Analyse de la variance

ABDELOUAHAB A

Position du problème :
Comparaison de deux moyennes
→ test de l'écart réduit
Comparaison de plusieurs moyennes ?

Solution :
Comparaison deux à deux :
Inconvénients : méthode longue
Vue d'ensemble flou
Analyse de la variance :bon reflet de dispersion autour de la moyenne.
Test approprié : test de Fisher-Snedecor(analyse de la variance)

Principe de solution :
Il repose sur la **décomposition de la variation de la variable** X
1) La variation totale(D totale) de la variable X est mesurée par la dispersion de toutes les valeurs (N valeurs) autour de la moyenne générale
Cette variation dépend de deux sources de variations :

2) La variation entre les classes (variation inter-groupe) (D inter)Chaque classe est caractérisée par sa moyenne ($\overline{x_1}$, ..., $\overline{x_k}$) qui s'écarte plus ou moins de la moyenne générale ($\overline{x}$).
Effet du facteur étudié : variable indépendante

3) La variation à l'intérieur de chaque classe (intra-groupe,résiduelle)(D intra)

4) La relation entre les variations

D totale = D inter + D intra

Principe du test:

On test le rapport de deux variances
variance **inter-groupes**(variance factorielle Vf)
variance **intra-groupes**(variance résiduelle Vr)

Ho teste donc l'hypothèse de l'homogénéité des k moyennes. On dit aussi que le facteur (à partir duquel on a construit les k groupes) n'a pas d'influence sur la variable X
Le rapport suit une loi F de Fisher -Snedecor

1-variance factorielle Vf

$Vf = D_{inter} / k-1$

K: nombre d'échantillons

D_{inter} Dispersion inter-groupes :analyse la variabilité **entre** les groupes

$D_{inter} = n_1 (\overline{X_1}-\overline{X})^2 + n_2(\overline{X_2}-\overline{X})^2 + n_3(\overline{X_3}-\overline{X})^2 + ...+n_k(\overline{X_k}-\overline{X})^2$

formule pratique $D_{inter} = n_1(\overline{X_1})^2+n_2(\overline{X_2})^2+.....-N(\overline{X})^2$

$\overline{X_i}$ = les moyennes respectives des K échantillons

$\overline{X}$ = moyenne générale = $\Sigma Xi/ n1+n2+n3+.....+nk$

N=n1 + n2+ n3+……..+nk

2- variance résiduelle Vr :

$V r = D_{intra} /N-k$

N= somme des tailles de tous les échantillons :

N=n1 + n2+ n3+……..+nk

D_{intra}: Dispersion intra-groupes :analyse la variabilité **à l'intérieur** des groupes

$D_{intra} = \Sigma(Xi_1 -\overline{X_1})^2 + \Sigma(Xi_2-\overline{X_2})^2+ \Sigma(Xi_3-\overline{X_3})^2+....+ \Sigma(Xi_k-\overline{X_k})^2$

Formule pratique :

$D_{intra} = \Sigma X^2i - (n_1.\overline{X}^2_1+n_2.\overline{X}^2_2+....+n_k.\overline{X}^2_k)$

$\Sigma x^2i = \Sigma x^2i1+ \Sigma xi^22+ \Sigma xi^23+.....+ \Sigma xi^2k$

3- la table de Fisher, donne directement F (valeur critique) selon le risque α choisi

↓

l'intersection de

ddl=K-1 ddl=N-K

Procédure

1)Détermination du risque α
2)Formulation des hypothèses H0et H1
3)Calcul de la dispersion inter –groupe
4)Calcul de la dispersion intra –groupe
5)Calcul de la variance factorielle Vf
6)Calcul de la variance résiduelle Vr
7)Conclusion : comparaison du rapport Vf/Vr avec F de la table:

Si Vf/Vr ≥ F la différence est significative au risque α
→ H0est rejetée
Effet de la variable indépendante sur la variable dépendante.
La variation totale du comportement est principalement expliquée par la variable indépendante.

Si Vf/Vr < F la différence n'est significative au risque α
→ H0 est retenue
Pas d'effet de la variable indépendante sur la variable dépendante.

Conditions d'utilisation

1-L'indépendance des échantillons.

2- L'homogénéité des variances (c'est-à-dire égalité des variances). Il convient alors de réaliser un test d'homogénéité :

- test de **Bartlett** si les effectifs des échantillons sont différents ;
- test de **Cochran** si les effectifs sont égaux.

3-Normalité de la distribution des mesures

Le non-respect de ces conditions peut entraîner des résultats erronés soit en acceptant à tort *H*0 soit en rejetant à tort *H*0.

Lorsque les variances sont hétérogènes, l'analyse de variance n'est pas utilisable et on recours à un test de **Kruskal-Wallis**

Exemple :
5 milieux de culture de BCG
10 tubes par milieu de culture
on observe le nombre de colonies par tubes
question :ces milieux sont ils équivalents?

Milieu de culture	A	B	C	D	E
	10	11	7	12	7
	12	18	14	9	6
	8	12	10	11	10
	10	15	11	10	7
	6	13	9	7	7
	13	8	10	8	5
	9	15	9	13	6
	10	16	11	14	7
	8	9	7	10	9
	9	3	9	11	6
Moyenne	**9.5**	**13.0**	**9.7**	**10.5**	**7.0**

1) $\alpha = 5\%$
2)H0 : les milieux sont équivalents
3) Dinter = $n1(\bar{x}1)^2+n2(\bar{x}2)^2+\ldots\ldots-N(\bar{x})^2=10(9.5)^2+10(13)^2+\ldots.-50(9.94)^2=$**185.7**
4)Dintra = $\Sigma x^2i - (n1.\bar{x}^21+n2.\bar{x}^22+\ldots.nk.\bar{x}^2k)=$ $5083-(10 . 9,5^2+10.13^2+\ldots\ldots+10.7^2)=$**225.1**
5)Vf=185.7/5-1=**46.4**
6)Vr=225.1/50-5=**5**
7) conclusion :Vf /Vr=46.4/5=**9.3**
F de la table=**2.5** (ddl_1=4 et ddl_2=45)
le rapport est supérieur à F critique
H0 est rejetée.

Comparaison de deux variances

ABDELOUAHAB A

Intérêt :
1) comparer deux moyennes dans le cas des petits échantillons, le test de Student n'est valide qu'au conditions suivantes:
-la variable étudiée doit être normale
-les variance des deux populations ne doivent pas être significativement différentes
2) comparer la précision de deux méthodes

Principe:
Comparaison du rapport $F=S^2_A/S^2_B$ (S^2_A est supérieure à S^2_B) et F de la table de Fisher.
F critique: intersection de ddl=n_A-1 et ddl=n_B-1

Exemple:

	A	B
	37	38
	36	39
	37	38
	38	37
	37	36
		39
		39
		38
		37
Moyenne	**37**	**38**
Variance	**0.5**	**1.11**

F calculé=1.11 /0.5=**2.22**
F table=**6.04**(intersection de 9-1 et 5-1)
Fc < Ft pas de différence significative entre les deux variances

Procédure:
Les étapes à suivre sont:
1) Détermination du risque α
2) construction de l'hypothèse nulle
3) Calcul de la variance testée $F=S^2A/S^2B$
4) Détermination de la valeur critique F de la table de Fischer, en fonction du risque α et les degrés de liberté V1=n1-1 et V2=n2-1
5) Conclusion :

Fc≥ Ft la différence est significative au risque α et H0 est rejetée

Fc < Ft la différence n'est pas significative au risque α et H0 est retenue

Exemple 2 :

Trois méthodes pédagogiques sont utilisées sur trois groupes d'adultes de même niveau initial. Après l'enseignement les performances sont les suivantes :

Enseignement	Effectif	Total	Moyenne
livresque	18	198	11
audiovisuel	20	240	12
ordinateur	22	308	14

Sachant que $\sum x^2 = 9967$, peut-on apporter la preuve d'une différence d'efficacité entre ces trois méthodes à p = 0,01 ?

Source de variation	Somme des carrés des écarts	Nombre de ddl	Carrés Moyens (variances)	F
entre les groupes (Inter)	94.73	3 - 1 = 2	47.36	4.52
à l'intérieur des groupes (intra)	597.00	60 - 3 = 57	10.47	
Total	691.73	59		

Lecture du F de Snédécor : k - 1 = 2 et N - k = 57

pour p = 0,01 F lu = 5,18 F calculé < F lu

donc Acceptation de Ho

La preuve d'une différence ne peut être apportée (à p = 0,01).

L'étude de la distribution du taux de cholestérol dans deux groupes de sujets adultes adonné les résultats en g/l:

	Taille	variance
Echantillon1	50	0.90
Echantillon2	62	0.79

Tester l'égalité des deux variances:

F calculé=1.15

F de la table=1.60(5%, 49, 61)