nuage des points) à l'examen d'autres types de relations possibles, notamment les relations sous forme de U inversé ou sous forme de courbe.

A noter également, même si la corrélation reste de loin le test le plus utilisé pour mesurer la force de la relation entre variable quantitative, d'autres alternatives existent comme êta carré (η^2), le V de Cramer et le coefficient de corrélation de Spearman. Par ailleurs, même si dans le langage courant on utilise le terme « corrélation » pour désigner les liens entre phénomènes ou entre variables, ce terme est approprié uniquement pour le cas des variables quantitatives. Enfin, on ne le dira jamais assez, corrélation n'est pas causalité. Une corrélation, même parfaite, entre deux variables, n'implique pas nécessairement qu'une variable est la cause de l'autre. La corrélation n'est qu'une analyse préliminaire et ne permet nullement de conclure à une relation de cause à effet entre deux variables, même s'il n'y a pas de causalité sans corrélation. Pour besoin d'étude de la causalité, compléter l'analyse de la corrélation par d'autres types d'analyse, notamment la régression linéaire.

Références bibliographiques

Broc, G. et Caumeil, B. (2018). *Analyse de données.* Louvain-la-Neuve : Deboeck Supérieur.

Costa, R. (2013). Deux variables quantitatives. In Masuy-Stroobant, G. et Costa, R. *Analyser les données en sciences sociales. De la préparation des données à l'analyse multivariée.* Bruxelles : P.I.E. Peter Lang, pp.117-141.

Dagnelie, P. (1998). *Statistique théorique et appliquée.* Tome 2. Inférence statistique à une et à deux dimension. Paris et Bruxelles : De Boeck et Larcier.

Dancey, C. P et Reidy, J. (2016). *Statistiques sans maths pour les psychologues* (2è édition française). Louvain-la-Neuve : Deboeck Supérieur.

Fox, W. (1999). *Statistiques sociales.* (3ième édition) (L. Imbeau, Traduction). Bruxelles : DeBoeck Université. (Travail original publié en 1998).

Ghewy, P. (2010). *Guide pratique de l'analyse de données. Avec applications sous IBM SPSS Statistics et Excel. Questionnez, analysez…et décidez.* Bruxelles : De Boeck.

Larose, D. T et Larose, D. C. (2018). *Data mining. Découverte de connaissances dans les données.* 2ième édition. (T. Vallaud, Traduction). Paris : Vuibert. (Travail original publié en 2014).

Mangalu, M.A. (2004). *Relations entre vieillissement démographique et développement socio-économique dans les pays du Sud.* (Mémoire de DEA). Université catholique de Louvain, Louvain-la-Neuve.

Martin, O. (2009). *L'enquête et ses méthodes. L'analyse de données quantitatives* (2è édition). Paris : Armand Colin.

Py, B. (2007). *La statistique sans formule mathématique. Comprendre la logique et maîtriser les outils.* Paris : Pearson Education.

Rosnow, R. L. & Rosenthal, R. (1996). Computing contrasts, effect sizes and counternulls on other people's published data: general procedures for consumers' research. *Psychological Methods,* 1(14), 331-340.

Wiley, J. (1991). *Statistiques. Economie-Gestion-Sciences-Médecine.* 4ième édition. (T.H. Wonnacott & R.J. Wonnacott, Traduction). Paris : Economica.

Annexe 1 : Table de la loi de Fisher-Snedecor, $\alpha = 5\%$

Annexe 2 : Indicateurs du vieillissement démographique (en %) et du développement (en US$) des pays du Sud autour de l'an 2000

Pays	Prop_15ans	Prop15_59ans	Prop60ans+	Age_Moyen	RDE	PIB/Hab en US$
Afghanistan	48,6	47,7	3,7	15,6	1,1	1083
Afrique du sud	29,7	62,2	8,1	25,2	0,61	9678
Algérie	27,1	66,1	6,8	26	0,51	7643
Angola	47,8	48,4	3,8	16	1,07	5201
Arabie saoudite	30,7	65	4,3	26,1	0,54	21430
Argentine	24,9	60,5	14,7	30,3	0,65	15501
Arménie	20,5	65,8	13,7	31,6	0,52	5112
Azerbaïdjan	22,7	69,4	7,9	28,6	0,44	8890
Bangladesh	31,7	61,6	6,8	24	0,63	1568
Benin	43,4	52,1	4,5	18,1	0,92	1428
Bhutan	29,8	63,5	6,7	24,4	0,57	5096
Bolivie	36	56,9	7,1	21,7	0,76	4499
Botswana	34,3	60,4	5,3	22	0,66	12939
Brésil	25,5	64,3	10,2	29	0,56	10278
Burkina Faso	46	50,1	3,9	16,8	1	1149
Burundi	43,9	52,2	3,8	17,7	0,91	533
Cambodge	31,8	60,9	7,2	23,5	0,64	2080
Cameroun	43,4	51,7	4,9	18	0,93	2090
Cap Vert	47,7	47,8	4,5	22,7	1,09	3616
Chili	22,1	64,8	13,1	32,1	0,54	15272
Colombie	28,8	62,6	8,6	26,8	0,6	8861
Comores	42,2	53,3	4,5	19,1	0,88	980
Congo	42,2	52,6	5,1	18,9	0,9	3885
Corée du Sud	16,2	68,2	15,6	37,8	0,47	27541
Costa-Rica	24,9	65,6	9,5	28,4	0,52	10732
Côte d'Ivoire	47,7	47,8	4,5	18,8	1,09	1581
Cuba	17,3	65,7	17	38,4	0,52	
Djibouti	34,1	60	5,8	22	0,66	2087
Egypte	31,5	60	8,5	24,4	0,67	5547
El Salvador	32,1	58,5	9,4	23,1	0,71	6032
E.A.U	13,9	85,4	0,7	28	0,17	42293
Equateur	31	60,3	8,8	25,2	0,66	7443
Erythrée	43	53,3	3,7	18,3	0,88	516
Ethiopie	44,4	50,5	5,1	17,5	0,98	979
Gabon	38,6	53,8	7,5	20,5	0,86	13998
Gambie	46	50,3	3,7	16,9	0,99	1873
Ghana	39	55,6	5,4	20,2	0,8	1652
Guatemala	41,5	52	6,4	18,8	0,92	4351
Guinée	42,8	52,2	5	18,3	0,92	990
Guinée Bissau	41,9	53,4	4,7	18,8	0,87	1097

Guinée Equat.	39,3	56,3	4,4	20,2	0,78	32026
Haïti	36,2	57,3	6,5	21,5	0,75	1034
Honduras	36,8	57	6,2	20,9	0,75	3566
Indonésie	29,8	62,6	7,6	26,9	0,6	4094
Iran	23,6	69	7,4	27	0,45	4094
Iraq	41,2	53,9	4,8	19,1	0,85	3412
Israël	27,2	57,9	14,9	30,1	0,73	26720
Jamaïque	29	60,4	10,6	27	0,66	7074
Jordanie	35,1	59,7	5,3	22,5	0,68	5269
Kazakhstan	24,9	65,3	9,8	28,9	0,53	2126
Kenya	42,6	53,3	4,1	18,5	0,88	1507
Kirghizstan	30	63,6	6,3	23,8	0,57	2126
Koweït	25,2	70,9	3,9	28,4	0,41	47935
Lesotho	37,6	56	6,3	20,1	0,78	1504
Liban	23,7	64,4	11,9	28,5	0,55	12900
Liberia	43,3	51,9	4,8	18,3	0,93	506
Libye	29,4	63,9	6,7	25,6	0,56	15361
Madagascar	43,4	52,2	4,4	18	0,92	853
Malaisie	27,7	64,6	7,8	26,1	0,55	13672
Malawi	45,8	49,3	4,9	16,9	1,03	805
Mali	46,8	48,8	4,4	16,5	1,05	964
Maurice	21,2	66,6	12,2	33,3	0,5	12737
Mauritanie	40,6	54,6	4,9	19,5	0,83	2255
Mexique	30	61,3	8,6	26	0,63	12776
Mongolie	27	67,4	5,7	25,7	0,49	4178
Maroc	28,1	64,6	7,3	26,2	0,55	4373
Mozambique	45,3	49,7	5	17,2	1,01	861
Myanmar	26,1	66,2	7,7	27,8	0,51	
Namibie	37,7	57,1	5,3	20,3	0,75	5986
Népal	37,1	55,5	7,4	21,3	0,8	1102
Nicaragua	34,5	59,3	6,2	22	0,69	2597
Niger	49,8	45,9	4,3	15,1	1,18	642
Nigeria	44	51,4	4,5	17,9	0,94	2221
Oman	27,4	68,6	4	25,1	0,46	25330
Ouganda	48,9	47,4	3,7	15,5	1,11	1188
Ouzbékistan	29,8	64,1	6,2	24,1	0,56	2903
Pakistan	35,4	58,2	6,4	21,6	0,72	2424
Palaos	36,8	57,6	5,6	20,3	0,74	13176
Panama	29,3	61	9,7	27,1	0,64	13788
Paraguay	33,5	58,8	7,7	23,1	0,7	4752
Pérou	30	61,3	8,7	25,5	0,63	9049
Philippines	35,3	58,9	5,9	22,3	0,7	3631
RCA	40,6	53,6	5,8	19,3	0,87	716
RDC	45,5	50	4,5	17,1	1	329
Rép. dominicaine	31,2	60,3	8,6	25	0,66	8651
Rwanda	44,7	51,5	3,8	17,8	0,94	1097
Sao Tomé et Prin.	41,6	53,5	4,9	19	0,87	1805

Sénégal	43,6	51,7	4,6	17,9	0,93	1737
Seychelles	22,3	67,2	10,5	31,6	0,49	23172
Sierra Leone	42,2	53,4	4,4	18,8	0,87	769
Singapour	17,3	68,6	14,1	37,3	0,46	53591
Sri Lanka	25,1	63,1	11,8	30,5	0,58	4929
Soudan	42,1	53,1	4,9	18,7	0,89	1878
Swaziland	38,8	56	5,2	19,3	0,79	5349
Tadjikistan	35,9	59,3	4,8	21,2	0,69	2052
Tanzanie	44,8	50,3	4,9	17,4	0,99	1334
Tchad	48,8	47,3	3,8	15,5	1,11	1343
Thaïlande	19,3	67,7	12,9	35,4	0,48	7633
Togo	42,1	53,5	4,4	18,7	0,87	914
Trinite et Tobago	20,7	66,8	12,4	31,9	0,5	22671
Tunisie	23,5	66,6	9,9	29	0,5	8258
Turkménistan	29,2	64,7	6,1	24,5	0,55	8055
Turquie	26,7	63,2	10,1	28,3	0,58	13466
Uruguay	22,5	59,1	18,4	33,7	0,69	13315
Venezuela	29,5	61,9	8,6	26,1	0,62	11258
Viet Nam	23,5	67,6	8,9	28,5	0,48	3013
Yémen	42	53,6	4,5	18,2	0,87	2060
Zambie	46,9	49,1	4	16,5	1,04	1423
Zimbabwe	41,2	53,1	5,6	18,5	0,88	