Créer un outil pour les non-professionnels de la statistique : Récit d'une expérience

Stéphane HERAULT
contact@adscience.fr

I. INTRODUCTION

Docteur en Biologie-Santé, les quelques années pendant lesquelles j'ai travaillé dans le secteur de la recherche en milieux hospitalier, universitaire et pharmaceutique (à Paris et en province) m'ont donné l'occasion d'apprécier comment les médecins, chercheurs et scientifiques que j'ai côtoyés, utilisent l'outil statistique. La remarque principale est que la plupart des logiciels de statistiques, étant tous des instruments de calcul particulièrement performants et très complets, ne sont par nature accessibles qu'aux professionnels de la statistique ; ce que ne souhaitent pas nécessairement devenir les acteurs de la recherche en France. J'ai listé ci-après les principaux problèmes auxquels ces utilisateurs non-professionnels de la statistique sont confrontés.

Il faut toutefois préciser que, contrairement à bon nombre de mes collègues, j'ai pour ma part "attrapé" le goût de la statistique, grâce aux enseignements des statisticiens du Centre d'Océanologie de Marseille (Station Marine d'Endoume – Université de la Méditerranée).

II. CONSTAT

1. Le choix du test statistique

Lors de ma thèse, ce goût pour la statistique m'a valu ce qui arrive souvent dans un laboratoire de recherche : réaliser l'analyse des données des collègues chercheurs. L'essentiel de mes interventions consistait à les conseiller quant au choix du test statistique le plus approprié, puis à le réaliser et à leur expliquer les résultats. Cette étape du choix du test qui, en théorie, doit être réglée en même temps que la mise au point du protocole, n'est en pratique presque jamais fixée avant que l'utilisateur ne soit confronté au problème. Il m'est également arrivé de recevoir la visite d'un professeur de médecine dont la publication scientifique avait été rejetée par le "reviewer" du journal pour cause de "mauvaise statistique" (i.e. mauvais choix du test statistique utilisé).

Outre un retard des publications, cette habitude de faire l'analyse statistique "comme on peut" ou avec les moyens du bord, peut avoir des conséquences délétères sur la crédibilité qu'engagent le chercheur et son laboratoire vis-à-vis de la communauté scientifique. Par ailleurs, il ne s'agit pas d'une spécificité de la recherche publique puisque je l'ai également rencontrée au sein de laboratoires privés de petite taille, ne disposant pas d'un statisticien à demeure, ni des moyens nécessaires pour s'offrir les services d'une entreprise spécialisée dans l'analyse des données (type CRO : Contract Research Organization).

2. La vérification des conditions d'utilisation

Par ailleurs, dans le cas des tests paramétriques (comparaisons de moyennes, régression linéaire…), la vérification des conditions d'utilisation des tests est une étape ignorée de façon quasi systématique ; non pas par paresse intellectuelle, mais tout simplement parce que beaucoup d'utilisateurs ignorent quelles sont ces conditions (quand ils savent qu'il existe des conditions à respecter). La conséquence est que cette phase est très souvent omise, au risque d'aboutir à une conclusion erronée.

3. L'aide à l'interprétation des résultats

A la décharge de ces utilisateurs non-professionnels de la statistique, qui préfèrent légitimement consacrer leur temps à la recherche plutôt qu'à se plonger dans des ouvrages de statistique, j'ai pu constater que les outils informatiques les plus répandus ne leur facilitent pas la tâche. Une très grande majorité des logiciels de statistique sont conçus pour des professionnels de la statistique et/ou de la programmation informatique. Bien peu prennent en compte le fait que beaucoup d'utilisateurs sont demandeurs d'un accompagnement quasi pédagogique dans la réalisation de leurs tests statistiques.

Ainsi, nombreux sont ceux qui ignorent la façon d'interpréter le résultat d'un test qui se limite à "$p<0.05$". Une simple phrase, exprimant en termes littéraux ce qu'il faut conclure du test, suffirait à renseigner efficacement n'importe quel utilisateur.

En outre, cet accompagnement serait particulièrement bienvenu dans des situations où la conclusion d'un test nécessite une analyse visuelle de la part de l'utilisateur, comme c'est le cas au cours des diagnostics consécutifs à une régression linéaire.

4. Simplicité d'utilisation

Enfin, les chercheurs étant des gens pragmatiques, ils ne veulent pas perdre de temps à appréhender un nouveau logiciel (nouvelle interface, complexité des boîtes de dialogue…), à transformer leurs "data" pour pouvoir les transférer vers le logiciel de statistique et à vérifier que la transformation s'est bien opérée (ex : caractère décimal)…

Ils veulent du "clé en main" avec un outil en mesure de s'adapter à eux, plutôt que l'inverse.

III. PROPOSITIONS

Plusieurs pistes sont envisageables pour répondre aux problèmes présentés ci-dessus. En voici une qui s'efforce de résoudre un maximum de ces difficultés. Il s'agit d'une application logicielle :

- utilisable sous Excel car :
  - tableur le plus fréquemment utilisé pour stocker les données,
  - pas de transfert des données vers le logiciel de statistique (donc pas de risque de transformation),
  - simple d'utilisation et bien connu des scientifiques,

- qui leur permet d'effectuer l'analyse statistique de leurs données même s'ils ne connaissent pas ou plus suffisamment les concepts mathématiques et statistiques des tests à pratiquer,

- en les aidant, s'ils le souhaitent, à sélectionner le test *ad hoc* en fonction des conditions expérimentales de leur protocole,

- en les accompagnant dans la procédure de sélection des données à analyser,

- en leur assurant la vérification du respect des conditions d'utilisation des tests et, le cas échéant, en leur proposant un test de substitution,

- en leur apportant une aide à l'interprétation des résultats,

- en utilisant un double niveau de langage :
  - un niveau technique prenant en compte les détails du calcul statistique (pour les plus curieux),
  - un niveau plus basique pour les conclusions du test, exprimées dans un vocabulaire explicite et compréhensible par tous,

- le tout en français, pour ne pas ajouter une difficulté linguistique à la compréhension des résultats de l'analyse statistique par les chercheurs francophones.

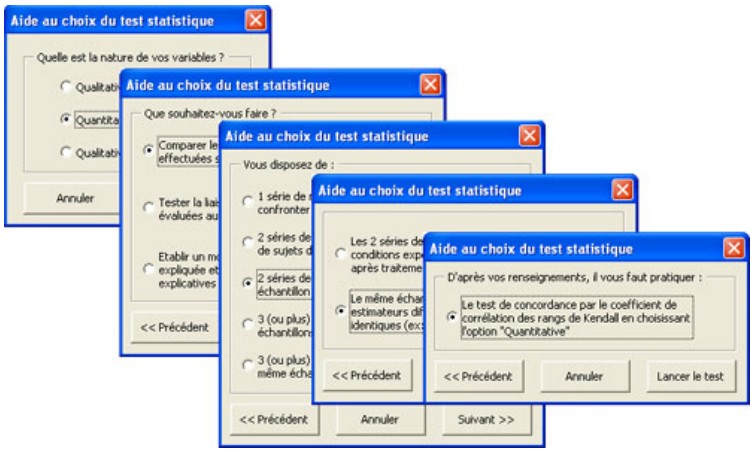
Les conséquences et avantages de l'utilisation d'un tel outil devant être :

- de permettre aux chercheurs de faire enfin eux-mêmes l'analyse de leurs résultats sans faire appel à des intermédiaires pour les tests les plus couramment utilisés et ce, en toute sécurité,
- d'accélérer le travail d'analyse des données et donc la publication des travaux de recherche,
- de réduire les risques de rejets d'articles pour simple cause de "mauvaise statistique" et donc d'assurer la crédibilité des chercheurs impliqués et de l'ensemble du laboratoire,
- de familiariser l'utilisateur avec les procédures des tests statistiques (intérêt pédagogique),
- de faciliter la tâche des personnes qui utilisent l'outil statistique régulièrement.

IV. CONCRETEMENT

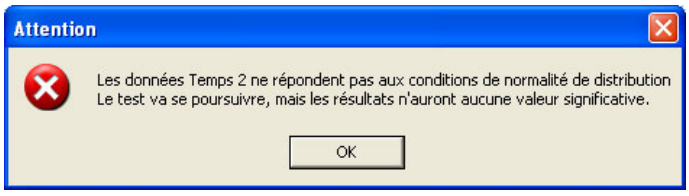
1. Aide au choix du test statistique

Une série de questions, destinées à préciser la nature du protocole expérimental, peut être posée par une succession de boîtes de dialogues. Chaque question est à réponse unique et une réponse à l'étape N modifie la question de l'étape N+1. Au terme de cette procédure, l'utilisateur obtient le nom du test statistique le plus adapté à ce qu'il veut étudier en fonction de ses données.



2. Vérification des conditions d'utilisation

En fonction de la nature du test, certaines conditions d'utilisation sont à respecter afin d'assurer la crédibilité de la conclusion du test. Ainsi, le logiciel vérifie ces conditions à la place de l'utilisateur. Le non-respect d'une des conditions d'utilisation du test est précisé à l'utilisateur par un message d'alerte, mais sans toutefois bloquer la réalisation du test en question.



Enfin, sur la feuille des résultats – laquelle est nécessairement une nouvelle feuille qui s'insère automatiquement dans le fichier Excel – l'information du respect ou du non-respect des conditions d'utilisation est rappelée à l'utilisateur. En cas de non-respect, un test de substitution lui est proposé, à charge pour l'utilisateur de pousser plus loin l'étude ou non.

Conclusions du test de comparaison des moyennes de 2 séries appariées :
Les conditions ne sont pas respectées pour pratiquer le test (distributions normales, homogénéité des variances)
Les résultats sont présentés à titre informatif mais n'ont aucune valeur significative
Il est conseillé de procéder à un test non paramétrique de Wilcoxon pour comparer la moyenne de ces 2 groupes
La valeur du t permet de rejeter l'hypothèse H0
Les moyennes des 2 séries de mesures sur le même groupe sont significativement différentes
Le seuil de significativité du test (ou risque d'erreur) est p < 4,54188119222202E-03

3. Aide à l'interprétation des résultats

a. Soutien pédagogique

Afin de remplir sa vocation pédagogique, le logiciel ne se contente pas de fournir uniquement le résultat numérique du calcul du test. Il propose systématiquement et quel que soit le test choisi :

- les informations de statistiques descriptives élémentaires,
- les hypothèses de travail (H0 et H1),
- le type de calcul utilisé,
- les résultats numériques des calculs intermédiaires et finaux,
- une conclusion littérale complétant le résultat numérique,
- une mise en avant des éléments essentiels (par une police de couleur),
- le cas échéant, une illustration graphique (modifiable à volonté car élaborée de façon automatique avec les outils Excel) :

Résultats	Groupe A	Groupe B
Nb sujets	26	23
Moyenne	53,55	66,17391
Ecart-Type	4,795557	6,88643
Médiane	55,5	65
Quartile 25	56	63
Quartile 75	63	68,5
Inter Quart	7	5,5

Hypothèses de travail :

H0 : les moyennes des 2 échantillons étudiés ne sont pas différentes (ou les 2 échantillons sont issus d'une même population).

H1 : les moyennes des 2 échantillons étudiés sont différentes (ou les 2 échantillons sont issus de 2 populations différentes).

Résultats de la comparaison de 2 moyennes :

Les effectifs sont insuffisants pour baser le test sur le calcul de l'écart-réduit, on utilisera la valeur de t.

t :	3,6056458
t limite :	2,0195409
p <	0,0008364

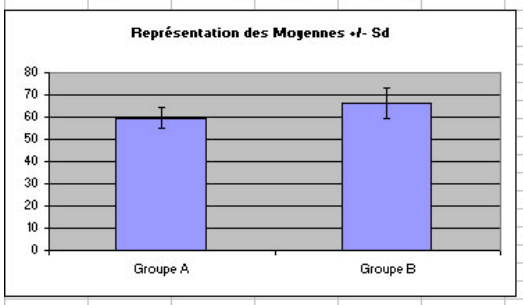
Conclusions du test de comparaison de 2 moyennes :

Les conditions d'applications du test sont respectées (distributions normales, homogénéité des variances)

La valeur du t permet de rejeter l'hypothèse H0

Les moyennes des 2 groupes sont significativement différentes

Le seuil de significativité du test (ou risque d'erreur) est p < 8,36383717531137E-04



Basculer l'affichage en Médiane +/- Quartiles

☐ Afficher les valeurs sur le graphique

b. Cas particulier des diagnostics consécutifs à une régression linéaire

Outre les calculs classiques nécessaires à une étude de régression linéaire (coefficients, significativité, intervalle de confiance à 95%, coefficients de détermination et de détermination ajusté), le logiciel accompagne l'utilisateur en lui détaillant la procédure d'analyse de la qualité du modèle.
Celui-ci lui précise comment il doit visuellement considérer les résultats de la régression (étude des résidus, des leviers) et les remèdes à apporter en cas de divergence avec les hypothèses du modèle.

Taille (X)	Performance (Y)	Y calculés	Résidus	Rés. Studentisés	Leviers	IC Sup(95%)	IC Inf(95%)
1,73	2,32	2,32	-0,004	-0,134	0,224	2,36	2,29
1,73	2,31	2,32	-0,014	-0,491	0,224	2,36	2,29
1,78	2,33	2,34	-0,011	-0,365	0,121	2,36	2,32
1,83	2,4	2,36	0,042	1,367	0,063	2,37	2,34
1,84	2,4	2,36	0,039	1,251	0,057	2,38	2,35
1,84	2,4	2,36	0,039	1,251	0,057	2,38	2,35
1,84	2,37	2,36	0,009	0,279	0,057	2,38	2,35
1,85	2,37	2,36	0,005	0,167	0,053	2,38	2,35
1,85	2,37	2,36	0,005	0,167	0,053	2,38	2,35
1,85	2,36	2,36	-0,005	-0,156			2,35
1,85	2,28	2,36	-0,085	-2,745			2,35
1,86	2,37	2,37	0,002	0,056			2,35
1,87	2,36	2,37	-0,012	-0,377			2,36
1,88	2,36	2,38	-0,015	-0,488			2,36
1,91	2,41	2,39	0,025	0,802	0,066	2,40	2,37
1,94	2,39	2,40	-0,006	-0,188	0,097	2,42	2,37
1,94	2,35	2,40	-0,046	-1,513	0,097	2,42	2,37
1,96	2,45	2,40	0,047	1,601	0,127	2,43	2,38
2	2,42	2,42	0,004	0,134	0,209	2,45	2,39
2,01	2,4	2,42	-0,020	-0,706	0,234	2,45	2,39

Résidu studentisé élevé (|Résidu|>2)

Etude de la qualité du modèle de régression :

Un modèle de régression linéaire est basé sur les hypothèses suivantes à vérifier :

1) Linéarité du modèle

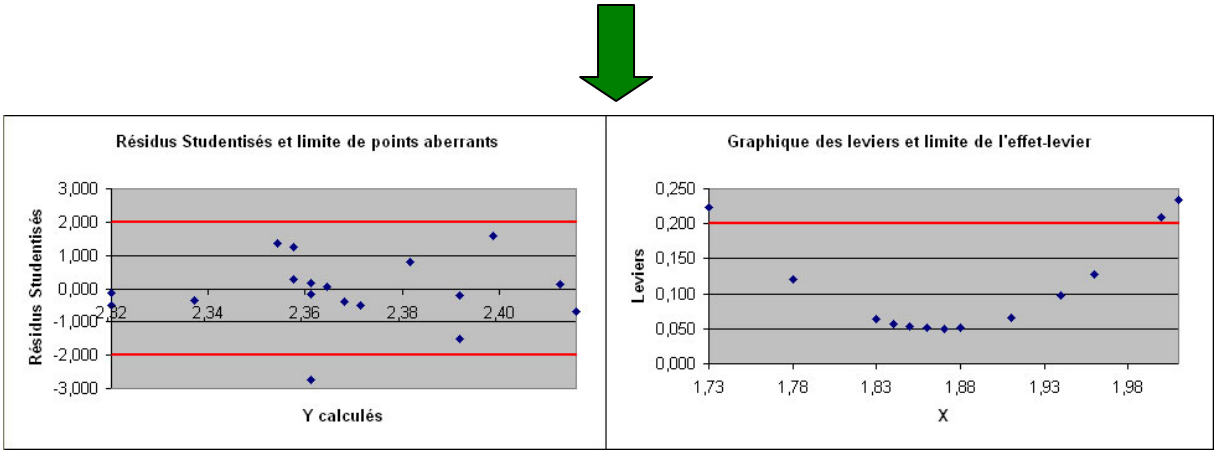
a) Observez le graphique des résidus en fonction des Y calculés.
Si le modèle choisi est adéquat, les résidus sont répartis uniformément sans représenter une forme particulière.
Sinon, il vous faut procéder à une transformation de vos données (ln, exp, ², ...) et renouveler le test ou ajouter une variable dans le modèle et procéder à une analyse de régression linéaire multiple.
b) L'étude des résidus studentisés montre que les couples d'observation suivants présentent des résidus qui s'écartent de 2 : (1,85;2,28),
Cela confirme la présence de point(s) aberrant(s) en Y. (cf. graphique des Résidus studentisés en fonction des Y calculés)
Les observations en question sont à examiner.
c) Le test des signes montre que l'hypothèse H0:'la répartition des observations est linéaire' n'est pas rejetée.
La liaison entre les variables X et Y semble donc bien être de nature linéaire.

2) Homoscédasticité (la variance des résidus doit être constante)

a) Observez le graphique des résidus en fonction des X.
Si le modèle choisi est adéquat, les résidus sont répartis uniformément sans représenter une forme particulière.
Sinon, il vous faut procéder à une transformation de vos données (ln, exp, ², ...) et renouveler le test ou ajouter une variable dans le modèle et procéder à une analyse de régression linéaire multiple.
b) L'étude de l'influence des observations sur le modèle montre que les couples d'observation suivants présentent un levier très important (> 4/N): (1,73;2,32), (1,73;2,31), (2;2,42), (2,01;2,4),
Ceci montre que les couples d'observation sus-cités jouent un rôle susceptible d'influencer le modèle linéaire. (cf. graphique des Leviers en fonction des X)
Les observations en question sont à examiner.

3) Normalité des résidus :

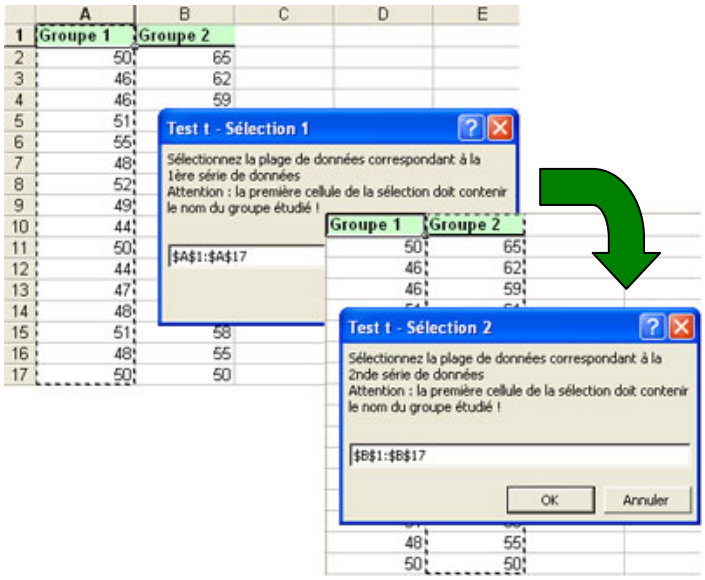
a) L'application du test de normalité sur les résidus montre que leur distribution semble normale.
b) Observez le graphique Q plot des résidus, en cas de normalité des résidus, ceux-ci doivent être à peu près alignés.



4. Simplicité d'utilisation

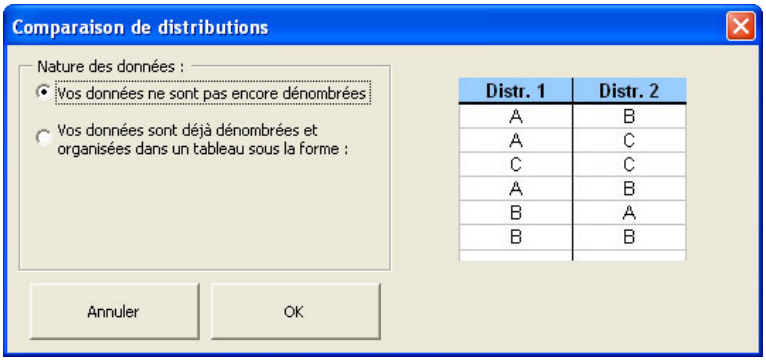
a. Aide à la sélection des données

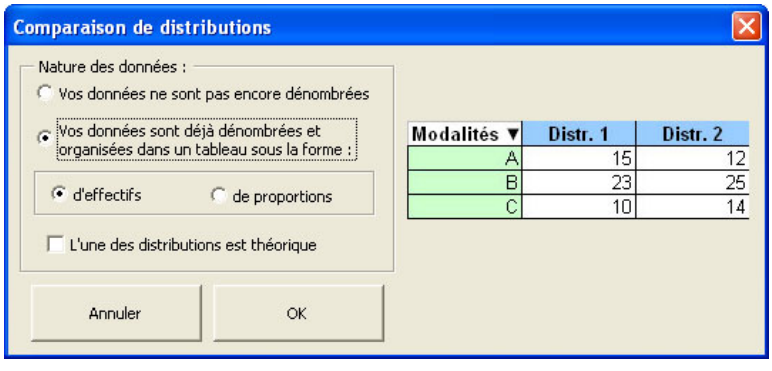
Afin d'accompagner l'utilisateur dans la sélection des données nécessaires à la réalisation du test choisi, celle-ci se fait de façon progressive. En outre, le logiciel vérifie "en temps réel" la nature et le nombre des données afin d'assurer l'utilisateur qu'il a "rentré les bonnes données dans les bonnes cases".

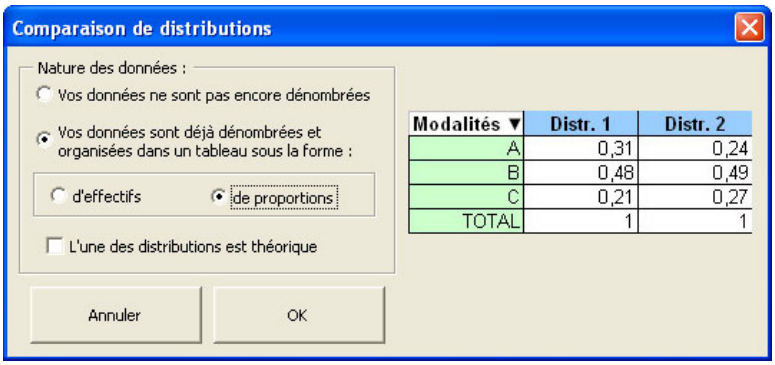


b. Adaptabilité aux données de l'utilisateur

Dans la mesure du possible, le logiciel est en mesure de procéder à l'analyse d'un même jeu de données, quelle qu'en soit la forme de présentation (ex : données dénombrées ou données brutes) afin d'éviter à l'utilisateur d'avoir à réorganiser tout son fichier Excel.







V. CONCLUSION

La finalité d'un tel outil n'est bien évidemment pas de proposer des applications statistiques nouvelles mais plutôt d'apporter un autre type de service à l'utilisateur.

Nous avons essayé de mettre en œuvre ces principes dans un logiciel en cours de développement, dénommé StatEL, et qui présente actuellement 3 modules :

- StatEL_Base pour les tests statistiques généralistes (étude de normalité, comparaison de moyennes, régressions linéaires, calcul de N, Chi²…),
- StatEL_Med adapté pour les tests et analyses de données biomédicales (Sensibilité/Spécificité, ROC, études de survie…),
- StatEL_AD destiné à l'analyse factorielle des données (ACP, AFC, Classification, AFD).

Cette gamme est bien sur appelée à s'enrichir, afin de proposer d'autres tests moins courants que ceux déjà développés, mais toujours avec la volonté de se rendre accessible et compréhensible pour l'utilisateur non-professionnel de la statistique, tout en insistant sur le versant pédagogique.

Les personnes intéressées par la version "Démonstration", et/ou souhaitant apporter leurs remarques, conseils ou recommandations sur ce logiciel, sont invitées à écrire à contact@adscience.fr