Analyse de la variance

adapté du cours de S. Ducay

F. Wlazinski

Licence d'économie

Dans ce qui suit, on considère un entier $k \geq 2$.
L'analyse de la variance, ANOVA en abrégé pour **an**alysis **o**f **va**riance, permet de comparer les moyennes de plusieurs populations à partir d'échantillons indépendants, afin de tester l'influence d'un ou de plusieurs facteurs.

1 Comparaison de k variances : test de Bartlett

Principe

Dans $k(\geq 2)$ populations P_1, P_2, ..., P_k, on étudie le même caractère.
Soient X_1, X_2, ..., X_k des variables aléatoires représentant le caractère dans chaque population, de moyennes respectives μ_1, μ_2, ..., μ_k, d'écart-types respectifs σ_1, σ_2, ..., σ_k.
Pour tout $i = 1, \ldots, k$, on extrait de P_i un échantillon $E_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n_i})$ de taille n_i de X_i.
La moyenne de chaque échantillon E_i est $\overline{x_i} = \dfrac{1}{n_i}\displaystyle\sum_{j=1}^{n_i} x_{i,j}$.
Et sa variance corrigée est $s_{c,i}^2 = \dfrac{n_i}{n_i - 1} s_i^2$ avec $s_i^2 = \overline{x_i^2} - (\overline{x_i})^2$.
On suppose que les échantillons E_i sont indépendants et que les X_i suivent les lois normales $\mathcal{N}(\mu_i; \sigma_i)$.
On cherche à évaluer l'égalité des variances.

Méthode

On teste H_0 contre H_1 où l'hypothèse nulle H_0 est "toutes les variances sont égales" c'est-à-dire $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 = \sigma^2$ et $H_1 = \overline{H_0}$ est "au moins deux de ces variances ne sont pas égales" c'est-à-dire il existe un couple (i, j) de $[\![1; k]\!]^2$ tel que $\sigma_i^2 \neq \sigma_j^2$.
On pose $n = \displaystyle\sum_{i=1}^{k} n_i$ et la valeur $s_r^2 = \dfrac{1}{n-k}\displaystyle\sum_{i=1}^{k} (n_i - 1)\, s_{c,i}^2 = \dfrac{1}{n-k}\displaystyle\sum_{i=1}^{k} n_i s_i^2$ est appelée la *variance résiduelle (ou intragroupe)* des échantillons et sera une estimation de σ^2 sous l'hypothèse H_0.
On calcule $b = \dfrac{1}{\lambda}\left[(n-k)\ln s_r^2 - \left(\displaystyle\sum_{i=1}^{k}(n_i - 1)\ln s_{c,i}^2\right)\right]$ avec $\lambda = 1 + \dfrac{1}{3(k-1)}\left[\left(\displaystyle\sum_{i=1}^{k}\dfrac{1}{n_i - 1}\right) - \dfrac{1}{n-k}\right]$.
Grâce à la table du χ^2, en prenant la valeur α fournie ($0,05$ en cas d'absence de donnée) et $d.d.l. = k-1$ (c'est-à-dire $k-1$ degrés de liberté), on détermine le réel $b_{\max}$.
Et on décide que :

- si $b < b_{\max}$, alors on ne peut pas rejeter H_0.
- si $b \geq b_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper.

Remarque 1.1

Exemple 1.2

Les enseignants en mathématiques de Licence 2 de trois universités différentes ont comparé leurs résultats annuels. Le bilan partiel est résumé par le tableau suivant :

Université	Effectif	Moyenne	Ecart-type corrigé
Amiens	140	$9,4$	5
Besançon	128	$10,3$	6
Calais	180	$8,2$	$6,2$

On cherche à savoir si on peut supposer les écarts-types des notes annuelles égaux.

2 Analyse de la variance

En toute rigueur, l'analyse de la variance n'est valable que pour des échantillons extraits de populations gaussiennes de même variance. En général, le non-respect de ces conditions n'a pas trop d'influence sur la validité du test (méthode robuste). L'erreur introduite est cependant d'autant plus forte que les effectifs des échantillons sont faibles et inégaux. On verra dans un prochain chapitre comment procéder en précence de petits échantillons non nécessairement gaussiens.

2.1 Comparaison de k moyennes : analyse de la variance à un facteur

Principe

On reprend la situation de la partie **??**.
Les échantillons E_i sont supposés indépendants. On suppose de plus que les X_i suivent les lois normales $\mathscr{N}(\mu_i;\sigma_i)$ et que $\sigma_1^2=\sigma_2^2=\cdots=\sigma_k^2=\sigma^2$.
Cette dernière hypothèse peut être validée en effectuant le test d'égalité des variances décrit à la partie **??**.
En général, les k populations correspondent aux k modalités d'un caractère contrôlé (par exemple k groupes de malades, chaque groupe recevant un traitement différent).
On cherche à évaluer l'égalité des moyennes.

Méthode

On teste H_0 contre H_1 où l'hypothèse nulle H_0 est "toutes les moyennes sont égales" c'est-à-dire $\mu_1=\mu_2=\cdots=\mu_k$ et $H_1=\overline{H_0}$ est "au moins deux de ces moyennes ne sont pas égales" c'est-à-dire il existe un couple (i,j) de $[\![1;k]\!]^2$ tel que $\mu_i\neq\mu_j$.

On désigne par $\overline{x}$ et s_t^2 la moyenne et la variance corrigée de la réunion des k échantillons.

C'est-à-dire $n=\displaystyle\sum_{i=1}^{k}n_i$, $\overline{x}=\dfrac{1}{n}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}x_{i,j}=\dfrac{1}{n}\sum_{i=1}^{k}n_i\overline{x_i}$ et $s_t^2=\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{i,j}-\overline{x})^2$.

La variance résiduelle (ou intragroupe) $s_r^2=\dfrac{1}{n-k}\displaystyle\sum_{i=1}^{k}(n_i-1)\,s_{c,i}^2=\dfrac{1}{n-k}\sum_{i=1}^{k}n_is_i^2$ caractérise la dispersion des valeurs à l'intérieur des échantillons.
La variance résiduelle s_r^2 est une estimation de σ^2 d'après l'hypothèse d'égalité des variances.

La *variance factorielle (ou intergroupe)* définie par $s_f^2=\dfrac{1}{k-1}\displaystyle\sum_{i=1}^{k}n_i\left(\overline{x_i}-\overline{x}\right)^2$ caractérise la dispersion des valeurs d'un échantillon à l'autre, i.e., la variation due à l'influence du facteur étudié.
La variance factorielle s_f^2 est une estimation de σ^2 sous l'hypothèse H_0.

De plus, on a $(n-1)\,s_t^2=(n-k)\,s_r^2+(k-1)\,s_f^2$. Autrement dit, s_t^2 est une moyenne pondérée de s_r^2 et s_f^2. On peut donc, si nécessaire et suivant les données d'un problème, utiliser cette relation pour déterminer s_f ou s_r en fonction de l'autre et de s_t.

On calcule $f=\dfrac{s_f^2}{s_r^2}$ et on détermine $f_{\max}$ grâce aux tables de Fisher-Snédécor avec $v_1=k-1$ et

$v_2 = n - k$.

Et on décide que :

• si $f < f_{\max}$, alors on ne peut pas rejeter H_0.
• si $f \geq f_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper, i.e., que l'on attribue une influence significative au facteur étudié.

Exemple 2.1

Dans un but d'harmonisation inter-universités, les enseignants en mathématiques de Licence 2 d'économie et gestion des trois universités de l'exemple **??** ont revu leurs résultats annuels. Le nouveau bilan est résumé par le tableau suivant :

Université	Effectif	Moyenne	Ecart-type corrigé
Amiens	140	$9,5$	$5,1$
Besançon	128	$10,1$	6
Calais	180	9	$5,7$

On cherche à savoir si les moyennes dépendent de l'université.

2.2 Analyse de la variance à deux facteurs

Principe

On considère deux facteurs : un facteur A à r modalités et un facteur B à q modalités. Ces deux facteurs déterminent $k = rq$ populations $P_{i,j}$ avec $1 \leq i \leq r$ et $1 \leq j \leq q$.
Dans les k populations $P_{i,j}$ on étudie le même caractère. Soit $X_{i,j}$ la variable aléatoire représentant le caractère dans la population $P_{i,j}$, de moyennes respectives $\mu_{i,j}$ et d'écart-type respectifs $\sigma_{i,j}$.
De chaque population $P_{i,j}$, on extrait un échantillon $E_{i,j} = \left(x_{i,j,1}, x_{i,j,2}, \ldots, x_{i,j,n_{i,j}}\right)$ de taille $n_{i,j}$.
La moyenne de chaque échantillon $E_{i,j}$ est alors $\overline{x_{i,j}} = \dfrac{1}{n_{i,j}} \displaystyle\sum_{k=1}^{n_{i,j}} x_{i,j,k}$.
Et sa variance corrigée est $s^2_{c,i,j} = \dfrac{n_{i,j}}{n_{i,j}-1} s^2_{i,j}$ avec $s^2_{i,j} = \dfrac{1}{n_{i,j}} \displaystyle\sum_{k=1}^{n_{i,j}} x^2_{i,j,k} - \overline{x_{i,j}}^2$.
Les échantillons $E_{i,j}$ sont supposés **indépendants et de même taille** n (pour tous les i,j, $n_{i,j} = n$).
On suppose de plus que les $x_{i,j}$ suivent les lois normales $\mathcal{N}(\mu_{i,j}; \sigma)$. Autrement dit, toutes les variances $\sigma^2_{i,j}$ sont égales à σ^2.
L'analyse de la variance à deux facteurs permet de comparer les moyennes des $k = rq$ échantillons et de tester l'influence du facteur A seul, l'influence du facteur B seul et l'influence de l'interaction des deux facteurs (il y a interaction lorsque l'influence d'un facteur sur la moyenne des populations est différente en l'absence ou en la présence de l'autre facteur). Il y aura donc trois tests d'égalité des moyennes.

Méthode

On désigne respectivement par $\overline{x}$ et s^2_t, la moyenne et la variance corrigée de la réunion des $k = rq$ échantillons (réunion de taille nrq).
On a $\overline{x} = \dfrac{1}{nrq} \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{q}\sum_{k=1}^{n_{i,j}} x_{i,j,k} = \dfrac{1}{rq} \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{q} \overline{x_{i,j}}$ et $s^2_t = \dfrac{1}{nrq-1} \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{q}\sum_{k=1}^{n_{i,j}} (x_{i,j,k} - \overline{x})^2$.
On définit la variance résiduelle $s^2_r = \dfrac{1}{rq} \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{q} s^2_{c,i,j}$, qui caractérise la dispersion des valeurs à l'intérieur des échantillons, et la variance factorielle $s^2_f = \dfrac{1}{rq-1} \displaystyle\sum_{i=1}^{r}\sum_{j=1}^{q} n\left(\overline{x_{i,j}} - \overline{x}\right)^2$, qui caractérise la dispersion des valeurs d'un échantillon à l'autre, i.e. la variation due à l'influence des facteurs étudiés.
On a alors $(nrq-1)\, s^2_t = (n-1)\, rq\, s^2_r + (rq-1)\, s^2_f$. Ainsi, s^2_t est une moyenne pondérée de s^2_r et s^2_f.

Pour étudier l'influence de chacun des deux facteurs et de leur interaction, on définit :

$\overline{x_{i,\star}} = \dfrac{1}{q}\sum_{j=1}^{q} \overline{x_{i,j}}$, moyenne conditionnelle à la $i^{ème}$ modalité du facteur A

$\overline{x_{\star,j}} = \dfrac{1}{r}\sum_{i=1}^{r} \overline{x_{i,j}}$, moyenne conditionnelle à la $j^{ème}$ modalité du facteur B

$s_A^2 = \dfrac{nq}{r-1}\sum_{i=1}^{r} \left(\overline{x_{i,\star}} - \overline{x}\right)^2$, variance factorielle due au facteur A seul

$s_B^2 = \dfrac{nr}{q-1}\sum_{j=1}^{s} \left(\overline{x_{\star,j}} - \overline{x}\right)^2$, variance factorielle due au facteur B seul

$s_{AB}^2 = \dfrac{n}{(r-1)(q-1)}\sum_{i=1}^{r}\sum_{j=1}^{q} \left(\overline{x_{i,j}} - \overline{x_{i,\star}} - \overline{x_{\star,j}} + \overline{x}\right)^2$, variance factorielle due à l'interaction de A et B.

En particulier, s_f^2 s'obtient grâce à la formule $(rs-1)\,s_f^2 = (r-1)\,s_A^2 + (s-1)\,s_B^2 + (r-1)\,(s-1)\,s_{AB}^2$. Ainsi, s_f^2 est une moyenne pondérée de s_A^2, s_B^2 et s_{AB}^2.

Test 1

On teste H_0 contre H_1 où $H_0 = H_{0,A}$ est "le facteur A n'a pas d'influence sur la moyenne des populations" et $H_1 = \overline{H_{0,A}}$.

On calcule s_r^2, s_A^2 et $f_A = \dfrac{s_A^2}{s_r^2}$.

On détermine $f_{\max}$ grâce aux tables de Fisher-Snédécor avec $v_1 = r-1$ et $v_2 = (n-1)rq$.

Et on décide que :

• si $f_A < f_{\max}$, alors on ne peut pas rejeter H_0.
• si $f_A \geq f_{\max}$, alors on rejette H_0 avec une probabilité α de se tromper, i.e., que l'on attribue une influence significative au facteur étudié.

Test 2

On teste H_0 contre H_1 où $H_0 = H_{0,B}$ est "le facteur B n'a pas d'influence sur la moyenne des populations" et $H_1 = \overline{H_{0,B}}$.

On calcule s_r^2, s_B^2 et $f_B = \dfrac{s_A^2}{s_r^2}$.

On détermine $f_{\max}$ grâce aux tables de Fisher-Snédécor avec $v_1 = q-1$ et $v_2 = (n-1)rq$.

On conclut comme au test 1.

Test 3

On teste H_0 contre H_1 où $H_{0,AB}$ est "il n'y a pas d'interaction entre les facteurs A et B" et $H_1 = \overline{H_{0,AB}}$.

On calcule s_r^2, s_{AB}^2 et $f_{AB} = \dfrac{s_{AB}^2}{s_r^2}$.

On détermine $f_{\max}$ grâce aux tables de Fisher-Snédécor avec $v_1 = (r-1)(q-1)$ et $v_2 = (n-1)rq$.

On conclut comme au test 1.

Exemple 2.2

Dans une expérience, on présente à chaque sujet soit oralement soit par écrit des mots qui sont soit familiers, soit non familiers. Après une période d'attente, on interroge le sujet et on calcule le nombre de syllabes non significatives mémorisées. 24 sujets ont participé, répartis en 4 groupes de 6. Les résultats sont les suivants :

Oral	Familier	20	17	18	23	14	16
Oral	Non familier	16	14	8	10	9	12
Ecrit	Familier	10	12	18	16	17	14
Ecrit	Non familier	10	17	14	11	12	8

On cherche à tester les hypothèses suivantes :

Hypothèse 1 : le mode de présentation (oral ou écrit) n'a pas d'effet sur la mémorisation.
Hypothèse 2 : la nature des mots présentés (familier ou non) n'a pas d'effet sur la mémorisation.
Hypothèse 3 : il n'y a pas d'effet d'interaction sur la mémorisation.

Remarque 2.3

Si chaque échantillon ne comporte qu'une seule observation, i.e. $n = 1$, alors les $s^2_{i,j}$ sont nuls et les $s^2_{c,i,j}$ ne sont pas définis. De plus, s^2_r est nul et les quotients f_A, f_B et f_{AB} ne sont pas définis.
Dans ce cas, on effectue les tests de la manière suivante :

Test 1 : $H_{0,A}$ contre $H_1 = \overline{H_{0,A}}$.

On compare $f_A = \dfrac{s^2_A}{s^2_{AB}}$ à $f_{\max}$ obtenu dans la table de Fisher-Snédécor pour $v_1 = r - 1$ et $v_2 = (r-1)(q-1)$.

Test 2 : $H_{0,B}$ contre $H_1 = \overline{H_{0,B}}$.

On compare $f_B = \dfrac{s^2_B}{s^2_{AB}}$ à $f_{\max}$ obtenu dans la table de Fisher-Snédécor pour $v_1 = q - 1$ et $v_2 = (r-1)(q-1)$.

Et il n'y a pas de Test 3.

Exemple 2.4

Pour tester la fiabilité de 4 laboratoires d'analyse, on utilise 3 solutions ayant le même titre de glucose dans du sérum physiologique additionné de quantités variables de galactose. Chaque laboratoire reçoit un échantillon de chaque solution et fournit le résultat de ses mesures. L'ensemble des résultats, exprimés en grammes de glucose pour 10 litres de solution, est regroupé dans le tableau suivant :

Solution \ Laboratoire	L1	L2	L3	L4
S1	10,5	11,5	10,8	11,3
S2	11,2	11,5	11,1	10,9
S3	10,2	11,0	10,4	10,5

On cherche à savoir, au risque de 5%, si :
- le choix du laboratoire a une influence sur la mesure du taux de glucose.
- la quantité de galactose présente dans la solution a une influence sur la mesure du taux de glucose.