People's Democratic Republic Of Algeria
Ministry Of Higher Education And Scientific Research

Centtre Univercitaire Abdelhafid Boussouf Mila

Analyse de variance à un facteur (ANOVA)

Dans les cours précédants, on a vu le test d'écart réduit Z qui permet de comparer deux moyennes théorique $\mu_1; \mu_2$ à partir de deux échantillons, ou pour comparer une moyenne observée et une moyenne théorique . Dans cette partie on va voir un nouveau test dite ANALYSE DE LA VARIANCE (ANOVA).

On va voir deux cas :
La comparaison de moyennes de plusieurs (deux ou plus) populations à partir d'échantillons de ces populations.
la compariason de variances de deux populations à partir de deux échantillons de ces populations.

Condition de validité du test :
Les populations doivent etre normales.
Les populations ayant la meme variances.

Exemple

On se limite à quatre variétés V_1, V_2, V_3, V_4 qui offrent une bonne résistance aux maladies et sont donc fréquement cultivées dans la région étudiée. On concidère des haricots issus de sols comparables et de techniques culturales proche. On prélève des échantillons aléatoires de chacune des variétés et l'on observe les résultats indiqués sur le tableau suivant oú sont mentionnés les diamètres en cm.

Exemple

On donnent : $V_1 = 8.8; 7.1; 3.7; 4.5; 8.3; 9.2; 7.5; 4.9; 5.5; 5.5;$
$7.8; 10; 5.7; 8.1; 5.8; 7.3; 6.0; 8.6; 6.4; 6.8; 7.0$
$V_2 = 6.8; 3.5; 5.5; 6.2; 6.0; 8.0; 6.3; 6.3; 8.0; 7.7; 5.9; 8.2; 7.5$
$5.7; 6.2; 3.0; 7.0; 7.0; 3.5; 7.8; 4.0; 7.5; 4.2; 7.3; 4.3; 5.9;$
$4.4; 5.7; 4.6; 5.8; 4.8; 5.9; 5.0; 5.0; 6.1; 5.1$
$V_3 = 5.2; 6.3; 5.3; 6.4; 5.4; 6.5; 5.5; 6.6; 4.8; 6.7$
$; 5.0; 5.8; 5.3; 5.7; 5.5; 6.5; 5.6; 6.7; 3.2; 3.0; 3.1; 6.1; 6.8; 6.6; 8.6; 6.9; 6.9;$
$8.6; 7.6; 4.8; 5.7; 6.7; 7.7; 7.4; 4.1; 9.9; 8.8; 5.6; 5.9; 4.3; 7.7; 5.4$
$V_4 = 6.1; 6.8; 6.6; 8.6; 6.9; 6.9; 8.6; 7.6; 4.8; 5.7; 6.7; 7.7; 7.4; 4.1;$
$9.9; 8.8; 5.6; 5.9; 4.3; 7.7; 5.4; 6.0; 9.0; 8.0; 6.0; 5.0; 6.0; 10; 6.2; 8.0; 8.6; 6.4; 8.2$

On sait qu'une étude préalable a permis d'accepter l'hypothèse de la normalité ainsi que l'hypothèse de l'égalité des variances des variables aléatoires (diamètre des haricots verts) pour les quatre variétés.
Peut-on concidérer qu'en moyenne les quatre variétés ont le meme diamètre ?

Exemple

Notation générale.
$E_i (i = 1, 2, 3, 4)$ est un échantillon aléatoire de la population V_i de moyenne μ_i.
Le nombre d'échantillons E_1, E_2, E_3 et E_4 est $k = 4$.
La taille de l'échantillon $E_i (i = 1, 2, 3, 4)$ est noté N_i :

$$N_1 = 21, N_2 = 35, N_3 = 42, N_4 = 33$$

Exemple

x_{ij} est l'observation numéro j de l'échantillon numéro i, par exemple

$$\left| \begin{array}{l} x_{11} = 8.8, x_{14} = 4.5, x_{121} = 7.0, x_{26} = 8.0 \\ x_{210} = 7.7, x_{32} = 6.3, x_{35} = 5.4, x_{45} = 5.0 \end{array} \right|$$

$\bar{X}_i (i = 1, 2, 3, 4)$ est la moyenne de l'échantillon E_i :

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}.$$

$\bar{X}$ est la moyenne générale observée sur l'ensemble des observations de tous les échantillons :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{4} \sum_{j=1}^{N_i} x_{ij} = \frac{1}{N} \sum_{j=1}^{4} N_i \bar{x}_i$$

Exemple

La somme des carrés totale ou dispersion totale, notée SCE_t, est définie par

$$SCE_t = \sum_{j=1}^{4} \sum_{j=1}^{N_i} (x_{ij} - \bar{x})^2 .$$

La somme des carrés résiduelle ou dispersion intra-échantillons, noté SCE_T, est définie par

$$SCE_r = \sum_{j=1}^{4} \sum_{j=1}^{N_i} (x_{ij} - \bar{x}_i)^2 .$$

La somme des carrés factorielle ou dispersion inter- échantillons, noté SCE_f, est définie par

$$SCE_{fa} = \sum_{j=1}^{4} N_i (\bar{x}_i - \bar{x})^2 .$$

Exemple

x_{ij} est l'observation numéro j de l'échantillon numéro i, par exemple

$$\left| \begin{array}{l} x_{11} = 8.8, x_{14} = 4.5, x_{121} = 7.0, x_{26} = 8.0 \\ x_{210} = 7.7, x_{32} = 6.3, x_{35} = 5.4, x_{45} = 5.0 \end{array} \right|$$

$\bar{X}_i (i = 1, 2, 3, 4)$ est la moyenne de l'échantillon E_i :

$$\bar{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}.$$

$\bar{X}$ est la moyenne générale observée sur l'ensemble des observations de tous les échantillons :

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{4} \sum_{j=1}^{N_i} x_{ij} = \frac{1}{N} \sum_{j=1}^{4} N_i \bar{x}_i$$

Exemple

Remarque On peut montrer les formules suivantes, qui sont mieux adaptées pour les calculs

$$SCE_t = \sum_{j=1}^{4} \sum_{j=1}^{N_i} (X_{ij})^2 - \frac{X_{\bullet\bullet}^2}{N}$$

$$SCE_{fa} = \sum_{j=1}^{4} \frac{X_{ij}^2}{N_i} - \frac{X_{\bullet\bullet}^2}{N}$$

$$SCE_r = \sum_{j=1}^{4} \sum_{j=1}^{N_i} (x_{ij})^2 - \sum_{j=1}^{4} \frac{X_{i\bullet}^2}{N_i}$$

On rappelle les notations

$$X_{\bullet\bullet} = \sum_{i=1}^{4} \sum_{j=1}^{N_i} X_{ij} \text{ et } X_{i\bullet} = \sum_{j=1}^{N_i} X_{ij}$$

La variance interclasse ou carré moyen factoriel, notée $CM_{\mathrm{fa},}$, est définie par

$$CM_{fa} = \frac{SCE_{fa}}{k-1} = \frac{SCE_{fa}}{3}.$$

La variance intra-classe ou carré moyen residuel, notée CM_r, est définie par

$$CM_T = \frac{SCE_T}{N-k} = \frac{SCE_r}{127}.$$

Equation de l'analyse de la variance On peut montrer l'équation suivante, appelée équation de la variance.

$SCE_t = SCE_r + SCE_{fa}$

Variabilité totale $=$ Variabilité résiduelle (intra-échantillons) $+$ variabilité factorielle (inter-échantillons)

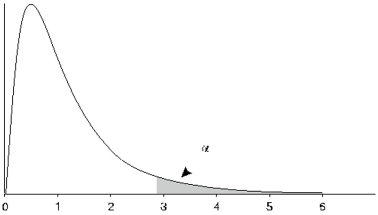
Les étapes du test. 1)- Hypothèse nulle $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$, l'hypothèse alternative H_1 : au moins deux moyennes ne sont pas égales. ici $\mu_i (i = 1, 2, 3, 4)$ est le diamètre moyen de la variété i. 2) Le seuil critique est à calculer sur la loi de Fisher $F_{3,127}$ de ddl (v_1, v_2) où $v_1 = k - 1 = 3$ et $v_2 = N - k = 127$. Ce seuil est le quantile $F_{3,127,0.95}$ d'ordre $1 - 0.05 = 0.95$:

Exemple

Dans la table de Fisher (table 5), on a pas la valeur exacte de $F_{3,127,0.95}$, mais on a $F_{3,100,0.95} = 2.70$ et $F_{3,200,0.95} = 2.65$. On prend la moyenne :

$$\mathbf{F}_{3,127,0.95} = \frac{F_{3,100,0.95} + F_{3,200,0.95}}{2} = \frac{2.70 + 2.65}{2} = \mathbf{2.675}$$

Table de Fisher Snédécor



$$P\left(F_{\vartheta_1,\vartheta_2} \geq c\right) = 0.05$$

	ν_2											
ν_1	**18**	**20**	**22**	**24**	**26**	**28**	**30**	**40**	**50**	**60**	**100**	**200**
1	4.41	4,35	4.30	4,26	4.23	4.20	4.17	4,08	4.03	4,00	3.94	3,89
2	3,55	3,49	3,44	3,40	3,37	3,34	3,32	3,23	3,18	3,15	3.09	3,04
3	3.16	3,10	3,05	3,01	2,98	2,95	2,92	2,84	2,79	2,76	2.70	2,65
4	2.93	2,87	2.82	2,78	2.74	2.71	2.69	2,61	2.56	2,53	2.46	2,42
5	2.77	2,71	2.66	2,62	2.59	2.56	2.53	2,45	2.40	2,37	2.31	2,26
6	2,66	2,60	2,55	2,51	2.47	2.45	2.42	2,34	2.29	2,25	2.19	2,14
7	2.58	2,51	2.46	2,12	2.39	2.36	2.33	2,25	2.20	2,17	2.10	2,06
8	2,51	2,45	2.40	2,36	2.32	2.29	2.27	2,18	2.13	2,10	2.03	1,98
9	2.46	2,39	2.34	2,30	2.27	2.24	2.21	2,12	2.07	2,04	1.97	1,93
10	2.41	2,35	2.30	2,25	2.22	2.19	2.16	2,08	2.03	1,99	1.93	1,88

FIGURE –

Exemple

3)- La statistique de test observée T_O est donnée par la formule

$$T_O = \frac{CM_{fa}}{CM_r} = \frac{17.53}{2.17} = \mathbf{8.08}$$

4)- Décision. Puisque

$$T_O = \mathbf{8.08} > F_{3,127,0.99} = \mathbf{2.675},$$

on rejette l'hypothèse nulle H_0
et on accepte l'hypothèse alternative H_1 : au moins les diamètres moyens de deux variétés ne sont pas égaux.

Ce test concerne la comparaison des variances σ_1 et σ_2 de deux populations. La condition nécessaire pour réaliser ce test est la normalité des deux populations. Présentation du test On se propose de comparer les variances σ_1 d'une population X_1 et σ_2 d'une population X_2 en utilisant les varianes, S_1^2 d'un échantillon aléatoire de X et S_2^2 d'un échantillon aléatoire de X_2. Les deux échantillons de X_1 et X_2 sont indépendants, de tailles n_1 et n_2 respectivement. Le test qu'on va présenter est valide sous la condition de normalité des lois des deux populations X_1 et X_2. Etapes du test. 1. Hypothèse nulle $H_0 : \sigma_1 = \sigma_2$ contre l'hypothèse alternative H_1 : a. $\sigma_1 \neq \sigma_2$ (pour un test bilatéral), b. $\sigma_1 > \sigma_2$ (pour un test unilatéral à droite) ou c. $\sigma_1 < \sigma_2$ (pour un test unilatéral à gauche). 2. La statistique de test observée T_O est égale au rapport des variances des échantillons :

$$T_O = \frac{S_1^2}{S_2^2}$$

Test de comparaison de deux variances de Fisher-Snedecor

3. La région d'acceptation de l'hypothèse nulle. a. Test bilatéral : la région d'acceptation de H_0 est l'intervalle $[a, b]$, où $a = F_{u,v,\frac{a}{2}}$ est le quantile d'ordre $\frac{\alpha}{2}$ de la loi de Fisher de degrés de liberté $(u = n_1 - 1, v = n_2 - 1)$ $b = F_{u,V,1-\frac{a}{2}}$ est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Fisher de degrés de liberté $(u = n_1 - 1, v = n_2 - 1)$.

b. Test unilatèral à droite : la région d'acceptation de H_0 est l'intervalle $[-\infty, d]$, où $d = F_{u,v,1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher de degrés de liberté $(u = n_1 - 1, v = n_2 - 1)$ c. Test unilatèral à gauche : la région d'acceptation de H_0 est l'intervalle $[e, +\infty]$, où $e = F_{u,v,\alpha}$ est le quantile d'ordre α de la loi de Fisher de degrés de liberté
$(u = n_1 - 1, v = n_2 - 1)$.
Utilisation de la table 5 de la loi de Fisher Les tables de Fisher-Snédécor donne les valeurs des quantiles de la loi de Fisher pour $1 - \alpha(\alpha < 0.5)$.

pour calculer des quantiles d'ordre inférieur à 0.5 on utilise la relation suivante. Le quantile d'ordre α de la loi de Fisher-Snédécor de $ddl' = (u, v)$ est égal à l'inverse du quantile d'ordre $1 - \alpha$ de la loi de Fisher-Snédécor de $ddl = (v, u)$.

$$F_{u,v,\alpha} = \frac{1}{F_{v,u,1-\alpha}}.$$

Par exemple

$$\begin{aligned} F_{30,10,0.05} &= \frac{1}{F_{10,30,0.95}} \\ &= \frac{1}{2.16} = 0.46 = \frac{1}{2.16} = 0.46 \end{aligned}$$