

ANALYSE DE LA VARIANCE

Source : Wannacott et Wannacott, Statistique : Economie, Gestion, Sciences , Médecine, Economica, 3° éd., 1984

L'analyse de la variance à un facteur

Le test

L'analyse de la variance recouvre un ensemble de technique de tests et d'estimation destinés à apprécier l'effet de variables qualitatives sur une variable numérique. Dans le cas gaussiens, il revient à comparer plusieurs moyennes d'échantillon.

1° CAS : MÊME NOMBRE D'OBSERVATIONS PAR ÉCHANTILLON

Posons que $X_{i,t}$ est l'observation t (on suppose à ce stade qu'il y a n observations par échantillon) pour l'échantillon i (le jeu de données étant composé de c échantillons). On définit \bar{X}_i **comme la moyenne de l'échantillon i** et \hat{X} **comme la moyenne des différents échantillons**. L'hypothèse nulle est celle de l'égalité des moyennes des différents échantillons ($m_1 = m_2 = \dots m_c$). n_i est le nombre d'individus de l'échantillon i et c est le nombre d'échantillons.

La **variance entre les échantillons** est fondée sur la prise en compte des variations entre la moyenne des échantillons et la moyenne globale : $s_{\bar{X}}^2 = \frac{1}{c-1} \sum (\bar{X}_i - \hat{X})^2$. On parle de variance interclasse.

La **variance au sein des échantillons** est quant à elle fondée sur l'addition des variations au sein de chaque échantillon : $s_p^2 = \frac{1}{c(n-1)} \sum_{i=1}^c \sum_{t=1}^n (X_{i,t} - \bar{X}_i)^2$.

Le test de Fisher est alors obtenu par la comparaison de ces deux grandeurs. Le tableau d'analyse de la variance prend dès lors la forme suivante :

Source de variation	Variation (ou somme des carrés) (1)	degrés de liberté (2)	Variance (3)	Fisher
Entre échantillon	$SC_{\text{echant}} = n \sum_{i=1}^c (\bar{X}_i - \hat{X})^2$	(c-1)	$n s_{\bar{X}}^2 = (1)/(2)$	$F = \frac{ns_{\bar{X}}^2}{s_p^2}$
Au sein des échantillons (ou résiduelle)	$SC_{\text{res}} = \sum_{i=1}^c \sum_{t=1}^n (X_{i,t} - \bar{X}_i)$	c(n-1)	$s_p^2 = (1)/(2)$	
Total	$SC_{\text{tot}} = \sum_{i=1}^c \sum_{t=1}^n (X_{i,t} - \hat{X})^2$	$nc-1 = (c-1)+c(n-1)$		

On notera que la variation totale (ou somme des carrés totales) est bien la somme de la variation entre les échantillons et de celle au sein des échantillons. Le test de Fisher, dont les degrés de liberté sont

précisés dans la colonne 2, n'est autre que le rapport entre la variance expliquée (par la variable qualitative qui a permis la formation des échantillons) et la variance inexpliquée.

2° CAS : NOMBRES DIFFÉRENTS D'OBSERVATIONS PAR ÉCHANTILLON

L'adaptation du tableau ANOVA est la suivante :

Source de variation	Variation (ou somme des carrés) (1)	degrés de liberté (2)	Variance (3)	Fisher
Entre échantillon	$SC_{\text{echant}} = \sum_{i=1}^c n_i (\bar{X}_i - \hat{X})^2$	$(c-1)$	$s_X^2 = (1)/(2)$	$F = \frac{s_X^2}{s_p^2}$
Au sein des échantillons (ou résiduelle)	$SC_{\text{res}} = \sum_{i=1}^c \sum_{t=1}^n (X_{it} - \bar{X}_i)^2$	$\sum_{i=1}^c (n_i - 1)$	$s_p^2 = (1)/(2)$	
Total	$SC_{\text{tot}} = \sum_{i=1}^c \sum_{t=1}^n (X_{it} - \hat{X})^2$	$\left[\sum_{i=1}^c (n_i) \right] - 1$		

On soulignera enfin que l'on suppose que les variances des différents échantillons étaient égales. Cette hypothèse est dans la pratique peu testée. Un test possible est celui de Bartlett. Si s_i^2 est la variance de l'échantillon i et si la variance des différents échantillons sont égales (hypothèse nulle), alors la grandeur :

$$(n - c) \ln \left(\frac{\sum_{i=1}^c (n_i - 1) s_i^2}{n - c} \right) - \sum_{i=1}^c (n_i - 1) \ln(s_i^2)$$

suit approximativement une distribution du Chi2 à $c-1$ degrés de liberté.

L'analyse de la variance à deux facteurs

L'extension à l'étude de deux facteurs est lourde en termes de notation mais représente une généralisation directe des propositions faites pour l'analyse à un facteur. Le tableau des données étant composé de r lignes (facteur 1) et de c colonnes (facteur 2)¹ et en adoptant la notation $\bar{X}_{i.}$ ² pour la moyenne des colonnes et $\bar{X}_{.j}$ pour la moyenne de chaque ligne, le tableau ANOVA prend la forme suivante :

¹ Les éléments du tableau des données sont notés X_{ij} où i est l'indice des colonnes et j est l'indice des lignes.

² $\bar{X}_{i.} = \frac{\sum_{j=1}^c X_{ij}}{r}$

Source de variation	Variation (ou somme des carrés) (1)	degrés de liberté (2)	Variance (3)	Fisher
Entre colonnes	$SC_{col} = r \sum_{i=1}^c (\bar{X}_{.i} - \hat{X})^2$	(c-1)	$VAR_{col} = (1)/(2)$	$F = \frac{VAR_{col}}{VAR_{res}}$
Entre lignes	$SC_{ligne} = c \sum_{j=1}^r (\bar{X}_{.j} - \hat{X})^2$	(r-1)	$VAR_{ligne} = (1)/(2)$	$F = \frac{VAR_{ligne}}{VAR_{res}}$
Résiduelle	$SC_{res} = \sum_{i=1}^c \sum_{j=1}^r (X_{ij} - \bar{X}_{.i})^2$	(c-1)x(r-1)	$VAR_{res} = (1)/(2)$	
Total	$SC_{tot} = \sum_{i=1}^c \sum_{t=1}^n (X_{it} - \hat{X})^2$	cr-1		

On soulignera qu'on a implicitement supposé qu'il n'y a pas d'interactions entre les deux facteurs et que leurs influences respectives sont additives. La prise en compte des interactions supposent plusieurs observations pour chaque croisement de facteur et le développement d'un modèle plus complexe.