Contrat de diffusion

Pour citer (créditer) ce document

L. Desquilbet, 2019. Introduction aux tests statistiques multiples. [En ligne, disponible à : https://loicdesquilbet.wixsite.com/biostat-epidemio]

Introduction aux tests statistiques multiples

Loïc Desquilbet
Département des **S**ciences **B**iologiques et **P**harmaceutiques
Ecole **n**ationale **v**étérinaire d'**A**lfort

Version 3.1
24/03/2019

EnvA
École nationale vétérinaire d'Alfort

Table des matières

I. Remarque préliminaire

Ce tutoriel est fortement inspiré d'un article 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, section 3. 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 vous aussi 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. Certains chercheurs n'utilisent pas les méthodes de correction en situation de tests statistiques multiples, alors même qu'ils savent qu'ils sont dans cette situation, car ces derniers ne connaissent que des méthodes peu puissantes statistiquement, et donc ne sont pas enclin à les utiliser. La correction la plus connue probablement, est celle de Bonferroni. C'est malheureusement une méthode très peu puissante, et certains chercheurs ont même fortement décrié cette méthode de correction, au mieux quasiment inutile selon eux [6]. Cependant, nous allons voir dans ce document qu'il existe une méthode de correction plus efficace que celle de Bonferroni (la correction de Holm), et ne demandant aucun logiciel de statistique (Excel suffira), et qui gagnerait à être davantage connue [7] !

III. 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 *à tort* l'hypothèse nulle H_0. En d'autres termes, si dans la population, il n'existe aucune association entre une caractéristique individuelle 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 (au seuil α) entre cette caractéristique et cette maladie. Et ce, en l'absence d'aucun biais – le hasard *seul* fait *croire* dans α% des cas à une vraie association lorsqu'en vrai, il n'existe aucune association réelle. Le risque d'erreur α est très souvent fixé à 5%. Dans la suite de ce document, je fixerai α=5%.

B. Présentation de la problématique sur un exemple

Supposons la maladie « être infecté par le FLV » chez le chat, et 20 caractéristiques individuelles du propriétaire 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 au propriétaire sur ces 20 caractéristiques, il y a $1 - (1 - 0{,}05)^{20} = 0{,}64 = 64\%$[1] de risques de trouver *au moins une* des 20 caractéristiques individuelles testées significativement associées au FLV. Il a donc 64% de risques de dire de sacrées bêtises : le risque d'erreur de 1ère espèce n'est donc plus du tout de 5%, mais bien supérieur, 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 : parmi *k* paramètres indépendants les uns des autres et par ailleurs non associés dans la population à une maladie M, la probabilité d'observer (à tort) dans un échantillon au moins une association significative ($p \leq 0{,}05$) entre un paramètre et la maladie parmi *k* paramètres testés est égale à $1 - (1 - 0{,}05)^k$, en prenant $\alpha = 0{,}05$.

C. Commentaires

La formule que j'ai utilisée ci-dessus pour obtenir 0,64 fait l'hypothèse que chacun des 20 tests statistiques testant l'association avec le FLV était indépendant 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 les 20 tests statistiques qui seront effectués, 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, 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* paramètres *non* indépendants, et dans la situation où aucun de ces paramètres n'est associé à la maladie, la probabilité d'observer (à tort) au moins un de ces paramètres significativement associé à la maladie est *moins* élevée que $1 - (1 - 0{,}05)^k$. L'erreur commise en rejetant H_0 est donc moindre dans le cas de paramètres non indépendants que dans le cas de paramètres indépendants les uns des autres. (Cela dit, elle reste bien supérieure à 5% si aucune correction n'est effectuée ! ☺)

IV. 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 paramètre est testé en association avec une « maladie » (au sens le plus général du terme) 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 d'une intuition médicale d'un clinicien (à ce moment-là, le clinicien devra 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 avec un paramètre sans hypothèse *a priori*, puis interprétation *a posteriori*). Toutes les situations où *k* paramètres sont testés 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) = $1 - \alpha = 1 - 0{,}05$. Donc, P(ne pas observer du tout d'associations significatives parmi les 20 testées) = $(1 - 0{,}05)^{20}$. CQFD

V. 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 nombreux paramètres et la maladie, en allant « à la pêche aux données » (« data fishing » en anglais).

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 et 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. 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 (l'association est alors « significative », et l'on rejette H_0), le rejet de H_0 est effectué avec un risque d'erreur inconnu (et non plus α=5%), donc potentiellement grand. Par conséquent, dans une étude exploratoire, il est *interdit* d'être convaincu de l'existence d'une association réelle si elle est significative dans l'échantillon. La conclusion doit être énoncée avec toutes les précautions possibles, en écrivant que des études *confirmant* ce résultat sont nécessaires.

C. Tests statistiques multiples et étude de confirmation

Dans une étude de confirmation, la règle est simple : si chaque test statistique testant l'association entre un paramètre et une « maladie » est conduit parce qu'il confirme l'hypothèse selon laquelle *ce* paramètre et *cette* maladie ont de bonnes raisons d'être associés, 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 *un* paramètre, mais sur *une famille* de paramètres (ces paramètres dépendant donc les uns des autres), dont on ne sait *a priori* pas celui (ou ceux) associé(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 paramètre testé.

VI. Que doit-on faire en cas situation de tests 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 (entre autres) ont été collectées via un questionnaire :

- Race : un paramètre en 3 classes (chiens de berger et de chasse ; terriers et dogues ; chiens issus d'un croisement) ;
- Environnement du chien (3 paramètres 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 (3 paramètres binaires) : repas *ad libitum* (*versus* fragmenté), réponse verbale si le chien quémande (*versus* pas de réponse verbale), retirer les jouets du chien quand il joue (*versus* ne pas les retirer) ;
- Agressivité : de nombreuses questions dans le questionnaire permettent de créer un score allant de 0 (chien non agressif) à 100 (chien très agressif).

Supposons dans le fichier de données les degrés de signification suivants, testant l'association entre chaque paramètre parmi les 7 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 6 autres paramètres[5]) :

Paramètres	Degré de signification *p*
Race	0,041
Environnement	
Repas en présence d'humains	0,002
Accès aux chambres	0,078
Lieu de repos réservé	0,009
Education	
Repas *ad libitum*	0,022
Réponse verbale	0,006
Retirer les jouets	0,243

Une première stratégie consiste à considérer les 7 paramètres comme indépendants les uns des autres. Soit S_7 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 un seul paramètre à tester), et trois tests statistiques dans les deux autres familles.

[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 le Dr 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] 6 paramètres binaires avec une « maladie » (le score d'agressivité) quantitative, donc 6 comparaisons de deux médianes à l'aide du test de Mann-Whitney.

B. Méthodes générales 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 [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 correction de Bonferroni et la méthode de Holm.

1. *Correction de Bonferroni*

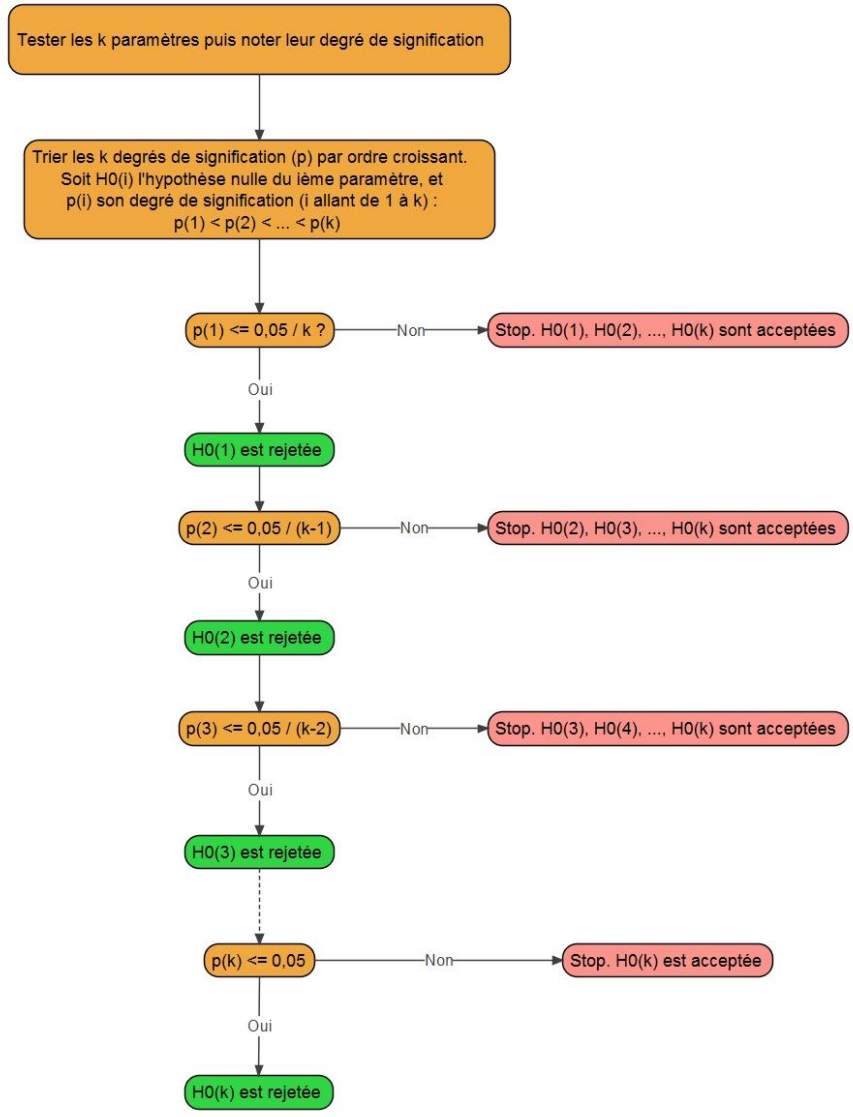
La correction de Bonferroni est une correction qui doit être utilisée quand on veut être sûr (au risque d'erreur α=5%) que H_0 est fausse en testant une multitude (*k*) de paramètres, *a priori* indépendants les uns des autres [6]. En d'autres termes, on utilise la correction de Bonferroni quand on veut conclure avec force qu'il existe au moins un paramètre parmi *k* qui soit associé à une maladie. La formule est simple : chaque test statistique est dit « significatif » si le degré de signification de chacun des *k* tests statistiques est inférieur à α/k. Si les paramètres ne sont pas indépendants, la méthode de Bonferroni est trop conservative, c'est-à-dire qu'elle va conduire moins souvent que cela devrait l'être au rejet de H_0. (Vous pouvez relire la partie « III.C Commentaires » ci-dessus.)

Prenons l'exemple sur l'agressivité du chien. Dans la stratégie S_7, puisque l'on effectue 7 tests statistiques, le seuil de signification est de 0,05/7, soit 0,007. Ainsi, les 2 paramètres « repas en présence d'humains » et « réponse verbale » sont significativement associés à l'agressivité du chien, après correction de Bonferroni. Dans la stratégie S_3, les seuils de signification sont de 0,05, 0,05/3=0,017, et 0,05/3=0,017, respectivement pour les familles « race », « environnement », et « éducation ». Dans cette stratégie S_3, les 4 paramètres « race », « repas en présence d'humains », « lieu de repos réservé », et « réponse verbale » sont significativement associés à l'agressivité du chien, après correction de Bonferroni.

Il existe une méthode plus puissante statistiquement que la méthode de Bonferroni (c'est-à-dire qu'elle permet plus facilement de mettre en évidence statistiquement une association réelle) : il s'agit de la méthode de Holm [8], qui respecte bien évidemment aussi le risque d'erreur de 1ère espèce après correction [7].

2. *Correction de Holm*

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



Un fichier Excel, disponible ici[6], permet d'établir la significativité d'une association en utilisant la méthode de Holm décrite ci-dessus, après avoir effectué les *k* tests statistiques qui fourniront les degrés de signification.

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

Prenons l'exemple sur l'agressivité du chien. Le tableau ci-dessous présente les degrés de signification triés par ordre croissant, et les seuils de signification calculés selon la méthode de Holm, en utilisant la stratégie S_7.

Paramètres	Degré de signification *p*	Seuil auquel comparer le *p*
Repas en présence d'humains	0,002	0,05/7=0,007
Réponse verbale	0,006	0,05/6=0,008
Lieu de repos réservé	0,009	0,05/5=0,010
Repas *ad libitum*	0,022	0,05/4=0,013
Race	0,041	0,05/3=0,017
Accès aux chambres	0,078	0,05/2=0,025
Retirer les jouets	0,243	0,05

Dans cette stratégie, les 3 paramètres « repas en présence d'humains », « réponse verbale », et « lieu de repos réservé » sont significativement associés à l'agressivité du chien, après correction de Holm.

Le tableau ci-dessous présente les degrés de signification triés par ordre croissant, et les seuils de signification calculés selon la méthode de Holm, en utilisant la stratégie S_3.

Paramètres	Degré de signification *p*	Seuil auquel comparer le *p*
Race	0,041	0,05
Environnement		
Repas en présence d'humains	0,002	0,05/3=0,013
Lieu de repos réservé	0,009	0,05/2=0,025
Accès aux chambres	0,078	0,05
Education		
Réponse verbale	0,006	0,05/3=0,013
Repas *ad libitum*	0,022	0,05/2=0,025
Retirer les jouets	0,243	0,05

Dans cette stratégie S_3, les 5 paramètres « race », « repas en présence d'humains », « lieu de repos réservé », « réponse verbale », et « repas *ad libitum* » sont significativement associés à l'agressivité du chien, après correction de Holm.

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

1. *La comparaison de plusieurs moyennes ou médianes*

Je n'ai volontairement pas insisté sur l'association entre la race du chien et l'agressivité du chien. En effet, j'ai utilisé le test de Kruskal-Wallis pour savoir si la race du chien était significativement associée à l'agressivité du chien. Mais ce serait naïf de penser que l'on souhaiterait « seulement savoir si, globalement, la race est associée, sans rechercher si une race en particulier est plus ou moins agressive qu'une autre », n'est-ce pas ?

Da façon plus générale, 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 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, 9]. 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 [1].

Si, en revanche, vous avez des hypothèses selon lesquelles une classe est *a priori* différentes des autres (parce qu'il s'agit par exemple d'une classe de « référence »), et si vous n'êtes pas dans une situation d'essai clinique avec différentes doses à tester par rapport à un groupe placebo, alors il n'est pas irraisonnable de ne pas utiliser de méthode de correction du risque d'erreur de 1ère espèce…

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 [1].

VII. 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 mauvais 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 paramètres que je vais tester peuvent-ils se regrouper en « famille », dont il existe une ou plusieurs hypothèse selon la ou lesquelles cette famille pourrait être associée à la maladie étudiée. » Nous avons vu en effet que la stratégie de « famille » (stratégie S_3 dans l'exemple) est plus puissante que la stratégie « on teste les paramètres les uns indépendamment des autres » (stratégie S_7 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 correction de Bonferroni est moins puissante que la méthode de Holm (nous avons vu en effet que la première des deux méthodes conduisant à moins fréquemment montrer une différence significative que la seconde). Des situations spécifiques peuvent être plus efficaces que la méthode de Holm, et ont été décrites dans l'article de Bender [1]. Cependant, la correction de Holm fonctionne toujours, et reste facile d'utilisation. C'est donc cette dernière que je vous recommande.

VIII. Références

1. Bender R, Lange S. Adjusting for multiple testing--when and how? *J Clin Epidemiol* 2001,**54**:343-349.
2. O'Brien PC, Fleming TR. A Multiple Testing Procedure for Clinical Trials. *Biometrics* 1979,**35**:549-556.
3. Bauer P. Multiple testing in clinical trials. *Stat Med* 1991,**10**:871-889; discussion 889-890.
4. Dmitrienko A, D'Agostino R, Sr. Traditional multiplicity adjustment methods in clinical trials. *Stat Med* 2013,**32**:5172-5218.
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. *Stat Med* 1995,**14**:1659-1682.
6. Perneger TV. What's wrong with Bonferroni adjustments. *BMJ* 1998,**316**:1236-1238.
7. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health* 1996,**86**:726-728.
8. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Statist* 1979,**6**:65-70.
9. Altman DG, Bland JM. Comparing several groups using analysis of variance. *BMJ* 1996,**312**:1472-1473.