Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

Cours 12 – Analyse de la variance

1 Dans ce cours nous étudions comment l'analyse de la variance de Y permet de tester l'égalité des moyennes conditionnelles de cette variable numérique dans les sous-populations induites par X ; dans cette problématique, X est appelée la `variable explicative`, ou le facteur explicatif, et Y la `variable expliquée`.

2 Dans la formule de décomposition de la variance, `variance totale = variance intra + variance inter`, la variance intra, moyenne des variances conditionnelles, quantifie la part de la variabilité intrinsèque de Y, et la variance inter, variance des moyennes conditionnelles, quantifie l'hétérogénéité des sous-populations.

Variance expliquée

3 déf **La variance expliquée par la variable** X est égale à la variance inter divisée par la variance globale de Y ; c'est un nombre compris entre 0 et 1 puisque les variances sont des nombres positifs ou nuls, et que la variance inter est une part de la variance globale.
La variance expliquée est une mesure du lien entre le facteur X et la mesure numérique Y, pour apprécier comment Y dépend du fait d'appartenir à une sous-population ou à une autre.

Exemple du nombre d'enfants de 0 à 6 ans selon la structure familiale (exemple 1 du cours 9) :

Arrondi 4

	Effectif	$\sum y_i$	$\sum y_i^2$	$\overline{y_i}$	σ_i^2	$n_{i.}\overline{y_i}^2$	$n_{i.}\sigma_i^2$	
Mono	*1035*	*1669*	*3199*	1,6126	0,4903	2691,4955	507,4605	
Couples	*6536*	*11036*	*22802*	1,6885	0,6376	18634,3468	4167,3536	
Total	**7571**	12705	26001			21325,8423	4674,8141	
Moyenne		**1,6781**	3,4343			2,8168	0,6175	
		Variance	**0,6183**		**Inter**	**0,0008**	**0,6175**	**Intra**

Les moyennes dans les deux sous-populations diffèrent assez peu (de l'ordre de 5%), la variance inter est très faible ($\approx 0{,}001$) de même que la variance expliquée par le facteur "structure familiale" ($\approx 1/1000$).

4 **Si la variance expliquée est égale à 1,** la variance intra vaut 0, ce qui entraîne que toutes les variances conditionnelles sont nulles (la variance intra étant une somme de nombres positifs ou nuls, elle ne peut valoir 0 que si chaque terme est nul). Par conséquent, les individus de chaque sous-population ont tous la même mesure Y. Dans ces conditions, le facteur X détermine entièrement la mesure Y : il suffit de connaître la sous-population dans laquelle l'individu se trouve pour connaître sa mesure Y ; et inversement il suffit de connaître la mesure Y d'un individu pour savoir dans quelle sous-population il se trouve, en supposant que les sous-populations ont toutes des valeurs de Y différentes.

5 **Si la variance expliquée est égale à 0,** la variance inter est nulle, ce qui revient à dire que les moyennes conditionnelles de Y sont identiques : Y donne globalement les mêmes mesures sur toutes les sous-populations. Cela ne signifie pas à proprement parler que X et Y sont indépendantes (il faudrait pour cela qu'en plus de l'égalité des moyennes conditionnelles on ait l'égalité des distributions conditionnelles), mais que le facteur X n'a pas d'*effet global* sur la mesure de Y.

Inversement, si X et Y sont indépendantes, les moyennes conditionnelles de Y sont identiques, la variance inter est donc nulle et la variance expliquée est égale à 0.

6 **Si la variance expliquée est proche de 0,** la variance inter l'est également. Les moyennes de Y dans les sous-populations sont peu différentes les une des autres ; alors de deux choses l'une :

- ou bien ces différences sont un effet des fluctuations d'échantillonnage pour des sous-populations de même moyenne, ce qui revient à considérer que le facteur X n'a pas d'effet sur la variable Y.
- ou bien au contraire, elles sont dues au fait que X a un effet sur la Y, sans doute faible, mais suffisant pour rendre les mesures de Y différentes dans les sous-populations.

Ces observations ont donné l'idée d'un test pour apprécier si des petites différences entre les moyennes de Y dans les sous-populations peuvent être interprétées ou non comme l'effet du facteur X.

Test de l'égalité des moyennes par l'analyse de la variance

7 Soit T la valeur $(n-k)*\frac{\text{var inter}}{\text{var intra}}$ calculée à partir de l'observation d'un échantillon de taille n, k étant le nombre de sous-populations (le nombre de modalités de X) ; dans l'hypothèse (H) où les moyennes de Y sont identiques dans les sous-populations, les valeurs de T calculées sur différents échantillons vont varier au voisinage de 0, puisque du fait des fluctuations d'échantillonnage, les variances inter vont être proches de 0 ; plus précisément, on sait par des travaux mathématiques que dans cette hypothèse (H), les valeurs de T varient approximativement comme la distribution du χ^2 à $k-1$ degrés de liberté.

8 Le principe du test est identique à celui du χ^2 pour l'indépendance statistique :

1. on détermine d'abord un seuil s au-delà duquel les valeurs de T sont considérées comme improbables si l'hypothèse (H) était vraie, en utilisant une table des distributions du χ^2 qui est la distribution de T dans ce cas : on prendra par exemple le 95ème centile, ce qui revient à considérer que si (H) est vraie 5% seulement des échantillons peuvent donner pour T une valeur supérieure à s.
2. on calcule la valeur t de T pour l'échantillon observé.
3. si t fait partie des valeurs improbables, de deux choses l'une : ou bien (H) est fausse (les moyennes de Y sont différentes dans les sous-populations), et il n'y a rien d'étonnant à ce que t soit une "grande" valeur ; ou bien (H) est vraie et l'échantillon observé est un des rares échantillons atypiques qui donnent un t élevé sous l'hypothèse (H) ; en estimant qu'il n'y a pas de raison sérieuse pour que l'échantillon observé soit atypique, on choisira de rejeter l'hypothèse (H), en considérant que le facteur X a un effet global sur Y.
4. si t ne fait pas partie des valeurs improbables, tout se passe comme si (H) était vraie, et on ne rejettera donc pas cette hypothèse, en considérant que le facteur X n'a pas d'effet global sur Y.

Exemple du nombre d'enfants de 0 à 6 ans selon la structure familiale :

- si la moyenne du nombre d'enfants est identique dans les deux structures familiales (Hypothèse H), la variable T suit une loi du χ^2 à $2-1=1$ degré de liberté ; le 95ème centile vaut 3,84, ce qui signifie que dans cette hypothèse (H) 5% seulement des échantillons donnent une valeur de T supérieure à $s=3{,}84$.
- pour l'échantillon observé, la valeur t calculée pour T est égale à $(7571-2)*\frac{0{,}0008}{0{,}6175} \approx 9{,}8$; si (H) était vraie, l'échantillon serait un de ces 5% d'échantillons atypiques, ce que nous n'avons pas de raison de penser : nous rejetons donc cette hypothèse en admettant que le facteur "structure familiale" a un (faible) effet global sur le nombre d'enfants, puisque les moyennes sont considérées comme différentes dans les deux sous-populations.

On remarquera en passant que la faible valeur de la variance inter conduit quand même à considérer comme significative la différence entre les moyennes des deux sous-populations (elle n'est

pas l'effet des fluctuations d'échantillonnage) ; cela est dû à la grande taille de l'échantillon. En supposant celui-ci de taille 1000, t vaudrait $(1000-2) * \frac{0{,}0008}{0{,}6175} \approx 1{,}3$, ce qui nous aurait amené à ne pas rejeter l'hypothèse d'égalité des moyennes.

Programme de travail

Savoir définir :

- la variance expliquée.

Savoir faire :

- interpréter une valeur limite prise par la variance expliquée.
- comparer les moyennes conditionnelles par un test d'analyse de la variance.