Introduction aux tests statistiques multiples

Loic Desquilbet

Introduction aux tests statistiques multiples

Loïc Desquilbet, PhD en Santé Publique

Professeur en Biostatistique et en Epidémiologie Clinique
Département des Sciences Biologiques et Pharmaceutiques
Ecole nationale vétérinaire d'Alfort

Version 3.3 (04/10/2022)

Contrat de diffusion

Table des matières

I. Remarque préliminaire

Ce tutoriel est fortement inspiré d'un article de Bender et Lange très facile d'accès (1), intitulé « Adjusting for multiple testing—when and how? », et dont vous pouvez trouver le fichier pdf sur le site http://eve.vet-alfort.fr/course/view.php?id=353&section=3, (dans « Articles divers »). Par conséquent, une grande partie de ce document provient de cet article, en me focalisant sur les notions essentielles. Mais je vous invite à lire cet article pour aller plus loin sur cette thématique… Cependant, des articles plus « techniques » spécifiques aux essais cliniques peuvent aussi vous intéresser (2-4) !

II. Introduction

En médecine humaine et dans le cadre d'essais cliniques cherchant à prouver l'efficacité d'un médicament, l'utilisation des tests statistiques multiples est rigoureusement encadrée (5). La situation de tests statistiques multiples modifie la valeur du risque d'erreur de 1ère espèce. Dans ces situations, il est fortement recommandé d'utiliser des méthodes de « correction » de ce risque d'erreur. La méthode la plus connue est probablement celle de Bonferroni. C'est malheureusement une méthode très peu puissante (une méthode acceptant trop souvent à tort l'hypothèse nulle H_0 testée), et certains chercheurs ont même décrié cette méthode, au mieux quasiment inutile selon eux (6). Nous allons voir dans ce document qu'il existe une méthode plus puissante que celle de Bonferroni, qui ne demande aucun logiciel de statistique (Excel suffira), et qui gagnerait à être davantage connue : il s'agit de la méthode de Holm.

III. Définitions préliminaires

Je vais appeler « maladie » toute « chose » relative à l'état de santé d'un animal dont on cherche à étudier l'association avec une ou plusieurs « expositions ».

Je vais appeler « exposition » une caractéristique individuelle, extrinsèque (lieu de vie, environnement, traitement reçu, …) ou intrinsèque (âge, sexe, race, concentration en urée, …).

Dans la situation où l'on cherche à montrer dans une étude qu'une exposition est associée à une maladie, « l'hypothèse nulle » (notée H_0) du test statistique testant cette association dans un échantillon est la suivante : cette exposition et cette maladie ne sont absolument pas associées dans la population.

IV. Significativité statistique et risque d'erreur de 1ère espèce

A. Rappel sur le risque d'erreur de 1ère espèce α

La définition du risque d'erreur de 1ère espèce α est la suivante : c'est la probabilité de rejeter H_0 lorsque H_0 est vraie. En d'autres termes, si dans la population, il n'existe aucune association entre une exposition et une maladie (H_0 est alors « vraie »), il y a α% de risques d'observer dans un échantillon parfaitement tiré au sort de la population une association significative ($p \leq \alpha$) entre cette exposition et cette maladie, en l'absence de tout biais d'association. Autrement dit, le hasard *seul* fait *croire* dans α% des cas à une vraie association lorsqu'en vrai, il n'existe aucune association réelle. Attention, α n'est donc pas l'erreur que l'on commet lorsque l'on rejette H_0 (7). Le risque d'erreur α est très souvent fixé à 0,05. Dans la suite de ce document, je fixerai la valeur de α à 0,05.

B. Présentation de la problématique sur un exemple

Supposons la maladie « être infecté par le FLV » chez le chat, et 20 expositions du propriétaire qui ne sont absolument pas associées à l'infection par le FLV (par exemple, la couleur préférée du propriétaire, le nombre d'heures de sport effectuées par semaine en moyenne au collège, le fait d'avoir voyagé en Ecosse au cours de sa vie, etc.). Si l'on tire parfaitement au sort un échantillon de chats en posant des questions aux propriétaires sur ces 20 expositions, il y a $1 – (1 – 0,05)^{20} = 0,64 = 64\%$[1] de risques de trouver *au moins une* des 20 expositions significativement associée au FLV. Il a donc 64% de risques de dire de sacrées bêtises dans cette situation particulière là. Ainsi, le risque d'erreur de 1ère espèce n'est donc plus du tout de 5%, mais bien supérieur (64%), ce qui est inacceptable. On parle alors d' « inflation » du risque d'erreur de 1ère espèce, dans cet exemple passant de 0,05 à 0,64.

La formule générale est la suivante : soit *k* expositions indépendantes les unes des autres et par ailleurs non associées dans la population à une maladie, alors la probabilité d'observer (à tort) dans un échantillon une association significative ($p \leq 0,05$) entre au moins l'une de ces *k* expositions et la maladie est égale à $1 – (1 – 0,05)^k$, en prenant $\alpha=0,05$.

C. Commentaires

La formule générale ci-dessus fait l'hypothèse que les *k* expositions sont indépendantes les unes des autres. En pratique, c'est rarement vrai. Par exemple, si l'on souhaite savoir si au moins un paramètre cardiaque parmi 20 est associé à une décompensation cardiaque par la suite, il n'y a pas du tout indépendance entre ces 20 paramètres cardiaques, puisque tous les paramètres cardiaques testés sont en rapport avec … le cœur ! En effet, il y a de bonnes chances que si un paramètre cardiaque est associé à la décompensation cardiaque, un autre paramètre cardiaque le soit aussi – en tout cas plus de chances qu'un paramètre qui n'a rien à voir avec le cœur !

Dans la situation où les tests statistiques mettent en jeu *k* expositions *non* indépendantes les unes des autres, et dans la situation où aucune de ces expositions n'est associée à la maladie, la probabilité d'observer (à tort) au moins l'une de ces expositions significativement associée à la maladie est *moins* élevée que $1 – (1 – 0,05)^k$. Le risque d'erreur de 1ère espèce est donc moindre dans le cas d'expositions non indépendantes que dans le cas d'expositions indépendantes les unes des autres. (Cela dit, il reste bien supérieur à 5% si aucune correction n'est effectuée ! ☺)

V. Quand est-on dans la situation de tests statistiques multiples ?

Une étude *n'*est *pas* en situation de tests statistiques multiples si et seulement si *chaque* exposition est testée en association avec une maladie pour confirmer une hypothèse *a priori* de l'existence probable / certaine d'une telle association. Cette hypothèse peut provenir de la littérature (c'est préférable) mais aussi de votre intuition médicale (à ce moment-là, vous devrez argumenter dans l'article la raison de cette intuition, pour ne pas être « accusé » d'avoir fait les choses à l'envers : tester l'association entre la maladie étudiée et une exposition sans hypothèse *a priori*, puis interprétation *a posteriori*), ce qui correspond au *HARKing* (8). Toutes les situations où *k* expositions sont testées en association avec la maladie sans hypothèse *a priori* sont des situations de tests statistiques multiples.

[1] Démonstration : sous l'hypothèse qu'aucune des 20 caractéristiques n'est associée au FLV, P(Observer ≥ 1 association significative) = 1 – P(ne pas observer du tout d'associations significatives) = 1 – P(1ère caractéristique n'est pas associée significativement au FLV ET que la 2ème ne l'est pas non ET la troisième non plus ET etc.). Or lorsque les événements A et B sont indépendants, P(A et B) = P(A) x P(B). Or, ici, P(la caractéristique *i* n'est pas associée significativement au FLV) = 1 – α = 1 – 0,05. Donc, P(ne pas observer du tout d'associations significatives parmi les 20 testées) = $(1 – 0,05)^{20}$. CQFD

VI. Quand doit-on utiliser une méthode de correction du risque d'erreur de 1ère espèce ?

A. Etude exploratoire ou étude de confirmation ?

La toute première question à se poser est la suivante : « l'étude que je suis en train de mener est-elle une étude exploratoire ou bien une étude de confirmation ? » Pour répondre à cette question, voici la définition de ces deux types d'études.

Une étude exploratoire est une étude au sein de laquelle les tests statistiques sont réalisés sans aucune hypothèse *a priori* : on ne souhaite pas *confirmer* une hypothèse médicale que l'on avait avant de réaliser l'étude. Une étude exploratoire va *explorer* un grand nombre d'associations avec la maladie, c'est-à-dire que les investigateurs de l'étude vont tester l'association entre de nombreuses expositions et la maladie, sans avoir aucune idée de l'association qui pourrait être significative dans leur étude.

Par opposition, même si je simplifie un peu les choses, une étude de confirmation est une étude qui n'est pas une étude exploratoire.

B. Tests statistiques multiples dans une étude exploratoire

Dans le cas d'une étude exploratoire, on est clairement dans la situation de tests statistiques multiples. Le gros problème dans un tel type d'étude est que la correction du risque d'erreur de 1ère espèce est tellement compliquée (car dépendant de nombreux paramètres inconnus) qu'aucune méthode de correction n'est acceptable (lire la partie « 3. When are adjustments for multiple tests necessary? » de l'article de Bender et Lange (1) pour plus d'explications). Ainsi, en cas d'étude exploratoire, il *faut* partir du principe que dans la conclusion à l'issue des tests statistiques pour lesquels le degré de signification *p* est ≤ 0,05, le risque d'erreur de 1ère espèce α n'est plus de 0,05 (5%), mais il devient inconnu, donc potentiellement grand. Par conséquent, dans une étude exploratoire, on n'a pas les moyens d'être convaincu de l'existence d'une association réelle si elle est « significative » dans l'échantillon après avoir réalisé de multiples tests statistiques. La conclusion doit être énoncée avec toutes les précautions possibles, en écrivant que des études *confirmant* ce résultat sont indispensables.

C. Tests statistiques multiples dans une étude de confirmation

Dans une étude de confirmation, la règle est simple : si chaque test statistique testant l'association entre une exposition et une maladie est conduit parce qu'il confirme l'hypothèse selon laquelle *cette* exposition et *cette* maladie ont de bonnes raisons d'être associées, alors une correction pour tests statistiques multiples n'est *pas* utile[2]. Dans tous les autres cas de figure, et si vous n'êtes pas dans la situation d'une étude exploratoire, alors une correction pour tests statistiques multiples est *indispensable*. Un exemple classique de situation de tests statistiques multiples nécessitant une correction du risque d'erreur de 1ère espèce est le suivant : si l'hypothèse repose non pas sur *une* exposition, mais sur *une famille* d'expositions (ces expositions dépendant donc les unes des autres), dont on ne sait *a priori* pas celle(s) associée(s) à la maladie.

[2] Ce n'est donc pas le nombre de tests statistiques qui détermine la situation de tests statistiques multiples, c'est la présence ou l'absence d'une hypothèse *a priori* pour chaque exposition testée.

VII. Que doit-on faire en cas situation de tests statistiques multiples ?

A. Présentation d'un exemple d'une étude de confirmation

Prenons l'exemple suivant (fortement inspiré d'une thèse vétérinaire[3]) : la littérature a suggéré des hypothèses selon lesquelles la race, l'environnement du chien ainsi que l'éducation que le chien a reçu dans la première année de sa vie seraient déterminants dans le fait que le chien soit agressif à l'âge adulte. Pour *confirmer* ces hypothèses, une étude a été menée auprès de propriétaires de chiens au sein de laquelle les informations ci-dessous ont été collectées via un questionnaire :

- Agressivité (la maladie étudiée) : score allant de 0 (chien non agressif) à 100 (chien très agressif) ;
- Race : exposition en 3 classes (chiens de berger et de chasse ; terriers et dogues ; chiens issus d'un croisement) ;
- Environnement du chien à partir de 3 expositions binaires : repas en présence d'humain (*versus* repas seul), accès aux chambres autorisé (*versus* accès interdit), et lieu de repos réservé au chien (*versus* non réservé au chien) ;
- Education du chien dans sa première année de vie à partir de 2 expositions binaires : repas *ad libitum* (*versus* fragmenté), et retirer les jouets du chien quand il joue (*versus* ne pas les retirer).

Supposons dans le fichier de données les degrés de signification suivants, testant l'association entre chaque exposition parmi les 6 et l'agressivité du chien, en comparant puis testant les différences de médianes du score d'agressivité (à l'aide du test de Kruskal-Wallis pour la race[4], et à l'aide du test de Mann-Whitney pour les 5 autres expositions[5]) :

Expositions	Degré de signification *p*
Race	0,041
Environnement	
Repas en présence d'humains	0,002
Lieu de repos réservé	0,009
Accès aux chambres	0,038
Education	
Repas *ad libitum*	0,032
Retirer les jouets	0,543

Une première stratégie consiste à considérer les 6 expositions comme indépendantes les unes des autres. Soit S_6 le nom de cette stratégie. Une seconde stratégie (S_3) consiste à créer 3 familles de tests statistiques, puisque la littérature propose trois hypothèses : la famille « race », la famille « environnement », et la famille « éducation ». Il y aura un seul test statistique dans la famille « race » (car une seule exposition à tester), trois dans la famille « environnement », et deux dans la famille « éducation ».

[3] Thèse vétérinaire de Sara Hoummady soutenue en 2013 à l'Ecole nationale vétérinaire d'Alfort, intitulée « Facteurs environnementaux et agressivité chez le chien », dirigée par la Pr Caroline Gilbert.

[4] La race est une variable en 3 classes, donc 3 médianes de scores (une par race) ont été testées avec le test de Kruskal-Wallis pour savoir si au moins l'une des trois est significativement différente des autres.

[5] 5 expositions binaires avec une maladie (le score d'agressivité) quantitative, donc 5 comparaisons de deux médianes à l'aide du test de Mann-Whitney.

B. Méthodes de correction du risque d'erreur de 1ère espèce

Il existe de nombreuses méthodes de correction du risque d'erreur de 1ère espèce. Elles sont présentées dans l'article de Bender et Lange (1). Parmi elles, je vais me focaliser sur deux méthodes, toutes les deux très simples d'emploi, ne nécessitant pas de logiciel de statistique (pour corriger le risque d'erreur – car il faut évidemment un logiciel de statistique pour effectuer un test statistique !) : la méthode de Bonferroni et la méthode de Holm.

L'utilisation de ces méthodes permet de maintenir le risque d'erreur de 1ère espèce à sa valeur initiale (0,05) dans la situation de tests statistiques multiples.

1. *Méthode de Bonferroni*

La méthode de Bonferroni doit être utilisée lorsque l'on veut tester l'association entre *k* d'expositions, *a priori* indépendantes les unes des autres, et la maladie étudiée (6). En utilisant cette méthode, chacun des *k* tests statistiques est dit « significatif » si le degré de signification du test statistique est inférieur à α/k. Si les expositions ne sont pas indépendantes, la méthode de Bonferroni est trop conservatrice (ou trop peu puissante), c'est-à-dire qu'elle va conduire moins souvent que cela ne devrait l'être au rejet de H_0.

Prenons l'exemple de l'étude sur l'agressivité du chien. Dans la stratégie S_6, puisque l'on effectue 6 tests statistiques, *k* est donc égal à 6, et le nouveau seuil de significativité est de 0,05/6, soit 0,008. Ainsi, seule l'exposition « repas en présence d'humains » était significativement associée à l'agressivité du chien, après avoir utilisé la méthode de correction de Bonferroni. Dans la stratégie S_3, les seuils de significativité sont de 0,05, 0,05/3=0,017, et 0,05/2=0,025, respectivement pour les familles « race » (*k*=1), « environnement » (*k*=3), et « éducation » (*k*=2). Dans cette stratégie S_3, les 3 expositions « race », « repas en présence d'humains », et « lieu de repos réservé » étaient significativement associées à l'agressivité du chien, après avoir utilisé la méthode de correction de Bonferroni.

Une autre façon de savoir si une association est significative en utilisant la méthode de Bonferroni consiste à corriger le degré de signification avec cette méthode. Pour cela, il suffit de multiplier le degré de signification obtenu à l'issue du test statistique par *k*, le nombre de tests statistiques multiples. Si la valeur du degré de signification corrigée (par la méthode de Bonferroni) est inférieure au risque d'erreur de 1ère espèce fixé *a priori* (0,05), alors l'association sera significative. Si cette multiplication par *k* conduit à une valeur supérieure à « 1 », on mettra la valeur de « 1 » pour le degré de signification corrigé. On obtient les valeurs ci-dessous pour l'étude de l'agressivité du chien.

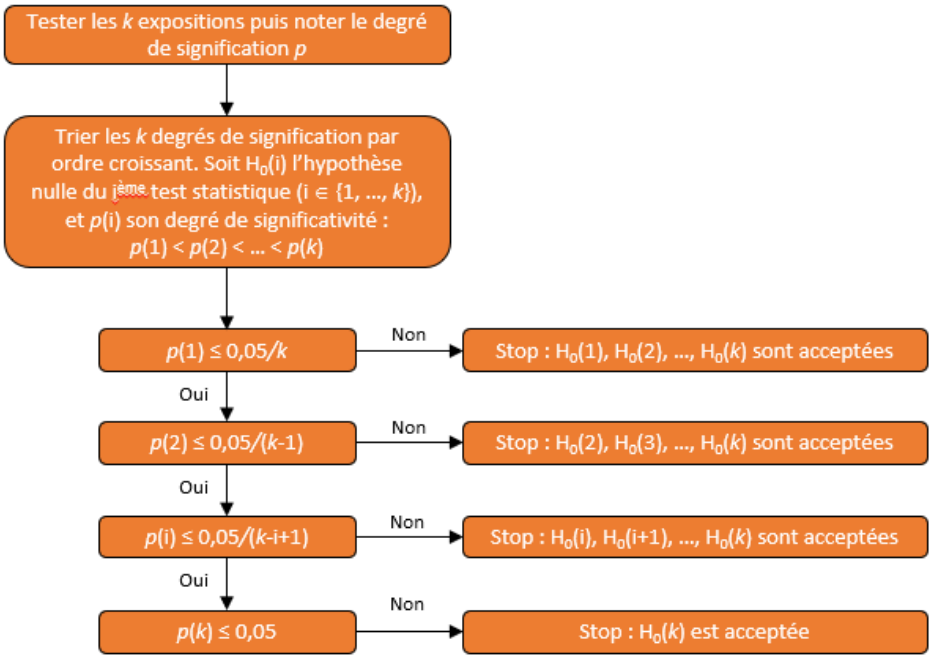
Paramètres	Degré de signification *p* avant correction	Degré de signification *p* après correction de Bonferroni	Association significative
Race	0,041	0,041	Oui
Environnement			
Repas en présence d'humains	0,002	0,002 x 3 = 0,006	Oui
Lieu de repos réservé	0,009	0,009 x 3 = 0,027	Oui
Accès aux chambres	0,038	0,038 x 3 = 0,114	Non
Education			
Repas *ad libitum*	0,032	0,032 x 2 = 0,064	Non
Retirer les jouets	0,543	0,543 x 2 > 1 → 1,000	Non

A partir des résultats du tableau ci-dessus, on obtient bien évidemment les mêmes associations significatives que précédemment, lorsque l'on comparait le degré de signification avant correction au nouveau seuil de significativité avec la stratégie S_3.

2. *Méthode de Holm*

Il existe une méthode plus puissante statistiquement que la méthode de Bonferroni (c'est-à-dire qu'elle permet plus facilement d'obtenir une association significative lorsque cette association existe réellement dans la population) : il s'agit de la méthode de Holm (9), qui respecte bien évidemment aussi le risque d'erreur de 1ère espèce après correction (10).

La méthode de Holm est très simple à mettre en place. Supposons que l'on soit dans la situation d'une étude de confirmation avec *k* expositions à tester avec une maladie. La démarche est la suivante :



Une autre façon de savoir si une association est significative avec la méthode de Holm est de calculer le degré de signification corrigé par la méthode de Holm, et de comparer cette nouvelle valeur de degré de signification à la valeur du risque d'erreur de 1ère espèce fixée *a priori* (0,05). Pour cela, un fichier Excel, disponible ici[6], permet de calculer ces degrés de signification corrigés.

Prenons l'exemple de l'étude sur l'agressivité du chien. Le tableau ci-dessous présente les degrés de signification triés par ordre croissant avant correction, et ces degrés de signification corrigés par la méthode de Holm à l'aide du fichier Excel mentionné précédemment, en utilisant la stratégie S_6.

Paramètres	Degré de signification *p* avant correction	Degré de signification *p* après correction de Holm	Association significative
Repas en présence d'humains	0,002	0,012	Oui
Lieu de repos réservé	0,009	0,045	Oui
Repas *ad libitum*	0,032	0,128	Non
Accès aux chambres	0,038	0,128	Non
Race	0,041	0,128	Non
Retirer les jouets	0,543	0,543	Non

Dans cette stratégie S_6, les deux expositions « repas en présence d'humains » et « lieu de repos réservé » étaient significativement associées à l'agressivité du chien, après utilisation de la méthode de correction de Holm. Rappelons que la méthode de Bonferroni ne conduisait qu'à une seule exposition significativement associée à l'agressivité du chien dans cette même stratégie S_6 (« repas en présence d'humains »).

[6] https://eve.vet-alfort.fr/course/view.php?id=353&section=4

Le tableau ci-dessous présente les degrés de signification triés par ordre croissant avant correction, et ces degrés de signification corrigés par la méthode de Holm à l'aide du fichier Excel mentionné précédemment, en utilisant cette fois-ci la stratégie S_3.

Paramètres	Degré de signification *p* avant correction	Degré de signification *p* après correction de Holm	Association significative
Race	0,041	0,041	Oui
Environnement			
Repas en présence d'humains	0,002	0,006	Oui
Lieu de repos réservé	0,009	0,018	Oui
Accès aux chambres	0,038	0,038	Oui
Education			
Repas *ad libitum*	0,032	0,064	Non
Retirer les jouets	0,543	0,543	Non

Dans cette stratégie S_3, les quatre expositions « race », « repas en présence d'humains », « lieu de repos réservé », et « accès aux chambres » étaient significativement associées à l'agressivité du chien, après avoir utilisé la méthode de correction de Holm. Rappelons que la méthode de Bonferroni ne conduisait qu'à trois expositions significativement associées à l'agressivité du chien dans cette stratégie S_3 (les trois premières citées ci-dessus).

C. Méthodes spécifiques de correction du risque d'erreur de 1ère espèce

1. *La comparaison de plusieurs moyennes ou médianes*

Quand on souhaite comparer deux à deux les moyennes ou les médianes entre trois groupes ou plus, il existe de nombreuses méthodes prenant en compte le fait que l'on soit en situation de tests statistiques multiples (1). Quand il n'y a que trois moyennes ou trois médianes à comparer deux à deux, une façon simple de procéder est la suivante : tester globalement si les moyennes ou les médianes sont significativement différentes les unes des autres à l'aide du test d'analyse de variance (ANOVA) ou de Kruskal-Wallis, respectivement, et si le degré de signification est ≤ 0,05, alors il est possible de tester deux à deux les moyennes ou les médianes, respectivement avec le test de Student ou de Mann-Whitney (3,11) sans utiliser de méthode de correction du risque d'erreur de 1ère espèce. Les situations impliquant plus de 3 moyennes ou médianes nécessitent des méthodes spécifiques plus compliquées, et je vous invite à lire des références citées dans l'article de Bender et Lange (1).

Dans l'étude sur l'agressivité du chien, la race est en trois classes. Dans la stratégie S_6, la race n'était pas significativement associée à l'agressivité du chien (quelle que soit la méthode, Bonferroni ou Holm). Ainsi, dans cette stratégie S_6, il n'est pas possible de comparer les races deux à deux (comparaison de deux médianes du score d'agressivité). Dans la stratégie S_3 en revanche, la race était significativement associée à l'agressivité du chien (quelle que soit la méthode, là encore). Ainsi, il est alors possible de comparer les races deux à deux pour savoir si une race avait une médiane de score d'agressivité significativement différente d'une autre race.

2. *Autres méthodes spécifiques*

D'autres situations spécifiques requièrent des méthodes de prise en compte de la situation de tests statistiques multiples telles que les analyses intermédiaires dans les essais cliniques, les analyses en sous-groupes de sujets, ou le fait qu'il y ait plusieurs critères de jugement pour évaluer, par exemple, l'efficacité d'un traitement dans une étude clinique. Toutes ces situations sont décrites dans l'article de Bender et Lange (1).

VIII. Conclusion

La toute première question qu'il faut se poser est la suivante : l'étude est-elle « exploratoire » ou bien « de confirmation ».

S'il s'agit d'une étude exploratoire, la bonne nouvelle est qu'il n'y a pas à utiliser de méthode de correction du risque d'erreur de 1ère espèce ; la mauvaise nouvelle est que vous vous trouvez dans une situation de tests statistiques multiples sans possibilité de corriger le risque d'erreur de 1ère espèce, avec pour conséquence directe une interdiction de conclure avec conviction en cas d'association significative ($p \leq 0,05$).

S'il s'agit d'une étude « de confirmation », vous devez alors vous poser la question suivante : « les expositions que je vais tester peuvent-elles se regrouper en « famille », dont il existe une ou plusieurs hypothèses selon la ou lesquelles chacune de ces familles pourrait être associée à la maladie étudiée ? » Nous avons vu en effet que la stratégie de « famille » (stratégie S_3 dans l'exemple de l'étude sur l'agressivité du chien ci-dessus) est plus puissante que la stratégie « on teste les paramètres les uns indépendamment des autres » (stratégie S_6 dans l'exemple), sans que cela ne remette en cause le risque d'erreur de 1ère espèce, une fois celui-ci corrigé par différentes méthodes telles que Bonferroni ou Holm (1).

La méthode de Bonferroni est moins puissante que la méthode de Holm (nous avons vu en effet que la méthode de Bonferroni conduisait à moins fréquemment montrer une différence significative que la méthode de Holm). Des situations spécifiques peuvent être plus efficaces que la méthode de Holm, et ont été décrites dans l'article de Bender et Lange (1). Cependant, la méthode de Holm fonctionne toujours, et reste facile d'utilisation. C'est donc cette dernière que je vous recommande.

IX. Références

1. Bender R, Lange S. Adjusting for multiple testing--when and how? Review. *J Clin Epidemiol*. 2001;54(4):343-9.
2. O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. *Biometrics*. 1979;35(3):549-56.
3. Bauer P. Multiple testing in clinical trials. Comparative Study

Meta-Analysis. *Stat Med*. 1991;10(6):871-89; discussion 889-90.

4. Dmitrienko A, D'Agostino R, Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med*. 2013;32(29):5172-218. doi:10.1002/sim.5990
5. Biostatistical methodology in clinical trials in applications for marketing authorizations for medicinal products. CPMP Working Party on Efficacy of Medicinal Products Note for Guidance III/3630/92-EN. Guideline. *Stat Med*. 1995;14(15):1659-82.
6. Perneger TV. What's wrong with Bonferroni adjustments. Research Support, Non-U.S. Gov't

Review. *BMJ*. 1998;316(7139):1236-8.

7. Desquilbet L. Enhancing Clinical Decision-Making: Challenges of making decisions on the basis of significant statistical associations. *Journal of the American Veterinary Medical Association*. 2020;256(2):187-193. doi:10.2460/javma.256.2.187
8. Kerr NL. HARKing: hypothesizing after the results are known. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc*. 1998;2(3):196-217. doi:10.1207/s15327957pspr0203_4
9. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Statist*. 1979;6(2):65-70.
10. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health*. 1996;86(5):726-8.
11. Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ*. 1996;312(7044):1472-3.