2019 ■ Vol. 15 ■ no. 1

Fisher contre Tukey, ou les paralogismes du contrôle alpha global pour évaluer la significativité statistique

Louis Laurencelle [a], ✉

[a]Université du Québec à Trois-Rivières

Abstract ■ Le principe d'un contrôle global du seuil de signification alpha dans le cas de tests statistiques multiples est remis en question, à la fois quant à l'aspect logique de la preuve de significativité et pour les paralogismes qu'il comporte. Le tout est illustré par des arguments de probabilité et grâce à un échantillon des principales techniques de contrôle en vogue (HSD, Bonferroni, Dunn-Sidàk, etc.). Tel qu'appliqué, le contrôle alpha global est inconsistant, contredit le principe de réplication en recherche et compromet ou fausse la puissance statistique. // The principle of a global control of the alpha significance threshold for a defined set of statistical tests is called into question, both as regards the logical aspect of the proof of significance and the paralogisms it entails. Illustrations are given based on probability arguments as well as through a sample of the main control techniques in vogue (HSD, Bonferroni, Dunn-Sidàk, etc.). The global alpha control as applied is inconsistent, contradicts the replication principle (for example, in meta-analysis) and compromises or distorts statistical power.

Keywords ■ Global alpha control, Test of statistical hypotheses, Bonferroni correction, Probabilistic proof's logic.

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

✉ louis.laurencelle@gmail.com

LL: na

10.20982/tqmp.15.1.p025

Introduction

Depuis la proposition de J. W. Tukey en 1983 à cet effet (*The problem of multiple comparisons*, voir Benjamini et Braun, 2002), l'idée s'est implantée, pour une expérimentation comportant plusieurs situations et comparaisons, de contrôler le risque d'une récolte surabondante de tests artificiellement significatifs. Cette idée a donné lieu au critère probabiliste HSD (ou WSD)[1] de Tukey, puis à ceux de Newman-Keuls (Winer, 1971), Welsch (1977), Dunn-Bonferroni (désigné Bonferroni, voir Dunn, 1961), Dunn-Sidàk (Sokal & Rohlf, 1981; Benjamini & Hochberg, 1995), et cætera, quelques-uns ayant aussi des variantes.

Cette idée, dans une étude comportant plusieurs tests statistiques et pour laquelle toutes les hypothèses nulles (H_0) seraient vraies, vise à imposer une valeur limite à la probabilité qu'au moins l'une d'entre elles soit erronément rejetée. Rappelons que l'H_0 reflète la condition théorique selon laquelle les variations constatées dans les données d'une étude sont entièrement le résultat d'influences aléatoires et reflètent en fait du 'bruit'. La probabilité limite mentionnée est le seuil alpha (α) global (ou risque alpha global, α_G), par opposition au seuil alpha par comparaison (α_C) appliqué à chacun des tests de l'ensemble. Évidemment, nous avons l'inégalité $\alpha_C \leq \alpha_G$.

Dans le système théorique de Neyman and Pearson (1928a, 1928b, 1936) élaboré pour donner un cadre à la

[1]Honestly Significant Difference ou Wholly Significant Difference (Benjamini & Braun, 2002).

procédure des tests d'hypothèses statistiques, chaque H_0 est couplée à une H_1, l'hypothèse de recherche (ou contre-hypothèse) contredisant H_0 et reflétant spécifiquement la variation attendue dans la situation testée. L'hypothèse nulle (H_0) serait non spécifique, reflétant le seul effet du hasard, mais accompagnée d'une H_1 spécifique déterminée par la situation testée. Dans un contexte qui comprend deux ou plusieurs situations testées et en vertu du contrôle global souvent préconisé, chacun des couples H_0 :H_1 est soumis au seuil global α_G mentionné et testé généralement[2] par le seuil α_C, calculé en vue de garantir un seuil global α_G.

Établi sous le principe d'un contrôle du risque alpha global, le seuil effectif permettant de décider de la significativité d'un résultat individuel, α_C, est plus exigeant, parfois beaucoup plus exigeant que le seuil nominal α ($= \alpha_G$) convenu, p. ex. $\alpha = 0,05$ ou $\alpha = 0,01$. Ainsi, en vertu de l'approche Dunn-Bonferroni, chacun des tests participant à un ensemble de k tests doit être soumis au seuil de probabilité $\alpha_C = \alpha/k$. Cette exigence, maintenant imposée presque partout sous une forme ou sous une autre pour la communication et la publication des résultats de recherche, est-elle justifiée, voire résiste-t-elle à l'analyse logique? C'est ce que nous examinerons tantôt, dans cet article. Il reste qu'un seuil plus exigeant entraîne automatiquement une baisse de puissance statistique, baisse parfois dramatique et qui peut compromettre la productivité du paradigme de la recherche scientifique, à savoir : (1) expérimenter et repérer une cause ou une relation potentielle, (2) la vérifier dans des expérimentations complémentaires, (3) une fois la relation confirmée par réplication, en faire l'étude paramétrique, et (4) la diffuser à titre de règle ou loi généralement applicable. L'exigence imposée par le contrôle alpha global attaque directement le germe de cette démarche, son pas no 1, en conduisant le chercheur à ignorer ou rejeter un résultat n'atteignant pas le seuil voulu, cela parce que ce résultat fait partie d'un contexte dans lequel d'autres résultats sont présents et qu'ils doivent supposément être contrôlés collectivement. Cette importante perte de puissance a poussé certains, chercheurs et statisticiens, à élaborer ou inventer d'autres approches supposées appliquer le contrôle global, ou un corollaire de ce contrôle, tout en essayant de circonscrire la perte de puissance encourue ou de l'atténuer : en témoignent les recensions de Klockars (1986), Hochberg (1987), Hsu (1996), Hsu (1996), Benjamini, Bretz, and Sarkar (2004) ainsi que, par exemple, les "améliorations" apportées au critère Bonferroni (Worsley, 1982; Simes, 1986; Hochberg, 1988). Il reste que, dans ces efforts, souvent ingénieux, parfois controuvés et parfois mathématiquement astucieux, la logique même du contrôle alpha global n'est jamais mise en cause.

Nous soulignerons plus bas d'autres bizarreries, voire des illogismes, qui s'attachent au principe et aux principales procédures du contrôle alpha global.

Pour fins de clarté dans cet essai, nous distinguerons l'H_0 globale et l'H_0 générale : la proposition des vocables "générale" (~universelle (Perneger, 1998), générique) et "globale" (~unique, applicable identiquement) est partiellement arbitraire et servira surtout à discriminer deux catégories d'expérimentations à situations multiples. L'H_0 générale s'accouple à différentes H_1 (qu'on peut noter $H_1^{(1)}, H_1^{(2)}, \ldots, H_1^{(k)}$) correspondant à k situations mises en jeu dans une expérimentation, alors que l'H_0 globale est associée à une H_1 unique, couvrant un ensemble de k situations identiques, lesquelles constituent en fait des réplications: nous fournirons quelques exemples importants de ce type de situations de recherche, le contrôle d'un alpha global pouvant s'appliquer légitimement dans ce cas.

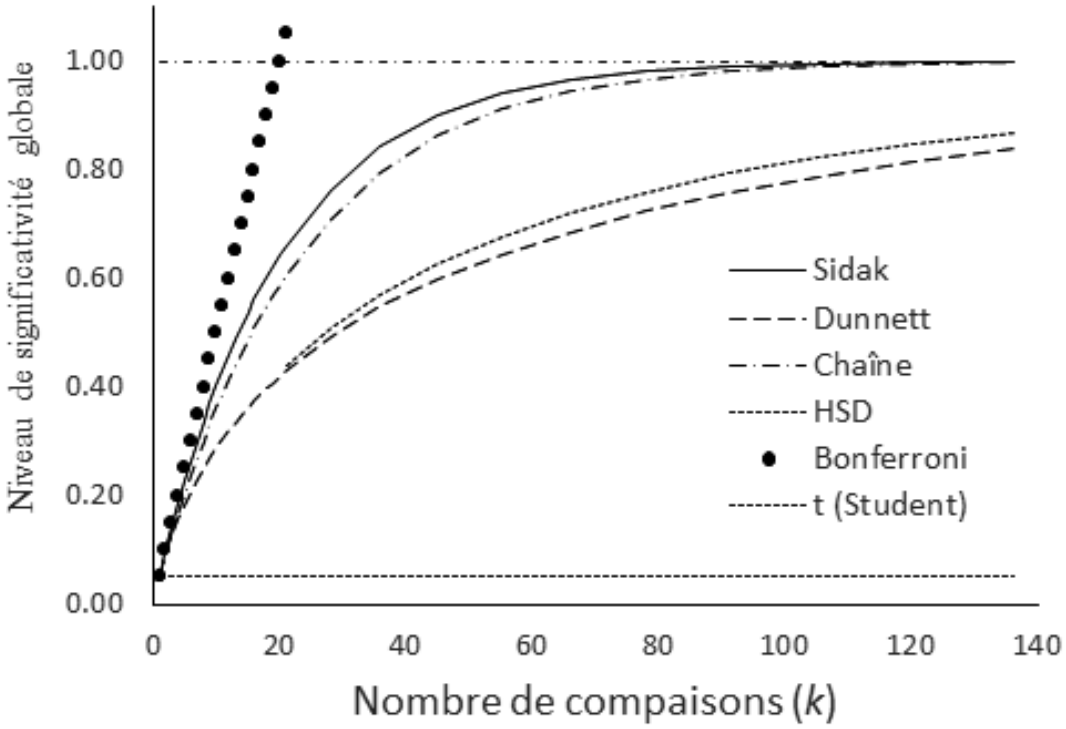
En épilogue, nous situerons la technique du test d'hypothèses statistiques concernée ici parmi d'autres approches de preuve concurrentes, soit l'intervalle de confiance, l'estimation bayésienne, la grandeur d'effet, en tentant d'identifier l'incidence et l'impact d'un "contrôle alpha global" sur les techniques mentionnées.

Le contrôle imposé au chercheur et ses dommages

D'abord, faisons un court rappel sur le principe du contrôle d'un risque alpha global, tel que les auteurs ont cherché à l'intégrer dans les principales procédures connues. Pour fins d'illustration, le contexte de référence sera celui de situations où l'on teste les différences de moyennes d'un groupe à l'autre, le test étant équitable en ce que les deux groupes comparés sont échantillonnalement équivalents, leurs données bien distribuées et le seul critère qui les différencie tient au traitement ou régime distinct qu'on a appliqué à chacun. L'hypothèse nulle (H_0) stipule que les seules variations présentes dans les données des deux groupes sont l'effet du hasard, c.-à-d. de conditions non contrôlées et homogénéisées d'un groupe à l'autre par échantillonnage ou randomisation. Sous H_0 et en vertu du test appliqué, la loi de distribution est connue, le t de Student, et elle gouverne la probabilité que la statistique du test déborde une valeur critique correspondant à un seuil de probabilité α, que ce soit en mode bilatéral ou unilatéral.

[2] Il existe au moins deux catégories de procédures visant à appliquer le contrôle global, certaines, dites hiérarchiques ou conditionnelles, utilisant une hiérarchie de seuils selon la structure (ordinale) des données comparées.

Figure 1 ■ Inflation de significativité sous H_0 selon le nombre (k) de comparaisons, en fonction du système de comparaisons et du mode de calcul appliqués, selon $\alpha_C = 0,05$.



L'inflation d'un risque d'erreur global sous comparaisons multiples. Soit le test t sur la différence de deux moyennes indépendantes, $t_{\bar{X}_1-\bar{X}_2}$, calculé par :

$$t_{\bar{X}_1-\bar{X}_2} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{2\,\widehat{\sigma}_e^2/\tilde{n}}},$$

où

$$\widehat{\sigma}_e^2 = \frac{(n_1-1)\,s_1^2 + (n_2-1)\,s_2^2}{n_1+n_2-2}$$

et $\tilde{n}$ est la moyenne harmonique de n_1 et n_2, répondant à la loi du t de Student, dotée ici de $\nu = n_1 + n_2 - 2$ degrés de liberté. Différents systèmes de comparaisons sont possibles, variant selon la structure de l'expérimentation et la nature des hypothèses à valider. Dans certains cas, les comparaisons de moyennes sont statistiquement indépendantes les unes des autres, étant basées sur des groupes de données différents. Dans d'autres cas, certaines comparaisons reprennent une moyenne déjà incluse dans une ou quelques autres comparaisons, entraînant ainsi une corrélation entre celles-ci : les $k =_m C_2$ comparaisons de m ($m > 2$) moyennes entre elles en sont l'exemple type[3]. La Figure **??** fait voir, pour six systèmes et modes de calcul différents (détaillés plus bas) la relation mesurée[4] entre le nombre (k) de comparaisons effectuées sous H_0 et la probabilité (sauf pour Bonferroni, voir plus bas) d'observer au moins une sanction positive du t de Student. Le seuil nominal de référence est $\alpha = 0,05$.

- t (Student) : pour un seuil α nominal de 0,05, le test t régulier produira un résultat significatif 5 fois sur cent dans chacune des comparaisons, à condition que H_0 soit vraie. Ainsi, la performance tracée en ligne pointillée au bas de graphique indique la significativité de référence, ici $\alpha = 0,05$, telle que conçue et ap-

[3] Par exemple, $m = 4$ moyennes (a, b, c, d) permettent de former ${}_4C_2 = 6$ comparaisons (ab, ac, ad, bc, bd, cd) et ${}_6C_2 = 15$ paires de comparaisons. Parmi ces paires, 3 sont "indépendantes", soit ab:cd, ac:bd et ad:bc, et 12 sont corrélées, partageant entre elles un élément, telle la paire ab:ac). Le quotient (τ) de paires corrélées est ainsi de 12/15, ou 0,80.

[4] Les données affichées à la Figure **??** sont obtenues par calcul pour la ligne Sidàk (ainsi que pour Student et Bonferroni) et par estimation Monte Carlo (1.000.000 échantillons) pour Dunnett, Chaîne et HSD. Noter que, pour le critère de Scheffé (réexaminé plus bas), le risque global de tests significatifs sous α_C est indéterminé et indénombrable, puisqu'il couvre théoriquement tous les contrastes linéaires (ou comparaisons) formables à partir des m moyennes.

pliquée selon le t classique de Student. Dans les cinq calculs illustrés qui suivent, cette valeur est appliquée à chaque comparaison ($\alpha_C = 0,05$) pour l'estimation d'un risque α global ($\alpha_G > 0,05$).

- Sidàk : le principe Sidàk (ou Dunn-Sidàk) suppose que les k comparaisons faites sont mutuellement indépendantes, étant basées sur $m = 2 \times k$ moyennes distinctes et avec corrélation nulle. Grâce à leur indépendance, la probabilité sous H_0 qu'au moins l'une d'elles soit significative est le complément de la probabilité qu'aucune ne le soit, c.-à-d. $P = 1 - (1 - \alpha)^k$. Le quotient τ de comparaisons corrélées dans ce cas est donc $0/k = 0$.
- Dunnett : le principe Dunnett concerne les comparaisons de k moyennes à une même moyenne de référence, comparaisons de forme a-b, a-c, a-d, etc., basées sur $m = k + 1$ situations. Ce système sous-tend une corrélation de $\rho = 1/2$ (pour n_j égaux) entre les k comparaisons effectuées. Le quotient de comparaisons corrélées sur le nombre total (k) de comparaisons est donc ici de $\tau = k/k = 1$.
- Chaîne : nous avons créé un système de k comparaisons enchaînées, de forme a-b, b-c, c-d, etc., basées sur $m = k + 1$ moyennes $\bar{X}_j$, chaque comparaison héritant de la seconde moyenne dans la comparaison précédente, soit $\bar{X}_j - \bar{X}_{j-1}$, pour $j = 1$ à k, et générant ainsi une corrélation entre celles-ci. Il se trouve donc k comparaisons corrélées par paires successives, d'où un quotient de paires corrélées de $\tau = (k-1)/{}_kC_2 = 2/k$.
- HSD : c'est ici le principe invoqué par Tukey et qui concerne l'ensemble des $k = {}_mC_2$ comparaisons pairées entre les moyennes d'un ensemble de m moyennes. Parmi les ${}_kC_2 = m(m-1)(m-2)(m+1)/8$ paires de différences, on compte $m(m-1)(m-2)/2$ paires corrélées à $\rho = 0,5$, pour un quotient de paires corrélées égal à $\tau = 4/(m+1) = 8/(3+\sqrt{1+8k})$.
- Bonferroni : tandis que les principes de calcul alpha évoqués ci-dessus, comme les procédures de contrôle qui leur sont associées, concernent la probabilité de produire par erreur au moins 1 sanction de significativité pour k tests, le principe exact dit de Bonferroni concerne le nombre moyen de sanctions attendu: si chaque test a une probabilité α_C d'être significatif, alors, selon la loi binomiale pour k événements (indépendants ou non), le nombre attendu est $k \times \alpha_C$: c'est la raison pour laquelle la "courbe" Bonferroni, dans la Figure **??**, défonce la ligne dénotant la probabilité asymptotique, $P = 1$, applicable aux autres critères.

Un double motif possible de l'invocation fréquente du critère Bonferroni exact dans les publications est que d'une part il est d'application simple et que, d'autre part, il approche le calcul de probabilité plus strict de Sidàk pour de petites valeurs de k. En effet, dans l'expression présentée plus haut, $P = 1 - (1 - \alpha)^k$, le binôme $(1 - \alpha)^k$ donne lieu au développement en série $1 - k \cdot \alpha + {}_kC_2 \cdot \alpha^2 - \ldots$ comprenant $k+1$ termes. Comme le troisième terme (comportant α^2) et les suivants ($\alpha^3, \alpha^4, \ldots$) s'évanouissent rapidement, on peut arrêter la série au second terme, obtenant alors $P \approx 1 - (1 - k \cdot \alpha) \approx k \times \alpha$, soit la valeur de Bonferroni.

Abstraction faite du calcul Bonferroni, il nous paraît plus qu'intéressant de noter à la Figure **??** la relation *négative* patente qu'il y a entre le degré d'intercorrélation (τ) des comparaisons et le taux de significativité enregistré. En effet, la gradation des courbes tracées, soit Dunnett < HSD < Chaîne < Sidàk, évolue a contrario par rapport à celle de la proportion de liens corrélationnels (τ) entre les comparaisons : toutes indépendantes dans le système Sidàk ($\tau = 0$), le taux d'intercorrélation augmente dès le système Chaîne, où chaque comparaison est liée à une autre, soit $\tau = 2/k$, progresse avec HSD ($\tau = 8/(3 + \sqrt{1+8k})$) et culmine avec Dunnett ($\tau = 1$) : la Figure **??** illustre cette relation, indiquant cette fois l'évolution de la portion de comparaisons indépendantes (non corrélées) pour les quatre systèmes principaux.

Il est donc légitime de conclure que l'influence du risque α est maximale dans un système de comparaisons indépendantes et qu'elle est freinée par la portion de comparaisons corrélées qu'il renferme.

Le contrôle du risque d'erreur global et ses critères. Les critères. Par contrôle du risque α global, la presque totalité des auteurs, avant et après Tukey (1983), entendent le fait de :

> Conserver sous α, un seuil de probabilité prescrit, la probabilité que H_0 soit rejetée au moins 1 fois sur un ensemble défini de k tests.[5]

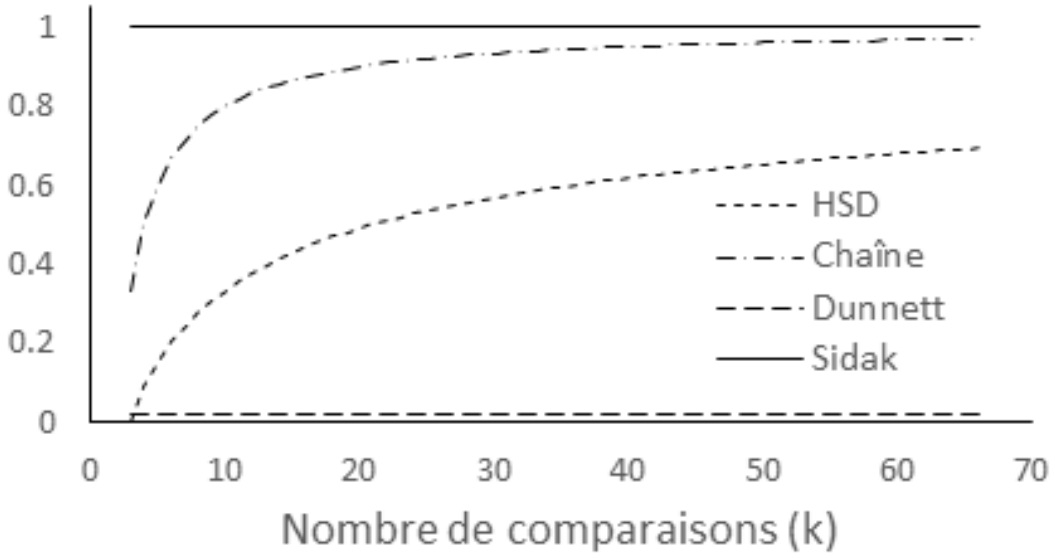
Pour qui voudrait contrôler le risque α global en le maintenant sous un seuil nominal α donné, il faut en principe faire le calcul inverse de celui représenté à la Figure **??**, de telle façon que la probabilité d'application du critère à un ensemble de k comparaisons respecte cet alpha global, soit, symboliquement :

$$P = \text{Calcul}\{\text{système}, k, \alpha_C\} \sim \alpha_G.$$

Il s'agit alors de déterminer le seuil individuel α_C et sa valeur critique appropriés au système de comparaisons,

[5] Sauf certains critères pour lesquels l'ensemble évoqué est bien défini (les ${}_mC_2$ comparaisons de m moyennes pour le HSD, les moyennes soumises au F global pour le LSD, etc.), la plupart des critères, notamment ladite règle de Bonferroni, réfèrent à un ensemble flou, pouvant inclure et devant théoriquement inclure tous les tests de la même hypothèse de recherche (H_1), ceux de l'expérimentation rapportée comme ceux déjà parus, voire à paraître, le seuil α_C pouvant alors être néantisé (Perneger, 1998; Moran, 2003)

Figure 2 ■ Proportion τ de comparaisons indépendantes (non corrélées) parmi k, selon le système de comparaisons appliqué.



tels que la probabilité d'obtenir sous H_0 au moins une sanction significative parmi les k comparaisons est bornée supérieurement par α_G, le seuil global. Ainsi, dans le cas très simple du critère Bonferroni, on aura $\alpha_C = \alpha_G/k$, de sorte qu'il suffit de trouver, par exemple, la valeur critique du t de Student appropriée, $\pm t_{v[1-\alpha_C/2]}$, en mode bilatéral par exemple. Quant au critère Dunn-Sidàk, on l'obtient facilement en inversant l'expression fournie plus haut, soit:

$$\alpha_C = 1 - (1 - \alpha_G)^{1/k},$$

et en trouvant la valeur t correspondante.[6] Pour le critère HSD de Tukey, il est déterminé non seulement par le seuil de probabilité α (= α_G), mais correspond surtout à la valeur critique de l'étendue "studentisée" de m normales (indépendantes). Pour m échantillons produisant les $k = {}_mC_2$ comparaisons et asymptotiquement sur la taille n des m moyennes, la variable servant de critère HSD est $q_m = (z_{[m,m]} - z_{[1,m]})/s_\nu$, où les $z_{[1,m]}$ et $z_{[m,m]}$ sont les statistiques d'ordre extrêmes d'une série normale standard de m éléments, s_ν est l'écart-type de distribution χ_ν et $q_{m[1-\alpha]}$ est le quantile $1-\alpha$ de q_m (David, 1981; Laurencelle & Dupuis, 2000). Quant au critère de Dunnett, il dérive aussi d'une valeur critique[7] (Dunnett, 1955; Laurencelle & Dupuis, 2000) et se présente en mode unilatéral comme bilatéral : il se calcule comme un t et sa valeur critique peut s'écrire $t_{D[k,\alpha]}$, La statistique q_m appliquée dans la procédure HSD peut elle aussi être exprimée sous forme d'un t, par la simple transformation $t = q_m/\sqrt{2}$. Mentionnons enfin le critère classique de Scheffé (1959); voir aussi Winer (1971) et Sokal and Rohlf (1981), basé sur le F de l'analyse de variance des k groupes et exprimable sous forme d'un t: ce critère "protège" tous les contrastes linéaires possibles entre les k moyennes, par exemple $c_1\bar{X}_1 + c_2\bar{X}_2 + \cdots + c_k\bar{X}_k$ tel que $\sum c_j = 0$ et $\sum c_j^2 = 1$.

Les valeurs critiques et leur impact sur chaque comparaison. À toutes fins utiles, le Tableau **??** présente des calculs illustrant le large spectre des valeurs critiques qui découlent des procédures de contrôle alpha mentionnées ainsi que leur impact sur la significativité et la puissance disponible pour chaque comparaison individuelle.

L'examen du Tableau **??** permet de comparer les différentes règles de décision sous contrôle global α, ce par rapport à la règle classique représentée par le simple t de Student. Un exemple-type de laboratoire, figuré par $m = 5$ groupes de $n = 10$ participants chacun, sert d'illustration, en regard de son horizon asymptotique ($n \to \infty$). Pour les comparaisons pairées entre les 5 moyennes, l'exigence de "valeur exceptionnelle" imposée par les principes de Bonferroni et Sidàk, leur sévérité relative, est de $0,05/0,005 = 10$, c.-à-d. généralement k fois plus élevée que celle du sim-

[6] Des tables de valeurs critiques existent, par exemple dans Rohlf and Sokal (1994) et Laurencelle and Dupuis (2000).

[7] La variable t_D est définie (au numérateur) comme le maximum de k normales standard à intercorrélation $\rho = 0,5$: voir Kotz, Balakrishnan, and Johnson (2000).

Table 1 ■ Valeurs critiques du critère de forme t et seuils α_C associés dans une situation-type présentant $m = 5$ groupes de $n \to \infty$ ($\nu = \infty$) ou $n = 10$ ($\nu = 45$ degrés de liberté) éléments chacun, selon différentes procédures de contrôle alpha et pour un test de mode bilatéral sous $\alpha = 0,05$, avec puissance estimée selon une grandeur d'effet $d = 0,8$.

Procédure	$\nu \to \infty$			$\nu = 45$		
	t	α_C	Puissance	t	α_C	Puissance
Sans contrôle alpha global - Tous tests t individuels						
t (Student)	1,960	0,0500	0,716	2,014	0,0500	0,697
Avec contrôle alpha global - Toutes comparaisons pairées						
HSD*	2,728	0,0064	0,422	2,841	0,0067	0,389
Dunn-Sidàk	2,800	0,0051	0,394	2,944	0,0051	0,352
Bonferroni	2,807	0,0050	0,391	2,952	0,0050	0,350
Avec contrôle alpha global - Comparaisons d'une moyenne à toutes les autres						
Dunnett	2,442	0,0146	0,535	2,531	0,0149	0,505
Dunn-Sidàk	2,491	0,0127	0,516	2,595	0,0127	0,480
Bonferroni	2,498	0,0125	0,513	2,602	0,0125	0,478
Avec contrôle alpha global - Tous contrastes possibles ($k \to \infty$)						
Scheffé**	3,080	0,0021	0,291	3,212	0,0024	0,265

Note.
* Selon la transformation $t_{dl} = q_{m,\nu}/\sqrt{2}$;
** Selon la transformation $t = \sqrt{(m-1)F_{m-1,dl_e}}$.

ple t, celle du critère de Scheffé est de 20,8 plus élevée, alors que le critère HSD, avec 7,4, ressort ici comme le plus recommandable parmi les critères généraux (Ury, 1976). Recommandable oui, principalement pour les $k = {}_mC_2$ comparaisons (family-wise) de m moyennes entre elles pour lesquelles son contrôle est exact, mais non pour k comparaisons toutes indépendantes, pour lesquelles seul le critère Sidàk est approprié et juste, comme l'est le critère Dunnett pour k comparaisons corrélées à $\tau = 1/2$.

D'autres critères, la plupart étant des variantes de ceux présentés ici, donnent lieu à des niveaux de sévérité intermédiaires, notamment les critères hiérarchiques tels ceux de Neuman-Keuls (Winer, 1971), Holm (1979), Hochberg (1988), Benjamini and Hochberg (1995), critères dont la mathématique parfois savante manque de transparence logique.

Il peut intéresser le lecteur d'apprendre que, en analyse de variance, une forme de contrôle global, dérangeante elle aussi, s'exerce dans le test F global appliqué à $m \geq 3$ moyennes. En effet, la significativité qui résulterait d'un test F sur la différence entre deux moyennes particulières s'étiole et tend à s'évanouir suite à l'inclusion de moyennes supplémentaires, la variance "significative" se trouvant diluée par l'ajout de degrés de liberté supplémentaires au numérateur du F (Saville, 1990; Laurencelle, 2012), comme c'est aussi le cas aussi en analyse de variance multivariée avec l'ajout de nouvelles variables dépendantes (Scheiner, 2001; Moran, 2003).

Les absurdités et contradictions du contrôle alpha global

Les absurdités.

En plus de la perte de puissance qui lui est associée, le contrôle d'un risque d'erreur α global appliqué à un contexte comprenant deux tests statistiques ou plus donne lieu à des conclusions pour le moins paradoxales, comme le montrent les exemples qui suivent.

En recherche expérimentale. Un chercheur en nutrition étudie les effets de la caséine, une protéine (présente dans le lait), sur le développement de la masse musculaire des souris. Il compare deux groupes de 10 souriceaux, les uns nourris par biberon durant quatre semaines avec un lait à teneur normale en caséine (32 mg/kg), les autres, avec un lait à teneur augmentée (40 mg/kg). À l'échéance, le poids maigre des souris est estimé, et le test t résultant (de mode bilatéral) donne une probabilité extrême de 0,030, une différence significative au seuil α = 0,05. Par souci de rigueur et parce qu'un résultat répliqué lui semble plus assuré, le chercheur refait l'expérience, employant 20 nouveaux souriceaux, et il obtient derechef un t significatif, de probabilité 0,030 : notons que l'H_0 considérée ici est globale, les deux situations testées présentant la même H_1.

Or, d'après le principe d'un contrôle α global, par exemple selon le critère de Bonferroni, le seuil individuel α_C à battre est α/k, ici $\alpha_C = 0,05/2 = 0,025$. En appliquant ce seuil corrigé, ni l'une ni l'autre recherche ne ressort significative, de sorte que ce qui, pour le chercheur, devait

être une confirmation de l'effet par réplication devient ici une dénonciation de l'effet. La même conséquence paradoxale se produirait dans un contexte autre, par exemple si, en plus de ses deux groupes à caséine, évalués une seule fois, le chercheur avait mis sur pied une expérience semblable, cette fois avec des teneurs en glucose différentes : l'hypothèse nulle serait ici dite générale, les deux H_1 concernées étant distinctes. Que l'une ou l'autre expérience (ou les deux) se révèle individuellement significative, leur concomitance dans le même contexte de recherche, sous le régime du contrôle alpha global, pourra les invalider.

En santé publique. Un centre de santé publique, menant une enquête sur le territoire sous sa juridiction, fait un relevé exhaustif de la prévalence de 4 maladies génétiques, dont les niveaux de prévalence connues sont, respectivement, de π_1, π_2, π_3 et π_4 ($0 < \pi < 1$).

La population se dénombrant à N personnes, on relève n_1, n_2, n_3 et n_4 personnes touchées respectivement. Les tests montrent que la maladie 3, à prévalence $p_3 = n_3/N$, excède sa valeur de référence π_3 avec une probabilité extrême de $P \approx 0,022$.

Doit-on s'alarmer? Selon le critère de Bonferroni et les 4 maladies considérées, le seuil global α de 0,05 exigerait, pour la significativité de chaque test, une probabilité individuelle de $\alpha_C = 0,05/4 = 0,0125$, voire le critère mieux calculé attribué à Dunn-Sidàk, soit de $1 - (1 - 0,05)^{1/4} \approx 0,0127$. D'après ce critère, les autorités de santé publique peuvent se rassurer : la prévalence de la maladie 3 n'est pas vraiment significative… et il faudrait conclure pareillement si l'une ou l'autre des 3 autres maladies donnait une prévalence "faussement significative" selon une probabilité $0,0125 < P < 0,05$! Moran (2003) présente aussi une expérience virtuelle pour laquelle la conclusion de significativité est logiquement et évidemment probante, mais que le contrôle à la Bonferroni écrase tout à fait.

Au casino. Situons-nous maintenant dans un environnement civil, le casino, où le joueur, qui se croit capable d'influencer le résultat, est confronté à un système dans lequel règne le hasard seul : quel est l'impact du critère Bonferroni dans ce contexte? Le joueur paie D unités monétaires son droit de jouer, son gain espéré est de G unités et π sa probabilité de gagner. Pour être honnête[8] et sous un postulat[9] de hasard pur, le système doit obéir à l'équation :

$$\pi \cdot G - D = 0,$$

d'où les relations $\pi = D/G$, $G = D/\pi$ et $D = \pi \cdot G$. Le joueur débourse D unités, par exemple des dollars; s'il est 'chanceux', il en remporte G, son gain net étant $G - D$. Pour illustration, supposons une roulette à 20 cases, le joueur devant risquer $D = 1$; la probabilité que la boule s'arrête à la case prédite est $\pi = 1/20 = 0,05$, et le gain espéré $G = 1/0,05 = 20$ unités. Le même joueur peut, s'il le veut, s'essayer $k = 3$ fois, au coût total $D_k = k \cdot D = 3$, pour des gains espérés (mais non garantis!) de $G_k = k \cdot G = 60$: un jeu toujours équilibré, avec un gain probable de $G_k \cdot \pi = 3$ égal à son coût D_k. Qu'arrive-t-il si, s'inspirant du principe de contrôle appliqué dans certains secteurs de recherche universitaires, le directeur du casino impose le critère Bonferroni aux clients déterminés à jouer k fois à la roulette, ce principe stipulant que la probabilité globale (π) consentie doit être répartie également entre les k coups? Ici, pour $k = 3$ coups, la probabilité par coup sera fixée à $\pi'_3 = \pi/3 = 1/60$, le client devant alors jouer sur la roulette à 60 cases. Le client déboursera encore un montant $D_k = k \cdot D$, ici $D_k = 3$, rêvant faire un gain de $G_k = 3 \times 20 = 60$ unités. Mais son gain probable est maintenant $G_3 = 1$, selon $G_k \cdot \pi'_k = 60 \times 1/60 = 1$, le gain probable net étant négatif, soit $-(k-1) \cdot D = -2$, ce à l'avantage de la banque, une situation qui vaudrait certainement un procès pour fraude contre le casino. Par analogie avec le contexte des tests d'hypothèses, le coût d'un jeu (D) représente la mise sur pied de chaque condition de recherche par le chercheur et les mesures et dépenses associées à sa mise en œuvre, le gain (G) espéré étant évidemment l'obtention d'une sanction positive, un test significatif pour chaque condition. Quant à la probabilité de succès (π), elle correspond au caractère remarquable, à l'exceptionnalité du résultat obtenu dans la condition testée (tout comme la chute de la boule dans la bonne case de la roulette). En quoi, pour la recherche et les tests d'hypothèses, serait-il plus 'honnête', plus légitime de proportionner la sévérité du test au nombre (k) de conditions testées, trahissant ainsi l'équilibre comptable du modèle aléatoire sous-jacent, alors qu'on parlerait de crime dans le cas d'un casino?

En contrôle de qualité. Une usine métallurgique fabrique des jetons (Face / Pile) destinés notamment aux casinos. L'ingénieur remarque une tendance des jetons à "tomber" du côté "Face". Sur 10 premiers lancers, il obtient 10 "Face"; pour des pièces non biaisées, selon $p(\text{Face}) = 1/2$, le résultat aurait une probabilité d'environ $(1/2)^{10} \approx 0,000977$, suggérant fortement un problème de biais. Or,

[8] Dans un casino réel, le jeu ne respecte pas le strict principe d'honnêteté mathématique imposé ici, à défaut de quoi les propriétaires feraient banqueroute. En général, la valeur de π, la probabilité de gain, est légèrement rabattue, par exemple en ajoutant une case neutre (non gagnante) à la roulette.

[9] Il s'agit ici d'un postulat plutôt que d'une hypothèse, étant donné que rien, sinon peut-être les pouvoirs paranormaux du joueur, ne peut restreindre les effets du hasard. Dans un contexte de recherche typique, au contraire, plusieurs influences incontrôlables peuvent être présentes et, surtout, l'intervention expérimentale elle-même qui combat expressément l'hypothèse nulle!

en répétant 53 fois l'expérience et appliquant la règle Bonferroni, avec $\alpha_C = 0,05/53 \approx 0,000943$, ce résultat et d'autres semblables cesseraient tout bonnement d'être trouvés "significatifs" au seuil global $\alpha_G = 0,05$!

Les contradictions. Outre les paradoxes et anomalies signalés déjà, le contrôle de la significativité d'un ensemble de tests par un seuil alpha global contrevient aux principes régissant d'autres procédures statistiques : nous en revoyons quelques-unes.

La première difficulté touche la construction d'un intervalle de confiance et son interprétation, par exemple avec l'indice d de Cohen (1988). Dans la comparaison de deux moyennes, chacune a son incertitude (mesurée par son erreur-type propre) et leur différence, sanctionnée ou non par un test t, donne lieu à une grandeur d'effet d comportant sa propre erreur-type, l'intervalle sur d couvrant aussi un domaine de probabilité $1-\alpha$. Si d'aventure l'étude concernée inclut $r-1$ autres comparaisons, faudra-t-il, en obéissant au principe Bonferroni, élargir le domaine d'incertitude au niveau $1-\alpha/r$, et voir grandir inopinément l'erreur sur nos différences de moyennes et le flottement de nos grandeurs d'effet? Et, sinon, en quoi consiste la différence entre la source d'erreur qu'il faut juguler dans le cas du test d'hypothèses et celle qui préside à la construction d'un intervalle de confiance?

Une seconde difficulté touchant l'application d'un contrôle alpha global à la Bonferroni concerne l'interprétation des tableaux de valeurs estimées en analyse multivariée. Citons par exemple les coefficients de régression bruts (b_j) ou standardisés (β_j) de la régression multiple et leurs sanctions de probabilité, le relevé des valeurs significatives dans un tableau d'intercorrélations, la lecture et l'interprétation des (multiples) coefficients et indices des analyses dites confirmatoires, tels l'analyse acheminatoire et les modèles en équations structurelles. Le principe Bonferroni n'a pas encore envahi tout à fait ce territoire, mais le fera-t-il bientôt?

Le calcul de puissance statistique, entrepris afin de pronostiquer la rentabilité d'une recherche, tombe aussi sous la coupe du contrôle alpha global. Comme on sait, les principaux facteurs qui affectent la puissance sont la grandeur d'effet escomptée, la taille d'échantillon et le seuil de signification α appliqué. Si, par malheur, la recherche envisagée est importante et comporte plusieurs conditions et comparaisons, la correction Bonferroni y imposera un seuil α_C ($= \alpha/k$, pour k comparaisons) plus rigoureux, la puissance s'en trouvant automatiquement amoindrie. Pour combler ce manque engendré par un contrôle soi-disant plus juste, le chercheur se trouve alors contraint de multiplier la taille d'échantillon prévue, sinon de renoncer.

Finalement, qu'en est-il de la stratégie de réplication, le troisième moment de la recherche, après l'énoncé d'une hypothèse explicative et l'obtention d'une indication favorable par un test statistique significatif. La réplication établit que la chaîne causale alléguée, démontrée une fois, est reproductible à volonté et est fiable. La réplication, obtenue par une ou plusieurs répétitions de l'expérience significative, confirme l'hypothèse et en établit une loi. C'est par ce trajet que, si et quand l'hypothèse est confirmée, la recherche progresse dans l'explication des phénomènes. C'est dans cette stratégie aussi, d'après un même raisonnement, que procèdent la méta-analyse et ses méthodes (Laroche, 2015), l'accumulation de résultats (statistiques) indépendants servant à augmenter la certitude sur un effet donné. Soit, par exemple, trois tests produisant les probabilités $p_j = 0,06, 0,07$ et $0,08$, toutes non significatives individuellement vis-à-vis d'un seuil prescrit de 0,05. Selon le principe Bonferroni et afin de garantir un contrôle alpha global à 0,05, il faudrait comparer les 3 p_j au seuil $\alpha_C = 0,05/3 \approx 0,017$ ou, en équivalence, comparer chaque quantité $3 \times p_j$ à α, par exemple $3 \times 0,06 = 0,18$ versus 0,05 : la valeur 0,06, tout juste non significative au départ, cesse alors de signaler une "tendance" et rejoint les deux autres. Or, l'agrégation de k valeurs de probabilité vers une variable khi-deux, une technique de méta-analyse attribuée à Fisher, s'obtient par khi-deux $= -2 \cdot \ln(p_1 \times p_2 \times \cdots \times p_k)$, le résultat étant distribué comme $\chi^2_{(dl=2k)}$. Pour notre exemple, nous avons khi-deux $= -2 \cdot \ln(0,06 \times 0,07 \times 0,08) \approx 16,00$, à comparer à $\chi^2_{[0,95]}(dl = 6) = 12,59$, une valeur largement significative (avec $p \approx 0,014$). Quelle que soit la technique méta-analytique employée, il est clair qu'elle va à contre-pied du principe Bonferroni, voire de la visée de la stratégie de réplication.

L'hypothèse nulle globale et son application correcte

En recherche empirique, il arrive fréquemment que le chercheur se trouve placé devant plusieurs résultats et qu'il doive se prononcer sur la crédibilité statistique de chacun, leur significativité, cela en confrontant chacun à un critère de crédibilité reconnu, le seuil α. Or, dans certains cas, les résultats de l'ensemble considéré réfèrent tous à la même hypothèse de recherche (H_1) mise au test, de sorte qu'ils sont mis en balance avec un couple d'hypothèses H_0 : H_1 unique. Chacun des tests, avec son niveau de probabilité p, contribue ainsi à apprécier la crédibilité de la même H_1. C'est ici qu'il est indiqué de désigner d'un nom particulier l'H_0 concernée, que nous appelons globale.

Prenons l'exemple d'un chercheur qui, d'emblée, met sur pied 5 expérimentations portant sur l'effet d'un traitement "T" imposé aux cobayes d'un groupe, lesquels seront

comparés à d'autres cobayes sans traitement, les témoins: chaque expérimentation est conduite par un agent de recherche différent. L'hypothèse nulle (H_0) est ici que le traitement T n'a aucun effet systématique sur les cobayes, leurs mesures reflétant une variation aléatoire typique (et normale!) chez les animaux non traités. Nous sommes ici en présence d'une H_0 globale et, dans ce cas, le contrôle alpha global est légitime, toutefois sans être nécessaire, à notre avis.

La première option qu'a le chercheur est d'effectuer séparément le test de chaque expérimentation. Sur les k situations analysées, chacune selon le seuil nominal α, il pourra enregistrer de 0 à k sanctions significatives et en tirer sa conclusion. Le cas de 0 sanction, ou d'un nombre de sanctions de l'ordre de $k \times \alpha$, le conduira sans doute à décider de la tolérance de H_0, tandis que l'obtention d'un nombre élevé de sanctions significatives le confortera dans sa confiance pour son hypothèse de recherche. Par exemple, selon un calcul binomial et sous H_0 et α_C = 0,05, si la probabilité d'obtenir au moins 1 test significatif sur 5 est de 0,226, elle tombe à 0,023 pour 2 tests significatifs et à 0,001 pour 3 tests : à lui de juger.

L'option d'un contrôle global procéderait comme suit. Au seuil α prescrit et global, parce qu'il doit couvrir l'ensemble des k tests de la recherche, le critère de probabilité Dunn-Sidàk impose un seuil individuel α_C égal à $1-(1-\alpha)^{1/k}$ qui, selon l'H_0 globale, maintient sous α la probabilité que 1 test ou plus ressorte significatif. D'après cette règle qui, rappelons-le, gouverne l'H_0 mise au test dans les k expérimentations, si au moins 1 des k tests se montre significatif, l'H_0 doit être rejetée pour les k tests, c.-à-d. pour l'ensemble de l'expérimentation, le probabilité-critère calculée concernant l'ensemble des k tests. Allant une étape plus loin, soit un seuil individuel α_C^t et, par extension de la règle de contrôle de probabilité Dunn-Sidàk, définissons[10]:

$$\alpha_C^t \text{ tel que } \alpha = 1 - \sum_{j=0}^{t-1} (1-\alpha_C^t)^{(k-j)} (\alpha_C^t)^j, 1 \leq t \leq k;$$

le seuil par la règle de Sidàk correspond à α_C^1. Le seuil individuel α_C^t assure que, sous une H_0 globale, la probabilité que t tests ou plus ressortent significatifs est maintenue sous α. Par exemple, pour $\alpha = 0,05$ et $k = 5$, nous obtenons $\alpha_C^1 = 0,01021$, $\alpha_C^2 = 0,07644$, $\alpha_C^3 = 0,18926$, $\alpha_C^4 = 0,34259$ et $\alpha_C^5 = 0,54928$. Supposons que, pour ses 5 expérimentations, notre chercheur obtient les sanctions de probabilité suivantes : $p_1 = 0,069$, $p_2 = 0,355$, $p_3 = 0,032$, $p_4 = 0,086$ et $p_5 = 0,402$. Selon l'application naïve du critère Sidàk, $\alpha_C^1 = 0,01021$, rien ne sera déclaré significatif, et le chercheur devrait abandonner son hypothèse. Cependant, selon l'application rigoureuse de la règle de probabilité, nous voyons qu'ici, les sanctions p_1, p_3 et p_4 tombent toutes sous le seuil d'ordre 3, $\alpha_C^3 = 0,18926$, un événement collectif dont la probabilité de se produire sous H_0 est inférieure au seuil $\alpha = 0,05$ imposé.[11] Moran présente un raisonnement binomial analogue pour disqualifier la règle dite de Bonferroni séquentielle (Holm, 1979; Rice, 1989). Pour notre chercheur, force lui est de rejeter l'H_0 globale, et il peut déclarer avoir démontré clairement l'efficacité probable du traitement. [12]

Une seconde interprétation légitime du rejet de H_0 en vertu d'un contrôle global à la Sidàk est offerte par Perneger (1998), qui parle alors d'une H_0 universelle, c'est-à-dire collective, en ce qu'elle regroupe les k hypothèses de l'ensemble considéré. Pour deux groupes comparés par le biais de $k = 20$ variables, je cite :

> Le taux d'erreur global concerne seulement l'hypothèse que les deux groupes sont équivalents sur chacune des 20 variables (l'hypothèse nulle universelle). Si l'une ou plus d'une des valeurs p est plus basse que 0,00256, l'hypothèse nulle universelle est rejetée. On peut alors affirmer que les deux groupes ne sont pas égaux pour les 20 variables, mais on ne peut pas dire laquelle, voire combien de variables diffèrent [d'un groupe à l'autre]. (Perneger, 1998, p. 1236)

Légitime ou non, le chercheur n'a que faire d'une telle conclusion, alors que son but est de repérer sur quelles variables les deux groupes se sont distingués, comme le conclut Perneger.

Le paralogisme du contrôle alpha global et de sa procédure d'application recommandée. Qu'en est-il d'un contrôle d'erreur global dans les situations de recherche habituelles, là où sont mis en œuvre plusieurs conditions ou traitements différents donnant lieu à plusieurs tests? Il s'agit de situations où l'hypothèse nulle est, pour chaque test, accouplée à une H_1 distincte et où

[10] Nous n'avons pas trouvé d'expression directe pour calculer la valeur de α_C^t au-delà de $t = 1$, sauf pour $t = k$, soit $\alpha_C^k = \alpha^{1/k}$.

[11] Le test d'agrégation des 5 probabilités sous le khi-deux (voir plus haut) vaut ici 21,03, versus $\chi^2_{[0,95]}(dl = 10) = 18,31$ de sorte que l'hypothèse nulle devrait être globalement rejetée pour cette expérimentation.

[12] Laurencelle (2012) développe un argument analogue en exploitant cette fois l'approche de l'"étendue studentisée" de Tukey, à savoir la statistique $q(i,j)_k = (X_{j:k} - X_{i:k})/s$, $j > i$ (voir David, 1981), la règle HSD utilisant l'étendue complète $q(1,k)_k$, avec $i = 1$, $j = k$. Selon l'hypothèse H_0, chaque fonction $q(i,j)_k$, comme $q(1,2)_5$ a sa propre distribution et ses propres valeurs extrêmes, pouvant donner lieu à un rejet raisonné de H_0. Par exemple, pour $k = 5$ et ν (degrés de liberté) $= 25$, $q(1,5)_{5[0,95]} = 4,153$ (le critère de Tukey), mais aussi $q(1,2) = 1,854$, voire $q(2,3) = 1,406$. Le test HSD de Tukey permet de rejeter l'H_0 globale si $(X_{(5:5)} - X_{(1:5)})/s \geq 4,153$ mais c'est aussi le cas pour le test $(X_{(2:5)} - X_{(1:5)})/s \geq 1,854$ ou $(X_{(3:5)} - X_{(2:5)})/s \geq 1,406$. Tukey n'est pas allé au bout de sa pensée.

elle ne peut donc être appelée globale. Pour distinguer cette espèce d'hypothèse nulle et aux fins de notre argument, nous la désignons "hypothèse nulle générale". Disons qu'un chercheur, autre que celui évoqué tantôt, entreprend une expérimentation qui met en œuvre 5 situations différentes, donnant lieu à 5 tests. Appliquons-y encore la règle naïve de Dunn-Sidàk, voire son approximation présentée sous le nom de Bonferroni[13]: garder sous α le risque de rejeter H_0 au moins une fois. Or, reprenant de nouveau notre argument en probabilité, si, au seuil global α de 0,05, l'un de ses 5 tests obtient une valeur $p_j \leq \alpha_C^1 = 0,01021$ (ou 0,01000 selon le calcul Bonferroni), l'hypothèse nulle, celle que la règle appliquée entendait contrôler, devrait être rejetée pour l'ensemble des 5 situations, *même alors que ces situations sont toutes différentes*.[14] Il s'agit là, à la fois, d'une déduction numérique rigoureuse de la règle binomiale présidant à la règle Dunn-Sidàk et d'une conclusion absurde, dont l'absurdité découle d'un quiproquo dans la structure paramétrique de la règle. En effet, dans les situations comportant k réalisations d'une même expérience (E) et donnant lieu à une H_0 globale, chacune des k sanctions p_j est, sous H_0, la réalisation d'un processus stochastique homogène (le même mécanisme aléatoire est en jeu dans les k situations, et il est confronté à la même influence systématique supposée par H_1). Les paramètres binomiaux en jeu sont uniques : ou bien (π_E, k) sous H_1, où π_E ($\pi_E < \alpha_C$) est la probabilité résultant du succès de l'expérience pour chaque test, ou bien (π_A, k) sous H_0 ($\pi_A = \alpha_C$), le niveau de probabilité admis en contexte purement aléatoire pour son insuccès. Dans ce processus homogène, un rejet de H_0 par une règle globalisante comme celle de Dunn-Sidàk, voire par toutes les règles de contrôle global, entraînerait le rejet global de H_0 pour toute l'expérience, c.-à-d. toutes ses k réalisations, puisque c'est ce que le calcul de cette règle suppose. Dans une recherche à conditions ou traitements différents, par ailleurs, chaque condition, chaque test repose sur un couple H_0:H_1 distinct, et le rejet ou la tolérance de H_0 pour une condition donnée n'a aucun impact mathématique[15] ni surtout logique sur la sanction donnée aux autres conditions. La propagation à toutes les conditions du rejet de H_0 obtenu dans l'une d'elles, une conséquence pourtant nécessaire du calcul de contrôle global, montre qu'il s'agit là d'un paralogisme dans l'application de la règle. Le contrôle alpha global, illustré ici par le règle Dunn-Sidàk, n'a de légitimité que dans les situations à conditions répliquées, celles tombant sous la coupe d'une hypothèse nulle unique et globale.

La logique de la recherche et le rôle heuristique de l'hypothèse nulle

L'univers dans lequel évolue la recherche empirique est stochastique : tous les phénomènes, notamment les phénomènes biologiques et culturels, sont soumis à d'innombrables influences parmi lesquelles le chercheur tente d'isoler un facteur ou quelques facteurs pour en déterminer l'importance relative et pour en contrôler éventuellement l'effet. Par conséquent, dans un contexte de recherche contrôlée, l'hypothèse nulle n'est jamais vraie (Rothman, 1990), celle d'après laquelle les variations observées sont entièrement attribuables à un hasard mathématique. Le chercheur, la personne intelligente, sait que n'importe quel facteur (traitement, condition, catégorie) a un impact sur ce qu'il ou elle mesure, et son but n'est pas tant de vérifier si cet impact existe, mais de déterminer s'il est assez puissant pour qu'une différence visible émerge, au-delà du brouillard aléatoire qui enveloppe le phénomène.

Spécifiquement, il vaut la peine de le rappeler, l'hypothèse nulle n'est qu'une hypothèse, sauf peut-être dans les situations de jeux de hasard où sa véracité est structurellement et mathématiquement contrôlée. Une hypothèse donc, à l'effet que seules des variations aléatoires sont à l'œuvre, de sorte que le calcul de significativité, le p, est en fait une mesure de vraisemblance, non de probabilité : la valeur p dénote le niveau de crédibilité à accorder à l'H_0, cette valeur p étant associée aux résultats d'une expérimentation et découlant de leur confrontation à un modèle idéalisé à purs effets aléatoires. Qui plus est, dans la plupart des secteurs de recherche (sciences humaines, écologie, médecine, économie), la difficulté d'isoler et d'étudier un facteur d'influence fait que la puissance statistique qui le caractérise, dénotant l'efficacité causale relative de ce facteur, est souvent faible (Perneger, 1998; Moran, 2003; Nakagawa, 2004), et l'obtention d'un effet avec $p < \alpha$ est souvent une victoire, un pas en avant dans la connaissance et qu'il serait ruineux de négliger.

Dans la démarche d'une recherche, l'obtention d'une $p < \alpha$ n'est pas le terme mais seulement un signal d'intérêt suggérant au chercheur que le facteur étudié, la variable, la relation mérite plus ample investigation. Cette relation

[13] Le mathématicien Carlo Emilio Bonferroni (1892-1960) serait sans doute le premier surpris de l'usage fait de son nom et de son œuvre.

[14] Le critère Dunn-Sidàk, rappelons-le, est basé sur l'idée que les k situations sont chacune gouvernées par le hasard, à un niveau de crédibilité de $1 - \alpha_C$, la probabilité qu'elles tombent toutes dans la zone du hasard étant alors $\alpha_G = 1 - (1 - \alpha_C)^k$, d'où le seuil $\alpha_C = 1 - (1 - \alpha_G)^{1/k}$. L'obtention d'une sanction $p \leq \alpha_C$ rejette donc le bloc H_0 couvrant les k situations, en affirmant précisément et seulement qu'il est peu probable (au seuil α_G) que les k hypothèses nulles considérées soient crédibles.

[15] Certaines comparaisons pairées donnent lieu à un lien formel de corrélation statistique, p. ex. dans les différences $A - B$, $A - C$, corrélées à $\tau = 1/2$ sous H_0, mais vraisemblablement non ou peu corrélées sous H_1.

devra être confirmée par réplication et ne sera validée que lorsqu'on l'aura intégrée dans un modèle explicatif et prédictif plus vaste du phénomène étudié. L'énoncé "$p < \alpha$" n'est pas une conclusion ni un jugement, seulement un argument soulignant le sérieux d'un résultat. La négation d'un tel résultat ou sa mise au rancart ne sont pas heuristiques mais n'ont que des effets pervers, celui par exemple de dévaluer l'hypothèse de recherche ou encore de ne pas rendre un résultat intéressant accessible à la communauté des chercheurs (Nakagawa, 2004), ou encore d'amener le chercheur à émietter ses résultats en les publiant un par un afin de les soustraire au pilon du contrôle alpha global (Perneger, 1998).

Saville (1990) rejoint d'autres critiques (Rothman, 1990; Perneger, 1998; Moran, 2003) du principe de contrôle alpha global en le rejetant complètement, par exemple en gardant le simple t de Student plutôt même que le t protégé (selon la règle LSD, Least Significant Difference) évoqué par Fisher (Howell, 1998; voir par contre Hayter, 1986) ou que toute autre règle restrictive. Chaque résultat a droit à sa chance contre le hasard, à sa valeur p et à sa confrontation à un seuil α non collectivement contraint. Concluant avec K. J. Rothman :

> La prémisse d'une hypothèse nulle universelle est celle que les sciences empiriques réfutent systématiquement. La science comporte une multitude de comparaisons, et ce simple fait ne doit pas être une cause de souci [pour le chercheur]. (Rothman, 1990, p. 46).

Épilogue : les autres approches statistiques

Suite à une expérimentation par laquelle deux ou quelques conditions de traitement ont été imposées à des cobayes (animaux ou humains), le chercheur doit rendre compte de ses résultats et déclarer d'une manière ou d'une autre les effets constatés. Le test d'hypothèses, avec sa probabilité p attachée et sa décision de significativité selon un seuil α convenu, est l'une[16] de ces manières, la plus courante, voire encore la seule technique de décision appliquée en sciences physiques[17]. L'intervalle de confiance (p. ex. pour la valeur estimée d'un paramètre ou pour la différence entre deux moyennes paramétriques) en est une autre, celle-ci basée sur un niveau de crédibilité de $1 - \alpha$. La 'technique' de la grandeur d'effet[18], popularisée par Cohen (1988), exprime la valeur de la différence observée (par rapport à la valeur stipulée sous H_0) en unités abstraites, sans indication directe de probabilité. Les techniques d'inspiration bayésienne (voir notamment l'article polémique de O'Connor, 2017) sont aussi dans la course, nez à nez avec l'intervalle de confiance. Sur toutes ces approches sauf la grandeur d'effet, voir Kendall et Stuart (1979), chapitres 17-24, notamment la 'discussion' présentée en p. 165-171.

Quel serait l'impact d'un contrôle alpha global sur ces approches, si l'on optait pour l'appliquer? Aucun impact sur la mesure de la grandeur d'effet, cette mesure étant affranchie de toute considération probabiliste. Par exemple, qu'un d de Cohen soit petit (~0,2) ou grand (~0,8) n'implique par soi-même nulle information de crédibilité statistique. Pour ajouter une telle information, il faut y incorporer la ou les tailles n associées et un modèle de probabilité, et l'on obtient alors un test d'hypothèses classique (p. ex. un t sur la différence entre deux moyennes), susceptible lui-même d'un contrôle alpha global. L'intervalle de confiance, une extension intéressante du test d'hypothèses, tombe aussi sous un contrôle possible, à travers le niveau de crédibilité $1 - \alpha'$ (où $\alpha' = \alpha/k$ selon le calcul Bonferroni, où k dénote le nombre d'intervalles concomitamment calculés), contrôle qui élargirait lesdits intervalles. Cet impact touche aussi, et pour la même raison, les intervalles bayésiens (nonobstant O'Connor, 2017[19]), lesquels ont un comportement semblable, souvent identique, à celui des intervalles classiques (Kendall & Stuart, 1979).

Remerciements

L'auteur tient à remercier Denis Cousineau pour les corrections et judicieuses suggestions ayant contribué à rendre l'exposé plus clair.

References

Benjamini, Y., & Braun, H. (2002). John tukey's contributions to multiple comparisons. *Annals of Statistics*, *30*, 1576–1594.

[16] Nous incluons dans cette catégorie les tests dits permutationnels (ou distributionnels), discutés par Fisher dès 1935 (Fisher, 1971) comme contrepartie de son principe de randomisation expérimentale et magnifiquement illustrés par Bradley (1968) et Edgington (1980), lesquels reproduisent voire forment, selon Fisher, l'assise conceptuelle et la justification mathématique de la technique des tests d'hypothèses et des intervalles de confiance afférents.

[17] D'autres techniques peuvent survenir après la décision de significativité, telle l'étude paramétrique (modélisation des relations démontrées et calibrage).

[18] La grandeur d'effet d proposée par Cohen (1988) est une variante (impropre) du paramètre de non-centralité δ d'une loi de probabilité dite non-centrale, p. ex. les lois normale, t non centrale, χ^2 et F non centrales (Johnson, Kotz, & Balakrishnan, 1994).

[19] O'Connor (2017, p. 170) allègue que l'intervalle bayésien n'est pas soumis à un contrôle probabiliste imposé par la multiplicité des tests, "parce que la distribution multidimensionnelle postérieure reste inchangée lorsqu'examinée sous différentes perspectives", contrairement à l'intervalle de confiance classique qui, lui, y serait soumis. Argument rhétorique, assertif, fondé sur une dissimilitude non indiquée et non avérée entre le contexte de l'intervalle classique (basé sur ses données propres et son modèle de probabilité) et le contexte bayésien (basé sur les mêmes intrants).

Benjamini, Y., Bretz, F., & Sarkar, S. (2004). Recent developments in multiple comparison procedures. *Monograph Series, no*, *47*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, *57*, 289–300.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs (NJ): Prentice-Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2e edition)*. Hillsdale (NJ): Lawrence Erlbaum.

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

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52–64.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, *50*, 1096–1121.

Edgington, E. S. (1980). *Randomization tests*. New York: Marcel Dekker.

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 least significant difference test. *Journal of the American Statistical Association*, *81*, 1000–1004.

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

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

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.

Hsu, J. C. (1996). *Multiple comparisons theory and methods*. London: Chapman & Hall.

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions (vol. 1 et 2) (2e édition)*. New York: Wiley.

Kendall, M., & Stuart, A. (1979). *The advanced theory of statistics. vol. 2 inference and relationship (4e edition)*. New York: McMillan.

Klockars, A. J. (1986). *Multiple comparisons*. Beverley Hills: Sage.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous myltivariate distributions (2e édition)*. Wiley: New York.

Laroche, P. (2015). *La méta-analyse : méthodes et applications en sciences sociales*. Louvain-la-Neuve: de Boeck.

Laurencelle, L. (2012). Faut-il contrôler l'erreur de type i dans le cas de comparaisons de moyennes multiples ? *The Quantitative Methods for Psychology*, *8*, 88–95.

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

Moran, M. D. (2003). Arguments for rejecting the sequential bonferroni in ecological studies. *Oikos*, *100*, 403–405.

Nakagawa, S. (2004). A farewell to bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, *15*, 1044–1045.

Neyman, J., & Pearson, E. S. (1928a). On the use and interpretation of certain test criteria for purposes of statistical inference, part I. *Biometrika*, *20*, 175–240.

Neyman, J., & Pearson, E. S. (1928b). On the use and interpretation of certain test criteria for purposes of statistical inference, part II. *Biometrika*, *20*, 263–294.

Neyman, J., & Pearson, E. S. (1936). Contributions to the theory of testing statistical hypotheses. *Statistical Research Memoirs*, *1*, 1–37.

O'Connor, B. P. (2017). A first steps guide to the transition from null hypothesis significance testing to more accurate and informative Bayesian analyses. *Canadian Journal of Behavioral Science*, *49*, 166–182.

Perneger, T. V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal*, *316*, 1236–1238.

Rice, W. R. (1989). Analyzing tables of statistical tests. *43*, 223–225.

Rohlf, F. J., & Sokal, R. R. (1994). *Statistical tables (3e edition)*. New York: Freeman.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*, 43–46.

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

Scheffé, H. (1959). *The analysis of variance*. New York: Wiley.

Scheiner, S. M. (2001). Manova. In S. M. Scheiner & J. Gurevitch (Eds.), *Design and analysis of ecological experiments (2 édition)*. Oxford: Oxford University Press.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, *73*, 751–754.

Sokal, R. R., & Rohlf, F. J. (1981). *Biometry (2e edition)*. New York: Freeman.

Tukey, J. W. (1983). The problem of multiple comparisons. In *The collected works of John W. Tukey: vol. VIII Multiple comparisons 1948-1983*. New York: Chapman and Hall.

Ury, H. K. (1976). A comparison of four procedures for multiple comparisons among means (pairwise contrasts) for arbitrary sample sizes. *Technometrics*, *18*, 89–97.

Welsch, R. E. (1977). Stepwise multiple comparison procedures. *Journal of the American Statistical Association*, *72*, 566–575.

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

Worsley, K. J. (1982). An improved Bonferroni inequality and applications. *Biometrika*, 297–302.

Citation

Laurencelle, L. (2019). Fisher contre Tukey, ou les paralogismes du contrôle alpha global pour évaluer la significativité statistique. *The Quantitative Methods for Psychology*, *15*(1), 25–37. doi:10.20982/tqmp.15.1.p025

Received: 14/08/2018 ∼ Accepted: 01/11/2018