LES TESTS STATISTIQUES : INTERET, PRINCIPE ET INTERPRETATIONS

R. MICHEL, L. OLLIVIER-GAY, A. SPIEGEL, J-P. BOUTIN

Med Trop 2002 ; **62** : 561-563

RESUME • La comparaison de données observées sur des échantillons nécessite l'utilisation des tests statistiques. Dans cet article sont présentés, à l'aide d'un exemple pris dans la revue Médecine Tropicale, l'intérêt d'utiliser les tests statistiques, le principe général de ces tests ainsi qu'une interprétation de la valeur de la probabilité p associée aux conclusions de ces tests.

MOTS-CLES • Statistics - Data interpretation.

STATISTICAL TESTS : VALUE, PRINCIPLE, AND INTERPRETATION

ABSTRACT • Comparison of data groups requires the use of statistical tests. This article utilizes an example taken from the Revue Médecine Tropicale to illustrate the value of statistical testing as well as an interpretation of the probability value « p » associated with test results.

KEY WORDS • Statistics - Data interpretation.

Dans un article paru dans un précédent numéro de la revue Médecine Tropicale (Fargier *et Coll - Med Trop* 1999 ; **59** : 151-156), les auteurs ont comparé, sur un échantillon de 84 personnes, l'efficacité de l'artéméther et de la quinine dans le traitement du paludisme grave à *Plasmodium falcipanum*. Le critère de jugement de l'efficacité était le temps moyen de disparition de la parasitémie. Ce dernier était de 35 heures avec l'artéméther et de 42 heures avec la quinine. La différence était significative (p = 0,0004). Les auteurs ont par ailleurs étudié le temps de normalisation de la conscience. Il était de 31,2 heures sous artéméther et de 30 heures sous quinine. Cette différence était non significative (p = 0,6). Pourquoi les auteurs ont-ils utilisé un test statistique ? Que signifie la valeur de « p » ?

En nous appuyant sur les résultats de l'étude de Fargier et al. nous présentons de façon très synthétique le principe général des tests statistiques ainsi qu'une interprétation de la valeur de la probabilité « p » associée aux conclusions de ces tests.

POURQUOI UTILISER UN TEST STATISTIQUE ?

Un échantillon, même tiré au sort, n'est pas le reflet exact de la population dont il est issu. Ici, le temps moyen de disparition de la parasitémie sous artéméther était de 35 heures. Si l'on avait réalisé l'étude sur un autre échantillon issu de la même population de malades, le temps de disparition de la parasitémie aurait probablement été différent. L'écart entre le temps de disparition de la parasitémie observé sur un échantillon et sa vraie valeur dans la population est lié aux fluctuations d'échantillonnage. En raison de ces fluctuations, il est impossible de connaître la vraie valeur du temps de disparition de la parasitémie sous artéméther ou sous quinine. On ne peut donc répondre avec certitude à la question : l'artéméther est-il plus efficace que la quinine ? Néanmoins, la réponse à cette question peut être apportée par l'utilisation d'un test statistique avec un risque d'erreur connu.

L'ARTEMETHER EST-IL PLUS EFFICACE QUE LA QUININE ?

Fargier *et Coll* ont utilisé ici le test t de Student. Il existe de nombreux tests statistiques ayant chacun leurs indications et leurs contre-indications. Le choix du test est guidé par la nature des données à comparer (proportions, moyennes, etc.) et la nécessité de respecter ses conditions d'utilisation. Cependant, quel que soit le test utilisé, les étapes et l'interprétation des résultats sont les mêmes.

Le principe général de tous les tests statistiques est un raisonnement par l'absurde. En effet, pour déterminer s'il existe une différence d'efficacité entre les deux traitements, on suppose qu'il n'y a pas de différence (cette hypothèse est appelée hypothèse nulle). On calcule alors la probabilité d'observer les résultats obtenus sous cette hypothèse. Si cette probabilité est inférieure à un seuil fixé au départ (en général 5 pour 100) on rejettera l'hypothèse nulle.

• *Travail du Service de médecine des collectivités (R.M., L.O.G., Assistants des hôpitaux des armées ; A.S.et J-P.B., Professeurs agrégés du SSA), Institut de Médecine Tropicale du Service de Santé des Armées, Marseille, France* •

• *Correspondance : R. MICHEL, Service de Médecine des collectivités, Institut de Médecine Tropicale du Service de Santé des Armées, Marseille, France • Fax: +33 (0) 4 91 52 26 07 e-mail : medco.imtssa@wanadoo.fr* •

• *Article sollicité.*

La démarche adoptée dans les tests statistiques peut être résumée en quatre étapes successives. Il s'agit de (i) énoncer l'hypothèse nulle, (ii) déterminer la vraisemblance de notre observation sous cette l'hypothèse, (iii) choisir un seuil de décision et (iv) définir une règle de décision.

Enoncer l'hypothèse nulle

Pour déterminer si l'artéméther est plus efficace que la quinine dans le traitement du paludisme grave à *Plasmodium falcipanum,* on va poser comme hypothèse qu'il n'existe pas de différence d'efficacité entre ces deux traitements. Cette hypothèse appelée hypothèse nulle est notée H0 (lire H zéro). L' hypothèse alternative est notée H_1 (lire H un). Les auteurs ont réalisé un test unilatéral. L'hypothèse alternative est : l'artéméther est plus efficace que la quinine dans le traitement du paludisme grave à *Plasmodium falciparum.* Lorsque l'on veut tester l'efficacité ou la tolérance d'un nouveau médicament, on veut savoir s'il est meilleur que les traitements déjà disponibles. Il est donc intéressant d'utiliser un test unilatéral. Il était cependant possible de réaliser un test bilatéral. L' hypothèse alternative aurait été : l'efficacité de l'artéméther dans le traitement du paludisme grave à *Plasmodium falciparum* est différente de celle de la quinine.

Déterminer la vraisemblance de notre observation sous l'hypothèse nulle

La vraisemblance de notre observation sous l'hypothèse nulle est mesurée par la valeur de la probabilité « p » calculée à partir du résultat du test statistique et dont nous ne détaillerons pas ici les calculs. La probabilité « p » est appelée degré de signification. Elle indique la probabilité d'obtenir dans l'échantillon un écart d'au moins 7 heures si l'artéméther (35 heures) et la quinine (42 heures) avaient la même efficacité.

Choisir un seuil de décision

Comme toute décision fondée sur les observations d'un échantillon, et quelque soit le test utilisé, la conclusion d'un test statistique comporte un risque d'erreur.

Le risque α, encore appelé seuil de décision ou seuil de signification, est le risque de rejeter l'hypothèse H_0 alors que celle-ci est vraie. C'est ici le risque de conclure à tort que l'artéméther est plus efficace que la quinine.

Dans le domaine biomédical, on fixe habituellement la valeur du risque α à 5 pour 100 ($\alpha = 0,05$) mais il ne s'agit que d'un risque d'erreur acceptable par convention qui peut être modifié selon le type ou les objectifs de l'étude.

Définir une règle de décision

Dans notre exemple le seuil de signification α a été fixé à 5 pour 100. Si la valeur de p est inférieure à 5 pour 100, on rejettera l'hypothèse nulle. En revanche, si elle lui est supérieure ou égale, on ne rejettera pas l'hypothèse nulle.

COMMENT INTERPRETER LES RESULTATS D'UN TEST STATISTIQUE ?

Interpréter une différence significative p < 0,05

Rejeter H_0 consiste à dire que, si les deux traitements avaient la même efficacité, la probabilité d'observer, sur un échantillon, une différence au moins égale à celle observé est trop faible.

Dans notre exemple, p = 0,0004. Cela signifie que si l'artéméther et la quinine avaient la même efficacité, la probabilité d'observer, dans notre échantillon un écart de clairance d'au moins 7 heures serait de 4 sur 10 000. Cette probabilité étant inférieure au seuil de signification α fixé à 5 pour 100, on rejette H0 et l'on dit que la différence observée est statistiquement significative. Le degré de signification p est ici de 4 pour 10 000.

Comment interpréter une différence non significative p 0,05 ?

H_0 ne peut être rejetée. Nous n'avons pas mis en évidence de différence significative au risque α. La différence observée sur les données des échantillons peut être expliquée par les fluctuations d'échantillonnage.

Dans l'article de Fargier *et Coll*, le temps de normalisation de la conscience était de 31,2 heures sous artéméther et de 30 heures sous quinine. La valeur de p était de 0,6. Cette différence était donc non significative. Cela signifie que si, le temps de normalisation de la conscience était le même sous artéméther et sous quinine, la probabilité d'obtenir, du simple fait du hasard, un écart au moins aussi grand que celui observé dans l'échantillon (31,2-30 = 1,2 heures) serait de 60 chances sur 100. Cette probabilité étant supérieure au seuil de signification α fixé à 5 pour 100, la différence observée dans l'échantillon était jugée non significative.

Attention, cela ne veut pas dire qu'il n'existe pas de différence en réalité mais seulement que l'on n'a pas observé de différence !

La comparaison des données observées sur l'échantillon comporte le risque de conclure à tort que le temps de normalisation de la conscience est le même sous artéméther et sous quinine. Ce risque appelé risque β correspond à un manque de puissance du test, c'est à dire une incapacité à montrer que H_0 est fausse. La quantité (1 - β) mesure la capacité d'un test à mettre en évidence une différence lorsque celle-ci existe vraiment. Cette quantité est appelée puissance du test.

La puissance d'un test peut être comparée à celle d'une loupe : si on perçoit un signe, on peut affirmer son existence ; si on ne le perçoit pas, on ne peut pas affirmer qu'il n'existe pas, peut être serait-il perceptible avec une loupe plus puissante (D. Schwartz).

La puissance étant en partie liée au nombre de sujets, elle peut être améliorée en augmentant les effectifs d'un échantillon. L'étude a été menée sur 84 dossiers. L'inclusion d'un plus grand nombre de patients aurait permis d'augmenter la puissance du test et peut être de mettre en évidence une différence entre le temps de normalisation de la conscience sous artéméther et sous quinine.

Les fluctuations d'échantillonnage rendent impossible la comparaison de données observées sur des échantillons sans l'utilisation d'un test statistique. Ces tests comportent un risque de conclure à tort à qu'une différence n'existe pas. Ils comportent également le risque de ne pas mettre en évidence une différence qui existe en raison d'un manque de puissance. Cependant, la plupart des études étant réalisées sur des échantillons, les tests statistiques sont des outils incontournables dès lors que l'on veut comparer des données observées ■

QUELQUES LECTURES CONSEILLEES

1 - GOLDBERG M - L'épidémiologie sans peine. 2e éd. Frison Rocheed, Paris, 1998.

2 - SCHWARTZ D - Méthodes statistiques à l'usage des médecins et des biologistes. 4e éd. Médecine-Sciences Flammarion ed, Paris, 1996.

3 - BOUYER J - Méthodes statistiques. Médecine biologie. 2e éd. Les éditions INSERM ed, 1997.