Chapitre 12

Analyse de la variance à simple entrée

L'analyse de la variance [1] permet de généraliser le test de comparaison de plusieurs échantillons au problème suivant : **la comparaison des moyennes de plusieurs échantillons indépendants**. Ainsi, comme son nom ne l'indique pas, l'analyse de la variance permet de **comparer des moyennes** – c'est la méthode à utiliser lorsqu'il faut comparer plusieurs moyennes (à partir de trois). Il faut ajouter que le procédé qui consiste à tester l'égalité des moyennes de chaque couple n'est pas satisfaisant. Aussi, la nécessité d'une procédure permettant de **tester globalement** l'ensemble de tous les échantillons est fournie par la **théorie de l'analyse de la variance**. Le but de cette théorie est d'étudier la variabilité d'un objet en fonction d'un ensemble de facteurs que l'on peut contrôler systématiquement, et que l'on souhaite dissocier la part revenant à chaque facteur.

L'An.O.Va. permet d'étudier la dépendance d'une variable quantitative à une ou deux variables qualitatives. Plus généralement, les variables qualitatives sont appelées **facteurs**. Le facteur contrôlé peut intervenir dans des conditions qui diffèrent : 1. soit par leur nature, 2. soit par leur intensité. De plus, le facteur contrôlé peut être : 1. soit à effets fixes, 2. soit à effets aléatoires.

La variable dépendante (V.D.) est une variable quantitative continue. Les variables indépendantes (V.I.) correspondent aux facteurs. De fait, l'étude d'un facteur est une analyse bivariée, tandis que, avec au moins deux facteurs, l'analyse est multivariée. L'An.O.Va. établit si la dépendance étudiée est significative pour le facteur considéré. Pour y répondre, il faut tester si la moyenne de la variable quantitative d'étude est homogène sur l'ensemble des modalités de la variable qualitative. Il faut rejeter l'hypothèse nulle H_0 d'égalité des moyennes par l'analyse de la variance. Le test utilisé est le test F de Fisher consistant à comparer la

1. En anglais : Analysis of Variance (An.O.Va.)

variance inter-échantillon à la variance intra-échantillon. On tente d'expliquer la **cause** de la diversité des informations par l'analyse de leur variance.

Il faut noter que l'analyse de la variance n'est valable en toute rigueur que pour des **échantillons tirés de populations normales et de même variance**. En général, le non-respect de ces conditions n'a pas trop d'influence sur la validité du test. Dit autrement, l'analyse de la variance est une **méthode robuste**. L'erreur introduite est toutefois d'autant plus forte que les effectifs des échantillons sont faibles et inégaux.

L'**analyse de la variance à simple entrée** correspond à un seul facteur qui est contrôlé, les autres facteurs étant regroupés sous le nom « facteurs non contrôlés ».

12.1 Série statistique des observations

On suppose que le facteur contrôlé A prend k modalités A_i (Tab. 12.1).

Facteur A	A_1	$\dots$	A_i	$\dots$	A_k
Échantillons observés	x_1^1	$\dots$	x_i^1	$\dots$	x_k^1
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	x_1^j	$\dots$	x_i^j	$\dots$	x_k^j
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$x_1^{n_1}$	$\dots$	$x_i^{n_i}$	$\dots$	$x_k^{n_k}$
Moyenne de l'échantillon	$\bar{x}_1$	$\dots$	$\bar{x}_i$	$\dots$	$\bar{x}_k$
Nombre d'observations dans l'échantillon	n_1	$\dots$	n_i	$\dots$	n_k

TABLE 12.1 – Facteur contrôlé prenant k modalités

On note N le nombre total d'observations, $\bar{x}_i$ la moyenne des observations pour la modalité i, et $\bar{x}$ la moyenne générale des observations.

$$N = \sum_{i=1}^{k} n_i \tag{12.1}$$

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_i^j \tag{12.2}$$

$$\bar{x} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_i^j \tag{12.3}$$

N.B. Dans ce chapitre, l'exposant j sera un indice, sauf mention contraire.

12.2 Mise en œuvre du test

On suppose que le facteur contrôlé agit sur les moyennes, mais n'agit pas sur les variances, ce qui, en toute rigueur, devrait être contrôlé.

La loi de la variable aléatoire parente X_i est, pour toutes les valeurs de l'indice i, une loi normale $N(\mu_i, \sigma)$. Chaque observation s'écrit, en désignant par ξ une fluctuation aléatoire normale :

$$x_i^j = \mu_i + \xi_i^j \tag{12.4}$$

avec $\mathbb{E}(\xi) = 0$ et $\mathbb{V}(\xi) = \sigma^2$.

Les hypothèses à tester sont :

$$\begin{cases} H_0 : \forall i, \mu_i = \mu \\ H_1 : \exists i, j \backslash \mu_i \neq \mu \end{cases} \tag{12.5}$$

Sous H_0, la population est homogène, le facteur contrôlé n'exerce aucune influence sur la production. On peut alors comparer toutes les observations à la moyenne générale $\bar{x}$.

12.3 Définition des variations

La statistique T de la variation totale est la somme des carrés des écarts par rapport à la moyenne générale, elle est définie par :

$$T = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_i^j - \bar{X}\right)^2 \tag{12.6}$$

Le quotient $S^2 = \frac{T}{N}$ est la **variance totale**.

L'écart $X_i^j - \bar{X}$ peut s'écrire : $X_i^j - \bar{X} = \left(X_i^j - \bar{X}_i\right) + \left(\bar{X}_i - \bar{X}\right)$. La différence $X_i^j - \bar{X}_i$ correspond aux écarts des observations par rapport à la moyenne pour chaque modalité du facteur contrôlé. La différence $\bar{X}_i - \bar{X}$ correspond aux écarts des différentes moyennes par rapport à la moyenne générale.

La statistique $A = \sum_{i=1}^{k} n_i \left(\bar{X}_i - \bar{X}\right)^2$ est la **variation due au facteur contrôlé**. Le quotient ${S_A}^2 = \frac{A}{N}$ est la **variance due au facteur contrôlé** (ou variance inter-échantillon, ou variance factorielle, ou variance inter-groupe).

La statistique $R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(X_i^j - \bar{X}_i\right)^2$ est la **variation résiduelle**. Le quotient ${S_R}^2 = \frac{R}{N}$ est la **variance résiduelle** (ou variance intra-échantillon ou variance intra-groupe).

Un calcul facile conduit au résultat suivant :

$$T = A + R \tag{12.7}$$

ou

$$S^2 = {S_A}^2 + {S_R}^2 \tag{12.8}$$

La variance totale S^2 est égale à la somme de la variance des moyennes et de la moyenne des variances.

12.3.1 Étude de la statistique ${S_R}^2$

Par hypothèse, les variables aléatoires X_i suivent des lois normales $N(\mu_i, \sigma)$, d'où la statistique :

$${S_i}^2 = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(X_i^j - \bar{X}_i\right)^2 \tag{12.9}$$

est telle que $\frac{n_i {S_i}^2}{\sigma^2}$ suit la loi du $\chi^2(n_i - 1)$.

La variance résiduelle qui est égale à :

$${S_R}^2 = \frac{1}{N} \sum_{i=1}^{k} n_i {S_i}^2 \tag{12.10}$$

est telle que $\frac{N {S_R}^2}{\sigma^2}$ suit la loi du $\chi^2(N - k)$ par l'application de la propriété d'additivité du χ^2. On en déduit que ${S_R}^2$ est une estimation de la variance σ^2 à $(N - k)$ degrés de liberté.

12.3.2 Étude de la statistique S^2

Si l'hypothèse H_0 est vraie, les variables X_i suivent la même loi normale $N(\mu, \sigma)$. La statistique $\frac{NS^2}{\sigma^2}$ suit alors la loi du $\chi^2(N - 1)$.

12.3.3 Étude de la statistique ${S_A}^2$

Cette statistique peut être considérée comme la variance de l'échantillon formé par les k moyennes $\bar{X}_i$ pondérées par les effectifs n_i. On en déduit que la statistique $\frac{N {S_A}^2}{k-1}$ suit la loi du $\chi^2(k - 1)$.

De l'équation de l'analyse de la variance H_0, on en déduit que les statistiques ${S_A}^2$ et ${S_R}^2$ sont indépendantes. On en déduit le test suivant : si H_0 est vraie, et, d'après les définitions de ${S_A}^2$ et ${S_R}^2$, on a :

$$\frac{{S_A}^2}{k-1} \times \frac{N-k}{{S_R}^2} = F\left(k-1, N-k\right) \tag{12.11}$$

On rejette H_0 si :

$$\frac{{S_A}^2}{k-1} \times \frac{N-k}{{S_R}^2} > f_\alpha \tag{12.12}$$

la valeur critique f_α est lue dans les tables de Fisher et dépend évidemment du seuil α choisi. Ce résultat signifie que le facteur contrôlé a une **influence significative**.

Remarque. Plus F est grand, moins H_0 est crédible.

12.3.4 Calcul rapide des différentes statistiques

Les résultats suivants sont faciles à démontrer après le développement des différents termes carrés :

$$NS^2 = T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_i^j\right)^2 - \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_i^j\right)\right)^2 \tag{12.13}$$

avec $N-1$ degré de liberté. Le calcul S^2 est la moyenne des carrés moins le carré de la moyenne. La quantité $\Delta = \frac{1}{N}\left(\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_i^j\right)\right)^2$ est un terme correctif. On peut alors écrire ${NS_A}^2$ et ${NS_R}^2$.

${NS_A}^2 = \sum_{i=1}^{k}\frac{1}{n_i}\left(\sum_{j=1}^{n_i} X_i^j\right)^2 - \Delta$ possède $k-1$ degrés de liberté.

${NS_R}^2 = NS^2 - {NS_A}^2$ possède $N-k$ degré de liberté.

Tous ces résultats sont résumés dans le tableau d'analyse de la variance (Tab. 12.2) qui permet un contrôle commode des calculs.

Variation	Somme des carrés	Degré de liberté	Quotient
Variation due au facteur	${NS_A}^2$	$k-1$	$v_A = \frac{{NS_A}^2}{k-1}$
Variation résiduelle	${NS_R}^2$	$N-k$	$v_R = \frac{{NS_R}^2}{N-k}$
Variation totale	NS^2	$N-1$	

TABLE 12.2 – Synthèse des variations d'une analyse de la variance à simple entrée

La variance résiduelle au sein des colonnes est inexpliquée parce qu'elle est la variation aléatoire ou contingente, qui ne peut être expliquée systématiquement.

F est le rapport entre la variance expliquée et la variance inexpliquée :

$$F = \frac{\upsilon_A}{\upsilon_R} \tag{12.14}$$

Remarque. Le modèle le plus efficace demeure celui où l'on fait en sorte que tous les échantillons aient la même taille n.

12.3.5 Conclusion

Au seuil de confiance α, on accepte H_0 lorsque le facteur contrôlé n'a pas d'influence, donc dans le cas d'une population homogène, c'est-à-dire si :

$$\frac{\upsilon_A}{\upsilon_R} < F_\alpha\left(k-1, N-k\right) \tag{12.15}$$

Dans ce cas, on prend comme estimation de μ, la moyenne générale des observations, et comme estimation de la variance υ_R. De plus, on peut donner un intervalle de confiance pour la valeur μ, car la variable aléatoire $t = \frac{\mu-\bar{x}}{\sqrt{N{S_R}^2}}\sqrt{N-k}$ est la variable de Student à $N-k$ degrés de liberté.

On refuse H_0 lorsque le facteur exerce une influence, donc pour une population non homogène, c'est-à-dire si

$$\frac{\upsilon_A}{\upsilon_R} > F_\alpha\left(k-1, N-k\right) \tag{12.16}$$

Dans ce cas, les observations s'écrivent sous la forme $x_i^j = \mu_i + \xi_i^j$ dans laquelle le terme μ_i est une correction correspondant au niveau i et le terme ξ_i^j est une fluctuation aléatoire suivant la loi normale $N\left(0, \sigma\right)$, la variance étant indépendante du niveau choisi. Une estimation de chaque terme μ_i est donnée par chaque moyenne $\bar{x}_i$.

12.4 Exemple d'analyse de la variance à simple entrée

On veut comparer l'usure de quatre types de pneumatique P_1, P_2, P_3 et P_4. Sur chacun d'eux, on fait un certain nombre d'essais, 4 ou 5 ; les coefficients d'usure mesurés sont fournis (Tab. 12.3). Peut-on considérer que les quatre types de pneumatiques sont équivalents ?

On commence par calculer les statistiques des différents résultats :

— Pour la pneumatique P_1, sa moyenne vaut $\bar{x}_1 = 3,75$ et sa variance vaut ${s_1}^2 = 0,6875$.

Numéro de l'essai	P_1	P_2	P_3	P_4
I	3	1	2	3
II	3	1	5	3
III	4	2	6	2
IV	5	4	4	1
V			4	4
Total	15	8	21	13

TABLE 12.3 – Exemple. Les coefficients d'usure de quatre types de pneus

— Pour la pneumatique P_2, sa moyenne vaut $\bar{x}_2 = 2,00$ et sa variance vaut ${s_2}^2 = 1,5000$.

— Pour la pneumatique P_3, sa moyenne vaut $\bar{x}_3 = 4,20$ et sa variance vaut ${s_3}^2 = 1,7600$.

— Pour la pneumatique P_4, sa moyenne vaut $\bar{x}_4 = 2,60$ et sa variance vaut ${s_4}^2 = 1,0400$.

Puis, on effectue le test sur l'égalité des variances en comparant les estimations des variances des échantillons III et IV (la plus petite et la plus grande). Si ces variances peuvent être considérées comme égales, le rapport $\frac{5\times1,7600}{4} \times \frac{3}{4\times0,6875} = 2,40$ est la réalisation d'une variable de Fisher $F(4,3)$ telle que $\Pr(F(4,3) > 9,12) = 0,05$. On ne peut pas rejeter l'hypothèse d'égalité des variances des échantillons III et IV.

De fait, les quatre variances peuvent être considérées comme égales. Les paramètres de l'analyse des variances sont :

1. le nombre total d'observations : $N = 18$;
2. la somme de tous les termes : $15 + 8 + 21 + 13 = 57$;
3. la variation totale : $NS^2 = 3^2 + 3^2 + \ldots + 1^2 + 4^2 - \frac{57^2}{18} = 36,5$;
4. la variation due au facteur : ${NS_A}^2 = \frac{15^2}{4} + \frac{8^2}{4} + \frac{21^2}{5} + \frac{13^2}{5} - \frac{57^2}{18} = 13,75$;
5. la variation résiduelle : ${NS_R}^2 = 36,5 - 13,75 = 22,75$.

ce qui permet d'obtenir le tableau de l'analyse de la variance (Tab. 12.4).

Variation	Somme des carrés	Degré de liberté	Quotient
Variation due au facteur	13,75	3	$v_A = 4,58$
Variation résiduelle	22,75	14	$v_R = 1,625$
Variation totale	36,50	17	

TABLE 12.4 – Synthèse des variations d'une analyse de la variance à simple entrée

On obtient $\frac{v_A}{v_R} = 2,82$ et $\Pr(F(3,14) > 3,34) = 0,95$. On peut admettre que la population est homogène. Il n'y a pas de différence entre les quatre types de pneumatiques. L'estimation de l'usure moyenne est égale à $\frac{57}{18} = 3,17$, celle que l'on avait appelée moyenne générale. L'estimation de la variance au quotient est $v_R = 1,625$.

12.5 Liaison entre une variable quantitative et une variable qualitative

Pour étudier la liaison entre une variable quantitative Y et une variable qualitative X, définies sur un ensemble de n individus, on utilise le rapport de corrélation.

12.5.1 Rapport de corrélation théorique

Le rapport de corrélation η^2 de la variable Y en la variable X est donné par :

$$\eta_{Y\setminus X}{}^2 = \frac{\mathbb{V}(E(Y \setminus X))}{\mathbb{V}(Y)} \tag{12.17}$$

Sa valeur est comprise entre 0 et 1.

Pour $\eta_{Y\setminus X}{}^2 = 1$, si $\mathbb{V}[\mathbb{E}(Y \setminus X)] = \mathbb{V}(Y)$. Dans ces conditions, $\mathbb{E}[\mathbb{V}(Y \setminus X)] = 0$. Comme $\mathbb{V}(Y \setminus X) = 0$ presque sûrement, ou, en d'autres termes, à une valeur de X fixée, $\mathbb{V}(Y \setminus X) = 0$, Y ne prend qu'une valeur. Cela signifie que la variable aléatoire Y est liée de manière fonctionnelle à la variable aléatoire X : $Y = f(X)$.

Pour $\eta_{Y\setminus X}{}^2 = 0$, si $\mathbb{V}[\mathbb{E}(Y \setminus X)] = 0$. Il en résulte que $\mathbb{E}(Y \setminus X)$ est presque sûrement une constante. Cela signifie que la variable Y est non corrélée avec la variable aléatoire X.

12.5.2 Rapport de corrélation empirique

La variable qualitative X prend k modalités. On note : n_i l'effectif observé pour la variable Y lorsque la variable X prend la modalité i, $\bar{Y}_i$ la moyenne des n_i valeurs prises par la variable Y pour la modalité i de la variable X, et $\bar{Y}$ la moyenne générale des valeurs prises par la variable Y. Le rapport empirique de corrélation e^2 est donné par :

$$e^2 = \frac{1}{n}\frac{\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2}{s_Y{}^2} \tag{12.18}$$

avec $s_Y{}^2 = \sum_{i=1}^{k}\sum_{j=1}^{l}\left(Y_{ij} - \bar{Y}\right)^2$.

Le coefficient e^2 est compris entre 0 et 1.

— Si $e^2 = 0$, pour toutes les valeurs de l'indice i, on a $\bar{y}_i = \bar{y}$. Il n'existe pas de dépendance en moyenne.

— Si $e^2 = 1$, pour une modalité i de la variable X, tous les individus ont la même valeur, et ceci pour toutes les valeurs de l'indice i :

$$i^2 = 1 \Rightarrow \frac{1}{n}\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = {s_Y}^2 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l}\left(Y_{ij} - \bar{Y}\right)^2 \qquad (12.19)$$

— Si $e^2 \neq 0$, on utilise les résultats de l'analyse de la variance à simple entrée. La suite du texte en présente les liens.

La quantité $\frac{1}{n}\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{y}\right)^2 = e^2 {s_Y}^2 = {S_A}^2$ représente les variations entre les différentes modalités, c'est-à-dire la variation expliquée par le facteur contrôlé. Il s'agit de la réalisation d'une variable χ^2 à $\upsilon = (k-1)$ degrés de liberté.

La variation totale est représentée par la quantité ${s_Y}^2$, et la variation résiduelle par la différence :

$${s_Y}^2 - e^2 {s_Y}^2 = \left(1 - e^2\right){s_Y}^2 = {S_R}^2 \qquad (12.20)$$

Le rapport $F = \frac{\frac{{S_A}^2}{k-1}}{\frac{{S_R}^2}{n-k}} = \frac{\frac{e^2}{k-1}}{\frac{1-e^2}{n-k}}$ suit une loi de Fisher à $\upsilon = (k-1, n-k)$ degrés de liberté sous l'hypothèse H_0, $\eta^2 = 0$ correspondant au test de l'analyse de la variance. Si le rapport $F = \frac{\frac{{S_A}^2}{k-1}}{\frac{{S_R}^2}{n-k}}$ est supérieur à la valeur critique, pour un seuil donné α, d'une variable de Fisher $F\left(k-1, n-k\right)$, on rejette H_0.

Remarque importante. Pour appliquer ces résultats, il faut supposer que, pour chaque modalité du facteur contrôlé, les distributions de Y suivent des lois normales de même espérance et de même variance.

Bibliographie