Analyses bivariées

Michel Calciu

Cours à l'Université de Lille 1 - 2003/2004

Introduction

Objet

Les traitements bivariés ont pour objet de mettre en évidence les relations éventuelles qui existent entre deux variables analysées simultanément.

Dans la plupart des cas l'analyste cherchera à expliquer une des deux variables - dite variable à expliquer (Y) - à l'aide de l'autre variable - dite variable explicative (X). Expliquer une variable à l'aide d'une autre revient à repérer dans quelle mesure les différentes valeurs que peut prendre la variable explicative ont une conséquence sur les valeurs prises par la variable à expliquer.

Exemple:

Exemple:

Le fait de changer de conditionnement a-t-il un effet sur le niveau des ventes d'un produit donné ? Le conditionnement joue ici le rôle de variable explicative et le niveau des ventes de variable à expliquer
Le fait de posséder un four à micro-ondes dépend-il de l'âge ou de la taille. La possession du produit est la variable à expliquer; l'âge ou la taille de la famille sont des variables explicatives.

Comme dans le cas des traitements univariés, le mode d'analyse utilisable va dépendre de la nature des variables étudiées: quantitatives, ordinales ou nominales. Ces analyses seront à nouveau présentées ici dans le cadre des études marketing par questionnaires: il s'agira donc du traitement des tri-croisés.

Traitement des tris croises et nature des donnés

- **Présentation des tris croisés**

Dans la mesure où une question peut relever fondamentalement de trois niveaux de mesure différents, on comptera neuf types de croisements possibles entre les questions Qi et Qj Les plus fréquentes sont présentés ci-dessous.

2 - **Qi quantitatif x Qj quantitatif:** étude des relations entre deux séries de n chiffres s'il y a n questionnaires. Ces deux séries de chiffres n'apparaissent généralement pas explicitement. Leurs relations sont matérialisées par différents indicateurs.

Exemple:

Dans le questionnaire CAMIP, étude des relations entre le proportion des achats par catalogue (question 1) et le revenu de la personne (question 11). L'appartenance à une catégorie de revenu plus élevé entraîne-t-elle une plus grande proportion des achats par catalogue ?

2 - **Qj nominal x Qj nominal** : croisement le plus fréquent qui se traduit par la création d'un **tableau de contingence** où, en ligne, figurant les modalités de la variable à expliquer et en colonne, celles de la variable explicative. Lorsque le tableau de contingence est traduit en pourcentages de colonne, pour chaque modalité de la variable explicative apparaîtra une distribution de fréquences des modalités de la variable à expliquer.

Exemple:

Croisement entre la question 22 sur la préférence pour un type de magasin ou une modalité d'achat et la question 26: le fait de préférer d'acheter par catalogue ou par une autre modalité, dépende-t-elle de la situation civile du répondant?

3 - **Qi ordinal X Qj ordinal** : mise en correspondance de deux classements au niveau de chaque individu interrogé ou sur l'ensemble de l'échantillon si une procédure d'agrégation des rangs a été utilisé.

Exemple:

Croisement entre l'ordre de préférence pour les différentes catégories d'équipement de bureau (micro-ordinateurs, imprimantes, scanners..) exprimé par chaque répondant et un classement a priori *correspondant à l'importance accordé à chaque catégorie d'équipement dans le catalogue (exprimé par exemple en nombre de pages).*

4 - Qi **quantitatif x Qj nominal** : correspond à un tri-à-plat de la variable quantitative pour chacune des modalités de la variable nominale qui joue le rôle de variable explicative.

Exemple:

Croisement entre une question ouverte concernant le nombre d'objets achetés par catalogue et la question 26. La situation civile influence-t-elle le nombre d'objets achetés par catalogue?

5 - **Qi ordinal x Qj nominal** : repérage des rangs donnés à la question Qi pour différentes classes d'une variable Qj nominale explicative.

Exemple:

Croisement entre un classement de préférence et la question 26. Le fait de relever d'un statut familial donné entraîne-t-il des préférences pour une catégorie de produit bureautiques ?

- **Analyse des tris croisés**

Le tableau 2.1. donne les principaux tests utilisables dans l'analyse des tris croisés. Seuls les tests correspondant aux croisements les plus fréquents y sont indiqués.

Tableau 2.1 .: Tests des tris croisés selon la nature des variables

En colonne figurant les variables à expliquer et en ligne les variables explicatives.

• Traitement des variables quantitatives

Discussion

Le croisement de deux variables quantitatives peut être effectué dans des circonstances très variées :

- comparaison des résultats obtenus pour une variable sur deux ou plusieurs populations indépendantes;

- comparaison des résultats obtenus pour une variable sur deux échantillons appariés;

- comparaison des résultats obtenus pour deux variables quantitatives différentes pour la même population.

- **Une variable et deux populations indépendantes a) Le test de comparaison de moyennes**

En ce qui concerne le cas où une seule variable est étudiée, le **test de comparaison de moyennes** est la statistique classique lorsque deux populations sont concernées.

L'analyste dispose des données suivantes :

- deux populations A et B respectivement d'effectifs N_a et N_b;
- la moyenne de la variable étudiée est a dans la population A et b dans la population B.
- la variance de la variable analysée est s_a^2 pour A et s_b^2 pour B.

Dans la mesure où l'on estime que $\overline{X}_a$ et $\overline{X}_b$ suivent une loi normale, respectivement de moyenne μ_a et μ_b et de variance s_a^2 et s_b^2, on montre que la différence

$D = \overline{X}_a - \overline{X}_b$

suit également une loi normale de moyenne $\mu_a - \mu_b$ et de variance:

$\sigma_D^2 = [\sigma_a^2 /Na + \sigma_b^2 /Nb] \approx [s_a^2 /Na + s_b^2 /Nb]$

L'intervalle de confiance de la différence de moyenne au risque a est donné par :

$$\mu_a-\mu_b = \overline{X}_a - \overline{X}_b \pm z_{\alpha/2}[s_a^2/Na+s_b^2/Nb]^{1/2} \qquad (18)$$

L'hypothèse nulle Ho correspond au cas où la différence D = μa - μb de moyennes est nulle. Sous Ho, la variable réduite devient :

$$Z\ (a-b)/[s_a^2/Na+s_b^2/Nb]^{1/2} \qquad (19)$$

La valeur z ainsi calculée doit être comparée avec la valeur lue dans la table normale réduite pour le seuil de confiance désiré et compte tenu du caractère unilatéral ou bilatéral du t(est. Pour un test bilatéral par exemple, Ho sera rejeté au seuil de risque de 5 % si **|z| > 1,96.**

Exemple:

Supposons que dans le cadre de l'étude CAMIP, on a ajouté une question sur la nombre de disquettes d'ordinateur achetées par an qu'on croise avec la question 27(le fait d'être homme ou femme) a fait apparaître les résultats suivants:

les hommes (A): Na = 155; nombre moyen de disquettes = 10, avec $Sa^2 = 64$;

les femmes (B): Nb = 75; Xb = 3, Sb = 25.

La variance des différences de moyennes est donnée par:

64/155 + 25/75 = 0,74.

L'écart type de D est alors = 0,86. L'hypothèse nulle pour laquelle il n'existe pas de différence dans les quantités de disquettes achetées par les homme et par les femmes peut être rejetée, puisque z = (10 - 3)/0,86 est supérieur à 1 96

- **b) Autres tests** **1) test de Student**

- Pour des petits échantillons (Na et Nb < 30), on utilisera le test **de Student.** Dans la mesure où la variance des Xa et des Xb est estimée la variance de la distribution des différences de moyennes est approchée par l'expression :

$$\sigma_D^2 = [(N_a-1)s_a^2 + (N_b-1)\, s_b^2]$$

$$/(N_a + N_b - 2)(1/N_a + 1/N_b) \quad (20)$$

La différence D suit alors une loi de Student à $(N_a + N_b - 2)$ degrés de liberté.

- **2) (test en F et Kruskal - Wallis)**

- Quand plus de deux populations sont concernés, on aura recours au **test** F. Si l'hypothèse de normalité évoquée plus haut n'est pas respecté on pourra employer le **test de Kruskal-Wallis.**

- **Une variable et deux échantillons appariées**

Dans le cas d'échantillons appariés, à chaque individu d'un premier groupe est associé un individu du second groupe (le groupe-témoin) offrant les mêmes caractéristiques.

Pour chaque couple i de deux individus appariés, une différence $D_i = X_{ai} - X_{bi}$ est calculée. Sur l'ensemble n des couples étudiés, la différence moyenne est donnée par $\overline{D} = \Sigma d_i/n$ et la variance des différences est alors $s_D^2 = \Sigma(Di - \overline{D})^2/(n - 1)$

On montre que $\overline{D}$ est distribué selon une loi normale de moyenne µa - µb et de variance

$$s^2 = [\Sigma(Di - \overline{D})^2/(n - 1)]/n. \quad (21)$$

- **Deux variables et la même population**

Les relations entre deux variables quantitatives sur la même population sont généralement analysées à l'aide du **coefficient de corrélation** de Pearson qui sera étudié au cours du chapitre portant sur la régression linéaire

• Traitement des variables nominales - Analyse des tableaux de contingence

- **a) Test du caractère significatif de la relation entre les variables**

Dans le chapitre précédent on a vu une application du test du Khi-Deux pour l'évaluation de la qualité de l'ajustement d'une distribution de fréquences observées à une distribution théorique.

De façon plus générale, ce test est employé pour analyser les tableaux de contingence et repérer le caractère statistiquement significatif de l'association entre deux variables nominales.

La statistique du χ^2 est donnée, en ce qui concerne les tableaux de contingence, par la formule suivante:

$$\chi^2 = \Sigma\Sigma(N_{ij} - \Theta_{ij})^2 / \Theta_{ij} \quad (22)$$

où $\mathbf{N_{ij}}$ = nombre d'observations dans la case ij; $\mathbf{\Theta ij}$ = nombre théorique associé à la case ij = (total de la ligne i)·(Total de la colonne j)/Nombre total d'observations.

Pour un tableau comportant **C** colonnes et **L** lignes, la valeur ainsi calculée est comparée à la valeur critique $\boldsymbol{\chi^2}$ lue sur la table du Khi-Deux pour un seuil de confiance 1-α, et pour un nombre de degrés de liberté égal à (C - 1)(L - 1) .

La table du Khi-Deux donne la distribution de probabilité des valeurs de χ^2 obtenues dans un tableau lorsque l'hypothèse nulle est vraie, c'est-à-dire dans le cas d'indépendance entre les deux variables étudiées. Par exemple, au seuil de 5 % et pour deux degrés de liberté, le χ^2 lu sur la table vaut 5,99 : ceci veut dire que sous Ho il n'y a que 5 % de tableaux à deux degrés de liberté pour lesquels on pourrait calculer un x2 supérieur ou égal à 5,99. Si le χ^2 calculé est plus fort, il y a donc moins de cinq chances sur cent de se tromper en rejetant Ho.

Exemple

Exemple

Le tableau 2.2.a donné les résultats d'un éventuel résultat du croisement des questions 3 (le fait d'avoir commandé) et 27 (être homme où femme). Le tableau 2.2. fournit les valeurs théoriques ": par exemple, pour la case 1, la valeur 53,91 = (80)(155)/230. Les différences (Nij-Θij;) apparaissent sur le tableau 2.2.c, et le χ^2 associé à chaque case, sur le tableau 2.2.d. Au total, le χ^2 calculé s'élève à 28,64: il dépasse le %2 critique (5,99). La qualité d'abonné est donc liée significativement au motif du voyage**.**

Tableau 2.2.: Application du Khi-Deux à un tableau de contingence

a) Croisement question Q3 et Q27: valeurs observées

	Homme	**Femme**	**Total**
Jamais	**70**	**10**	**80**
l'année dernière	**35**	**15**	**50**
cette année	**50**	**50**	**100**
Total	**155**	**75**	**230**

b) Valeurs théoriques

Jamais	**53,91**	**26,09**	**80**
l'année dernière	**33,70**	**16,30**	**50**
cette année	**67,39**	**32,61**	**100**
Total	**155,00**	**75,00**	**230**

c) Différences entre valeurs observées et valeurs théoriques

Jamais	**16,09**	**-16,09**	**0,00**
l'année dernière	**1,30**	**-1,3**	**0,00**
cette année	**-1 7,39**	**1 7,39**	**0,00**
Total	**0,00**	**0,00**	**0 00**

d) Croisement questions Q3 et Q27: calcul du Khi-Deux

	Homme	**Femme**	**Total**
Jamais	**4,80**	**9,92**	**14,72**
l'année dernière	**0,05**	**010**	**015**
cette année	**4,49**	**9,28**	**13,76**
Total	**9,34**	**19,30**	**28,64**

Khi-Deux calculé: 28,64 nombre ddl: 2

Khi-Deux critique: 5,99 risque: 5 %

Le test est fourni couramment par les logiciels; il souffre cependant de certaines limitations:

- l'effectif sur lequel porte le tableau doit être suffisamment important on ne doit pas trouver plus de 20 % de cases avec un effectif inférieur à 5 ;

- le χ^2 est calculé à partir des valeurs absolues; il est donc très sensible à la taille des effectifs considérés;

le χ^2 permet de repérer le caractère significatif de la relation entre les deux variables, mais pas l'intensité de cette relation.

- **b) Test du degré d'association entre les variables**

Plusieurs coefficients peuvent repérer le degré d'association entre les deux variables étudiées. Pour les tableaux de contingence de taille 2 x 2, on peut déduire du χ^2, un coefficient **d'association Φ**, tel que :

$$\Phi = [\chi^2/n]^{1/2} \qquad (23)$$

Φ présente l'avantage d'être indépendant de la taille de l'échantillon, et de varier entre 0 et 1 .

Pour les tableaux plus grands, c'est un coefficient de contingence C **qui sera utilisé** :

$$C = [\chi^2/(\chi^2 + n)]^{1/2} \qquad (24)$$

Plus C est élevé et plus forte est l'association entre les deux variables concernées. Le minimum de ce coefficient est 0 (indépendance totale des variables avec $\chi^2 = 0$). Par contre, le maximum ne peut jamais s'élever jusqu'à 1. Dans le cas d'un tableau 2 x 2, on montre, par exemple, que le maximum est de = 0,707.

- **Test de comparaison de fréquences**

Lorsque l'on considère deux modalités de la variable explicative, les effectifs associés à une modalité donnée de la variable à expliquer peuvent être traduits en fréquence relative. Il s'agit de vérifier dans quelle mesure la différence de fréquences observées est significative.

Soient Pa la fréquence relative associée à la modalité A (effectif Na) de la variable explicative et Pb celle qui est associée à la modalité B (effectif Nb)

La variance des différences de proportions est donnée par $\boldsymbol{\sigma_D^2} = [\pi_a(1-\pi_a)/Na + \pi_b(1-\pi_b)/Nb]$ où π_a etπ_b sont les proportions réelles. Dans la mesure où les véritables fréquences π_a et π_b ne sont pas connues, on utilisera comme estimateur de $\boldsymbol{\sigma_D^2}$ l'expression Pc(l- Pc)[1/Na + 1/Nb] où Pc est la fréquence moyenne observée sur l'ensemble des deux groupes avec Pc = (Na Pa + Nb Pb)/(Na + Nb).

Les intervalles de confiance et le test des différences de proportions s'obtiennent dans les mêmes conditions qu'en ce qui concerne les moyennes.

Exemple:

Exemples:

Avec le croisement des questions 3 et 27 du questionnaire CAMIP, il apparaît que sur 155 hommes, 70 n'ont jamais commandé, soit 45,16 % et seulement 10 sur 75 femmes , soit 13,33 % La fréquence moyenne pondérée est donc p = [155(0,4516) + 75(0,1333)]/230= 0,3478. La variance des différences de proportions est alors $\sigma_D^2 = [0,3478(1 - 0,3478)]\ [1/155 + 1/75] =$ 0,00449 et σ_D = 0,067
Dans ces conditions l'intervalle de confiance des différences de proportions au seuil de risque de 5 % est donné par:
$\pi_a - \pi_b = 0\ 4516 - 0,1333 \pm 1\ 96\ (0,067) = [0,4496\ \ 0,187]$
La valeur 0 ne figurant pas dans cet intervalle de confiance, la différence de proportions apparaît significative.

• Traitement des donnes ordinales•> Relations entre deux variables ordinales

- a) Test de corrélation des rangs de Spearman

Ce test permet de repérer le caractère significatif de la relation qui existe entre deux classements. Il est également utilisé pour montrer la relation éventuelle qui existe entre deux variables quantitatives.

Soit n le nombre d'items à classer; Xi est le rang de l'item i dans un premier classement et Yi son rang dans un second. Dj = |Xi - Yi| est la différence de rangs observés entre les deux classements. Le coefficient de corrélation des rangs de Spearman a pour expression :

$$R_s = 1 - 6\ \Sigma\ D_i^2/[n(n^2 - 1)] \qquad (25)$$

Plus R_s est proche de 1 et plus les deux classements sont proches; à la limite, ils sont complètement identiques si $R_s = 1$. Au contraire, plus R_s est proche de 0 et plus les deux classements sont indépendants.

La signification statistique de Rs obtenue peut être testée à partir de la relation:

$$t = R_s[n - 2]^{1/2}/[1 - R_s^2]^{1/2} \qquad (26)$$

qui suit une loi de Student à n - *2* degrés de liberté

Exemple:

Exemple:

Croisement entre l'ordre de préférence pour les différentes catégories d'équipement de bureau (micro-ordinateurs, imprimantes, scanners..) exprimé par chaque répondant et un classement a priori correspondant à l'importance accordé à chaque catégorie d'équipement dans le catalogue (exprimé par exemple en nombre de pages).Le tableau 2.3. montre un exemple de calcul du coefficient de Spearman avec le traitement d'une question concernant le classement d'intérêt d'achat pour les catégories de produits (A - micro-ordinateurs, B - imprimantes, C - scanners et D fax) par l'individu 124. On constate une corrélation R. = 1 - (6)(38)/(5)(25-1)= - 0,9 entre les choix de la direction et ceux de l'individu n° 128. Cette corrélation est significative puisque t = (-0,9)( )/ (= - 3,576. Cette valeur est supérieure au t de Student pour 3 degrés de liberté et un seuil de confiance de 95 % (3,182).
L'application de cette procédure à l'ensemble de l'échantillon permettrait de repérer quels sont les individus qui manifestent des préférences conformes à l'importance donnée dans le catalogue à ces catégories .

Tableau 2.3.: Application du coefficient de Spearman

Propositions	**Classe-**	**Classement**	**|Dj|**	**|Dj|²**

	ment individu n° 124	CAMIP		
A	2	5	3	9
B	1	**4**	3	9
C	3	3	0	0
D	5	1	4	16
E	4	2	2	4
				$\Sigma\|D_i\|^2$ =38

R_s= -0,9

t = -3,576

- b) Tau de Kendall

Le tau de Kendall est un autre indicateur du caractère éventuellement significatif de la relation qui existe entre deux classements.

Les n items sont tout d'abord rangés dans l'ordre conforme au premier classement. Puis pour tout couple d'items dont l'ordre dans le premier classement est concordant avec celui du second on attribue un score de + 1; dans le cas de discordance on attribue un score de - 1. S est la somme de ces scores sur l'ensemble des n(n - 1)/2 comparaisons possibles.

Le tau de Kendall est alors :

$\tau = 2\ S/(n^2 - n)$ (27)

Le test de ce coefficient s'effectue à l'aide de tables. Pour n > 10 une approximation par la loi normale est envisageable .

Exemple:

Exemple:

Le tableau 2.4. donne une application du τ de Kendall pour l'analyse des résultats de la question précédente pour l'étude CAMIP. On constate un t de - 0,8 entre les choix de la direction et ceux de l'individu n° 124. En effet sur un ensemble de 10 comparaisons possibles, une seule est concordante et neuf sont discordantes.

Tableau 2.4.: Application du tau de Kendall

Propositions	Classement Individu n° 128	Classement Catalogue CAMIP
B	1	4
A	2	5
C	3	3
E	4	2
D	5	**1**

	Classe-ments concor-dants	Classements discordants	Score
Paires B	BA	BC, BE, BD	-2
Paires A		AC, AD, AE	-3
Paires C		CD, CE	-2
Paire E		ED	-1
			Total S = - 8

Nb. combinaisons = n(n - 1)/2= 10 Tau de Kendall = - 0.8