Tutorials in Quantitative Methods for Psychology
2012, Vol. 8(2), p. 88-95.
.

DOI: 10.20982/tqmp.08.2.p088

Faut-il *contrôler* l'erreur de type I dans le cas de comparaisons de moyennes multiples?
Must we *over-control* the type I error rate in post anova multiple comparison procedures?

Louis Laurencelle
Université du Québec à Trois-Rivières

Almost since the creation of analysis of variance by Fisher in the years 1920's, interpretation of its results and the multiple comparisons of means it entailed have raised the problem of the type I error rate (α) and its control. Fisher himself, then Tukey and many others have contributed to the question, finally stockpiling a plethora of principles, methods and suggestions, all aimed at keeping the effective α level within prescribed bounds, and all equally attractive to the naïve user. We revisit this controversial question, from the standpoint of the empirical researcher, and propose a severe stripping down of statistical-probabilistic complications, in order to give back to the researcher just what he needs to drive out and appraise the significant results in his data.

Dès la création, par R. A. Fisher vers les années 1920, de cet outil merveilleux qu'est l'analyse de variance (voir Fisher, 1971), le problème des « comparaisons multiples de moyennes » s'est posé. La méthode de Fisher permettait de comparer entre elles, non seulement 2 moyennes, mais 3 moyennes, 4 ou davantage, et donnait une sanction de significativité globale : il y a, ou non, une « variation significative » entre les moyennes. Mais, avec un test global significatif pour 3 moyennes ou plus, comment repérer, et décider, quelle moyenne diffère de quelle autre, et ce, en conservant un certain contrôle sur le risque de déclarations significatives fallacieuses, risque symbolisé par le fameux seuil α? Fisher (op. cit.) lui-même s'est penché sur le problème, puis surtout Tukey (1953) et d'autres, de sorte que la documentation sur le sujet a connu et connaît encore un foisonnement de principes, méthodes et suggestions, chacun prétendant régler à sa manière le problème du contrôle du risque d'erreur dans les comparaisons multiples. Principes probabilistes et méthodes statistiques, donc, entre lesquels le chercheur doit faire son chemin et opter, souvent à la faveur des méthodes que lui offre son logiciel d'analyse coutumier, voire en reprenant la méthode appliquée couramment dans son secteur de recherche. Nous re-soulevons ici la question du contrôle du risque dans les comparaisons multiples, en nous plaçant du point de vue du chercheur qui compare différentes conditions d'expérimentation et veut déterminer quels en sont les effets significatifs sur son phénomène observé, tandis que le statisticien, quant à lui, prend en considération une variable aléatoire et des échantillons au hasard simple et conçoit des principes pour garantir un non-rejet de l'hypothèse nulle au-delà d'un niveau de probabilité prescrit d'avance. Nous offrons notre analyse et nos propositions à titre polémique, dans l'espoir de raviver le débat sur la question et, si possible, de redonner aux chercheurs le contrôle de leurs méthodes de preuve.

Je prendrai comme référence un contexte de recherche expérimentale comparant $k = 5$ groupes de $n = 6$ sujets chacun, d'où un Carré moyen intragoupe ayant $dl = 25$ degrés de liberté. Les groupes sont échantillonnalement équivalents et se distinguent par le traitement spécifique que chacun a reçu. Je recourrai généralement à un seuil de

signification α de 5% (sans préjugé).

Mes arguments principaux sont les suivants :

(a) Si l'hypothèse à vérifier porte sur une structure intégrant les *k* moyennes, alors un test structuré sur ces *k* moyennes s'impose, au seuil de probabilité α. Si ce test n'existe pas tel quel, alors il faut l'inventer ou le simuler par expérimentation Monte Carlo.

(b) Si les données d'expérience ont été produites en vertu de prédictions théoriques précises, alors, pour chaque prédiction couvrant un sous-ensemble de moyennes, un test particulier peut être appliqué, le plus souvent en mode unilatéral, (les tests eux-mêmes n'ayant pas à être statistiquement indépendants), sans considération d'un taux d'erreur collectif.

(c) Si les données obtenues (par expérimentation ou par simple collecte) n'émanent pas d'un protocole à démonstration théorique mais d'une enquête descriptive, d'une étude « pour voir », dotée seulement d'anticipations de résultats plutôt que d'hypothèses proprement dites, alors les tests sont appliqués comme « à la pêche » et leurs conclusions doivent être interprétées comme des suggestions ou des hypothèses à vérifier expérimentalement, sans force de généralisation. Dans ce cas, la rigueur du seuil α doit être proportionnée à la grosseur de la moisson statistique attendue ; l'application d'un contrôle d'erreur global, p. ex. le taux d'erreur par expérience, n'a finalement pour effet net que de réduire, parfois exagérément, le seuil α par comparaison.

Démonstration 1 : Dilution de significativité par addition de groupes dans l'anova

Soit un modèle d'anova à *k* groupes aléatoires de $n = 6$ participants chacun, et un terme d'erreur unitaire ($CM_e = 1$), avec $dl_2 = 25$, et soit un « effet » E attaché au groupe 1, sous la forme $\overline{X}_1 = C + E$, les autres groupes $\overline{X}_i$ étant tous égaux à C.

Prenons d'abord E = 2.

Pour $k = 2$ groupes, $F = CM_G = 12{,}0$. Avec $dl_1 = 1$ et $dl_2 = 10$, $F_{1,10[0,95]} = 4{,}965$ et $F_{1,10[0,99]} = 10{,}04$, le test étant significatif aux deux niveaux.

Pour $k = 3$ groupes, $F = 8$. Avec $dl_1 = 2$ et $dl_2 = 15$, $F_{[0,95]} = 3{,}682$ et $F_{[0,99]} = 6{,}359$, le test étant encore significatif aux deux niveaux.

Pour $k = 10$ groupes, $F = 2{,}4$. Avec $dl_1 = 9$ et $dl_2 = 50$, $F_{[0,95]} = 2{,}073$ et $F_{[0,99]} = 2{,}785$, le test n'est plus significatif au seuil de 1%, tout en le demeurant au seuil de 5%.

Pour $k = 15$ groupes, $F = 1{,}6$. Avec $dl_1 = 14$ et $dl_2 = 75$, $F_{[0,95]} = 1{,}860$ et $F_{[0,99]} = 2{,}394$, le test n'est plus significatif ni à 1%, ni à 5%.

On constate ici que, pour un même effet E affectant le groupe 1, la significativité vient à se perdre, comme en se diluant à travers le nombre croissant de moyennes, et ce, malgré l'accroissement des degrés de liberté de l'erreur (et donc, de la puissance).

Prenons ensuite E = 1,5.

Pour $k = 2$ groupes, $F = CM_G = 6{,}75$. Avec $dl_1 = 1$ et $dl_2 = 10$, $F_{1,10[0,95]} = 4{,}965$ et $F_{1,10[0,99]} = 10{,}04$, le test n'est significatif qu'au seuil de 5%, et il cesse de l'être avec $k = 5$ groupes ($F = 2{,}7$ vs $F_{4,25[0,95]} = 2{,}759$).

Conclusion possible : Ceci démontre que l'addition de groupes en anova a pour effet de répartir la variance significative (attachée à l'effet E) sur un plus grand nombre de degrés de liberté (au numérateur du *F*) et, par ce biais, d'en diluer et éventuellement annuler la significativité. Dans cette perspective, et en supposant que l'effet expérimental ciblé est attaché au groupe 1, il serait logiquement invalide de conditionner les tests de comparaison des moyennes au résultat significatif du *F*, tel que le préconisent certaines procédures, notamment l'approche LSD (Least Signifiant Difference) dite de Fisher.

Démonstration 2 : la logique du Studentised range

Préambule

Parmi *k* moyennes obtenues, basées chacune sur *n* données, on peut repérer la plus petite et la plus grande, que nous dénoterons $\bar{X}_{(1)}$ et $\bar{X}_{(k)}$ respectivement, $\bar{X}_{(j)}$ représentant la j^e plus petite parmi les *k* moyennes. L'écart entre ces deux extrêmes est dénommé *étendue*, d'où, en divisant par une estimation indépendante de l'erreur-type (s_e) basée sur dl_e degrés de liberté, nous obtenons le « Studentised range », ou *étendue studentisée* q[1] :

$$q_{1,k\,:\,k} = n\,(\bar{X}_{(k)} - \bar{X}_{(1)}) \,/\, s_e. \qquad (1)$$

Cette statistique *q*, que nous noterons $q_{1,k}$, fait partie de la famille des écarts entre statistiques d'ordre $q_{i,j\,:\,k}$, évidemment définis par :

$$q_{i,j\,:\,k} = n\,(\bar{X}_{(j)} - \bar{X}_{(i)}) \,/\, s_e \quad \{\, 1 \le i < j \le k \,\}, \qquad (2)$$

dont la distribution générale ne semble pas avoir été tabulée. Soit *w*, l'écart entre les statistiques d'ordre $\bar{X}_{(i)}$ et $\bar{X}_{(j)}$, alors (David 1981) la densité de *w* est donnée par :

$$f(w) = C_{ij} \int_{-\infty}^{\infty} P^{i-1}(x)p(x)\,[P(x+w) - P(x)]^{j-i-1} \times p(x+w)\,[1 - P(x+w)]^{n-j}\,dx$$

[1] Nous avons simplifié la notation en écrivant $X_{(i)}$ pour $X_{(i\,:\,k)}$ afin de désigner la i^e statistique d'ordre dans une série de *k*, de sorte que, p. ex., $X_{(2)} - X_{(1)}$ dénote explicitement $X_{(2\,:\,k)} - X_{(1\,:\,k)}$ et $q_{1,2}$ dénote $q_{1:k,\,2:k}$, que nous résumons aussi par $q_{1,2\,:\,k.}$

Tableau 1. Centiles supérieurs de l'écart studentisé $q_{i,j:5}$

	$i, j =$	1 , 2 4 , 5	2 , 3 3 , 4	1 , 3 3 , 5	2 , 4	1 , 4 2 , 5	1 , 5
ν = 25	P = 0,95	1,854	1,406	2,538	2,171	3,199	4,153
	P = 0,99	2,602	2,001	3,325	2,848	4,048	5,144
ν = ∞	P = 0,95	1,771	1,344	2,397	2,052	2,994	3,858
	P = 0,99	2,422	1,867	3,046	2,610	3,661	4,603

où $C_{ij} = n! / \{ (i-1)!(j-i-1)!(n-j)! \}$, et $P(x)$ et $p(x)$ sont respectivement la fonction de répartition (f.r.) et la densité de la variable visée, ici une variable normale standard; la f.r. de w est simplement :

$$F(W) = \int_0^W f(W)\, dw.$$

Pour des quantités q basées sur un nombre fini de degrés de liberté, il faut convoluer la f.r. ci-dessus avec la densité de l'erreur-type s_e, lequel obéit à la loi du χ, plus précisément $\chi / \sqrt{\nu}$, avec $\nu = dl$ degrés de liberté, soit :

$$H(q) = \int_0^\infty g(s) \times F(q \cdot s)\, ds, \qquad (3)$$

où : $g(s) = 2(\nu/2)^{\nu/2}[\Gamma(\nu/2)]^{-1}s^{\nu-1}\cdot\exp(-\nu s^2/2)$.

Le tableau 1 fournit, à titre illustratif, les valeurs critiques de l'écart studentisé $q_{i,j:k}$, k = 5, pour v (= dl) = 25 et ν = ∞, aux rangs centiles P = 0,95 et 0,99, telles que $H(q) = P$.

Argument et contre-argument sur le critère de l'étendue studentisée

Que le chercheur aille à la pêche ou non dans ses données, c.-à-d. qu'il ait mis sur pied sa collecte sur la base de prédictions théoriquement orientées ou dans le seul but de « voir ce qu'il se passe », il reste que, parmi les comparaisons de moyennes de type « $\bar{X}_{(j)} - \bar{X}_{(i)}$ », celle qui a le plus de chances d'être significative *au test t normal* est celle donnant la plus grande différence, i.e. $\max_{i,j} \{ \bar{X}_{(j)} - \bar{X}_{(i)} \}$, soit celle correspondant à l'étendue, $\bar{X}_{(k)} - \bar{X}_{(1)}$. Puisque l'hypothèse nulle globale affirme que les k moyennes sont issues d'une population ayant même moyenne commune (μ), la différence $\bar{X}_{(k)} - \bar{X}_{(1)}$ devrait se comporter comme l'étendue de k variables normales, et la statistique $q_{1,k}$, obéir à la loi du Studentised range (3) : ceci est la base de la procédure de test de Tukey, dénommée HSD (Honestly Significant Difference). Comme chacune des moyennes (tirées au hasard d'une même population) aurait pu être la plus petite ou la plus grande, la procédure est appliquée à toutes les différences, et la même valeur critique $q_{1,k:k[1-\alpha]}$ sert à juger tous les tests de forme $q_{i,j:k}$.

CEPENDANT, si l'hypothèse nulle globale est vraie, elle est vraie partout et, en particulier, elle l'est pour chacune des différences $\bar{X}_{(j)} - \bar{X}_{(i)}$, $1 \le i < j \le k$. Ainsi, par exemple, la différence entre les deux premières statistiques d'ordre, $\bar{X}_{(2)} - \bar{X}_{(1)}$, se comporte sous H_0 comme $q_{1,2:k}$, de sorte que toute valeur obtenue telle que $q_{1,2:k} > q_{1,2:k[1-\alpha]}$ infirmerait H_0. En d'autres mots, si un facteur de variation systématique est à l'œuvre dans les données de telle façon à imprimer à celles-ci des écarts « plus que normaux », ce facteur peut affecter la plus petite différence observée aussi bien, sinon plus, que la plus grande différence, ce qui devrait se refléter aussi dans l'écart $\bar{X}_{(2)} - \bar{X}_{(1)}$ et pouvoir être avéré par le test avec $q_{1,2:k[1-\alpha]}$.

Tant qu'à faire, la plus petite différence de moyennes parmi k ($k \ge 3$) est forcément plus petite qu'aucune des k–1 différences de type $\bar{X}_{(i+1)} - \bar{X}_{(i)}$, et elle obéit à une autre loi de distribution. Nous avons mis sur pied un modèle Monte Carlo pour simuler cette loi et en estimer les centiles 95 et 99 : le tableau 2 en présente certaines valeurs approchées de cette statistique, baptisée $q_{\min,k}$. Rappelons que, pour k = 2 moyennes, la valeur critique présentée est essentiellement celle du t de Student. Or, le tableau 2 rend tout de suite apparent que cette valeur classique, $t_{k=2}$, est beaucoup *trop exigeante* vis-à-vis d'une statistique qui serait basée sur la différence minimale parmi 3 moyennes ou davantage. On constate aussi que, si le chercheur veut déterminer si ses k traitements ont réussi à élargir les différences entre des groupes constitués sur un mode échantillonnal équivalent, le test de la *différence minimale* est un argument au moins aussi convaincant, voire davantage, que n'est le test de la différence la plus grande, c.-à-d. le HSD de Tukey.

En résumé, dans une perspective de comparaisons a posteriori entre les moyennes, si on admet que le test de différence entre les moyennes les plus écartées parmi k viole les conditions d'application du simple t de Student et doit être rapporté au Studentised range ($q_{1,k:k}$), alors, la même admission doit être formulée pour le test de différence entre les deux moyennes les plus proches, lequel serait rapporté aux valeurs critiques telles qu'au tableau 2, ou encore, pour le test de différence entre des moyennes de rangs respectifs i et j (parmi k), lequel se rapporterait aux valeurs critiques

Tableau 2. Centiles supérieurs approximatifs de l'écart minimal parmi k moyennes, $q_{\min,k}$

	k =	2	3	4	5	7	10
$\nu = 25$	P = 0,95	2,913	1,25	0,701	0,449	0,229	0,112
	P = 0,99	3,914	1,74	0,996	0,650	0,337	0,168
$\nu = \infty$	P = 0,95	2,771	1,19	0,6761	0,430	0,219	0,107
	P = 0,99	3,643	1,62	0,933	0,606	0,316	0,158

Note : Le centile P pour $k = 2$ correspond au $t_{\nu[P]} \times \sqrt{2}$

telles qu'au tableau 1. La préséance logique du critère HSD de Tukey est donc toute relative et, vu son important impact négatif sur la puissance des comparaisons, son contrôle astringent de l'erreur de type I mérite d'être reconsidéré.

Démonstration 3 : Le contrôle des taux d'erreur

Dès que Tukey (1949, 1953) en eût jeté les bases, le « problème des comparaisons multiples » en analyse de variance s'est décliné en termes du contrôle du taux d'erreur, en fait l'erreur de type I consistant à rejeter à tort l'hypothèse de non-différence, et de la base de généralisation de ce taux À l'instar d'autres auteurs, et selon une terminologie variable, on distingue couramment :

α_C, le taux d'erreur par comparaison ;

α_E, le risque d'erreur par expérience (*experimentwise*) ou par famille de comparaisons (*familywise*) ;

τ_E, le taux d'erreur par expérience (*per experiment*).

Un court exemple illustrera la nature et les différences de ces différents indicateurs. Reprenons notre expérience dans laquelle $k = 5$ groupes sont comparés, via leurs moyennes ; les *dl* d'erreur correspondants seraient 25. En vue de comparer les groupes 1 et 2, le test de la différence $(\bar{X}_1 - \bar{X}_2)$ utilise un quotient $t_{dl[\alpha]}$ classique, au seuil α (disons, bilatéral) = 0,05 ; la valeur critique sera ici $t_{25[0,975]} = \pm 2{,}060$. Ce test encourt un taux d'erreur par comparaison de $\alpha_C = 0{,}05$, c.-à-d. que, si H_0 est vraie, il y a une probabilité de 0,05 que le test nous conduise à tort à l'infirmer. En théorie, cela signifierait que, sous 100 expériences indépendantes dans lesquelles H_0 est vraie, nous obtiendrions en moyenne 5 sanctions de significativité fausses *pour cette comparaison*. Cependant, l'expérience décrite comporte $k = 5$ moyennes, entraînant la formation possible de $nc = {}_5C_2 = 10$ comparaisons deux à deux. Or, si nulle variation systématique n'a pu être produite dans l'expérience invoquée et que l'hypothèse nulle globale est vraie, toutes les différences constatées seront imputables au hasard. Si, pour chacune des nc comparaisons possibles, il y a une probabilité α_C qu'elle soit faussement déclarée significative, la probabilité de rejeter faussement l'hypothèse nulle globale, c.-à-d. d'affirmer qu'il s'est produit une variation systématique dans l'expérience, repose quant à elle sur l'ensemble des nc comparaisons, et elle est (approximativement[2]) :

$$\alpha_E = 1 - (1 - \alpha_C)^{nc}\ , \qquad (4)$$

soit, pour $\alpha_C = 0{,}05$ et $nc = 10$, $\alpha_E \approx 0{,}40$. Il y aurait donc, ici, un risque de 40% à l'effet que l'hypothèse nulle globale soit rejetée, ce sur la base d'une ou de quelques différences pouvant être trouvées significatives parmi les 10 possibles. Enfin, comme nous avons ici $nc = 10$ comparaisons, chacune étant testée au seuil α_C de 0,05, le taux d'erreur par expérience sera simplement $\tau_E = nc \times \alpha_C = 0{,}50$, indiquant le nombre moyen de fausses déclarations attendu pour un ensemble de comparaisons donné.

Les auteurs, surtout des statisticiens (plutôt que des chercheurs), se sont intéressés à ces différents taux d'erreur, et plusieurs ont mis leur dévolu sur le risque d'erreur par expérience (α_E) et son contrôle. En effet, en première approximation, nous voyons que $\alpha_E \approx \tau_E = nc \times \alpha_C$, augmentant très rapidement vers 1 : utilisant $\alpha_C = 0{,}05$, nous calculons p. ex. $\alpha_E \approx 0{,}51$ pour $nc = 14$ (environ $k = 4$ moyennes), 0,90 pour $nc = 45$ (5 moyennes), 0,995 pour $nc = 105$ (6 moyennes).

Éprouvant une certaine panique intellectuelle devant cette quasi certitude de fausse déclaration *globale*, les auteurs ont proposé de contrôler primordialement ce risque d'erreur par expérience, en le plafonnant à une valeur

[2] L'approximation tient au fait que les comparaisons ne sont pas statistiquement indépendantes, puisqu'elles utilisent les mêmes termes. P. ex. pour $k = 4$ moyennes, les comparaisons 1 vs 2 et 1 vs 3 sont liées (par le recours commun à 1), engendrant une corrélation de ½, tandis que 1 vs 2 avec 3 vs 4 sont non corrélés (mais non indépendantes, en raison de l'utilisation du même terme d'erreur, au dénominateur du quotient de comparaison). Pour k moyennes, le nombre de couples de comparaisons est de $k(k–1)[k(k–1) – 2] / 8$, et la fraction de comparaisons corrélées est simplement $4 / (k + 1)$, d'où un niveau moyen de corrélation de $2 / (k + 1)$. Cette corrélation aurait peu d'effet sur le risque d'erreur par expérience (α_E) mais plutôt sur sa variance (Tukey 1991).

conventionnelle, p. ex. 0,05. Le principe de Bonferroni,

$$\alpha_C = \alpha_E / nc \,, \qquad (5a)$$

ou celui, presque équivalent, de Dunn-Sidak,

$$\alpha_C = 1 - (1 - \alpha_E)^{1/nc} \qquad (5b)$$

donnent une approche : elle consisterait ici à tester chacune des $nc = 10$ comparaisons par un t classique, en recourant à un seuil de signification α_C plus exigeant, soit 0,0050 (5a) [$t_{[\alpha C]} \approx \pm 3{,}078$] ou 0,0051 (5b) [$t_{[\alpha C]} \approx \pm 3{,}069$].

D'autres approches ont vu le jour, parmi lesquelles il est intéressant de citer :

- la procédure HSD de Tukey, déjà citée, et qui consiste à rapporter chaque différence $\bar{X}_i - \bar{X}_j$ en référence à la différence maximale parmi k moyennes, grâce à la loi du Studentised range (valeur utilisée : $q_{k,\, dl\,[1-\alpha]} / \sqrt{2} = q_{4,25[0,95]} / \sqrt{2} = 4{,}153 / \sqrt{2} \approx 2{,}937$) ;
- la procédure de Scheffé (Winer, 1971), qui couvrirait le risque émanant de toutes les formes possibles de contrastes entre les moyennes[3], en appliquant au test t la valeur critique $\{(k-1)\, F_{k-1,dl[1-\alpha]}\}^{½}$, ici $\{ 4 \times F_{4,25[0,95]} \}^{½} = \{ 4 \times 2{,}759 \}^{½} \approx 3{,}322$.
- la procédure de Newman-Keuls (Winer, 1971), une procédure hiérarchique très en vogue dans les années 1970, qui consiste à tester d'abord la différence la plus grande via le Studentised range de plein rang (i.e. q_k), puis, si celui-ci est significatif, de procéder aux différences de rang inférieur en se repliant sur un critère de rang correspondant (i.e. q_{k-1}), ainsi de suite[4]. Les valeurs applicables seraient donc $q_r / \sqrt{2} = 2{,}937$, 2,751, 2,491 et 2,060 pour r allant de 5 à 2. Cette procédure ne respecterait pas sa promesse de plafonner le risque d'erreur par expérience (cf. p. ex. Einot et Gabriel, 1985), et elle est plus ou moins tombée en discrédit depuis quelques décennies.
- la procédure de Ryan modifiée (voir Toothaker, 1991), similaire à celle de Neuman-Keuls mais avec un contrôle rigoureux de α_E. Soit r, l'empan ordinal de la différence, depuis $r = k$ jusqu'à $r = 2$: le test, non hiérarchique, consiste à appliquer une valeur critique issue du Studentised range $q_{r[\alpha']}$, où $\alpha' = \alpha_E$ pour $r = k$ et $k-1$, et $\alpha' = 1 - (1 - \alpha_E)^{r/k}$ pour $r \leq k - 2$. Dans notre cas et ramenant le tout à la forme du test t par $q / \sqrt{2}$, pour les différences d'empan $r = 5$, 4, 3 et 2, nous aurions les valeurs critiques respectives $q / \sqrt{2} \approx 2{,}937$, 2,751, 2,721 et 2,478.
- la procédure de Hayter (1986), qui vise à contrôler le risque d'erreur par expérience maximal (en référence au concept de différentes hypothèses nulles partielles), procédure qui se ramène à remplacer le critère q_k de HSD à un critère moins exigeant, à savoir q_{k-1}, applicable à toutes les différences, soit, ici, $q_{4[0,95]} / \sqrt{2} \approx 2{,}751$.
- la procédure de Dunnett (voir Winer, 1971), dans laquelle la k^e moyenne (« condition de contrôle ») est comparée à chacune des $k-1$ autres moyennes, le taux d'erreur global étant plafonné à α_E. Le « t de Dunnett » tient compte de la corrélation de +½ existant entre les $k-1$ comparaisons effectuées. Pour nos données, avec $k = 5$, $dl = 25$ et en mode bilatéral, $t_D = 2{,}607$.

Ces différents critères, on le voit, sont plus exigeants que le t normal (avec valeur critique de 2,060), lequel commettrait une erreur de type I au seuil de 0,0500, le seuil d'erreur par comparaison tombant jusqu'à 0,0070 (pour le HSD), 0,0050 (pour le critère Bonferroni), voire 0,0028 (pour le critère de Scheffé). Pour le chercheur qui, la plupart du temps, se plaint d'un manque de puissance plutôt que d'un excès de sensibilité expérimentale, ces procédures se justifient mal. De plus, pour reprendre un argument de Saville (1990), ces procédures sont inconsistantes en ce que, pour deux conditions expérimentales bien définies, la différence $\bar{X}_1 - \bar{X}_2$ sera significative si elle est testée en solo, et elle cessera de l'être au même seuil si elle est incluse dans un ensemble de comparaisons. Sans parler de la perte de puissance statisique lorsque le seuil effectif de significativité, α_C, devient dramatiquement ténu.

D'ailleurs, si on pousse la logique du contrôle du taux un peu plus loin, pourquoi ne pas limiter à α_E le risque de rejeter à tort une hypothèse nulle globale pour la totalité des tests d'une expérimentation, laquelle peut comporter plusieurs ensembles de variables et plusieurs séries d'analyses, voire, pour une série d'expérimentations sur la même question, ce qui acculerait éventuellement à zéro le seuil de rejet (α_C) par comparaison? Ce raisonnement *ad absurdum* nous oblige à rappeler le sens premier du test statistique, tel qu'appliqué sur une situation donnée : est-ce que *ce résultat* aurait pu être produit par le hasard échantillonnal, quelle est la probabilité que le hasard seul ait amené cette différence? Le fait que la probabilité soit très petite, ou qu'elle atteigne un seuil prédéterminé d'invraisemblance, nous permet de singulariser la dite différence, la rendre remarquable et concluante. Que ce test et cette différence soient inscrits dans un groupe d'autres tests semblables n'enlève rien au caractère exceptionnel du résultat. Bien sûr, on répliquera qu'il est plus facile de trouver un résultat « exceptionnel » si on observe 10

---

[3] Par exemple, on compte 6 contrastes simples pour $k = 3$ moyennes (1-2, 1-3, 2-3, 1+2-3, 1+3-2, 2+3-1), 25 pour $k = 4$, 90 pour $k = 5$, et 301, 966, 3025, 9330 et 28501 pour k de 6 à 10. Noter que le critère de Scheffé couvre aussi, en principe, *toute espèce de contraste*, p. ex. pour $k = 3$, tous les contrastes de forme $a_1M_1 + a_2M_2 - (a_1+a_2)M_3$, pour tout $a_1 > 0$ et $a_2 > 0$ (et non seulement $a_1 = a_2$, tels que posés dans le calcul présenté), donc un nombre *infini* de contrastes.

[4] Noter que, selon notre expérience, le principe *hiérarchique* (et conditionnel) de cette procédure est rarement respecté.

événements plutôt que seulement 1 ou 2. Malgré tout, le résultat exceptionnel n'en reste pas moins remarquable, et peut être dit « significatif », même s'il fait partie d'un grand ensemble. Ce résultat dit « significatif » n'est pas une preuve d'un effet expérimental certain : c'est seulement un argument, qui devra être corroboré par réplication expérimentale, comme le soulignait déjà Fisher (1971), et par d'autres démonstrations empiriques et logiques. La réduction du taux d'erreur par comparaison (α_C) à presque rien n'est pas une recette pour la certitude, mais au contraire un moyen pervers de miner la sensibilité et la puissance de nos dispositifs d'expérience.

Démonstration 4 : La force des comparaisons planifiées et globales

Pour la présente section, nous utiliserons encore le contexte d'une expérimentation sur $k = 5$ groupes de $n = 6$ participants chacun. Le Carré moyen intragroupe (CM_e), avec $dl_e = 25$, vaut 1,00, et les 5 moyennes sont :

1 : 4,1 2 : 4,7 3 : 4,6 4 : 5,2 5 : 5,6 .

À toutes fins utiles, la variance des moyennes est 0,333 et le Carré moyen groupes (CM_G), 1,998. Le test F global, 1,998 / 1,00 = 1,998, se compare au F avec $dl = 4$ et 25, soit $F_{4,25[0,95]} = 2{,}759$, et n'est pas significatif. Remarquons aussi que rien non plus ne ressort significatif au HSD de Tukey : p. ex., la différence la plus grande, 5,6 – 4,1, obtient $q = (5{,}6 - 4{,}1) / \sqrt{(1{,}00 / 6)} \approx 3{,}674$, contre $q_{1,5\,:5[0,95]} = 4{,}153$.

Test de régression linéaire (de premier degré). Supposons maintenant que les cinq conditions correspondent à des degrés d'une variable indépendante qui croissent linéairement (p. ex. la température, la charge pondérale, etc.) et à intervalles égaux. Nous pouvons alors tester si les moyennes obéissent aussi à cette progression linéaire, en appliquant la technique appropriée de calcul de contraste (désignée *trend analysis* : voir Winer, 1971). Dans le cas présent, les coefficients de contraste sont –3, –1, 0, 1 et 3, et le Carré moyen (CM_{Lin}) correspondant, $6 \times 5{,}0^2 / 20 = 7{,}50$, d'où un quotient $F = 7{,}50 / 1{,}00 = 7{,}50$, à comparer à $F_{1,25[0,95]} = 4{,}244$, soit une variation linéaire significative[5].

Test de variation monotone. Souvent, les conditions comparées correspondent à une variation croissante (ou décroissante) de la variable indépendante, sans qu'une métrique linéaire soit observée ou stipulée. En effet, soit la variable indépendante elle-même est mal contrôlée ou peu contrôlable (p. ex. taux d'humidité, degré d'entraînement ou d'expertise, etc.), soit la « théorie » ne va pas jusqu'à prédire un lien linéaire entre les variables indépendante et dépendante. La prédiction précise seulement une relation monotone, de niveau ordinal, telle que $\bar{X}_1 \leq \bar{X}_2 \leq \ldots \leq \bar{X}_k$. À cette situation, convient le test de variation monotone, utilisant par exemple la statistique $\bar{E}^2$ de Barlow, Bartholomew, Bremner et Brunk (1972 : voir aussi Laurencelle et Dupuis, 2000). Pour les données citées en exemple, la série comporte une interversion, entre les 2e et 3e conditions, qu'il y a lieu d'« amalgamer » (dans le vocabulaire des auteurs), soit fusionner, en en prenant la moyenne, obtenant la série monotone $\bar{X}_j^*$ = 4,1 4,65 4,65 5,2 5,6. Le Carré moyen du modèle (CM_{Mod}), basé sur $r = 4$ valeurs distinctes, est ici $6 \times s^2(\bar{X}_j^*) \times (k-1)/(r-1) = 2{,}654$, et, pour la statistique $\bar{E}^2$, nous obtenons $\bar{E}^2 = (r-1) \times CM_{Mod} / [(k-1) \times CM_G + dl_E \times CM_E] = 3 \times 2{,}654 / [4 \times 1{,}998 + 25 \times 1{,}00] = 0{,}2413$. Pour le modèle d'« ordre simple », approprié ici, la valeur critique du $\bar{E}^2$ avec $k = 5$, $dl_e = 25$ et $P = 0{,}95$ est 0,1708 (par interpolation harmonique, dans Laurencelle et Dupuis, 2000). Le test, significatif au seuil de 5%, indique bel et bien que les données respectent le modèle de progression monotone.

Test de l'égalité des différences successives. Comme exemple d'un autre test global, proposons celui consistant à vérifier si les différences successives de moyennes, $\bar{X}_{j+1} - \bar{X}_j$, sont « statistiquement » égales ou non (voir Annexe). Utilisant les $k = 5$ moyennes données en illustration, nous obtenons les 4 différences : 0,6, –0,1, 0,6 et 0,4. La variance de ces différences est 0,109167. Notant $\bar{X}_j = \mu + j \times \Delta + \bar{e}_j$, où Δ est la différence constante, nous avons en général $d_i = \bar{X}_{j+1} - \bar{X}_j = \Delta + (\bar{e}_{j+1} - \bar{e}_j)$, et $\mathrm{var}(d_i) = 2k/(k-1) \times \hat{\sigma}^2(\bar{e})$ ou $\frac{1}{2}n(k-1)/k \times \mathrm{var}(d_i) = \hat{\sigma}^2(e)$. Construisant notre Carré moyen des différences comme $CM_{Dif} = \frac{1}{2} \times 6 \times 4/5 \times 0{,}109167 \approx 0{,}262$, nous pouvons le tester par $F_{Dif} = CM_{Dif} / CM_E = 0{,}262 / 1{,}00 = 0{,}262$; le test, évidemment non significatif, ne rencontre pas la valeur critique $F_{Dif\,:\,k=5,\,dl=\,25[0,95]} = 3{,}09$ (trouvée par estimation Monte Carlo) et ne permet donc pas de rejeter l'hypothèse (globale) que les différences successives entre les 5 moyennes (dans l'ordre d'observation prédéterminé) sont constantes[6].

---

[5] Dans un tel contexte, il est souvent utile de montrer que les conditions ne démontrent pas de variation autre que linéaire, ce qui peut se faire en effectuant le test du résidu de variance. Ici, le total de variance entre les conditions est $4 \times 1{,}998 = 7{,}992$, dont on soustrait la variation linéaire (7,50), la différence 0,492 étant rapportée aux 3 degrés de liberté restants, d'où CM résiduel = 0,164, et un $F = 0{,}164 / 1{,}00 = 0{,}164$, manifestement non significatif.

[6] Ce résultat n'est pas surprenant ici, compte tenu du résultat antérieur relatif à la régression linéaire (sur intervalles égaux), un résultat non seulement significatif mais qui exprime aussi une proportion importante de la variance disponible (soit 7,50 sur 7,992, ou 94%). De façon générale, cependant, l'anti-concordance des deux tests n'est pas de règle.

En guise de conclusion

Tukey (1991), dans un texte assez touffu sur « The philosophy of multiple comparisons », nous formule deux rappels importants : le principe de réplication de Fisher, voulant que le caractère possiblement accidentel d'un résultat significatif a besoin d'être confirmé par des réplications expérimentales avant que le résultat soit considéré prouvé, et le fait qu'un intervalle de confiance en dit plus long, est plus informatif, que le test de significativité Oui/Non correspondant. Tukey, cependant, affirmant que « Comparisons build multiplicity fast » (1991, p. 104), plaide encore pour un plafonnement du risque d'erreur par expérience (α_E), possiblement par sa procédure HSD.

À l'opposé de Tukey (1991), qui évoque des exemples comportant k = 5, 10, jusqu'à 1415 conditions et un bassin virtuel énorme de comparaisons, Howell (1998) et Saville (1990) nous ramènent sur le terrain des expérimentations réelles, dans lesquelles la comparaison de 2, 3 ou parfois 4 conditions est d'usage, et où chaque test de comparaison s'appuie sur une hypothèse préalable. Les deux auteurs concourent pour revenir loin en arrière et reprendre la technique simple du LSD, ici un LSD non conditionnel, consistant à appliquer à chaque comparaison $\bar{X}_j - \bar{X}_i$ un simple test t (utilisant le CM_e comme estimateur de variance) : (1) ce seul test respecte le principe de consistance, à savoir que la différence sera jugée semblablement, quel que soit le contexte de conditions supplémentaires dans lequel elle est évaluée ; (2) ce seul test garde son rapport d'équivalence classique avec les intervalles de confiance par comparaison, un argument qu'a ignoré Tukey (1991) ; (3) ce test et cette procédure assurent une sensibilité expérimentale contrôlable pour chaque comparaison, au lieu de presque anéantir la puissance au profit d'un contrôle du risque d'erreur pour toutes les comparaisons d'une expérience. Par ailleurs, si en effet il y a un bon nombre de conditions à comparer et un nombre encore plus grand de comparaisons possibles, le chercheur reste libre de resserrer son seuil de détection α_C vers un niveau plus exigeant (p. ex. 0,01 ou 0,001) ou, mieux encore, de garder un seuil sensible (p. ex. 0,05) et de considérer sa récolte de résultats significatifs comme l'indication d'effets à confirmer ou d'hypothèses à envisager, tels que le suggèrent à la fois Howell (1998), Saville (1990) et Tukey (1991).

Références

Barlow, R. E., Bartholomew, D. J., Bremner, J. M., Brunk, H. D. (1972). *Statistical inference under order restrictions*. New York : Wiley.

David, H. A. (1981). *Order statistics* (2e édition). New York : Wiley.

Einot, I., Gabriel, K. R. (1985). A study of the powers of several methods of multiple comparisons. *Journal of the American Statistical Association*, *70*, 574-583.

Fisher, R. A. (1970). *Statistical methods for the research workers* (14e édition). New York: Hafner.

Fisher, R. A. (1971). *The design of experiments* (9e édition). New York : Hafner Press.

Hayter, A. J. (1986). The maximum familywise error rate of Fisher's Lest Significant Difference test. *Journal of the American Statistical Association*, *81*, 1000-1004.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, *75*, 800-802.

Hochberg, Y., Tamhane, A. C. (1987). *Multiple comparison procedures*. New York : Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65-70.

Howell, D. C. (1998). *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck Université.

Howell, D. C. (1999). *Fundamental statistics for the behavioural sciences* (4e édition). Pacific Grove (CA) : Duxbury Press.

Laurencelle, L., Dupuis, F. A. (2000) *Tables statistiques expliquées et appliquées* (2e édition). Québec : Le Griffon d'argile.

Miller, R. G. Jr. (1981). *Simultaneous statistical inference* (2e édition). New York : McGraw-Hill.

Ryan, T. A. (1960). Significance tests for multiple comparisons of proportions, variances, and other statistics. *Psychological Bulletin*, *57*, 318-328.

Saville, D. J. (1990). Multiple comparison procedures : The practical solution. *The American Statistician*, *44*, 174-180.

Toothaker, L. E. (1991). Multiple comparisons for researchers. Newbury Park (CA) : Sage.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics, *5*, 99-114.

Tukey, J. W. (1953). The problem of multiple comparisons. Document inédit (mais très circulé), Princeton University.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, *6*, 100-116.

Winer, B. J. (1971). *Statistical principles in experimental design* (2e édition). New York : McGraw-Hill.

Manuscript received 25 November 2011.
Manucript accepted 28 November 2011.

Appendix follows.

Annexe : La variance des différences successives, vd

Propriétés de base

La statistique proposée est :

$$\text{vd} = \frac{\sum_1^{k-1} \left(d_i - \bar{d}\right)^2}{2(k-2)} \times \frac{k-1}{k}, \qquad \text{(A1)}$$

où :

$$d_i = X_{i+1} - X_i \ , \ i = 1 \text{ à } k-1 \ . \qquad \text{(A2)}$$

Le développement de la somme, au numérateur de (A1) donne :

$$\sum_1^{k-1} \left(d_i - \bar{d}\right)^2 = 2 \sum_2^{k-1} X_i^2 + \left(X_1^2 - X_1^2/(k-1)\right) + \left(X_k^2 - X_k^2/(k-1)\right) - 2 \sum_1^{k-1} X_i \, X_{i+1} + 2X_1 X_k/(k-1) \qquad \text{(A3)}$$

En supposant des variables aléatoires centrées (i.e. $\mu_X = 0$, $\sigma_X^2 = \sigma^2$, $E\{X_i X_j = 0\}$), l'espérance de (A3) est $2\sigma^2 \, k(k-2)/(k-1)$, d'où évidemment $E(\text{vd}) = \sigma^2$.

Nous n'avons pas obtenu d'expressions algébriques pour les moments 2 (σ^2), 3 (γ_1) et 4 (γ_2) de vd. En approximation, par expérimentation Monte Carlo sur des données normales standard et pour des R^2 de 0,999 (établis sur $k = 3$ à 30) ou mieux, nous proposons :

$$\sigma^2 (\text{vd}) \approx 2\sigma^4 / (k-2)^{0,88} \qquad \text{(A4)}$$

$$\gamma_1 (\text{vd}) \approx 4 / \sqrt{(k-1)} \qquad \text{(A5)}$$

$$\gamma_2 (\text{vd}) \approx 24 / (k-1), \qquad \text{(A6)}$$

ces deux dernières expressions suggérant le comportement d'une variable χ^2 dotée de $½(k-1)$ degrés de liberté.

Test des différences successives pour l'anova

Nous avons produit par échantillonnage Monte Carlo les distributions d'une statistique mettant en jeu la variance des différences successives de k moyennes basées chacune sur n données, d'où un Carré moyen d'erreur (intragroupe) muni de $k(n-1)$ degrés de liberté : le test sur vd a la forme $n \times \text{vd} / \text{CM}_e$. Dans un premier temps, nous avons obtenu des valeurs critiques approximatives, dont le tableau 3 donne un extrait.

Dans un second temps, nous avons étudié les relations mutuelles entretenues par ce test (vd), le test F global de l'anova (constitué par $\text{CM}_G / \text{CM}_e$, i.e. le quotient du Carré moyen groupes sur le Carré moyen intragroupe), et le test HSD de Tukey, à savoir si un test rejette H_0 concomitamment à un autre test. Les statistiques obtenues sont basées sur 1000000 échantillons, ayant donc une précision de ±0,001. Par exemple, au seuil supérieur de 5%, le taux de rejets conjoints des tests vd et F est de 0,019, produisant une corrélation $\phi \approx 0,34$; la concomitance de vd avec HSD donne $\phi \approx 0,33$, alors qu'entre F et HSD, nous observons $\phi \approx 0,80$; au seuil de 1%, les données comparatives sont $\phi \approx 0,19$, 0,19 et 0,80. Sans dénier qu'il existe en fait un lien stochastique entre ces tests, il reste que le test des différences successives permet de vérifier une hypothèse distincte, et qu'on peut supposer qu'il sera plus sensible que les autres, et plus puissant, dans des conditions qui contredisent expressément cette hypothèse.

La distribution du test, $n \times \text{vd} / \text{CM}_e$, et ses propriétés sont, pour le moment, parfaitement inconnues.

Tableau 3. Centiles 95 et 99 (approximatifs) du test de différences successives (données normales)

	$n = 3$	5	7	10	20	∞
Centile 95						
$k = 3$	5,98	4,75	4,42	4,21	4,00	3,84
4	4,51	3,68	3,45	3,31	3,17	3,04
5	3,80	3,20	3,02	2,91	2,81	2,70
7	3,09	2,70	2,58	2,50	2,42	2,36
10	2,59	2,33	2,24	2,19	2,14	2,09
Centile 99						
$k = 3$	13,8	9,38	8,28	7,67	7,09	6,63
4	8,84	6,42	5,81	5,43	5,08	4,79
5	6,87	5,24	4,81	4,54	4,29	4,07
7	5,05	4,10	3,84	3,66	3,50	3,36
10	3,88	3,30	3,13	3,03	2,92	2,83