Le rapport de corrélation : mesurer la liaison entre une variable qualitative et une variable quantitative.

Frédéric Santos
CNRS, UMR 5199 PACEA
Courriel : frederic.santos@u-bordeaux.fr

11 mars 2015

Résumé

Le rapport de corrélation est une mesure de la force de la liaison existant entre une variable catégorielle et une variable numérique continue. Son usage et son interprétation sont similaires à ceux du coefficient de corrélation linéaire de Pearson. Un test de nullité du rapport de corrélation permet de détecter les liaisons « significatives ».

1. Variances inter-groupes et intra-groupes

§1.1. Rappels sur la notion de variance. — On rappelle que la variance empirique s^2 d'une série de n valeurs $x_1, \cdots, x_n$ est définie comme *la moyenne des carrés des écarts à la moyenne*, c'est-à-dire :

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} \tag{1}$$

où $\bar{x}$ est la moyenne empirique de la série statistique.

En effet, chaque terme $(x_i - \bar{x})^2$ représente le *carré de l'écart* entre la i-ème valeur de la série statistique et la moyenne ; et la variance s^2 donne donc une idée de la *dispersion* de la série, en ce sens qu'elle représente l'écart (au carré) *moyen* entre les valeurs de la série et sa moyenne.

Dans la suite de ce document, s^2 sera appelée *variance totale* de la série statistique.

§1.2. Définitions. — On suppose à présent que l'on dispose de n individus répartis en p groupes. L'effectif de chaque groupe est noté n_k (avec $k \in [\![1, p]\!]$), de telle sorte que $n_1 + \cdots + n_p = n$.

Une même variable X a été mesurée sur chaque individu, et on note x_{ij} la valeur obtenue sur le i-ème individu du j-ème groupe.

On note toujours $\overline{x}$ la moyenne totale des n individus étudiés, sans distinction de groupes :

$$\overline{x} = \frac{1}{n}\left(\sum_{j=1}^{p}\sum_{i=1}^{n_j} x_{ij}\right) \tag{2}$$

On note $\bar{x_j}$ la moyenne des valeurs appartenant au j-ème groupe :

$$\bar{x_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} = \frac{x_{1j} + \cdots + x_{n_j j}}{n_j} \tag{3}$$

et s_j^2 la variance des valeurs appartenant au j-ème groupe :

$$s_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (x_{ij} - \bar{x_j})^2 \tag{4}$$

Définition 1 (Variance intra-groupes). — On appelle alors, pour l'ensemble des individus étudiés, *variance intra-groupes*, la moyenne pondérée des p variances $(s_j^2)_{j=1}^p$ mesurées à l'intérieur de chaque groupe :

$$B = \text{VarIntra} = \frac{1}{n} \sum_{j=1}^{p} n_j s_j^2 = \frac{n_1 s_1^2 + \cdots + n_p s_p^2}{n_1 + \cdots + n_p} \tag{5}$$

La variance intra-groupes donne donc une idée de la variabilité *à l'intérieur de chaque groupe* : si cette variance est plutôt faible alors chacun des groupes est constitué d'individus relativement homogènes ; si elle est plutôt élevée alors les groupes rassemblent en leur sein des individus assez peu semblables.

Définition 2 (Variance inter-groupes). — On appelle, pour l'ensemble des individus étudiés, *variance inter-groupes*, la variance (pondérée) de la série statistique $(\bar{x_j})_{j=1}^p$:

$$W = \text{VarInter} = \frac{1}{n} \sum_{j=1}^{p} n_j (\bar{x_j} - \overline{x})^2 = \frac{n_1 (\bar{x_1} - \overline{x})^2 + \cdots + n_p (\bar{x_p} - \overline{x})^2}{n_1 + \cdots + n_p} \tag{6}$$

La variance inter-groupes donne quant à elle une idée de la variabilité *entre les différents ensembles (groupes)* : la différenciation entre les groupes est d'autant plus prononcée que cette variance est élevée.

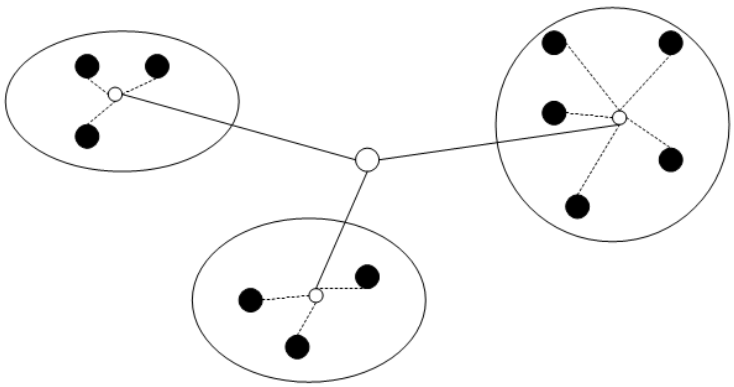

Figure 1. — Illustration graphique des variances intra et inter-groupes

§1.3. Le théorème de Huygens. —

Théorème 1 (Décomposition de la variance). — La variance totale (*i.e.* calculée sur l'ensemble de tous les individus sans distinction de groupes) est égale à la somme de la variance inter-groupes et de la variance intra-groupes. ♢

Démonstration. — Simples transformations d'écriture :

$$\begin{aligned}
B+W &= \text{VarIntra} + \text{VarInter} \\
&= \left(\frac{1}{n}\sum_{j=1}^{p} n_j s_j^2\right) + \left(\frac{1}{n}\sum_{j=1}^{p} n_j(\bar{x_j}-\overline{x})^2\right) \\
&= \left(\frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij}-\bar{x_j})^2\right) + \left(\frac{1}{n}\sum_{j=1}^{p} n_j(\bar{x_j}-\overline{x})^2\right) \\
&= \left(\frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{n_j} x_{ij}^2 - 2x_{ij}\bar{x_j} + \bar{x_j}^2\right) + \left(\frac{1}{n}\sum_{j=1}^{p} n_j\bar{x_j}^2 - 2n_j\bar{x_j}\overline{x} + n_j\overline{x}^2\right) \\
&= \sum_{j=1}^{p}\left(\sum_{i=1}^{n_j}\frac{x_{ij}^2}{n} - \frac{2\bar{x_j}}{n}\sum_{i=1}^{n_j} x_{ij} + \frac{n_j}{n}\bar{x_j}^2\right) + \frac{1}{n}\sum_{j=1}^{p} n_j\bar{x_j}^2 - \overline{x}^2 \\
&= \sum_{j=1}^{p}\left(\sum_{i=1}^{n_j}\frac{x_{ij}^2}{n}\right) - \frac{1}{n}\sum_{j=1}^{p} n_j\bar{x_j}^2 + \frac{1}{n}\sum_{j=1}^{p} n_j\bar{x_j}^2 - \overline{x}^2 \\
&= \sum_{j=1}^{p}\sum_{i=1}^{n_j}\frac{x_{ij}^2}{n} - \overline{x}^2 \\
&= \text{VarTotale}
\end{aligned}$$

Ce qui démontre finalement la formule de Huygens de décomposition de la variance. ♢

2. Rapport de corrélation empirique

§2.1. Idée générale. — Le rapport de corrélation empirique offre une mesure simple de la liaison entre une variable qualitative et une variable quantitative. Considérons que la variable qualitative possède p modalités. On obtient alors une partition naturelle de notre échantillon de données en p groupes : chaque individu appartient au groupe naturellement défini par la modalité de la variable qualitative qu'il présente.

Par exemple, considérons le jeu de données suivant :

Étudiant	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Note	15	14	12	8	16	13	10	17	4	8	11	6	8	7	12
Assiduité	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C

Table 1. — Notes obtenues par 15 étudiants à un examen terminal en fonction de leur assiduité en cours (A = parfaite assiduité ; B = assiduité correcte ; C = nombreuses absences).

Ici, la variable quantitative est la note obtenue, et la variable qualitative (à 3 modalités) est l'assiduité de l'étudiant. Le niveau d'assiduité définit naturellement une partition des étudiants en trois « groupes » distincts. Le rapport de corrélation permet d'obtenir une mesure descriptive simple de la force de liaison existant entre la présence en cours et la note obtenue à la fin du semestre.

§2.2. Aspects calculatoires. — Le théorème de Huygens est essentiel puisqu'il permet de comprendre que la variabilité totale dans notre échantillon est la somme de la contribution des variations à l'intérieur des groupes et entre les groupes. En d'autres termes, ici, la variabilité entre les notes obtenues dans toute la promotion dépend de deux facteurs : le fait que les étudiants assistent ou pas aux cours, et le fait qu'à assiduité égale (*i.e.* à l'intérieur d'un même « groupe d'assiduité ») les étudiants n'ont pas le même niveau.

La rapport de corrélation permet de savoir lequel de ces deux facteurs est prédominant pour expliquer la variabilité des notes dans toute la promo.

Le poids relatif des variances intra et inter est déterminant pour comprendre la structure et la pertinence d'un découpage en groupes : si la variance intra-groupes est nettement plus élevée que la variance inter-groupes, on est dans un cas où les groupes sont en moyenne assez semblables entre eux, mais où chacun d'eux abrite en son sein une énorme variabilité inter-individuelle. Les groupes sont donc vraisemblablement mal définis, et ne correspondent pas à une réalité (physique, biologique, sociale, ...) bien définie.

À l'inverse, si la variabilité inter-groupes est nettement plus élevée que la variabilité intra, nous sommes alors en présence de groupes bien différenciés les uns des autres et bien homogènes en leur sein : le découpage en groupes est pertinent et correspond à une réalité concrète.

L'idée du rapport de corrélation est tout simplement de mesurer le poids de la contribution de la variance inter-groupes dans la variance totale (ce qui revient donc à mesurer le poids relatif de la variance inter et de la variance intra) :

Définition 3 (Rapport de corrélation empirique). — Soient X une variable quantitative et Y une variable qualitative à p modalités. Ces deux variables sont mesurées sur n individus, et on suppose que chacune des p modalités de Y est présente sur au moins deux individus. Les individus sont alors naturellement répartis en p groupes correspondant aux p modalités de Y. Le rapport de corrélation entre X et Y, noté $\widehat{\eta}^2$, est le rapport de la variance inter sur la variance totale [Sap06] :

$$\widehat{\eta}^2 = \frac{\text{VarInter}}{\text{VarTotale}} = \frac{\text{VarInter}}{\text{VarInter} + \text{VarIntra}} \tag{7}$$

Proposition 2. — Le rapport de corrélation est toujours compris entre 0 et 1. Une valeur de 0 signifie qu'il n'y a aucun lien entre les deux variables. Plus la valeur est proche de 1 et plus les variables sont liées.

Démonstration. — Évident en regardant l'équation 7 de la définition : $\widehat{\eta}^2$ est le rapport de deux quantités positives et est donc toujours supérieur ou égal à 0. Il est également inférieur ou égal à 1 puisque son numérateur (la variance inter-groupes) est toujours inférieur ou égal à son numérateur (la variance totale, qui est la somme de la variance inter et de la variance intra). Enfin, plus la valeur est proche de 1 et plus la variance inter-groupes a un poids important par rapport à la variance intra-groupes, et donc plus Y est en lien avec X. ♢

Proposition 3 (Test de significativité). — On admettra que sous l'hypothèse $\mathcal{H}_0$ de nullité du rapport de corrélation, la quantité $K = \big(\widehat{\eta}^2(n-p)\big) \,/\, \big((p-1)(1-\widehat{\eta}^2)\big)$ suit une loi de Fisher $F(p-1; n-p)$ [Cha04]. Pour déterminer si la valeur $\widehat{\eta^2}$ est significativement différente de 0 (avec un risque d'erreur de 5%), il suffit donc de comparer la statistique de test K au quantile d'ordre 0.95 de la loi de Fisher à $p-1$ et $n-p$ degrés de liberté.

3. Un exemple avec le logiciel R

On reprend les données de la table 1, que l'on peut saisir dans R par les commandes suivantes (à copier-coller dans une console R pour l'exemple) :

```
note = c(15, 14, 12, 8, 16, 13, 10, 17, 4, 8, 11, 6, 8, 7, 12)
assi = factor(c(rep("A",5), rep("B", 5), rep("C", 5)))
dat = data.frame(note, assi)
dat
```

On peut définir de la façon suivante une fonction calculant la variance inter-groupes d'une variable `X` où les groupes sont donnés par une variable qualitative `Y` :

```
VarInter <- function(X, Y) {
 moyennes = tapply(X, INDEX=Y, FUN=mean)
 effectifs = tapply(X, INDEX=Y, FUN=length)
 res = (sum(effectifs * (moyennes - mean(X))^2))/length(X)
 return(res)
}
```

et enfin une fonction calculant la variance populationnelle (i.e. à dénominateur égal à n et non $n-1$) :

```
VarTot <- function(X) {
 return(sum((X - mean(X))^2)/length(X))
}
```

Le rapport de corrélation se calcule alors comme suit, suivant la formule donnée en équation 7 :

```
VarInter(note, assi) / VarTot(note)
```

On obtient alors une valeur du rapport de corrélation $\widehat{\eta}^2 \approx 0.215$, ce qui correspond à une liaison faible entre l'assiduité aux cours et les notes à l'examen terminal. En d'autres termes, la variabilité observée sur les étudiants de toute la promo ne s'expliquent que par des différences de niveau « intrinsèques » entre étudiants, l'assiduité aux cours ne semblant pas être le facteur déterminant ici.

Références

[Cha04] S. Champély : *Statistique vraiment appliquée au sport : cours et exercices.* De Boeck, 2004.

[Sap06] G. Saporta : *Probabilité, statistique et analyse de données.* Technip, 2e édition, 2006.