REVUE DE STATISTIQUE APPLIQUÉE

J. LELLOUCH

Quelques aspects du problème des comparaisons multiples

Revue de statistique appliquée, tome 14, n° 1 (1966), p. 25-37

<http://www.numdam.org/item?id=RSA_1966__14_1_25_0>

QUELQUES ASPECTS DU PROBLÈME DES COMPARAISONS MULTIPLES

J. LELLOUCH *

I - INTRODUCTION

L'homogénéité d'un groupe de k populations s'étudie habituellement par un test global (χ^2, F...) qui, si son résultat est significatif, indique seulement que l'ensemble est hétérogène sans aucune autre précision. Cependant les problèmes où cette simple conclusion suffit sont rares ; d'où la nécessité d'étudier d'autres situations.

Ces autres situations peuvent être les suivantes :

1°) - L'ordre des comparaisons est imposé par la nature du problème : un exemple en est donné par les enquêtes médicales où disposant de plusieurs groupes témoins et d'un groupe malade, on est logiquement conduit à tester l'homogénéité de l'ensemble des groupes témoins, et si cette homogénéité est acceptée, à comparer le groupe malade a cet ensemble.

2°) - Une des populations joue un rôle particulier. Exemple : Comparaison de plusieurs traitements à un même témoin - un cas particulier de ce problème est celui où :

3°) - On compare plusieurs variances d'effets orthogonaux, à une même variance résiduelle, dans les modèles d'analyse de la variance.

4°) - On essaie de préciser après un test d'ensemble significatif où se trouvent les différences.

Le point 1°) est cité pour mémoire. Il ne sera pas étudié ici. Le point 2°) sera étudié à propos des deux points suivants.

Notre but est de donner une revue générale des méthodes existantes, en renvoyant aux articles originaux pour leur démonstration et leur utilisation pratique.

II - CAS OU PLUSIEURS FACTEURS SONT ETUDIES SIMULTANEMENT DANS UNE MEME EXPERIENCE

Quand on étudie dans une même expérience plusieurs facteurs orthogonaux, on fait des tests F_i qui ne sont pas indépendants, puisqu'ils ont tous la variance résiduelle comme dénominateur commun. On montre [24] que le coefficient de corrélation entre $F_\nu^{\nu_1}$ et $F_\nu^{\nu_2}$ (les numérateurs

---

* Unité de Recherches Statistiques de l'Institut National de la Santé et de la Recherche Médicale.

étant indépendants) est pour $\nu \geqslant 5$.

$$\rho = \frac{1}{\sqrt{\left(1 + \frac{\nu - 2}{\nu_1}\right)\left(1 + \frac{\nu - 2}{\nu_2}\right)}}$$

qui tend très vite vers O quand le nombre de d. d. l. ν de la résiduelle augmente.

Cependant même si ρ est nul, se pose un problème. Si on a par exemple l'habitude de travailler au seuil 5 %, doit-on faire chaque test au seuil α = 5 %, ou choisir des α tels que le risque de trouver au moins une différence significative sous toutes les hypothèses nulles soit égal à 5 % ? Autrement dit, doit-on contrôler le pourcentage de conclusions erronées, ou le pourcentage des expériences où une ou plusieurs conclusions sont erronées ? La première voie semble universellement adoptée : elle ne paraîtra discutable que quand on rencontre des interactions d'ordre élevé, souvent difficiles à interpréter. Même si les F sont correlés, chaque test doit être fait au seuil de 5 % ; le pourcentage de différences nulles proclamées significatives sera alors exactement 5 %.

Si on adopte la deuxième voie on doit trouver des f_i tels que Pr $\{F_i \leqslant f_i$ pour tout $i\} = 1 - \alpha$, ce qui peut se faire en étudiant la distribution de $\frac{\chi^2_{max}}{\chi^2_{res}}$ du moins si le nombre de d. d. l. des numérateurs des F_i sont les mêmes. [15]

On peut également étudier Pr $\{F_i \geqslant f_i$ pour tout $i\}$ par exemple pour vérifier que le modèle d'analyse de variance que l'on a choisi est correct, quand on a trouvé des variances d'effets plus petites que la variance résiduelle. [17]

Ces derniers tests sont aussi utiles pour comparer du point de vue de la variance, plusieurs traitements à un même témoin.

III - COMPARAISONS MULTIPLES ENTRE PLUSIEURS POPULATIONS

Il s'agit de préciser les sous-ensembles qui sont hétérogènes. Nous étudierons particulièrement le cas des moyennes de plusieurs lois normales ; mais dirons également un mot des pourcentages et des variances.

Le traitement du problème dépend du type d'erreur que l'on veut contrôler.

- si on veut contrôler le pourcentage de conclusions erronées, on comparera 2 à 2 toutes les populations par le test t habituel en prenant le seuil α = 5 % pour toutes les comparaisons. On a ainsi la méthode qualifiée habituellement de "t - multiple".

- si on veut contrôler le pourcentage des expériences où une ou plusieurs erreurs de première espèce sont commises, il faut auparavant définir les hypothèses nulles que l'on peut rencontrer au cours d'une expérience.

Or ce nombre croît très vite avec le nombre de populations à comparer. Ainsi si on a 5 moyennes à comparer 2 à 2, soit 10 comparaisons au total, il y a :

10	H_0 du type	H_{12}	$\mu_1 = \mu_2$
10		H_{123}	$\mu_1 = \mu_2 = \mu_3$
5		H_{1234}	$\mu_1 = \mu_2 = \mu_3 = \mu_4$
1		H_{12345}	$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
10		$H_{12,345}$	$\mu_1 = \mu_2 ; \mu_3 = \mu_4 = \mu_5$
30		$H_{12,34}$	$\mu_1 = \mu_2 ; \mu_3 = \mu_4$

à chaque H_0 correspond un certain nombre de conclusions erronées possibles : (par exemple sous H_{123} on peut commettre 1, 2, ou 3 erreurs).

On peut ensuite définir, avec DUNCAN [2], l'erreur de première espèce associée à chacune des H_0 ci-dessus ; ainsi à H_{12} on associe la probabilité

γ_{12} = Pr {de conclure $\mu_1 \neq \mu_2 / \mu_1 = \mu_2$}. Cette fonction pouvant dépendre comme on le verra plus loin des vraies valeurs de $\mu_1 = \mu_2$; μ_3 , μ_4 μ_5 on définit

$\bar{\gamma}_{12}$ = Pr_{max} (de conclure $\mu_1 \neq \mu_2 / \mu_1 = \mu_2$) le maximum étant pris sur tous les μ. De même $\bar{\gamma}_{123}$ = Pr_{max} {conclure $\mu_1 \neq \mu_2$, ou $\mu_1 \neq \mu_3$, ou $\mu_2 \neq \mu_3 / \mu_1 = \mu_2 = \mu_3$} .

A - Méthode contrôlant le pourcentage de conclusions erronées (t-multiple)

Si on a à comparer 5 moyennes de même variance supposée connue c'est-à-dire basée sur un nombre infini de d. d. l., et égale, disons à 1, on conclura que μ_i diffère de μ_j (au seuil 5 %) si

$$\frac{|m_i - m_j|}{\sqrt{2}} \geqslant 1,96 \text{ soit } |m - m| \geqslant 2,77$$

Avec cette règle $\gamma_{12} = 0,05$

$$\gamma_{123} = \text{Prob. } \{\text{de conclure à au moins une différence } / H_{123}\} = \Pr \{R_3 > 2,77\} = 12,2\ \%$$

où R_k désigne le "range" d'un échantillon de k observations.

de même

$$\gamma_{1234} = \Pr \{R_4 > 2,77\} = 20,3\ \%$$

et

$$\gamma_{12345} = \Pr \{R_5 > 2,77\} \# 25\ \%$$

Quant à $\gamma_{12,34}$ il vaut, puisque les comparaisons (12) et (34) sont indépendantes

$$\gamma_{12,34} = 1 - \overline{0,95^2} = 9,75\ \%$$

et

$$\gamma_{12,345} = 1 - 0,95 \times 0,878 = 17,59\ \%$$

Les probabilités des erreurs définies ci-dessus croissent donc très vite avec le nombre k de moyennes à comparer.

B - Procédés contrôlant le pourcentage des expériences où une conclusion au moins est erronée

On peut utiliser différents critères pour classer ces procédés.

Certains sont basés sur des sommes de carrés, d'autres sur le "range" des observations ; d'autres sont non paramétriques.

Les uns sont déterminés à l'avance, alors que les autres se font par étapes, c'est-à-dire que la décision de poursuivre les calculs dépend des résultats des calculs antérieurs.

Le tableau ci-après donne une classification de certains des tests disponibles.

	Somme des carrés	Range	non paramétriques
Test d'ensemble	F	T	KRUSKAL et WALLIS
Procédés déterminés d'avance	méthode S de SCHEFFE	méthode T de TUKEY	STEEL
Cas particulier : Cf à un même témoin	DUNNET	-	STEEL
Procédés par étapes	NEWMAN - KEULS	NEWMAN - KEULS	NEWMAN-KEULS-STEEL

1) Méthode S de SCHEFFE [19]

a) Elle est basée sur le résultat très général suivant :

Etant donnés le modèle d'analyse de la variance $y = X\beta + \varepsilon$, où X est une matrice connue, β le vecteur des paramètres inconnus, $\varepsilon = N(O, \sigma^2)$, et un ensemble de q fonctions linéaires estimables φ linéairement indépendantes, il y a une probabilité $1 - \alpha$ pour que, pour les q fonctions φ et toutes leurs combinaisons linéaires Ψ, on ait

$$\hat{\Psi} - Ss\hat{} < \Psi < \hat{\Psi} + Ss\hat{}_{\psi}$$

avec $\hat{\Psi}$ estimation des moindres carrés de Ψ, $s^2_{\hat{\psi}}$ estimation de la variance de $\hat{\Psi}$ et $S^2 = q\ {}_{\alpha}F^{q}_{\nu}$, γ désignant le nombre de d.d.l. de la variance résiduelle, et ${}_{\alpha}F^{q}_{\nu}$, la limite supérieure au risque α d'une variable F avec q et ν d.d.l.

On conclut donc que $\hat{\Psi}$ est différente de O si $\frac{|\hat{\Psi}|}{s\hat{}_{\hat{\psi}}} > S$. Et si toutes les φ (donc Ψ) sont nulles on n'a qu'un risque α d'émettre une (ou plus) conclusion erronée.

Dans le cas où on a à comparer 2 à 2, un ensemble de k moyennes, le nombre de formes linéaires $\mu_i - \mu_j$ indépendantes est $k - 1$ et $S^2 = (k - 1).\ {}_{\alpha}F^{k-1}_{n-k}$ (n : somme des effectifs des populations).

Dans l'exemple des 5 moyennes de variance unité on conclut à la différence $\mu_i - \mu_j$ (au seuil 0.05) si

$$\frac{m_i - m_j}{\sqrt{2}} > \sqrt{4\ F^4_{\infty}} = \sqrt{\chi^2_4} = 3,08$$

donc si

$$|m_i - m_j| \geqslant 4,36$$

Si la différence $m_i - m_j$ n'est pas significative on dira que le couple (ij) est homogène.

Mais la méthode S permet d'aller plus loin et de considérer non seulement les couples, mais également tous les sous-ensembles de moyennes. En effet, elle est directement liée au test F par la relation suivante : si F est significatif, c'est-à-dire si on conclut à l'hétérogénéité des k moyennes, alors une au moins des fonctions linéaires des $\mu_i - \mu_j$, (c'est-à-dire en fait un "contraste", fonction linéaire $\Sigma c_i \mu_i$ telle que $\Sigma c_i = 0$) est significativement différente de 0 par la méthode S. Réciproquement si un contraste au moins est non nul au seuil α, alors F est significativement trop grand au même seuil. Il y a donc équivalence entre l'existence d'un contraste non nul et l'hétérogénéité globale des k populations.

Si on considère maintenant un sous-ensemble E de moyennes, il sera considéré comme homogène si tous les contrastes de ces moyennes sont nuls par la méthode S, et hétérogène dans le cas contraire. GABRIEL [7] montre que pour tester si tous les contrastes d'un sous-ensemble sont nuls, donc le sous-ensemble homogène, il suffit de comparer S_E^2 à $s^2(k-1)\ {}_\alpha F^{k-1}_{n-k}$ où $s_E^2 = \sum_{i \in E} n_i (\bar{x}_i - \bar{x}_E)^2$ désigne la "somme des carrés entre moyennes" et s^2 la variance résiduelle.

Le procédé est transitif en ce que si E est homogène et si $E' \subset E$ alors E' est homogène et inversement si E est hétérogène et si $E' \supset E$, E' est hétérogène. Cependant E peut être hétérogène alors que tous ses sous-ensembles sont homogènes, et a fortiori, E peut être hétérogène alors qu'aucune différence $m_i - m_j$ n'est différente de 0.

Exemple : Soient les 5 moyennes observées $m_1 = 0$, $m_2 = 3$, $m_3 = 6$, $m_4 = 7$, $m_5 = 9$ de variance unité.

Il y a $2^k - k - 1 = 26$ sous-ensembles, un sous-ensemble E est considéré comme hétérogène au seuil 5 % si $S_E^2 > \chi_4^2 = 9,49$. Après calculs on trouve que sont homogènes les ensembles

(234), (345), (12) et leurs sous-ensembles.

Les différences 2 à 2 significatives sont : (13), (14), (15) et (25)

b) Evaluation des risques de première espèce.

On peut les définir de deux façons : d'abord comme nous l'avons fait plus haut, par les quantités dénommées γ. Par exemple

$\gamma^s_{12345} = \Pr\{\text{de conclure à au moins une différence } \mu_i - \mu_j / \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5\}$

mais on peut aussi définir les quantités δ, telles que, par exemple

$\delta^s_{12345} = \Pr\{\text{de conclure que l'ensemble 12345 est hétérogène} / \mu_1 = \ldots = \mu_5\}$.

Ces deux probabilités ne sont pas égales d'après la remarque précédente qu'un ensemble peut être hétérogène, sans qu'il existe de différence entre aucun couple de moyennes. Evidemment $\gamma \leqslant \delta$.

Ainsi

$\gamma^s_{12345} = \Pr\{R_5 > 4,36\} \# 2\ \%$ $\quad \gamma^s_{1234} \doteq \Pr\{R_4 > 4,36\} \# 1\ \%$ etc..

δ^s_{12345} vaut par définition 5 %.

Pour calculer δ^s_{1234} par exemple, il suffit de remarquer [7] qu'on prend comme limite de signification de $\frac{S_E^2}{S^2}$, qui est distribué comme $3 F^3_{n-k}$, la valeur $4\ {}_\alpha F^4_{n-k}$ d'où la relation

$3\delta_{1234} F^{3}_{n-k} = 4_{0.05} F^{4}_{n-k}$ soit dans le cas particulier $\delta_{1234} \chi^2_3 = {}_{0.05}\chi^2_4$

et

$$\delta^{S}_{1234} \# 3\ \%\ \text{etc.....}$$

c) La méthode est extrêmement générale puisqu'elle s'applique à toute comparaison de paramètres de modèles d'analyse de la variance mais on voit que, du fait de sa généralité, elle conduit à des intervalles très grands, donc à des tests peu puissants.

d) Elle s'étend immédiatement à la comparaison de pourcentages et plus généralement de variables normales ou asymptotiquement normales. Dans le cas des pourcentages on a le théorème suivant :

soient des pourcentages p_i linéairement indépendants et $\varphi = \Sigma\ c_i\ p_i$, q formes linéaires linéairement indépendantes de ces pourcentages. On a, si les effectifs sont grands,

$$P\left\{\frac{\varphi - \hat{\varphi}}{s_{\hat{\varphi}}} \leqslant S \text{ pour toutes les } \varphi \text{ et leurs combinaisons linéaires}\right\} = 1 - \alpha$$

avec

$$S^2 = q\ {}_{\alpha}F^{q}_{\infty} = {}_{\alpha}\chi^2_q$$

Si on a k pourcentages à comparer 2 à 2, $S^2 = {}_{\alpha}\chi^2_{k-1}$

On peut également passer par l'intermédiaire de arc sin $\sqrt{p}$.

2) Méthode T de TUKEY [voir 19 ou 24]

a) Elle est basée sur le théorème suivant qui est immédiat :

Si $\theta_1\ \theta_2\ \ldots\ \theta_k$ sont des paramètres inconnus dont les estimations $\hat{\theta}_1\ \hat{\theta}_2 \ldots\ \hat{\theta}_k$ suivent des lois normales de moyenne θ_i et de même variance $a^2\sigma^2$, si s^2 est une estimation de σ^2 basée sur ν d. d. l. et indépendante en probabilité des $\hat{\theta}_i$, alors

$P\{\hat{\theta}_i - \hat{\theta}_j - Ts \leqslant \theta_i - \theta_j \leqslant \hat{\theta}_i - \hat{\theta}_j + Ts$, pour les couples (ij) $\} = 1 - \alpha$, avec $T = a_{\alpha} q^{k}_{\nu}$, limite supérieure au risque α du range Studentisé.

Dans la comparaison de k moyennes 2 à 2, on conclura que $\mu_i \neq \mu_j$ si $\frac{|m_i - m_j|}{s_m} > {}_{\alpha}q^{k}_{\nu}$

Ce théorème se généralise pour tous les contrastes:

si $\Psi = \Sigma\ c_i\ \theta_i$ avec $\Sigma c_i = 0$, alors pour toutes les Ψ

$$Pr\left\{\hat{\Psi} - Ts\left(\frac{1}{2}\Sigma\ |c_i|\right) < \Psi < \hat{\Psi} + Ts\left(\frac{1}{2}\Sigma\ |c_i|\right)\right\} = 1 - \alpha.$$

Remarquons que cette méthode peut être utilisée comme test de l'hypothèse $\mu_1 = \mu_2 = \ldots = \mu_k$ au même titre que le test F dont elle ne possède cependant ni les qualités optimales, si ce n'est la non distorsion au sens de NEYMAN-PEARSON [16], ni la robustesse. On dira donc qu'un ensemble de moyennes est hétérogène si son "étendue" est significativement trop grande. Il y a donc avec cette méthode, équivalence entre l'hétérogénéité d'un ensemble et la différence de 2 au moins de ses moyennes. Elle a les mêmes propriétés de transitivité que la méthode, S avec en plus ce dernier point que tout ensemble hétérogène contient au moins un sous-ensemble hétérogène.

Dans l'exemple des 5 moyennes de variance 1 on conclut à la différence $\mu_i - \mu_j$ si

$$|m_i - m_j| > {}_{0.05}q^5_\infty = 3,86$$

On aboutit ainsi à des intervalles plus courts qu'avec la méthode S. Les différences 2 à 2 significatives sont celles obtenues par la méthode S, plus le couple (24). Mais la méthode a théoriquement une limitation importante : les m_i doivent avoir la même variance, c'est-à-dire, en particulier, être calculées sur les mêmes effectifs.

b) Calcul des risques de première espèce.

On a évidemment

$\gamma^T_{12345} = 5\ \%$ $\qquad$ $\gamma^T_{123} = Pr\ [R_3 > 3,86] \neq 2\ \%$

$\gamma^T_{1234} = Pr\ [R_4 > 3,86] \neq 3\ \%$ $\qquad$ $\gamma^T_{12} = Pr\ [R_2 > 3,86] \neq 0,75\ \%$ etc.

qui sont bien tous inférieurs à 5 %. On peut également vérifier qu'ils sont supérieurs aux risques auxquels conduit la méthode S.

c) L'extension à des pourcentages $p_1 \ldots\ldots p_k$ calculés sur k populations de même effectif n est immédiate. Deux pourcentages p_i p_j seront considérés comme différents si

$$\frac{|p_i - p_j|}{\sqrt{\frac{pq}{n}}} > {}_\alpha q^k_\infty$$

où p est estimé sur l'ensemble des k populations qui, dans l'hypothèse nulle, sont homogènes. On peut aussi comparer

$$\frac{|\text{arc sin } \sqrt{p_i} - \text{arc sin } \sqrt{p_j}|}{\sqrt{\frac{1}{4n}}} \text{ à } {}_\alpha q^k_\infty$$

d) Dans le cas où les variances sont inégales et les observations corrélées, KRAMER [10 - 11] propose la méthode suivante :

$\hat{\theta}_i$ et $\hat{\theta}_j$ de variance s_i^2 et s_j^2 seront considérés comme différents si

$$\frac{\hat{\theta}_i - \hat{\theta}_j}{\sqrt{\frac{1}{2}[s_i^2 + s_j^2 - 2\text{cov}\,(\hat{\theta}_i\ \hat{\theta}_j)]}} \geqslant {}_\alpha q^k_\infty$$

Ce procédé se ramène à la méthode T si $\hat{\theta}_i$ et $\hat{\theta}_j$ ont la même variance et sont indépendants. On peut montrer qu'il est conservatif c'est-à-dire que les erreurs de première espèce sont en fait plus petites que celles calculées dans l'hypothèse d'égalité des variances et de non corrélation. Il y a donc perte de puissance, mais dont on peut penser qu'elle est faible si les variances ne sont pas trop différentes et les corrélations pas trop fortes.

3) Procédures de NEWMAN - KEULS [9 - 14]

Elles sont basées sur la règle suivante :

Tout sous-ensemble de moyennes est hétérogène si tous les sous-ensembles qui le contiennent (donc en particulier lui-même) sont hétérogènes ; l'homogénéité d'un sous-ensemble de p moyennes étant testée soit par une somme de carrés à p - 1 d. d. l., soit par le range de p moyennes, les seuils de signification étant tous égaux à α.

3 - 1 La méthode du F multiple [1 - 2]

a) soient les 5 moyennes observées $m_1 < m_2 < m_3 < m_4 < m_5$. On conclut par exemple à la différence de (25) si les sous-ensembles suivants sont hétérogènes :

(12345) par un test ${}_\alpha F^4_\nu$, (1235), (1245), (2345) par les tests ${}_\alpha F^3_\nu$, (125), (235), (245) par les tests ${}_\alpha F^2_\nu$ et enfin (25) par un test ${}_\alpha F^1_\nu$.

De même l'ensemble (123) sera dit hétérogène si les ensembles (12345) (1234) (1235) et (123) sont hétérogènes par des tests F à 4, 3, 3, 2 d.d.l. Bien entendu (123) peut être hétérogène sans qu'aucune des différences 2 à 2 ne soit différente de O.

Dans le cas présent, seuls sont homogènes l'ensemble (345) et ses sous-ensembles. En particulier les différences (12), (13), (14), (15), (23), (24), (25) sont significatives.

b) Calcul des risques de première espèce :

On définit comme pour la méthode S, deux sortes d'erreur : δ, risque de trouver hétérogène un ensemble en fait homogène et γ, risque de trouver au moins une différence de 2 moyennes significative dans un ensemble en fait homogène.

Le point nouveau est que ces erreurs ne sont plus constantes, mais dépendent des vraies valeurs des moyennes qui ne font pas partie du sous-ensemble considéré. Il est intéressant d'en trouver des bornes inférieure et supérieure . [7]

En ce qui concerne les δ, $\delta_{12345} = \alpha$ par définition. Quant à δ_{1234}, il est maximum quand on a une probabilité 1 de tester effectivement l'homogénéité de ce sous-ensemble, ce qui arrive quand μ_5 est extrêmement différent des autres moyennes. Alors δ_{1234} vaut . De même toutes les autres bornes supérieures $\bar{\delta}$ valent exactement α.

Pour ce qui est d'une borne inférieure des δ, il suffit de remarquer que si un sous-ensemble est trouvé significatif par la méthode S il le sera nécessairement par ce procédé. Le risque d'erreur est donc plus grand avec le procédé par étapes, de sorte que $\delta^s \leqslant \delta \leqslant \alpha$.

Les $\bar{\delta}$ sont bien inférieurs au risque choisi sauf ceux de la forme $\bar{\delta}_{12,34}$ qui font intervenir 2 sous-ensembles distincts. C'est ainsi par exemple que si le nombre de d.d.l. de la résiduelle est infini, $\bar{\delta}_{12,34} = 2\alpha - \alpha^2 = 9,75\ \%$.

3 - 2 La méthode du range multiple [9 - 2]

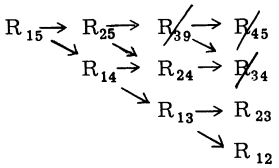
a) Dans le cas des 5 moyennes $m_1 < m_2 < m_3 < m_4 < m_5$, la différence (25) sera significative si les sous-ensembles déjà écrits sont hétérogènes par le test du range, c'est-à-dire si

$R_{15} > {}_\alpha q^5_\infty$; $R_{15} > {}_\alpha q^4_\infty$; $R_{25} > {}_\alpha q^4_\infty$; $R_{15} > {}_\alpha q^3_\infty$; $R_{25} > {}_\alpha q^3_\infty$; $R_{25} > {}_\alpha q^2_\infty$,

conditions qui se réduisent évidemment à

$$R_{15} > {}_\alpha q^5_\infty, \quad R_{25} > {}_\alpha q^4_\infty.$$

Pour trouver pratiquement les couples hétérogènes on fait le schéma ci-dessous. Dès qu'un R est non significatif on le barre ainsi que tous ceux relatifs à des ensembles intérieurs.



En effet, pour 5 %, la table du range (avec un nombre de d.d.l. infini) nous donne :

q^5	q^4	q^3	q^2
3, 86	3, 63	3, 31	2, 77

Le tableau des étendues étant :

	m_5	m_4	m_3	m_2
m_1	9_S	7_S	6_S	3_S
m_2	6_S	4_S	3_S	
m_3	3_{NS}	1_{NS}		
m_4	2_{NS}			

la conclusion est la même qu'avec le F multiple. (Les symboles S et NS signifiant significatif et non significatif au seuil 5 %)

b) Calcul des risques

Le raisonnement fait précédemment s'applique sans aucun changement, de sorte que

$$\gamma^T \leqslant \gamma \leqslant \alpha$$

Tous les γ sont inférieurs au seuil choisi, sauf ceux qui font intervenir des sous-ensembles distincts.

c) L'extension à des variables normales ou asymptotiquement normales se fait sans aucune difficulté si les variances sont les mêmes (pourcentages calculés sur des mêmes effectifs).

d) Cas où les variances sont inégales et les échantillons corrélés.

On pourrait penser opérer exactement de la même façon en utilisant les approximations de KRAMER vues à propos de la méthode T. Cependant si les moyennes extrêmes m_1, m_5 ont une très grande variance, R_{15} ne sera pas trouvé significatif, ce qui conduirait alors à conclure $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$, même si les différences réelles entre ces moyennes sont grandes, et le procédé sera très peu puissant.

Aussi DUNCAN [3] propose-t-il la méthode suivante :

Tout sous-ensemble de p moyennes $m_1\ m_2 \ldots\ m_p$ est homogène si la plus grande des valeurs

$$R_{ij} = \frac{|m_i - m_j|}{\sqrt{\frac{1}{2}\left[s_i^2 + s_j^2 - 2\ \mathrm{cov}\ (m_i\ m_j)\right]}}$$

est inférieure à la valeur critique du range de p moyennes ${}_{\alpha}q_{\nu}^{p}$.

Deux moyennes non contenues dans le même ensemble homogène sont considérées comme différant significativement ; inversement deux moyennes contenues dans le même ensemble homogène ne diffèrent pas significativement. On trouvera un exemple complètement traité en [3].

3 - 3 Considérations sur le choix des risques des tests successifs

On a jusqu'à présent fait chacun des tests successifs au même risque α. Mais ceci n'est évidemment pas une obligation théorique ; et on peut être amené dans certains cas, à choisir des seuils de signification, donc s'il s'agit de tests de range, des $\bar{\gamma}$ différents.

C'est ainsi que pour limiter à 5 % tous les risques de première espèce dans la comparaison de 5 populations on choisira $\bar{\gamma}_{12345} = \bar{\gamma}_{1234} = 5\ \%$; $\bar{\gamma}_{123} = \bar{\gamma}_{12} = 2,5\ \%$.

Dans cette catégorie rentrent les tests de TUKEY (voir [2]) et de DUNCAN [2].

Ce dernier propose de choisir les $\bar{\gamma}$ de telle sorte que

$$\bar{\gamma}_{12\ldots j} = 1 - (1 - \bar{\gamma}_{12})^{j-1}$$

Ainsi si

$$\bar{\gamma}_{12} = 5\ \%,\ \bar{\gamma}_{123} = 9,75\ \% ;\ \bar{\gamma}_{1234} = 14,3\ \% \text{ et } \bar{\gamma}_{12345} = 19,5\ \%$$

La raison de ce choix est la suivante : une comparaison de j moyennes peut se décomposer en j - 1 comparaisons indépendantes, qui conduisent à des tests indépendants si le nombre de d. d. l. est infini. Si pour tester ces j - 1 comparaisons, on faisait j - 1 expériences indépendantes, on n'aurait aucune hésitation à appliquer à chacune un seuil de 5 % ce qui au total donnerait un risque de première espèce de $1 - 0.95^{j-1}$ valeur que l'on choisit pour $\bar{\gamma}_{12\ldots j}$. Quand j devient grand $\bar{\gamma}_{12\ldots j}$ tend vers 100 %, ce qui n'est pas grave si on admet que la probabilité a priori que des moyennes soient toutes égales diminue quand le nombre de ces moyennes augmente. On peut montrer [7] que les bornes inférieures des différents risques sont égales aux valeurs des risques d'un test T de TUKEY, effectué avec un seuil de signification $\alpha = \bar{\gamma}_{12\ldots j}$.

4 - Comparaison des méthodes décrites

Les méthodes basées sur la somme des carrés sont plus générales mais moins puissantes que celles basées sur le range. C'est ainsi que dans l'exemple étudié deux moyennes seront différentes par la méthode S si leur différence est supérieure à 4,36 et seulement 3,86 par la méthode T. Si on ne s'intéresse qu'aux différences 2 à 2 et non à tous les contrastes, les méthodes basées sur le range doivent être préférées.

Les méthodes par étapes sont les plus avantageuses : elles sont en effet plus puissantes. De plus on peut prendre à chaque étape des risques variables et satisfaisant le mieux au problème posé : ces risques peuvent être choisis grands pour un nombre important de moyennes car si on admet, ce qui peut être discutable, que la probabilité a priori d'une totale homogénéité diminue avec le nombre de moyennes, il n'y a pas lieu de se prémunir contre un danger qui n'existe pas.

5 - Autres procédés

Ils sont décrits par TUKEY [23]. Ils supposent que les moyennes sont calculées sur des effectifs égaux.

1) - Test des intervalles successifs (Tukey) : on teste si l'intervalle qui sépare deux moyennes observées successives n'est pas trop grand - s'il est significativement trop grand, les deux moyennes appartiennent à deux sous-groupes différents. La difficulté est de trouver la distribution du plus long intervalle entre des observations rangées par ordre croissant. Dans le cas de k moyennes de même variance, TUKEY propose d'approcher cette distribution par celle du t de la différence de deux variables normales. On a,ce faisant,un test conservatif. Du fait de l'ignorance où on est de la distribution du plus grand intervalle, on ne peut calculer les différentes erreurs de première espèce.

2) - Test de la valeur extrême (Nair - Tukey) [13 - 23] : on compare l'observation qui s'écarte le plus de la moyenne générale, à cette moyenne générale. NAIR a donné des tables de la "déviation extrême studentisée", et TUKEY propose l'approximation suivante :

$$\frac{m - \bar{m}}{s_m} = N\left[\mu = \frac{6}{5}\log_{10} k\ ;\ \sigma = 3\left(\frac{1}{4} + \frac{1}{\nu}\right)\right] \text{ si } k > 3$$

et

$$\frac{m - \bar{m}}{s_m} = N\left[\mu = \frac{1}{2}\ ;\ \sigma = 3\left(\frac{1}{4} + \frac{1}{\nu}\right)\right] \text{ si } k = 3$$

(où ν est le nombre de d. d. l. de la variance résiduelle).

Si $|m - \bar{m}|$ est trop grand, m est considérée comme différente des autres moyennes.

3) - Tests des valeurs extrêmes successifs

Si le procédé précédent sépare une moyenne de l'ensemble on peut le répéter sur les k - 1 moyennes restantes et ainsi de suite. Supposons que sur 12 moyennes on soit arrivé à la disposition suivante, après séparation des moyennes extrêmes de l'ensemble $m_5\ m_6\ m_7\ m_8$.

$$m_1,\ m_2\ ,\ m_3,\ m_4\ (m_5\ m_6\ m_7\ m_8),\ m_9,\ m_{10},\ m_{11}$$

on doit bien entendu tester l'homogénéité de $(m_1\ m_2\ m_3\ m_4)$ et $(m_9,\ m_{10},\ m_{11})$. On leur applique le même procédé et on aboutit à des sous-ensembles disjoints.

4) - Combinaison des procédés précédents

TUKEY propose d'utiliser successivement le procédé 1, puis le procédé 3 et enfin un des tests décrits précédemment à chacun des sous-ensembles disjoints ainsi obtenus.

6 - Comparaisons multiples de variances

On peut appliquer les méthodes qui précèdent à $\log s_i^2$ qui a une distribution pas trop écartée de la normale. [19]

L'analogue du test T de TUKEY pour les comparaisons de variances de lois normales et le test F_{max} de HARTLEY [8] ; si s_i^2 est l'estimation de σ_i^2, σ_i^2 et σ_j^2 seront considérées comme différentes si

$$\frac{\text{Max}\ (s_i^2,\ s_j^2)}{\text{Min}\ (s_i^2,\ s_j^2)} > F_{max}$$

où Max $(s_j^2,\ s_j^2)$ désigne la plus grande des deux quantités s_i^2 et s_j^2.

Considéré comme test de comparaison d'un ensemble de plusieurs variances, ce test est comme la méthode T, sans distorsion. [16]

7 - Comparaison de plusieurs populations $P_1 \ldots\ldots P_k$ à un même standard P_o

Ce problème est un cas particulier du problème général traité jusqu'ici. Les risques de première espèce sont définis de la façon suivante

$$\gamma_i = Pr[\, P_i \neq P_o / P_i = P_o]$$

$$\gamma_{ij} = Pr[\text{ conclure } P_i \neq P_o, \text{ ou } P_j \neq P_o / P_i = P_j = P_o\,] \quad \text{etc...}$$

DUNNET [4 - 6] a donné une solution de ce problème quand il s'agit de moyennes de lois normales. Il donne des tables de t_i telles que

$$Pr\left[\frac{m_o - m_1}{s_{m_o - m_i}} < t_i \;/\; P_o = P_i = \ldots. P_k\right] = 1 - \alpha$$

à la fois pour le test unilatéral et bilatéral.

8 - Méthode non paramétriques

STEEL [20 - 21 - 22] a donné des procédés non paramétriques, basés sur la somme des rangs, identiques dans leur principe à ceux de SCHEFFE, TUKEY, NEWMAN - KEULS et DUNNET.

IV - CONCLUSION GENERALE

Nous nous sommes efforcé de classer et de décrire les différentes méthodes proposées dans la littérature pour trouver les sous-ensembles homogènes d'un ensemble de k moyennes dont un test global a montré l'hétérogènéité. Le grand nombre de ces procédés montre d'emblée qu'aucun n'est entièrement satisfaisant. Le plus grave de leurs inconvénients est qu'ils aboutissent souvent à des conclusions extrêmement difficiles voire impossibles à interpréter concrètement du fait que leur transivité n'est pas parfaite. Dans l'ensemble de trois moyennes (m_1, m_2, m_3), m peut être différent de m_3, et m_2 non différent de m_1 et m_3. Le statisticien sait bien que "non différent" signifie seulement que la différence n'a pas pu être prouvée. Mais une conclusion de ce genre aboutit nécessairement à une erreur de deuxième espèce. Aussi devrait-on essayer de formuler le problème en termes de décision plutôt qu'en termes de test classique où sont contrôlées les erreurs de première espèce.DUNNET [5] a abordé le problème de trouver la plus grande de k moyennes et LAZAR [12] étudie la partition d'un groupe hétérogène, en sous-groupes homogènes et distincts ce qui évite en particulier toutes les incohérences.

REFERENCES

[1] DUNCAN D.B. (1951) - A significance test for differences between ranked treatments in an analysis of variance. Virginia Journal of Science 2, n° 3, 171-189.

[2] DUNCAN D.B. (1955) - Multiple range and multiple F tests. Biometrics 11, 1-42.

[3] DUNCAN D.B. (1957) - Multiple range tests for correlated and heteroscedastic means. Biometrics 13, 164-176.

[4] DUNNET C.W. (1955) - A multiple comparison procedure for comparing several treatments with a control. J. Amer. Statist. Assoc 50, 1096-1121.

[5] DUNNET C.W. (1960) - On selecting the largest of k normal populations means. J.R.S.S. B, 22 140.

[6] DUNNET C.W. (1964) - New tables for multiple comparisons with a control. Biometrics 20. 482-491.

[7] GABRIEL K.R. (1964) - A procédure for testing the homogeneity of all sets of means in analysis of variance. Biometrics 20, 458-477.

[8] HARTLEY H.O. (1950) - The maximum F. ratio as a short cut test for heterogeneity of variance. Biometrika 37, 308-312.

[9] KEULS M. (1952) - The use of the "Studentized range" in connection with an analysis of variance. Euphytica 1, 112-122.

[10] KRAMER C.Y. (1956) - Extension of multiple range tests to group means with unequal numbers of replications. Biometrics 12, 307-310.

[11] KRAMER C.Y. (1957) - Extension of multiple range tests to group correlated ajusted means. Biometrics 13, 13-18.

[12] LAZAR P. (1966) - Partition d'un groupe hétérogène en sous-groupe homogènes. Revue de statistique appliquée (Même numéro).

[13] NAIR K.R. (1948) - The distribution of the extrêmme deviate from the sample mean and its Studentized form. Biometrika 35, 118-144.

[14] NEWMAN D. (1939) - The distribution of the range in samples from the normal population, expressed in terms of an independent estimate of standard deviation. Biometrika 31, 20-30.

[15] RAMACHADRAN K.V. (1956) - On the simultaneous analysis of variance test. Ann. Math. Statist. 27, 521-528.

[16] RAMACHADRAN K.V. (1956) - On the Tukey test for the equality of means and the Hartley test for the equality of variance. Ann. Math. Statist. 27, 825-831.

[17] RAMACHADRAN K.V. (1958) - On the Studendized smallest chi-square. J. Amer. Statist. Assoc. 53, 868-872.

[18] RIVES M. (1959) - Sur la comparaison des moyennes dans les essais varietaux. Ann. de l'Amelioration des plantes, 357-376.

[19] SCHEFFE H. (1959) - The Analysis of Variance - New York - John Wiley.

[20] STEEL R.G.D. (1959) - A multiple comparison rank sum test: Treatments versus control. Biometrics 15, 560-572.

[21] STEEL R.G.D. (1960) - A rank sum test for comparing all pairs of treatments. Technometrics 2, 197-207.

[22] STEEL R.G.D. (1961) - Some rank sum multiple comparisons tests. Biometrics 17, 539-552.

[23] TUKEY J.W. (1949) - Comparing individual means in the analysis of variance. Biometrics 2, 99-114.

[24] ULMO J. (1959) - Etude fondamentale de la regression linéaire multiple. Cours de l'I.S.U.P. - Paris.