Notions clés :

Chapitre 03 : Analyse de la variance d'un facteur — ANOVA

I. Définition et exemple :

1) Définition et principe de l'ANOVA

l'ANOVA à un facteur permet de tester l'effet d'un facteur qualitatif (x) ayant k modalités ou niveaux ($k \geq 3$) sur une variable quantitative y

=> objectif : Comparer les moyennes de la variable y pour chaque modalité du facteur

Le principe de l'ANOVA repose sur la dispertion des données autour de la moyenne (variance)

Cette dispertion peut avoir deux origines :

L'effet du facteur étudié :

Dans ce cas, une partie de la dispertion est imputable aux modalités du facteur étudié. Cette partie de dispertion est appellée variabilité factorielle ou variabilité interclasse

Variabilité résiduelle ou intra-classe :

Il s'agit de la variabilité qui reste lorsque la variabilité factorielle est soustraite à la variabilité total, elle correspond à la part qui n'est pas expliqué

ainsi le principe de l'ANOVA est de determiner à l'aide d'un teste statistique que si les moyennes de la variable quantitative y sont toutes égales ou non.

2) Exemple :

On considère une étude sur la plantation d'arbre dans 3 forêts, on se propose de comparer les hauteurs moyennes des arbres. Pour ce faire on dispose de

forêt 1	forêt 2	forêt 3
23,3	18,9	22,5
24,4	21,1	22,9
24,6	21,1	23,7
24,9	22,1	24,0
25,0	22,5	24,0
26,2	23,5	24,5

soient les forêts les variable qualitatives ayant 3 modalités 1, 2, 3.

La hauteur des arbres : la variable quantitative y

Problématique : peut-on considérer que la hauteur moyenne des arbres diffère selon les forêts ?

II Modèles statistique et hypothèses :

1) Modèle statistique :

ANOVA est un modèle de regression linéaire qui s'écrit

$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \forall \begin{cases} i = 1, \dots, k \\ j = 1, \dots n_i \end{cases}$$

Avec : μ_i : est le paramètre qui représente l'effet de la modalité i de la variable qualitative x

ε_{ij} : ce sont les résidus c.à.d les écart entre les observations et les moyennes des groupes auxquelles elles sont relative $\varepsilon_{ij} = y_{ij} - \overline{y_i}$

Scanné avec CamScanner

2) Hypothèses de l'ANOVA :

H_1 : les résidus ε_{ij} sont indépendants

H_2 : les résidus sont homogènes ou encore homoscédastiques

$$\mu_{ij} = \mu_i + \varepsilon_{ij} \quad \forall \begin{cases} i = 1, \ldots, 3 \\ j = 1, \ldots, 6 \end{cases}$$

III Notation et calcul de la dispersion totale :

1) Notation :

* Le facteur étudiés (A_i ici, les forêts) comporte 3 modalités

+ le nombre d'observation pour chacune des modalité est noté n_i (ici, $n_1 = n_2 = n_3 = 6$)

* le nombre total d'observations est notés n tel que $n = \sum_{i=1}^{3}$,
(ici, $n = 6+6+6 = 18$)

* les observations sont notés y_{ij},
(ici y_{ij} correspond à la hauteur des arbres)

c'est l'observation relative à la hauteurs du $j^{ème}$ arbre dans les modalités

* i : est l'indice des modalités telle que
$i = 1, \ldots, b$ (ici $i = 1 \ldots 3$)

* j : est l'indice de l'observation au sein d'une modalité tel que $j = 1, \ldots, n_i$ (ici $j = 1, \ldots 6$)

les moyennes des observations de chaque modalité sont noté $\overline{y_i}$ $\Rightarrow$ $\overline{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$

* la moyenne générale ou globale des observations est noté $\bar{y}$

$$\Rightarrow \bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij}) = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{y}_i$$

2) Calcul de la dispertion totale :

Pour mesurer la dispertion / variabilité totale on utilise la somme des carrés totale SCT qui correspond à la somme des ~~distance~~ au carré entre chaque valeur ~~~~ observé et la moyenne générale ou globale. La SCT se calcule

$$\boxed{SCT = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (y_{ij} - \bar{y})^2}$$

IV Decomposition de la variance ou de la dispertion totale

La variabilité totale mesurer par la SCT ~~est~~ décomposer en

* la variabilité factorielle (due au facteur qualitatif) mesuré par la somme des carrées factorielle

$$\Rightarrow \boxed{SCF = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2}$$

$\Rightarrow$ variabilité interclasse.

* la variabilité résiduelle ou intraclasse menée par la somme des carrés résiduelles noté SCR

$$\Rightarrow \boxed{SCR = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \varepsilon_{ij}^2}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

l'équation de l'ANOVA s'écrit comme suit :

$$SCT = SCF + SCR \Rightarrow \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_i - \bar{y})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

V Testes d'hypothèses : Teste de Fisher

1) Tableau d'analyse de la variance

Source de variation	Somme des carrés	d.d.l	Carrés moyens	Statistique F de Fisher	
				calculé	théorique
Facteur	$SCF = \sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2$	$k-1$	$CM_F = \frac{SCF}{k-1}$	$F_c = \frac{CM_F}{CM_R}$	$F_\alpha(k-1, n-k)$
Résidu	$SCR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$	$n-k$	$CM_R = \frac{SCR}{n-k}$		
Total	$SCT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2$	$n-1$	variance empirique $= \frac{SCT}{n-1}$		

2) Teste de Fisher :

* Hypothèses :

$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ contre H_1 : il existe au moins un μ_i différents des autres.

* Statistique et loi :

Sous H_0 vraie on a $F_c = \dfrac{SCF/k-1}{SCR/n-k} = \dfrac{\sum_{i=1}^{k} n_i(\bar{y}_i - \bar{y})^2/k-1}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2/n-k} \rightsquigarrow F_\alpha(k-1, n-k)$

* Règle de décision

- si $F_c < F_t = F_\alpha(k-1, n-k)$ alors H_0 est vraie
- si $F_c > F_c$ alors H_1 est vraie

$\frac{n}{m}$ 1 2 3 4

VI Application à l'exemple :

* moyenne de chaque modalité :

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij})$$

* $\bar{y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} (y_{1j})$

$$\Rightarrow \boxed{\bar{y}_1 = 24,733}$$

* $\bar{y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} (y_{2j})$

$$\Rightarrow \boxed{\bar{y}_2 = 21,533}$$

* $\bar{y}_3 = \frac{1}{n_3} \sum_{j=1}^{n_3} (y_{3j})$

$$\boxed{\bar{y}_3 = 23,6}$$

* moyenne générale des observations :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij})$$

$$= \frac{1}{n} \sum_{i=1}^{k} n_i \times \bar{y}_i$$

$$\Rightarrow \bar{y} = \frac{1}{18} \left(n_1 \times \bar{y}_1 + n_2 \times \bar{y}_2 + n_3 \times \bar{y}_3 \right)$$

$$= \frac{1}{18} \left(6 \times 24,733 + 6 \times 21,533 + 6 \times 23,6 \right)$$

$$\boxed{\bar{y} = 23,289}$$

* ~~Règle de décision~~ :
  - ~~si $F_c \ll F_t = F_\alpha(k-1,$~~

* Équation d'analyse de la variance

$$SCT = SCF + SCR \iff$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

* $SCF = \sum_{i=1}^{k} n_i(\bar{y}_i-\bar{y})^2$

$$= \sum_{i=1}^{3} n_i(\bar{y}_i-\bar{y})^2$$

$$\Rightarrow SCF = n_1(y_1-\bar{y})^2 + n_2(\bar{y}_2-\bar{y})^2 + n_3(\bar{y}_3-\bar{y})^2$$

$$\boxed{SCF = 31,592}$$

* $SCR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$

$$= \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

$$\Rightarrow \boxed{SCR = 19,707}$$

* $SCT = SCF + SCR$

$$= 31,592 + 19,707$$

$$= 51,3$$

* Tableau d'analyse de la variance :

Source de variation	Somme des carrés	d.d.l	Carrés moyens	Statistique de Fisher calculée	Statistique de Fisher théorique
facteur	$SCF = 31,592$	2	$CM_F = 15,796$	$F_c = 12,021$	$F_t = F_{5\%}(2,15) = 3,68$
résidu	$SCR = 19,707$	15	$CM_R = 1,314$		
totale	$SCT = 51,3$	17	3,018		

* Teste de Fisher :

* Hypothèses

$H_0 : \mu_1 = \mu_2 = \mu_3$ contre H_1 : il existe au moins un μ_i différent des autres

* Sous H_0 vraie, on a :

$$F_0 = \frac{SCF / k-1}{SCR / n-k} \sim F_\alpha (k-1, n-k) = F_{5\%}(2,15)$$

* Règle de décision

si $F_c \leq F_t$ alors H_0 est vraie sinon H_1 est vraie

Conclusion :

on a $F_c = 12,021 > F_t = 3,68$

Donc H_1 est vraie $\Rightarrow$ oui, la hauteur moyenne des arbres diffère selon la forêt.

Question 2 ex chapitre 2 :

Test d'hypothèse jointes

* $$\frac{\hat{\beta}_1 - 2\hat{\beta}_2}{\sqrt{\hat{V}(\hat{\beta}_1 - 2\hat{\beta}_2)}} \sim st(T-K) \simeq st(20)$$

$$\Rightarrow t^{T-K}_{\frac{\alpha}{2}} = t^{20}_{0,025} = 2,086$$

. $$\hat{V}(\hat{\beta}_1 - 2\hat{\beta}_2) = \hat{V}(\hat{\beta}_1) + 4\hat{V}(\hat{\beta}_2) - 4\,\hat{cov}(\hat{\beta}_1, \hat{\beta}_2)$$

$=$

$$\hat{\Omega}_{\hat{\beta}} = \hat{Var}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

matrice $\hat{\Omega}_\beta$

et $\hat{\sigma}^2 = \frac{SCR}{T-K}$

et

au

$$R^2 = 1 - \frac{SCR}{SCT} \iff SCR = (1 - R^2) \times SCT$$

$$= (1 - 0{,}937) \times 317{,}46$$

$$= 20$$

$$\Rightarrow \hat{\sigma}^2 = \frac{20}{20} = 1$$

$$\Rightarrow \hat{\Omega}_{\hat{\beta}} = \hat{\sigma}^2 (X'X)^{-1}$$

$$\Rightarrow \hat{\Omega}_{\hat{\beta}} = \begin{pmatrix} \hat{V}(\hat{\beta}_1) & \hat{cov}(\hat{\beta}_1, \hat{\beta}_2) & \hat{cov}(\hat{\beta}_1, \hat{\beta}_0) \\ & V(\hat{\beta}_2) & \hat{cov}(\hat{\beta}_2, \hat{\beta}_0) \\ & & V(\hat{\beta}_0) \end{pmatrix}$$

$$\hat{y}_t = 0{,}51 x_{1t} \mp 0{,}35 x_{2t} + 27{,}3$$

$$\boxed{\begin{aligned} \hat{V}(\hat{\beta}_1) &= 0{,}009 \\ \hat{V}(\hat{\beta}_2) &= 0{,}0025 \\ \hat{Cov}(\hat{\beta}_1, \hat{\beta}_2) &= -0{,}08 \end{aligned}}$$

* Si $t_c = \left| \frac{\hat{\beta}_1 - 2\hat{\beta}_2}{\sqrt{\hat{V}(\hat{\beta}_1 - 2\hat{\beta}_2)}} \right| < t^{T-K}_{\frac{\alpha}{2}}$ alors H_1 est vraie

- sinon, H_1 est vraie $\Rightarrow t^{T-K}_{\frac{\alpha}{2}} = t^{20}_{0{,}025} = 2{,}086$

- $\hat{V}(\hat{\beta}_1 - 2\hat{\beta}_2) = 0{,}33$

$$t_c = \left| \frac{0{,}51 + 2 \times 0{,}35}{\sqrt{0{,}33}} \right| = 2{,}106 > 2{,}086$$

$\Rightarrow$ on accepte $H_1 : \beta_1 \neq 2\beta_2$