Liaisons entre variables non numÃ©riques

Soumis par Ã‰ric Raufaste
DerniÃ¨re mise Ã jour : 04-01-2013

Objectifs. Ã‰tendre le concept de liaison entre variables, de corrÃ©lation, aux cas oÃ¹ les variables considÃ©rÃ©es ne sont pas toutes numÃ©riques.

PrÃ©requis.

- Les deux articles DÃ©pendances entre variables et Liaisons entre variables : la corrÃ©lation linÃ©aire sont essentiels pour comprendre cette leÃ§on.

-
Au plan de la technique mathÃ©matique, il est aussi nÃ©cessaire d'avoir vu l'article sur la somme algÃ©brique.

RÃ©sumÃ©. Le principe gÃ©nÃ©ral de corrÃ©lation ayant Ã©tÃ© vu dans le cas numÃ©rique, le prÃ©sent article prÃ©sente des extensions Ã un petit nombre de cas particuliers importants : 2 variables ordinales (rho de Spearman et Tau de Kendall);
2 variables nominales... Â Â

Â Â

1. Liaison entre deux variables ordinales

Â

Rappelons que les variables ordinales sont des variables telles que leurs valeurs respectent une relation d'ordre (on peut les ordonner de la plus petite Ã la plus grande ou rÃ©ciproquement) mais sans que la propriÃ©tÃ© d'Ã©galitÃ© des intervalles n'ait de sens. Par exemple, le 1 est mieux classÃ© que le deuxiÃ¨me de la classe qui lui mÃªme est mieux classÃ© que le troisiÃ¨me. Mais rien ne permet de garantir que l'Ã©cart de performance entre le premier et le deuxiÃ¨me soit Ã©quivalent Ã l'Ã©cart de performance qui existe entre le deuxiÃ¨me et le troisiÃ¨me.Â

Â

Il existe principalement deux extensions de la corrÃ©lation linÃ©aire aux cas des variables ordinales. Ce sont le rho de Spearman et le Tau de Kendall.

2.1. Le coefficient de corrÃ©lation des rangs de Spearman

Le coefficient de Spearman se calcule exactement comme la coefficient de corrÃ©lation linÃ©aire Ã un point prÃ¨s : on commence par recoder les deux variables en termes de rangs avant de faire le calcul de la corrÃ©lation. Pour illustrer cela, imaginons les donnÃ©es suivantes, pour lesquels nous voulons calculer une corrÃ©lation de rangs entre les variables V1 et V2 :

Â Â Â Â Â Â Â Â Â SujetÂ Â Â Â Â Â V1Â Â Â

Â 1Â Â Â Â Â Â

Â 2Â Â Â Â Â Â

Â 3Â Â Â Â Â

Â 4Â Â Â Â Â Â

Â 5Â Â Â Â Â Â

Â 6Â Â Â Â Â Â

Ce recodage en termes de rangs se fait selon la logique suivante : les observations sont ordonnÃ©es de la plus petites Ã la plus grande (ou l'inverse). La plus forte prend le rang 1, la deuxiÃ¨me le rang 2, etc... Pour ce qui concerne lesÂ ex-aequo, on remplace simplement leur valeur par la moyenne des rangs obtenus par cette valeur. Dans l'exemple prÃ©cÃ©dent, la variable V1 donne le reclassement suivant (du plus grand au plus petit) :Â Â

Â SujetÂ Â Â Â Â Â

Â 1Â Â Â Â Â Â

Â 2Â Â Â Â Â Â

Â 3Â Â Â Â Â

Â 4Â Â Â Â Â Â

Â 5Â Â Â Â Â Â

Â 6Â Â Â Â Â Â

La valeur 3 prend les rang 2 et 3, ce qui fait un rang moyen de 2,5 pour chacune de ces deux valeurs et finalement nous avons le recodage suivant :

Â SujetÂ Â Â Â Â Â

Â 1Â Â Â Â Â

Â 2Â Â Â Â Â

Â 3Â Â Â Â Â

Â 4Â Â Â Â Â

Â 5Â Â Â Â Â

Â 6Â Â Â Â Â

Â On applique la mÃªme logique sur la variable V2 et on obtient une variable recodÃ©e V2R:

Â SujetÂ Â Â Â Â V2

Â 1Â Â Â Â Â

Â 2Â Â Â Â Â

Â 3Â Â Â Â Â

Â 4Â Â Â Â Â

Â 5Â Â Â Â Â

Â 6Â Â Â Â Â

Si nous calculons alors la corrÃ©lation linÃ©aire classique entre les variables V1R et V2R, nous trouvons la valeurÂ 0,161 qui est le rho de Spearman.

Le rho de spearman s'interprÃ¨te exactement comme une corrÃ©lation classique, qui varie entre -1 et +1, avec 0 signifiant l'absence de corrÃ©lation.

Â

2.2. Le Ï„ (Tau) de Kendall

Pour calculer le Tau de Kendall, il faut commencer par trier les observations par ordre croissant (ou dÃ©croissant de l'une des deux variables). Cela nous donne donc un ordre parfait (aux ex-aequo prÃ¨s) sur la premiÃ¨re variable mais pas sur la deuxiÃ¨me variable. On va alors comparer toutes les paires possibles de valeurs Ã l'intÃ©rieur de la seconde variable. On appelera paire concordante une paire qui ira dans le mÃªme sens que la variable nÂ°1 et paire discordante toute paire qui ne va pas dans le mÃªme sens que la variable nÂ°1. Le tau de Kendall s'obtient alors par la formule suivante :

{moslatex}\Large{$\tau = \frac{n_c-n_d}{\frac{1}{2}{n(n-1)}}$}{/moslatex}Â

oÃ¹ nc est le nombre de paires concordantes et nd le nombre de paires discordantes.

Voyons cela sur un exemple. Nous partons des donnÃ©es suivantes :

Â SujetÂ Â Â Â Â Â Â V1Â Â Â

Â 1Â Â Â Â Â

Â 2Â Â Â Â Â

Â 3Â Â Â Â Â

Â 4Â Â Â Â Â

Â 5Â Â Â Â Â

Â 6Â Â Â Â Â
Â

Â
En retriant ces donnÃ©es selon la variable V1, nous obtenons :

Â SujetÂ Â Â Â Â V1Â Â Â

Â 1Â Â Â Â Â

Â 6Â Â Â Â Â Â

Â 3Â Â Â Â Â

Â 4Â Â Â Â Â

Â 5Â Â Â Â Â

Â 2Â Â Â Â Â

Nous nous intÃ©ressons maintenant aux paires de donnÃ©es de la variable V2. Nous allons compter comme discordantes toutes les paires d'observations oÃ¹ la seconde valeur sera plus petite que la premiÃ¨re (elles vont dans le sens inverse de l'ordre dÃ©fini pour la variable V1) et concordantes les paires telles que la deuxiÃ¨me valeur est plus grande que la premiÃ¨re. Ainsi, si nous comparons les sujets 1 et 6, nous voyons que la premiÃ¨re valeur (3,4) est plus grande que la seconde (1,1) et donc la paire est discordante puisqu'elle ne va pas dans le mÃªme sens que l'ordre de la variable V1 (du plus petit au plus grand). Nous comptons encore 1 discordance pour la comparaison des sujets 1 et 3, etc... la paire sujet 1 - sujet 2 par contre est concordante (puisque 3,4 est plus petit que 6,5). Nous continuons avec la paire sujet 6 - sujet 3 qui est concordante). Nous examinons ainsi toutes les paires possibles, ce qui nous donc nc= 8; nd = 7.
Si nous appliquons la formule plus haut, nous trouvons donc

{moslatex}\Large{$\tau = \frac{8-7}{\frac{1}{2}{(6\times 5)}}$}{/moslatex}Â

Â

soit 0,067.

Notons que le dÃ©nominateur de la formuleÂ

{moslatex}\large{$\frac{1}{2}n(n-1)$}{/moslatex}Â

Â reprÃ©sente le nombre total de paires possibles. Il y a en effetÂ

{moslatex}\Large{$C_n^2=\frac{n!}{2!(n-2)!}=\frac{n\times (n-1)\times \ldots\times 2}{2\times (n-2)\times (n-3)\times \ldots\times 2}$}{/moslatex}Â

combinaisons de deux Ã©lÃ©ments pami n. AprÃ¨s simplification, cela fait,

{moslatex}\Large{$\frac{n\times (n-1)}{2}$}{/moslatex}Â

Â

On voit donc que le coefficient de corrÃ©lation de Kendall reprÃ©sente en fait le degrÃ© auquel les deux variables sont rangÃ©es dans le mÃªme ordre. Il faut cependant noter que les ex-aequo ne sont ni des paires concordantes, ni des paires discordantes et qu'il faut donc introduire un facteur de correction en cas d'ex-aequo. Â

Â

{mospagebreak heading=1. CorrÃ©lation entre variables ordinales &title=2. Liaison entreÂ deux variables nominales}

Â Â

2. Liaison entre deux variables nominales

Rappelons que les variables nominales sont des variables telles que
leurs valeurs rcorrespondent Ã des catÃ©gories discrÃ¨tes mutuellement exclusives, mais qu''il nexiste pas de relation d'ordre entre ces modalitÃ©s. On ne peut mÃªme pas les classer entre elles. Tout ce que l'on peut faire c'est donc de compter combien d'individus tombent dans chaque modalitÃ©.

Dans ces conditions, comment peut-on Ã©valuer la liaison entre deux variables ? Un premier Ã©lÃ©ment de solution consiste Ã s'interroger sur ce que signifierait au contraire l'indÃ©pendance, l'idÃ©e Ã©tant que si l'on sait Ã©valuer Ã quel degrÃ© des variables sont indÃ©pendantes, cela nous donne de facto une indication sur le degrÃ© auquel elles sont liÃ©es. Dans la grande leÃ§on sur la Statistique descriptive, nous avons vu que
des variables peuvent Ãªtre considÃ©rÃ©es indÃ©pendantes si les valeurs

prises par une variable ne sont pas affectÃ©es par les valeurs prises sur l'autre variable. Puisqu'ici nous travaillons avec des variables nominales, cela revient Ã dire que la distribution des effectifs dans les modalitÃ©s d'une variable ne sera pas affectÃ©e par la distribution des effectifs dans l'autre variable. C'est l'idÃ©e sous-jacente au Khi-2.

2.1. Le Khi-2 d'indÃ©pendance (ou Khi-deux, ou Chi-2, ou encore Ï‡2).

Ã€ titre d'exemple, commenÃ§ons par considÃ©rer deux variables nominales dichotomiques, une variable Sexe avec deux modalitÃ©s, Homme et Femme, et une variable Salaire Ã deux modalitÃ©s, haut salaire et bas salaire. On s'intÃ©resse Ã savoir s'il y a un lien entre le sexe et le salaire. On a un Ã©chantillon de sujets, et le tri croisÃ© donne le tableau de donnÃ©es suivant :

{moslatex}\footnotesize{\begin{tabular}{cccc}

Â Â Â Â Â Â Â Â Â Â & Bas salaire & Haut salaire &Â Â Â Â Â Total \\

Â Â Â Â Homme &Â Â Â Â Â Â Â Â 250 &Â Â Â Â Â Â Â Â 150 &Â Â Â Â Â Â Â Â 400 \\

Â Â Â Â Femme &Â Â Â Â Â Â Â Â 128 &Â Â Â Â Â Â Â Â 102 &Â Â Â Â Â Â Â Â 230 \\

Â Â Â Â Total &Â Â Â Â Â Â Â Â 378 &Â Â Â Â Â Â Â Â 252 &Â Â Â Â Â Â Â Â 630 \\

\end{tabular}}{/moslatex}Â

Â

La premiÃ¨re chose que nous remarquons en regardant les marges du tableau, c'est-Ã -dire la colonne Total et la ligne Total, c'est que dans cet Ã©chantillon la proportion des femmes est plus faible que la proportion des hommes. Par ailleurs nous voyons aisÃ©ment que la proportion des bas salaires est plus importante que celle des haut salaires. Seulement voilÃ , il ne s'agit lÃ que de considÃ©rations relatives chacune Ã une variable prise isolÃ©ment, sans considÃ©ration de l'autre variable. Or, ce qui nous intÃ©resse c'est de comprendre la liaison Ã©ventuelle entre les variables, donc de considÃ©rer les deux variables SIMULTANÃ‰MENT.

Â

Ce sont les cellules Ã l'intÃ©rieur du tableau qui nous donnent ces informations. Ainsi, par exemple, la case Ã l'intersection de la colonne 1 et de la
ligne 1, nous renseigne sur les hommes ayant un bas salaire. Il va donc
falloir comparer l'effectif observÃ© dans cette case avec l'effectifÂ thÃ©orique qu'on
aurait en cas d'indÃ©pendance des deux variables.

Â

Comment obtenir l'effectif thÃ©orique ? Pour traiter cette question, nous allons considÃ©rer que les marges du tableau nous donnent des informations sur les taux de base, c'est-Ã -dire les distributions de frÃ©quences a priori de notre Ã©chantillon. Pour calculer l'effectif thÃ©orique de la case "Hommes Ã bas salaires", il nous suffit alors de multiplier le nombre total d'homme (case Total de la ligne 1) par le nombre total de personnes Ã bas salaires (case Total de la colonne 1), puis Ã diviser le rÃ©sultat par l'effectif gÃ©nÃ©ral (ici le contenu de la case en bas Ã droite, soit 630) pour que la somme des effectifs thÃ©oriques des 4 cases, donne un effectif thÃ©orique total qui soit identique Ã l'effectif observÃ©.

Autrement dit, dans le cas gÃ©nÃ©ral, en notant Li le total de la ligne i et Ci le total de la colonne j, l'effectif thÃ©orique de la cellule ij est donnÃ©e par la formule

Â {moslatex}\Large{$t_{ij}=\frac{L_i\times C_j}{n}$}{/moslatex}Â

Â

Cela nous donne alors le tableau des effectifs thÃ©oriques :

{moslatex}\footnotesize{\begin{tabular}{cccc}

Â Â Â Â Â Â Â Â Â Â & Bas salaire & Haut salaire &Â Â Â Â Â Total \\

Â Â Â Â Homme &Â Â Â Â Â Â Â 240 &Â Â Â Â Â Â Â Â 160 &Â Â Â Â Â Â Â Â 400 \\

Â Â Â Â Femme &Â Â Â Â Â Â Â Â 138 &Â Â Â Â Â Â Â Â 92 &Â Â Â Â Â Â Â Â 230 \\

Â Â Â Â Total &Â Â Â Â Â Â Â Â 378 &Â Â Â Â Â Â Â Â 252 &Â Â Â Â Â Â Â Â 630 \\

\end{tabular}}{/moslatex}Â

Â

Pour chaque case, nous sommes alors en mesure de calculer les Ã©carts Ã la valeur thÃ©orique en faisant simplement la diffÃ©rence entre les cellules correspondantes. Par exemple dans la case hommes Ã bas salaires, nous avons observÃ© 250 sujets alors que les effectifs thÃ©oriques n'Ã©taient que de 240. Autrement dit, nous avons un Ã©cart de +10 pour cette case : nous observons 10 hommes Ã bas salaires de plus que ce que nous aurions attendu sous l'hypothÃ¨se d'indÃ©pendance des variables.

Â

Afin d'Ã©viter les compensations, on Ã©lÃ¨ve chaque diffÃ©rence au carrÃ©. Il paraÃ®t cependant clair que la signification d'un sur-reprÃ©sentation (il y a plus d'individus qu'attendu thÃ©oriquement)
ou d'une sous-reprÃ©sentation (il y a moins d'individus qu'attendu
thÃ©oriquement) est Ã relativiser en fonction de l'effectif thÃ©orique de
la case :Â Le mÃªme Ã©cart de 10 sujets n'a pas le mÃªme sens selon qu'on
attendait 20 ou 3000 sujets dans la case. On divisera donc le carrÃ© de l'Ã©cart par l'effectif thÃ©orique afin d'obtenir un carrÃ© pondÃ©rÃ© de l'Ã©cart.

Â

Il ne reste alors qu'Ã sommer les carrÃ©s pondÃ©rÃ©s des Ã©carts obtenus dans chaque case pour obtenir le Ï‡2. Ainsi, dans le cas gÃ©nÃ©ral oÃ¹ la variable en ligne a L modalitÃ©s et la variable en colonne a C modalitÃ©s, et en notant oij l'effectif observé d'une case Ã l'intersection de la ligne i et de la colonne j, et tij son effectif thÃ©orique, nous obtenons la formuleÂ

{moslatex}\Large{$\chi^2=\Sigma_{i=1}^L\Sigma_{j=1}^C\frac{(o_{ij}-t_{ij})^2}{t_{ij}}$}{/moslatex}Â

Â

Dans l'exemple prÃ©cÃ©dent, cela nous donne donc Ï‡2=0,42+0,63+0,72+1,09 = 2,85.

Â

On voit donc immÃ©diatement que le Ï‡2 peut dÃ©passer 1. Il suffit aussi de considÃ©rer la formule pour se persuader qu'il ne saurait Ãªtre nÃ©gatif.Â Le Ï‡2 ne peut donc pas Ãªtre assimilÃ© Ã un coefficient de corrÃ©lation. En fait, comme nous l'avons dans l'article sur les distributions, la variable Ï‡2 suit une distribution particuliÃ¨re et c'est l'application de techniques statistiques fondÃ©es sur cette distribution qui nous permet ensuite de dÃ©cider si l'hypothÃ¨se d'indÃ©pendance est plausible ou doit Ãªtre rejetÃ©e.Â

Â

Il faut noter que l'utilisation du Ï‡2 est dÃ©conseillÃ©e lorsque l'effectif thÃ©orique de certaines cases est petit (plus petit que 5). Il faut donc disposer d'autres indices.

2.2. L'indice Phi (Î¦)

C'est un indice d'association que l'on rencontre frÃ©quemment dans la littÃ©rature, dÃ¨s lors que les deux variables sont dichotomiques (elles n'ont chacune que deux modalitÃ©s). Sa formule est la suivante et se calcule Ã partir du Ï‡2 :

{moslatex}\Large{$\Phi=\sqrt{\frac{\chi^2}{N}}$}{/moslatex}

Notons qu'une autre faÃ§on d'obtenir le coefficient Î¦ consiste Ã recoder chaque variable en 0 et 1,Â puis Ã calculerÂ un simple coefficient de corrÃ©lation de Pearson sur ces variables recodÃ©es, ce qui montre bien la similaritÃ© de nature entre le coefficient Î¦ et une corrÃ©lation.

2.2. L'indice V de CramÃ©r

Le V de CramÃ©r est une gÃ©nÃ©ralisation du coefficient Î¦ qui permet de traiter les cas oÃ¹ il y a plus de deux modalitÃ©s. On commence alors par calculer la valeurÂ

{moslatex}\large{$k=min(L,C)$}{/moslatex}

Â

Autrement dit on prend pour k la plus petite des deux valeurs L et C, c'est-Ã -dire, selon la convention vue plus haut, le nombre de lignes L ou le nombre de colonnes C. L'indice V est alors donnÃ© par

{moslatex}$V=$\Large{\sqrt{\frac{\chi^2}{N(k-1)}}$}{/moslatex}

Â

Â